Asymptotic behavior of pre-test procedures

Pieta Boon
Faculty of Mathematical Sciences
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

ISBN 90-36513685

ASYMPTOTIC BEHAVIOR OF PRE-TEST PROCEDURES

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F. A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 17 december 1999 te 13.15 uur

door

Petronella Christine Boon

geboren op 2 december 1972
te Hengelo (O)

Dit proefschrift is goedgekeurd door de promotor
prof. dr. W. Albers,

en de assistent-promotor
dr. W. C. M. Kallenberg.

Voorwoord

Het voor u liggende proefschrift vormt het resultaat van vier jaar onderzoek, waar niet alleen ik, maar zeker ook vele anderen direct of indirect bij betrokken waren. Hen wil ik graag bedanken.

Daarbij wil ik beginnen met Wim Albers en Wilbert Kallenberg, die behalve inhoudelijk, ook op andere manieren veel bijgedragen hebben. Altijd waren zij beschikbaar voor vragen en stonden ze klaar om stukken tekst te corrigeren. Met hun steeds positieve instelling hebben ze mij zeker in de laatste fase van mijn AiO-periode enorm geholpen.

Graag wil ik de commissieleden bedanken, met name Willem Schaafsma voor zijn enthousiasme en zijn opbouwende kritiek, en Chris Klaassen en Erik van Doorn voor hun nuttige opmerkingen naar aanleiding van het concept.

Verder een woord van dank voor Gjerrit Meinsma, die aanbood het concept door te nemen en daarin vele kleine verbeteringen in het Engels en in de opmaak met Latex heeft aangebracht.

Alle collega's wil ik graag bedanken voor hun gezelligheid. Daarbij wil ik met name Gerda Arts noemen, met wie ik drie jaar lang een kamer heb gedeeld, waar het na haar vertrek toch een stuk stiller werd.

Van onschatbare waarde was de steun van Wilbert ('mijn' Wilbert) gedurende de afgelopen jaren, waarin hij mij met veel liefde en geduld heeft bijgestaan. Hij stond me met raad en daad terzijde, o.a. bij het gebruik van Latex en Matlab, waardoor ik tijd kon besparen, maar hem de tijd soms door de vingers glipte. Maar vooral op de momenten waarop ik zelf even niet meer wist hoe het verder moest, en in de laatste drukke weken voordat het concept klaar moest zijn, was zijn steun onmisbaar. Ook mijn familie en vrienden wil ik bedanken voor hun steun en medeleven, vooral mijn ouders op wie ik altijd kon rekenen.

Pieta Boon *Hengelo (O), november 1999*

Contents

Chapter 1

Introduction

1.1 Pre-test procedures

This thesis discusses pre-test procedures, which may be described as follows. Suppose we want to test a given hypothesis, but we doubt whether we may assume a restricted, but possibly incorrect model, or have to resort to a larger and thus less precise model. To settle this issue, we perform a preliminary test on the adequacy of the restricted model. If this test fails to reject, then we feel free to stick to the restricted model and use a test specifically suitable for that model. Otherwise, we use an alternative main test which is more general and appropriate for the large model, but less powerful than the first test when the restricted model holds. The schematic picture below illustrates how the procedure works.

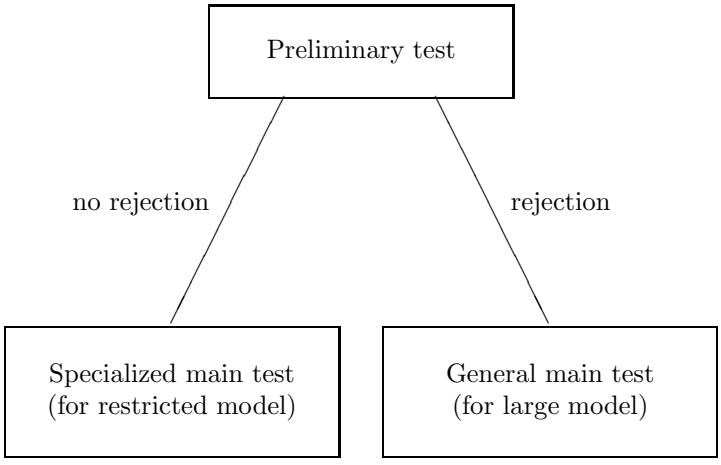

Figure 1.1.1 Pre-test procedure

The underlying idea of the above procedure is indeed appealing: if the restricted model is incorrect, the specialized test is not valid and should not be used. But always using the general and typically less powerful test would be a waste of power if the restricted model is correct after all. The specialized main test might also be preferred because it is simpler to use or explain. As it is not known beforehand whether the restricted model is applicable, it seems very natural to settle this simply through a preliminary test.

But thinking further about this attractive approach, some doubts may arise. Like any other test, the preliminary test, which should help us to decide which of the two main tests is most appropriate to test our main hypothesis, will make errors of first and second kind. This may lead to application of the specialized main test when in fact the restricted model is inadequate, or to application of the general main test when it is not necessary.

As a consequence, the behavior of the combined procedure, which we call a *pre-test procedure*, may be disappointing. Moreover, repeated use of the same data for the subsequent tests may introduce correlations which influence the behavior of the procedure. Hence, the idea of getting higher power if possible, while being protected against violation of the level if necessary, may well be too optimistic.

After these considerations the questions arise what is known about these drawbacks and how people deal with them in practice. The answers seem to be rather simple: little is known, and most people ignore them. Still, almost everyone uses such procedures, although many people do not even realize it. People who do, in most cases assume that these procedures are good and that the possible problems sketched above, will not be severe. However, in the articles from statistical literature in which attempts were made to attack the problem, things turned out to be very complicated. This makes it impossible to confirm the implicit optimism or to give guidelines for responsible use of such procedures.

One might argue that it would be better to advise to refrain from pre-testing. However, this would not be very helpful, since pre-test procedures are widely used in practice and their use only grows in view of the steadily increasing computing facilities, which make such an approach easier and easier. Hence, it does not seem likely that the problem will vanish by ignoring it. It will rather increase and therefore deserves attention.

In Section 1.2 we give an impression of the widespread use of pre-testing, not only when the final inference concerns a testing problem, as in the pre-test procedures we defined, but also when the main issue is an estimation problem. To this end, we present several recognizable examples from literature. Next, in Section 1.3 we put pre-test procedures in a wider perspective by discussing some related approaches, which have the common feature that the data are repeatedly used, first to determine better procedures for final inference based on the data under consideration, secondly for the final inference itself. Finally, in Section 1.4 we give an outline of the material covered in this thesis.

1.2 Examples from literature

The starting point for the statistical analysis of specific data from an investigation is usually a statistical model. This model may be specified before the data are collected, based on theoretical knowledge or previous experience with similar data. This is what is called an 'unconditionally specified model' in the bibliographies by Bancroft and Han (1977) and Han, Rao, and Ravichandran (1988), see also Bancroft and Han (1980). In contrast to this and of particular interest to us, is a 'conditionally specified model', or an 'incompletely specified model involving preliminary tests' as it was previously called, see Bancroft (1964). In this case the collected data are used to carry out preliminary tests to aid in the choice of a reasonable model. Subsequently, the same data are used for inferences regarding the parameter(s) of interest in the determined model.

The final inference following the preliminary test mostly concerns an estimation or testing problem. In the case of an estimator based on the outcome of a preliminary test, the term 'testimator' is used. In literature 'testimation' problems turn up more often than 'testitesting' problems. Indeed, the study of testing problems is more complicated than that of estimation problems. The quality of estimators is usually measured by the mean squared error (MSE), allowing for some bias in exchange for a smaller variance. On the other hand, in testing problems, people do prefer a high power, but they still tend to adhere to the principal requirement that the size of the procedure should not exceed the nominal level. However, it is unrealistic to require this in pre-test procedures because the possibility that the preliminary test wrongly 'accepts', cannot be excluded. This leads to application of an inappropriate test and consequently to violation of the level. Hence, in the pre-test testing problem a small exceedance of the nominal level should be allowed. Nevertheless, the level should still be controlled and the power maximized, while in the pre-test estimation problem bias and variance are combined in the MSE, in which they are exchangeable. Also from a technical point of view, testing problems are more complicated than (point) estimation problems. While the analysis of estimation problems concentrates on the first two moments of the final estimator, for the evaluation of size and power of testing procedures that are based on a preliminary test, the whole distribution is needed. This also makes the study of interval estimation after a preliminary test (considered by a few authors, for instance Arabatzis, Gregoire, and Reynolds Jr. (1989) and Giles and Srivastava (1993)) more difficult than point estimation. Arabatzis, Gregoire, and Reynolds Jr. (1989) study the interval estimation of a mean from a sample of a normal population. If the preliminary test rejects the null hypothesis that $\mu = \mu_0$, then the mean μ is estimated by $\bar{X}$ and a confidence interval is calculated, otherwise μ_0 is taken. They study the actual coverage probability of a nominal or unconditional interval after rejection by the preliminary test. The results are not uniform and the authors do not advise pre-testing.

Many examples from literature involve the question of whether two or more samples come from the same underlying model or can be pooled to make inference about a parameter occurring in the population distribution of one of the samples.

A very famous example is the situation where we want to test whether the means of two independent normal samples are equal, but where there is some uncertainty about the validity of the usual assumption of equal variances. Then sometimes a preliminary F-test is recommended to decide whether the ordinary two-sample t-test with the pooled variance estimator may be used, or whether another method for testing equality of means under heterogeneous variances should be used (the well-known Behrens-Fisher problem, for references see e.g. Nikulin and Voinov (1995)). In popular statistical packages like SAS and SPSS, preliminary variance tests have been incorporated into procedures for the two-sample means tests. However, Moser and Stevens (1992) conclude that the practice of preliminary variance tests is not appropriate and that instead of the emphasis on equality of variances, more effort should be spent on the qualities of an alternative Welch-Satterthwaite test, when teaching the two-sample means test. However, most textbooks still treat the two-sample t-test and recommend the F-test to check the assumption of equal variances. The following quotation, from a well-known textbook in statistics, illustrates this (Hoel (1984), p. 296, p. 300):

> "It will be recalled that it was necessary to assume that $\sigma_x = \sigma_y$ in order to apply the t distribution to testing the difference between two means. For the purpose of checking on this assumption, a density function that can be used for testing the equality of two variances will be derived. [...] The sample value of $F = 2.63$ is therefore not significant. This result implies that the assumption of equal variances is a reasonable one and that the significant value of t obtained in connection with this problem when testing the hypothesis $\mu_x = \mu_y$ may not be reasonably attributed to a lack of the assumption $\sigma_x = \sigma_y$ being satisfied."

Other references in which the effect of variance pre-testing before comparing two normal means was studied, are Gurland and McCullough (1962), Wehrhahn and Ogawa (1978), Markowski and Markowski (1990), Moser, Stevens, and Watts (1989). A short remark might be made here. Fisher (1990) criticized the previous point of view (see p. 125 of 'Statistical Methods for Research Workers'), as he preferred to consider the equality of variances as a part of the hypothesis to be tested, namely that the two samples are drawn from the same normal population, rather than as an assumption. However, he admits that one may also be interested in the question of whether the samples have been drawn from a (possibly different) normal population with the same mean, as we are.

Another example is the following. Consider two normal samples with the same variance but possibly different means μ_1 and μ_2. It may be suspected that $\mu_1 = \mu_2$, but one is mainly interested in μ_1. If the hypothesis of equality of the means cannot be rejected by a preliminary test, then the two samples are often pooled to get a better estimate of μ_1 or to get a better test for testing a given hypothesis concerning μ_1. If the preliminary test rejects, only the first sample is used. Some numerical work for this situation was done by Arnold (1970).

Numerous variations on these examples are possible, for example extensions to more than two samples, pooling problems for variances instead of means (Mehta and

Gurland (1969a)), or situations with uncertainty about the independence of the two samples. In the latter case the preliminary test is used to test whether the correlation equals zero (Mehta and Gurland (1969b)).

The problem of pooling is also frequently discussed in an analysis-of-variance (ANOVA) context. Bancroft and Han (1980) give an instructive overview with references of the use of preliminary tests in ANOVA models. They consider fixed, random and mixed models separately, and discuss results for the cases where either estimation or testing is the final inference. Special ANOVA models are considered by e.g. Rao and Saxena (1981) who studied the power of a 'sometimes pool procedure', first by numerical evaluation of the power function with a (for those years) sophisticated computer, and then by deriving an approximation which can be evaluated easier, though still numerically. Gupta and Srivastava (1963) derived an upper bound for the size of a sometimes pool procedure in the limited case that also the null hypotheses of the preliminary tests are true. Bancroft and Han (1980) also discuss the choice of the significance level for the preliminary test, as it is observed that 0.05 is unacceptable in most cases. In many cases they recommend to take at least 0.25 for the level of the preliminary test.

Some other authors too considered the choice of the significance level for the preliminary test, all recommending a larger value than the common value 0.05. In most cases this problem is treated for some procedure where the final inference is estimation, in which case the bias, arising when the significance level is chosen smaller than 1, can be weighted against the variance, cf. e.g. Gun (1969), Sawa and Hiromatsu (1973), Farebrother (1975), Toyoda and Wallace (1976).

Procedures involving preliminary tests are furthermore often employed if the uncertainty in the model specification concerns the inclusion of a parameter in a given model. A practical example is the question whether an interaction term should be included in an analysis-of-variance model. A preliminary test is used to test whether the interaction term equals zero. If this test is not significant, the main effects are estimated and tested in the model without interaction term. On the other hand, if the preliminary test rejects, we are faced with an interpretation problem for the main effects. This example is considered by Fabian (1991). In the light of his negative conclusions, he does not even specify what may be done after rejection by the preliminary test. He analyzes an improved version of the above procedure which also takes into account the estimated (on the basis of the power of the preliminary test) error due to neglecting interactions. He concludes however, that the improved procedure, and hence also the original procedure, is absolutely unsatisfactory.

Regression models also appear frequently as a subject of study. Consider the simple bivariate regression model $Y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ with the ε's independently identically normally distributed with expectation zero. Suppose that one is interested in the influence of x_1 on the response variable Y and hence in inference about β_1, but that one is uncertain about whether the term with x_2 should be incorporated in the model. Then a common procedure is to test the hypothesis $\bar{H}_0 : \beta_2 = 0$ vs. $\bar{H}_1 : \beta_2 \neq 0$. If the preliminary test for this hypothesis is not significant at some given level, then the term with x_2 is omitted and the reduced model $Y = \beta_1 x_1 + \varepsilon$ is

fitted to make inference about β_1. Otherwise it is incorporated in the model and the complete model is fitted.

This example occurred already in the first article in which the effect of preliminary testing on subsequent inference was studied (Bancroft (1944)), but it also has appeared in many articles published since then. For example, more recently Giles and Srivastava (1993) derived the distribution function of the conditional estimator $\hat{\beta}_1$ of β_1, given by either the restricted (by $\beta_2 = 0$) ordinary least squares (OLS) estimator in case the preliminary test did not reject, or by the unrestricted OLS estimator after rejection. The result turns out to be very complicated thus requiring numerical evaluation. Moreover, it is data-dependent and depends on all the unknown parameters in the problem.

Saleh and Sen (1983) and Saleh and Sen (1984) do not restrict attention to one particular situation as most authors do, but consider the more general multivariate linear model $\boldsymbol{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$, where a test of hypothesis on a subset $\boldsymbol{\beta_1}$ of the parameters follows a preliminary test on the complementary subset $\boldsymbol{\beta_2}$. For a class of likelihood ratio and rank order tests, they derive first-order asymptotic approximations for size and power of the pre-test procedure. These approximations involve the joint distribution of non-central χ^2-statistics which can only be evaluated by some elaborate expansion. For some special cases (like $\boldsymbol{\beta_2} = \mathbf{0}$) they give an upper bound of the asymptotic size that depends on the levels of the separate tests and the power of the preliminary test. As regards the asymptotic power of the pre-test procedure, they observe that it lies between that of the two main tests, being more (less) efficiency-robust than that of the restricted (unrestricted) test when $\boldsymbol{\beta_2} \neq \mathbf{0}$ may not hold. For comparison of the ranks vs. least squares solutions, numerical evaluation is needed since the results are too complicated.

Easterling and Anderson (1978) study by means of simulations the effect of preliminary goodness-of-fit tests for normality on subsequent inference for the mean. They initiated their study with the feeling that pre-testing on normality is not only conventional, but good. However, their simulation results are not at all supportive. Particularly for asymmetric distributions, passing a goodness-of-fit test does not provide protection against errors due to wrongly using normal distribution theory for obtaining a confidence interval for the mean.

1.3 Repeated use of data in a wider perspective

Besides preliminary tests followed by the main statistical analysis on the same data, there are other possibilities to deal with uncertainty in the specification of a model. One might divide the available data into two parts, one part to perform the preliminary test on, the other part to make the final inference that one is interested in. In this situation, a recommendation regarding the sizes of the sub-samples would be needed. A technique of data-splitting is encountered in the theory of semi-parametric models for the construction of efficient estimators (Bickel, Klaassen, Ritov, and Wellner (1993), Sec. 7.8). Unfortunately, its practical use is limited, since for finite samples it

is not that clear how to split the sample. Another possibility is to use non-parametric or distribution-free methods instead of preliminary tests. The idea is that such tests are so robust that they will be valid even under the largest possible model, and thus selection of the model becomes superfluous. However, use of these methods may lead to unnecessary loss of efficiency. Also Bayesian procedures might be considered, but a difficulty then is that they require a priori information. Cohen (1974) remarks that preliminary tests form a sort of compromise between standard procedures and Bayesian procedures.

In practice, people often use less formal methods involving repeated use of data. In exploratory data analysis, for example, graphical techniques like box plots and stem-and-leaf plots are applied to get an impression of the data. After fitting a regression model, probability plots and other plots of residuals are made to check usual assumptions like normality and homogeneous variances. However, the judgment of these plots is mostly done by eye and not by objective criteria. People often just try a few models and choose the one that best fits their data. In this way they 'check' the assumptions underlying their model, and after convincing themselves that the data do not conflict the model, they are satisfied and proceed as if they haven't done anything yet.

The use of such 'trial and error' methods to find an appropriate model has grown with the increased computing facilities, which has made it much easier to try several variants. However, even if one would try, it would hardly be possible to describe the behavior of people analyzing data with a computer, in a systematic way. Many decisions taken during the process are subjective or ad hoc, which makes it impossible to analyze the consequences of such behavior on final inference.

The problem however is, that the conclusions resulting from use of a model or method that has been selected with the same data, also depend on the preceding steps and should be interpreted with care. In practice, people often do not realize this or they assume that it won't be so bad. And if they are aware of the problem, they propose to use new data for the application of the method. However, as these are in most cases not available, we typically are left without a satisfactory solution.

Some other approaches which are related to procedures incorporating a preliminary test in the sense that they also use the available data to select a suitable model or method, are worth mentioning. In this respect one can think of so-called 'adaptive methods'. Typically a class of procedures is available, each of which is appropriate for a special distribution, but not really bad for others. Then the value of a selector statistic is computed, on the basis of which one of the available procedures, or a weighted combination of these, is chosen. Such adaptive methods are used to improve the robustness of the resulting procedure.

As an example we mention the estimation of the center of a distribution. Here one may use a trimmed mean with the trimming factor based on the observed sample, chosen such that it minimizes what is essentially the standard error of the trimmed mean after observing the sample. Another possibility is to consider a Stein-like estimator which differs from a usual pre-test estimator by the fact that it does not choose one of two possible estimators, but attaches weights according to the value of the test

statistic of the preliminary test.

In the field of 'model selection' the problem is to choose the best model from an appropriate class of models. To get a good compromise between simplicity and accuracy, selection criteria are applied. These criteria measure the discrepancy between the fitted and the actual distribution, but put a penalty on the number of parameters in the selected model, to guard against over-fitting. Akaike's criterion (AIC) and the Bayesian information criterion (BIC) are well-known examples.

Sometimes the term 'data-driven procedures' is used. The last few years data-driven goodness-of-fit tests (see Ledwina (1994)), where a goodness-of-fit test is based on a model that is selected by a selection rule applied to the data at hand, have received a lot of attention. Another example of a data-driven procedure might be the use of available data to estimate the optimal bandwidth in density estimation problems.

1.4 Outline of this thesis

Since repeated use of data is very common in practice, there is definitely a need for more insight in its consequences on final inference. In this thesis we restrict ourselves to the well-defined pre-test procedures defined in Section 1.1, in which a preliminary test determines which of two subsequent main tests will be used. Although some numerical work has been done for this relatively simple looking procedure (compared to some other procedures involving repeated use of data), insight is still lacking. The controversy between the optimistic view of most of the textbooks and the recommendations in statistical literature based on numerical work, is not yet settled. However, classical first-order asymptotics does not give suitable answers. With second-order asymptotics we will throw light on the behavior of pre-test procedures. We present transparent expressions for the actual size and power of pre-test procedures, thus offering a clear picture of their behavior in a qualitative and quantitative sense. At this point, I would like to mention the following quotations from Lehmann (1999), p. 2, p. 583, which support our idea:

> "... approximations tend to be much simpler than the exact formulas and, as a result, provide a basis for insight and understanding that often would be difficult to obtain otherwise."
> "... numerical work can be greatly strengthened by consideration of higher order asymptotics. These theoretical results paint a more general picture than that obtained through the snapshots provided by simulation."

In Chapters 2 and 3 we study the famous normal two-sample problem, mentioned before. In order to get a better understanding, we first treat the corresponding one-sample problem, in which a preliminary variance test may be used to decide on use of the Gauss test or the t-test for a subsequent test about the mean. In Chapter 2 we begin by deriving simple approximations for the sample sizes needed such that the preliminary test has sufficient power to detect such deviations from the restricted

model that would lead to unacceptable differences between the size of the specialized main test and its level. The idea behind this approach of course is that if the size of the pre-test procedure is too large, then this will be due to the fact that the size of the main test for the restricted model is unacceptable, in combination with a power of the preliminary test which is too low to prevent use of the specialized test. In Chapter 3 we study the behavior of the procedure as a whole, thus taking into account the correlations between the preliminary test and the subsequent main tests. We derive approximations for the actual size and power of the pre-test procedure for the one- and two-sample problem. From the structure of the approximations it follows that the error in the size of the pre-test procedure is not as large as expected from the approach in Chapter 2. This is due to the dependence taken into account. Chapter 2 is based on the paper by Albers, Boon, and Kallenberg (1998), which is essentially based on the two internal reports Albers, Boon, and Kallenberg (1997b) and Albers, Boon, and Kallenberg (1997a). Chapter 3 closely follows the text in Albers, Boon, and Kallenberg (2000a).

In Chapter 4 we extend our work for the normal case to general families of densities with two parameters, viz. a parameter of interest and a nuisance parameter. The preliminary test then tests whether the nuisance parameter equals some prescribed value, and is followed by a main test on the parameter of interest, either in the restricted model with the nuisance parameter known to be equal to the given value, or in the larger model where the nuisance parameter is assumed to be unknown. For the pre-test procedure resulting when first-order asymptotically optimal tests from a general class of tests are used for each testing problem, we derive a second-order asymptotic approximation for the power gain or size difference from the pre-test procedure compared to the general main test. This approximation is very transparent and hence gives us the desired insight in the relationship between the power gain (or size difference) as a function of all the underlying parameters in the problem. The class of tests turns out to be involved in the result through only one quantity, while the family of distributions appears in only three additional quantities. The result is applied to several examples from a two-parameter exponential family and from a symmetric location-scale family of distributions. For the latter, the approximation reduces to that for the normal case, considered in the previous chapter, multiplied by a constant which may differ from distribution to distribution. However, for the classes of tests considered, the hope for power gain together with protection against a large size turns out to be false for most values of the underlying unknown nuisance and interest parameter. Chapter 4 is essentially based on both the internal report Albers, Boon, and Kallenberg (1997c) and the more general paper Albers, Boon, and Kallenberg (2000b), in which fewer details are presented.

Chapter 5 deals with an extension of the classes of tests considered in Chapter 4. In the extended class the alternative main test is robust in a larger model than the model against which the preliminary test aims to provide protection. This happens for example when after a preliminary test which only protects against particular deviations from normality, the sign test is used as an alternative test for testing the median, instead of the t-test, which is optimal in a normal model. In such a situation,

the approximation derived shows that power gain is indeed possible without severe violation of the level. Chapter 5 is based on Albers, Boon, and Kallenberg (2001).

Chapter 2

The effectiveness of a preliminary test of variance for the normal one- and two-sample problem

2.1 Introduction

Consider the problem of testing the equality of the means of two independent samples from normal distributions. This is usually done with the standard two-sample t-test, which however requires that the variances are equal. In practice, quite often a preliminary F-test is applied to test the equality of the variances involved. This is close to suggesting a two-step procedure: if the F-test does not reject the hypothesis of equality of the variances, apply the two-sample t-test; otherwise apply a more robust test, for example Welch's or Satterthwaite's test or even a non-parametric test. The rationale behind this approach apparently is that once the F-test has not rejected, it will be safe to use the standard t-test, which presupposes equality of variances. More precisely, variances which differ so markedly that the level of the t-test would deviate intolerably from its nominal value, are supposed to be detected with high power by the F-test.

However, using a combination of simulation and numerical evaluation, Markowski and Markowski (1990) reveal that unfortunately we have a case of wishful thinking here. They demonstrate that for moderate sample sizes (5-40 for both samples), the actual picture rather is the reverse: variance ratios leading to grossly wrong levels of the t-test are only detected by the F-test with very low power. Hence Markowski and Markowski (1990) rightly conclude that on this basis one should certainly not advocate the use of a preliminary F-test. The same conclusion can be drawn from

Wehrhahn and Ogawa (1978).

A drawback of these papers is that the conclusions drawn are based on simulations and numerical evaluations of the rather complicated exact probabilities. However, in addition to numerical work, it would be nice if we could get insight in a qualitative way why this failure occurs and quantitatively how large it is. Hence, we would like to have some simple approximations that can be used to gain more insight in the functional relationships between various parameters.

A first step is to apply some standard asymptotics, from which the following picture evolves. Although the sample sizes tend to infinity in an asymptotic approach, the F-test should attain a prescribed power for a given level, and hence we cannot consider fixed alternatives, since then the power would tend to one as the sample sizes tend to infinity, but we have to consider Pitman alternatives. For such nearby alternatives, the variance ratio tends to one as the sample sizes increase. But under these circumstances, the level of the t-test will nicely tend to the prescribed nominal value and the problem resolves. Hence simple asymptotics, just as the intuitive reasoning mentioned above, completely overlooks the only too real difficulties established in practice.

In this chapter we show that application of second-order asymptotics fortunately does suffice to make matters transparent. We derive a simple expression for the minimal sample size that would be required to trust the intuitive justification for a preliminary F-test. These expressions make clear the relations between sample sizes, variance ratios and levels involved, thus giving the qualitative insight. Moreover, quantitatively it is also evident from these expressions that the sample sizes required are indeed typically large (and quite often huge).

Before dealing with the two-sample case in Section 2.3, we begin by studying in Section 2.2 the simpler one-sample version of the problem. Although not that often mentioned in textbooks or literature, a similar situation occurs there. If the variance is known to be equal to a certain value, one might prefer to use the Gauss test instead of the ordinary one-sample t-test for testing the hypothesis of main interest concerning the mean. A preliminary χ^2-test can be used to test the assumption about the variance.

It turns out that for the power of the preliminary test, first-order methods still suffice, but that in particular the approximation for the level of the preferred main test has to be improved by use of second-order asymptotics. Some figures are added to demonstrate that the approximations work very well.

2.2 The one-sample case

In this section we consider the one-sample counterpart of the more familiar two-sample case mentioned. From a practical point of view, this case is only marginally interesting. However, as it is considerably simpler, its study will help our understanding of what is going on, which is precisely our aim.

Let $X_1, \ldots, X_n$ be independent identically distributed (i.i.d.) random variables

(r.v.'s) from the $N(\mu, \sigma^2)$-distribution. Suppose we are interested in the mean, and want to test the main null hypothesis $H_0 : \mu = \mu_0$ against one-sided alternatives $H_1 : \mu > \mu_0$ at level α. To decide what test we shall use for this testing problem, we wonder whether we may assume the variance to be equal to some known value σ_0^2. This assumption, which clearly forms the analogue of the equality-of-variances assumption in the two-sample case, can be tested by a two-sided χ^2-test at level δ, say. This test rejects $\bar{H}_0 : \sigma = \sigma_0$ in favor of $\bar{H}_1 : \sigma \neq \sigma_0$ for either small or large values of the test statistic based on the sample variance S^2, i.e. when

$$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1,1-\delta/2} \quad \text{or} \quad > \chi^2_{n-1,\delta/2}, \tag{2.2.1}$$

where for example $\chi^2_{n-1,\delta/2}$ denotes the upper $(\delta/2)$-point of the chi-squared distribution with $(n-1)$ degrees of freedom (df).

If this preliminary test does not reject, we feel encouraged to apply the Gauss test for the main testing problem on the mean. Then H_0 is rejected when

$$\frac{n^{1/2}(\bar{X} - \mu_0)}{\sigma_0} > u_\alpha, \tag{2.2.2}$$

where $\bar{X}$ is the sample mean and $u_\alpha = \Phi^{-1}(1-\alpha)$ is the upper α-point of the standard normal distribution function Φ.

If, however, the preliminary test rejects $\bar{H}_0$, then an alternative test for the mean has to be used. An obvious choice is the t-test, which uses an estimator S for σ, and which hence rejects when

$$\frac{n^{1/2}(\bar{X} - \mu_0)}{S} > t_{n-1,\alpha} \tag{2.2.3}$$

with $t_{n-1,\alpha}$ denoting the upper α-point of the t-distribution with $(n-1)$ df.

For the procedure described above we study the power of the preliminary test in relation to the actual size of the Gauss test. We first specify what we consider to be an intolerable size for the Gauss test, and give an exact expression for the corresponding variance ratio by which this is caused. Subsequently, we consider the power of the preliminary test for these variance ratios and give a criterion to quantify when it is considered to be sufficiently large. The exact sample size needed to meet this criterion is evaluated numerically. After this, we derive approximations to get the desired theoretical insight.

The actual size of the Gauss test exactly equals

$$\tilde{\alpha} = 1 - \Phi(u_\alpha / r^{1/2}), \tag{2.2.4}$$

where $r = \sigma^2/\sigma_0^2$. Suppose we define a tolerable deviation from the nominal level by stipulating that the relative error $h(r) = (\tilde{\alpha} - \alpha)/\alpha$ in the size should satisfy

$$|h(r)| \leq |\varepsilon|, \tag{2.2.5}$$

for some given, small ε. Taking e.g. $\varepsilon = \pm 0.4$ means that at a nominal level of 5% a range from 3% to 7% is allowed. Clearly, as h is an increasing function of r, attention will focus on the just tolerable values r_e, for which $h(r_e) = \varepsilon$ holds exactly. It follows immediately that

$$r_e = u_\alpha^2 / u_{\alpha(1+\varepsilon)}^2. \tag{2.2.6}$$

Hence, for variance ratios r larger than the value of r_e corresponding to positive ε, or smaller than the value r_e corresponding to negative ε, the Gauss test is considered unacceptable.

Now consider the power of the preliminary χ^2-test for $\bar{H}_0 : \sigma = \sigma_0$. For a given level δ it can be considered as a function of the sample size n and the variance ratio r and will be denoted by $\pi(n, r)$. It equals

$$\pi(n,r) = 1 - P(\chi_{n-1}^2 < \chi_{n-1,\delta/2}^2/r) + P(\chi_{n-1}^2 < \chi_{n-1,1-\delta/2}^2/r), \tag{2.2.7}$$

where χ_{n-1}^2 is a r.v. with a chi-squared distribution with $(n-1)$ df. Again a bound has to be prescribed. Suppose we define sufficiently high power for this test to mean that for all variance ratios r for which the size of the Gauss test violates (2.2.5), we should have a power of at least π_0, i.e.

$$\pi(n,r) \geq \pi_0, \tag{2.2.8}$$

for some given, sufficiently large π_0. A minimal requirement would seem to be $\pi_0 \geq 0.50$. Since the variance ratios violating (2.2.5) satisfy $|r-1| > |r_e - 1|$, it follows from the monotonicity of the power in $|r-1|$ that it suffices to require (2.2.8) for r_e. Hence n should simply satisfy

$$\pi(n, r_e) \geq \pi_0. \tag{2.2.9}$$

Numerically, it is straightforward to obtain $n_e = n_e(\alpha, \delta, \varepsilon, \pi_0)$, the smallest integer-valued number n for which (2.2.9) holds. This is exactly the minimum sample size, needed to ensure a sufficiently high power (in the sense of (2.2.9)) in cases where deviations of the size of the Gauss test from the desired nominal level are no longer tolerable (in the sense of (2.2.5)). In Figure 2.2.1, among others, values of n_e are presented for various α, δ, ε and π_0. From these results we readily infer that very large sample sizes are required. This is in agreement with the assertion of Markowski and Markowski (1990) for the two-sample case they studied. Moreover, n_e is seen to vary widely as a function of its parameters. This makes it even more desirable to gain theoretical insight into the functional relationships involved, which is not given by the numerical results.

To this end, we now derive some simple approximations based on asymptotics. For the probability in (2.2.4) that the Gauss test rejects, we can easily write down a Taylor expansion for r about 1. This leads to a second-order approximation for the actual size of the Gauss test, reading

$$\tilde{\alpha} = 1 - \Phi(u_\alpha / r^{1/2}) = \alpha + \tfrac{1}{2}(r-1) u_\alpha \varphi(u_\alpha) + O\left((r-1)^2\right), \tag{2.2.10}$$

where $\varphi = \Phi'$ denotes the density of the standard normal distribution. The fact that we expand here in terms of $(r-1)$ is justified by the fact that $(r-1)$ has to be of order $n^{-1/2}$ (Pitman alternatives for the preliminary test). It will be immediately clear from expression (2.2.17) below, that for a different order the χ^2-test cannot achieve a prescribed power as n tends to infinity. Note that the result in (2.2.10) contains the more simple first-order approximation $\tilde{\alpha} = \alpha + O(|r-1|)$. However, this is clearly useless in the present context: it gives $\tilde{\alpha} \approx \alpha$ and thus $h(r) \approx 0$, which means that (2.2.5) is never violated. Hence we really need the second-order version $\tilde{\alpha} \approx \alpha + \frac{1}{2}(r-1)u_\alpha\varphi(u_\alpha)$ from (2.2.10), which leads to $h(r) = (\tilde{\alpha} - \alpha)/\alpha \approx \frac{1}{2}(r-1)u_\alpha\varphi(u_\alpha)/\alpha$. Replacing $h(r)$ by this approximation in solving $h(r) = \varepsilon$ leads to the approximation

$$r_1 - 1 = \frac{2\varepsilon\alpha}{u_\alpha\varphi(u_\alpha)} \tag{2.2.11}$$

for $r_e - 1$ with r_e from (2.2.6). (Obviously, the step from (2.2.6) to (2.2.11) could have been made by direct analysis, but that would have been less illuminating).

Next consider the power of the preliminary test. In order to approximate $\pi(n, r)$, we can for example use that

$$\begin{aligned} P\left(\frac{\chi^2_n - n}{\sqrt{2n}} \leq z\right) &= \Phi(z) - n^{-1/2}\frac{\sqrt{2}}{3}(z^2-1)\varphi(z) + O(n^{-1}) \\ &= \Phi(z) + O(n^{-1/2}). \end{aligned} \tag{2.2.12}$$

The first approximation in (2.2.12) is a one-step Edgeworth expansion (cf. e.g. Feller (1971), p. 542), whereas the latter obviously is the ordinary normal approximation for a standardized χ^2_n r.v. Inversion gives the corresponding approximation for the percentile points, namely

$$\begin{aligned} \frac{\chi^2_{n,\delta/2} - n}{\sqrt{2n}} &= u_{\delta/2} + n^{-1/2}\frac{\sqrt{2}}{3}(u^2_{\delta/2} - 1) + O(n^{-1}) \\ &= u_{\delta/2} + O(n^{-1/2}). \end{aligned} \tag{2.2.13}$$

An alternative possibility is to use the Wilson-Hilferty approximation, based on the normal approximation for a transformation of the original variable (see Johnson and Kotz (1970), p. 176):

$$P(\chi^2_n \leq z) \approx \Phi\left(\left\{\left(\frac{z}{n}\right)^{1/3} - 1 + \frac{2}{9n}\right\}\sqrt{\frac{9n}{2}}\right). \tag{2.2.14}$$

This leads to

$$\chi^2_{n,\delta/2} \approx n\left(u_{\delta/2}\sqrt{\frac{2}{9n}} + 1 - \frac{2}{9n}\right)^3 \tag{2.2.15}$$

as an approximation for the upper $(\delta/2)$-point.

We begin by simply applying the normal approximation from (2.2.12), Using $(r-1) = O(n^{-1/2})$ and writing the critical value, properly standardized under alternatives $r \neq 1$, as

$$\frac{\chi^2_{n-1,\delta/2}/r - (n-1)}{\sqrt{2(n-1)}} = u_{\delta/2} - (r-1)\sqrt{\frac{n}{2}} + O(n^{-1/2}), \tag{2.2.16}$$

it follows that we may approximate the power $\pi(n,r)$ of the preliminary test by

$$1 - \Phi\left(u_{\delta/2} - \sqrt{\frac{n}{2}}(r-1)\right) + \Phi\left(-u_{\delta/2} - \sqrt{\frac{n}{2}}(r-1)\right) + O(n^{-1/2}). \tag{2.2.17}$$

If we use (2.2.17) to solve $\pi(n, r_e) = \pi_0$ (cf. (2.2.9)), we obtain

$$n \approx n_1(r_e) = \frac{2(u_{\delta/2} - u_{\pi_0})^2}{(r_e - 1)^2} \tag{2.2.18}$$

under the typically reasonable assumption that one of the tails in (2.2.17) is negligible. Here we use "$\approx$" to indicate that the number of observations n should be an integer, and hence equals the smallest integer above the value $n_1(r_e)$.

Combination of this result with the approximation r_1 from (2.2.11) for the just tolerable variance ratio, leads to the following approximation for n_e

$$n_1 = n_1(r_1) = \frac{1}{2}\left\{\frac{(u_{\delta/2} - u_{\pi_0})u_\alpha \varphi(u_\alpha)}{\varepsilon\alpha}\right\}^2. \tag{2.2.19}$$

Note that (2.2.19) indeed makes transparent the way in which the sample size depends on α, δ, ε and π_0. This is partly due to the fact that most of the parameters are separated. The expression does not only show that the required sample size increases as a function of the desired power π_0 and decreases as a function of the nominal levels α and δ and the allowed deviation $|\varepsilon|$ for the level of the Gauss test (as is intuitively clear), it also clarifies its behavior in a quantitative way. Moreover, the occurrence of $\varepsilon^2\alpha^2$ in the denominator of (2.2.19) explains why such large values occur. As mentioned before, Figure 2.2.1 presents n_e for a number of configurations of the underlying parameters. To judge the performance of (2.2.19), we evaluated n_1 for these cases as well. Comparison to the exact values shows that n_1 follows these widely varying values remarkably well. In fact, the simple n_1 ignores skewness effects and thus is an even function of ε (cf. (2.2.19)), while the exact n_e typically is somewhat larger for positive ε than for negative ε. It turns out that the approximations nicely lie near the average of the corresponding two exact values. Only for the parameter combination $\delta = 0.5$, $\pi_0 = 0.7$ this does not hold. A closer look at the derivation of the approximation reveals the reason: the nominal level δ is very large compared to the required power π_0 and therefore, neglecting one of the tails is not justified and leads to overestimation (cf. (2.2.17) and (2.2.18)).

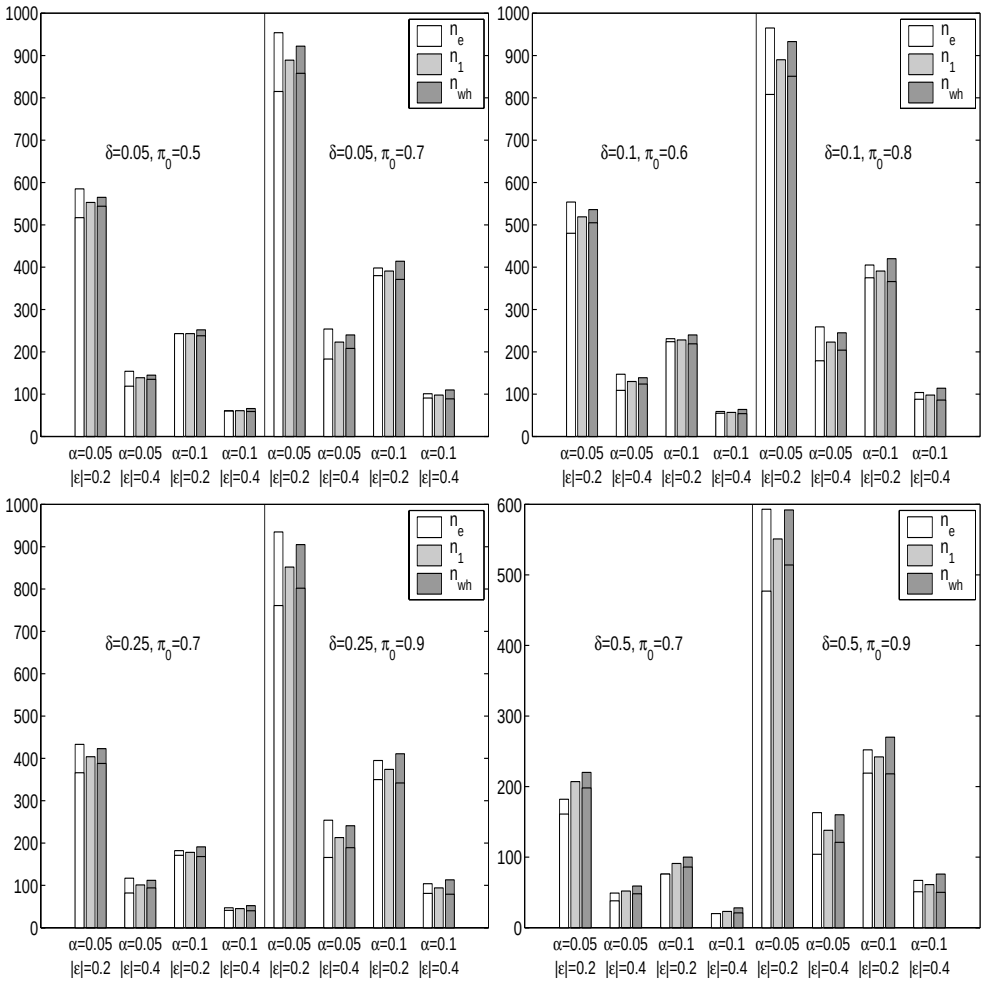

Figure 2.2.1 Minimal sample sizes for which $\pi(n, r_e) \geq \pi_0$, with $h(r_e) = \varepsilon$, for various π_0, ε and levels α and δ of Gauss test and χ^2-test, respectively. Values of n_e and n_{wh} corresponding to positive and negative ε are denoted in the same bar, where the value for positive ε is the larger one; n_1 is the same for positive and negative ε.

Looking back, we observe that for the actual size of the Gauss test in (2.2.10) the second-order approximation was really necessary, whereas for the power of the preliminary test we could do with the first-order approximation in (2.2.12). Nevertheless, we did investigate whether further improvement could be achieved by using the refinement from (2.2.12) or the Wilson-Hilferty approximation from (2.2.14) after all. As the Wilson-Hilferty approximation coincides (not by accident!) with the one-step Edgeworth approximation to $O(n^{-1})$, we only present the results for this first

possibility. The difference is completely negligible.

If $r > 1$ (which corresponds to a positive value for ε), then we ignore the lower tail probability and solve

$$1 - P(\chi^2_{n-1} < \chi^2_{n-1,\delta/2}/r) = 1 - \Phi(u_{\pi_0}). \tag{2.2.20}$$

Using (2.2.14) and (2.2.15) gives a quadratic equation in $\sqrt{9(n-1)/2}$, with positive root

$$-f + \sqrt{1+f^2}, \quad \text{where} \quad f = \frac{u_{\delta/2} - r^{1/3}u_{\pi_0}}{2(1-r^{1/3})}. \tag{2.2.21}$$

Similarly, when $r < 1$ (or ε negative), then we neglect the upper tail probability and find the same expression with f replaced by $-f$ (since $u_{1-\gamma} = -u_\gamma$). Noting that the first term ($-f$ or $+f$) is positive in both cases, substitution of r_1 into f from (2.2.21) gives as an approximation for the required sample size

$$n_{wh} - 1 = n_{wh}(r_1) - 1 = \frac{2}{9}\left(|f| + \sqrt{1+f^2}\right)^2. \tag{2.2.22}$$

The results, also shown in Figure 2.2.1, are mixed. For $\alpha = 0.05$, the values of n_{wh} for negative and positive ε lie between the corresponding values for n_e, with n_1 lying nicely in between. Hence for $\alpha = 0.05$ improvement indeed resulted. But for $\alpha = 0.10$, the value we are primarily interested in, namely the maximum of the two, is generally overestimated by n_{wh}, with a larger error than the error of n_1, which underestimates n_e. However, more important to us is the fact that the expressions for n_{wh} are much more complicated than the one for n_1. Hence, in view of the much greater simplicity of (2.2.19), refinements are not advised.

2.3 The two-sample case

We return to the main case, which is the two-sample situation. Consider two independent samples: $X_1, \ldots, X_m$ are i.i.d. r.v.'s from a $N(\mu_1, \sigma_1^2)$-distribution and $Y_1, \ldots, Y_n$ are i.i.d. r.v.'s from a $N(\mu_2, \sigma_2^2)$-distribution. The question here is whether in testing the equality of means, we may trust the assumption that the variances are equal. To test this assumption, i.e. to test $\bar{H}_0 : \sigma_1 = \sigma_2$ against $\bar{H}_1 : \sigma_1 \neq \sigma_2$, we apply a two-sided F-test at level δ, based on the ratio S_1^2/S_2^2 of sample variances. This test rejects $\bar{H}_0$ when

$$\frac{S_1^2}{S_2^2} < F_{m-1,n-1,1-\delta/2} \quad \text{or} \quad > F_{m-1,n-1,\delta/2}, \tag{2.3.1}$$

with $F_{m,n,\gamma}$ denoting the upper γ-point of the F-distribution with m and n df in numerator and denominator, respectively.

If no rejection is necessary, we dare to use the ordinary two-sample t-test, which assumes equality of variances. Then we reject $H_0 : \mu_1 = \mu_2$ in favor of $H_1 : \mu_1 > \mu_2$ if

$$\frac{(\bar{X} - \bar{Y})}{S} \left(\frac{1}{m} + \frac{1}{n} \right)^{-1/2} > t_{N-2,\alpha}, \tag{2.3.2}$$

where $N = m + n$, $\bar{X}$ and $\bar{Y}$ are the sample means, $S^2 = \left\{ (m-1)S_1^2 + (n-1)S_2^2 \right\} /(N-2)$ is the pooled variance estimator, and $t_{N-2,\alpha}$ is the upper α-point of the t-distribution with $(N-2)$ df.

If the preliminary F-test rejects $\bar{H}_0$, then one can use a modification of the two-sample t-test to test the equality of means, for example the one proposed by Welch (1937) and Satterthwaite (1946). Then H_0 is rejected when

$$(\bar{X} - \bar{Y}) \left(\frac{S_1^2}{m} + \frac{S_2^2}{n} \right)^{-1/2} > t_{\nu,\alpha}, \tag{2.3.3}$$

with ν given by

$$\nu = \left(\frac{S_1^2}{m} + \frac{S_2^2}{n} \right)^2 \Bigg/ \left(\frac{S_1^4}{m^2(m-1)} + \frac{S_2^4}{n^2(n-1)} \right). \tag{2.3.4}$$

Some straightforward modification suffices to apply the approach from the previous section. Denote $r = \sigma_1^2/\sigma_2^2$, let $\tilde{\alpha}$ now be the actual size of the two-sample t-test instead of the Gauss test and let $\pi(N, r)$ refer to the power of the F-test instead of the χ^2-test. Then we can start looking again for approximations for the total sample size N required to detect harmful deviations from the equality-of-variances assumption with high probability.

First determine for which variance ratios the size of the two-sample t-test is unsatisfactory. Using the independence between $(\bar{X}, \bar{Y})$ and (S_1^2, S_2^2), we obtain that

$$\tilde{\alpha} = 1 - E\Phi \left(t_{N-2,\alpha} \left\{ S^2 \left(\frac{1}{m} + \frac{1}{n} \right) \Bigg/ \left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \right) \right\}^{1/2} \right), \tag{2.3.5}$$

which is indeed more complicated than its one-sample counterpart $1 - \Phi(u_\alpha / r^{1/2})$. Fortunately, (2.3.5) can be simplified by expanding S^2 around its expected value and by using that $t_{N-2,\alpha} = u_\alpha + O(N^{-1})$. Combined this leads to

$$\tilde{\alpha} = 1 - \Phi \left(u_\alpha \left\{ \frac{(1-\lambda)r + \lambda}{(1-\lambda) + \lambda r} \right\}^{1/2} \right) + O(N^{-1}), \tag{2.3.6}$$

where $\lambda = n/N$. Moreover, the correspondence between the one- and the two-sample case becomes visible: the choice $\lambda = 1$ reproduces $1 - \Phi(u_\alpha / r^{1/2})$ in (2.3.6), while it transforms (2.3.2) into $m^{1/2}\bar{X}/\sigma_2 > u_\alpha$, which parallels (2.2.2). Note that this approximation does not depend on the sample sizes anymore, and that we got rid

of the expectation. This simplifies the problem in such a way that from now on, we can proceed in the same way as in the one-sample case. Hence, we may separate the two problems of, firstly, finding variance ratios r for which the size of the ordinary t-test behaves unsatisfactorily, and secondly, by substitution of the just tolerable value r_e according to (2.2.5) into $\pi(N, r)$, finding the minimum sample sizes N needed to achieve sufficiently large power in the sense of (2.2.9). As regards the error term in (2.3.6), note that it is of a smaller order than $\tilde{\alpha} - \alpha = O(|r - 1|) = O(N^{-1/2})$ (cf. (2.2.10) and (2.2.17)). Indeed, a numerical check confirms that the error due to ignoring this $O(N^{-1})$-term is negligible for the present purpose.

Solving for r_e from $h(r_e) = \varepsilon$ while using this simplification and thus ignoring the $O(N^{-1})$-term in (2.3.6), leads through the equation $\tilde{\alpha} = \alpha(1 + \varepsilon)$ to

$$1 - \Phi\left(u_\alpha \left\{\frac{(1-\lambda)r_e + \lambda}{(1-\lambda) + \lambda r_e}\right\}^{1/2}\right) = 1 - \Phi(u_{\alpha(1+\varepsilon)}), \tag{2.3.7}$$

and thus to

$$\frac{(1-\lambda)(r_e - 1) + 1}{1 + \lambda(r_e - 1)} = B, \tag{2.3.8}$$

where $B = u^2_{\alpha(1+\varepsilon)}/u^2_\alpha$. Hence

$$r_e - 1 = \frac{(B-1)}{(1-2\lambda) - (B-1)\lambda}\,. \tag{2.3.9}$$

(Note that $\lambda = 1$ leads to $r_e = B^{-1}$, as in (2.2.6).) In analogy to (2.2.9), let $N_e = N_e(\alpha, \delta, \varepsilon, \pi_0, \lambda)$ be the smallest N such that $\pi(N, r_e) \geq \pi_0$. Just like r_e, this N_e will be identified with the truly exact solution, as the difference is negligible.

Next we turn to the problem of how to derive simple approximations to r_e, and, in particular, to N_e. As $(B - 1)$ is small, the right-hand side of (2.3.9) may be approximated by $(B - 1)/(1 - 2\lambda)$, unless $(1 - 2\lambda)$ is small as well. However, it means no loss at all to exclude this possibility: when λ is close to $\frac{1}{2}$, the level of the t-test remains close to α. (Note that for $\lambda = \frac{1}{2}$, i.e. for equal sample sizes, the test statistics of the t-test and of the Satterthwaite test, which has approximately the correct level, coincide.) The problem we are interested in, of large deviations of level combined with small power will occur only for $|1 - 2\lambda|$ bounded away from 0. As a second step, we expand $(B - 1)$: as $B = u^2_{\alpha(1+\varepsilon)}/u^2_\alpha$, we observe that $\alpha(1 + \varepsilon) = 1 - \Phi(u_\alpha B^{1/2}) = \alpha - \frac{1}{2}(B - 1)u_\alpha\varphi(u_\alpha) + O((B - 1)^2)$. Hence $(B - 1)$ approximately equals $-(2\varepsilon\alpha)/(u_\alpha\varphi(u_\alpha))$. Together with the first step, substitution now gives (cf. (2.2.11))

$$r_1 - 1 = \frac{\varepsilon\alpha}{(\lambda - \frac{1}{2})u_\alpha\varphi(u_\alpha)}\,. \tag{2.3.10}$$

For the power $\pi(N, r)$ of the preliminary F-test we again begin by applying the straightforward normal approximation. Let $F_{m,n}$ be a r.v. which has an F-distribution with m and n df in numerator and denominator, respectively, then

$F_{m,n}$ (and of course also $F_{m-1,n-1}$) asymptotically has a $N(1, 2(m^{-1} + n^{-1}))$ $= N(1, 2\{\lambda(1-\lambda)N\}^{-1})$-distribution. Hence $\pi(N, r)$ is approximated by (cf. approximation (2.2.17) for the power in the one-sample case)

$$1 - \Phi\left(u_{\delta/2} - \left\{\frac{N\lambda(1-\lambda)}{2}\right\}^{1/2}(r-1)\right) + \Phi\left(-u_{\delta/2} - \left\{\frac{N\lambda(1-\lambda)}{2}\right\}^{1/2}(r-1)\right). \tag{2.3.11}$$

Assuming that one of the tails may be neglected, this leads to (cf. (2.2.18))

$$N \approx N_1(r_e) = \frac{2(u_{\delta/2} - u_{\pi_0})^2}{\lambda(1-\lambda)(r_e - 1)^2} \tag{2.3.12}$$

as an approximate solution for the total sample size satisfying $\pi(N, r_e) = \pi_0$. Combination of (2.3.10) and (2.3.12) gives as an approximation for N_e (cf. (2.2.19)):

$$N_1 = N_1(r_1) = \frac{2}{\lambda(1-\lambda)}\left\{\frac{(\lambda - \frac{1}{2})(u_{\delta/2} - u_{\pi_0})u_\alpha\varphi(u_\alpha)}{\varepsilon\alpha}\right\}^2. \tag{2.3.13}$$

Just as in the one-sample case, the dependence on the underlying parameters is made quite clear, as well as the occurrence of large values of N_e. In particular, note how $N_1/n_1 = (2\lambda - 1)^2/\{\lambda(1-\lambda)\}$ reflects the influence of the unbalancedness of the design. Figure 2.3.1, as a counterpart of Figure 2.2.1, depicts N_1 as well as N_e. Since both N_1 and N_e are symmetric around $\lambda = \frac{1}{2}$, it suffices to consider values of $\lambda < \frac{1}{2}$. The conclusion is similar: N_1 manages to follow the wild behavior of N_e quite well, as an approximate average of the exact values for positive ε and negative ε.

As concerns possible refinements, we point out that a well-known improvement consists of replacing $F_{m,n}$ by $\log F_{m,n}$. Using that $\log(F/r) \approx F - 1 - \log r$, together with the normal approximation for $F - 1$, it is easily verified that this entails replacing $(r-1)$ in (2.3.11) by $\log r$, and thus replacing (2.3.12) by

$$N_{log}(r_e) = \frac{2(u_{\delta/2} - u_{\pi_0})^2}{\lambda(1-\lambda)\log^2 r_e}, \tag{2.3.14}$$

which through combination with the approximate variance ratio r_1 from (2.3.10) leads to an approximation $N_{log} = N_{log}(r_1)$. However, it turns out that the simple N_1 outperforms this N_{log}, rather than the other way around. This surprising phenomenon is easily explained by closer inspection of the errors involved. The inaccuracy of r_1 w.r.t. r_e is largely compensated by the inaccuracy of $N_1(r_e)$ w.r.t. N_e. Such compensation is lacking for the more accurate approximation $N_{log}(r_e)$ of N_e. Of course, $N_{log}(r_e)$ itself, in combination with (2.3.9), also gives an explicit approximation for N_e, and this one is very accurate as can be seen from Figure 2.3.1. However, it is much less simple and transparent than N_1.

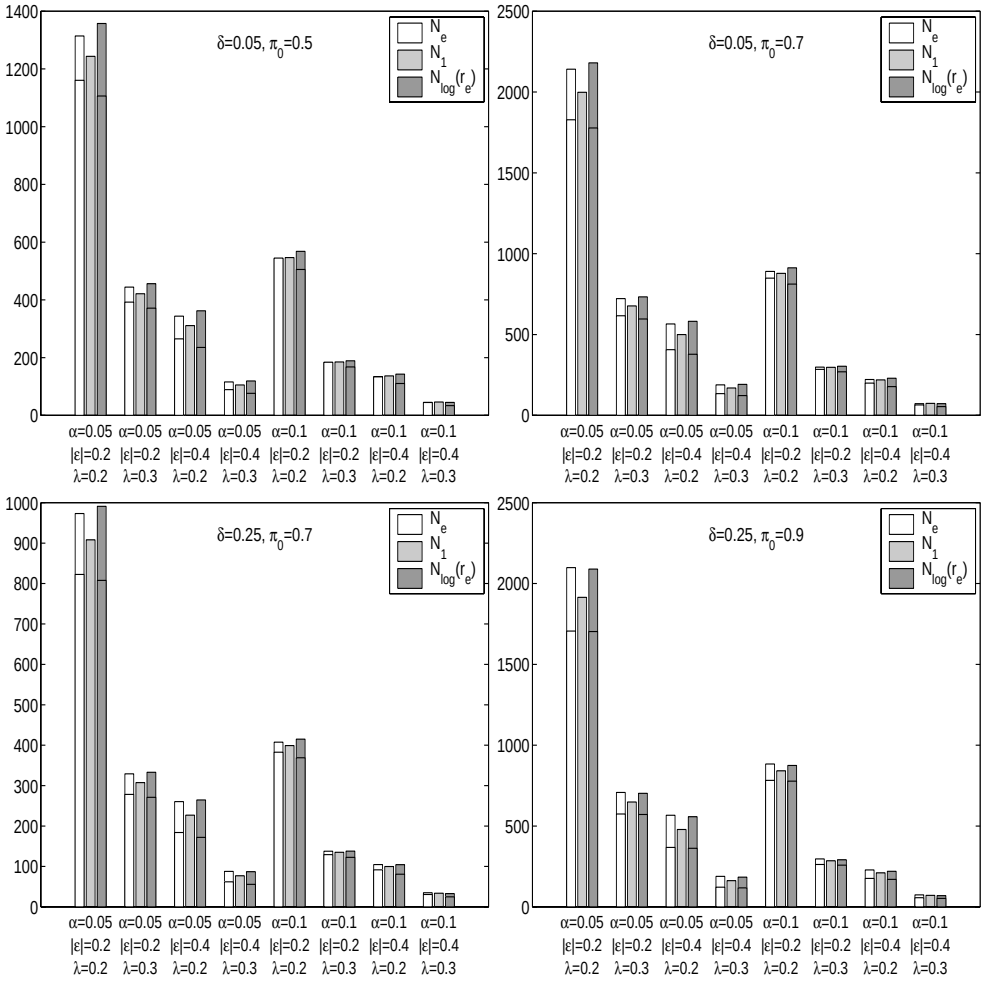

Figure 2.3.1 Minimal sample sizes for which $\pi(N, r_e) \geq \pi_0$, with $h(r_e) = \varepsilon$, for various π_0, ε and levels α and δ of t-test and F-test, respectively, and sample size ratios $\lambda < \frac{1}{2}$. Values of N_e and $N_{log}(r_e)$ corresponding to positive and negative ε are denoted in the same bar, where the value for positive ε is the larger one; N_1 is the same for positive as for negative ε.

Chapter 3

Size and power of the pre-test procedure for the normal one- and two-sample problem

3.1 Introduction

In this chapter we further analyze the pre-test procedure for the normal one- and two-sample problem from Chapter 2. Let us briefly recall the procedure. In the two-sample case, a preliminary F-test is applied in the first step to check whether the variances of the two normal samples are equal. If the F-test does not reject, then in the second step the equality of variances is taken for granted and the simple two-sample t-test is applied to test the equality of means. If the F-test does reject, then we use the more complicated Welch-Satterthwaite test. The idea is quite clear: there is a natural inclination to stick to the standard and efficient t-test as long as possible. The F-test is supposed to single out deviations from the homogeneity-of-variances assumption that would lead to unacceptable deviations from the nominal level of the t-test. In the one-sample case the F-, t- and Welch-Satterthwaite test are replaced by the χ^2-, Gauss and one-sample t-test, respectively.

Although the method described above is mentioned in numerous textbooks on statistics, almost no attention is paid to the behavior of the pre-test procedure. This suggests that most writers think that the procedure works well (and their readers too!). However, there are some questionable aspects about such pre-test procedures. The preliminary F-test may fail to reject, leading to application of the two-sample t-test while the essential assumption of equal variances is violated, and hence to serious deviations in size with respect to the prescribed level. Furthermore, the repeated use of the same data introduces correlations which influence size and power of the combined procedure.

Contrary to the implicit optimistic view of textbooks, there are some papers on the

subject that are in sharp contrast with this optimism. Related to the first aspect are the numerical results of Markowski and Markowski (1990), which show that for sample sizes occurring in practice, typically the F-test has low power against deviations that are sufficiently large to cause unacceptable departures from the nominal level of the t-test. Hence this first look at matters is not exactly reassuring. This phenomenon has been further analyzed and explained using second-order asymptotics in the previous chapter.

The preceding results are related indirectly to the problem, since they only look at the first aspect, the relation between failing of the preliminary test and the resulting violation of the level of the main test. More direct results, which also incorporate the second aspect, the correlation between the tests, are also available. Moser, Stevens, and Watts (1989) have done considerable numerical work considering a variety of combinations of the parameters involved. They conclude that whenever the t-test is unreliable with respect to size, the same may hold for the combined procedure, and therefore they advocate to always use the alternative test. Their numerical results are very interesting, but do not give much insight in the behavior of actual size and power as functions of all the parameters involved. Their conclusion is clear ("always apply the Welch-Satterthwaite test"), but still many practitioners will prefer the simpler t-test if anyhow possible. To answer questions like "how wrong can it be to follow the textbooks" and "what is the best level for the preliminary test", a simple and transparent formula for the actual size and power of the combined procedure is needed.

In view of the above, it seems worthwhile to obtain insight in a qualitative way why the combined procedure fails to live up to expectations as indicated by the papers mentioned above. Moreover, it would also be nice to have quantitative guidelines on where and to what extent problems will occur.

Instead of doing additional numerical work and trying to get the underlying structure in an experimental way, we apply asymptotic methods to express the actual size and power as functions of the parameters involved: the nominal sizes of the preliminary test and the other tests, the total number of observations and the ratio of the sample sizes of the two samples, the ratio of the variances of the two samples and (for the power) the distance between the means of the two samples. With so many parameters it is far from easy (if at all possible) to uncover the structure using only numerical results.

To begin with, however, we emphasize that first-order asymptotics will be of little help either. In the Pitman case of nearby alternatives for the preliminary F-test, the variance ratios tend to one as the sample sizes increase. In that case the standard t-test asymptotically agrees with the Welch-Satterthwaite test. In particular, its size will tend to the intended nominal value. Hence, no problem seems to be present.

In this chapter we show that second-order asymptotics suffices to make clear what happens. An explicit second-order approximation for the size and power of the combined pre-test procedure is derived in Section 3.2. To make the underlying structure as clear as possible, we first consider the simpler one-sample analogue. This facilitates the treatment and understanding of the two-sample case which subsequently follows. In both cases, the resulting expression shows the dependence on the under-

lying parameters in a straightforward manner, thus providing the desired qualitative insight. It also serves quite well for quantitative purposes, to study the magnitude of the deviations involved. It turns out that the correlation between the tests, which was not taken into account in the analysis of the previous chapter, partly (but not sufficiently!) compensates the effect of using the wrong test after an inappropriate decision (failure to reject) by the preliminary test.

The consequences with respect to the actual size of the pre-test procedure are discussed in Section 3.3, whereas the effects on the power are the subject of Section 3.4. In these sections the applicability and the accuracy of the approximations are also illustrated by solving different quantitative questions. The chapter is concluded with Section 3.5, where some comments are given about the methods used, as well as about possible refinements.

3.2 Main results: difference in size and power

3.2.1 The one-sample case

First we consider the one-sample analogue of the two-sample case from Section 3.1. As it is simpler, the desired insight is obtained rather easily, after which generalization to the case of main interest follows in a straightforward manner.

Let $X_1, \ldots, X_m$ be i.i.d. r.v.'s from the $N(\mu, \sigma^2)$-distribution. Under the assumption $\sigma = \sigma_0$, we test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ at level α by applying the Gauss test, which rejects if

$$\frac{m^{1/2}(\bar{X} - \mu_0)}{\sigma_0} > u_\alpha, \tag{3.2.1}$$

where $\bar{X}$ is the sample mean and $u_\alpha = \Phi^{-1}(1-\alpha)$ is the upper α-point of the standard normal distribution function Φ. Without such an assumption about σ, our alternative procedure will of course be the one-sample t-test, which rejects if

$$\frac{m^{1/2}(\bar{X} - \mu_0)}{S} > t_{m-1,\alpha}, \tag{3.2.2}$$

where S^2 is the sample variance and $t_{m-1,\alpha}$ is the upper α-point of the t-distribution with $(m-1)$ df. As a preliminary test for $\bar{H}_0 : \sigma = \sigma_0$, we apply a two-sided χ^2-test at level δ, which rejects if

$$\frac{(m-1)S^2}{\sigma_0^2} > \chi^2_{m-1,\delta/2} \quad \text{or} \quad < \chi^2_{m-1,1-\delta/2}, \tag{3.2.3}$$

where $\chi^2_{m-1,\delta/2}$ ($\chi^2_{m-1,1-\delta/2}$) is the upper (lower) $(\delta/2)$-point of the χ^2-distribution with $(m-1)$ df. If this preliminary test rejects $\bar{H}_0$, we subsequently use the t-test; otherwise the Gauss test is applied.

Let $P_{\mu,\sigma}$ denote that X_i, $i = 1, \ldots, m$ has a $N(\mu, \sigma^2)$-distribution, and denote the probability that the just described pre-test procedure rejects H_0 by π^*. (Of

course, it is a function of the actual values of the mean and variance. Later we will explicitly express this dependence as a function of the parameters which define the Pitman alternatives for μ and σ.) Clearly π^* is the following sum of two bivariate probabilities

$$\pi^* = P_{\mu,\sigma}\left(\frac{m^{1/2}(\bar{X}-\mu_0)}{\sigma_0} > u_\alpha, A\right) + P_{\mu,\sigma}\left(\frac{m^{1/2}(\bar{X}-\mu_0)}{S} > t_{m-1,\alpha}, A^c\right), \tag{3.2.4}$$

where

$$A = \{\chi^2_{m-1,1-\delta/2} < \frac{(m-1)S^2}{\sigma_0^2} < \chi^2_{m-1,\delta/2}\} \tag{3.2.5}$$

is the acceptance region of the χ^2-test. To study the behavior of π^*, we relate it to the probabilities of rejection of the Gauss, the t- and the χ^2-test, which we will denote by π, $\tilde{\pi}$ and $\bar{\pi}$, respectively. In particular, we consider the difference $\pi^* - \tilde{\pi}$ between the rejection probabilities of the combined procedure and the t-test. Under H_0 this is precisely the deviation of the size of the combined procedure from the required level α, since the size of the t-test is correct regardless of the actual value of σ, i.e. $\tilde{\pi}$ equals α throughout H_0 (as opposed to $\pi = \alpha$ being true for $H_0 \cap \bar{H}_0$ only). The difference $\pi^* - \tilde{\pi}$, which can of course straightforwardly be obtained by subtracting $\tilde{\pi}$ from the expression for π^* given in (3.2.4), can be rewritten in a more convenient way by

$$\pi^* - \tilde{\pi} = P_{\mu,\sigma}\left(\frac{m^{1/2}(\bar{X}-\mu_0)}{\sigma_0} > u_\alpha, A\right) - P_{\mu,\sigma}\left(\frac{m^{1/2}(\bar{X}-\mu_0)}{S} > t_{m-1,\alpha}, A\right). \tag{3.2.6}$$

This will be useful in the proof of the main result, given in the next theorem. In the theorem we give an asymptotic approximation for the difference $\pi^* - \tilde{\pi}$ as a function of the actual underlying mean and variance, the sample size m, and the nominal levels α and δ for the two main tests and for the preliminary test, respectively.

To this end, we consider Pitman alternatives, not only for $H_0 : \mu = \mu_0$, but also for $\bar{H}_0 : \sigma = \sigma_0$. To be more precise, we let without loss of generality $\mu_0 = 0$ and $\sigma_0 = 1$ and we suppose that for $m \to \infty$ we have

$$\mu = bm^{-1/2}, \quad \sigma = 1 + c(2m)^{-1/2}, \tag{3.2.7}$$

for some constants b and c with $b \geq 0$. The normalization for σ is chosen for convenience in the result. Then we obtain, with $\varphi = \Phi'$

Theorem 3.2.1 *Under* (3.2.7)*, the difference* $\pi^* - \tilde{\pi}$ *from* (3.2.6) *satisfies*

$$\pi^*(b,c) - \tilde{\pi}(b,c) = (2m)^{-1/2} u_\alpha \varphi(u_\alpha - b) h(c, u_{\delta/2}) + O(m^{-1}), \tag{3.2.8}$$

where

$$h(x,y) = x\{\Phi(y-x) - \Phi(-y-x)\} - \{\varphi(y-x) - \varphi(-y-x)\}. \tag{3.2.9}$$

Proof. To shorten notation, write the acceptance region A (cf. (3.2.5)) as $\{a_L < S < a_U\}$ (hence e.g. $a_U = \{\chi^2_{m-1,\delta/2}/(m-1)\}^{1/2}$). First conditioning on S in (3.2.6) and taking the expectation with respect to S, it follows that

$$\pi^* - \tilde{\pi} = E\left\{\Phi\left(\frac{t_{m-1,\alpha}S - b}{\sigma}\right) - \Phi\left(\frac{u_\alpha - b}{\sigma}\right)\right\} I_{(a_L,a_U)}(S). \tag{3.2.10}$$

Now we expand the part in brackets. As $t_{m-1,\alpha} = u_\alpha + O(m^{-1})$, we obtain in view of (3.2.7), according to which $\sigma = 1 + O(m^{-1/2})$

$$\Phi\left(\frac{t_{m-1,\alpha}S - b}{\sigma}\right) - \Phi\left(\frac{u_\alpha - b}{\sigma}\right) = u_\alpha\varphi(u_\alpha - b)(S-1) + O(m^{-1} + (S-1)^2). \tag{3.2.11}$$

Combination of (3.2.10) and (3.2.11) leads to

$$\pi^* - \tilde{\pi} = u_\alpha\varphi(u_\alpha - b)E(S-1)I_{(a_L,a_U)}(S) + O(m^{-1} + E(S-1)^2). \tag{3.2.12}$$

The factor $E(S-1)I_{(a_L,a_U)}(S)$ can be approximated by using the standard normal approximation to the distribution of $\tilde{S} = (S - ES)/\sqrt{\mathrm{var}(S)}$. Then we get, with $\tilde{a}_L$ and $\tilde{a}_U$ the boundaries of the acceptance region written in terms of $\tilde{S}$ (e.g. $\tilde{a}_U = (a_U - ES)/\sqrt{\mathrm{var}(S)}$)

$$\begin{aligned} &E(S-1)I_{(a_L,a_U)}(S) = E\left(\tilde{S}\sqrt{\mathrm{var}(S)} + ES - 1\right) I_{(\tilde{a}_L,\tilde{a}_U)}(\tilde{S}) \\ &= (ES-1)\left\{\Phi(\tilde{a}_U) - \Phi(\tilde{a}_L) + O(m^{-1/2})\right\} \\ &-\sqrt{\mathrm{var}(S)}\left\{\varphi(\tilde{a}_U) - \varphi(\tilde{a}_L) + O(m^{-1/2})\right\}. \end{aligned} \tag{3.2.13}$$

Remains to evaluate this expression. For evaluation of the moments, we may use the following expansion

$$\frac{S}{\sigma} = \left\{1 + \left(\frac{S^2}{\sigma^2} - 1\right)\right\}^{1/2} = 1 + \frac{1}{2}\left(\frac{S^2}{\sigma^2} - 1\right) - \frac{1}{8}\left(\frac{S^2}{\sigma^2} - 1\right)^2 + O\left(\left|\frac{S^2}{\sigma^2} - 1\right|^3\right).$$

Noting that $E\left(S^2/\sigma^2 - 1\right) = 0$, $E\left(S^2/\sigma^2 - 1\right)^2 = 2(m-1)^{-1}$ and $E\left|S^2/\sigma^2 - 1\right|^3 = O(m^{-3/2})$ and hence $E(S/\sigma) = 1 - (4m)^{-1} + O(m^{-3/2})$, $\mathrm{var}(S/\sigma) = (2m)^{-1} + O(m^{-3/2})$, we find in view of (3.2.7) that

$$\begin{aligned} &ES = \sigma + O(m^{-1}) = 1 + c(2m)^{-1/2} + O(m^{-1}), \\ &\sqrt{\mathrm{var}(S)} = (2m)^{-1/2} + O(m^{-1}). \end{aligned} \tag{3.2.14}$$

From the normal approximation of the χ^2-percentile points (cf. also (2.2.13)), it is easy to derive the approximations for a_L and a_U. For a_U we get $a_U = 1 + u_{\delta/2}(2m)^{-1/2} + O(m^{-1})$, for a_L the same expression with $u_{\delta/2}$ replaced by $-u_{\delta/2}$. Hence

$$\tilde{a}_L = -u_{\delta/2} - c + O(m^{-1/2}) \quad \text{and} \quad \tilde{a}_U = u_{\delta/2} - c + O(m^{-1/2}). \tag{3.2.15}$$

Finally, combination of (3.2.12) and (3.2.13) with (3.2.14) and (3.2.15) yields the desired result. □

3.2.2 The two-sample case

Next we return to the situation of main interest, which is the two-sample case. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent r.v.'s, the X_i from $N(\mu_1, \sigma_1^2)$, $i = 1, \ldots, m$ and the Y_j from $N(\mu_2, \sigma_2^2)$, $j = 1, \ldots, n$. A two-sided F-test, based on the ratio S_1^2/S_2^2 of sample variances, is applied at level δ to test $\bar{H}_0 : \sigma_1 = \sigma_2$. This test rejects $\bar{H}_0$ when

$$\frac{S_1^2}{S_2^2} < F_{m-1,n-1,1-\delta/2} \quad \text{or} \quad > F_{m-1,n-1,\delta/2}, \tag{3.2.16}$$

with $F_{m-1,n-1,\delta/2}$ ($F_{m-1,n-1,1-\delta/2}$) denoting the upper (lower) $(\delta/2)$-point of the F-distribution with $(m-1)$ and $(n-1)$ df in numerator and denominator, respectively. As long as no rejection is needed, we use the standard two-sample t-test, which rejects $H_0 : \mu_1 = \mu_2$ at level α in favor of $H_1 : \mu_1 > \mu_2$ if

$$\frac{(\bar{X} - \bar{Y})}{S} \left(\frac{1}{m} + \frac{1}{n} \right)^{-1/2} > t_{N-2,\alpha}, \tag{3.2.17}$$

where $N = m + n$, $\bar{X}$ and $\bar{Y}$ are the sample means and $S^2 = \{(m-1)S_1^2 + (n-1)S_2^2\}/(N-2)$. However, if $\bar{H}_0$ has to be rejected, we resort to the approximate Welch-Satterthwaite t-test, which rejects if

$$\frac{\bar{X} - \bar{Y}}{(S_1^2/m + S_2^2/n)^{1/2}} > t_{\nu,\alpha}, \tag{3.2.18}$$

where

$$\nu = \left(\frac{S_1^2}{m} + \frac{S_2^2}{n} \right)^2 \Bigg/ \left(\frac{S_1^4}{m^2(m-1)} + \frac{S_2^4}{n^2(n-1)} \right).$$

Again π^* is the power of the combined procedure, which can easily be written down analogous to expression (3.2.4) for the one-sample case, but now π, $\tilde{\pi}$ and $\bar{\pi}$ stand for the powers of the two-sample t-, the Welch-Satterthwaite and the F-test. Without loss of generality we let $\sigma_2 = 1$ and, in analogy to (3.2.7), assume that for some constants b and c with $b \geq 0$

$$\mu_1 - \mu_2 = b(\kappa N)^{-1/2}, \quad \sigma_1 = 1 + c(2\kappa N)^{-1/2}, \tag{3.2.19}$$

where $\kappa = \lambda(1-\lambda)$, with $\lambda = n/N$ (and thus $(\kappa N)^{-1/2} = (1/m + 1/n)^{1/2}$). Then we have that

Theorem 3.2.2 *Under (3.2.19), the difference $\pi^* - \tilde{\pi}$ satisfies*

$$\pi^*(b,c) - \tilde{\pi}(b,c) = (2\lambda - 1)(2\kappa N)^{-1/2} u_\alpha \varphi(u_\alpha - b) h(c, u_{\delta/2}) + O(N^{-1}), \tag{3.2.20}$$

with h as in (3.2.9).

Proof. We extend the proof of Theorem 3.2.1. From (3.2.17) and (3.2.18) it follows that we obtain in analogy to (3.2.10) that $\pi^* - \tilde{\pi}$ equals

$$\begin{aligned} E\Bigg\{ &- \Phi\left(t_{N-2,\alpha} \left[\frac{S^2}{(\sigma_1^2/m + \sigma_2^2/n)} \left(\frac{1}{m} + \frac{1}{n} \right) \right]^{1/2} - \frac{b(\kappa N)^{-1/2}}{(\sigma_1^2/m + \sigma_2^2/n)^{1/2}} \right) \\ &+ \Phi\left(t_{\nu,\alpha} \left[\frac{S_1^2/m + S_2^2/n}{\sigma_1^2/m + \sigma_2^2/n} \right]^{1/2} - \frac{b(\kappa N)^{-1/2}}{(\sigma_1^2/m + \sigma_2^2/n)^{1/2}} \right) \Bigg\} I_{(a_L, a_U)} \left(\frac{S_1^2}{S_2^2} \right), \end{aligned} \tag{3.2.21}$$

where now a_U and a_L correspond to the upper and lower $(\delta/2)$-points of the $F_{m-1,n-1}$-distribution (hence e.g. $a_U = F_{m-1,n-1,\delta/2}$). A step similar to (3.2.11), but slightly more complicated, shows that the leading term of $\pi^* - \tilde{\pi}$ will be contained in

$$\tfrac{1}{2}(2\lambda - 1) u_\alpha \varphi(u_\alpha - b) E\left(\frac{S_1^2}{S_2^2} - 1 \right) I_{(a_L, a_U)} \left(\frac{S_1^2}{S_2^2} \right). \tag{3.2.22}$$

The desired result (3.2.20) then follows by applying (3.2.13) with S replaced by S_1^2/S_2^2 and by using some straightforward results like

$$\begin{aligned} &ES_1^2/S_2^2 = \sigma_1^2/\sigma_2^2(1 + O(N^{-1})) = 1 + c\{2/(\kappa N)\}^{1/2} + O(N^{-1}), \\ &\mathrm{var}(S_1^2/S_2^2) = 2/(\kappa N)(1 + O(N^{-1/2})), \\ &a_U = 1 + u_{\delta/2}\{2/(\kappa N)\}^{1/2} + O(N^{-1}), \end{aligned} \tag{3.2.23}$$

which follow from expectation, variance and normal approximation of an F-distributed variable, respectively. □

Remark 3.2.1 Comparison of (3.2.19) and (3.2.20) to (3.2.7) and (3.2.8), respectively, shows that the two-sample case is indeed closely related to the more transparent one-sample case: $(\kappa N)^{-1} = m^{-1} + n^{-1}$ replaces m^{-1} from the former case, while the factor $(2\lambda - 1)\kappa^{-1/2}$ in (3.2.20) represents the unbalance of the design. In fact, the limiting situation $n = \infty$ exactly coincides with the one-sample case. Note that the two-sample t-test corresponds to the Gauss test in the one-sample case, rather than to the one-sample t-test.

Remark 3.2.2 Up to now we have restricted attention to the one-sided test of the main null hypothesis H_0. It is easily verified that for the two-sided version of the test, results similar to (3.2.8) and (3.2.20) will hold. To be precise, it suffices to put in a factor 2 (to account for both tails) and to replace α by $\alpha/2$. (Hence the expression for the relative error $(\pi^*(0,c)-\alpha)/\alpha$ of the size in the two-sided case is the same as in the one-sided case with therein α replaced by $\alpha/2$.)

Remark 3.2.3 We should bear in mind that the approximation (3.2.8) or (3.2.20) is no goal in itself, but only serves to provide answers about questions concerning when and to what extent deviations can occur. Hence we shall postpone details about numerical accuracy mainly to the next section. Here we merely remark that the approximations seem sufficiently adequate for our purposes. For example, (3.2.8) leads for $m = 25$ and $\alpha = 0.05$ to relative errors in the size of about 3% to 4%. Moreover, the magnitude of the error indeed behaves like order m^{-1}, as suggested by (3.2.8).

3.3 Consequences for the actual size

In this section we consider the null hypothesis case, where b from (3.2.7) and (3.2.19) equals zero and $\pi^* - \tilde{\pi}$ from (3.2.8) and (3.2.20) thus reduces (up to the order given) to $\pi^*(0,c) - \alpha$, the departure from the nominal level. We begin by noting that the expressions in (3.2.8) and (3.2.20) nicely separate the effects of the parameters involved. Clearly, the error $\pi^*(0,c) - \alpha$ decreases at rate $m^{-1/2}$ or $(\kappa N)^{-1/2}$ as the sample sizes increase. It also tends to 0 for $\alpha \downarrow 0$, but $u_\alpha \varphi(u_\alpha)/\alpha$ behaves like u_α^2 in this case, and thus the relative error increases. For the two-sample case, the factor $(2\lambda - 1)\kappa^{-1/2}$ in (3.2.20) represents the unbalancedness of the design. In particular, it follows that for $m = n$ the error is 0 to second order as well. This is in agreement with the conclusion from Markowski and Markowski (1990) and Moser, Stevens, and Watts (1989), according to which the problem vanishes for (nearly) equal sample sizes. Moreover, $(2\lambda - 1)\kappa^{-1/2} = 2(\lambda - \frac{1}{2})\{\frac{1}{4} - (\lambda - \frac{1}{2})^2\}^{-1/2}$ and hence, the farther away λ is from $\frac{1}{2}$, the larger this factor is.

After these simple relations for m, n, and α, it remains to study the dependence on δ and c, which is contained in $h(c, u_{\delta/2})$ from (3.2.8). As our starting point we shall use the following link to the preliminary approach by Markowski and Markowski (1990), which was further analyzed in Chapter 2. If the dependence between the preliminary test and the subsequent tests is ignored, the right-hand side of (3.2.6) reduces to $(\pi - \tilde{\pi})(1 - \bar{\pi})$, which in the present situation equals $(\pi(0,c) - \alpha)(1 - \bar{\pi}(0,c))$. Hence this indeed suggests that control of the error $\pi^*(0,c) - \alpha$ requires that either the error $\pi(0,c) - \alpha$ in the size of the Gauss test or two-sample t-test, respectively, is small, or that the power $\bar{\pi}$ is large for $c \neq 0$.

The connection to the present approach now simply is the following: write $h =$

$h_1 + h_2$, in which

$$\begin{aligned} h_1(x,y) &= x\{\Phi(y-x) - \Phi(-y-x)\}, \\ h_2(x,y) &= \varphi(y+x) - \varphi(y-x), \end{aligned} \tag{3.3.1}$$

then $(\pi - \tilde{\pi})(1 - \bar{\pi})$ is nothing but the expression in either (3.2.8) or (3.2.20) if we replace h by h_1. Hence the thus ignored contribution by h_2 represents the correction required by the dependence between the test statistics. The behavior of h and the role of h_1 and h_2 can be analyzed as follows. Note that $h_1(c, u_{\delta/2})$ and $h_2(c, u_{\delta/2})$ are odd in c, so it suffices to consider $c \geq 0$. For all $c \geq 0$ and $u_{\delta/2} \geq 0$, we have $h_1 \geq 0$ and $h_2 \leq 0$. Since $h(c, 0) = h_1(c, 0) = h_2(c, 0) = 0$ for all c, and $h = h_1 + h_2$ increases in $u_{\delta/2}$ for $c \geq 0$, the sum $h \geq 0$ for all $u_{\delta/2} \geq 0$ and $c \geq 0$. Hence, $0 \leq h \leq h_1$. Moreover, since $h(0, u_{\delta/2}) = 0$ and $\lim_{c\to\infty} h(c, u_{\delta/2}) = 0$ while $h(c, u_{\delta/2}) \geq 0$ for all $u_{\delta/2}$ and $c \geq 0$, the function h has a maximum for some $c \geq 0$. This also applies to h_1.

The statistical implications of these properties are illustrated in Figure 3.3.1. If the Gauss test or two-sample t-test, respectively, is directly applied without preliminary test ($\delta = 0$), then we have $h(c, u_{\delta/2}) = c$ which would lead to dramatic deviations of the size, since then the deviation grows linearly in c. Application of the pre-test procedure ($\delta > 0$) improves matters. Then the violation of the level is smaller than that of the Gauss or two-sample t-test, for two reasons. The first reason for this is that the factor $(1 - \bar{\pi})$ starts to play a role. Instead of the linear function c, we now get the function h_1, in which c is multiplied by the acceptance probability of the preliminary test $(1-\bar{\pi})$ which is smaller than 1. Contrary to the linear function c, the function h_1 reaches a maximum for some $c \geq 0$, and then decreases, reflecting that at some point the power of the preliminary test becomes large enough to detect harmful deviations from the assumptions about the variance(s). Secondly, the effect of the dependence between the tests (which was not taken into account previously) leads to a further reduction of the error through replacement of h_1 by h. In Figure 3.3.1 the functions are sketched for $\delta = 0.05$.

Apart from the qualitative properties, derived in the foregoing from (3.2.8) or (3.2.20), also many quantitative questions can be answered from these results. We mention a few typical questions of increasing complexity and present also some numerical results to illustrate the accuracy of the approximations. Note that $\tilde{\pi}(0, c) = \alpha + O(N^{-1})$.

Question 1. What is the error of the size if the nominal level $\alpha = 0.05$, $m = 20$, $n = 40$, $\sigma_1 = 1.3$, $\sigma_2 = 1$ and $\delta = 0.05$?

We get $\lambda = \frac{2}{3}$, $\kappa = \frac{2}{9}$, $N = 60$, $u_\alpha = 1.645$, $c = 1.55$ and $u_{\delta/2} = 1.96$. Inserting this in (3.2.20) yields $\pi^*(0, 1.55) \approx 0.0572$. The numerical value of $\pi^*(0, 1.55)$ equals 0.0582. Replacing $\alpha = 0.05$ by $\alpha = 0.01$ we get $\pi^*(0, 1.55) \approx 0.0126$ with corresponding numerical value 0.0137.

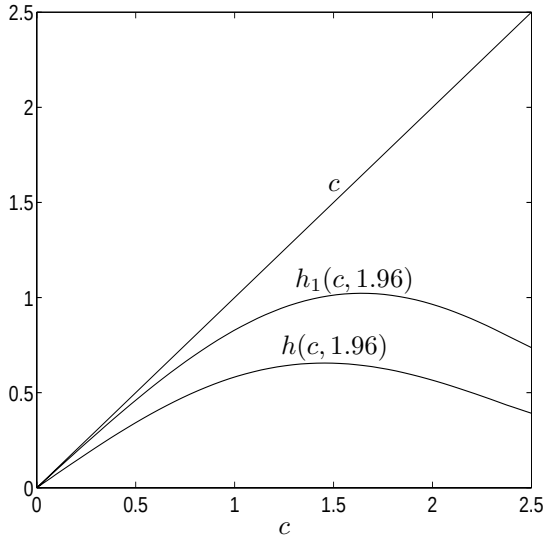

Figure 3.3.1 Functions c, h_1 and h as a function of c for $\delta = 0.05$.

Question 2. How large do we have to take N in order that the absolute error of the size is at most 0.01 for all λ with $0.2 \leq \lambda \leq 0.8$, when $c = 1.55$, $\alpha = 0.05$ and $\delta = 0.05$?

The approximation of the absolute error, using the right-hand side of (3.2.20), is an increasing function of $|\lambda - \frac{1}{2}|$. Hence, we should consider $\lambda = 0.2$ or 0.8. We get $|\pi^*(0, 1.55) - \alpha| \approx 0.1179\ N^{-1/2}$, implying $N = 139$. The numerical value for $m = 28$, $n = 111$ of $\pi^*(0, 1.55)$ is 0.0612, while for $m = 111$ and $n = 28$ we get 0.0407 numerically.

Question 3. How large do we have to take δ in order that the (absolute) error of the size is at most 0.005 for $\alpha = 0.05$, $m = 20$, $n = 40$, $\sigma_1 = 1.3$ and $\sigma_2 = 1$?

Instead of δ being prescribed (cf. Question 1), we now know the maximum deviation from the size which is allowed, and have to choose δ. Substituting $\lambda = \frac{2}{3}$, $\kappa = \frac{2}{9}$, $N = 60$, $u_\alpha = 1.645$ and $c = 1.55$ in (3.2.20), it follows that we have to solve $0.01095\ h(1.55, u_{\delta/2}) \leq 0.005$. This can be done numerically, but also some further approximations can be applied. Firstly, we may ignore $\Phi(-u_{\delta/2} - 1.55)$ and $\varphi(-u_{\delta/2} - 1.55)$ (from the lower tail of the acceptance probability of the preliminary test), since they will be very small. Secondly, for $u_{\delta/2}$ not too far from 1.55 we may replace $\Phi(u_{\delta/2} - c)$ by its first-order Taylor expansion $\Phi(0) + (u_{\delta/2} - c)\varphi(0)$ and also $\varphi(u_{\delta/2} - c)$ by $\varphi(0)$. Hence, we solve $0.01095[1.55\{\frac{1}{2} + (u_{\delta/2} - 1.55)\varphi(0)\} - \varphi(0)] \leq 0.005$, which is linear in $u_{\delta/2}$, yielding $u_{\delta/2} \leq 1.68$. The numerical value of $\pi^*(0, 1.55)$ for $m = 20$, $n = 40$, $\sigma_1 = 1.3$, $\sigma_2 = 1$, $\alpha = 0.05$ and $\delta = 0.093$ (corresponding to

$u_{\delta/2} = 1.68$) is 0.0561.

Question 4. What is the *maximal* error of the size if $\alpha = 0.05$, $m = 20$, $n = 40$ and $\delta = 0.05$? (cf. Question 1).

Now we have to deal with $\max_c h(c, 1.96)$. The maximum equals 0.6581 and the maximum is attained at $c = 1.4583$, cf. also Figure 3.3.1. Therefore the error, maximized over c, equals 0.0072 according to (3.2.20). The numerical value of $\pi^*(0, c) - \alpha$ at $c = 1.4583$ is 0.0081. Replacing $\delta = 0.05$ by $\delta = 0.25$ we get $\max_c h(c, 1.15) = 0.1922$, which is attained at $c = 1.1431$. The resulting error, according to (3.2.20), equals 0.0021. The numerical value of $\pi^*(0, c) - \alpha$ at $c = 1.1431$ is 0.0027.

Question 5. What is the best level for the preliminary test if $|(\pi^*(0, c) - \alpha)/\alpha|$ should not exceed some given $\varepsilon > 0$ (e.g. 0.20 leading to an actual size of either 2% or 3% for a nominal level $\alpha = 0.025$) for all c?

Let $c^* = c^*(u_{\delta/2})$ be the c that maximizes $h(c, u_{\delta/2})$ for given δ and $h^*(u_{\delta/2}) = h(c^*, u_{\delta/2})$ the maximum value. Note that, as h itself decreases in δ, so does h^*. By (3.2.20), δ should be sufficiently large to ensure that

$$h^*(u_{\delta/2}) \le \frac{\varepsilon(2\kappa N)^{1/2}\alpha}{|2\lambda - 1|u_\alpha \varphi(u_\alpha)}\,. \tag{3.3.2}$$

To evaluate $h^*(u_{\delta/2})$ we may use the following further approximation, cf. also Question 3. For $u_{\delta/2} \ge 1$ we have $c^* \ge 1.1$ (which can be checked empirically) and hence we ignore $c\Phi(-u_{\delta/2} - c)$ and $\varphi(-u_{\delta/2} - c)$. Then $c^* = u_{\delta/2} - g(1/u_{\delta/2})$ with $g = (\varphi/\Phi)^{-1}$ and $h^*(u_{\delta/2}) = \{u_{\delta/2}[u_{\delta/2} - g(1/u_{\delta/2})] - 1\}\varphi(g(1/u_{\delta/2}))$. It turns out that for $1 \le u_{\delta/2} \le 2.5$ a good linear approximation is obtained by taking $g(x) = \frac{3}{2}(1 - \frac{5}{4}x)$ and $h^*(x) = \frac{3}{5}x - \frac{1}{2}$. Taking $m = 20$, $n = 80$, $\alpha = 0.025$ and $\varepsilon = 0.2$, (3.3.2) reads as $\frac{3}{5}u_{\delta/2} - \frac{1}{2} \le 0.4115$, yielding $\delta \ge 0.129$. The numerical value of $\pi^*(0, c)$ at $c = u_{\delta/2} - \frac{3}{2}(1 - 5/(4u_{\delta/2})) = 1.2534$ equals 0.0315.

It is seen from these examples that pretty good and very useful answers can be achieved for many questions in an easy way using Theorem 3.2.2. In a similar way Theorem 3.2.1 can be applied.

3.4 Consequences for the actual power

Here we can be quite brief, by virtue of the simplicity of the situation. From (3.2.20) it follows that for all c

$$\frac{\pi^*(b, c) - \tilde{\pi}(b, c)}{\pi^*(0, c) - \alpha} = \frac{\varphi(u_\alpha - b)}{\varphi(u_\alpha)} + O(N^{-1/2}). \tag{3.4.1}$$

Hence to first order the power difference is nothing but the size difference, inflated by a factor $\varphi(u_\alpha - b)/\varphi(u_\alpha)$. Clearly, this factor is always positive. It will typically be larger than 1, and it will reach its maximal value for $b = u_\alpha$, e.g. 3.9 if $\alpha = 0.05$.

As expected, a gain in power is thus achieved by exceeding the nominal level under the null hypothesis. Likewise, if the procedure happens to be conservative, then a loss of power is sustained. However, it is not known beforehand which of the two will be the case.

This pattern is of a general nature. For example, take the simpler situation of a single test statistic which has a distribution that converges to a normal one with expectation b and variance 1. For comparison one may think of the one-sample t-test in the one-sample case. Then we can also gain power by allowing a larger size, (which would be fair in the comparison with the pre-test procedure if we do allow it there), since a larger size will automatically lead to a larger power. Suppose we tolerate a size of $\alpha(1+\varepsilon)$ instead of α, with $\varepsilon = O(N^{-1/2})$. Then relation (3.4.1) also applies. However, the difference is that then the size will always be as large as $\alpha(1+\varepsilon)$ (and the power difference will be in correspondence with this according to (3.4.1)), while in the pre-test procedure the actual size will typically be somewhere between α and $\alpha(1+\varepsilon)$. In our application we can choose δ according to (3.3.2) to ensure that $\alpha(1+\varepsilon)$ will be the worst possible case, attained for c^*.

To conclude this section, we answer a typical question (cf. the approach in Section 3.3).

Question 6. What is the power difference $\pi^*(b,c) - \tilde{\pi}(b,c)$ if the nominal level $\alpha = 0.05$, $m = 80$, $n = 20$, $c = -1$, $b = 1$ and $\delta = 0.05$?

We get $\lambda = 0.2$, $\kappa = 0.16$, $N = 100$, $u_\alpha = 1.645$, $\sigma_1 = 0.823$ and $u_{\delta/2} = 1.96$. Inserting this in (3.2.20) yields $\pi^*(1,-1) - \tilde{\pi}(1,-1) \approx 0.0330$. The numerical value of $\pi^*(1,-1) - \tilde{\pi}(1,-1) = 0.0384$.

3.5 Discussion of the methods used

In Section 3.2 we presented the main results, followed by the implications for size (Section 3.3) and power (Section 3.4). Here we conclude this chapter with some comments on the methods used and a few possible refinements.

From the fact that the expression on the right-hand side of (3.2.8) is of order $m^{-1/2}$, it is evident that first-order asymptotics fails. Note that using (3.2.6) provides a considerable shortcut compared to using (3.2.4) together with separate second-order approximations for π, $\tilde{\pi}$ and $\bar{\pi}$. By taking the difference as in (3.2.6) the first-order terms, which are the same for π and $\tilde{\pi}$, immediately cancel. For $1-\bar{\pi}$ the first-order result $\Phi(u_{\delta/2} - c) - \Phi(-u_{\delta/2} - c)$ suffices. Being multiplied with the expression for the difference between the main tests, this still leads to a second-order approximation for the whole procedure. By taking $\delta = 0$ it follows that the leading term of $\pi - \tilde{\pi}$ equals the expression in (3.2.8), with $h(c, u_{\delta/2})$ replaced by c. Hence the part of (3.2.8) corresponding to the first of the two terms of h, stands for $(\pi - \tilde{\pi})(1 - \bar{\pi})$. This we encountered in Section 3.3 as an approximation to (3.2.6), obtained by ignoring the dependence between the t-test and the χ^2-test. Obviously, this defect has been repaired in (3.2.8) by taking the second term of h into account. This interaction term

vanishes for both $\delta = 0$ (i.e. when $\bar{\pi} = 0$ and thus $\pi^* = \pi$) and $\delta = 1$ ($\bar{\pi} = 1$ and $\pi^* = \tilde{\pi}$).

Although the expression in (3.2.8) is asymptotically negligible from the point of view of first-order asymptotics, it is typically of a larger order of magnitude than the ordinary difference $\pi - \tilde{\pi}$ when no heterogeneity of variances occurs. Indeed, for $\delta = 0$ and $c = 0$, we get $\pi - \tilde{\pi} = O(m^{-1})$ from (3.2.8). In fact, the deficiency of the t-test with respect to the Gauss test is known to equal $\frac{1}{2}u_\alpha^2$ (cf. Hodges and Lehmann (1970)). This effect, be it rather small, is visible in practice. In this light, it becomes understandable that the present effects, which are of order $m^{-1/2}$ rather than of order m^{-1}, do cause problems for sample sizes occurring in practice.

Without problem the term of order m^{-1} can be obtained in (3.2.8) by applying the Edgeworth expansion for the distribution function of S and expanding one step further in (3.2.11)-(3.2.13). However, the resulting expression, being a linear combination of $\Phi^{(k)}(\pm u_{\delta/2} - c)$, $k = 0, 1, \ldots, 4$, is much more complicated than (3.2.8) and hence we refrain from giving it here, as our aim is to provide a simple description. A similar comment applies to attempts to use a better approximation to the χ^2-distribution, such as the one due to Wilson and Hilferty (see Johnson and Kotz (1970), p. 176). They suggest to treat

$$V = \left\{ \left(\frac{S}{\sigma} \right)^{2/3} - \left(1 - \frac{2}{9(n-1)} \right) \right\} \sqrt{\frac{9(n-1)}{2}}$$

as a standard normal r.v. (cf. also (2.2.14)). Unfortunately, this leads through (3.2.11) and (3.2.12) to dealing with a truncated moment of

$$\frac{S}{\sigma} = \left\{ 1 + \sqrt{\frac{2}{9(n-1)}} V - \frac{2}{9(n-1)} \right\}^{3/2}.$$

If we approximate this in turn by $1 + \{2(n-1)\}^{-1/2} V$, we end up with (3.2.8) again.

As concerns bounds, note the following. If we do not approximate $t_{m-1,\alpha}$ and σ as we did in (3.2.11), then we can approximate (3.2.10) by

$$\varphi\left(\frac{u_\alpha - b}{\sigma} \right) E\left\{ t_{m-1,\alpha} \frac{S}{\sigma} - \frac{u_\alpha}{\sigma} \right\} I_{(a_L, a_U)}(S). \tag{3.5.1}$$

instead of (3.2.12). Since for x and y both nonnegative (nonpositive), $\Phi(y) - \Phi(x) - (y-x)\varphi(x)$ is nonpositive (nonnegative), this approximation serves in such situations as a bound for the power gain (or size difference) $\pi^* - \tilde{\pi}$. For example, under H_0 we have $b = 0$ and (3.5.1) is an upper bound for $\pi^*(0) - \alpha$. If b is sufficiently large then $(t_{m-1,\alpha}\ S - b)/\sigma$ is nonpositive except for an exponentially small probability and (3.5.1) can be used to obtain a lower bound for $\pi^* - \tilde{\pi}$. Note that for evaluation of (3.5.1) both $EI_{(a_L, a_U)}(S) = 1 - \bar{\pi}$ and $E(S/\sigma) I_{(a_L, a_U)}(S)$ can be found from χ^2- or gamma-tables.

Chapter 4

Size and power of pre-test procedures for parametric models

4.1 Introduction

The analysis in Chapters 2 and 3 of the pre-test procedure for the normal one- and two-sample problem gave us considerable insight in its behavior and in the dependence of size and power on the several variables involved (sample sizes, nominal levels and real underlying means and variances). However, more general insight in the behavior of pre-test procedures for other situations is still lacking. In this chapter we extend our knowledge to pre-test procedures for general parametric densities.

Suppose we have a density with two parameters, θ and τ, say. The main testing problem concerns a testing problem on θ, while τ is a nuisance parameter. The pre-test procedure consists of a preliminary test on τ, to test whether it equals a given value or not, followed by a suitable test in the restricted family with the given value of τ in case the preliminary test does not reject, while otherwise the test on θ is performed in the complete family including the nuisance parameter. For each of the three testing problems in the given models, we define an appropriate class of tests, containing all standard first-order optimal tests for the problem under consideration.

The idea of the procedure just described is of course that people prefer the test in the restricted model as long as possible, either because of the convenience of greater simplicity of the test in the smaller model, or because of a possibly higher power of the test for the restricted model compared to the alternative test, due to "knowing the value of the nuisance parameter". It is the aim in this chapter to find an attractive expression that reveals the differences in size and power between the pre-test procedure and the one-stage test in the complete family.

Concerning the methods we use to arrive at meaningful and transparent results,

a few things can be noticed in advance. In the special case of normal distributions, which gives back precisely the one-sample problem of the previous chapter, the χ^2-test and the Gauss or one-sample t-test are asymptotically independent (to first order). In general, independence between the preliminary test and the main tests does not necessarily hold. However, if the preliminary test is already almost a test for the main problem, a two-step procedure is not very appealing: the two steps should not be mixed up too much. And more than that, if the correlations between the test statistic applied for the preliminary test and those for the main tests are not small, then the size of the pre-test procedure varies wildly and unacceptable violations of the prescribed level cannot be avoided. This conclusion follows already from first-order asymptotics and is presented in Section 4.2. Hence, pre-test procedures are only of interest if the correlations between the preliminary test and the main tests are small.

Furthermore, if the correlation parameter is small, first-order asymptotics with respect to the sample size is of little help anymore, as will be explained in Section 4.2. Note that this was also the case in the special situation of normal distributions considered in Chapter 3. There the correlation parameter equals zero. Fortunately, second-order asymptotics makes clear what is going on. However, second-order asymptotic analysis with respect to the sample size for a fixed value of the correlation parameter, is rather hopeless and will not lead to useful expressions. This problem is tackled by applying asymptotics not only with respect to the sample size, but also with respect to the correlation parameter. The more complicated second-order analysis is made more transparent by taking into account that the two main tests are tests for essentially the same testing problem (although in different models) and therefore have some common part. The remaining difference is small, partly due to considering a small correlation parameter, which is no restriction, as pointed out. This makes it possible to reduce the rather large number of terms in the second-order asymptotic expansions substantially. This argument is worked out in the first lemma in Section 4.2 and may be of independent interest.

In the same section notations and conditions are gathered, and the classes of tests for each of the three testing problems are defined. Also a technical lemma is presented, due to Götze (1987) on Edgeworth expansions for bivariate U-statistics when dealing with probability measures depending on the number of observations. The use of second-order asymptotics implies that we have to deal with U-statistics, rather than with sums of i.i.d. r.v.'s. The evaluation of the power under local alternatives induces the dependence on the sample size for the involved probability measures.

Section 4.3 contains the main results formulated in two theorems which are proved using the lemmas of Section 4.2. The (bivariate) Edgeworth expansions of the test statistics are presented, followed by a transparent expression for the difference in size and power between the pre-test procedure and the one-stage test in the complete family. This expression gives much insight in a qualitative and quantitative sense.

It turns out that for all members of the class of test statistics the same expression holds, except for one quantity coming from the main test in the simpler model. In particular, it means that (up to the considered order) there is no difference between applying e.g. the locally most powerful test with the restricted or unrestricted max-

imum likelihood estimator of the nuisance parameter inserted, the likelihood ratio test, Rao's efficient score test or Wald's test in the preliminary test on τ or in the testing problem on θ in the complete family.

The implications of the main results are discussed in Section 4.4. It starts with a general discussion on the consequences for the actual size and power. It is noted that the family of distributions and the class of tests are involved through only four parameters. Regarding the hope for power gain, it turns out that it is in most cases not realistic, since the power change is often mainly nothing else but a factor times the size change. This implies that a higher power is almost only obtained if the size exceeds the nominal level. Moreover, a higher power at some value of the nuisance parameter goes hand in hand with a lower power at another value of the nuisance parameter, while the true value is unknown!

After the general discussion, two important classes of distributions are considered: two-parameter exponential families and symmetric location-scale families. In a general symmetric location-scale family the approximation for the deviation from the nominal level of the pre-test procedure is very simple: it is the same as that for the normal case, except for a multiplicative constant. For several examples from these two classes of distributions, the accuracy of our approximations is seen from comparison with some numerical results.

4.2 Notation, assumptions and preliminaries

Let $\theta \in \mathbb{R}$, $\tau \in \mathbb{R}$, and let $X_1, \ldots, X_n$ be i.i.d. r.v.'s with density $f(x;\theta,\tau)$ with respect to some measure μ on the measurable space $(\mathcal{X}, \mathcal{A})$. Our main testing problem concerns testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. If the nuisance parameter τ would be known and be equal to τ_0, say, we would test $\theta = \theta_0$ in the family $f(x;\theta,\tau_0)$. If we are not sure that $\tau = \tau_0$, then we use a pre-test procedure and proceed with the following two steps. First, we perform a preliminary test of $\bar{H}_0 : \tau = \tau_0$ against $\bar{H}_1 : \tau \neq \tau_0$. Secondly, if $\bar{H}_0$ is not rejected, then we test H_0 in the family $f(x;\theta,\tau_0)$; otherwise, we test H_0 against H_1 in the family $f(x;\theta,\tau)$.

To define the tests of the separate testing problems we give some notation and state some regularity conditions. Without loss of generality let $\theta_0 = 0$ and $\tau_0 = 0$. As usual (θ,τ) will denote the true value of the parameter as well as a variable in $\mathbb{R}^2$. Its meaning is clear from the context.

The regularity conditions should hold *on an open subset* Ω_0 *of the parameter space* $\Omega \subset \mathbb{R}^2$*, containing* $(0,0)$. Since we only deal with local alternatives, attention is restricted to Ω_0. The regularity conditions are of the same type as those used in classical large sample theory, see e.g. Lehmann and Casella (1998), p. 462.

(Ri) The set $A = \{x \in \mathcal{X} : f(x;\theta,\tau) > 0\}$ is independent of (θ,τ), i.e. the distributions $P_{\theta,\tau}$ of X_i have common support.

(Rii) For every $x \in A$, $(\partial^{i+j}/\partial\theta^i\partial\tau^j)f(x;\theta,\tau)$ exists for $i,j = 0,1,2,3$, $(i,j) \neq$

$(0,0)$, and we write

$$\psi^*_{ij}(x;\theta,\tau) = \frac{(\partial^{i+j}/\partial\theta^i\partial\tau^j)f(x;\theta,\tau)}{f(x;\theta,\tau)}\,.$$

(Riii) For $i,j = 0,1,2,3$, $(i,j) \neq (0,0)$, there exist functions H_{ij} such that

$$\sup_{(\theta,\tau)\in\Omega_0}\left|\frac{\partial^{i+j}}{\partial\theta^i\partial\tau^j}\,f(x;\theta,\tau)\right| \leq H_{ij}(x)$$

with

$$\int_A H_{ij}(x)d\mu(x) < \infty.$$

(Riv) The Fisher information matrix under $(\theta,\tau) = (0,0)$

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} = \begin{pmatrix} E(\psi^*_{10})^2 & E\psi^*_{10}\psi^*_{01} \\ E\psi^*_{10}\psi^*_{01} & E(\psi^*_{01})^2 \end{pmatrix}$$

is finite and positive definite.

In order to simplify notation, we write ψ^*_{ij} or $\psi^*_{ij}(X)$ for $\psi^*_{ij}(X;0,0)$. Expectations under the model $f(x;\theta,\tau)$ are denoted by $E_{\theta,\tau}$, but are abbreviated to E under $(\theta,\tau) = (0,0)$. Furthermore, let

$$\psi_{ij}(x;\theta,\tau) = \psi^*_{ij}(x;\theta,\tau)I_{11}^{-i/2}I_{22}^{-j/2}.$$

By application of the dominated convergence theorem it follows from (Rii) and (Riii) that

$$E_{\theta,\tau}\psi_{ij}(X;\theta,\tau) = 0 \quad \text{for} \quad i,j = 0,1,2,3, \quad (i,j) \neq (0,0). \tag{4.2.1}$$

This will be used many times with $(\theta,\tau) = (0,0)$. The correlation coefficient of $\psi_{10}(X_i)$ and $\psi_{01}(X_i)$ under $(\theta,\tau) = (0,0)$ is given by

$$\rho = I_{12}(I_{11}I_{22})^{-1/2}. \tag{4.2.2}$$

Define the so-called score functions under $(0,0)$

$$S = n^{-1/2}\sum_{i=1}^n \psi_{10}(X_i) \quad \text{and} \quad T = n^{-1/2}\sum_{i=1}^n \psi_{01}(X_i). \tag{4.2.3}$$

The statistic S corresponds to the locally most powerful (LMP) test for testing $\theta = 0$ in the family $f(x;\theta,0)$. Other candidates for testing H_0 are e.g. the likelihood ratio (LR) test, the test based on the maximum likelihood estimator (MLE) and

Wald's test. These are first-order optimal tests, all having the same leading term, and can (up to order $n^{-1/2}$) be written in the form

$$SK = S + n^{-1/2} S \left\{ n^{-1/2} \sum_{i=1}^{n} k(X_i) \right\} \tag{4.2.4}$$

with $Ek = 0$. Therefore, we use this class of test statistics for testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$, rejecting for large values of SK. (Here "K" refers to the known value of τ.) In particular, we have for the tests mentioned

$$\begin{aligned}
&\text{LMP} : k = 0 \\
&\text{LR} \quad : k = \tfrac{1}{2} \left\{ \psi_{20} - (\psi_{10}^2 - 1) \right\} + (\tfrac{1}{3} E\psi_{10}^3 - \tfrac{1}{2} E\psi_{10}\psi_{20})\psi_{10} \\
&\text{MLE} : k = \psi_{20} - (\psi_{10}^2 - 1) + (E\psi_{10}^3 - \tfrac{3}{2} E\psi_{10}\psi_{20})\psi_{10} \\
&\text{Wald} : k = \psi_{20} - (\psi_{10}^2 - 1) + \tfrac{1}{2}(E\psi_{10}^3 - E\psi_{10}\psi_{20})\psi_{10}.
\end{aligned} \tag{4.2.5}$$

Note that Rao's efficient score test coincides with the LMP test. In (4.2.5) the expectations $E\psi_{10}^3$ and $E\psi_{10}\psi_{20}$ appear. Their existence will be amply guaranteed by the additional regularity conditions (Rv) and (Rvi), respectively, which will be needed at the end of this section. This remark will also apply to the expectations occurring in the second main test and the preliminary test.

These tests are more familiar in case of two-sided alternatives. The one-sided forms as needed here are e.g. presented as directed versions in Barndorff-Nielsen and Cox (1994), p. 82. A brief justification for LR is as follows. (We use "$\doteq$" for approximately equal up to order $n^{-1/2}$.) Here we deal with testing in the family $f(x; \theta, 0)$ and we write $\hat{\theta}_0$ for the corresponding MLE of θ. The likelihood equations yield

$$\begin{aligned}
0 = n^{-1/2} \sum_{i=1}^{n} \psi_{10}(X_i; \hat{\theta}_0, 0) &\doteq S - n^{1/2} \hat{\theta}_0 I_{11}^{1/2} \\
&+ n^{-1/2} \left[n^{1/2} \hat{\theta}_0 I_{11}^{1/2} \left\{ n^{-1/2} \sum_{i=1}^{n} z(X_i) \right\} + an\hat{\theta}_0^2 I_{11} \right],
\end{aligned} \tag{4.2.6}$$

where $z = \psi_{20} - (\psi_{10}^2 - 1)$ and $a = E\psi_{10}^3 - \frac{3}{2} E\psi_{10}\psi_{20}$. The directed LR test statistic is given by

$$sgn(\hat{\theta}_0) \sqrt{2\{l(\hat{\theta}_0) - l(0)\}} \quad \text{with} \quad l(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta, 0).$$

By Taylor expansion we get

$$l(\hat{\theta}_0) - l(0) \doteq \hat{\theta}_0 n^{1/2} I_{11}^{1/2} S + \tfrac{1}{2} \hat{\theta}_0^2 I_{11} \left\{ \sum_{i=1}^{n} z(X_i) - n \right\} + \tfrac{1}{6} n \hat{\theta}_0^3 2a I_{11}^{3/2}. \tag{4.2.7}$$

In view of (4.2.6) we obtain

$$n^{1/2}\hat{\theta}_0 I_{11}^{1/2} \doteq S + n^{-1/2} S \left\{ n^{-1/2} \sum_{i=1}^{n} z(X_i) + aS \right\}, \tag{4.2.8}$$

which by the way gives (4.2.4) and (4.2.5) for the MLE. Inserting (4.2.8) in (4.2.7) gives

$$l(\hat{\theta}_0) - l(0) = \tfrac{1}{2} S^2 \left\{ 1 + n^{-1/2} \left[n^{-1/2} \sum_{i=1}^{n} z(X_i) + \tfrac{2}{3} aS \right] \right\}$$

and (4.2.4) and (4.2.5) for the LR statistic now easily follow.

It is easily seen that under $f(x; 0, cI_{22}^{-1/2} n^{-1/2})$ the test statistic SK converges in distribution to a normal r.v. with expectation $c\rho$ and variance 1. Therefore, if ρ is not small, the actual size of SK under $f(x; 0, cI_{22}^{-1/2} n^{-1/2})$ with $c \neq 0$ differs drastically from the prescribed level. So, even small departures for τ from 0 lead to unacceptable deviations in size.

The statistics of the preliminary test, defined later on, to first order equal $[T - \rho S](1 - \rho^2)^{-1/2}$ and converge in distribution to a normal r.v. with expectation $c\sqrt{1-\rho^2}$ and variance 1 under $f(x; 0, cI_{22}^{-1/2} n^{-1/2})$. Since $c\sqrt{1-\rho^2}$ is of the same order as $c\rho$, the preliminary tests do not have sufficient power to protect against using the test based on SK under these deviations. Consequently, the pre-test procedure will have the same disorder in size, unless ρ is small. Hence, for ρ not small the situation is clear: the pre-test procedure is unacceptable. For a practical application see Example 4.4.2.

As a consequence it is only of interest to consider small ρ, as will be assumed from now on. This implies that the preliminary tests, (mainly) based on $\psi_{01} - \rho\psi_{10}$, and the main tests, (essentially) based on ψ_{10} and $\psi_{10} - \rho\psi_{01}$, are at most weakly dependent and have, so to say, different aims.

With respect to n, first-order asymptotics is not sufficient. This is clearly seen in the special case that $\rho = 0$. The test statistics for testing $\theta = 0$ when τ is unknown, are in first order equal to $[S - \rho T](1 - \rho^2)^{-1/2}$ (cf. also (4.2.9) below). And hence, for $\rho = 0$ they equal SK to first order. This corresponds to the notion of adaptation, which implies that we can do as well not knowing the nuisance parameter as knowing it, and which is possible when $\rho = 0$, see Bickel, Klaassen, Ritov, and Wellner (1993), Sec. 2.4. This would imply that, based on first-order asymptotics, for $\rho = 0$ there is no problem with the pre-test procedure, since the two main tests and the pre-test procedure all behave in the same way. However, in the normal case, where we do have $\rho = 0$, the actual size of the pre-test procedure may differ substantially from its nominal level, see also Chapters 2 and 3, which deal with pre-test procedures for tests on normal means as well in the one-sample as in the two-sample problem, and Moser, Stevens, and Watts (1989) and Markowski and Markowski (1990), which discuss the latter testing problem. Therefore, to make clear what is going on, second-order asymptotics in n will be applied.

If τ is unknown, a test for $H_0 : \theta = 0$ is in principle based on the efficient score function for θ with nuisance parameter τ. The corresponding test statistic has the form $[S-\rho T](1-\rho^2)^{-1/2}$, and is asymptotically efficient (Pfanzagl (1980), Sec. 11-13). Similar to (4.2.4) we consider a class of first-order optimal tests, namely the class of tests based on

$$\left[S - \rho T + n^{-1/2}\left\{Sn^{-1/2}\sum_{i=1}^{n} q(X_i) + Tn^{-1/2}\sum_{i=1}^{n} r^*(X_i)\right\}\right](1-\rho^2)^{-1/2} \tag{4.2.9}$$

with $Eq = Er^* = 0$. However, the (second-order) limiting distribution of these statistics under $(0, cI_{22}^{-1/2}n^{-1/2})$ depends on c, since their expectation and variance under $(0, cI_{22}^{-1/2}n^{-1/2})$ are, up to order $n^{-1/2}$ and ρ^2 but ignoring terms of order $\rho n^{-1/2}$,

$$n^{-1/2}\left\{Eq\psi_{10} + Er^*\psi_{01} + c^2(\tfrac{1}{2}E\psi_{10}\psi_{02} + Er^*\psi_{01})\right\}$$

and

$$1 + cn^{-1/2}\left\{E\psi_{10}^2\psi_{01} + 2\left[Eq\psi_{01} + Er^*\psi_{10}\right]\right\},$$

respectively. Standardizing and plugging in $n^{1/2}\hat{\tau}I_{22}^{1/2}$, or, equivalently, T as estimator of c solves the problem. As a result we use as test statistic SU, which is (4.2.9) in which r^* is replaced by

$$r = r^* - \tfrac{1}{2}\left\{E\psi_{10}^2\psi_{01} + 2\left[Eq\psi_{01} + Er^*\psi_{10}\right]\right\}\psi_{10} - \left(\tfrac{1}{2}E\psi_{10}\psi_{02} + Er^*\psi_{01}\right)\psi_{01}. \tag{4.2.10}$$

(The "U" in SU refers to τ unknown.) The class of test statistics SU is a natural extension of the class of test statistics SK. Test statistics of the form (4.2.9), up to the considered order, are e.g. the LMP test with the given τ replaced by the restricted or unrestricted MLE, the LR test, the test based on the MLE of θ in the unrestricted model, Wald's test and Rao's efficient score test. In view of the fact that our final results do not depend on q and r we do not present the specific q and r of the beforementioned tests, except for the LMP test with the unrestricted MLE for τ, which will be used in the examples in Section 4.4. For this case, we have

$$\begin{aligned}
\text{LMP}: \quad & q = 0, \quad r^* = \psi_{11} - \psi_{10}\psi_{01} + \rho + \tfrac{1}{2}d_1\psi_{01}, \\
& r = \psi_{11} - \psi_{10}\psi_{01} + \rho + \tfrac{1}{2}d_1\psi_{01} - \tfrac{1}{2}(2E\psi_{11}\psi_{10} - E\psi_{10}^2\psi_{01} + d_1\rho)\psi_{10}, \\
& \text{where} \quad d_1 = E(-2\psi_{11}\psi_{01} - \psi_{10}\psi_{02} + 2\psi_{10}\psi_{01}^2),
\end{aligned} \tag{4.2.11}$$

where we used the expression for r from (4.2.10) to get from r^* to r.

Similarly, to test $\bar{H}_0 : \tau = 0$ (with θ unknown) we start with (4.2.9), where T and S are interchanged. Because here we have a two-sided testing problem, the small

bias in expectation does not affect the size of the test up to the order $n^{-1/2}$, but for the variance a correction is needed. Therefore, the corresponding correction for r^* is given by

$$r = r^* - \tfrac{1}{2}\left\{E\psi_{10}\psi_{01}^2 + 2\left[Eq\psi_{10} + Er^*\psi_{01}\right]\right\}\psi_{01}$$

and the test statistic is called TU. Although we use the same notation, the functions q and r appearing in SU may be different from those in TU. Which q and r are meant, from SU or TU, will be clear from the context. For the LMP case, the functions q and r turn out to be the same as those for SU, but with ψ_{10} replaced by ψ_{01}, ψ_{01} by ψ_{10} and ψ_{02} by ψ_{20}.

Having defined the classes of test statistics in the pre-test procedure, we can analyze the behavior of the resulting procedure. The remaining part of this section is devoted to two generally formulated lemmas which will facilitate the derivation of the main result for our pre-test procedure in Section 4.3.

Conceptually, a straightforward approach to investigate the pre-test procedure is to derive an expression for its (asymptotic) power and to analyze that, in particular by comparing it with the (asymptotic) power of SU, which we should use when no preliminary test is applied. We do not follow this path, but take a more subtle approach. It is seen from (4.2.4) and (4.2.9) that SK and SU are not that much different if n is large and ρ is small: S is the leading term of SK and SU. One of the basic technical tools of this chapter is to use this similarity from the very beginning. In this way we avoid a large number of terms which would cancel afterwards anyhow. In a more abstract form this argument is presented in the lemma below.

Let Φ denote the standard normal distribution function and $\Phi^{(j)}$ its j^{th} derivative. Instead of $\Phi^{(1)}$ we also write φ for the standard normal density. Moreover, let $\Phi(\cdot,\cdot;\rho)$ be the distribution function of the bivariate normal $N(0,0,1,1,\rho)$-distribution. (Note that here ρ is not the quantity defined in (4.2.2), but the usual name for the correlation coefficient in the bivariate normal distribution. The ρ occurring in Lemma 4.2.1 is in general simply the name of a variable and hence should not be identified with the quantity defined in (4.2.2), until Lemma 4.2.1 is applied in the proof of Theorem 4.3.2, when it is the quantity from (4.2.2). In view of this application it is convenient to use this notation in Lemma 4.2.1.)

Lemma 4.2.1 *Let (U_{1n}, V_n) and (U_{2n}, V_n) be (standardized) sequences of r.v.'s (possibly depending on ρ) admitting Edgeworth expansions of the following form*

$$\begin{aligned} Pr(U_{in} \le u, V_n \le v) &= \Phi(u,v;\rho_i) + n^{-1/2}\sum_{j=0}^{3} c_{ij}\Phi^{(j)}(u)\Phi^{(3-j)}(v) \\ &+ O(\rho^3 + \rho n^{-1/2}) + o(n^{-1/2}) \end{aligned} \tag{4.2.12}$$

as $\rho \to 0$, $n \to \infty$ and $i = 1, 2$, uniformly for (u, v) in each compact set in $\mathbb{R}^2$, where c_{ij}, $i = 1, 2, j = 0, \ldots, 3$, are constants and ρ_1, ρ_2 are functions of ρ and n. Let

$\rho_1 - \rho_2 = O(\rho + n^{-1/2})$ and $\rho_1 = O(n^{-1/2})$. Then for $u_{1n} = u_0 + O(\rho + n^{-1/2})$, $v_n = v_0 + O(\rho^2 + n^{-1/2})$ and $u_{2n} = u_0 + O(\rho^2 + n^{-1/2})$,

$$
\begin{aligned}
&Pr(U_{1n} \le u_{1n}, V_n \le v_n) - Pr(U_{2n} \le u_{2n}, V_n \le v_n) \\
&= \varphi(u_0)\left\{(u_{1n} - u_{2n})\Phi(v_0) - \tfrac{1}{2}(u_{1n} - u_{2n})^2 u_0 \Phi(v_0)\right. \\
&\qquad \left. + (\rho_1 - \rho_2)\varphi(v_0) - \tfrac{1}{2}(\rho_1 - \rho_2)^2 u_0 v_0 \varphi(v_0)\right\} \\
&+ n^{-1/2} \sum_{j=1}^{3} (c_{1j} - c_{2j}) \Phi^{(j)}(u_0) \Phi^{(3-j)}(v_0) \\
&+ O(\rho^3 + \rho n^{-1/2}) + o(n^{-1/2}).
\end{aligned}
\tag{4.2.13}
$$

Before proving Lemma 4.2.1 we emphasize that $O(\rho^3 + \rho n^{-1/2}) + o(n^{-1/2})$ in expressions like (4.2.12) and (4.2.13) is understood as follows. There exist a constant C and a sequence $\{a_n\}$ with $\lim_{n\to\infty} a_n = 0$ such that the difference between the expressions on the left-hand side and the right-hand side is bounded by $C(\rho^3 + \rho n^{-1/2}) + a_n n^{-1/2}$ for all ρ and n. Hence, our results are *uniformly* valid in ρ and n, but of course they are only meaningful (e.g. as approximations) for $n \to \infty$ and $\rho \to 0$.

Proof. Write $\Phi(u_{1n}, v_n; \rho_1) - \Phi(u_{2n}, v_n; \rho_2)$ as

$$\{\Phi(u_{1n}, v_n; \rho_1) - \Phi(u_{2n}, v_n; \rho_1)\} + \{\Phi(u_{2n}, v_n; \rho_1) - \Phi(u_{2n}, v_n; \rho_2)\}.$$

For the first part we use

$$
\begin{aligned}
&\Phi(u_1, v; \rho_1) - \Phi(u_2, v; \rho_1) = (u_1 - u_2)\varphi(u_2)\{\Phi(v) + O(\rho_1)\} \\
&+ \tfrac{1}{2}(u_1 - u_2)^2 \varphi(u_2)\{-u_2 \Phi(v) + O(\rho_1)\} + O(|u_1 - u_2|^3),
\end{aligned}
\tag{4.2.14}
$$

which leads to

$$
\begin{aligned}
&\Phi(u_{1n}, v_n; \rho_1) - \Phi(u_{2n}, v_n; \rho_1) \\
&= \varphi(u_0)\{(u_{1n} - u_{2n})\Phi(v_0) - \tfrac{1}{2}(u_{1n} - u_{2n})^2 u_0 \Phi(v_0)\} \\
&+ O(\rho^3 + n^{-1} + \rho n^{-1/2}).
\end{aligned}
$$

For the second part we get, with $\varphi(u, v; \rho)$ denoting the density of the bivariate normal $N(0, 0, 1, 1, \rho)$-distribution,

$$
\begin{aligned}
&\Phi(u, v; \rho_1) - \Phi(u, v; \rho_2) \\
&= \ (\rho_1 - \rho_2)\varphi(u, v; \rho_1) - \tfrac{1}{2}(\rho_1 - \rho_2)^2 uv\varphi(u)\varphi(v) + O(|\rho_1 - \rho_2|^3)
\end{aligned}
\tag{4.2.15}
$$

and hence, using $\varphi(u,v;\rho_1) = \varphi(u)\varphi(v) + O(\rho_1)$, we obtain

$$\begin{aligned}&\Phi(u_{2n}, v_n; \rho_1) - \Phi(u_{2n}, v_n; \rho_2)\\ &= \varphi(u_0)\varphi(v_0)\{(\rho_1 - \rho_2) - \tfrac{1}{2}(\rho_1 - \rho_2)^2 u_0 v_0\}\\ &+ O(\rho^3 + n^{-1} + \rho n^{-1/2}).\end{aligned}$$

Note that by considering the marginal distribution of V_n it follows that $c_{10} = c_{20}$. The proof is now easily completed. □

One should think of U_{1n}, U_{2n} and V_n as the test statistics SK, SU and TU, respectively, standardized under local alternatives with u_{in} and v_n as their critical values, shifted and rescaled by the standardization of the test statistics. The resemblance of U_{1n} and U_{2n} in our application is represented in Lemma 4.2.1 by the closeness of u_{1n} and u_{2n} to each other.

The straightforward approach, mentioned before, would imply an expansion of $\Phi(u,v;\rho_i)$ for ρ_i around 0 in (4.2.12) and expansions of *all* the terms in (4.2.12) for u_{in} around u_0 and v_n around v_0, where the ρ-, ρ^2- and $n^{-1/2}$-terms of u_{1n}, u_{2n} and v_n are made explicit. This would give a large number of terms. Presumably, also the c_{ij} will be given explicitly in such an approach, which again gives a lot of terms. Carefully gathering all these terms would show that many of them are the same for U_{1n} and U_{2n} and hence cancel in taking the difference.

Replacing the expansion of $\Phi(u,v;\rho_i)$ for ρ_i around 0 and the expansion of all the terms in (4.2.12) for u_{in} around u_0 and v_n around v_0 by the expansions for the differences given in (4.2.14) and (4.2.15), yields an enormous reduction in the number of terms and shows moreover what is *really* needed: for u_{in} their first-order term u_0 and furthermore only the *difference* $u_{1n} - u_{2n}$, while for v_n even the first-order term v_0 suffices! This is due to the similarity between U_{1n} and U_{2n} and the small correlation of V_n and U_{in}. Finally, for the c_{ij} also only the differences $c_{1j} - c_{2j}$ are involved, which gives again a large reduction of terms. The application of Lemma 4.2.1 in Section 4.3 to SK, SU and TU clearly shows that the more subtle approach using (4.2.14) and (4.2.15) is very profitable compared to the straightforward one, sketched above.

In the application of Lemma 4.2.1 we need an Edgeworth expansion of the form (4.2.12). Since SU and TU are U-statistics and since we need the expansion under local alternatives, we should have an Edgeworth expansion for bivariate U-statistics when dealing with a probability measure depending on n. Such a result is given in Götze (1987), p. 215. (The formulation of Corollary 1.18 in Götze (1987) is not very transparent. The following lemma resulted from personal communication with Götze.) Denote by $\circ$ the inner product in $\mathbb{R}^2$ and by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^2$. Let $\varphi_{0,W}$ be the density of the bivariate normal distribution with expectations 0 and covariance matrix W. Write

$$f(z) \circ D = f^{(1)}(z)\frac{\partial}{\partial x} + f^{(2)}(z)\frac{\partial}{\partial y}$$

for $f(z) = (f^{(1)}(z), f^{(2)}(z))$.

Lemma 4.2.2 (Götze). *Let $Z_1, \ldots, Z_n$ be i.i.d. r.v.'s with common distribution P_n. Expectations and covariances in this lemma are under P_n and denoted by E_n, cov_n. Let*

$$T_n = n^{-1/2} \sum_{i=1}^{n} \{f_0(Z_i) + n^{-1/2} f_1(Z_i)\} + n^{-3/2} \sum \sum_{1 \le i < j \le n} f_2(Z_i, Z_j),$$

where f_0, f_1, f_2 denote $\mathbb{R}^2$-valued functions such that $E_n f_0(Z_1) = E_n f_1(Z_1) = 0$, $E_n(f_2(Z_1, Z_2)|Z_2) = 0$ a.s and $f_2(x, y)$ is symmetric. Assume there exist a constant $a > 0$ and continuous non-increasing functions $\chi_j : [0, \infty) \to [0, \infty)$, $j = 1, 2$, satisfying

$$\begin{aligned}
&\lim_{x \to \infty} \chi_1(x) = 0, \\
&0 < \chi_2(x) \le 1 \quad \textit{for every } x \ge a, \\
&E_n \|f_2(Z_1, Z_2)\|^3 I(\|f_2(Z_1, Z_2)\| \ge x) \le \chi_1(x) \quad \textit{for every } x > 0, \\
&|E_n \exp[it \circ f_0(Z_1)]| \le 1 - \chi_2(\|t\|) \quad \textit{for every } \|t\| \ge a > 0, \\
&E_n \|f_0(Z_1)\|^3 I(\|f_0(Z_1)\| \ge x) \le \chi_1(x) \quad \textit{for every } x > 0, \\
&E_n \|f_1(Z_1)\|^{3/2} I(\|f_1(Z_1)\| \ge x) \le \chi_1(x) \quad \textit{for every } x > 0.
\end{aligned}$$

If furthermore, $\lim_{n \to \infty} cov_n(f_0)$ is positive definite, then there exists a sequence $\epsilon_n \downarrow 0$ and a constant K depending on χ_j and $\lim_{n \to \infty} cov_n(f_0)$ only, such that

$$\sup_{(u,v) \in \mathbb{R}^2} |F_n(u, v) - G_n(u, v)| \le K \epsilon_n n^{-1/2}, \ n \ge 2,$$

where F_n is the distribution function of T_n and

$$G_n(u, v) = \int_{-\infty}^{u} \int_{-\infty}^{v} \{1 + \tfrac{1}{6} n^{-1/2} \kappa_3(-D)\} \varphi_{0, W_n}(x, y) dx dy,$$

with $W_n = cov_n(f_0 + n^{-1/2} f_1)$ and κ_3 the cumulant differential operator given by

$$\kappa_3(D) = E_n(f_0(Z_1) \circ D)^3 + 3E_n \Big\{ [f_2(Z_1, Z_2) \circ D][f_0(Z_1) \circ D][f_0(Z_2) \circ D] \Big\}.$$

□

To apply Lemma 4.2.2 we add the following regularity condition

(Rv) $$\begin{aligned}
&\lim_{x \to \infty} \sup_{(\theta, \tau) \in \Omega_0} E_{\theta, \tau} |\psi_{ij}(X_1)|^3 I(|\psi_{ij}(X_1)| \ge x) = 0, \text{ for } (i, j) = (0, 1), (1, 0), \\
&\lim_{x \to \infty} \sup_{(\theta, \tau) \in \Omega_0} E_{\theta, \tau} |k(X_1)|^3 I(|k(X_1)| \ge x) = 0,
\end{aligned}$$

$$\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau}|q(X_1)|^3 I(|q(X_1)|\geq x) = 0,$$
$$\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau}|r^*(X_1)|^3 I(|r^*(X_1)|\geq x) = 0,$$
$$\lim_{\|t\|\to\infty} \sup_{(\theta,\tau)\in\Omega_0} |E_{\theta,\tau}\exp[it_1\{\psi_{10}(X_1) - r_1\psi_{01}(X_1)\} + it_2\{\psi_{01}(X_1) - r_2\psi_{10}(X_1)\}]| < 1$$

for $(r_1, r_2) = (0,\rho), (\rho,\rho)$.

If k, q, and r are given in terms of functions ψ_{ij}, then the conditions for k, q and r may be replaced by conditions on the corresponding expectations for those functions. For example for the tests based on LR, MLE or Wald (see (4.2.5)), the condition on k can be replaced by the conditions

$$\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau}|\psi_{20}(X_1)|^3 I(|\psi_{20}(X_1)|\geq x) = 0,$$
$$\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau}|\psi_{10}^2(X_1)|^3 I(|\psi_{10}(X_1)|\geq x) = 0.$$

Note that

$$\sup_{(\theta,\tau)\in\Omega_0} |\psi_{ij}(x)|^3 f(x;\theta,\tau) \leq L_{ij}(x) \quad \text{with} \quad \int_A L_{ij}(x)d\mu(x) < \infty$$

implies (e.g. by the dominated convergence theorem) that

$$\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau}|\psi_{ij}(X_1)|^3 I(|\psi_{ij}(X_1)|\geq x) = 0.$$

This may be helpful in verifying (Rv). Condition (Rv) also guarantees the existence of the third moments of the functions given, in particular the existence of $E\psi_{10}^3$ and $E\psi_{01}^3$, which we needed already in the definition of the standard first-order optimal tests in (4.2.5).

The following condition, finally, takes care of the existence of the expectations used in the definition of the tests and in the calculations of the next section. It is needed in the next section to justify that the remainder terms in the calculation of moments under (θ,τ), are of the order given.

(Rvi) $$\sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau}\Big(|\psi_{10}(X_1)|^3 + |\psi_{01}(X_1)|^3 + |k(X_1)|^2 + |q(X_1)|^2 + |r^*(X_1)|^2 + 1\Big)|\psi_{ij}(X_1;\theta,\tau)| < \infty \quad \text{for } i,j = 0,1,2,3.$$

Note that the conditions for the functions q and r^* should hold for the q and r^* corresponding to SU as well as for those corresponding to TU.

4.3 Main results: difference in size and power

Our main aim in this chapter is to reveal the difference in size and power of the pre-test procedure and the one-stage test SU. We consider local alternatives of the form (θ_n, τ_n) with

$$\theta_n = bI_{11}^{-1/2}n^{-1/2},\ \tau_n = cI_{22}^{-1/2}n^{-1/2} \text{ with } b \geq 0 \text{ and } c \in \mathbb{R}. \tag{4.3.1}$$

(Note that θ_n and τ_n are described in $I_{11}^{-1/2}$ and $I_{22}^{-1/2}$ "units". In this way the redundant parameters I_{11} and I_{22} are absorbed and I_{12} reduces to ρ.) As usual in this kind of analysis the (composite) null hypothesis H_0 is represented by sequences $(0, cI_{22}^{-1/2}n^{-1/2})$, $\bar{H}_0$ by $(bI_{11}^{-1/2}n^{-1/2}, 0)$ and H_1 of course by $(bI_{11}^{-1/2}n^{-1/2}, cI_{22}^{-1/2}n^{-1/2})$.

We start with the Edgeworth expansions of the simultaneous distribution of (SK, TU) and (SU, TU) under (θ_n, τ_n).

Theorem 4.3.1 *Suppose that* (Ri)–(Rvi) *hold. Write*

$$U_{1n} = \frac{SK - \mu_{1n}(b,c)}{\sigma_{1n}(b,c)},\ U_{2n} = \frac{SU - \mu_{2n}(b,c)}{\sigma_{2n}(b)} \text{ and } V_n = \frac{TU - \mu_n(b,c)}{\sigma_n(c)}$$

with

$$\begin{aligned}
\mu_{1n}(b,c) &= b + c\rho + \tfrac{1}{2}\{b^2E\psi_{10}\psi_{20} + 2bcE\psi_{10}\psi_{11} + c^2E\psi_{10}\psi_{02} \\
&\qquad + 2(1+b^2)Ek\psi_{10} + 2bcEk\psi_{01}\}n^{-1/2}, \\
\sigma_{1n}(b,c) &= 1 + \tfrac{1}{2}\{bE\psi_{10}^3 + cE\psi_{10}^2\psi_{01} + 4bEk\psi_{10} + 2cEk\psi_{01}\}n^{-1/2}, \\
\mu_{2n}(b,c) &= b - \tfrac{1}{2}b\rho^2 + \tfrac{1}{2}\{b^2E\psi_{10}\psi_{20} + 2bcE\psi_{10}\psi_{11} - bcE\psi_{10}^2\psi_{01} \\
&\qquad - E\psi_{10}\psi_{02} + 2(1+b^2)Eq\psi_{10}\}n^{-1/2}, \\
\sigma_{2n}(b) &= 1 + \tfrac{1}{2}b(E\psi_{10}^3 + 4Eq\psi_{10})n^{-1/2}, \\
\mu_n(b,c) &= c - \tfrac{1}{2}c\rho^2 + \tfrac{1}{2}\{c^2E\psi_{01}\psi_{02} + 2bcE\psi_{01}\psi_{11} + b^2E\psi_{01}\psi_{20} \\
&\qquad - bcE\psi_{10}\psi_{01}^2 + 2(1+b^2)Er\psi_{10} + 2(1+c^2)Eq\psi_{01}\}n^{-1/2}, \\
\sigma_n(c) &= 1 + \tfrac{1}{2}c(E\psi_{01}^3 + 4Eq\psi_{01})n^{-1/2}.
\end{aligned} \tag{4.3.2}$$

Then, uniformly for $(u,v) \in \mathbb{R}^2$,

$$\begin{aligned}
&P_{\theta_n,\tau_n}(U_{in} \leq u, V_n \leq v) = \Phi(u,v;\rho_i) + n^{-1/2}\sum_{j=0}^{3} c_{ij}\Phi^{(j)}(u)\Phi^{(3-j)}(v) \\
&+ O(\rho^4 + \rho n^{-1/2}) + o(n^{-1/2})
\end{aligned} \tag{4.3.3}$$

with

$$\begin{aligned}&\rho_1 = n^{-1/2} r_1(b,c),\ \rho_2 = -\rho + n^{-1/2} r_2(b,c)\\ &\textit{for some functions } r_1(b,c),\ r_2(b,c) \textit{ satisfying}\\ &r_2(b,c) - r_1(b,c) = -\tfrac{1}{2} bE\psi_{10}^2\psi_{01} - bEk\psi_{01} - cE\psi_{10}\psi_{02},\end{aligned} \tag{4.3.4}$$

and

$$\begin{aligned}&c_{10} = c_{20},\ c_{11} - c_{21} = -\tfrac{1}{2}E\psi_{10}\psi_{02},\ c_{12} - c_{22} = -\tfrac{1}{2}E\psi_{10}^2\psi_{01} - Ek\psi_{01},\\ &c_{13} = -\tfrac{1}{6}E\psi_{10}^3 - Ek\psi_{10},\ c_{23} = -\tfrac{1}{6}E\psi_{10}^3 - Eq\psi_{10}.\end{aligned} \tag{4.3.5}$$

Hence (4.2.5) holds.

Proof. We prove (4.3.3) for U_{2n} in detail and make a few remarks about the U_{1n}-case. The proof for U_{1n} is quite similar and even slightly more simple. We apply Lemma 4.2.2 with $P_n = P_{\theta_n,\tau_n}$ and therefore we rewrite SU as follows (we write E_n for E_{θ_n,τ_n}), with $g = \psi_{10} - \rho\psi_{01}$ and hence $Eg^2 = 1 - \rho^2$:

$$\begin{aligned}SU = \Big[&n^{-1/2}\sum_{i=1}^n \{g(X_i) - E_n g\}\\ &+ E_n\psi_{10}\, n^{-1/2}\sum_{i=1}^n \{q(X_i) - E_n q\} + E_n q\, n^{-1/2}\sum_{i=1}^n \{\psi_{10}(X_i) - E_n\psi_{10}\}\\ &+ E_n\psi_{01}\, n^{-1/2}\sum_{i=1}^n \{r(X_i) - E_n r\} + E_n r\, n^{-1/2}\sum_{i=1}^n \{\psi_{01}(X_i) - E_n\psi_{01}\}\\ &+ n^{-3/2}\sum_{i=1}^n \{z(X_i) - E_n z\}\\ &+ n^{-3/2}\sum_{1\le i\ne j\le n}\sum \{\psi_{10}(X_i) - E_n\psi_{10}\}\{q(X_j) - E_n q\}\\ &+ n^{-3/2}\sum_{1\le i\ne j\le n}\sum \{\psi_{01}(X_i) - E_n\psi_{01}\}\{r(X_j) - E_n r\}\\ &+ n^{1/2}E_n g + n^{1/2}E_n\psi_{10}E_n q + n^{1/2}E_n\psi_{01}E_n r + n^{-1/2}E_n z\Big]\Big/(Eg^2)^{1/2},\end{aligned} \tag{4.3.6}$$

with

$$z = (\psi_{10} - E_n\psi_{10})(q - E_n q) + (\psi_{01} - E_n\psi_{01})(r - E_n r).$$

By direct calculation we get

$$\begin{aligned}
n^{1/2}E_n g &= b - b\rho^2 + \tfrac{1}{2}(b^2 E\psi_{10}\psi_{20} + 2bcE\psi_{10}\psi_{11} + c^2 E\psi_{10}\psi_{02})n^{-1/2} \\
&\quad + O(n^{-1} + \rho n^{-1/2}), \\
n^{1/2}E_n\psi_{10}E_n q &= (b^2 Eq\psi_{10} + bcEq\psi_{01})n^{-1/2} + O(n^{-1} + \rho n^{-1/2}) \\
n^{1/2}E_n\psi_{01}E_n r &= (bcEr\psi_{10} + c^2 Er\psi_{01})n^{-1/2} + O(n^{-1} + \rho n^{-1/2}) \\
n^{-1/2}E_n z &= (Eq\psi_{10} + Er\psi_{01})n^{-1/2} + O(n^{-1}).
\end{aligned}$$

Adding these, using $Er\psi_{10}+Eq\psi_{01} = -\frac{1}{2}E\psi_{10}^2\psi_{01}+O(\rho)$ and $Er\psi_{01} = -\frac{1}{2}E\psi_{10}\psi_{02}+O(\rho)$ (which follows immediately from r of (4.2.10)), and dividing by $(Eg^2)^{1/2}$ gives the expression for μ_{2n} in (4.3.2) up to order $O(\rho^4 + \rho n^{-1/2} + n^{-1})$. For the variance of the first three lines (the last four lines of SU do not contribute to the variance up to the order considered), we get

$$\begin{aligned}
&\mathrm{var}_n\,(g + (E_n\psi_{10})(q - E_n q) + (E_n q)(\psi_{10} - E_n\psi_{10}) \\
&\qquad + (E_n\psi_{01})(r - E_n r) + (E_n r)(\psi_{01} - E_n\psi_{01})) \\
&= Eg^2 + (bE\psi_{10}^3 + cE\psi_{10}^2\psi_{01})n^{-1/2} + 2n^{-1/2}bEq\psi_{10} + 2n^{-1/2}cEr\psi_{10} \\
&\quad + 2(bEq\psi_{10} + cEq\psi_{01})n^{-1/2} + O(n^{-1} + \rho n^{-1/2}) \\
&= Eg^2 + (bE\psi_{10}^3 + 4bEq\psi_{10})n^{-1/2} + O(n^{-1} + \rho n^{-1/2}),
\end{aligned} \tag{4.3.7}$$

using in the last step again that $Er\psi_{10} + Eq\psi_{01} + \frac{1}{2}E\psi_{10}^2\psi_{01} = O(\rho)$. Hence, we may write

$$\begin{aligned}
U_{2n} &= n^{-1/2}\sum_{i=1}^{n}\Big\{f_0^{(1)}(X_i) + n^{-1/2}f_1^{(1)}(X_i)\Big\} + n^{-3/2}\sum_{1\le i<j\le n}\sum f_2^{(1)}(X_i, X_j) \\
&\quad + O(\rho^4 + n^{-1} + \rho n^{-1/2})
\end{aligned} \tag{4.3.8}$$

with

$$\begin{aligned}
f_0^{(1)} &= (g - E_n g)/\sigma_n^*, \\
f_1^{(1)} &= \big\{(E_n\psi_{10})n^{1/2}(q - E_n q) + (E_n q)n^{1/2}(\psi_{10} - E_n\psi_{10}) \\
&\quad + (E_n\psi_{01})n^{1/2}(r - E_n r) + (E_n r)n^{1/2}(\psi_{01} - E_n\psi_{01}) + n^{-1/2}(z - E_n z)\big\}/\sigma_n^*, \\
f_2^{(1)}(x, y) &= \big\{(\psi_{10}(x) - E_n\psi_{10})(q(y) - E_n q) + (\psi_{10}(y) - E_n\psi_{10})(q(x) - E_n q) \\
&\quad + (\psi_{01}(x) - E_n\psi_{01})(r(y) - E_n r) + (\psi_{01}(y) - E_n\psi_{01})(r(x) - E_n r)\big\}/\sigma_n^*,
\end{aligned}$$

where

$$\sigma_n^* = (Eg^2)^{1/2}\sigma_{2n}(b).$$

Similarly, but now with the functions q and r corresponding to TU instead of to SU, we get

$$V_n = n^{-1/2}\sum_{i=1}^{n}\left\{f_0^{(2)}(X_i) + n^{-1/2}f_1^{(2)}(X_i)\right\} + n^{-3/2}\sum\sum_{1\le i<j\le n} f_2^{(2)}(X_i, X_j)$$

$$+ O(\rho^4 + n^{-1} + \rho n^{-1/2})$$

with $f_0^{(2)}, f_1^{(2)}, f_2^{(2)}$ obtained from $f_0^{(1)}, f_1^{(1)}, f_2^{(1)}$ by interchanging ψ_{10} and ψ_{01} (also in the function g), and replacing $\sigma_{2n}(b)$ in σ_n^* by $\sigma_n(c)$.

To apply Lemma 4.2.2 we verify the conditions of the lemma. It is easily seen that $E_n f_0 = E_n f_1 = E_n(f_2(X_1, X_2)|X_2) = 0$ and that f_2 is symmetric. The conditions on the moments and characteristic function follow from (Rv). Finally,

$$\lim_{n\to\infty} \mathrm{cov}_n(f_0) = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix},$$

which is positive definite. Using (4.3.7), direct calculation gives

$$W_n = \mathrm{cov}_n(f_0 + n^{-1/2}f_1) = \begin{pmatrix} 1 & -\rho + n^{-1/2}r_2(b,c) \\ -\rho + n^{-1/2}r_2(b,c) & 1 \end{pmatrix} + O(n^{-1} + \rho n^{-1/2}),$$

where

$$r_2(b,c) = \lim_{\rho\to 0}\lim_{n\to\infty}\big[n^{1/2}\{\mathrm{cov}_n(f_0^{(1)}, f_0^{(2)}) - \mathrm{cov}(f_0^{(1)}, f_0^{(2)})\} + \mathrm{cov}(f_0^{(1)}, f_1^{(2)}) + \mathrm{cov}(f_1^{(1)}, f_0^{(2)})\big].$$

By the similarity of U_{1n} and U_{2n} we get

$$r_2(b,c) - r_1(b,c) = \lim_{\rho\to 0}\lim_{n\to\infty}\mathrm{cov}(f_{1,U_{2n}}^{(1)} - f_{1,U_{1n}}^{(1)}, f_0^{(2)})$$
$$= \lim_{\rho\to 0}\lim_{n\to\infty}[(E_n\psi_{10})n^{1/2}(Eq\psi_{01} - Ek\psi_{01}) + (E_n\psi_{01})n^{1/2}(Er\psi_{01}) + (E_n r)n^{1/2}]$$
$$= -\tfrac{1}{2}bE\psi_{10}^2\psi_{01} - bEk\psi_{01} - cE\psi_{10}\psi_{02},$$

where the last equality follows from direct calculation, using the expressions for $f_1^{(1)}$ (corresponding to U_{2n} and U_{1n}, respectively) and $f_0^{(2)}$ (which corresponds to V_n), and substituting $Er\psi_{10} = -\frac{1}{2}E\psi_{10}^2\psi_{01} - Eq\psi_{01} + O(\rho)$ and $Er\psi_{01} = -\frac{1}{2}E\psi_{10}\psi_{02} + O(\rho)$ with r from (4.2.10).

Application of Lemma 4.2.2 and noting that replacement of E_n by E in κ_3 in the application of this lemma gives an error of order $O(n^{-1})$, yields (4.3.3) with

$$\sum_{j=0}^{3} c_{2j}\Phi^{(j)}(u)\Phi^{(3-j)}(v) = \int_{-\infty}^{u}\int_{-\infty}^{v} \tfrac{1}{6}\kappa_3(-D)\varphi(x)\varphi(y)dxdy,$$

where κ_3 is given by

$$\kappa_3(D) = E[(\psi_{10}(X_1), \psi_{01}(X_1)) \circ D]^3 + \kappa_3^*(D)$$

with

$$\kappa_3^*(D) = 3E\Big\{[f_2(X_1, X_2) \circ D][(\psi_{10}(X_1), \psi_{01}(X_1)) \circ D] \\ [(\psi_{10}(X_2), \psi_{01}(X_2)) \circ D]\Big\}.$$

Hence, noting that $f_2^{(2)}(x, y)$ is the same for U_{1n} and U_{2n}, we get

$$\sum_{j=0}^{3}(c_{2j} - c_{1j})\Phi^{(j)}(u)\Phi^{(3-j)}(v) = \int_{-\infty}^{u}\int_{-\infty}^{v} \tfrac{1}{6}\tilde{\kappa}_3(-D)\varphi(x)\varphi(y)dxdy$$

with

$$\tilde{\kappa}_3(D) = 3E\Big\{[\tilde{f}_2(X_1, X_2) \circ D][(\psi_{10}(X_1), \psi_{01}(X_1)) \circ D] \\ [(\psi_{10}(X_2), \psi_{01}(X_2)) \circ D]\Big\}$$

and

$$\tilde{f}_2(x, y) = (f^{(1)}_{2,U_{2n}}(x, y) - f^{(1)}_{2,U_{1n}}(x, y), 0).$$

In view of the marginal distributions it follows that $c_{10} = c_{20}$, and it remains to calculate $c_{21} - c_{11}$, $c_{22} - c_{12}$ and c_{13} and c_{23}. Direct calculation gives, up to $O(\rho + n^{-1/2})$, $c_{21} - c_{11} = \frac{1}{2}E\psi_{10}\psi_{02}$, $c_{22} - c_{12} = \frac{1}{2}E\psi_{10}^2\psi_{01} + Ek\psi_{01}$, $c_{13} = -\frac{1}{6}E\psi_{10}^3 - Ek\psi_{10}$ and $c_{23} = -\frac{1}{6}E\psi_{10}^3 - Eq\psi_{10}$. This completes the proof of Theorem 4.3.1. □

Now we can derive the critical values for the tests based on the given test statistics. The test based on SK is meant for testing $\theta = 0, \tau = 0$ against $\theta > 0, \tau = 0$. With $\sigma_{1n}(0, 0) = 1$ and $\mu_{1n}(0, 0)$ and c_{13} from (4.3.2) and (4.3.5), respectively, it follows that

$$P_{0,0}(SK \le u) = \Phi(u - \mu_{1n}(0, 0)) + n^{-1/2}c_{13}\Phi^{(3)}(u) + o(n^{-1/2}),$$

where it is easy to understand that under $(0, 0)$ we only get the remainder term $o(n^{-1/2})$, since SK and its distribution do not depend on ρ. Of course, this result

can also easily be obtained from standard Edgeworth expansions. Hence $\theta = 0, \tau = 0$ is rejected when

$$SK > u_\alpha - c_{13} n^{-1/2}(u_\alpha^2 - 1) + \mu_{1n}(0,0)$$

with $u_\alpha = \Phi^{-1}(1-\alpha)$, giving size $\alpha + o(n^{-1/2})$ under $(0,0)$.

The test based on SU is meant for testing H_0 against H_1. With $\sigma_{2n}(0) = 1$ and $\mu_{2n}(0,c)$ and c_{23} from (4.3.2) and (4.3.5), respectively, it follows from Theorem 4.3.1 that

$$P_{0,\tau_n}(SU \le u) = \Phi(u - \mu_{2n}(0,c)) + n^{-1/2} c_{23} \Phi^{(3)}(u) + O(\rho n^{-1/2}) + o(n^{-1/2})$$

and hence H_0 is rejected when

$$SU > u_\alpha - c_{23} n^{-1/2}(u_\alpha^2 - 1) + \mu_{2n}(0,c),$$

giving size $\alpha + O(\rho n^{-1/2}) + o(n^{-1/2})$ under $(0,\tau_n)$. Note that we dropped the $O(\rho^4)$-terms. These turned out to cancel under $(0,\tau_n)$.

The test based on TU is meant for testing $\bar{H}_0$ against $\bar{H}_1$. With $\sigma_n(0) = 1$, it follows from Theorem 4.3.1 that

$$P_{\theta_n,0}(TU \le v) = \Phi(v - \mu_n(b,0)) + n^{-1/2} c_{10} \Phi^{(3)}(v) + O(\rho n^{-1/2}) + o(n^{-1/2}).$$

Since $\mu_n(b,0) = O(n^{-1/2})$, the two-sided test which rejects $\bar{H}_0$ when

$$|TU| > u_{\delta/2},$$

gives size $\delta + O(\rho n^{-1/2}) + o(n^{-1/2})$ under $(\theta_n, 0)$. Note that $n^{-1/2}$-terms which occurred in the critical values for SK and SU are not necessary here, since a shift of order $n^{-1/2}$ of both upper and lower critical value in the same direction does not affect the size of the test.

The pre-test procedure is defined as to reject H_0 if

$$SK > u_\alpha + n^{-1/2}\left(\tfrac{1}{6} E\psi_{10}^3 (u_\alpha^2 - 1) + Ek\psi_{10} u_\alpha^2\right) \text{ and } |TU| \le u_{\delta/2}$$

or

$$SU > u_\alpha + n^{-1/2}\left(\tfrac{1}{6} E\psi_{10}^3 (u_\alpha^2 - 1) + Eq\psi_{10} u_\alpha^2 - \tfrac{1}{2} E\psi_{10}\psi_{02}\right) \text{ and } |TU| > u_{\delta/2}.$$

The probability that the pre-test procedure rejects, depends on the local alternatives (θ_n, τ_n) parameterized by b and c, and is denoted by $\pi^* = \pi^*(b,c)$. The probability that SU rejects will be denoted by $\tilde{\pi} = \tilde{\pi}(b,c)$. The next theorem is our main result. It gives an attractive expression for the difference $\pi^* - \tilde{\pi}$, from which much insight can be obtained for the comparison of the pre-test procedure with the test based on SU.

Theorem 4.3.2 provides a good approximation of $\pi^* - \tilde{\pi}$ if n is large and ρ is small. The results are *uniformly* valid in ρ and n, but of course they are only meaningful for

$n \to \infty$ and $\rho \to 0$. The latter is no serious restriction, because if ρ is not small the situation is clear: in that case the pre-test procedure is unacceptable (see Section 4.2). Moreover, Theorem 4.3.2 gives the rate of convergence, $O(\rho^3 + \rho n^{-1/2}) + o(n^{-1/2})$, which demonstrates the (high) accuracy of the approximation and its dependence on n and ρ. If we would add the terms of order $\rho n^{-1/2}$ to the approximation a large number of terms would come in. When we really would use these terms, the expression would become less insightful due to too many complications. The same holds for the ρ^3-terms. Ignoring terms of order $O(\rho^3 + \rho n^{-1/2}) + o(n^{-1/2})$ seems to be the right compromise between needed accuracy and transparency. A further illustration of the accuracy obtained in this way is seen in the numerical results of Section 4.4, while for the technical meaning of $O(\rho^3 + \rho n^{-1/2}) + o(n^{-1/2})$ we refer to the remark just before the proof of Lemma 4.2.1. Note that for $\rho = 0$ the error term in Theorem 4.3.2 reduces to $o(n^{-1/2})$.

Theorem 4.3.2 *Suppose that* (Ri)–(Rvi) *hold. Then*

$$\begin{aligned}
&\pi^*(b,c) - \tilde{\pi}(b,c) \\
&= \varphi(u_\alpha - b)\left\{h_1(c,u_{\delta/2})[\rho + \tfrac{1}{2}\rho^2\{(b/c) + (u_\alpha - b)c\} + m(c,u_\alpha)n^{-1/2}]\right. \\
&\qquad + h_2(c,u_{\delta/2})[\rho + \tfrac{1}{2}c\rho^2(u_\alpha - b) + m(c,u_\alpha)n^{-1/2}] \\
&\qquad \left. + h_3(c,u_{\delta/2})[\rho^2(u_\alpha - b) - E\psi_{10}\psi_{02}n^{-1/2}]\right\} \\
&+ O(\rho^3 + \rho n^{-1/2}) + o(n^{-1/2}) \quad \textit{as } \rho \to 0 \ \textit{ and } \ n \to \infty,
\end{aligned}$$

where

$$\begin{aligned}
h_1(x,y) &= x\{\Phi(y-x) - \Phi(-y-x)\}, \\
h_2(x,y) &= \varphi(y+x) - \varphi(y-x), \\
h_3(x,y) &= \tfrac{1}{2}y\{\varphi(y+x) + \varphi(y-x)\}, \\
m(x,y) &= \tfrac{1}{2}\left\{xE\psi_{10}\psi_{02} + y\left(E\psi_{10}^2\psi_{01} + 2Ek\psi_{01}\right)\right\}.
\end{aligned}$$

Proof. Let

$$\begin{aligned}
u_{1n} &= \{u_\alpha - c_{13}n^{-1/2}(u_\alpha^2 - 1) + \mu_{1n}(0,0) - \mu_{1n}(b,c)\}/\sigma_{1n}(b,c), \\
u_{2n} &= \{u_\alpha - c_{23}n^{-1/2}(u_\alpha^2 - 1) + \mu_{2n}(0,c) - \mu_{2n}(b,c)\}/\sigma_{2n}(b), \\
v_n^U &= \{u_{\delta/2} - \mu_n(b,c)\}/\sigma_n(c), \\
v_n^L &= \{-u_{\delta/2} - \mu_n(b,c)\}/\sigma_n(c).
\end{aligned}$$

with μ_{in} and σ_{in} from (4.3.2), and c_{13} and c_{23} from (4.3.5). Then

$$\pi^*(b,c) - \tilde{\pi}(b,c) = P_{\theta_n,\tau_n}(U_{1n} \le u_{1n}, V_n \le v_n^L) - P_{\theta_n,\tau_n}(U_{2n} \le u_{2n}, V_n \le v_n^L)$$

$$- \{P_{\theta_n\tau_n}(U_{1n} \le u_{1n}, V_n < v_n^U) - P_{\theta_n,\tau_n}(U_{2n} \le u_{2n}, V_n < v_n^U)\}.$$

By Theorem 4.3.1 we get (4.2.12). It is easily seen that the other conditions of Lemma 4.2.1 are satisfied with $u_0 = u_\alpha - b$, $v_0 = -u_{\delta/2} - c$ when v_n^L is used and $v_0 = u_{\delta/2} - c$ in case of v_n^U. Noting that

$$\begin{aligned} u_{1n} - u_{2n} = &- c\rho - \tfrac{1}{2}b\rho^2 + (u_0^2 - 1)(Ek\psi_{10} - Eq\psi_{10})n^{-1/2} \\ &- cm(c, u_\alpha)n^{-1/2} + O(n^{-1} + \rho n^{-1/2}) \end{aligned}$$

and

$$\rho_1 - \rho_2 = \rho + n^{-1/2}\{\tfrac{1}{2}bE\psi_{10}^2\psi_{01} + bEk\psi_{01} + cE\psi_{10}\psi_{02}\},$$

straightforward calculation gives the result. □

It is remarkable that $\pi^*(b,c) - \tilde{\pi}(b,c)$ does not depend (up to the considered order) on the q's and r's occurring in SU and TU and only through $Ek\psi_{01}$ on k. This property is related to the phenomenon that "first-order efficiency implies second-order efficiency", cf. Bickel, Chibisov, and van Zwet (1981) or Pfanzagl (1980), p. 62. Note that due to this phenomenon for testing $\theta = 0$ the power at $(bI_{11}n^{-1/2}, 0)$ of SK is the same up to order $n^{-1/2}$, irrespective of the choice of k.

For the tests mentioned in Section 4.2 we get, ignoring ρ-terms, cf. (4.2.5),

$$\begin{aligned} \text{LMP} &: Ek\psi_{01} = 0 \\ \text{LR} &: Ek\psi_{01} = \tfrac{1}{2}(E\psi_{20}\psi_{01} - E\psi_{10}^2\psi_{01}) \\ \text{MLE and Wald} &: Ek\psi_{01} = E\psi_{20}\psi_{01} - E\psi_{10}^2\psi_{01}. \end{aligned} \tag{4.3.9}$$

A special case which is of particular interest is when $\rho = 0$. The following corollary gives this as an immediate consequence of Theorem 4.3.2.

Corollary 4.3.3 *Suppose that* (Ri)–(Rvi) *hold and that* $\rho = 0$. *Then*

$$\pi^*(b,c) - \tilde{\pi}(b,c)$$

$$= \varphi(u_\alpha - b)[h(c, u_{\delta/2})m(c, u_\alpha) - h_3(c, u_{\delta/2})E\psi_{10}\psi_{02}]n^{-1/2} + o(n^{-1/2})$$

as $n \to \infty$, *where* $h = h_1 + h_2$.

4.4 Consequences for the actual size and power

Under the null hypothesis $H_0 : \theta = 0$, the difference $\pi^* - \tilde{\pi}$ reduces to $\pi^*(0, c) - \alpha + o(n^{-1/2})$. Hence the departure from the nominal level of the pre-test procedure follows from Theorem 4.3.2 and approximately equals

$$\begin{aligned}\varphi(u_\alpha)\big[h(c, u_{\delta/2})\left\{\rho + \tfrac{1}{2}u_\alpha c\rho^2 + m(c, u_\alpha)n^{-1/2}\right\} \\ + h_3(c, u_{\delta/2})\left\{u_\alpha\rho^2 - E\psi_{10}\psi_{02}n^{-1/2}\right\}\big],\end{aligned} \tag{4.4.1}$$

where $h = h_1 + h_2$. Firstly, we note that the family of distributions and classes of tests are involved through only four parameters: ρ, $E\psi_{10}\psi_{02}$, $E\psi_{10}^2\psi_{01}$ and $Ek\psi_{01}$. If ρ tends to 0 and $n \to \infty$, the error $\pi^*(0, c) - \alpha$ tends to 0. Also for $\alpha \downarrow 0$ the error tends to 0.

Secondly, we analyze the behavior of h and h_3. They are shown in Figure 4.4.1. As $h(c, u_{\delta/2})$ is odd in c and $h_3(c, u_{\delta/2})$ is even in c, we only consider $c \geq 0$. The function h_3, to begin with, is strictly positive for all c. The function h increases in $u_{\delta/2}$, and since $h(c, 0) = 0$ for all c, we have $h(c, u_{\delta/2}) \geq 0$ for all $c \geq 0$ and $u_{\delta/2} \geq 0$. Since $h(0, u_{\delta/2}) = 0$, $\lim_{c\to\infty} ch(c, u_{\delta/2}) = 0$ and $\lim_{c\to\infty} h_3(c, u_{\delta/2}) = 0$, there exists a $c^* = c^*(u_{\delta/2})$ for which the error, given by (4.4.1), is maximal.

If we ignore the ρ^2-terms (which includes of course the case $\rho = 0$ presented in Corollary 4.3.3), and assume that $E\psi_{10}\psi_{02} = 0$ (which holds in many examples, see Sections 4.4.1 and 4.4.2), then (4.4.1) reduces to

$$\varphi(u_\alpha)h(c, u_{\delta/2})\left\{\rho + u_\alpha\left(\tfrac{1}{2}E\psi_{10}^2\psi_{01} + Ek\psi_{01}\right)n^{-1/2}\right\}, \tag{4.4.2}$$

which depends on c only through the function h. Expression (4.4.2) may be interpreted as follows. If SK is used without a preliminary test, we have $\delta = 0$ and $h(c, u_{\delta/2}) = c$ and hence the error $\pi^*(0, c) - \alpha$ grows linearly in that case. For $\delta > 0$ the function c is replaced by the re-descending function h, sketched for $\delta = 0.05$ in Figure 4.4.1. Note that the function h consists of two parts. The first part is the function h_1 which equals c multiplied by the first-order approximation of the acceptance probability of the preliminary test. This acceptance probability is at most 1 and hence a first reduction of the error is obtained. The larger c, the smaller the acceptance probability, and thus the larger the reduction. The second part is the function h_2, which is negative for $c \geq 0$ and gives a further reduction. Hence the replacement of c by $h(c, u_{\delta/2})$ indeed yields a substantial reduction of the error, as is seen in the figure.

Again ignoring terms of order ρ^2, it is immediately seen from Theorem 4.3.2 that the power difference is nothing but the size difference, inflated by a factor $\varphi(u_\alpha - b)/\varphi(u_\alpha)$. For $\alpha = 0.05$ this factor runs from 1 at $b = 0$ to its maximal value 3.9 at $b = u_\alpha$ and, being a multiple of φ, it then decreases. Here, it is seen that the idea mentioned in the introduction of getting higher power due to "knowing the value of the nuisance parameter" does not come true. If there is a gain in power, it is due to the difference between size and level, possibly blown up by some factor.

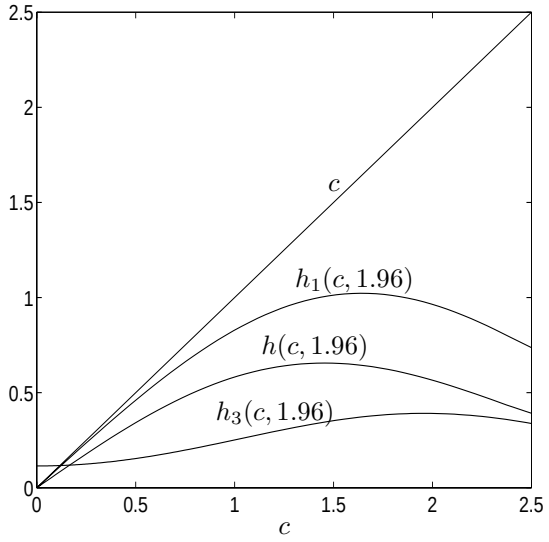

Figure 4.4.1 Functions c, h_1, h and h_3 as a function of c for $\delta = 0.05$.

Moreover, since $h(c, u_{\delta/2})$ is odd in c, a gain in power at c will as a rule imply a loss in power at $-c$. An exception should be made for (very) small values of c, because in that case ρ^2-terms are dominant and should not be ignored. Consider the special case $c = 0$. The approximation given by Theorem 4.3.2 now reads as

$$\begin{aligned}\varphi(u_\alpha - b)\,\big[\tfrac{1}{2}b\rho^2\left\{\Phi(u_{\delta/2}) - \Phi(-u_{\delta/2}) - 2u_{\delta/2}\varphi(u_{\delta/2})\right\}\\ + u_{\delta/2}\varphi(u_{\delta/2})\{u_\alpha\rho^2 - E\psi_{10}\psi_{02}n^{-1/2}\}\big]\,.\end{aligned} \tag{4.4.3}$$

Since the coefficient of $b\rho^2$ is positive for all $u_{\delta/2}$, the term with $b\rho^2$ gives some (small) gain in power not due to the difference between size and level. Note however that this does not mean that the size may not exceed α. If $E\psi_{10}\psi_{02} = 0$ or small enough, then the second line of (4.4.3) is always positive due to the fact that $u_{\delta/2}\varphi(u_{\delta/2})$ is strictly positive, and hence the size is certainly larger than α.

In the remainder of this section we consider some examples, gathered in two important classes: two-parameter exponential families and symmetric location-scale families. In all examples we use LMP tests in which the nuisance parameter, if present, is estimated by the MLE in the unrestricted model.

In Example 4.4.1 numerical calculations are performed, while in the other examples simulation results are presented. Each of the simulations is based on 100 000 repetitions. Hence, the standard deviations of the simulated power differences are at

most $(100\,000)^{-1/2} = 0.0032$. This reduces to 0.0019 if at least one of the probabilities π^* or $\tilde{\pi}$ is at most 0.05 and to 0.001 if both probabilities are at most 0.05.

4.4.1 Two-parameter exponential families

Let μ be some probability measure on $\mathbb{R}^2$ and assume that the moment generating function of μ exists in some open neighborhood Ω_0 of $(0,0)$. With respect to this measure, define a family of densities

$$f(x;\theta,\tau) = \exp\Big\{\theta x^{(1)} + \tau x^{(2)} - \omega(\theta,\tau)\Big\} \tag{4.4.4}$$

with $x = (x^{(1)}, x^{(2)}) \in \mathbb{R}^2$ and ω a normalizing constant. This representation is a standard representation (cf. Barndorff-Nielsen (1978), p. 115) of the densities of the two-parameter exponential family with respect to the measure μ. It is easily seen that the regularity conditions (Ri)-(Riv) hold. Assume that also the regularity conditions (Rv) and (Rvi) hold. From exponential family theory it follows rather easily that

$$\begin{aligned}
\psi_{10}(x) &= \frac{x^{(1)} - EX^{(1)}}{\sqrt{\mathrm{var}X^{(1)}}}, \qquad & \psi_{01}(x) &= \frac{x^{(2)} - EX^{(2)}}{\sqrt{\mathrm{var}X^{(2)}}}, \\
\psi_{20}(x) &= \frac{(x^{(1)} - EX^{(1)})^2}{\mathrm{var}X^{(1)}} - 1, \quad & \psi_{02}(x) &= \frac{(x^{(2)} - EX^{(2)})^2}{\mathrm{var}X^{(2)}} - 1, \\
\psi_{11}(x) &= \frac{(x^{(1)} - EX^{(1)})(x^{(2)} - EX^{(2)})}{\sqrt{\mathrm{var}X^{(1)}\mathrm{var}X^{(2)}}} - \rho,
\end{aligned} \tag{4.4.5}$$

see for example Lehmann and Casella (1998), Sec. 1.5.

First note that all these functions may be written in terms of ψ_{10} and ψ_{01}. This may greatly simplify the expressions for k, q and r which determine the test statistics. Since $\psi_{20} = \psi_{10}^2 - 1$, the first part of the expressions for k cancels for the LR, MLE and Wald's test (see (4.2.5)) and only the term with ψ_{10} remains. For Wald's test this term also vanishes, since $E\psi_{10}\psi_{20} = E\psi_{10}^3$. For the quantity $Ek\psi_{01}$ we get $Ek\psi_{01} = 0$ (up to order ρ) for LMP, LR, MLE and Wald's test, cf. also (4.3.9). This means that for the two-parameter exponential family, our approximation from Theorem 4.3.2 yields the same result for all these tests!

In the following examples, we compare the approximation given by Theorem 4.3.2 with simulation results for the pre-test procedure with LMP tests. In the LMP case we have $k = q = 0$, and using the relations $\psi_{11} = \psi_{10}\psi_{01} - \rho$ and for the expectations $E\psi_{11}\psi_{01} = E\psi_{10}\psi_{02} = E\psi_{10}\psi_{01}^2$ and $E\psi_{11}\psi_{10} = E\psi_{01}\psi_{20} = E\psi_{10}^2\psi_{01}$, it follows that r from (4.2.11) reduces (up to order ρ) to

$$r = -\tfrac{1}{2}(E\psi_{10}\psi_{01}^2)\psi_{01} - \tfrac{1}{2}(E\psi_{10}^2\psi_{01})\psi_{10}.$$

Note that the ignored term with coefficient $-\frac{1}{2}d_1\rho$ indeed does not contribute to the final result, in which terms of order $\rho n^{-1/2}$ are ignored. The function r is (up to order

ρ) the same both for SU and TU, since interchange of ψ_{10} and ψ_{01} does not make any difference.

Example 4.4.1 Problem of the reference set for normal means. Suppose we have a sample from a normal distribution with expectation θ and variance 1. Further, a second sample is available from a normal distribution, with (known) variance v^2, but we are not sure whether this second sample has the same expectation as the first one. We denote this expectation by $\theta - \tau$. If the second sample has the same mean, we might want to use it together with the first sample in order to test whether $\theta = 0$. A preliminary test is performed to investigate the equality of the means.

Arnold (1970) numerically analyzed the pre-test procedure for this example with $v = 1$, $\alpha = 0.05$ and $\delta = 0.25$, showing it to be biased. He compares the pre-test procedure with the (unbiased) generalized likelihood ratio test with the same size, and concludes that the latter is more powerful in a rather large part of the parameter space. He furthermore notes that biasedness is not only a feature in this example, but may be encountered frequently in pre-test procedures. He recommends to make power comparisons between the pre-test procedure and available unbiased tests before deciding to use the pre-test procedure.

If Y has a $N(\theta, 1)$-distribution and Z a $N(\theta - \tau, v^2)$-distribution, independent of Y, then their joint density with respect to the Lebesgue measure on $\mathbb{R}^2$ equals

$$\begin{aligned}&\frac{1}{2\pi v}\exp\Big\{-\frac{1}{2}(y-\theta)^2 - \frac{1}{2}\left(\frac{z-\theta+\tau}{v}\right)^2\Big\}\\ &= C(\theta,\tau)\exp\Big\{\theta\left(y+\frac{z}{v^2}\right)+\tau\left(-\frac{z}{v^2}\right)\Big\}\,g(y,z)\end{aligned}$$

for some function g and a normalizing constant C. Hence, in the standard representation (4.4.4) the probability measure μ corresponds to the distribution of $(X^{(1)}, X^{(2)}) = (Y + Z/v^2, -Z/v^2)$ under $(\theta, \tau) = (0, 0)$, i.e. for a measurable set $B \in \mathbb{R}^2$, $\mu(B) = Pr(X \in B)$ with $X = (X^{(1)}, X^{(2)})$ distributed according to the $N(0, 0, 1 + 1/v^2, 1/v^2, 0)$-distribution.

Here we have $\rho = -1/\sqrt{v^2+1}$ and hence the larger v, the smaller $|\rho|$. Furthermore, $E\psi_{10}^2\psi_{01} = E\psi_{10}\psi_{01}^2 = 0$, and hence also $r = 0$ for SU and TU. Note that $SU = n^{-1/2}\sum_{i=1}^n\{X_i^{(1)} + X_i^{(2)}\}$. This means that SU is the standardized sample mean of the first sample, as it should be, since for unknown τ the second sample is of no use for testing about θ.

Because SK, SU and TU are exactly normally distributed, it is no surprise that the $n^{-1/2}$-terms cancel in this case. Application of Theorem 4.3.2 yields

$$\begin{aligned}\pi^*(b,c) - \tilde{\pi}(b,c) = \varphi(u_\alpha - b)\big[&h(c,u_{\delta/2})\left\{\rho + \tfrac{1}{2}(u_\alpha - b)c\rho^2\right\}\\ &+ h_1(c,u_{\delta/2})\tfrac{1}{2}(b/c)\rho^2\\ &+ h_3(c,u_{\delta/2})(u_\alpha - b)\rho^2\big] + O(\rho^3).\end{aligned} \tag{4.4.6}$$

Suppose that we want to know the error of the size of the pre-test procedure if the nominal level $\alpha = 0.05$, $c = 1$, $v = 4$ and $\delta = 0.05$. Approximation (4.4.6) with $b = 0$ yields -0.0092. The numerical value of $\pi^*(0, c) - \alpha$ equals -0.0099. Furthermore, consider the power difference $\pi^*(b, c) - \tilde{\pi}(b, c)$ if the nominal level $\alpha = 0.05$, $c = -1$, $b = 1$, $v = 5$ and $\delta = 0.05$. Approximation (4.4.6) yields 0.0466. The numerical value of $\pi^*(1, -1) - \tilde{\pi}(1, -1)$ equals 0.0473. This illustrates the accuracy of the approximations. □

Example 4.4.2 Gamma distribution. In this example the main testing problem concerns the scale parameter in an exponential distribution. However, we are not sure about the model and therefore the idea is to perform a preliminary test to decide whether the exponential distribution is appropriate against the alternative of gamma-distributions.

The gamma-density with scale parameter $1 - \theta$ and second parameter $\tau + 1$ reads as

$$\frac{(1-\theta)^{\tau+1} z^{\tau} \exp\{-(1-\theta)z\}}{\Gamma(\tau+1)} = \frac{(1-\theta)^{\tau+1}}{\Gamma(\tau+1)} \exp\{\tau \log z + \theta z - z\}$$

with respect to the Lebesgue measure. For $\tau = 0$ this reduces to an exponential density with parameter $(1 - \theta)$. The measure μ in this example corresponds to the distribution of $(X^{(1)}, X^{(2)}) = (Z, \log Z)$ on $\mathbb{R}^2$ with Z exponentially distributed with parameter 1. In this example we have

$$\rho = \mathrm{cov}(Z, \log Z) / \sqrt{\mathrm{var} Z \mathrm{var}(\log Z)} = \sqrt{6}/\pi = 0.78.$$

As argued in Section 4.2, this large value of the correlation parameter ρ indicates that the size of the pre-test procedure is unacceptable. Indeed, simulation of the pre-test procedure with LMP tests for $n = 25$, $\alpha = 0.05$, $\delta = 0.05$ and $c = 1$, yields a simulated value of $\pi^*(0, 1)$ equal to 0.1808, which differs too much from the prescribed level. □

Example 4.4.3 Normal one-sample problem. Suppose we have a sample from a normal distribution. The main testing problem is to test expectation 0 against a positive expectation. A preliminary test is performed to test the hypothesis that the variance equals 1 against the alternative of a variance not equal to 1. For convenience we start with a $N(\nu, \sigma^2)$-distribution. Then the density with respect to Lebesgue measure equals

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{z-\nu}{\sigma}\right)^2\right\} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} z^2 + \frac{\nu}{\sigma^2} z - \frac{\nu^2}{2\sigma^2}\right\}.$$

Reparameterizing such that $(\theta, \tau) = (0, 0)$ corresponds with $(\nu, \sigma) = (0, 1)$, we get $\theta = \nu/\sigma^2$, $\tau = \frac{1}{2}(1 - 1/\sigma^2)$. Hence $f(x; \theta, \tau)$ represents a normal distribution with expectation $\theta/(1-2\tau)$ and variance $(1-2\tau)^{-1}$, written in standard form with respect to the measure μ, which corresponds to the distribution of $(X^{(1)}, X^{(2)}) = (Z, Z^2)$ on $\mathbb{R}^2$ when Z is standard normally distributed.

In view of (4.4.5) we get $\psi_{10} = X^{(1)}$ and $\psi_{01} = (X^{(2)} - 1)/\sqrt{2}$. Since $EZ^3 = 0$, we have $\rho = 0$ and moreover, $E\psi_{10}\psi_{01}^2 = 0$, because $EZ^5 = 0$. Further, $E\psi_{10}^2\psi_{01} = EZ^2(Z^2-1)/\sqrt{2} = \sqrt{2}$ and hence $r = -\psi_{10}/\sqrt{2}$. Substituting this in the expressions for SK, SU and TU gives back the Gauss test for SK, an asymptotically equivalent form of the one-sample t-test for SU and the well-known χ^2-test statistic for TU.

Application of Corollary 4.3.3 gives

$$\pi^*(b,c) - \tilde{\pi}(b,c) = \varphi(u_\alpha - b)h(c, u_{\delta/2})u_\alpha/\sqrt{2n} + o(n^{-1/2}), \tag{4.4.7}$$

which approximation was also presented in Theorem 3.2.1. (Note that there the local alternatives were parameterized as $\nu = bn^{-1/2}$, $\sigma = 1 + c(2n)^{-1/2}$, which leads through the reparameterization to $\theta = \nu/\sigma^2 = bn^{-1/2} + O(n^{-1})$ and $\tau = \frac{1}{2}(1 - 1/\sigma^2) = c(2n)^{-1/2} + O(n^{-1})$, corresponding to $bI_{11}^{-1/2}n^{-1/2}$ and $cI_{22}^{-1/2}n^{-1/2}$, respectively, in this example.)

Suppose we want to know the error in size if $\alpha = 0.05$, $n = 50$, $c = -1$ and $\delta = 0.05$. Application of (4.4.7) with $b = 0$ yields -0.0099 while the simulated value equals -0.0113. Next consider the power if $\alpha = 0.05$, $n = 50$, $c = 1$, $b = 2$ and $\delta = 0.05$. Approximation (4.4.7) gives 0.0359 and the simulated value equals 0.0462.

Finally, suppose that we want to know how wrong the size can be if $\alpha = 0.05$, $n = 50$ and $\delta = 0.05$. Then we have to deal with $\max_c h(c, 1.96)$. The maximum equals 0.6581 and is attained at $c = 1.4583$, cf. also Figure 4.4.1. Therefore, according to (4.4.7), the error maximized over c equals 0.0112. The simulated value of $\pi^*(0,c) - \tilde{\pi}(0,c)$ at $c = 1.4583$ is 0.0157, while the simulated value of $\pi^*(0,c) - \alpha$ at $c = 1.4583$ equals 0.0024. Consequently, there is a rather large discrepancy between the simulated value of $\tilde{\pi}(0,c)$, which equals $0.0524 - 0.0157 = 0.0367$, and the level $\alpha = 0.05$ which should be achieved. Note that $c = 1.4583$ corresponds to 1.4118 as value of the variance of the normal distribution. Hence the variance is rather far away from 1. The resulting conservatism of the test based on SU is also the reason for the difference between the approximated and simulated value of $\pi^*(0,c) - \tilde{\pi}(0,c)$. If we replace SU by the (exact) t-test in the pre-test procedure, $\pi^*(0,c) - \alpha$ can be calculated numerically. Then the result is 0.0119 and the approximation according to (4.4.7), 0.0112, is quite close to it. □

It is seen from the preceding examples that many interesting situations can be written in the form of a two-parameter exponential family. This way of presenting makes application very easy.

4.4.2 Symmetric location-scale families

Let f_0 be a given probability density with respect to the Lebesgue measure on $\mathbb{R}$ with $f_0(x) > 0$ for all $x \in \mathbb{R}$. Consider the location-scale family defined by

$$f(x;\theta,\tau) = \frac{1}{1+\tau} f_0\left(\frac{x-\theta}{1+\tau}\right)$$

with (θ, τ) in some open neighborhood of $(0, 0)$. Suppose in addition that f_0 is symmetric: $f_0(x) = f_0(-x)$ for all $x \in \mathbb{R}$. Finally, assume that (Ri)–(Rvi) hold.

The main testing problem in the pre-test procedure concerns the location parameter θ. The preliminary test is used to determine whether it may be assumed that the scale parameter $1 + \tau$ equals 1.

Direct calculation gives

$$\begin{aligned}
\psi_{10}^*(x) &= -\frac{f_0'}{f_0}(x), & \psi_{01}^*(x) &= -1 - x\frac{f_0'}{f_0}(x), \\
\psi_{20}^*(x) &= \frac{f_0''}{f_0}(x), & \psi_{02}^*(x) &= 2 + 4x\frac{f_0'}{f_0}(x) + x^2\frac{f_0''}{f_0}(x), \\
\psi_{11}^*(x) &= 2\frac{f_0'}{f_0}(x) + x\frac{f_0''}{f_0}(x).
\end{aligned} \tag{4.4.8}$$

The symmetry of f_0 implies that f_0'/f_0 is odd and f_0''/f_0 even, and therefore ψ_{10}^* and ψ_{11}^* are odd, while ψ_{01}^*, ψ_{20}^* and ψ_{02}^* are even. This implies in the first place that $\rho = 0$ and $E\psi_{10}\psi_{02} = 0$. Hence, Corollary 4.3.3 can be applied with $E\psi_{10}\psi_{02} = 0$, which leads to the simple approximation

$$\begin{aligned}
\pi^*(b, c) - \tilde{\pi}(b, c) = \tfrac{1}{2}\varphi(u_\alpha - b)h(c, u_{\delta/2})u_\alpha \left(E\psi_{10}^2\psi_{01} + 2Ek\psi_{01}\right) n^{-1/2} \\
+ o(n^{-1/2}).
\end{aligned} \tag{4.4.9}$$

This means that the approximation for the difference in a general symmetric location-scale family is the same as that for the normal case (see (4.4.7)), corresponding to $f_0(x) = \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$, except for the multiplicative constant $E\psi_{10}^2\psi_{01} + 2Ek\psi_{01}$, which may differ from family to family. For the normal distribution $E\psi_{10}^2\psi_{01} = \sqrt{2}$.

Furthermore, due to the oddness of ψ_{10} and ψ_{11} and evenness of the other ψ's, we have in addition to $\rho = E\psi_{10}\psi_{02} = 0$ that $E\psi_{10}^3 = E\psi_{10}\psi_{01}^2 = E\psi_{10}\psi_{20} = E\psi_{11}\psi_{01} = 0$. Again this may reduce the calculations of expectations occurring in the test statistics and critical values.

For the LMP tests which are used in the following examples, we have already $k = q = 0$, and get, using the simplifications mentioned

$$\begin{aligned}
&\text{for } SU\text{: } r = \psi_{11} - \psi_{10}\psi_{01} - \tfrac{1}{2}(2E\psi_{11}\psi_{10} - E\psi_{10}^2\psi_{01})\psi_{10} \\
&\text{for } TU\text{: } r = \psi_{11} - \psi_{10}\psi_{01} + \tfrac{1}{2}E(-2\psi_{11}\psi_{10} - \psi_{01}\psi_{20} + 2\psi_{01}\psi_{10}^2)\psi_{10}
\end{aligned}$$

for the symmetric location-scale family.

Example 4.4.4 Logistic distribution. Let f_0 be the logistic distribution, i.e.

$$f_0(x) = e^{-x}(1 + e^{-x})^{-2}.$$

Standardizing ψ_{10}^* and ψ_{01}^* from (4.4.8), we get

$$\begin{aligned}
\psi_{10}(x) &= \sqrt{3}(1 - e^{-x})(1 + e^{-x})^{-1}, \\
\psi_{01}(x) &= 3(\pi^2 + 3)^{-1/2}\{x - 1 - (x + 1)e^{-x}\}(1 + e^{-x})^{-1}.
\end{aligned}$$

It follows that

$$E\psi_{10}^2\psi_{01} = 3(\pi^2+3)^{-1/2} \tag{4.4.10}$$

and hence the power or size difference of (4.4.9) in the logistic case is $3(\pi^2+3)^{-1/2}/\sqrt{2} = 0.591$ times the difference in the normal case (cf. (4.4.7)).

Suppose we want to know how large we should take δ in order that the relative error of the size $(\pi^*(0,c)-\alpha)/\alpha$ is at most ε for some given $\varepsilon > 0$ (e.g. $\varepsilon = 0.2$ leading to 2% or 3% if $\alpha = 0.025$) for all c. Remember that $c^* = c^*(u_{\delta/2})$ is the c that maximizes the approximation of $\pi^*(0,c)-\alpha$, given in (4.4.1). In view of (4.4.9) c^* is the c that maximizes $h(c, u_{\delta/2})$ for given δ. Let $h^*(u_{\delta/2}) = h(c^*, u_{\delta/2})$ be the maximum value. Note that, as h itself decreases in δ, so does h^*. By (4.4.9) and (4.4.10), δ should be sufficiently large to ensure

$$h^*(u_{\delta/2}) \le \frac{2\varepsilon\alpha(\pi^2+3)^{1/2}n^{1/2}}{3u_\alpha\varphi(u_\alpha)}\,. \tag{4.4.11}$$

To evaluate $h^*(u_{\delta/2})$ we use the same further approximation as on p. 33, namely $h^*(x) = \frac{3}{5}x - \frac{1}{2}$. Taking $n = 25$, $\alpha = 0.025$ and $\varepsilon = 0.2$ and using this approximation in (4.4.11), we get $\frac{3}{5}u_{\delta/2} - \frac{1}{2} \le 0.5220$, which yields $\delta \ge 0.089$. The simulated value of $\pi^*(0,c)$ at $\delta = 0.089$ and $c = u_{\delta/2} - \frac{3}{2}\{1 - 5/(4u_{\delta/2})\} = 1.3041$ (which is the approximate solution for c^* according to p. 33), equals 0.0303, just as it should be.□

Example 4.4.5 Normal mixture. Let f_0 be a mixture of two normal distributions, as often used e.g. in robustness studies (cf. Huber (1981), p. 2)

$$f_0(x) = 0.95\varphi(x) + 0.05\varphi(x/3)/3.$$

We get

$$E\psi_{10}^2\psi_{01} = 1.104$$

and hence the difference in this mixture model is $1.104/\sqrt{2} = 0.781$ times the difference in the normal case.

Suppose we want to know how large we should take n in order that the (absolute) error of the size is at most 0.01 when $c = 1.5$, $\alpha = 0.05$ and $\delta = 0.05$. Inserting this in (4.4.9) with $b = 0$ we get $0.0616\, n^{-1/2} \le 0.01$ implying $n \ge 37.9$. The simulated value for $n = 38$ of $\pi^*(0, 1.5)$ equals 0.0673 and that of $\pi^*(0,1.5) - \tilde{\pi}(0,1.5)$ is 0.0111, which is close to the required 0.01. □

It is seen from the examples that pretty good and very useful answers are achieved for many questions in an easy way using the approximations given in Theorem 4.3.2 and Corollary 4.3.3.

Chapter 5

Power gain by pre-testing?

5.1 Introduction

Pre-test procedures, consisting of a preliminary test followed by either a basic main test or, in case of rejection, by an alternative, second main test, are often used for one of the following two reasons. First of all, in practice people strongly prefer simple procedures and hence like to stick to the basic main test as long as possible. Only if this causes really unacceptable deviations in the size of the main test (which should be detected by the preliminary test), one shifts to the alternative main test, which may be more complicated. The second nice aspect is that in principle a higher power can be achieved if stronger assumptions can be made. Hence again, if possible one wants to apply the basic main test, because one feels that always using the robust second main test implies an unnecessary loss of power in case the model assumptions for the basic main test hold. As a typical example, consider the use of a preliminary test to choose between the t-test and the sign test for testing the main hypothesis of a zero median. The main issue of this chapter is to investigate in what kind of situations such power gain by the pre-test procedure indeed comes true.

First, let us summarize what we have done so far. In Chapters 2 and 3, we studied the behavior of the pre-test procedure for the normal one- and two-sample case. The one-sample case can be considered as a special case of the more general situation treated in Chapter 4. There we considered a parametric family of densities $f(x;\theta,\tau)$, with θ the parameter of interest and τ the nuisance parameter. The preliminary test concerns the nuisance parameter τ and is, depending on its outcome, followed by a main test on θ either in a restricted family with the given value of τ, or in the complete family where τ is unknown. For this situation, a general class of tests (including the standard first-order optimal tests) was considered for each of the three testing problems in the pre-test procedure. We obtained transparent expressions for both the deviation in size and the gain in power of the pre-test procedure compared to the main test in the complete family. These expressions demonstrated that in general no substantial power gain can be realized without unacceptable violation of the

prescribed level. The key to the explanation of this phenomenon is the correlation between the test statistic of the preliminary test on the one hand, and the test statistics of the main tests on the other hand (represented by a correlation parameter ρ). Wild and uncontrollable variation in the size of the pre-test procedure can only be avoided if this correlation is really small (like in the previously studied normal case, where in fact it is zero). But then the two main tests, being first-order optimal in their respective models, are quite close, and their power difference will be small.

Hence, in this situation the virtue of the pre-test procedure is really largely based on the aforementioned simplicity: usually no rejection by the preliminary test occurs (at least, that is what one hopes) and one can stick to the simple basic main test. Only if rejection occurs, one is forced to perform another, possibly more complicated, test.

In the above, not only θ, but τ as well, has been a one-dimensional parameter. However, for higher-dimensional τ, the situation essentially remains the same: as long as optimal tests are used for each of the three components of the problem, we either have a stable size but closely related main tests with little power difference, or we get really different main tests, which opens the way to power gain, but only at the expense of unacceptable deviations in size.

Nevertheless, interesting situations involving power gain may arise, and these are the subject of the present chapter. The starting point now is not the parametric family $f(x;\theta,\tau)$ where the nuisance parameter τ may or may not be known, but the basic main test. This test is concerned with the main hypothesis on θ and is appropriate in a restricted (parametric) model $f(x;\theta)$. However, this model may be incorrect. To check its validity a preliminary test is applied. The preliminary test picks up a particular property of the family $f(x;\theta)$ and explores whether the data are in agreement with it. For instance, if $f(x;\theta)$ represents a symmetric location model with location parameter θ, then the mean and median in this model coincide and the preliminary test can be devised for testing whether the difference between mean and median equals zero. More generally, the preliminary test protects against a limited number of directions in which the density at hand may deviate from the restricted model. For simplicity, we assume that the deviations against which the preliminary test effectively protects, can be parameterized by a single parameter τ, giving back the restricted family $f(x;\theta)$ for some given value of τ. Hence, the preliminary test is appropriate in a model $f(x;\theta,\tau)$ and tests whether τ equals some prescribed value.

If the null hypothesis regarding τ is not rejected by the preliminary test, then the basic main test in the restricted model $f(x;\theta)$ is carried out as intended. Otherwise, the main hypothesis on θ is tested in a non-parametric model in the following sense. If, for example, θ can be interpreted as the median in the family $f(x;\theta)$, then in the non-parametric case we may also test the corresponding null hypothesis about the median, and apply the sign test for this purpose. Hence, the interpretation of θ in the restricted model determines the second main test. In principle, there is a choice which property to take, since usually θ can be interpreted in several ways. The particular choice determines the second main test. Note that in contrast to the previous chapter, where the main hypothesis on θ was tested in a parametric family, the second main

test in the present setting is often not that complicated, thus resulting in an attractive pre-test procedure from the point of view of simplicity.

Because the preliminary test protects (mainly) against alternatives of the form $f(x;\theta,\tau)$, we analyze the pre-test procedure in this model. The reason for possible power gain of the pre-test procedure compared to the second main test, becomes now transparent: typically the second main test provides robustness of validity in a much larger model, resulting in some loss of power in the model $f(x;\theta,\tau)$. (Compare again the sign test, which indeed guarantees the correct level for all continuous densities.)

It is our purpose not only to reveal the properties of the pre-test procedure, but in particular to clarify when power gain is obtained while still controlling the size. Although the practical application of the pre-test procedure is straightforward, its analysis is not trivial. Two quantities play a very important role: (i) the correlation (ρ) between the score function (ψ_{10}) for the parameter of interest and the score function (ψ_{01}) for the nuisance parameter, and (ii) the correlation between the basic main test and the second main test.

- *The correlation between the score functions for the parameter of interest and for the nuisance parameter.* The validity of the pre-test procedure, measured by its size, can only be controlled if the correlation between the score functions for the parameter of interest and the nuisance parameter is small. Otherwise, the size of the pre-test procedure varies wildly and unacceptable violations of the prescribed level cannot be avoided. For an explanation of this effect we refer to Section 4.2 or 5.2, noting that it is mainly caused by the interplay of the basic main test and the preliminary test, which remain the same as in Chapter 4. But even in case of a small correlation, non-negligible departures from the nominal level can occur. While first-order asymptotics does not reveal this feature, the use of second-order asymptotics in Theorem 5.3.2 makes clear what is going on.

- *The correlation between the basic main test and the second main test.* Having controlled the size of the pre-test procedure, it is indeed possible to get power gain for the pre-test procedure compared to always using the second main test. This will happen when the basic main test and the second main test are not too highly correlated. In principle, the smaller the correlation, the higher the gain in power. This qualitative statement is made more precise in Theorem 5.3.2, where a transparent approximation also gives the possible (asymptotic) power gain in a quantitative way. By this result the aim to show whether it is possible to get power gain in pre-testing and if so, to give insight in when this occurs and to what extent, is achieved.

The organization of this chapter is as follows. In Section 5.2 the framework for our analysis is set. Notation is given, the various models are specified and the classes of statistics to be considered are given, together with some natural assumptions.

In Section 5.3 we present and discuss the approximation (given by Theorem 5.3.2) for the size and power difference of the pre-test procedure with respect to the robust second main test. This approximation gives us the required qualitative and quanti-

tative insight in the behavior of the pre-test procedure. To arrive at this approximation, we use second-order asymptotics in n, together with assumptions about the two correlations mentioned. That these assumptions are necessary and that first-order asymptotics is generally insufficient to explain deviations of the size of the procedure, is demonstrated in Section 5.2.

In Section 5.4 the theory is exemplified by considering the pre-test procedure with essentially the t-test as basic main test for the null hypothesis of a zero median, and the sign test as second main test. The accuracy of the approximations is illustrated by comparison with simulation results.

5.2 Notation, assumptions and preliminaries

Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s with density f with respect to a measure μ on the measurable space $(\mathcal{X}, \mathcal{A})$. Suppose that we conjecture that this density can be parameterized as $f(x;\theta)$. A short remark on notation should be made. When we write f, an unknown density is meant, while $f(x;\theta)$ indicates a member of a family of densities with given (and hence known) function f and unknown parameter θ. To avoid too many different symbols we use the same notation f with different interpretation; which interpretation is the right one is however obvious from the context. For simplicity of presentation we take θ one-dimensional, but an extension to higher dimensions is possible. Suppose that, according to the conjectured parameterization, we are interested in testing the main hypothesis $H_0 : \theta = 0$ against $H_1 : \theta > 0$. This can be done by a test in the restricted model $f(x;\theta)$. But in order to determine whether this model is suitable, a preliminary test is performed. This test investigates whether there are deviations from the model $f(x;\theta)$ into a direction parameterized by a parameter τ (also one-dimensional) giving back the restricted model for $\tau = 0$. Hence the preliminary test aims at testing $\bar{H}_0 : \tau = 0$ against $\bar{H}_1 : \tau \neq 0$ in a model $f(x;\theta,\tau)$ with $f(x;\theta,0) = f(x;\theta)$. If the preliminary test does not reject, the main hypothesis on θ is tested by the basic main test in the restricted model $f(x;\theta)$. If the preliminary test rejects, however, then a well-defined property that corresponds to $\theta = 0$ in the restricted family, is tested without making restrictive assumptions on the density f. This property is often written in the form $\int \xi f d\mu = 0$ for some given function ξ, e.g. $\xi(x) = sign(x)$ for the case of the median. As $f(x;\theta,\tau)$ is an extension of the restricted family $f(x;\theta)$, we should of course also have that $\theta = 0$ in this model gives $\int \xi f d\mu = 0$, or, equivalently, that $\int \xi(x) f(x;0,\tau) dx = 0$ for all τ. The hypotheses of the main testing problem are rewritten in the form $H_0^* : \int \xi f d\mu = 0$ against $H_1^* : \int \xi f d\mu > 0$. Without further assumptions on the form of the distribution, it is asymptotically optimal to base the test simply on the empirical estimator of $\int \xi f d\mu$, given by $n^{-1} \sum \xi(X_i)$ (Van der Vaart (1998), p. 368, 385).

So, we consider three models. The smallest one is denoted by $f(x;\theta)$. Here, we test $H_0 : \theta = 0$ against $H_1 : \theta > 0$. This model may be incorrect. This is checked by a preliminary test which is appropriate in the somewhat larger model $f(x;\theta,\tau)$ with $f(x;\theta,0) = f(x;\theta)$. Here we test $\bar{H}_0 : \tau = 0$ against $\bar{H}_1 : \tau \neq 0$.

Finally, if the preliminary test rejects, the main testing problem is replaced by testing $H_0^* : \int \xi f d\mu = 0$ against $H_1^* : \int \xi f d\mu > 0$ with the function ξ chosen such that $\theta = 0$ gives $H_0^* : \int \xi(x) f(x;0) d\mu(x) = 0$. Since H_0^* is the extension of H_0 in a larger model, including at least the model $f(x;\theta,\tau)$, we have also $\int \xi(x) f(x;0,\tau) d\mu(x) = 0$ for all τ. Because the preliminary test protects only against alternatives of the form $f(x;\theta,\tau)$, we analyze the pre-test procedure in this model.

Before defining the test statistics, we will give some further notation and conditions. By $P_{\theta,\tau}$ we denote that X_i has density $f(x;\theta,\tau)$. Expectations under this model are denoted by $E_{\theta,\tau}$ or shortly by E if $E_{0,0}$ is meant. We assume that the regularity conditions (Ri)–(Riv) of the previous chapter, Section 4.2, p. 39 hold. Conditions (Rv) and (Rvi) (needed for Götze's Lemma 4.2.2 and for the existence of expectations and remainder terms) need to be adapted for the tests used in this chapter. They will be as follows (where the functions k, q, r^* which will appear in the basic main test and the preliminary test, are the same as in Chapter 4, and where ξ and w are functions appearing in the second main test, defined later in this section):

(Rv)
$$\begin{aligned}
&\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau} |\psi_{ij}(X_1)|^3 I(|\psi_{ij}(X_1)| \geq x) = 0, \text{ for } (i,j) = (0,1),(1,0),\\
&\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau} |\xi(X_1)|^3 I(|\xi(X_1)| \geq x) = 0,\\
&\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau} |k(X_1)|^3 I(|k(X_1)| \geq x) = 0,\\
&\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau} |q(X_1)|^3 I(|q(X_1)| \geq x) = 0,\\
&\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau} |r^*(X_1)|^3 I(|r^*(X_1)| \geq x) = 0,\\
&\lim_{x\to\infty} \sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau} |w(X_1)|^3 I(|w(X_1)| \geq x) = 0,\\
&\lim_{\|t\|\to\infty} \sup_{(\theta,\tau)\in\Omega_0} |E_{\theta,\tau} \exp[it_1\psi_{10}(X_1) + it_2\{\psi_{01}(X_1) - \rho\psi_{10}(X_1)\}]| < 1,\\
&\lim_{\|t\|\to\infty} \sup_{(\theta,\tau)\in\Omega_0} |E_{\theta,\tau} \exp[it_1\xi(X_1) + it_2\{\psi_{01}(X_1) - \rho\psi_{10}(X_1)\}]| < 1.
\end{aligned}$$

(Rvi)
$$\begin{aligned}
\sup_{(\theta,\tau)\in\Omega_0} E_{\theta,\tau} \Big(&|\psi_{10}(X_1)|^3 + |\psi_{01}(X_1)|^3 + |\xi(X_1)|^3 + |k(X_1)|^2 + |q(X_1)|^2 \\
&+ |r^*(X_1)|^2 + |w(X_1)|^2 + 1 \Big) |\psi_{ij}(X_1;\theta,\tau)| < \infty
\end{aligned}$$
for $i,j = 0,1,2,3$.

The score functions $\psi_{ij}^*(x;\theta,\tau)$, under (0,0) denoted by ψ_{ij}^* or $\psi_{ij}^*(x)$, and the elements I_{11}, I_{12} and I_{22} of the Fisher information matrix are as given in Section 4.2, and also the definitions of the standardized versions $\psi_{ij}(x)$ and the correlation coefficient ρ are the same. In addition to

$$S = n^{-1/2} \sum_{i=1}^{n} \psi_{10}(X_i) \quad \text{and} \quad T = n^{-1/2} \sum_{i=1}^{n} \psi_{01}(X_i) \tag{5.2.1}$$

we now define

$$\Xi = n^{-1/2} \sum_{i=1}^{n} \xi(X_i).$$

The basic main test is meant for testing $H_0 : \theta = 0$ in the family $f(x; \theta, 0)$, thus assuming that the form of the underlying distribution is known except for the parameter of interest itself. A class of test statistics for this problem is given by

$$SK = S + n^{-1/2} S \left\{ n^{-1/2} \sum_{i=1}^{n} k(X_i) \right\} \tag{5.2.2}$$

for some function k with $Ek = 0$. The null hypothesis $H_0 : \theta = 0$ is rejected in favor of $H_1 : \theta > 0$ for large values of SK. The functions k corresponding to the LMP test, Rao's efficient score test, the LR test, the test based on the MLE of θ, and Wald's test, are given in (4.2.5).

The only interesting case is the situation where the correlation coefficient ρ is small, which will be assumed hereafter. This may be explained as follows. Since the basic main test and the preliminary test are meant for testing problems in the same parametric models as in Chapter 4, the corresponding test statistics SK and TU (given below) have the same forms, and hence we may use the same arguments as there: Under $f(x; 0, cI_{22}^{-1/2} n^{-1/2})$ the distribution of SK converges to a normal one with expectation $c\rho$ and variance 1. At the same time, TU is asymptotically normal with expectation $c\sqrt{1-\rho^2}$ and variance 1. Hence, for the preliminary test to have large power, it is necessary that c is large. But then the error in the size of the basic main test, and together with it the error in the size of the total procedure, grows very large, unless ρ is small. Therefore, the procedure is unacceptable for large ρ and it is only natural to restrict attention to the case that ρ is small.

Although the preceding argument could be based on first-order asymptotics in n, this does not suffice for our analysis. For the present procedure, this can be seen by considering the sizes of the two main tests and of the combined two-stage procedure. Under $f(x; bI_{11}^{-1/2} n^{-1/2}, cI_{22}^{-1/2} n^{-1/2})$, the test statistic for the basic main test converges in distribution to a normal r.v. with expectation $b + c\rho$ and variance 1. The test statistic for the second main test will be based on Ξ to first order, and converges to a normal distribution with expectation $bE\xi\psi_{10}$ and variance 1. In situations where $\rho = 0$, the size of the two main tests is thus the same to first order. Hence, to first order, the size of the pre-test procedure is equal to the nominal level if $\rho = 0$. Nevertheless, numerical work shows that differences in size between the pre-test procedure and the nominal level are certainly not negligible in practice. To explain these differences, we need higher-order asymptotics. Note however, that in the present situation (contrary to that of Chapter 4), the test statistics of the two main tests do differ in first order, also when $\rho = 0$. We will see that this causes differences in power between the pre-test procedure and the (second) main test, while the size of the procedure still equals the nominal level to first order.

The fact that small ρ is the only interesting case justifies that in the derivation of the results, we use expansions not only with respect to n, but also with respect to ρ, ignoring terms of order ρ^3 and $\rho n^{-1/2}$. In this way the results are not only accurate, but also transparent, which would not have been the case if ρ were taken fixed.

The class of preliminary tests for $\bar{H}_0 : \tau = 0$ against (two-sided) alternatives in the parametric model $f(x;\theta,\tau)$ with θ unknown, is based on (cf. Section 4.2)

$$\left[T - \rho S + n^{-1/2}\left\{Tn^{-1/2}\sum_{i=1}^{n} q(X_i) + Sn^{-1/2}\sum_{i=1}^{n} r^*(X_i)\right\}\right](1-\rho^2)^{-1/2} \tag{5.2.3}$$

with functions q and r^* satisfying $Eq = Er^* = 0$. In order to get a well-defined statistic, studentizing is needed, since under $f(x; bI_{11}^{-1/2}n^{-1/2}, 0)$ the variance of (5.2.3) depends on the unknown b. Ignoring terms of order $\rho n^{-1/2}$ and n^{-1}, this leads to a statistic TU of the given form (5.2.3), where r^* is replaced by

$$r = r^* - \tfrac{1}{2}\left\{E\psi_{10}\psi_{01}^2 + 2[Eq\psi_{10} + Er^*\psi_{01}]\right\}\psi_{01}, \tag{5.2.4}$$

The preliminary test rejects for large values of $|TU|$.

Finally, we have to specify the second main test. This test is based on a non-parametric model and will therefore differ from the corresponding test in the previous chapter (although we still call it SU).

In that model, the null hypothesis is written as $H_0^* : \int \xi f d\mu = 0$. Since $f(x;\theta,\tau)$ is a sub-model of the large model, $f(x;0,\tau)$ belongs to H_0^*, implying $E_{0,\tau}\xi = 0$ for all τ and hence $E\xi = E\xi\psi_{01} = E\xi\psi_{02} = 0$. Moreover, $f(x;\theta,0)$ belongs to H_1^* for $\theta > 0$, implying $E_{\theta,0}\xi > 0$ for $\theta > 0$. Hence, it is natural to assume that $E\xi\psi_{10} \geq 0$. This indicates that ξ and ψ_{10} are positively correlated, as they should be, since tests based on these functions are meant for the same hypothesis, written as H_0 or H_0^*. Without loss of generality let also $E\xi^2 = 1$. Then it follows from the Cauchy-Schwarz inequality that $E\xi\psi_{10} = E\xi(\psi_{10} - \rho\psi_{01}) \leq \sqrt{1-\rho^2}$. Hence, by the assumption that $E\xi\psi_{10} \geq 0$, we may write $E\xi\psi_{10}$ as $E\xi\psi_{10} = \sqrt{(1-\rho^2)(1-\gamma^2)}$ for some γ.

In order to keep convenient clarity in the results, it will be assumed that terms of order $\gamma n^{-1/2}$, with γ from $E\xi\psi_{10} = \sqrt{(1-\rho^2)(1-\gamma^2)}$, are negligible. For γ small, this is of course justified. If γ is not small, it is also justified because then the first-order terms which depend on γ, will dominate. As a consequence of this assumption, we may replace ξ by ψ_{10} in $n^{-1/2}$-terms with factors of the form $E\xi h$, $E\xi^2 h$, $E\xi^3$ for regular functions h. This follows from the Cauchy-Schwarz inequality, e.g. for the first form: $E(\xi - \psi_{10})h \leq \sqrt{E(\xi-\psi_{10})^2 Eh^2} = \sqrt{O(\rho^2+\gamma^2)Eh^2} = O(\rho+\gamma)$, provided the second moments under $(0,0)$ exist. For factors of the form $E\xi^2 h$ or $E\xi^3$ we need fourth moments in the application of Cauchy-Schwarz. The existence of second moments of k, r, q and w and fourth moments of ψ_{10} and ψ_{01} already follows from (Rvi). In addition to this, we assume that the following condition holds

(Rvii) $\quad E\xi^4 < \infty, \quad E\psi_{20}^2 < \infty, \quad E\psi_{11}^2 < \infty, \quad E\psi_{02}^2 < \infty.$

Since $E\xi\psi_{02} = 0$, condition (Rvii) also implies that terms with $E\psi_{10}\psi_{02}n^{-1/2}$ are of order $\rho n^{-1/2} + \gamma n^{-1/2}$, and may be ignored.

The negligibility of terms of order $\rho n^{-1/2}$ and $\gamma n^{-1/2}$ will simplify the calculation of the second-order part of the approximate power gain. By considering the power difference right from the start, and using the similarity of the second-order parts of the two main tests, we can avoid a lot of terms that we would have to deal with if we would calculate the powers of the two-stage procedure and of the second main test separately, and take differences afterwards.

The empirical estimator for $\int \xi f d\mu$ equals $n^{-1/2}\Xi$. This can be extended to a class of test statistics based on

$$\Xi + n^{-1/2}\Xi\left\{n^{-1/2}\sum_{i=1}^{n} w^*(X_i)\right\} \tag{5.2.5}$$

with $Ew^* = 0$. Under the assumptions concerning ξ, the expectation under $f(x; 0, cI_{22}^{-1/2}n^{-1/2})$ does not depend on the unknown c up to the considered order, but again we need to replace the unknown c in the variance by a consistent estimator in order to get a well-defined test statistic. The variance under $f(x; 0, cI_{22}^{-1/2}n^{-1/2})$ equals up to order $n^{-1/2}$

$$1 + cn^{-1/2}\{E\xi^2\psi_{01} + 2Ew^*\psi_{01}\}.$$

Studentizing leads to a statistic SU of the form (5.2.5) with w^* replaced by

$$w = w^* - \tfrac{1}{2}\left\{E\xi^2\psi_{01} + 2Ew^*\psi_{01}\right\}\psi_{01}. \tag{5.2.6}$$

Note that studentizing Ξ by means of its sample variance, corresponds to $w^* = -\frac{1}{2}(\xi^2 - 1)$. Then no further correction is needed. Also ignoring terms of order $\rho n^{-1/2} + \gamma n^{-1/2}$, we may replace ξ by ψ_{10} and use

$$w = w^* - \tfrac{1}{2}\left\{E\psi_{10}^2\psi_{01} + 2Ew^*\psi_{01}\right\}\psi_{01}. \tag{5.2.7}$$

The second main test rejects for large values of SU.

Now we know from (5.2.3) and (5.2.4), from (5.2.2), and from (5.2.5) and (5.2.6) or (5.2.7) the form of the test statistics TU, SK and SU for the preliminary test and the two main tests, respectively. We will investigate the power (or size) difference between the pre-test procedure consisting of these tests and the second main test based on SU. The latter has the correct level (at least to $O(\rho n^{-1/2} + \gamma n^{-1/2}) + o(n^{-1/2})$), since it does not use possibly incorrect information about the form of the distribution.

5.3 Main results: difference in size and power

In this section we will give an approximation for the difference in size and power between the pre-test procedure and the main test based on SU. For this approximation we need a bivariate Edgeworth expansion for the simultaneous distribution of the

statistics corresponding to the preliminary test and each of the main tests. Since we consider local alternatives of the form (θ_n, τ_n) with

$$\theta_n = bI_{11}^{-1/2}n^{-1/2},\ \tau_n = cI_{22}^{-1/2}n^{-1/2} \text{ with } b \geq 0 \text{ and } c \in \mathbb{R}, \tag{5.3.1}$$

this expansion must be valid under probability measures depending on n. Theorem 5.3.1 below gives the required expansion and is the analogue of Theorem 4.3.1 of the previous chapter. From that expansion we can derive the critical values for the three testing problems. Then we have all the ingredients to state the main result in Theorem 5.3.2. For notational convenience, define $\lambda = \sqrt{1-\gamma^2}$.

Theorem 5.3.1 *Suppose that* (Ri)–(Rvii) *hold. Let*

$$U_{1n} = \frac{SK - \mu_{1n}(b,c)}{\sigma_{1n}(b,c)},\ U_{2n} = \frac{SU - \mu_{2n}(b,c)}{\sigma_{2n}(b)} \text{ and } V_n = \frac{TU - \mu_n(b,c)}{\sigma_n(c)}$$

with

$$\begin{aligned}
\mu_{1n}(b,c) &= b + c\rho + \tfrac{1}{2}\{b^2E\psi_{10}\psi_{20} + 2bcE\psi_{10}\psi_{11} + 2(1+b^2)Ek\psi_{10} \\
&\qquad + 2bcEk\psi_{01}\}n^{-1/2}, \\
\sigma_{1n}(b,c) &= 1 + \tfrac{1}{2}\{bE\psi_{10}^3 + cE\psi_{10}^2\psi_{01} + 4bEk\psi_{10} + 2cEk\psi_{01}\}n^{-1/2}, \\
\mu_{2n}(b,c) &= b\lambda(1 - \tfrac{1}{2}\rho^2) + \tfrac{1}{2}\{b^2E\psi_{10}\psi_{20} + 2bcE\psi_{10}\psi_{11} \\
&\qquad + 2(1+b^2)Ew\psi_{10} - bcE\psi_{10}^2\psi_{01}\}n^{-1/2} \\
\sigma_{2n}(b) &= 1 + \tfrac{1}{2}b(E\psi_{10}^3 + 4Ew\psi_{10})n^{-1/2} \\
\mu_n(b,c) &= c - \tfrac{1}{2}c\rho^2 + \tfrac{1}{2}\{c^2E\psi_{01}\psi_{02} + 2bcE\psi_{01}\psi_{11} + b^2E\psi_{01}\psi_{20} \\
&\qquad - bcE\psi_{10}\psi_{01}^2 + 2(1+b^2)Er\psi_{10} + 2(1+c^2)Eq\psi_{01}\}n^{-1/2}, \\
\sigma_n(c) &= 1 + \tfrac{1}{2}c(E\psi_{01}^3 + 4Eq\psi_{01})n^{-1/2}.
\end{aligned} \tag{5.3.2}$$

Then

$$\begin{aligned}
&P_{\theta_n,\tau_n}(U_{in} \leq u, V_n \leq v) = \Phi(u,v;\rho_i) + n^{-1/2}\sum_{j=0}^{3} c_{ij}\Phi^{(j)}(u)\Phi^{(3-j)}(v) \\
&+ O(\rho^4 + \rho n^{-1/2} + \gamma n^{-1/2}) + o(n^{-1/2})
\end{aligned} \tag{5.3.3}$$

uniformly for $(u,v) \in \mathbb{R}^2$, *with*

$$\begin{aligned}
&\rho_1 = n^{-1/2}r_1(b,c),\ \rho_2 = -\rho\lambda + n^{-1/2}r_2(b,c) \\
&\textit{for some functions } r_1(b,c),\ r_2(b,c) \textit{ satisfying} \\
&r_2(b,c) - r_1(b,c) = -\tfrac{1}{2}bE\psi_{10}^2\psi_{01} - bEk\psi_{01},
\end{aligned} \tag{5.3.4}$$

and

$$\begin{aligned} &c_{10} = c_{20},\ c_{11} - c_{21} = 0,\ c_{12} - c_{22} = -\tfrac{1}{2}E\psi_{10}^2\psi_{01} - Ek\psi_{01}, \\ &c_{13} = -\tfrac{1}{6}E\psi_{10}^3 - Ek\psi_{10},\ c_{23} = -\tfrac{1}{6}E\psi_{10}^3 - Ew\psi_{10}. \end{aligned} \tag{5.3.5}$$

Note that, except for the cancellation of $E\psi_{10}\psi_{02}n^{-1/2}$ and the order of the remainder term, only the leading terms of μ_{2n} and ρ_2 differ from those in Chapter 4, by the appearance of λ. All expectations, variances, correlations and skewness terms corresponding to both SK and TU remain the same, since the form of these test statistics is unchanged. Hence, also the proof of this theorem, which we do not present here, is very similar. It goes along the same lines as the proof of Theorem 4.3.1, except that it is even slightly more technical. This is due to the fact that in all expectations occurring in $n^{-1/2}$-terms, we replace ξ by ψ_{10}, taking account for this replacement by adding a remainder term of order $O(\rho n^{-1/2} + \gamma n^{-1/2})$. This is allowed if the necessary moments exist, as explained before condition (Rvii), which was added for this purpose. Owing to that replacement, we may again profit from the similarity of (the $n^{-1/2}$-terms of) the two main tests in the calculation of $r_2(b,c) - r_1(b,c)$ and the differences $c_{1j} - c_{2j}$. Therefore we get the same expressions for the $n^{-1/2}$-terms (except for the fact that terms with $E\psi_{10}\psi_{02}n^{-1/2}$ do not appear anymore).

From the marginal distributions under the corresponding null hypothesis it is easy to derive the critical values for the three separate testing problems. Since $\sigma_{2n}(0) = 1$, we have for the second main test

$$\begin{aligned} P_{0,\tau_n}(SU \le u) = \Phi(u - \mu_{2n}(0,c)) + n^{-1/2}c_{23}\Phi^{(3)}(u) + O(\rho n^{-1/2} + \gamma n^{-1/2}) \\ + o(n^{-1/2}) \end{aligned}$$

with $\mu_{2n}(0,c)$ and c_{23} from (5.3.2) and (5.3.5). Note that $\mu_{2n}(0,c) = Ew\psi_{10}n^{-1/2}$ does not depend on c. Hence, the one-sided test based on SU which rejects $H_0 : \theta = 0$ when

$$SU > u_\alpha - c_{23}n^{-1/2}(u_\alpha^2 - 1) + \mu_{2n}(0,c),$$

has size $\alpha + O(\rho n^{-1/2} + \gamma n^{-1/2}) + o(n^{-1/2})$ under $(0, \tau_n)$.

With the tests based on SK and TU the same as before, it follows that the pre-test procedure rejects $H_0 : \theta = 0$ against $H_1 : \theta > 0$ when

$$\begin{aligned} &SK > u_\alpha + \tfrac{1}{6}E\psi_{10}^3 n^{-1/2}(u_\alpha^2 - 1) + Ek\psi_{10}n^{-1/2}u_\alpha^2 \text{ and } |TU| \le u_{\delta/2} \\ &\text{or when} \\ &SU > u_\alpha + \tfrac{1}{6}E\psi_{10}^3 n^{-1/2}(u_\alpha^2 - 1) + Ew\psi_{10}n^{-1/2}u_\alpha^2 \text{ and } |TU| > u_{\delta/2}. \end{aligned} \tag{5.3.6}$$

With $\pi^*(b,c)$ and $\tilde{\pi}(b,c)$ the probabilities of rejection by the pre-test procedure and the second main test, the following theorem gives the main result: an expression

for the approximate power gain (or size difference) $\pi^*(b,c) - \tilde{\pi}(b,c)$ of the two-stage pre-test procedure compared to the second main test based on SU, which does not assume restrictive knowledge about the underlying distribution.

Theorem 5.3.2 *Suppose that* (Ri)–(Rvii) *are satisfied. Then*

$$\begin{aligned}
&\pi^*(b,c) - \tilde{\pi}(b,c) \\
&= \{\Phi(u_\alpha - b\lambda) - \Phi(u_\alpha - b)\}\{\Phi(u_{\delta/2} - c) - \Phi(-u_{\delta/2} - c)\} \\
&+ h_1(c, u_{\delta/2})[\rho\varphi(u_\alpha - b) + \tfrac{1}{2}\rho^2\{\frac{b\lambda}{c}\varphi(u_\alpha - b\lambda) + c(u_\alpha - b)\varphi(u_\alpha - b)\} \\
&\qquad + m(u_\alpha)\varphi(u_\alpha - b)n^{-1/2}] \\
&+ h_2(c, u_{\delta/2})[\rho\lambda\varphi(u_\alpha - b\lambda) + \tfrac{1}{2}c\rho^2\{\lambda^2(u_\alpha - b\lambda)\varphi(u_\alpha - b\lambda) \\
&\qquad + \Phi(u_\alpha - b) - \Phi(u_\alpha - b\lambda)\} + m(u_\alpha)\varphi(u_\alpha - b)n^{-1/2}] \\
&+ h_3(c, u_{\delta/2})[\rho^2\lambda^2(u_\alpha - b\lambda)\varphi(u_\alpha - b\lambda)] \\
&+ O(\rho^3 + \rho n^{-1/2} + \gamma n^{-1/2}) + o(n^{-1/2}) \text{ as } \rho \to 0 \text{ and } n \to \infty,
\end{aligned} \tag{5.3.7}$$

where

$$\begin{aligned}
h_1(x,y) &= x\{\Phi(y-x) - \Phi(-y-x)\}, \\
h_2(x,y) &= \varphi(y+x) - \varphi(y-x), \\
h_3(x,y) &= \tfrac{1}{2}y\{\varphi(y+x) + \varphi(y-x)\}, \\
m(y) \quad &= y\{\tfrac{1}{2}E\psi_{10}^2\psi_{01} + Ek\psi_{01}\}.
\end{aligned}$$

Proof. Rewrite the power gain (or size difference) $\pi^* - \tilde{\pi}$ as

$$\begin{aligned}
&\pi^*(b,c) - \tilde{\pi}(b,c) = P_{\theta_n,\tau_n}(U_{1n} \le u_{1n}, V_n \le v_n^L) - P_{\theta_n,\tau_n}(U_{2n} \le u_{2n}, V_n \le v_n^L) \\
&- \{P_{\theta_n,\tau_n}(U_{1n} \le u_{1n}, V_n < v_n^U) - P_{\theta_n,\tau_n}(U_{2n} \le u_{2n}, V_n < v_n^U)\}
\end{aligned} \tag{5.3.8}$$

with U_{1n} and U_{2n} the test statistics for the main tests based on SK and SU, standardized under local alternatives, u_{1n} and u_{2n} the standardized critical values, and V_n the standardized test statistic for the two-sided preliminary test, with corresponding lower and upper critical value v_n^L and v_n^U. With $\lambda = \sqrt{1-\gamma^2}$, $u_0 = u_\alpha - b$,

$v_0^U = u_{\delta/2} - c$, $v_0^L = -u_{\delta/2} - c$, and u_1^n and u_2^n denoting the $n^{-1/2}$-terms of u_{1n} and u_{2n}, these critical values equal

$$\begin{aligned}
u_{1n} &= u_0 - c\rho + n^{-1/2}u_1^n + O(\rho n^{-1/2} + n^{-1}),\\
u_{2n} &= u_\alpha - b\lambda + \tfrac{1}{2}b\lambda\rho^2 + n^{-1/2}u_2^n + O(\rho^2 n^{-1/2} + \gamma^2 n^{-1/2} + n^{-1}),\\
v_n^U &= v_0^U + \tfrac{1}{2}c\rho^2 + O(n^{-1/2}),\\
v_n^L &= v_0^L + \tfrac{1}{2}c\rho^2 + O(n^{-1/2}),
\end{aligned}$$

where the $n^{-1/2}$-terms of v_n^U and v_n^L will not contribute to the result, and where the difference between the $n^{-1/2}$-parts of u_{1n} and u_{2n} equals

$$u_1^n - u_2^n = (Ek\psi_{10} - Ew\psi_{10})(u_0^2 - 1) - cm(u_\alpha).$$

Theorem 5.3.1 gives expansion (5.3.3) with

$$\rho_1 = O(n^{-1/2}), \quad \rho_1 - \rho_2 = \rho\lambda + n^{-1/2}bm(u_\alpha)/u_\alpha, \quad \text{and}$$

$$c_{10} = c_{20},\ c_{11} = c_{21},\ c_{12} - c_{22} = -m(u_\alpha)/u_\alpha,\ c_{13} - c_{23} = Ew\psi_{10} - Ek\psi_{10}.$$

This expansion can be used to evaluate (5.3.8). However, we cannot directly apply Lemma 4.2.1 now, since the difference between u_{1n} and u_{2n} is of order $O(\rho+\gamma^2+n^{-1/2})$ instead of order $O(\rho + n^{-1/2})$. We have to be more careful when taking differences of terms with u_{1n} and u_{2n}. Distinguishing between terms with ρ, ρ^2, γ^2 and $n^{-1/2}$, and carefully gathering and ignoring terms of the orders in the remainder term of (5.3.7), leads to the result, as is shown next.

From expansion (5.3.3) and the rewritten form (5.3.8) of $\pi^* - \tilde{\pi}$, it follows that the first part of the power gain is given by the difference

$$\{\Phi(u_{1n}, v_n^L; \rho_1) - \Phi(u_{2n}, v_n^L; \rho_2)\} - \{\Phi(u_{1n}, v_n^U; \rho_1) - \Phi(u_{2n}, v_n^U; \rho_2)\}.$$

with $\rho_1 = O(n^{-1/2})$ and $\rho_1 - \rho_2 = O(\rho + n^{-1/2})$. Using this when expanding for ρ_1 and ρ_2 small, together with the facts that $u_{1n} - u_{2n} = O(\rho + \gamma^2 + n^{-1/2})$ and $v_n^U = v_0^U + O(\rho^2 + n^{-1/2})$, gives

$$\begin{aligned}
&\Phi(u_{1n}, v_n^U; \rho_1) - \Phi(u_{2n}, v_n^U; \rho_2)\\
&= \{\Phi(u_{1n}) - \Phi(u_{2n})\}\Phi(v_n^U) + \{\rho_1\varphi(u_{1n}) - \rho_2\varphi(u_{2n})\}\varphi(v_n^U)\\
&\quad - \tfrac{1}{2}\rho_2^2\varphi'(u_{2n})\varphi'(v_n^U) + O(\rho_1^2 + \rho_2^3) \qquad (5.3.9)\\
&= \{\Phi(u_{1n}) - \Phi(u_{2n})\}\Phi(v_n^U) - (\rho_2 - \rho_1)\varphi(u_{2n})\varphi(v_0^U)\\
&\quad - \tfrac{1}{2}(\rho_2 - \rho_1)^2\varphi'(u_{2n})\varphi'(v_0^U) + O(\rho^3 + \rho n^{-1/2} + \gamma^2 n^{-1/2} + n^{-1}).
\end{aligned}$$

In this expression, we expand u_{1n}, u_{2n}, v_n^U (up to the required order) for ρ small and n large around $u_0 = u_\alpha - b$, $u_\alpha - b\lambda$, $v_0^U = u_{\delta/2} - c$, respectively. Substituting the necessary terms of $\rho_2 - \rho_1$, gives for the first part

$$\begin{aligned}
&\{\Phi(u_\alpha - b) - \Phi(u_\alpha - b\lambda)\}\Phi(u_{\delta/2} - c) \\
&+ \Phi(u_{\delta/2} - c)[\{-c\rho - \tfrac{1}{2}c^2\rho^2(u_\alpha - b)\}\varphi(u_\alpha - b) - \tfrac{1}{2}b\lambda\rho^2\varphi(u_\alpha - b\lambda) \\
&\qquad + n^{-1/2}\{(Ek\psi_{10} - Ew\psi_{10})(u_0^2 - 1) - cm(u_\alpha)\}\varphi(u_\alpha - b)] \\
&+ \varphi(u_{\delta/2} - c)[\{\rho\lambda - \tfrac{1}{2}\rho^2\lambda^2(u_\alpha - b\lambda)(u_{\delta/2} - c)\}\varphi(u_\alpha - b\lambda) \\
&\qquad + n^{-1/2}b\frac{m(u_\alpha)}{u_\alpha}\varphi(u_\alpha - b) + \tfrac{1}{2}c\rho^2\{\Phi(u_\alpha - b) - \Phi(u_\alpha - b\lambda)\}].
\end{aligned} \tag{5.3.10}$$

For the second part we get (using $c_{10} = c_{20}$ and $c_{11} = c_{21}$)

$$\begin{aligned}
&n^{-1/2}\sum_{j=0}^{3}\{c_{1j}\Phi^{(j)}(u_{1n}) - c_{2j}\Phi^{(j)}(u_{2n})\}\Phi^{(3-j)}(v_n^U) \\
&= \; n^{-1/2}\sum_{j=2}^{3}(c_{1j} - c_{2j})\Phi^{(j)}(u_0)\Phi^{(3-j)}(v_0) + O(\rho n^{-1/2} + \gamma^2 n^{-1/2} + n^{-1}).
\end{aligned} \tag{5.3.11}$$

Combining these two parts shows that the $n^{-1/2}$-terms with $c_{13} - c_{23}$ and $Ek\psi_{10} - Ew\psi_{10}$ cancel, as well as those with $-b(c_{12} - c_{22})$ and $bm(u_\alpha)/u_\alpha$. Taking $u_{\delta/2}$ outside brackets and taking the difference with the expression with $-u_{\delta/2}$ yields the result. □

The result given by (5.3.7) depends on the family of distributions and the classes of tests only through the correlations ρ and λ, and through the expectations $E\psi_{10}^2\psi_{01}$ and $Ek\psi_{01}$ in the function m. Hence, only the function k, coming from the basic main test, appears in the result. The functions q, r and w, which define the $n^{-1/2}$-parts of the other tests, do not influence the final result.

The result $\pi^* - \tilde{\pi}$ may be decomposed into two parts. Let π denote the probability that the basic main test based on SK rejects H_0 and $\bar{\pi}$ the probability that the preliminary test rejects $\bar{H}_0$. Then the first part may be written as the product $(\pi - \tilde{\pi})(1 - \bar{\pi})$, which is the power difference of the two main tests multiplied by the acceptance probability of the preliminary test. This part is recognized in the first three lines (after the equality sign) of (5.3.7) together with the ρ^2-part in the second line of the term with h_2. The second part consists of the remaining terms, which are due to the correlation between the preliminary test and the main tests.

First consider the size of the procedure. For the size ($b = 0$) the approximation (5.3.7) reduces to

$$\begin{aligned}\varphi(u_\alpha)\{&h_1(c, u_{\delta/2})[\rho + \tfrac{1}{2}c\rho^2 u_\alpha + m(u_\alpha)n^{-1/2}] \\ &+ h_2(c, u_{\delta/2})[\rho\lambda + \tfrac{1}{2}c\rho^2\lambda^2 u_\alpha + m(u_\alpha)n^{-1/2}] \\ &+ h_3(c, u_{\delta/2})[\rho^2\lambda^2 u_\alpha]\}.\end{aligned} \tag{5.3.12}$$

For $\lambda = 1$, the coefficients for h_1 and h_2 are equal and the result agrees with the result in the previous chapter, see formula (4.4.1), p. 57. The behavior of the size may be deduced by analyzing the properties of the functions h_1, h_2 and h_3. This has been done on p. 31 and p. 57, 58, where Figure 4.4.1 shows for $\delta = 0.05$ the functions $h = h_1 + h_2$, h_1, h_3 and $c = h_1(c, \infty) = h(c, \infty)$ (to illustrate the effect of always using the basic main test ($\delta = 0$) without preliminary test). The function h_3, being even in c, is positive for all c. Regarding h_1 and h_2, the reasoning there shows that $0 \le h_1 + h_2 \le h_1 \le c$ for all $u_{\delta/2}$ and $c \ge 0$, and that $h = h_1 + h_2$ and h_1 have a maximum for some value of $c > 0$. The same holds when we replace h_2 by λh_2 or $\lambda^2 h_2$, since $0 \le \lambda \le 1$. (For $c \le 0$, inequalities are reversed since h_1 and h_2 are odd in c.) Therefore, the interpretation of these properties does not change much in the present situation. Using the basic main test for the restricted model without a preliminary test, leads to an error in the size that grows linearly with the deviation of τ (or c) from 0:

$$\varphi(u_\alpha)c[\rho + \tfrac{1}{2}c\rho^2 u_\alpha + m(u_\alpha)n^{-1/2}].$$

Compared to the application of the basic main test, pre-testing leads to a reduction of the error in the size (see (5.3.12)). The error is now redescending, and (for not too large ρ) smaller than the first line of (5.3.12) (the terms with h_1). This holds true because the deviation from the level due to the terms with h_1, is partly compensated by the terms with h_2. This compensation occurs as long as ρ is small enough in order that the ρ^2-terms in the term with h_3 cannot overrule this again.

Notice however, that in situations where $\rho = 0$, all first-order terms cancel. Hence, the $n^{-1/2}$-terms are essential to explain differences between the actual size of the pre-test procedure and the prescribed level.

Now let us pay attention to the power. In case $\lambda = 1$, power gain is (almost only) possible at the cost of a larger size. In the present situation, where $\lambda < 1$, the terms with $\Phi(u_\alpha - b\lambda) - \Phi(u_\alpha - b)$ do not cancel and lead to a possibly substantial increase of power without automatically exceeding the prescribed level.

If the correlation between the two main tests is smaller (λ smaller or ρ larger), indicating that the two main tests differ more, then for smaller λ in combination with a small ρ, a larger (possibly substantial) power gain of the pre-test procedure compared to the power of the second main test can be achieved without getting unacceptable

sizes. For $\rho = 0$ the result reduces to

$$\begin{aligned} &\{\Phi(u_\alpha - b\lambda) - \Phi(u_\alpha - b)\}\{\Phi(u_{\delta/2} - c) - \Phi(-u_{\delta/2} - c)\} \\ &+ \{h_1(c, u_{\delta/2}) + h_2(c, u_{\delta/2})\}m(u_\alpha)\varphi(u_\alpha - b)n^{-1/2} \end{aligned} \tag{5.3.13}$$

and λ occurs only in the leading term. Hence, in this case the difference between the two main tests does not influence the size at all. Therefore, for $\rho = 0$ one can first control the size, which then does not change anymore if the second test is chosen such that λ is smaller, and a substantial power gain can be achieved.

5.4 Numerical example: testing the median

The theory of the preceding section is now illustrated by means of the previously mentioned example, in which a preliminary test is performed to choose which of two main tests is used to test whether the median equals zero. For this example, we compare simulation results regarding power and size of the pre-test procedure with the approximation given in Theorem 5.3.2.

Let $X_1, \ldots, X_n$ be r.v.'s from a continuous distribution with corresponding density f with respect to Lebesgue measure. Consider the restricted model $f(x;\theta) = \varphi(x-\theta)$. The main testing problem in this model concerns testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. From this model it follows that $\psi_{10}(x) = x$. Hence, a first-order optimal test within this model is the Gauss test, corresponding to SK from (5.2.2) with $k = 0$. Its critical value equals u_α, which corresponds to (5.3.6), since $E\psi_{10}^3 = Ek\psi_{10} = 0$. The null hypothesis of a zero median is rejected when $n^{-1/2}\sum_{i=1}^n X_i > u_\alpha$.

Note that we use the Gauss test instead of the t-test. In principle the theory may be extended with additional nuisance parameters in the restricted model to account for an unknown variance, but for this chapter that is not relevant and might distract the attention from the present purpose: showing that power gain is possible and in what kind of situations. For this purpose, the difference between the Gauss test and the t-test is not essential.

In case the restricted model might not hold, we need a more general test. Interpreting θ in the restricted model as the median, we can for any continuous density f rewrite the hypothesis as $H_0^* : \int \xi f dx = 0$ with $\xi(x) = sign(x)$. Here $sign(x) = \pm 1$ for $x > 0$ or $x < 0$, respectively, and $sign(x) = 0$ for $x = 0$. The second main test becomes the sign test, based on the difference between the number of observations strictly larger than 0, and the number of observations strictly smaller than 0. Within the class of continuous distributions, this test is well-known to be uniformly most powerful for testing H_0^* (Lehmann (1986), p. 106, 107). We take $w^* = 0$ in (5.2.5). Since the variance of Ξ exactly equals 1, we do not explicitly need to studentize, and hence $w = w^* = 0$ in SU. This is in accordance with (5.2.6). Again we have $E\psi_{10}^3 = 0$ and $Ew\psi_{10} = 0$ in (5.3.6). So the critical value reduces to u_α again.

However, the test statistic Ξ has a lattice distribution with jumps of order $n^{-1/2}$. Hence, the remainder term of approximation (5.3.7) presented in Theorem 5.3.2 can-

not be of order $o(n^{-1/2})$ if the second main test rejects when $\Xi > u_\alpha$. But a minor modification suffices to achieve the order given by the theorem. An Edgeworth expansion for the distribution of the discrete variable remains valid to order $n^{-1/2}$ when the original distribution function is replaced by its polygonal approximant (which is continuous and equals the original distribution function at midpoints of the lattice and which equals the average of its upper and lower limits at lattice points), (see Feller (1966), p. 540). Using randomization, we can develop a test with the same size and power (to the order considered) as the test which would reject when a r.v., distributed according to the polygonal approximant of the distribution of Ξ, would exceed u_α.

Let $c_{\alpha,n} = u_\alpha n^{1/2} + 1$. Then the nearest lattice point of the distribution of $\sum_{i=1}^n \xi(X_i)$ below $c_{\alpha,n}$ is the integer $k_{\alpha,n}$ given by

$$k_{\alpha,n} = \begin{cases} 2\lfloor \frac{1}{2} c_{\alpha,n} \rfloor & \text{if } n \text{ is even} \\ 2\lfloor \frac{1}{2}(c_{\alpha,n}+1) \rfloor - 1 & \text{if } n \text{ is odd.} \end{cases} \tag{5.4.1}$$

If $p = \frac{1}{2}(c_{\alpha,n} - k_{\alpha,n})$ denotes the difference between the real-valued and the integer-valued "critical values", divided by 2 such that p lies between 0 and 1, then the test which rejects

$$\begin{cases} \text{with probability 1} & \text{if } \sum_{i=1}^n \xi(X_i) > k_{\alpha,n} \\ \text{with probability } 1-p & \text{if } \sum_{i=1}^n \xi(X_i) = k_{\alpha,n}, \end{cases} \tag{5.4.2}$$

has the same size and power as the test based on the "continuous version" of Ξ. The approximation given by Theorem 5.3.2 is applied with the sign test randomized in this way.

Finally, we have to provide the model $f(x;\theta,\tau)$ and the corresponding preliminary test which has to determine whether the Gauss test or the sign test is used to test the main null hypothesis of a zero median. For this testing problem there is considerable freedom of choice. Therefore, we consider a whole class of alternative models $f(x;\theta,\tau)$ for the deviations from the restricted model $f(x;\theta)$. Corresponding to this class, there is a class of preliminary tests that are (first-order) optimal for the considered model. The preliminary test aims to test $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$, where τ parameterizes the deviations from the restricted model against which protection should be offered. It is natural to consider deviations of the following form

$$\begin{aligned} f(x;\theta,\tau) &= (1-\tau)\varphi(x-\theta) + \tau g(x-\theta) \\ &= \varphi(x-\theta)\left\{1 + \tau\left(\frac{g}{\varphi}(x-\theta) - 1\right)\right\} \end{aligned} \tag{5.4.3}$$

for some given function g. If g is a density and τ is positive (between 0 and 1), then this may be interpreted as a contamination model.

There are two conditions which must be satisfied, as explained in Section 5.2. First, it is necessary that $E\xi\psi_{01} = 0$, since $f(x;0,\tau)$ should belong to H_0^*. Secondly,

the side condition of small ρ has to be respected, since otherwise the size of the procedure may deteriorate.

It is easy to construct a class of functions g for which these two conditions are satisfied. Let g be a density with corresponding distribution function G, and let h denote a symmetric density on (0,1) with distribution function H. We consider the class satisfying

$$\begin{aligned} &G = H \circ \Phi, \text{ hence } g = (h \circ \Phi)\varphi \\ &\text{with } h(t) = (2s+1)(2t-1)^{2s} \text{ for } 0 \le t \le 1,\ s = 1, 2, ... \end{aligned} \tag{5.4.4}$$

In order to achieve symmetry, the density h on (0,1) is an even power with its minimum at $t = \frac{1}{2}$. Then g, and hence also $\psi_{01}^*(x) = (g/\varphi)(x) - 1$, is even, which leads to $E\xi\psi_{01} = 0$ and $\rho = E\psi_{10}\psi_{01} = 0$ due to the oddness of both ξ and ψ_{10}. The boundedness of ψ_{01}^* assures that its moments do not deteriorate. Note that in order for $f(x;0,\tau) = \varphi(x)\{1 + \tau\left[(2s+1)(2\Phi(x)-1)^{2s} - 1\right]\}$ to be positive for all x, it is necessary and sufficient that $-1/(2s) \le \tau \le 1$.

From (5.4.3) and (5.4.4) it follows that $\psi_{01}^*(x) = (2s+1)(2\Phi(x)-1)^{2s} - 1$ and $\psi_{01}(x) = \psi_{01}^*(x) I_{22}^{-1/2}$ with $I_{22} = (2s)^2/(4s+1)$. The preliminary test is based on TU with ψ_{01} as given, and with $\rho = 0$, $q = 0$ and $r^* = 0$. Since $E\psi_{10}\psi_{01}^2 = 0$ by oddness of ψ_{10}, $r = r^* = 0$ after studentization. The null hypothesis $\bar{H}_0 : \tau = 0$ is therefore rejected when $|n^{-1/2}\sum_{i=1}^n \psi_{01}(X_i)| > u_{\delta/2}$.

Summarizing, first a preliminary test is performed based on TU with $\psi_{01}(x) = \{\sqrt{4s+1}/(2s)\}\{(2s+1)(2\Phi(x)-1)^{2s}-1\}$, $\rho = 0$ and $q = r = 0$. After acceptance, the main hypothesis is tested by the Gauss test, based on SK with $\psi_{10}(x) = x$ and $k = 0$. After rejection, the randomized sign test is used, based on SU with $\xi(x) = sign(x)$ and $w = 0$. The pre-test procedure (with these tests) is analyzed in the model given by (5.4.3) and (5.4.4).

Now we present numerical results for the case $s = 1$, and $\alpha = \delta = 0.05$. For $n = 25$, 50, 75 and 100, for different values of b between 0 and 1.5, and for c varying between -1 and 2, we performed 100 000 simulations. This implies that the standard deviations for the simulated power differences $\pi^* - \tilde{\pi}$ are at most $(100\,000)^{-1/2} = 0.0032$. If π^* or $\tilde{\pi}$ is at most 0.05, then this reduces to 0.0019, if they are both less than 0.05, then the standard deviation of the difference is not more than 0.001. Observations from the given density were generated using the acceptance-rejection method. As a majorizing function we used a multiple of the standard normal density, since $f(x;0,\tau) \le \varphi(x)\{1+|\tau|2s\}$ for all x. The approximate size difference and power gain are given by Theorem 5.3.2 with $E\psi_{10}^2\psi_{01} = \sqrt{15}/\pi$.

Figure 5.4.1 shows simulated and approximate values for the power or size difference $\pi^* - \tilde{\pi}$ as a function of c for $n = 50$ and for different values of b. For the other values of n, the pattern was not really different, except for the contribution of $n^{-1/2}$-terms. These are important for the size deviations, but are relatively small if power gain is considered.

The size difference is for all parameter combinations considered less than 0.014, and decreases for larger n, equalling at most 0.008 for $n = 100$ and $c \approx 1.5$. (Note

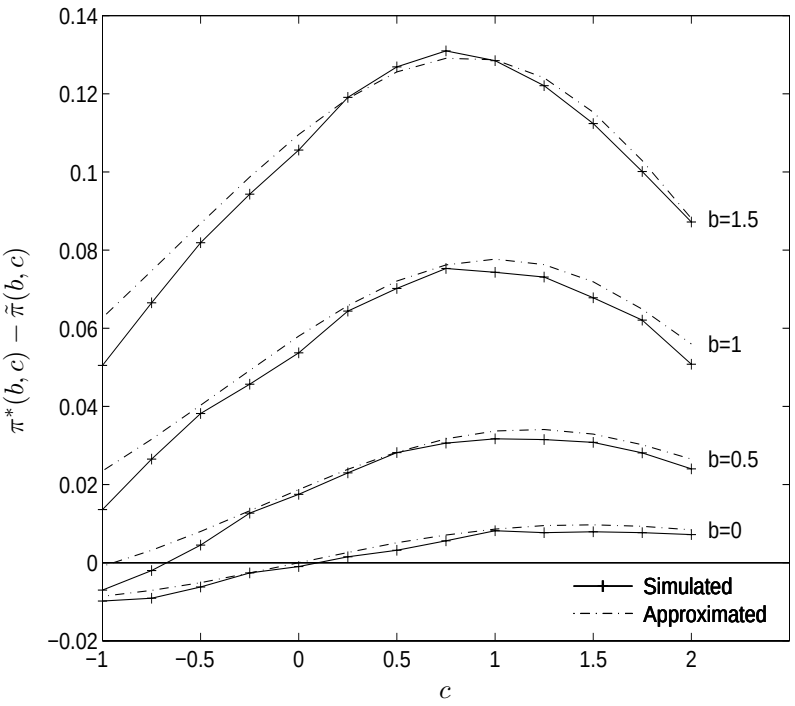

Figure 5.4.1 Power gain (or size difference) for several values of b, as a function of the deviation from the restricted model for $s = 1$, $\alpha = \delta = 0.05$ and $n = 50$.

that this is much better than the size difference of the Gauss test, which is monotone and equals more than 0.02 for $c = 2$ and $n = 100$, almost 0.03 for $n = 50$, and about 0.04 for $n = 25$.) The maximal error between the simulated and approximated value of the size varies between 0.001 for $n = 100$, and 0.002 for smaller n.

For the power gain, the maximal error of the approximate compared to the simulated value varies for $c > 0$ from 0.003 to 0.01 when n decreases from 100 to 25, and for $c < 0$ from 0.01 to 0.02 for decreasing n.

The main result is the substantial power gain which is achieved: for all n its maximal value for $b = 1.5$ equals about 0.13, while the simulated power $\tilde{\pi}$ was only 0.24 for $n = 25$ and 0.29 for $n = 100$. For $c = 0$, where there is no size deviation, the power gain equals for all n more than 0.10, compared to a power $\tilde{\pi}$ of less than 0.33 for the sign test. So indeed, there are situations in which it is possible to gain much power without getting unacceptable size deviations.

References

Albers, W., P. C. Boon, and W. C. M. Kallenberg (1997a). Asymptotic methods for the two-sample problem with a preliminary test of variance. Memorandum 1384, University of Twente.

Albers, W., P. C. Boon, and W. C. M. Kallenberg (1997b). Simple approximations for the size of the Gauss test and the power of a preliminary test of variance. Memorandum 1364, University of Twente.

Albers, W., P. C. Boon, and W. C. M. Kallenberg (1997c). Size and power of pre-test procedures. Memorandum 1421, University of Twente.

Albers, W., P. C. Boon, and W. C. M. Kallenberg (1998). Testing equality of two normal means using a variance pre-test. *Statist. Probab. Lett. 38*, 221–227.

Albers, W., P. C. Boon, and W. C. M. Kallenberg (2000a). The asymptotic behavior of tests for normal means based on a variance pre-test. *J. Statist. Plann. Inference*. To appear.

Albers, W., P. C. Boon, and W. C. M. Kallenberg (2000b). Size and power of pre-test procedures. *Ann. Statist*. To appear.

Albers, W., P. C. Boon, and W. C. M. Kallenberg (2001). Power gain by pre-testing? *Statist. Decisions 19*. To appear.

Arabatzis, A. A., T. G. Gregoire, and M. R. Reynolds Jr. (1989). Conditional interval estimation of the mean following rejection of a two sided test. *Comm. Statist. Theory Methods 18*(12), 4359–4373.

Arnold, B. C. (1970). Hypothesis testing incorporating a preliminary test of significance. *J. Amer. Statist. Assoc. 65*(332), 1590–1596.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Ann. Math. Statist. 15*, 190–204.

Bancroft, T. A. (1964). Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. *Biometrics 20*, 427–442.

Bancroft, T. A. and C.-P. Han (1977). Inference based on conditional specification: A note and a bibliography. *Int. Statist. Rev. 45*, 117–127.

Bancroft, T. A. and C.-P. Han (1980). Inference based on conditionally specified ANOVA models incorporating preliminary testing. In: P. R. Krishnaiah (ed.),

Handbook of Statistics, vol. 1: Analysis of Variance, pp. 407–441. Amsterdam: North-Holland Publishing Company.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families In Statistical Theory*. New York: Wiley.

Barndorff-Nielsen, O. E. and D. R. Cox (1994). *Inference and Asymptotics*. London: Chapman & Hall.

Bickel, P. J., D. M. Chibisov, and W. R. van Zwet (1981). On efficiency of first and second order. *Int. Statist. Rev. 49*, 169–175.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Johns Hopkins University Press.

Cohen, A. (1974). To pool or not to pool in hypothesis testing. *J. Amer. Statist. Assoc. 69*, 721–725.

Easterling, R. G. and H. E. Anderson (1978). The effect of preliminary normality goodness of fit tests on subsequent inference. *J. Statist. Comput. Simul. 8*, 1–11.

Fabian, V. (1991). On the problem of interactions in the analysis of variance. *J. Amer. Statist. Assoc. 86*(414), 362–375.

Farebrother, R. W. (1975). Minimax regret significance points for a preliminary test in regression analysis: Comment. *Econometrica 43*(5-6), 1005–1006.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications*, Volume II. New York: Wiley.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications* (2nd ed.), Volume II. New York: Wiley.

Fisher, R. A. (1990). *Statistical Methods, Experimental Design and Scientific Inference*. Oxford: Oxford University Press. A re-issue of 'Statistical Methods for Research Workers', 'The Design of Experiments', and 'Statistical Methods and Scientific Inference'.

Giles, D. E. A. and V. K. Srivastava (1993). The exact distribution of a least squares regression coefficient estimator after a preliminary t-test. *Statist. Probab. Lett. 16*, 59–64.

Götze, F. (1987). Approximations for multivariate U-statistics. *J. Multivariate Anal. 22*, 212–229.

Gun, A. M. (1969). On the significance level of preliminary tests in some "TE" procedures. *Ann. Inst. Statist. Math. 21*, 373–376.

Gupta, V. P. and S. R. Srivastava (1963). Upper bound for the size of a test procedure using preliminary tests of significance. *J. Indian Statist. Assoc. 7*, 26–29.

Gurland, J. and R. S. McCullough (1962). Testing equality of means after a preliminary test of equality of variances. *Biometrika 49*, 403–417.

Han, C.-P., C. V. Rao, and J. Ravichandran (1988). Inference based on conditional specification: A second bibliography. *Comm. Statist. Theory Methods 17*(6), 1945–1964.

Hodges, Jr., J. L. and E. L. Lehmann (1970). Deficiency. *Ann. Math. Statist. 41*(3), 783–801.

Hoel, P. G. (1984). *Introduction to Mathematical Statistics* (5th ed.). New York: Wiley.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Johnson, N. L. and S. Kotz (1970). *Distributions in Statistics. Continuous Univariate Distributions-1*. New York: Wiley.

Ledwina, T. (1994). Data driven version of the Neyman smooth test of fit. *J. Amer. Statist. Assoc. 89*, 1000–1005.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses* (2nd ed.). New York: Wiley.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. New York: Springer.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer-Verlag.

Markowski, C. A. and E. P. Markowski (1990). Conditions for the effectiveness of a preliminary test of variance. *Amer. Statist. 44*(4), 322–326.

Mehta, J. S. and J. Gurland (1969a). On utilizing information from a second sample in estimating variance. *Biometrika 56*(3), 527–532.

Mehta, J. S. and J. Gurland (1969b). Testing equality of means in the presence of correlation. *Biometrika 56*(1), 119–126.

Moser, B. K. and G. R. Stevens (1992). Homogeneity of variance in the two-sample means test. *Amer. Statist. 46*(1), 19–21.

Moser, B. K., G. R. Stevens, and C. L. Watts (1989). The two-sample t test versus Satterthwaite's approximate F test. *Comm. Statist. Theory Methods 18*(11), 3963–3975.

Nikulin, M. S. and V. G. Voinov (1995). On the problem of the means of weighted normal populations. *Questiio 19*, 93–106.

Pfanzagl, J. (1980). Asymptotic expansions in parametric statistical theory. In: P. R. Krishnaiah (ed.), *Developments in Statistics, vol. 3*, pp. 1–97. New York: Academic Press.

Rao, C. V. and K. P. Saxena (1981). On approximation of power of a test procedure based on preliminary tests of significance. *Comm. Statist. Theory Methods A10*(13), 1305–1321.

Saleh, A. K. Md. E. and P. K. Sen (1983). Asymptotic properties of tests of hypothesis following a preliminary test. *Statist. Decisions 1*, 455–477.

Saleh, A. K. Md. E. and P. K. Sen (1984). Nonparametric preliminary test inference. In: P. R. Krishnaiah and P. K. Sen (eds.), *Handbook of Statistics, vol. 4: Nonparametric Methods*, pp. 275–297. Amsterdam, North-Holland: Elsevier Science Publishers.

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

Sawa, T. and T. Hiromatsu (1973). Minimax regret significance points for a preliminary test in regression analysis. *Econometrica 41*(6), 1093–1101.

Toyoda, T. and T. D. Wallace (1976). Optimal critical values for pre-testing in regression. *Econometrica 44*(2), 365–375.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wehrhahn, K. and J. Ogawa (1978). On the use of the t-statistic in preliminary testing procedures for the Behrens-Fisher problem. *J. Statist. Plann. Inference 2*, 15–25.

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

Summary

In statistical practice, it frequently happens that one is interested in a certain parameter, but that there are doubts about the adequacy of the statistical model. In such a situation the assumptions of the model may be tested by a preliminary test. In this thesis we consider so-called *pre-test procedures*, in which the preliminary test on the adequacy of a given restricted model is followed by a main testing problem on a parameter of interest. If the preliminary test does not reject, then a basic main test that is suitable for the restricted model is performed. If the preliminary test does reject, then a more general main test is used.

The implicit idea behind these kinds of procedures is that researchers often want to use the main test for the restricted model as long as possible, either because that test is simpler or more familiar to them or to their clients, or because that test yields a higher power if the restricted model holds. The preliminary test should protect against inappropriate use of the basic main test. However, the preliminary test may fail to reject and moreover, test statistics for the preliminary test and the main tests may be dependent. Although in practice these two problems are mostly ignored, they do influence the size, which differs from the nominal level of the main test, and the power of the pre-test procedure.

It is the aim of this thesis to gain insight in the consequences of pre-testing for the size and power of the whole pre-test procedure as a function of the underlying parameters. As first-order asymptotics turns out to be insufficient, higher-order asymptotic methods are used to derive transparent expressions which provide the desired qualitative and quantitative insight.

In Chapters 2 and 3, we consider the normal one- and two-sample problem. Since the two-sample t-test for testing the equality of two normal means requires homogeneous variances, a preliminary F-test is often advised to check this assumption. After rejection by the F-test one may use the Welch-Satterthwaite test as an alternative. In the one-sample analogue, the χ^2-test is used as preliminary test for the variance, the Gauss test as basic main test, and the one-sample t-test as alternative main test for the mean.

First, we derive for the one- and two-sample problem approximations for the size of the basic main test and for the power of the preliminary variance test, evaluated under the more general model in which the assumption on the variance(s) does not need to be true. Solving from these approximations the number of observations needed such

that the preliminary test has a sufficiently large power to get adequate protection, shows that very large sample sizes are needed. It is observed that in order to get simple and accurate approximations for the sample size, first-order asymptotics suffices for the preliminary test, while for the main test second-order asymptotics is required.

In Chapter 3 we consider the procedure as a whole, thus also taking into account the dependence between the tests. A transparent approximation for the size and power difference of the pre-test procedure compared to that of the general main test (which has the correct level) is derived. Numerical examples show the usefulness and accuracy of the approximation. It turns out that the result can be decomposed into two parts. One part equals the product of the acceptance probability of the preliminary test and the difference of the rejection probabilities of the two main tests. This part reflects the idea of the previous chapter that in order to control the size of the procedure, the power of the preliminary test should be large when the size of the basic main test differs too much from the nominal level. The second part may be viewed as a correlation term and partly compensates the first part. The power difference turns out to be a multiple of the size difference and hence power gain is only possible at the cost of size violation. Furthermore, the power and size difference are antisymmetric about the null hypothesis value of the preliminary testing problem for the variance. This implies that power gain at a certain value of the variance goes hand in hand with power loss at another value, while the actual value is not known beforehand. For the two-sample case the differences vanish when the sample sizes are equal, and increase when the sample sizes differ more.

Chapter 4 discusses the pre-test procedure for a general family of densities with two parameters. For convenience we assume that they are one-dimensional, but this is not a serious restriction. The preliminary test investigates whether the nuisance parameter equals a given value or not. It is followed by a main test on the parameter of interest, either in the restricted model with the given value of the nuisance parameter, or in the complete model that contains the nuisance parameter. For a class of first-order optimal tests, the pre-test procedure is analyzed. In this analysis, the correlation between the score functions corresponding to the parameter of interest and the nuisance parameter plays a crucial role. Arguments based on first-order asymptotics already show that the pre-test procedure cannot work satisfactorily if this correlation parameter is not small. Hence, it only makes sense to consider the situation of a small correlation parameter. This is used in addition to second-order asymptotics to arrive at a transparent as well as accurate approximation for the size and power difference. In the result the classes of tests and the family of distributions are involved through only four parameters. The conclusions that may be drawn can be interpreted in a similar way as those in the normal one-sample case, which fits in the general situation with a correlation parameter of zero. Moreover, a small gain in power not due to violation of the size is possible now when the (unknown) nuisance parameter equals the supposed value. However, depending on the family of distributions, the size may still exceed the nominal level. Numerical results for several examples from an exponential family of distributions and a symmetric location-scale family illustrate the accuracy of the approximation.

In the final chapter, we consider a situation in which power gain by pre-testing is possible. We assume that the restricted model contains only the parameter of interest, and that the preliminary test tests on a particular property of this model. Deviations from this property are parameterized by a single (nuisance) parameter. After rejection by the preliminary test, we now use an alternative main test that is robust in a much larger model than the model of deviations against which the preliminary test may protect. The procedure is analyzed in the latter model for a class of tests that contains more general alternative main tests than the class of tests considered in Chapter 4. The expression for the power and size difference of the pre-test procedure compared to the general main test, shows that substantial power gain, not merely due to size violation, is possible if the alternative main test really differs from the basic main test. The smaller the correlation between the two main tests, the larger the power gain. As an example, the problem of testing the median is considered, with essentially the t-test as basic main test and with the sign test as alternative. For this example theoretical results are compared with simulation results.

Samenvatting

In de statistiek komt het vaak voor dat men geïnteresseerd is in een gegeven parameter, maar twijfels heeft over de geschiktheid van het statistische model. In zo'n situatie kunnen de aannames van het model getoetst worden met behulp van een voorafgaande toets, een pretoets. In dit proefschrift beschouwen we zogenaamde *pretoetsprocedures*, waarin de pretoets voor de geschiktheid van een gegeven beperkt model gevolgd wordt door een toetsingsprobleem voor de parameter waarin we geïnteresseerd zijn. Als de pretoets niet verwerpt, dan wordt een hoofdtoets uitgevoerd die toegesneden is op het beperkte model. Als de pretoets wel verwerpt, dan wordt een algemenere hoofdtoets gebruikt.

Het impliciete idee achter dit soort procedures is dat onderzoekers vaak de hoofdtoets voor het beperkte model willen gebruiken waar dat maar mogelijk is. Dit is enerzijds omdat die toets eenvoudiger is of omdat zij of hun klanten er beter bekend mee zijn, anderzijds omdat die toets een hoger onderscheidend vermogen haalt als het beperkte model geldt. De pretoets moet beschermen tegen onterecht gebruik van de speciale hoofdtoets. Helaas kan de pretoets wel eens falen, en bovendien kunnen de toetsingsgrootheden voor de pretoets en voor de hoofdtoetsen afhankelijk zijn. Hoewel deze twee problemen in de praktijk meestal genegeerd worden, beïnvloeden ze wel degelijk de onbetrouwbaarheid, die verschilt van de onbetrouwbaarheidsdrempel van de hoofdtoets, en het onderscheidend vermogen van de pretoetsprocedure.

Het doel van dit proefschrift is om inzicht te krijgen in de consequenties van pretoetsen voor de onbetrouwbaarheid en het onderscheidend vermogen van de gehele pretoetsprocedure, als functie van de onderliggende parameters. Aangezien eerste-orde-asymptotiek ontoereikend blijkt te zijn, wordt hogere-orde-asymptotiek gebruikt om doorzichtige uitdrukkingen af te leiden, die het gewenste kwalitatieve en kwantitatieve inzicht verschaffen.

In de hoofdstukken 2 en 3 beschouwen we het normale één- en twee-steekproevenprobleem. Omdat voor het toetsen op gelijkheid van twee verwachtingen uit normale verdelingen met behulp van de twee-steekproeven-t-toets homogene varianties nodig zijn, wordt vaak een voorafgaande F-toets aanbevolen om deze aanname te toetsen. Na verwerpen door de F-toets zou men de Welch-Satterthwaite-toets als alternatief kunnen gebruiken. In het één-steekproefanalogon wordt de χ^2-toets gebruikt als pretoets voor de variantie, de Gauss-toets als speciale hoofdtoets, en de één-steekproef-t-toets als alternatieve hoofdtoets voor de verwachting.

Eerst leiden we voor het één- en twee-steekproevenprobleem benaderingen af voor de onbetrouwbaarheid van de speciale hoofdtoets en voor het onderscheidend vermogen van de pretoets. We doen dit onder het algemenere model waarin de aanname met betrekking tot de variantie(s) niet waar hoeft te zijn. Lossen we uit deze benaderingen het benodigde aantal waarnemingen op zodanig dat de pretoets een onderscheidend vermogen heeft dat groot genoeg is om adequaat bescherming te kunnen bieden, dan blijken zeer grote steekproefomvangen nodig te zijn. Om simpele en accurate benaderingen voor de steekproefomvang te krijgen, volstaat eerste-orde-asymptotiek voor de pretoets, terwijl voor de hoofdtoets tweede-orde-asymptotiek vereist is.

In hoofdstuk 3 beschouwen we de pretoetsprocedure in z'n totaliteit, waarbij we dus ook de afhankelijkheid tussen de toetsen meenemen. Er wordt een doorzichtige benadering afgeleid voor het verschil in onbetrouwbaarheid en onderscheidend vermogen van de pretoetsprocedure ten opzichte van de algemene hoofdtoets (die de juiste onbetrouwbaarheid heeft). Numerieke voorbeelden illustreren de bruikbaarheid en nauwkeurigheid van de benadering. Het resultaat blijkt in twee delen te kunnen worden gesplitst. Eén deel is gelijk aan het produkt van de kans op niet-verwerpen door de pretoets en het verschil tussen de kansen waarmee de twee hoofdtoetsen verwerpen. Dit stuk weerspiegelt het idee uit het vorige hoofdstuk, dat om de onbetrouwbaarheid onder controle te houden, het onderscheidend vermogen van de pretoets groot moet zijn wanneer de onbetrouwbaarheid van de speciale hoofdtoets teveel afwijkt van de onbetrouwbaarheidsdrempel. Het tweede stuk kan gezien worden als een correlatieterm en compenseert gedeeltelijk het eerste stuk. Het verschil in onderscheidend vermogen blijkt een veelvoud te zijn van het verschil in onbetrouwbaarheid en daarom is winst in het onderscheidend vermogen alleen mogelijk ten koste van een grotere onbetrouwbaarheid. Verder zijn de verschillen in onderscheidend vermogen en onbetrouwbaarheid antisymmetrisch rondom de nulhypothesewaarde van het voorafgaande toetsingsprobleem voor de variantie. Dit impliceert dat winst in onderscheidend vermogen voor een bepaalde waarde van de variantie hand in hand gaat met verlies van onderscheidend vermogen voor een andere waarde, terwijl de werkelijke waarde van tevoren niet bekend is. Voor het twee-steekproevenprobleem verdwijnen de verschillen als de steekproefgroottes gelijk zijn, en stijgen ze wanneer de steekproefgroottes meer van elkaar verschillen.

Hoofdstuk 4 bespreekt de pretoetsprocedure voor een algemene familie van dichtheden met twee parameters. Voor het gemak nemen we aan dat deze ééndimensionaal zijn, maar dit is geen serieuze beperking. De pretoets onderzoekt of de nuisance parameter gelijk is aan een gegeven waarde of niet. Daarna volgt een hoofdtoets op de parameter waarin we geïnteresseerd zijn, ofwel in het beperkte model met de gegeven waarde van de nuisance parameter, ofwel in het volledige model waar de nuisance parameter nog in voorkomt. Voor een klasse van eerste-orde-optimale toetsen wordt de pretoetsprocedure geanalyseerd. In deze analyse speelt de correlatie tussen de score functies horend bij de parameter waarin we geïnteresseerd zijn en de nuisance parameter een cruciale rol. Argumenten gebaseerd op eerste-orde-asymptotiek laten al zien dat de pretoetsprocedure niet tot tevredenheid kan werken als deze correlatieparameter niet klein is. Daarom heeft het alleen maar zin de situatie met een

kleine correlatieparameter te bekijken. Dit gebruiken we in aanvulling op tweede-orde-asymptotiek om tot een zowel doorzichtige als accurate benadering te komen voor het verschil in onbetrouwbaarheid en onderscheidend vermogen. In het resultaat komen de klasse van toetsen en de familie van verdelingen voor via slechts vier parameters. De conclusies die we kunnen trekken, kunnen op gelijksoortige manier geïnterpreteerd worden als die in het normale één-steekproefgeval, dat in de algemene situatie past met een correlatieparameter van nul. Bovendien is nu een kleine winst in onderscheidend vermogen, niet als gevolg van een afwijking in de onbetrouwbaarheid, mogelijk, wanneer de (onbekende) nuisance parameter gelijk is aan de veronderstelde waarde. De onbetrouwbaarheid kan echter nog steeds de onbetrouwbaarheidsdrempel overschrijden, afhankelijk van de familie van verdelingen. Numerieke resultaten voor verschillende voorbeelden uit een exponentiële familie van verdelingen en een symmetrische locatie-schaalfamilie illustreren de nauwkeurigheid van de benadering.

In het laatste hoofdstuk beschouwen we een situatie waarin een groter onderscheidend vermogen door pretoetsen wel mogelijk is. We nemen aan dat het beperkte model alleen de parameter bevat waarin we geïnteresseerd zijn, en dat de pretoets op een speciale eigenschap van dit model toetst. Afwijkingen van deze eigenschap worden geparametriseerd door één nuisance parameter. Na verwerpen door de pretoets gebruiken we nu een alternatieve hoofdtoets die robuust is in een veel groter model dan het model van afwijkingen waartegen de pretoets kan beschermen. De procedure wordt in het laatstgenoemde model geanalyseerd voor een klasse van toetsen die algemenere alternatieve hoofdtoetsen bevat dan de klasse van toetsen in hoofdstuk 4. De uitdrukking voor het verschil in onderscheidend vermogen en onbetrouwbaarheid van de pretoetsprocedure ten opzichte van de algemene hoofdtoets, laat zien dat een substantiële winst in onderscheidend vermogen, niet alleen ten gevolge van een te grote onbetrouwbaarheid, mogelijk is als de alternatieve hoofdtoets echt verschilt van de speciale hoofdtoets. Hoe kleiner de correlatie tussen de twee hoofdtoetsen, hoe groter de winst in onderscheidend vermogen. Als voorbeeld beschouwen we het toetsen van de mediaan, met in essentie de t-toets als speciale hoofdtoets en met de tekentoets als alternatief. Voor dit voorbeeld worden theoretische resultaten vergeleken met simulatieresultaten.

Curriculum Vitae

Petronella Christine Boon, roepnaam Pieta, werd geboren op 2 december 1972 te Hengelo (O). Daar bezocht zij vanaf 1984 de O.S.G. Bataafse Kamp, waar zij in 1990 haar middelbare schooltijd afsloot met het diploma Gymnasium B.

Daarna ging zij Toegepaste Wiskunde studeren aan de Universiteit Twente. Ze besloot af te studeren bij de leerstoel Statistiek van prof. dr. W. Albers, onderdeel van de vakgroep Stochastiek en Operationele Research. Van september 1994 tot april 1995 voerde zij een gecombineerde stage- en afstudeeropdracht uit bij de sectie Biometrie van N.V. Organon te Oss. Daar werkte zij onder dagelijkse begeleiding van dr. C. B. Roes aan de analyse en efficiëntie van proefopzetten. Haar afstudeerverslag was getiteld "The three-period crossover design in phase I clinical research: analysis and efficiency issues". In juni 1995 studeerde ze cum laude af.

In september 1995 begon zij als Assistent in Opleiding aan het promotieonderzoek, dat uitgevoerd werd onder begeleiding van prof. dr. W. Albers en dr. W. C. M. Kallenberg en dat geresulteerd heeft in dit proefschrift.

Naast haar studie en werk heeft muziek steeds een belangrijke rol gespeeld. Zo studeerde zij naast haar wiskundestudie enkele jaren piano aan de Hogeschool Enschede, sector Conservatorium, waar zij het propedeutisch examen Docerend Musicus haalde. Tijdens haar AiO-schap was zij gedurende anderhalf jaar secretaris van pianovereniging Utopiano.

Na haar promotie treedt zij in dienst van N.V. Organon te Oss, wederom bij de sectie Biometrie.