UNIVERSITEIT
GENT

Faculteit
Landbouwkundige en
Toegepaste Biologische
Wetenschappen

Academiejaar 2000 – 2001

Nonparametrical tests based on sample space partitions

Niet-parametrische testen gebaseerd op partities van de steekproefruimte

door
ir. OLIVIER THAS

Thesis submitted in fulfillment of the requirements for the degree
of Doctor (Ph.D.) in Applied Biological Sciences

Proefschrift voorgedragen tot het bekomen van de graad van
Doctor in de Toegepaste Biologische Wetenschappen

op gezag van
Rector: **Prof. Dr. ir. J. WILLEMS**

Decaan:
Prof. Dr. ir. O. VAN CLEEMPUT

Promotor:
Prof. Dr. J.P. OTTOY

voor Ingeborg

Copyright

Gent, april 2001

De promotor:

Prof. Dr. J.P. Ottoy

De auteur:

ir. Olivier Thas

Dankwoord

Het schrijven van een doctoraatsthesis mag dan wel het resultaat van een heel individualistisch proces zijn, het zou ontegensprekelijk nooit tot stand gekomen kunnen zijn zonder de hulp en de steun van anderen. Mijn doctoraat is zeker niet verschillend in dit opzicht. Iedereen bedanken is echter bijna onmogelijk, maar de voornaamsten wens ik hier toch te vernoemen.

Vooreerst wil ik mijn promotor, Prof. Dr. J.P. Ottoy, uitdrukkelijk bedanken voor zijn onafgebroken steun. Hij heeft steeds in mij geloofd en me een grote wetenschappelijke vrijheid gegeven. Door me regelmatig de mogelijkheid te bieden naar internationale congressen te gaan, of werkbezoeken aan buitenlandse universiteiten te brengen, heb ik zeker onmisbare sleutelideeën opgedaan en interessante discussies gevoerd die tot het voorliggende onderzoeksresultaat bijgedragen hebben.
With respect to these working visits, I defenitely owe many thanks to Prof. Peter Diggle (Lancaster University, UK) and to Prof. Joe Romano (Stanford University, USA).

Ook wens ik Prof. Marc Aerts (LUC) te bedanken voor zijn enthousiasme en interesse die hij getoond heeft voor mijn onderzoek. Aan de andere docenten van het Masters of Science in Biostatistics programma van het LUC kan het ook geen dankwoord ontbreken; hun inzet en enthousiasme hebben er immers toe geleid dat de statistiek-microbe me gebeten heeft.

Uiteraard zou ik mijn inzet en optimisme niet hebben kunnen handhaven indien de sfeer binnen de Biomath groep niet zo goed geweest zou zijn. Daarom dank ik iedereen die de sinds 1995 de revue aldaar gepasseerd is. Van de “oude garde” wil ik zeker Lieven, Danny en Hans hartelijk bedanken. Van Lieven heb ik waarschijnlijk het meeste geleerd over de gastvrijheid binnen de assistentenjob. Danny herinnerde me er bijtijds aan dat de andere collega’s jaarlijks mijn naam vergeten ... en uit Hans’ energieke woordenvloed viel ook steeds wat te leren. Onder de huidige collega’s bevindt zich de meest frequente bezoeker van mijn voice-mail, voor wie zeker geen woordje van dank mag ontbreken; zij heeft me immers gedurende de laatste zware maanden van het schrijfwerk beschermd

tegen onwetende tijdrovers. En, last but not least, dank ik nog Annie, Ronny, Roos en Viv voor hun goede zorgen.

Na al deze professionele dankbetuigingen, zijn we toe aan de familie en de vrienden die ik allen hartelijk dank voor hun steun, begrip en gezelligheid.
Een bijzonder woordje van dank wil ik richten tot mijn ouders, die steeds in mijn “studies” zijn blijven geloven en die mij ononderbroken gesteund hebben. Mama, Papa, aangezien ik denk dat het nu wel mijn laatste thesis zal zijn, en er dus geen dankwoorden als deze meer zullen volgen, wil ik jullie nogmaals bedanken voor jullie vriendschap en alle mogelijkheden die jullie in mijn eerste 30 jaren geboden hebben.
Ik kan het ook niet nalaten mijn schoonouders te bedanken voor hun vriendschap, begrip en vertrouwen. Het is immers niet alleen door de lekkere wijn dat dit werk tot stand is gekomen.

Ingeborg, zonder jouw begrip, geduld en ondersteuning zou het me nooit gelukt zijn! Je geduld was onbegrensd wanneer ik weer eens tot diep in de nacht moest doorwerken, je begrip is onnavolgbaar voor een man die een “onbegrijpelijk boek” aan het schrijven was en je steun tijdens de spreekwoordelijke laatste loodjes hebben het me ongetwijfeld wat lichter gemaakt. Woorden schieten tekort om mijn dank en bewondering uit te drukken; aan jouw draag ik graag deze thesis op.

Mortsel, april 2001

Olivier

Contents

Abbreviations

AD	Anderson-Darling
AIC	Akaike's information criterion
BIC	Bayesian information criterion
CDF	cumulative distribution function
CL	Chernoff-Lehmann
CONS	complete orthonormal system
df	density function
EDF	empirical distribution function
GOF	goodness-of-fit
GQF	generalized quadratic form
H	Hoeffding
i.i.d.	identically and independently distributed
IQS	interquadrant statistic
KL	Kullback-Leibler
KS	Kolmogorov-Smirnov
LL	log-log
LR	likelihood ratio
pAIC	pseudo-AIC
PF	Pearson-Fisher
PS	Puri-Sen
RR	Rao-Robson
rv	random variable
SS	sample space
SSP	sample space partition
SW	Shapiro-Wilk
TL	Tukey-Lambda

CHAPTER 1

Introduction

This thesis is concerned with probably one of the oldest problems in statistical science: Goodness-of-Fit (GOF). One century ago, Pearson (1900) introduced the well known χ^2-statistic for testing the hypothesis that a given sample has arisen from a specified distribution. At that time, many of the statistical methods that were developed, were based on the assumption that the data are normally distributed. Also around that time, Karl Pearson, in the course of his "Mathematical Contributions to the Theory of Evolution", abandoned the assumption that biological populations are normally distributed. As an alternative he introduced a flexible system of distributions, known as the Pearson curves. Now that there was an alternative to the normal distribution, it was of practical importance to have a formal statistical method to decide whether the data possibly came from a given distribution or not. Statistical tests that address this type of problems, are GOF tests. Ever since, the literature on the topic has kept on growing, and many other statistical methods and tests have been published.

Today, GOF tests are still a very widely applied group of tests. In almost all popular statistical software, at least some of them are implemented. Even Pearson's original question, "is the data normally distributed?", is by many experimental researchers still applied regularly. The reason may be that most statistical methods that are implemented in the commercial statistical software, are only valid under the assumption of normality. Thus, a researcher who wants to do e.g. an independent samples t-tests, will first perform a GOF test on his data to assess the normality assumption. Of course there are many more situations in which a GOF test forms the solution, but these will be handled later.

Although nowadays the original Pearson χ^2-test is almost exclusively applied to discrete data, for which Pearson's test statistic is calculated as a "distance" between the observed frequencies in the sample, and the expected frequencies when the hypothesized distribution would be the true underlying distribution, in earlier days it was also applied to continuous data after having discretized the observations in categories, such that the corresponding frequencies could be computed from the data and Pearson's statistic be calculated. In the former setting, not many problems arise, but in the latter situation of continuous data, many questions to be solved remained: in how many groups must the data be categorized, where and how must the group boundaries be placed? During the time that most of these questions were being answered in the literature, other GOF statistical tests were proposed, specifically to handle continuous data without having to discretize them first. Many of these tests are based on the Empirical Distribution Function (EDF), e.g. the well known Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1939). Of particular interest for this thesis, are the EDF-based tests that are still clearly related to the original χ^2-test. These tests are specifically constructed for continuous data, but they may also be considered as a result of Pearson's statistics applied to data that are discretized according to a very specific system. The Anderson-Darling statistic (Anderson and Darling, 1952) is the most important example.
Almost independently of the previous mentioned evolution, (Neyman, 1937) constructed a GOF test for continuous data which he called "smooth". Two important properties are associated with Neyman's smooth test: first, it is in some sense optimal, and second, when the GOF null hypothesis is rejected, by studying the components of the test statistic some features of the difference between the true and the hypothesized distribution may be distinguished (e.g. the means are different, or the skewnesses are different, ...). For many years there has not been much scientific interest in these type of tests, until the papers by Thomas and Pierce (1979), Kopecky and Pierce (1979). Fundamental in this type of tests is that the hypothesized and the true, but unknown distribution are embedded in an exponential family of distributions, indexed by an s-dimensional parameter. The test statistic is correspondingly decomposed into s terms. The higher the order s is taken, the more distributions can be represented by the family and the more sensitive the test becomes to a broader range of alternatives, but on the other hand, the higher s the less powerful the test becomes for low-order alternatives. In order to overcome the problem of choosing an appropriate value for s, Ledwina (1994) proposed to make the test data-driven, i.e. the order s is estimated from the data. The last 5 years many such data-driven tests are developed, and they seem to have desirable properties.

Up to now we only discussed the oldest, original GOF problem of deciding whether or not a given sample may have been generated by a specified distribution. This problem is sometimes also known as the one-sample GOF problem. This is however not the only GOF problem. Another important member is the 2-sample GOF problem, where the question is to test the hypothesis that two independent samples come from the same population or have the same distribution. Again tests based on both Pearson's χ^2 statistic and on EDF-based statistics are constructed. Many of these methods are very similar to those used for the one-sample problem, e.g. once more a Kolmogorov-Smirnov and an Anderson-Darling (Scholz and Stephens, 1987) statistic exist, as well as Neyman's

smooth tests (Eubank, LaRiccia and Rosenstein, 1987) and its data-driven versions.
The tests named so far for the 2-sample problem are more or less omnibus tests, i.e. they are sensitive to almost all alternatives to the null hypothesis. Sometimes one is especially interested in a specific alternative, e.g. a shift in location or a change in dispersion between the distribution. Probably the best known test for testing a location shift is the classical t-test (under the assumption of normality and equality of variances), or the nonparametric Mann-Whitney test (Mann and Whitney, 1947; Wilcoxon, 1945).
The null hypothesis of equality of k distributions is referred to as the k-sample problem. Most of the above mentioned techniques are extended to this setting as well. Among the most recent methods, only the data-driven 2-sample test of Janic-Wróblewska and Ledwina (2000) is not yet extended.

Yet another problem, that is nowadays recognized as a GOF problem, is testing for independence, i.e. the null hypothesis is that 2 variables are independent. Originally, this type of question was addressed in the context of regression, even as soon as in the beginning of the 19th century. It was Galton, who lived from 1822-1911, who invented the correlation coefficient which is today known as the Pearson's product moment correlation coefficient, who discovered the relation between correlation and regression. In the early days the independence problem was studied completely in the bivariate normal distribution. Under this assumption, the first statistical test for testing whether or not the correlation coefficient is zero was proposed by Fisher. A general restriction of the use of the Pearson correlation coefficient is that it only measures a linear dependence between the two variables, and thus tests based on this measure are not omnibus.
Once the tests for the general GOF problem began to appear, one started to see the similarity between both problems, and analogous omnibus tests were developed. Again, most relevant to this thesis are those methods that are EDF-based but could be placed in the framework of testing independence between discrete variables. The tests of Hoeffding (1948) and Blum, Kiefer and Rosenblatt (1961) are such examples. Also smooth tests (e.g. Koziol, 1979; Eubank et al., 1987) and their data-driven versions (Kallenberg, Ledwina and Rafajlowicz, 1997; Kallenberg and Ledwina, 1999) are developed and will turn out to be important for the work presented in this thesis.

Most of the methods referred to in the previous paragraphs are omnibus tests, i.e. the tests are sensitive to almost all alternatives to the null hypothesis. Although this is considered to be a desirable property, it also directly implies that when the null hypothesis is rejected, still no answer is given towards the reason for the rejection, i.e. what the true alternative may be. E.g. when the 2-sample null hypothesis is rejected, it may be mainly caused by a difference in means or a difference in dispersion, etc. Test statistics that can be decomposed into interpretable components may be a solution.
Another important property from a practical point of view is that a test is nonparametric (distribution-free). This means that its null distribution does not depend on the hypothesized, nor on the true distribution. This makes the calculation of the critical values and the p-values much more easier. Many tests have this property asymptotically, others have it even for finite sample sizes.
Although already many tests exist with the above mentioned characteristics, there is still

ongoing research in this field. The reason for this is that within the most general omnibus setting, except for a few specific simple null hypotheses, there is no Uniformly Most Powerful (UMP) test available. Hence, for finite sample sizes, some tests will have better powers under some alternative hypotheses, and others will have better powers under other alternatives, but none has the highest power under all alternatives. This leaves the question open for a distribution-free omnibus consistent test with good "overall" power properties.

Despite the existence of other forms of GOF problems (e.g. the symmetry problem), in this work the attention is restricted to the three above mentioned hypotheses: the one-sample, the k-sample and the independence problem. We have constructed a new family of nonparametric omnibus consistent tests which are all based on the same principle of repeatedly partitioning the sample space, and applying χ^2-tests to the contingency tables that are induced by the partitioning. By changing the size of the sample space partition (SSP) the power characteristics of the test may change. Since the flexibility of letting the SSP size choose by the user is sometimes interpreted as a disadvantage, we have proposed a data-driven version as well. In this way a good partition size is chosen by the data. The tests are closely related to tests of the Anderson-Darling type, or, in some cases, may be considered as a generalization of them. Moreover, the new tests also partially fill in the gap between these Anderson-Darling type tests and Neyman's smooth tests.
The whole class of tests is referred to as Sample Space Partition tests (SSP tests).

In **Chapter 2** the three aforementioned GOF problems are explained in an informal way. Simultaneously, examples are introduced. These examples will be used later to illustrate the methods that are developed in this thesis.

Some basic terminology and notation, as well as a formal description of the GOF problem are given in the beginning of **Chapter 3**.

In **Chapter 4** an overview of the literature on tests for the three GOF problems is given. Since there is a very vast literature on the subject, it is definitely not our intention to give a complete survey of all GOF methods. We have chosen to present those techniques that are relevant for the methods that are developed in this work. Although there exist exploratory tools to assess GOF in an informal way (e.g. by means of a graphical exploration), the literature study is mainly limited to statistical tests. A distinction is made between three types of tests: χ^2-type tests, EDF-based tests and smooth tests.

A new test for the one-sample GOF problem is developed in **Chapter 5**: the SSP test. All concepts and principles that serve as the cornerstone of all SSP tests, are first carefully introduced. Next, the SSP test is constructed, and its asymptotic null distribution is proposed. A distinction between a simple and a composite null distribution must be made. The relation with the original Anderson-Darling test (Anderson and Darling, 1952) is shown. Apart from the more theoretical discussion of the SSP test, the results of a power study are shown as well. This is especially of interest to understand the behaviour of the test with finite sample sizes. Also the new tests are applied to the data sets introduced in Chapter 2. Further a data-driven version of the SSP test is constructed.

For the k-sample problem, a new SSP test is presented in **Chapter 6**. Since it is based on the same principles as those for the one-sample SSP test, the construction can be kept more concise. The SSP test now turns out to be a rank test. Both its asymptotic and its exact null distribution are presented. It is shown that the k-sample Anderson-Darling statistic (Scholz and Stephens, 1987) is a special case of the SSP statistic. Again a data-driven version is constructed. Further, in a limiting case, the SSP statistic is shown to be a Neyman's smooth test statistic (Eubank et al., 1987). A data-driven version in the sense of Ledwina (1994), has the potential of leading to a generalization of a recently published 2-sample data-driven smooth test of Janic-Wróblewska and Ledwina (2000) to the k-sample problem. A decomposition of the SSP test statistic is given, which may be considered to indicate an alternative when the null hypothesis is rejected.
The new tests are applied to the examples introduced earlier, and a power study is presented.

In **Chapter 7** the SSP test for independence is presented. Again it is based on the same central idea of repeatedly partitioning the sample space. Since the test is again a rank test, both a permutation test and a test based on its asymptotic null distribution are considered. A relation with the Anderson-Darling test and the tests of Hoeffding (1948), Blum et al. (1961) is shown. Also a data-driven version is defined. A power study is included to compare the new tests with classical tests.

Most of the theory that is developed in this thesis is restricted to univariate random variables. E.g. in Chapter 5 the SSP one-sample GOF test may be used to test univariate normality. It is however also briefly indicated how the methods can be extended to deal with multivariate situation, e.g. testing for multivariate normality.
Further, the core of the SSP statistics is here mostly taken to be a Pearson χ^2 statistic, but most of the theory remains valid when the core statistic is replaced by another member of the family of Power Divergence Statistics of Cressie and Read (1984). Also this extension is only briefly mentioned in the respective chapters.

Although in Chapters 5, 6 and 7 many of the properties of the SSP statistics are discussed in a rather theoretical way, we feel that the importance of the work that is presented in this thesis, is found in the ideas that lay beneath and its applicability to practical problems, which is shown here by means of examples and extended simulation studies. Moreover, the "proofs" that are presented in this thesis are actually meant as sketches of the proofs. Even, in some instances, heuristic proofs are given.

In the last chapter (**Chapter 8**) related topics for further research are presented, as well as some concluding remarks.

CHAPTER 2

Three Goodness-of-Fit Problems: Introduction through Examples

The purpose of this chapter is to give an informal introduction to the three types of GOF problems that will be handled in the remainder of this thesis. This will not be achieved by discussing the three types in sequence, but rather by introducing the examples one after the other, and where relevant the GOF problems will be discussed in terms of the respective examples. Most of the examples serve as illustration to more than one type of GOF problem. They will also be used in the subsequent chapters to illustrate the new statistical tests proposed in this work.

While presenting the examples, some of the most classical statistical methods will already be performed in order to get at least some insight in the data.

2.1 Chemical Concentration Data

In a study on the effect of environmental pollutants on animals, Risebrough (1972) gives data on the concentration of several chemicals in the yolk lipids of pelican eggs. The data considered here are the PCB (polychlorinated biphenyl) concentrations for 65 Anacapa birds. The complete data is presented in Table 2.1. The example will be referred to as the *Chemical Concentration Data*.

Table 2.1: Concentration of PCB in the yolk lipids of 65 pelican eggs.

Concentration												
452	184	115	315	139	177	214	356	166	246	177	289	175
324	260	188	208	109	204	89	320	256	138	198	191	193
305	203	396	250	230	214	46	256	204	150	218	261	143
132	175	236	220	212	119	144	147	171	216	232	216	164
199	236	237	206	87	205	122	173	216	296	316	229	185

In the original study the mean PCB concentration in Anacapa eggs was compared to the mean concentration in eggs of other birds, but here we will not use the data for that purpose. Many researchers, after having collected this type of data, would like to make some explorative summaries of the data. One of the first quantities that are often computed, is the sample mean of the observations, as well as a, say 95% confidence interval for the mean. When this type of analysis is done within standard software, or by hand based on formula's presented in most basic statistical handbooks for researchers, the confidence interval will most probably be given by

$$\left[\bar{X} - t_{0.025;n-1}\sqrt{\frac{S^2}{n}}, \bar{X} + t_{0.025;n-1}\sqrt{\frac{S^2}{n}}\right], \tag{2.1}$$

where $\bar{X}$ and S^2 are the sample mean and the sample variance of the n observations, respectively, and $t_{0.025;n-1}$ is the $100(1-0.025)\% = 97.5\%$ percentile of a t-distribution with $n-1$ degrees of freedom. For finite sample sizes the coverage of this confidence interval is however only correct when the data are normally distributed. Also, when in the original study the data for the Anacapa birds is compared to other data, a t-test might be used, for which the same assumption of normality must be satisfied.
One may make a point here that the parametric confidence interval of Equation 2.1 is not the only solution. There exist e.g. rank based intervals as well (e.g. Puri and Sen, 1971), or bootstrap intervals (e.g. Efron and Tibshirani, 1993; Davison and Hinkley, 1997), for which no normality assumption is needed. Indeed, this is true, but on the other hand, when the data is normally distributed, then the parametric interval of Equation 2.1 is the optimal interval, in the sense that it is the most narrow interval among all unbiased intervals. Furthermore, one might argue that with the Chemical Concentration Data, there is actually no need for testing normality as long as the purpose is to construct a confidence interval of the mean, for with 65 observations, by the central limit theorem, the distribution of the sample mean may have sufficiently converged to a normal distribition. Indeed, this argument may hold in some occasions, but when the data is e.g. highly skewed, 65 observations may still be too small (Boos and Hughes-Oliver, 2000).

Thus, a researcher will perform first a test for normality, e.g. a Kolmogorov-Smirnov (Kolmogorov, 1933; Smirnov, 1939), or a Shapiro-Wilk (Shapiro and Wilk, 1965) test. The testing problem just described is the **one-sample GOF problem**. More generally, the null hypothesis of the one-sample problem is

H_0: the data are distributed according to a *specific* distribution.

At this point, we would like to draw the attention to the meaning of *specific*. In the example at hand the specific distribution is *a* normal distribution, but a normal distribution is characterized by two parameters: the mean μ and the variance σ^2. Since it is however not known a priori what the true mean and variance are, it can not be specified in the null hypothesis, resulting in a null hypothesis which only states the form of the distribution. Hence the null distribution actually covers an infinite number of possible normal distributions, which are all of the same form. Such a set of distributions, parameterized by some parameters, is called a parametric family of distributions.
In practice the one-sample test is performed with the unknown mean μ and variance σ^2 replaced by their sample counterparts: the sample mean $\hat{\mu} = \bar{X}$ and the sample variance $\hat{\sigma}^2 = S^2$.

When the parameters would have been known a priori, the distribution that is specified under the null hypothesis would have been specified completely. In this case, the null hypothesis is called a **simple null hypothesis**. From a theoretical point of view this is the most easy situation. When, on the other hand, the parameters are unknown and must be estimated from the data first, and thus only a parametric family is specified under the null hypothesis, then the null hypothesis is called a **composite null hypothesis**. The latter situation is what occurs most frequently in practical situations, but the corresponding theory is heavier, i.e. the null distributions of the test statistics are often more difficult to find. E.g. the Kolmogorov-Smirnov statistic is under a simple null hypothesis distribution-free, i.e. its null distribution does not depend on the distribution specified under the null hypothesis. But when only a family is specified, then its null distribution depends both on the parametric family and on the parameters that must be estimated. For the present example of testing normality with unknown mean and variance, the null distribution is given by Lilliefors (1967).

The GOF of the chemical concentration data has been analyzed before by some authors (Best and Rayner, 1985; Rayner and Best, 1989; Thomas and Pierce, 1979).

A histogram of the data is shown in Figure 2.1(a). On the same graph a kernel density estimator of the density is shown as well. The kernel density estimation was performed with a Gaussian kernel with a bandwidth determined by means of the Unbiased Cross Validation method (Silverman, 1986; Venables and Ripley, 1997). Especially the histogram suggests that the distribution might be a bit skewed, but this must be interpreted with care since the form of the histogram, when based on a moderate sample size, depends strongly on the break points (it is indeed possible to construct a histogram which does not suggest skewness (figure not shown)). The skewness indication is not contradicted by the density estimation. The box plot (Figure 2.1(c)), however, shows three outliers with rather high PCB concentrations. This may at least partially explain the skewed histogram and density estimate. Apart from these outliers, the remainder of the box plots looks very symmetric. Also the QQ-plot (Figure 2.1(b)) shows the asymmetric form, but also here it may be caused by the three outliers.

The Kolmogorov-Smirnov test gives $p = 0.0521$, i.e. a nearly non-significant result at the

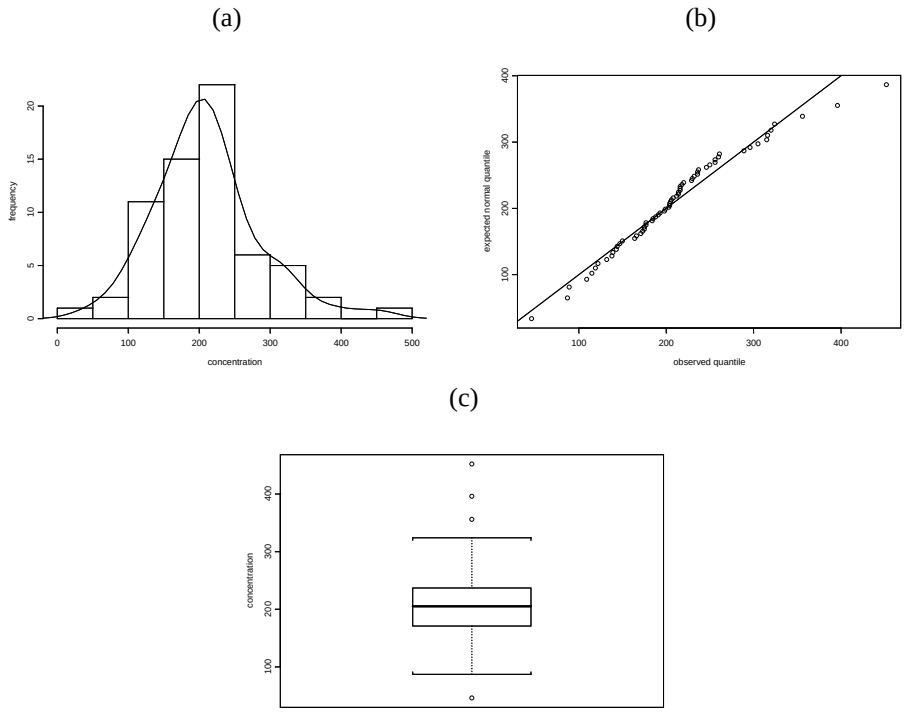


Figure 2.1: A histogram with a density estimation (a), the normal QQ-plot (b) and the box plot (c) of the Chemical Concentration Data.

traditional 5% level of significance. Nevertheless, $p = 0.0521$ still shows much evidence in the direction of nonnormality. Thomas and Pierce (1979) performed an ordinary χ^2 test with 8 and 10 equiprobably cells, both resulting in non-significance. On the other hand, they also performed a Neyman's smooth test which did give a significant result. Later Best and Rayner (1985), Rayner and Best (1989) applied a similar smooth test on the data, based on components 3 up to 6, which clearly showed a significant deviation from normality. In particular, since especially their first component was rather high, they concluded that the departure from normality is due to the asymmetry around the mean. Details on the tests referred to above, will be given in subsequent chapters.

2.2 Singer Data

A similar GOF problem arises with the Singer dataset, which is used by Cleveland (1993) to illustrate Trellis graphs. The dataset consists of heights of singers in the New York Choral Society. We only consider the first group of 35 alto. The data are presented in

Table 2.2

Table 2.2: Heights (in inches) of 35 Alto.

Height														
65	62	68	67	67	63	67	66	63	72	62	61	66	64	60
61	66	66	66	62	70	65	64	63	65	69	61	66	65	61
63	64	67	66	68										

Thus the question that has to answered is whether or not the heights of alto is normally distributed. Figure 2.2 shows the histogram with kernel density estimate (a), the QQ-plot (b), and the box plot (c) of the data. None of these graphs suggests any severe deviation from normality, especially considering that the data set contains only 35 observations.

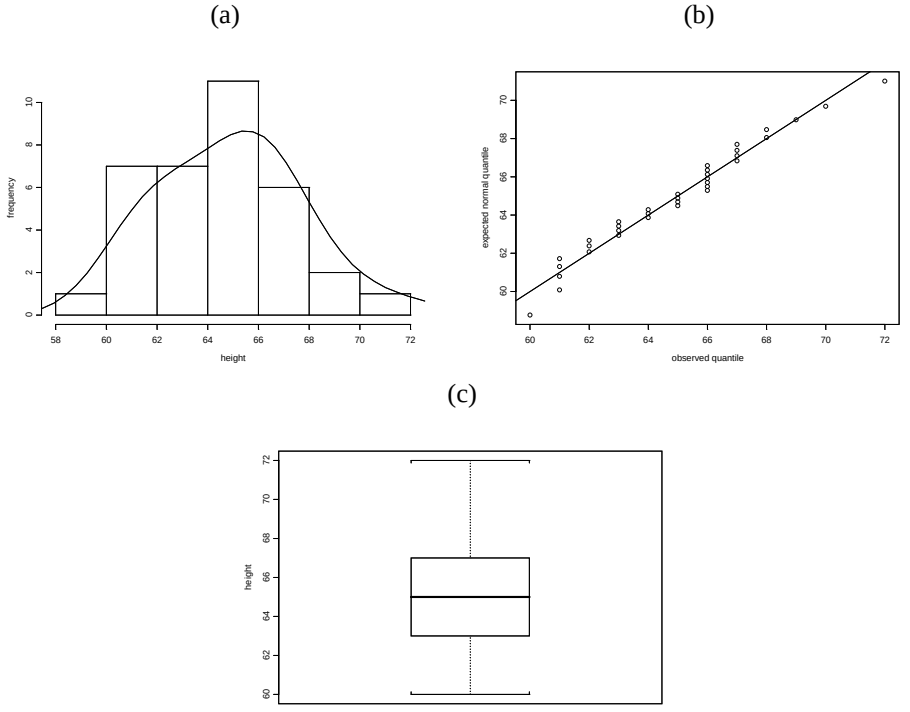


Figure 2.2: A histogram with a density estimation (a), the normal QQ-plot (b) and the box plot (c) of the Singer Data.

We performed both a Kolmogorov-Smirnov test (with Lilliefors correction) and a Shapiro-Wilk test on the data, resulting in $p = 0.309$ and $p = 0.379$, respectively. We also applied a data-driven smooth test of Inglot, Kallenberg and Ledwina (1997), Kallenberg and Ledwina (1997), which gave $p = 0.410$. Since none of the p-values is small, in conclusion,

there does not seem to be much evidence against the hypothesis of normality.

2.3 Cultivars Data

The Cultivars Dataset is taken from Karpenstein-Machen, Honermeier and Hartmann (1994), Karpenstein-Machan and Maschka (1996). It has also been analyzed by Piepho (2000). This example will be referred to as the *Cultivars Data*.

The data set contains the yields (in tons per hectare) of two triticale cultivars: Alamo and Modus. Yields on both cultivars are obtained in 19 different environments. For each environment, a fertility score ("Ackerzahl" (AZ)) was recorded. The data are presented in Table 2.3.

Table 2.3: Yields (in tons per hectare) of two cultivars, and the fertility score (AZ) from 19 environments.

	yield		
Environment	Alamo	Modus	AZ
1	98.250	96.200	61
2	112.950	115.400	60
3	66.875	69.175	39
4	106.500	123.900	82
5	64.800	53.750	30
6	82.900	88.350	55
7	96.433	101.033	35
8	78.950	82.650	75
9	74.200	80.000	28
10	71.600	79.300	42
11	88.550	86.250	28
12	93.650	95.550	42
13	75.000	71.300	54
14	94.450	100.450	80
15	95.033	98.067	85
16	84.150	80.150	33
17	93.350	97.200	50
18	64.650	60.000	24
19	67.750	70.600	45

The aim of the study was to assess if there is a difference between both cultivars, and to check whether or not there is an association between the AZ-score and the difference in yield. The former question may be solved by performing a paired t-test on the paired data. An assumption underlying a paired t-test is that the difference between the yields of the cultivars is normally distributed. Again this is a one-sample GOF problem with a composite null hypothesis.

Piepho (2000) performed a Shapiro-Wilk test on the difference in yields to assess normality. Since the corresponding $p = 0.2548$, the normality assumption was accepted. A histogram with a kernel density estimate, the QQ-plot and the box plot are shown in Figure 2.3, and they seem to support the conclusion. Note that some irregularities may be observed in the graphs, but this is not uncommon because of the natural variability in small samples ($n = 19$).

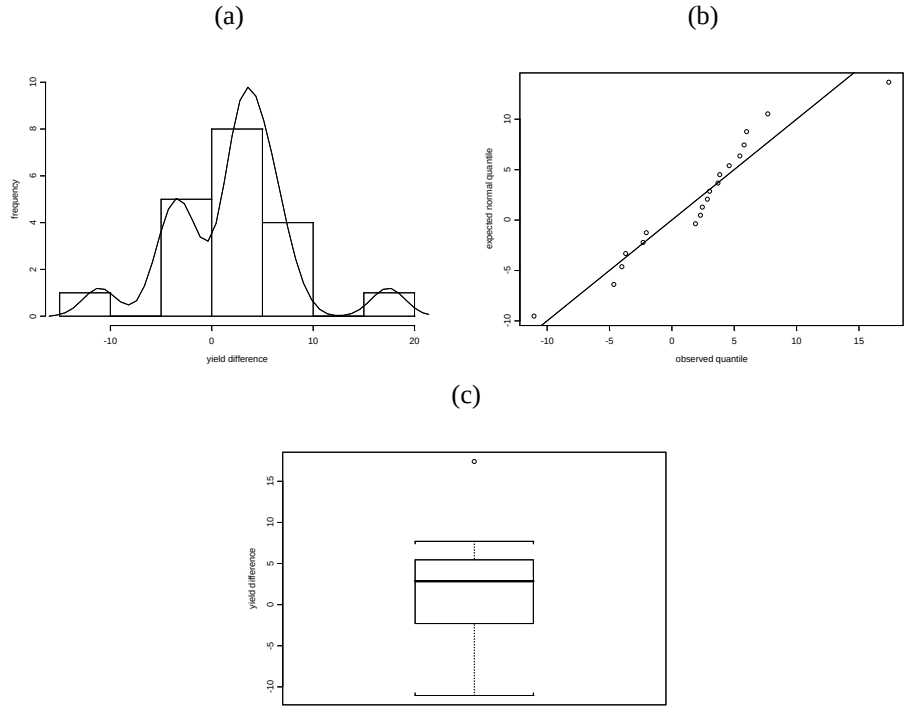


Figure 2.3: An histogram with a density estimation (a), the normal QQ-plot (b) and the box plot of the Cultivars Data.

The problem of assessing an association between two variables is the **independence problem**. Let X and Y denote the two variables, then the null hypothesis is

$$H_0\text{: } X \text{ and } Y \text{ are independent .}$$

Piepho (2000) performed hereto a linear regression of the yield difference on the AZ-score. His analysis resulted in a significant slope ($p = 0.0215$) at the 5% level, so it could be concluded that there is a significant linear relation between the AZ-score and the difference in cultivar yields. He also did some diagnostics on the regression fit, indicating no deviation from the linearity assumption (residual plots, and a comparison with a model which also included a quadratic term for AZ), no evidence against normally distributed

residuals (Shapiro-Wilk: $p = 0.2548$), and no deviation from the variance-constancy assumption (residual plot).
Thus, at first sight, there seems to be no critique on his analysis. Figure 2.4 shows a scatter plot of the yield difference against the AZ-score. On the graph the linear regression line is indicated. Regression diagnostics revealed however that observations 4 and 5 had a Cook's Distance greater than 0.2 (14 out of the 19 observations had Cook's Distances smaller than 0.1). Also their DFITS values are the most extreme of all observations. Thus, both observation could be considered as influential outliers. (See e.g. Belsey, Kuh and Welsch (1980), Cook and Weisberg (1982) for a discussion on diagnostics for influential outliers.) Therefore, we did the regression again, but on a reduced data set with observations 4 and 5 removed. The deleted observations and the new regression line are also shown in Figure 2.4. Now the p-value associated with the test for zero-slope is $p = 0.2335$, which is clearly far larger than the original 0.0215. This analysis shows that the original conclusion is strongly dependent on the presence (and the validity) of these influential outliers. The linear regression analysis does not seem to be robust to this phenomenon. Moreover, the linear regression was also performed by means of a robust regression procedure based on M-estimation (Heiberger and Becker, 1992), which also indicated a reduction in slope estimate after the two observations were deleted.

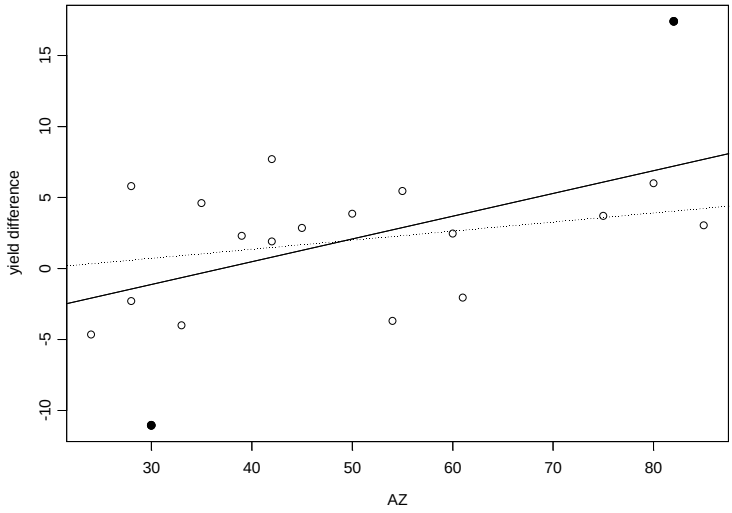


Figure 2.4: A scatter plot of the yield difference against the AZ-score. The full line and the dashed line represent the linear regression line of the complete and the outlier-deleted data set, respectively. The outliers are indicated by filled dots.

Apart from the robustness problems, there could have occurred other difficulties that made the analysis less straightforward. E.g. in many occasions the researcher does not known beforehand what type of dependence exists between the two variables that he is studying.

In the present example there does not seem to be an indication that the dependence is nonlinear, though the researcher did not have any a priori knowledge to support any such assumption. In this thesis the methods that are studied for testing independence are all omnibus tests. Omnibus methods have the characteristic that they are sensitive to almost all alternatives to the null hypothesis of independence. It may however be expected that, if in reality the dependence is linear, an omnibus test has lower power than a test that is constructed specifically for testing a linear dependence (a test of the latter kind is called a directional test). Thus, omnibus tests have the advantage that no a priori knowledge must be available, but on the other hand they have the disadvantage that they often have lower power as compared to a directional tests if the true dependence is the one for which the latter test is directional sensitive.

2.4 Gravity Data

The (American) National Bureau of Standards in Washington DC conducted between May 1934 and July 1935 eight series of experiments to determine the acceleration due to gravity (g). All experiments were conducted at the same place (Washington DC). The data are shown in Table 2.4, and will be referred to as the *Gravity Data*.

Table 2.4: The Gravity data for the 8 series of experiments. The data are given as deviation from $9.8ms^{-2}$ in units of $cms^{-2}.10^{-3}$.

series							
1	2	3	4	5	6	7	8
76	87	105	95	76	78	82	84
82	95	83	90	76	78	79	86
83	98	76	76	78	78	81	85
54	100	75	76	79	86	79	82
35	109	51	87	72	87	77	77
46	109	76	79	68	81	79	76
87	100	93	77	75	73	79	77
68	81	75	71	78	67	78	80
	75	62			75	79	83
	68				82	82	81
	67				83	76	78
						73	78
						64	78

The data have been analyzed before by e.g. Cressie (1982), Davison and Hinkley (1997), Rosenkrantz (2000).

The question that may be of interest here is whether the observations from the 8 series come from the same distribution. If this would not be the case then, of course, the National Bureau of Standards would not have found a standard, nor a standard method to measure

the gravity.

The question raised in the previous paragraph is the ***k*-sample problem**. Thus the null hypothesis is

$$H_0\text{: the observations from the } k \text{ samples have the same distribution .}$$

There are many alternatives to the null hypothesis possible. Probably the best known alternative is the location shift alternative. In this case one only wants to reject the null hypothesis if the means (or the medians) of the k distributions are not equal. This specific question will be referred to as the k-sample location shift problem. In other occasions one only wants to reject the null hypothesis if the variances are not equal (the k-sample dispersion problem). In this thesis, however, most of the time the most general k-sample problem will be considered, though, when the null hypothesis is rejected in favour of the omnibus alternative, it would be nice to have a method that indicates as well what the true situation may be (e.g. rejection is due to a location shift, or a change in dispersion).

The box plots for the Gravity Data are shown in Figure 2.5. Most apparent, these box plots show a difference in dispersion, though this conclusion based on a visual inspection must be interpreted with care since each of the 8 samples contains only between 8 and 13 observations. QQ-plots are presented in Figure 2.6. Although some of the QQ-plots may seem to indicate some deviation from normality, it is hard to formulate a conclusion based on the visual inspection, again because the samples are rather small. Kolmogorov-Smirnov tests for normality have been performed on each of the 8 series: only for the 7th series the test gave a significant result ($p = 0.0451$) (note the two outliers in Figure 2.5 for series 7).

Except for the one-sample GOF question for all 8 series, the specific k-sample question that will be of interest in this thesis involves only the first and the last series. Thus the problem is reduced to a 2-sample problem. A typical solution for testing the null hypothesis against the location shift alternative is an F-test in a classical ANOVA, or equivalently an unpaired t-test, for which however a normality and a variance-constancy assumption must hold. Although for both series 1 and 8 the Kolomogorov-Smirnov tests indicated no deviation from normality, it may still be possible that the residuals of the ANOVA model are more informative towards nonnormality (since the number of residuals is 8+13=21 is larger than the number of observations used for the KS tests (8 and 13), the power of the former test is expected to be larger). The normality assumption may be assessed by means of a QQ-plot. This is presented in Figure 2.7, which now clearly shows a systematic departure from normality. Moreover, Cook's distances (Cook and Weisberg, 1982) were calculated; they indicate that 3 observations are infuental outliers. Thus the p-value of the F-test ($p = 0.0175$) may be questionable. As an alternative approach, Wilcoxon's rank sum test is appied, resulting in $p = 0.1213$. Thus the evidence in favour of a location shift is not convincing. Further, Mood's test is computed for comparing the variances of the two series. With $p = 0.0171$ it is concluded that the two samples most probably are different in scale.

Finally, we would like to mention that Davison and Hinkley (1997) conducted a bootstrap

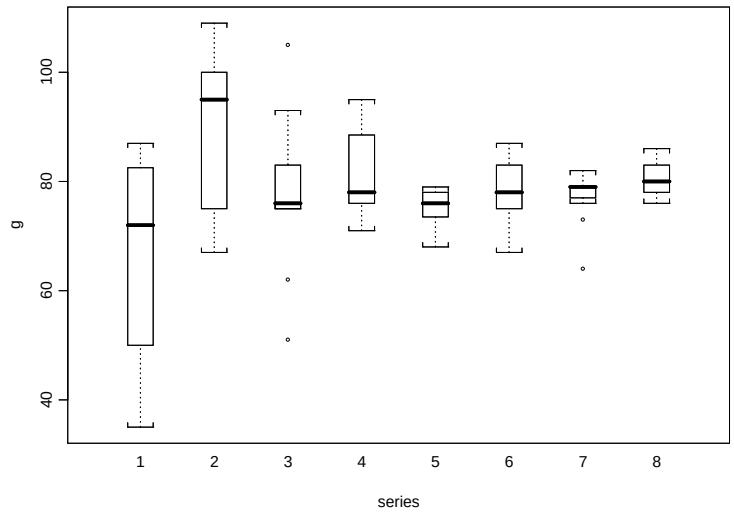


Figure 2.5: Box plots of the 8 series of gravity measurements as units in $cms^{-2}.10^{-3}$ from the reference value $9.8ms^{-2}$.

test to test the k-sample location shift hypothesis, taking the variance heterogeneity into account (i.e. a bootstrap test for the Behrens-Fisher problem). They found $p = 0.030$, pointing into the direction of a location shift.

2.5 Sleep Data

Allison and Cicchetti (1976) published the results from a study on the sleep behaviour of 62 mammals, in which they tried to relate them to some possible covariates. Here, we consider only three variables: the brain weight (measured in grams), the length of the gestation period (measured in days), and a categorical variable indexing the overall exposure to danger (5 categories). The data on the selected variables is reproduced in Tables 2.5 and 2.6. This example will be referred to as the *Sleep Data*.

A first question to be answered is whether there is a relation between the brain weight and the gestation, or not. This requires clearly a test for independence. Figure 2.8(a) shows a scatter plot of both variables. Inference based on a classical linear regression analysis (solid line) is not valid since the residual variance shows non-constancy. Moreover, there seems to be a systematic departure from normality, which may be to a large extent due to the two most extreme outliers (African and Asian Elephant). A robust linear regression (dashed line), still shows a positive dependence, but to a much lower extend, and still showing residual variance heterogeneity.

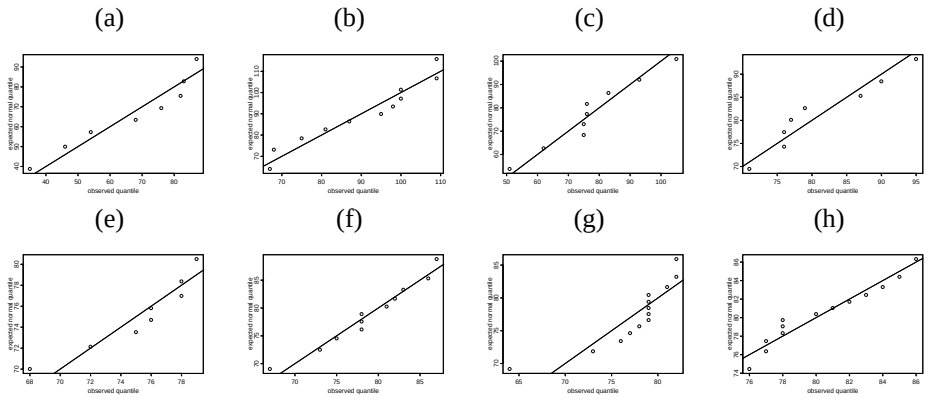


Figure 2.6: QQ-plots of the Gravity Data. Panels (a) up to (h) correspond to series 1 to 8, respectively.

The second question that needs to be answered is to assess any difference in brain weight between the 5 classes of danger exposure. Figure 2.8(b) shows the box plots. The Kolmogorov-Smirnov test reveals that only the brain weights of the 5th danger class may be considered to be normally distributed (all p-values ≤ 0.0001, except for class 5, $p \geq 0.5$). Levene's test for equality of variances resulted in $p = 0.007$, indicating a strong significant difference between the variances. Hence, the assumptions for an F-test for equality of means are not fulfilled. A Kruskal-Wallis test may be performed, but since the forms of the 5 distributions are definitely not equal, a rejection of the null hypothesis would not necessarily imply a difference in location, but rather the more general alternative of a difference in stochastic ordering. Kruskal-Wallis gives $p = 0.0027$. Thus, at least one of the distributions is stochastically larger or smaller than the others.

2.6 Ethanol Data

The data presented in this section, which will be referred to as the *ethanol data*, consists of 88 observations of the concentration of nitric oxides in engine exhaust and the equivalence ratio, which is a measure of the richness of the air-ethanol mix. The measurements are made on a single cylinder automobile test engine. The data are shown in Figure 2.9.

Originally one was interested in the relation between both variables; Simonoff (1996) used the data to illustrate smoothing methods. Here, we only want to test whether or not the two variables are dependent. Although one look at the scatterplot reveals immediately that most probably there is a relation (more or less quadratic), we included this example mainly to compare the test that will be constructed in this thesis with other tests of which it is generally known that they are insensitive to non-monotonic relations. E.g. Spearman's rank correlation test results in $p = 0.1873$ indicating no dependence. Kallenberg and Ledwina (1999) performed their V test on the same data, and calculated a p-value of 0,

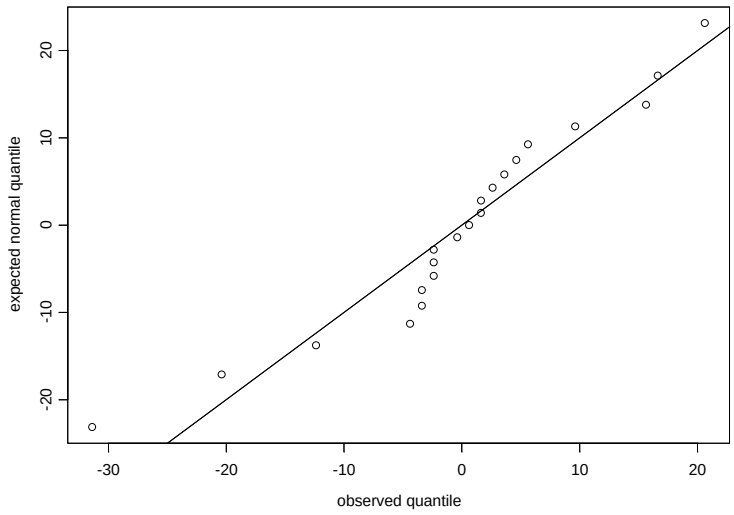


Figure 2.7: QQ-plot of the residuals of the ANOVA model of the Gravity Data.

concluding the expected, very strong dependence.

2.7 Some General Remarks

The data sets presented in the previous sections had at least one GOF problem that has to be solved. Although for most of these problems it may seem that we provided already a solution (e.g. F-tests, Kolmogorov-Smirnov, Kruskal-Wallis, t-test in least-squares linear regression, ...), we would like to stress here once more that in this thesis omnibus tests will be discussed. For such tests it will not be necessary to assess assumptions and to subsequently look for a method that works under the assessed assumptions (e.g. normality, linearity of a dependence). With the latter procedure not only the problem of multiplicity comes into play, but also the danger of ending up with a too strongly data-driven solution that might insufficiently guarantee the generalization of the conclusion from the observed pattern in the sample to the true structure in the population. Omnibus tests have the advantage of being more generally applicable (no prior knowledge needed).

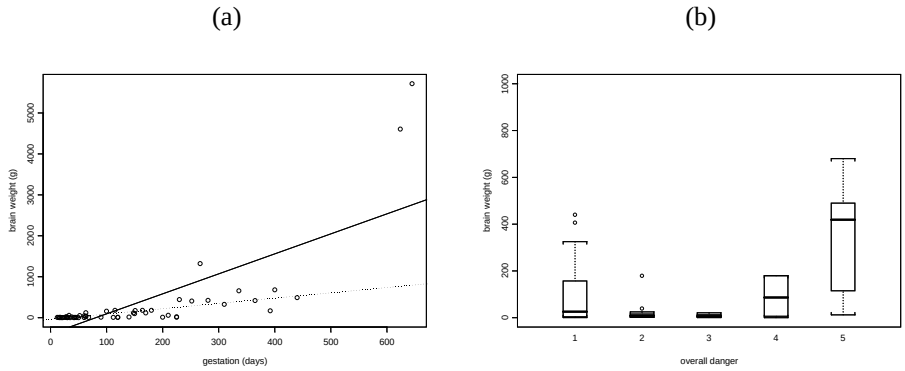


Figure 2.8: (a) A scatter plot of the brain weight against the gestation. The solid and the dashed line represent the least-squares linear regression and the M-estimation robust linear regression lines, respectively. (b) The box plots of the brain weight for each of the 5 danger exposure classes (the data for the African (danger=3) and Asian (danger=4) Elephant and for man (danger=1) are not shown).

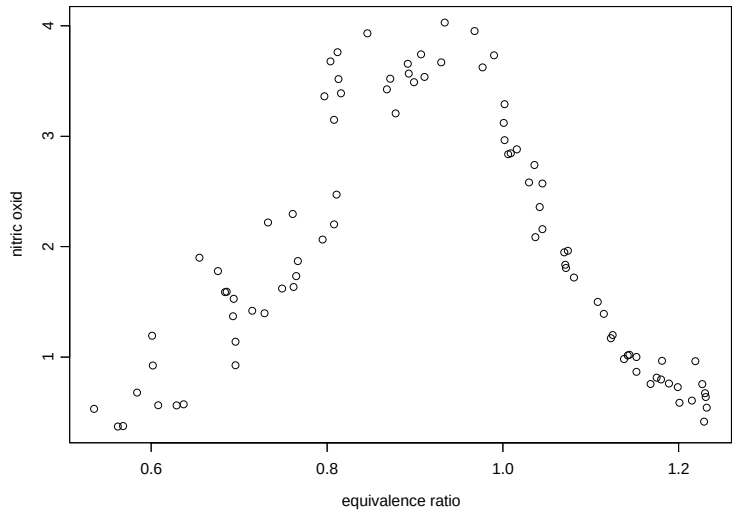


Figure 2.9: A scatterplot of the nitric oxide concentration and the equivalence ratio (ethanol data).

Table 2.5: The Sleep Data.

species	brain weight (g)	gestation (days)	danger index
African elephant	5712.00	645.00	3
African giant pouched rat	6.60	42.00	3
Arctic Fox	44.50	60.00	1
Arctic ground squirrel	5.70	25.00	3
Asian elephant	4603.00	624.00	4
Baboon	179.50	180.00	4
Big brown bat	.30	35.00	1
Brazilian tapir	169.00	392.00	4
Cat	25.60	63.00	1
Chimpanzee	440.00	230.00	1
Chinchilla	6.40	112.00	4
Cow	423.00	281.00	5
Desert hedgehog	2.40		2
Donkey	419.00	365.00	5
Eastern American mole	1.20	42.00	1
Echidna	25.00	28.00	2
European hedgehog	3.50	42.00	2
Galago	5.00	120.00	2
Genet	17.50		1
Giant armadillo	81.00		1
Giraffe	680.00	400.00	5
Goat	115.00	148.00	5
Golden hamster	1.00	16.00	2
Gorilla	406.00	252.00	1
Gray seal	325.00	310.00	1
Gray wolf	119.50	63.00	1
Ground squirrel	4.00	28.00	3
Guinea pig	5.50	68.00	4
Horse	655.00	336.00	5
Jaguar	157.00	100.00	1
Kangaroo	56.00	33.00	4
Lesser short-tailed shrew	.14	21.50	4
Little brown bat	.25	50.00	1
Man	1320.00	267.00	1
Mole rat	3.00	30.00	1
Mountain beaver	8.10	45.00	3
Mouse	.40	19.00	3
Musk shrew	.33	30.00	3
N. American opossum	6.30	12.00	1
Nine-banded armadillo	10.80	120.00	1

Table 2.6: The Sleep Data (continued).

species	brain weight (g)	gestation (days)	danger index
Okapi	490.00	440.00	5
Owl monkey	15.50	140.00	2
Patas monkey	115.00	170.00	4
Phanlanger	11.40	17.00	2
Pig	180.00	115.00	4
Rabbit	12.10	31.00	5
Raccoon	39.20	63.00	2
Rat	1.90	21.00	3
Red fox	50.40	52.00	1
Rhesus monkey	179.00	164.00	2
Rock hyrax (Hetero. b)	12.30	225.00	2
Rock hyrax (Procavia hab)	21.00	225.00	3
Roe deer	98.20	150.00	5
Sheep	175.00	151.00	5
Slow loris	12.50	90.00	2
Star nosed mole	1.00		2
Tenrec	2.60	60.00	2
Tree hyrax	12.30	200.00	3
Tree shrew	2.50	46.00	2
Vervet	58.00	210.00	4
Water opossum	3.90	14.00	1
Yellow-bellied marmot	17.00	38.00	1

CHAPTER 3

Three Goodness-of-Fit Problems: a Formal Introduction

In the introductory Chapters 1 and 2 an informal definition of the problem was given. Here, in Section 3.3 a formal definition of the GOF-problem will be given, but first, in Section 3.1, some basic terminology and notation is introduced and in Section 3.2 the meaning of a statistical model is explained. In the latter section also the distinction between parametrical and nonparametrical models is explained. Since some of the terms that are introduced in this section are basic terms of Probability Theory, we prefer not to give all the definitions in detail for these can be found in any introductory book on Probability Theory (e.g. Shorack, 2000).

3.1 Some Basic Definitions

Let $\boldsymbol{X} = (X_1, \ldots, X_p)'$ be a p-variate random variable (rv) defined over a sample space (SS) $\mathcal{S}_{\boldsymbol{x}} = \mathcal{S}$. A realization of $\boldsymbol{X}$ will be denoted by $\boldsymbol{x}$, which takes values in the sample space $\mathcal{S}$. $\boldsymbol{X}$ will be called continuous or discrete if all its components X_i ($i = 1, \ldots, p$) are continuous or discrete rv's, respectively. If some of its component are continuous and the other are discrete, then $\boldsymbol{X}$ is said to be a mixed continuous-discrete random variable. Throughout this thesis it is assumed that the multivariate distribution of $\boldsymbol{X}$ is proper in

the sense of Lauritzen and Wermuth (1989). If $\boldsymbol{X}$ is continuous, we will also assume that the cumulative distribution funtion (CDF), which is denoted by $F_{\boldsymbol{x}}(\boldsymbol{x})$, is differentiable. Hence, the density function (df) of $\boldsymbol{X}$ exists and is denoted by $f_{\boldsymbol{x}}(\boldsymbol{x})$. Moreover, we follow these authors in assuming that the density function of $\boldsymbol{X}$ is stricktly positive over the the sample space $\mathcal{S}$. These assumptions on countinuous rvs are summarized in Assumption A1.

Assumption (A1).
The CDF of $\boldsymbol{X}$ is differentiable, and the corresponding df is stictly positive over the sample space.

Sometimes the indices may be expanded, resulting in e.g. $f_{x_1 x_2 \ldots x_p}(.)$, or equivalently $f_{12\ldots p}(.)$. In case there is no confusion possible, the index may even be dropped. Marginal dfs and CDFs are notated similarly. Although for discrete rvs the term *density function* is often replaced by *probability function*, we will for the sake of generality use the former term in both occasions.

If the rv $\boldsymbol{X}$ is partitioned into $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, then the conditional df of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$ is denoted by $f_{\boldsymbol{x}_1|\boldsymbol{x}_2}(\boldsymbol{x}_1|\boldsymbol{x}_2)$, sometimes $f_{1|2}(.)$ for short. The notation for the conditional CDF is similar.

The sample of n observations $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ will be denoted by $\mathcal{S}_n$. This notation is used for both the sample of rvs $\boldsymbol{X}_i$ as for the realized sample, i.e. the sample of realized values $\boldsymbol{x}_i$ ($i = 1, \ldots, n$).

3.2 Statistical Models

In statistical literature the term *statistical model* can have several interpretations, depending on the context. In this thesis there is need for about the most general definition of a statistical model.

First, we will give a brief formal description of a probability space, from which implicitly the definition of a statistical model follows. Next, the distinction between parametric and nonparametric models is explained, and a brief discussion on parameter estimation in these models is given.

3.2.1 The Relation between a Statistical Model and a Probability Space

Formally, the CDF $F_{\boldsymbol{x}}(\boldsymbol{x})$ is actually uniquely related to a specifically constructed probability law P_F by choosing an appropriate σ-algebra $\mathcal{B}$ on $\mathcal{S}$. This probability law P_F, the σ-algebra $\mathcal{B}$ and the sample space $\mathcal{S}$ form together the probability space $(\mathcal{S}, \mathcal{B}, P_F)$. We will interpret the latter mathematical object in the same way as Lindsey (1996, p. 411): in an experiment, a point $\boldsymbol{x} \in \mathcal{S}$ is chosen at random according to the law P_F. This

process is called the underlying *data generating mechanism*. It is also considered as an abstract formulation of a statistical or probabilistic *model*. The chosen point determines the outcome of the experiment, E. For this $E \in \mathcal{B}$, $P_F(E)$ represents the probability of E. In this thesis, we will not use the probability space explicitely. Instead we will always work on the level of *distribution functions* and *density functions*. The aim of this paragraph is only to show the relation between both and their relation to statistical models. A *distribution* is briefly denoted by F when the CDF is $F_{\boldsymbol{x}}(.)$. Further, $\boldsymbol{X} \sim F$ is read as "the rv $\boldsymbol{X}$ is distributed according to F", i.e. the rv $\boldsymbol{X}$ obeys the probability law P_F within $(\mathcal{S}, \mathcal{B}, P_F)$, or, more loosely, $\boldsymbol{X}$ has CDF $F_{\boldsymbol{x}}(.)$.

Thus, a statistical model is basically the distribution that generates a sample that can be observed. In many contexts, both terms are therefore interchangeable.
A few examples are given to illustrate the meaning of a statistical model.

EXAMPLE 3.1. The number (rv X) of bacteria per unit of volume in the blood of an animal can typically be described by a Poisson distribution. The probability space $(\mathcal{S}, \mathcal{B}, P_F)$ is obtained by setting $\mathcal{S} = \mathbb{N}$ and $\mathcal{B} = \{\{0\}, \{1\}, \dots\}$. Then, each $x \in \mathcal{S}$ clearly determines the outcome $E = \{x\} \in \mathcal{B}$, and the probability law $P_F(E)$ is the Poisson probability function, i.e. $P_F(E) = \mathrm{P}[X = x] = \frac{e^{-\mu}\mu^x}{x!}$, where $\mu > 0$ is the constant mean. □

EXAMPLE 3.2. Well known statistical models include the traditional linear regression models. Consider the simple linear regression model

$$Y_i = \mu + x_i\beta + \epsilon_i, \tag{3.1}$$

($i = 1, \dots, n$) where μ, β and the explanatory variables x_i are given, and the ϵ_i are i.i.d. $N(0, \sigma^2)$ for some given σ. Based on some properties of the normal distribution, the model formulated in Equation 3.1 now becomes $Y_i \sim N(\mu + x_i\beta, \sigma^2)$ ($i = 1, \dots, n$). Thus, at every x_i at which an experiment is conducted resulting in Y_i, this rv behaves according the probability space $(\mathcal{S}, \mathcal{B}, P_{Fi})$, where now $\mathcal{S} = \mathbb{R}$, $\mathcal{B} = \{]-\infty, y] : y \in \mathcal{S}\}$. Every point $y \in \mathcal{S}$ determines uniquely the event $E =]-\infty, y] \in \mathcal{B}$, and the probability law $P_{Fi}(E) = \mathrm{P}[Y_i \in E] = F_{y_i}(y_i)$, where $F_{y_i}(.)$ is the CDF of $N(\mu + x_i\beta, \sigma^2)$.

Thus, strictly the above model is only defined at the given design points $x_1, \dots, x_n$, although one is often willing to adopt the assumption that the model remains true at intermittent values for x_i. A way to generalize the model is to consider the x_i as realizations of a rv X, then the discussion in the former paragraph remains valid, except that the probability law now becomes a function of x, i.e. $P_{Fx}(E) = \mathrm{P}[Y \in E | X = x] = F_{y|x}(y|x)$, where $F_{y|x}(.)$ is the conditional CDF of Y, given x, $N(\mu + x\beta, \sigma^2)$.

This example can serve as well to illustrate a connection between a statistical model and a deterministic model. Suppose that $\sigma = 0$ (actually the limit $\sigma \to 0$), then the normal distribution $N(0, \sigma^2)$ becomes degenerate, i.e. the distribution reduces to a point

probability at 0. The model in Equation 3.1 now reduces to $Y_i = \mu + x_i\beta$, in which no stochastic terms occur and which therefore can be called deterministic. □

3.2.2 Parametric versus Nonparametric Models

In the former section we defined a statistical model as a probability space $(\mathcal{S}, \mathcal{B}, P_F)$, which serves as the data generating mechanism. From a practical point of view the sample space $\mathcal{S}$, which is the set of all possible values that a realization of X can take, is actually determined by the experimental frame which is known to the researcher prior to the collection of the observations. Therefore, without loss of generality, we will suppose that the sample space is known, in both parametric and nonparametric models. Further, for all proper distributions that are considered in this thesis the σ-algebra can be chosen such that the corresponding probability law is a CDF, a probability function or a product of both when the rv is continuous, discrete or mixed continuous-discrete, respectively. (In order not to have to make the distinction between these three types of multivariate distributions, we will mostly use the term CDF in all occasions.) By this specific construction of the σ-algebra, the probability law is equivalent to the CDF.

What remains unknown at the start of the data collection or sampling is the probability law, or equivalently the CDF. Depending on the degree of knowledge about the CDF, three types of models are distinguished.

- **Fully specified parametric model.** Here, the CDF is completely specified, e.g. in Example 1 the model is fully specified when μ is set to a known constant. In Example 3.2.1 this is obtained when μ, β and σ are known constants.

- **Parametric family of models.** Sometimes only the functional form of the CDF is known. This function still depends on a finite number of parameters, say $\boldsymbol{\theta}$, that take values in a set Θ of permissible values, which is called the parameter space. Thus for all $\boldsymbol{\theta} \in \Theta$ the function defines indeed a proper CDF. The CDF and the corresponding distribution are now denoted as $F_x(x; \boldsymbol{\theta})$ and $F_{\boldsymbol{\theta}}$, respectively. The collection of $F_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta} \in \Theta$, or the corresponding CDFs, is called a parametric family of models, indexed by $\boldsymbol{\theta}$. This is also denoted as $\mathcal{F}_\Theta = \{F_\theta : \boldsymbol{\theta} \in \Theta\}$. Both the Examples 1 and 3.2.1 are parametric families, indexed by $\theta = \mu$ and $\boldsymbol{\theta} = (\mu, \beta, \sigma)$, respectively.

- **Nonparametric models.** When one is even not willing to assume any functional form for the CDF, but only that a proper CDF exists, the model is called nonparametric. The class of all proper CDF's will be denoted by $\mathcal{F}_\infty \supset \mathcal{F}_\Theta$.

  More formally, under certain regularity conditions, any CDF can be uniquely reconstructed from the set of all moments μ_i of whatever distribution. Hence, the moments can be considered as the parameters defining the model. For continuous rvs this means that in general an infinite number of parameters is needed. For discrete rvs a finite number of parameters may be sufficient. Thus, a nonparametric

model can be seen as a special case of a parametric family. Nevertheless, since its genesis is motivated by the absence of knowledge on the data generating mechanism in any sense, nonparametric models are conceptually completely different.

Yet another type of models are the *semi-parametric models*, which are are a special class of nonparametric models. Here, a part of the distribution is modelled as in a parametric family, and the remaining part is left unspecified as in a nonparametric model. Thus still an infinite number of nuisance parameters are involved.

EXAMPLE 3.3. Consider the model in Equation 3.1 of Example 3.2.1, but suppose now that the zero-mean distribution of ϵ_i is unknown. This model is called semiparametric. □

3.2.3 Parameter Estimation: a few Remarks

Of all the models defined above, it is only the fully specified parametric model that does not ask for parameter estimation. All other models have a parameter vector $\boldsymbol{\theta}$ that contains at least one parameter that must be estimated from the sample. The estimator of $\boldsymbol{\theta}$ will be denoted by $\hat{\boldsymbol{\theta}}$, or by $\hat{\boldsymbol{\theta}}_n$ when its dependence on the sample size n must be stressed.

In many parametric models the researcher is interested in only some of the parameters, say $\boldsymbol{\gamma}$, contained in $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\nu})$. The remaining parameters $\boldsymbol{\nu}$ are referred to as *nuisance* parameters. In many occasions estimates for $\boldsymbol{\gamma}$ can only be calculated simultaneously with those for the nuisance parameters $\boldsymbol{\nu}$, which makes the computations generally more heavy, and, more importantly, may reduce the efficiency of the $\boldsymbol{\gamma}$-estimators.

EXAMPLE 3.4. Consider the model in Example 1. The parametric family is the family of all Poisson distributions, indexed by the parameter μ, which is the mean of the Poisson distributed rv. If the researcher is interested in the mean concentration of bacteria in the blood of the animal, he must estimate μ based on a sample of observations. Since μ is the only parameter in the model, no nuisance parameter problems arise. □

EXAMPLE 3.5. Consider the model in 3.2.1. Suppose that the researcher needs to answer the question whether there is a linear regression relationship or not. Then he is only interested in the parameter β, but β can only be estimated simultaneously with μ, and for performing a t-test or for the calculation of a confidence interval for β he will also need an estimate for σ. Thus the parameter $\boldsymbol{\nu} = (\mu, \sigma)$ may then be called a nuisance parameter. □

For nonparametric models we restrict the discussion here to continuous rvs because this represents the worst case. Earlier it was mentioned that the parameter vector may consist of all moments of the distribution. Other infinite-dimensional parameterizations may be

considered. In general, however, most of these parameters do not have always a clear interpretation to the researcher. Moreover, from a practical point of view, he can only be interested in a finite number of parameters. Hence, an infinite number of nuisance parameters still remains. It is obvious that estimating an infinite number of parameters is problematic with only a finite sample at hand. The same discussion holds for semi-parametric models in which all the parameters in the nonparametric part are nuisance parameters.

3.3 Three Goodness-of-Fit Problems

In Chapter 2 some typical examples were given to illustrate three important GOF Problems. Basically, one wants to apply a GOF method when one has to assess whether a given sample of observations comes from a distribution with some specified characteristics. This thesis is mainly focussed on statistical tests for GOF.

3.3.1 The General Null Hypothesis

From the discussion in Section 3.2, every proper CDF can be denoted by $F_{\boldsymbol{\theta}} \in \mathcal{F}_\Theta \subseteq \mathcal{F}_\infty$, where the parameter vector $\boldsymbol{\theta} \in \Theta$ may possibly be infinite-dimensional. We will further assume that the parameterization is such that there is a one-to-one mapping from Θ to $\mathcal{F}_\Theta$. Then, the null hypothesis is stated as

$$H_0 : F_{\boldsymbol{x}} \in \mathcal{G}_0, \tag{3.2}$$

where $F_{\boldsymbol{x}}$ is the true, but unknown CDF of rv $\boldsymbol{X}$, and $\mathcal{G}_0 \subseteq \mathcal{F}_\Theta$ is a set of specified CDFs. The set $\mathcal{G}_1 = \mathcal{F}_\Theta \setminus \mathcal{G}_0$ specifies the distributions under the alternative hypothesis. Since there is a one-to-one mapping from Θ to $\mathcal{F}_\Theta$, the decomposition $\mathcal{F}_\Theta = \mathcal{G}_0 \cup \mathcal{G}_1$ implies a corresponding decomposition of Θ into the mutually exclusive subsets Θ_0 and Θ_1, respectively. Hence the null hypothesis formulated in Equation 3.2 is equivalent to

$$H_0 : \boldsymbol{\theta} \in \Theta_0, \tag{3.3}$$

where $\boldsymbol{\theta}$ is the parameter vector of the true distribution F. The symbol $F = F_{\boldsymbol{x}}$ will be used to denote the *true*, but unknown CDF of the rv $\boldsymbol{X}$, i.e. the observations at hand are a sample from this distribution, and for the elements of $\mathcal{G}_0$, $G = G_{\boldsymbol{x}}$ will be used. Of course, if F is known, there is no need anymore for a GOF method. Hence, we will suppose that in practice the distribution F is not completely known to the researcher. Furthermore, as a working hypothesis we will suppose that the sample spaces associated with both distributions coincide.

Assumption (A2).
The distributions F and G are defined on the same sample space $\mathcal{S}$.

3.3.2 The Alternative Hypothesis

In the most general setting, when the null hypothesis is not true, one does not know in what sense the true distribution F deviates from the specified distributions in $\mathcal{G}_0$. Therefore, the alternative hypothesis is just the complement of the null hypothesis, $\mathcal{G}_1 = \mathcal{F}_\Theta \setminus \mathcal{G}_0$. GOF tests, that are constructed for this situation, and that are thus sensitive to all types of deviation from the null hypothesis, are called *omnibus tests*. The price that has to be payed for this overall sensitivity is a loss in power for some specific alternatives as compared to GOF tests that are especially designed to detect deviations from H_0 in the direction of these specific alternatives.

3.3.3 The One-sample GOF Problem: Simple and Composite Null Hypotheses

The null hypothesis as given in Equation 3.2 or 3.3 is general in the sense that the set $\mathcal{G}_0$ still can be constructed in several ways. When the observations on $\boldsymbol{X}$ represent 1 sample which is assumed to be generated by one distribution, and when $\mathcal{G}_0$ contains just one distribution, or at most a parametric family of distributions, indexed by a finite dimensional parameter, then the null hypothesis in Equation 3.2 is said to be the one-sample problem null hypothesis.
Depending on the form of $\mathcal{G}_0$ two types of null hypothesis are distinguished.

- **Simple Null Hypothesis.** If $\mathcal{G}_0 = \{G_{\boldsymbol{x}}\}$, i.e. $\mathcal{G}_0$ contains only one element, then the corresponding null hypothesis is called *simple*. This CDF can also be completely specified by one specific, possibly infinite dimensional, parameter vector $\boldsymbol{\theta}_0$. Thus, $\Theta = \{\boldsymbol{\theta}_0\}$.

  The simple null hypothesis is often formulated as

  $$H_0 : F_{\boldsymbol{x}}(\boldsymbol{x}) = G_{\boldsymbol{x}}(\boldsymbol{x}), \tag{3.4}$$

  for all $\boldsymbol{x} \in \mathcal{S}$.

- **Composite Null Hypothesis.** If $\mathcal{G}_0$, or equivalently Θ_0, contains more than one element, we will suppose that $\mathcal{G}_0 = \mathcal{G}_{\Theta_0}$ represents a parametric family, indexed by at least a subvector of $\boldsymbol{\theta} \in \Theta_0$. In many cases, when there is nothing assumed on the form of the true distribution F, Θ will be infinite dimensional and the parameterization of F will be hard to interpret by the researcher. The specified parametric family $\mathcal{G}_{\Theta_0}$, on the other hand ,will often have another parameterization which is finite dimensional and has a clearer interpretation. In this case Θ_0 is not a subset of Θ, but we will suppose that Θ_0 is a subset of a parameter space, say Θ', which stands in a one-to-one relation to Θ. Then, $\Theta_1 = \Theta' \setminus \Theta_0$.

EXAMPLE 3.6. In the classical null hypothesis for GOF-tests for normality,

there is nothing assumed about the true distribution F, and $\mathcal{G}_0 = \mathcal{G}_{\Theta_0}$ is the set of all possible normal distributions. Then, $\mathcal{F}_\Theta = \mathcal{F}_\infty$, which has a moments parameterization indexed by $\boldsymbol{\theta} = (\theta_1, \dots)$ where θ_i is the ith moment; Θ is the set of all such θ. The null parameter space, on the other hand, is defined as $\Theta_0 = \{(\mu, \sigma, \theta_3, \theta_4, \dots) : \theta_3 = 0, \theta_4 = 3\sigma^4, \dots\}$, where μ and σ are the mean and the standard deviation, respectively. The relation between both parameterizations is given by $\theta_1 = \mu$, $\theta_2 = \sigma^2 + \mu^2$, and $\theta_3, \dots$ are uniquely determined by μ and σ. Thus Θ' is the set of all $\theta' = (\mu, \sigma, \theta_3, \dots)$. In the present context, μ and σ are nuisance parameters. □

For notational comfort we will not always make the distinction between Θ and Θ'.

3.3.4 The k-sample GOF Problem

A classical problem in statistics is testing the equality of k distributions. Consider an experiment in which data is independently collected from k different populations, then, of course, the experiment results in k independent samples. A frequently occurring question is then: " do the observations from these k samples have the same distribution ? ". It is referred to as the k-sample GOF problem.

There is a need here to make a distinction between the k samples and distributions.

Consider k samples $\mathcal{S}_{n_i} = \{\boldsymbol{X}_{i1}, \dots, \boldsymbol{X}_{in_i}\}$ of n_i p-variate observations $\boldsymbol{X}_{ij}$ ($j = 1, \dots, n_i$; $i = 1 \dots, k$). The corresponding sample spaces are denoted by $\mathcal{S}_i$ ($i = 1, \dots, k$). The true CDF of the observations from the ith sample is denoted by $F_{x,i} = F_i$. As before, the distributions may be embedded in a parametric family, indexed by a parameter $\boldsymbol{\theta}$. An additional index i may then be used to indicate the correspondence to distribution i.

Consider k independent samples $\mathcal{S}_{n_i}$ with corresponding CDFs F_i ($i = 1, \dots, k$). The null hypothesis is

$$H_0 : F_1(\boldsymbol{x}) = F_2(\boldsymbol{x}) = \ldots = F_k(\boldsymbol{x}) = G(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{S} \text{ and } G \in \mathcal{F}, \tag{3.5}$$

where G denotes the common CDF, which, in contrast to the GOF hypothesis discussed previously, must not be specified. Further, $\mathcal{S}$ is the common sample space under the null hypothesis. The expression $G \in \mathcal{F}$ stresses that the distribution belongs to some specified class of distributions, which possibly is the set $\mathcal{F}_\infty$ of all proper CDFs.

To see how this hypothesis fits into the general GOF framework, a working variable Y is introduced. Y is defined as a discrete variable that takes values in $\{1, \dots, k\}$ with probabilities $\mathrm{P}[Y = i] = \frac{n_i}{n}$, indicating the corresponding sample. The variable $\boldsymbol{X}$ is augmented with Y, resulting in $\boldsymbol{X}_a = (\boldsymbol{X}, Y)$ in such a way that for all $i = 1, \dots, k$, $\boldsymbol{X}_{a,ij} = (\boldsymbol{X}_{ij}, i)$. Let $\mathcal{S}_{xy}$ denote the sample space of the augmented rv; the joint CDF of $\boldsymbol{X}_a$ is denoted by F_{xy}. Then, the distribution of the augmented variable of the ith sample is simply $F_i = F_{x|i}(\boldsymbol{x}|y = i)$, i.e. the conditional distribution of $\boldsymbol{X}$, given $y = i$. The

k-sample null hypothesis in Equation 3.5 can now be expressed as

$$H_0 : F_{xy} \in \mathcal{G}_0 = \{F_{xy}(\boldsymbol{x}, y) = F_x(\boldsymbol{x})F_y(y), \text{ for all } (\boldsymbol{x}, y) \in \mathcal{S}_{xy}\},$$

where generally the common distribution $F_x = G$ is not specified. Hence $\mathcal{G}_0$ typically contains an infinite number of distributions, reflecting the nonparametric nature of the problem.

It may already be clear from the generality of the null hypothesis that there may be many types of alternative hypotheses.

The most general alternative hypothesis is just the negation of H_0, i.e.

$$H_1 : \exists i \neq j, \boldsymbol{x} : F_i(\boldsymbol{x}) \neq F_j(\boldsymbol{x}), F_i, F_j \in \mathcal{F}.$$

It is important to note that it is still required that all $F_i \in \mathcal{F}$ ($i = 1, \ldots, k$). The statistical problem related to this hypothesis will be called the *general* k-sample GOF problem, or simply the k-sample GOF problem. Tests that are constructed for testing H_0 against this general alternative hypothesis, are called *omnibus* tests.

Probably the most discussed k-sample problem is the one related to the translation-type hypothesis which represents a shift in location:

$$H_1 : F_i(\boldsymbol{x}) = G(\boldsymbol{x} + \boldsymbol{\delta}_i) \text{ for all } i = 1, \ldots, k \text{ and not all } \boldsymbol{\delta}_i \text{ equal, and } G \in \mathcal{F},$$

where $\mathcal{F}$ is the same set as specified in H_0, and where the $\boldsymbol{\delta}_i \in \mathbb{R}^p$ quantify the shift in location. If one restricts $\mathcal{F} = \{F : F(\boldsymbol{x}) = F_0(\boldsymbol{x} + \boldsymbol{\delta}), \boldsymbol{\delta} \in \mathbb{R}^p\}$, where F_0 is some specified CDF, then the null hypothesis is reduced to $H_0 : \boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \ldots = \boldsymbol{\delta}_k = \mathbf{0}$, and the alternative hypothesis is that at least two $\boldsymbol{\delta}$'s differ.
This setting will be referred to as the k-sample location problem.

Another important alternative hypothesis is given by

$$\begin{aligned} H_1 \quad : \quad & F_i(\boldsymbol{x}) = G\left((\boldsymbol{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\right) \text{ for all } i = 1, \ldots, k \\ & \text{and not all } \boldsymbol{\Sigma}_i \text{ equal, and } G \in \mathcal{F}, \end{aligned}$$

where $\boldsymbol{\mu}_i = \mathrm{E}_f[\boldsymbol{X}]$ and where the $\boldsymbol{\Sigma}_i$ are $p \times p$ positive definite matrices. Especially in the univariate case, where $\boldsymbol{\Sigma}_i = \sigma_i \in \mathbb{R}$, this alternative hypothesis is well known. It is called the k-sample scale or dispersion problem.

3.3.5 The Independence Problem

The independence problem is only defined in a multivariate setting. Let the p-variate rv under study be partitioned into two components $\boldsymbol{X}$ and $\boldsymbol{Y}$, which contain p_1 and p_2 univariate rvs, respectively. The sample space is denoted by $\mathcal{S}_{xy}$. Then, the null

hypothesis of independence is

$$H_0 : F_{xy} \in \mathcal{G}_0 = \{F_{xy}(\boldsymbol{x}, \boldsymbol{y}) = F_x(\boldsymbol{x})F_y(\boldsymbol{y}), \text{ for all } (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S}_{xy}\},$$

where F_x and F_y are generally not further specified. Thus, it is clear that this problem asks again for a nonparametric solution.

The negation of the null hypothesis expresses that there is some type of dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$. It is for this most general setting that the methods in this thesis will be constructed.

Of course there are many specific types of dependence that are in practice of particular interest. Joe (1997) gives a systematic treatment of types of dependence. Most of them are best understood in the bivariate case ($p = 2$). We mention two of them.

- **Positive quadrant dependence**
  (X, Y) is positive quadrant dependent if

  $$\mathrm{P}\left[X > x, Y > y\right] \geq \mathrm{P}\left[X > x\right] \mathrm{P}\left[Y > y\right] \text{ for all } x, y \in \mathcal{S}_{xy}.$$

- **Stochastic increasing positive dependence**
  Y is stochastically increasing in X if for all $y \in \mathcal{S}_y$ $\mathrm{P}\left[Y > y | X = x\right]$ increases as x increases.

CHAPTER 4

A Selective Overview of Some Tests for Goodness-of-Fit

This chapter is mainly meant as a literature-overview of some of the most popular GOF tests. It is however not the intention to be complete. Not only because a full listing of all published GOF tests with their most important characteristics would fill at least a book, but also because not all GOF-tests are relevant to the research presented in this thesis. The selection of the GOF tests presented here is mainly based on 4 criteria.

- Tests that have a practical value.
- Tests that are widely used.
- Tests that have good power characteristics.
- Tests that are in some sense related to the test that is proposed in this thesis.

In this thesis, a central role is given to Pearson's χ^2-test, which can be used both for discrete and continuous variables. These methods are discussed in Section 4.1. In particular, Pearson's test for the one-sample problem are discussed. At the end of the section also Pearson's test for independence in contingency tables is briefly given. In Section 4.2 an overview is given of techniques that are based on the empirical distribution function, and in Section 4.3 smooth tests are introduced. A comprehensive overview of many GOF tests is given in D'Agostino and Stephens (1986).

4.1 Pearson's χ^2-test

Probably the best known and oldest GOF test is Pearson's χ^2-test (Pearson, 1900). Originally, this test was constructed for discrete models resulting from multinomial, Poisson or product-multinomial sampling. For a simple null hypothesis Pearson's test is described in Section 4.1.1. The correct theory for a composite null hypothesis was first provided by Fisher (1924). Therefore, this test is often referred to as the Pearson-Fisher test, which is treated in Section 4.1.3.

Although for models for continuous data many GOF-tests are developed, there has always been a great interest in adapting the Pearson χ^2-test to the continuous case, even though many of the GOF-test that were specifically constructed for continuous variables have in general much higher powers. Pearson's test can only be applied to continuous data after the data has been *grouped* or *categorized*, which intuitively already suggests that information will be lost. Since the new methods that are developed and studied in this thesis are strongly related to a Pearson test based on an alternative way of grouping or categorizing the data, these modified Pearson's tests for grouped data will be discussed in detail in Section 4.1.4.

Since most of the test that are discussed in this section can be considered as special cases of a *general quadratic form* (GQF) (Moore, 1977), we will give some important results for the GQFs after the introduction to the original Pearson statistic (Section 4.1.1). In this way, the χ^2-type tests can be discussed very concisely. The tests will also be placed in their correct historical context.

4.1.1 Pearson's Original Formulation for Simple Null Hypotheses

In the easiest case, Pearson's statistic is constructed for a simple null hypothesis in the family of multinomial distributions $M(\pi_1, \ldots, \pi_k; n)$ for which n is known, and of course $\sum_{i=1}^k \pi_i = 1$. The family $\mathcal{F}_\Theta$ is thus indexed by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$. The simple null hypothesis is determined by $\boldsymbol{\Theta}_0 = \{\boldsymbol{\pi}_0\} = \{(\pi_{01}, \ldots, \pi_{0k})\}$, for which also $\sum_{i=1}^k \pi_{0i} = 1$ holds. Thus,

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0.$$

Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ denote an observation from the multinomial distribution with $\sum_{i=1}^k X_i = n$. Then, Pearson's test statistic is given by

$$\begin{aligned} X^2 &= \sum_{i=1}^k \frac{(X_i - n\pi_{0i})^2}{n\pi_{0i}} \\ &= n \sum_{i=1}^k \frac{(X_i/n - \pi_{0i})^2}{\pi_{0i}}. \end{aligned} \tag{4.1}$$

Since under H_0 the expectation of $\boldsymbol{X}$ is $n\boldsymbol{\pi}_0$, the test statistic is often generically formulated as

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{4.2}$$

where O_i and E_i refer to the observed and expected frequencies, respectively. In his 1900 paper, Pearson proved that under H_0 X^2 has a limiting χ^2_{k-1}-distribution, provided that in each of the k cells the expectations $\mathrm{E}[X_i]$ become infinitely large as $n \to \infty$. The latter condition makes the Pearson χ^2-test an asymptotic test. Later in this section, some remarks on exact GOF-tests and their relation to Pearson's χ^2-test will be given. We will not give here a complete rigourous proof, only a sketch that may be useful for further purposes. The sketch is in the line of Read and Cressie (1988, p.161).

Let $\boldsymbol{P} = \boldsymbol{X}/n$ represent the vector of k probabilities, and let $\boldsymbol{D}_0 = \mathrm{diag}(\pi_{01}, \ldots, \pi_{0k})$. Then the Pearson test statistic in Equation 4.1 can be written as

$$\begin{aligned} X^2_P &= \left(\sqrt{n}\,(\boldsymbol{P} - \boldsymbol{\pi}_0)'\right) \boldsymbol{D}_0^{-1} \left(\sqrt{n}\,(\boldsymbol{P} - \boldsymbol{\pi}_0)\right) \\ &= \boldsymbol{W}_n' \boldsymbol{D}_0^{-1} \boldsymbol{W}_n, \end{aligned} \tag{4.3}$$

where $\boldsymbol{W}_n = \sqrt{n}\,(\boldsymbol{P} - \boldsymbol{\pi}_0)$. In the proof it is first established that if for all $i = 1, \ldots, k$ $\mathrm{E}[X_i] \to \infty$ as $n \to \infty$, under H_0 $\boldsymbol{W}_n$ converges weakly to a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{D}_0 - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'$. Next, based on results from Rao (1973) and Bishop, Fienberg and Holland (1975) it is proven that X^2_P in Equation 4.3 is a quadratic form of a multivariate normal variate which in the present situation converges in distribution to a χ^2-distribution with $k - 1$ degrees of freedom.

Up to now Pearson's results are obtained for a multinomial observation. Under certain restrictions, the results apply as well for Poisson and for product-multinomial sampling (Bishop et al., 1975).

Pearson's test can also be obtained in some different ways, e.g. as a score test within the likelihood framework, or as a generalized Wald test. Since within the latter framework many other historically interesting χ^2-type tests can be easily constructed, it will be presented in the next section.

4.1.2 General Quadratic Forms

General Quadratic Forms were first introduced by Moore (1977). Suppose that the k probabilities in $\boldsymbol{\pi}$ are the true probabilities of a multinomial, which may depend on a p-vector of unknown parameters $\boldsymbol{\beta}$, i.e. $\boldsymbol{\pi}(\boldsymbol{\beta})$ and $\boldsymbol{W}_n(\boldsymbol{\beta}) = \sqrt{n}(\boldsymbol{P} - \boldsymbol{\pi}(\boldsymbol{\beta}))$. Let Y be a random variable taking values in $\mathcal{S}_Y$ ($\#\mathcal{S}_Y \geq k$) such that there exists a mapping of the probability space of Y on the probability space of the multinomial $M(\boldsymbol{\pi}; 1)$. This may be denoted by the k-vector valued function $\psi(Y) = \boldsymbol{X}$, where $\boldsymbol{X} \sim M(\boldsymbol{\pi}; 1)$. Then, there also exists a mapping Ψ of a sample $\mathcal{S}_{Y;n}$ of n i.i.d. random variables Y to the

multinomial $M(\boldsymbol{\pi}; n)$.

Further, suppose that an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is of the form

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(\boldsymbol{Y}_i) + \boldsymbol{r}_n, \tag{4.4}$$

where $\boldsymbol{r}_n \xrightarrow{p} \mathbf{0}$ and $h(\boldsymbol{Y})$ is a $\mathbb{R}^p$- valued function such that $\mathrm{E}\left[h(\boldsymbol{Y})|\boldsymbol{\beta}_0\right] = \mathbf{0}$ and $\mathrm{E}\left[h(\boldsymbol{Y})h(\boldsymbol{Y})'|\boldsymbol{\beta}_0\right] = \boldsymbol{L}$, where $\boldsymbol{L}$ is a finite $p \times p$ matrix. Then, the GQF is defined as

$$M_n^2 = \boldsymbol{W}_n(\hat{\boldsymbol{\beta}})' \boldsymbol{Q}_n \boldsymbol{W}_n(\hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{Q}_n$ denotes a possibly data-dependent $k \times k$ symmetric nonnegative definite matrix. First, based on a result of Moore and Spruill (1975) he shows that under mild conditions

$$\boldsymbol{W}_n(\hat{\boldsymbol{\beta}}) \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ may be of rank $s \leq k$. Thus the multivariate normal distribution may be singular.

Next, Moore gives a general form for the variance-covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{A}$ denote the $k \times p$ matrix with (i, j)th entry

$$\frac{\partial \pi_i}{\partial \beta_j}.$$

Let $\boldsymbol{\iota}(Y)$ be the k-vector with ith entry $\mathrm{I}\left[\psi(Y)_i = 1\right]$, and $\boldsymbol{W}(Y) = \boldsymbol{\iota}(Y) - \boldsymbol{\pi}$. The $\boldsymbol{\Sigma}$ can be decomposed as

$$\boldsymbol{\Sigma} = \boldsymbol{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}' - \boldsymbol{C},$$

where

$$\boldsymbol{C} = -\boldsymbol{A}\boldsymbol{L}\boldsymbol{A}' + \boldsymbol{A}\mathrm{E}\left[\boldsymbol{h}(\boldsymbol{Y})\boldsymbol{W}(Y)'\right] + \mathrm{E}\left[\boldsymbol{W}(Y)\boldsymbol{h}(\boldsymbol{Y})'\right]\boldsymbol{A}'.$$

Moore's main result can be summarized as follows. Suppose all aforementioned assumptions are true. If $\mathrm{rank}(\boldsymbol{\Sigma}) = s \leq k$ and $\boldsymbol{Q}_n = \boldsymbol{\Sigma}_n^-$ is a consistent estimator of the generalized inverse $\boldsymbol{\Sigma}^-$ of $\boldsymbol{\Sigma}$, then

$$M_n^2 = \boldsymbol{W}_n(\hat{\boldsymbol{\beta}})' \boldsymbol{Q}_n \boldsymbol{W}_n(\hat{\boldsymbol{\beta}}) \xrightarrow{d} \chi_s^2.$$

4.1.3 Pearson-Fisher Test for Composite Null Hypotheses

Often, the probabilities that are specified under the null hypothesis are still depending on an unknown parameter vector $\boldsymbol{\beta}$ that takes values in some p-dimensional set $\mathcal{R}$. These nuisance parameters need then to be estimated from the sample. The set $\mathcal{G}_0$ now repre-

sents a family of distributions, indexed by $\boldsymbol{\beta}$. The resulting composite null hypothesis is

$$H_0 : \boldsymbol{\pi} \in \Pi_0,$$

where $\Pi_0 = \{\pi_0(\boldsymbol{\beta}) = (\pi_{01}(\boldsymbol{\beta}), \ldots, \pi_{0k}(\boldsymbol{\beta})) : \sum_{i=1}^{k} \pi_{0i}(\boldsymbol{\beta}) = 1, \boldsymbol{\beta} \in \Theta_0 \subseteq \mathbb{R}\}$. $\boldsymbol{\beta}_0$ denotes the index of the true, but unknown member of Π_0 under H_0.

An obvious solution to handle the composite null hypothesis is to calculate the traditional Pearson test statistic with $\boldsymbol{\pi}_0$ replaced by $\hat{\boldsymbol{\pi}}_0 = \boldsymbol{\pi}_0(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is a consistent *restricted* estimator for $\boldsymbol{\beta}_0$. By "restricted" it is meant that $\hat{\boldsymbol{\beta}} \in \Theta_0$. Sometimes the notation $\hat{\boldsymbol{\beta}}_0$ is used to indicate a restricted estimator, but for notational comfort we will not use this notation unless it is needed. From the very beginning Pearson argued that this was the right solution and that the resulting test statistic has the same χ^2_{k-1} limiting null distribution as the one for the simple null hypothesis. In the early 1920's Fisher (1922, 1924), however, proved that the number of estimated parameters should be substracted from $k-1$ to derive the correct number of degrees of freedom. Thus χ^2_{k-p-1} is the correct asymptotic null distribution. (This famous controversy between Karl Pearson and Ronald Fisher is lively told by Box (1978).) Cochran (1952) provides an account on the early developments of the χ^2-test. Because of the important contribution of Fisher, this test is often referred to as the Pearson-Fisher test.

Formally, the above mentioned result for the Pearson-Fisher test is based on some conditions on the estimator $\hat{\boldsymbol{\beta}}$, on the vector function $\boldsymbol{\pi}_0(\boldsymbol{\beta})$ and on Θ_0. More specifically, the estimator must be *best asymptotically normal* (BAN), i.e. consistent, asymptotically normally distributed and asymptotically efficient. The vector function and Θ_0 must satisfy the Birch (1964) regularity conditions, which are within the present context clearly stated in Bishop et al. (1975) and Read and Cressie (1988, p.164). Basically the Birch regularity conditions assure the following asymptotic expansion

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}(\boldsymbol{P} - \boldsymbol{\pi}_0(\boldsymbol{\beta})) + o_p(n^{-1/2}), \tag{4.5}$$

where $\boldsymbol{A}$ is as before. The conditions also imply that the model really has p parameters.

Within the present context it is most easy to arrive at the appropriate test statistic by applying Moore's results for GQFs. First, suppose that $\boldsymbol{\beta}$ is estimated by its restricted maximum likelihood estimator $\hat{\boldsymbol{\beta}}$, which satisfies the expansion in Equation 4.5. Further, the multinomial $\boldsymbol{X}$ can be a result of the Ψ transformation with Y taking values in $\{1, \ldots, k\}$ of which the elements refer to multinomial classes such that $\sum_{i=1}^{n} \mathrm{I}[Y_i = j] = X_j$ $(j = 1, \ldots, k)$.

From Equation 4.5 it is seen that $\boldsymbol{h}(Y) = (\boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}$, and consequently $\boldsymbol{L} = (\boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}\boldsymbol{A})^{-1}$ and $\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'$. Also $\mathrm{rank}(\boldsymbol{\Sigma}) = \mathrm{rank}(\boldsymbol{D}_{\pi_0} - \boldsymbol{\pi}_0\boldsymbol{\pi}_0' - \boldsymbol{C}) = k - 1 - p$, and $\boldsymbol{\Sigma}^- = \boldsymbol{D}_{\pi_0}^{-1}$, which is consistently estimated by $\boldsymbol{D}_{\pi_0}(\hat{\boldsymbol{\beta}})^{-1}$. The GQF

now becomes

$$X^2_{PF} = \boldsymbol{W}_n(\hat{\boldsymbol{\beta}})' \boldsymbol{D}_{\pi_0}(\hat{\boldsymbol{\beta}})^{-1} \boldsymbol{W}_n(\hat{\boldsymbol{\beta}}),$$

and under H_0, Moore's result also gives $X^2_{PF} \xrightarrow{d} \chi^2_{k-p-1}$. This is exactly the Pearson-Fisher test.

4.1.4 Pearson's Test for Grouped Continuous Data

Despite the availability of many GOF-tests especially developed for the analysis of continuous data, both in the literature and in most of the statistical software, today many researchers still apply Pearson's test on grouped continuous data. One reason for this may be the intuitively appealing nature of Pearson's test and the fact that it is very easy to calculate.

Grouping Data

Pearson-type tests can only be applied to continuous data after having grouped or categorized the original continuous data in k groups. The grouping is accomplished by considering a partition $[A] = \{A_1, \ldots, A_k\}$ of the sample space $\mathcal{S}$. Each of the k elements of the partition is called a *cell*. The cells are determined by the position of the *cell boundaries* in $\mathcal{S}$. Each partition implies a multinomial observation $\boldsymbol{X} = (X_1, \ldots, X_k)$, where $X_i = \#\mathcal{S}_n \cap A_i$ $(i = 1, \ldots, k)$, with probabilities under a null hypothesis equal to

$$\pi_{0i}(\boldsymbol{\beta}) = \int_{A_i} dG_{\beta}(\boldsymbol{x}), \tag{4.6}$$

$(i = 1, \ldots, k)$ where $\boldsymbol{\beta} \in \Theta_0$. Cells that are constructed in this way are called *fixed* cells. Suppose for the moment that $\Theta_0 = \{\beta_0\}$ (simple null hypothesis).

Three questions are raised with this approach.

1. How many groups must be constructed?
2. Where to place the cell boundaries?
3. Is the Pearson χ^2-test still valid under these circumstances?

These are indeed important questions. Kempthorne (1967), e.g. showed that with different cell constructions different conclusions may be obtained. Fisher (1925) was probably one of the first to give a theoretically sound recommendation: the expected number of observations in each cell should be at least 5 under the simple null hypothesis. He argued that under these circumstances the null distribution of the test statistic would be

sufficiently well approximated by the asymptotic χ^2_{k-1} distribution. Based on another criterion Mann and Wald (1942) recommended that the cells must be chosen such that all cells have equal probability under the null hypothesis, i.e. for all $i, j = 1, \ldots, k$ $\int_{A_i} dG_{\boldsymbol{\beta}}(\boldsymbol{x}) = \int_{A_j} dG_{\boldsymbol{\beta}}(\boldsymbol{x})$. Cells that are constructed in this way are called *equiprobable cells*. Under these conditions they showed that the Pearson test is unbiased. Later, Cohen and Sackrowitz (1975) and Bednarski and Ledwina (1978) showed that in most cases Pearson's test is biased when applied to an unequiprobable grouping. Mann and Wald also give a formula to determine an appropriate number of equiprobable cells based on the requirement that a power of at least $1/2$ should be guaranteed for all alternatives no closer than some Δ to the equiprobable null hypothesis. Indeed, the intuitively appealing reasoning that the more cells are constructed the more information from the original sample of continuous data is retained and the higher the power will be, is however not always correct (Oosterhoff, 1985) because the increase of the number of cells implies both an increase in the non-centrality parameter of the non-central χ^2-distribution of the test statistic under an alternative hypothesis, and an increase of the variance of the limiting central χ^2-distribution under H_0. Whenever the second implication beats the first, an increase in power under partition refinements is not guaranteed anymore. Since the publication of the Mann and Wald paper, many more papers on the choice and the number of cells have appeared. In general it is concluded that the Mann-Wald number of cell is too high (e.g. Quine and Robinson, 1985) and may even reduce the power for some specific alternatives. A comprehensive and practical oriented summary can be found in Moore (1986). In most of the papers on the subject the authors agree with the initial recommendation of equiprobable cells, still it is important to recognize that some others have other recommendations. Kallenberg, Oosterhoff and Schriever (1985), for instance, suggest that for heavy-tailed distributions in the alternative hypothesis, smaller cells in the tails may result in better power characteristics.

For a composite null hypothesis, the boundaries of equiprobable cells might depend on the unknown parameter $\boldsymbol{\beta} \in \Theta_0$. A typical solution is to estimate $\boldsymbol{\beta}$ and to subsequently use this estimator to determine the cell boundaries. An important consequence of this approach is that now the cells are *random* or *data-dependent* as apposed to the fixed cells. Actually, random cells are not restricted to cells determined by the estimator, but may more generally be constructed as $[A](\mathcal{S}_n) = \{A_1(\mathcal{S}_n), \ldots, A_k(\mathcal{S}_n)\}$. The cell probabilities are then given by

$$\pi_{0i}(\boldsymbol{\beta}, \mathcal{S}_n) = \int_{A_i(\mathcal{S}_n)} dG_{\boldsymbol{\beta}}(\boldsymbol{x}),$$

and they do not necessarily determine a multinomial distribution. Moreover, it is now to be expected that this procedure would loose some of its nonparametric nature in the sense that it may become dependent on the underlying distribution. Fortunately, under some mild restrictions, it can be shown that this is not the case: suppose that the random cell boundaries converge in probability to the corresponding boundaries of a set of fixed cells. Then, the limiting boundaries will generally depend on $\boldsymbol{\beta}$. Under these conditions, Moore and Spruill (1975) and Pollard (1979) concluded that any statistic that has a limiting null

distribution that is not a function of $\boldsymbol{\beta}_0$ in the fixed-cell case, has the same distribution for any choice of converging random cells. Thus, given the aforementioned conditions on the random cell convergence, for all statistics with a limiting χ^2 distribution, it is not necessary in the sequel to make a distinction between fixed and random cells.

An interesting special case is the scale-location family, and the normal distribution in particular. Watson (1959) showed that for an equiprobable Pearson test the results for fixed and random cell are the same.

In the section on the GQFs a latent or hypothesized random variable Y was introduced such that $\psi(Y) \sim M(\boldsymbol{\pi}; 1)$. When the data is continuous, say $X \sim F$, then $Y = X$ and the function ψ is the vector-valued indicator function

$$\boldsymbol{\psi}(Y) = (\mathrm{I}[Y \in A_1], \ldots, \mathrm{I}[Y \in A_k])'.$$

Note that, depending on the way the partition $[A]$ is constructed, ψ can be a random function and may possibly depend on the unknown parameter $\boldsymbol{\beta}_0 \in \Theta_0$ when the null hypothesis is composite.

Simple Null Hypothesis

First, note that whenever a simple null hypothesis is imposed, there is actually no need for data dependent cells when equiprobable cell are to be constructed.

Since for fixed cells the probabilities calculated from Equation 4.6 (with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$) determine uniquely a true multinomial distribution, the null distribution of Pearson's X_P^2 statistic applied to this setting is χ^2_{k-1}, and thus independent of the underlying distribution. From the previous discussion it follows that the same result holds for converging random cell boundaries.

Composite Null Hypothesis

For the composite null hypothesis problem there is one important additional issue that needs some more attention: the p-dimensional parameter $\boldsymbol{\beta}$, indexing the null hypothesis, must be estimated from the data. How must they be estimated, and will this affect the null distribution of the test statistic? From Moore's GQFs it is seen that $\boldsymbol{C}$, and thus also $\boldsymbol{\Sigma}$, depends explicitly on the vector-function $\boldsymbol{h}$ which is specific for each type of estimator for $\boldsymbol{\beta}$. Here, the two most frequently used estimators will be considered:

- **Maximum Likelihood estimator based on grouped data**. Suppose that the continuous data is grouped first, resulting in k counts $\boldsymbol{\Psi}(\mathcal{S}_{Y;n}) = \boldsymbol{X}$. Then the parameter $\boldsymbol{\beta}$ can be estimated in the same way as for the Pearson-Fisher test in Section 4.1.3, i.e. the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$. Therefore, for fixed cells this method results in exactly the Pearson-Fisher test, which has a limiting null dis-

tribution χ^2_{k-p-1} which is independent of the underlying distribution. Hence the same test applies to the present situation.

- **Maximum Likelihood estimator based on ungrouped data**. When the original continuous data is available, a more natural estimator is of course the maximum likelihood estimator $\tilde{\boldsymbol{\beta}}$ based on the original sample. Cramér (1954) showed that regular restricted maximum likelihood estimators can be written as

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{J}^{-1} \frac{\partial \ln g(x_i; \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} + o_p(n^{-1/2}),$$

where $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\beta}_0)$ is the information matrix with (i,j)th entry $\mathrm{E}\left[-\frac{\partial^2 \ln g(X;\boldsymbol{\beta}_0)}{\partial \beta_{0i} \partial \beta_{0j}}\right]$. Thus the estimator is again clearly of the form of Equation 4.4. Suppose for the moment that all cells are fixed. Then $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{J}^{-1}\boldsymbol{A}$ and, provided that $\boldsymbol{J} - \boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}\boldsymbol{A}$ is positive definite, $\mathrm{rank}(\boldsymbol{\Sigma}) = k - 1$. The generalized inverse is given by $\boldsymbol{\Sigma}^- = \boldsymbol{D}_{\pi_0} - \boldsymbol{C}$. A consistent estimator for $\boldsymbol{\Sigma}^-$ is obtained by replacing all $\boldsymbol{\beta}_0$ in the expression for $\boldsymbol{\Sigma}^-$ by the restricted maximum likelihood estimator $\tilde{\boldsymbol{\beta}}$. The resulting GQF with $\boldsymbol{Q}_n = \boldsymbol{\Sigma}_n^-$ is

$$X^2_{RR} = \boldsymbol{W}_n(\tilde{\boldsymbol{\beta}})' \left(\boldsymbol{D}_{\pi_0}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{A}(\tilde{\boldsymbol{\beta}})\boldsymbol{J}(\tilde{\boldsymbol{\beta}})^{-1}\boldsymbol{A}(\tilde{\boldsymbol{\beta}})\right)^{-1} \boldsymbol{W}_n(\tilde{\boldsymbol{\beta}}), \tag{4.7}$$

or equivalently

$$X^2_{RR} = \boldsymbol{W}_n(\tilde{\boldsymbol{\beta}})' \left(\boldsymbol{D}_{\pi_0}(\tilde{\boldsymbol{\beta}}) + \boldsymbol{A}(\tilde{\boldsymbol{\beta}}) \left(\boldsymbol{J}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{A}(\tilde{\boldsymbol{\beta}})'\boldsymbol{D}_{\pi_0}(\tilde{\boldsymbol{\beta}})^{-1}\boldsymbol{A}(\tilde{\boldsymbol{\beta}})\right)^{-1} \boldsymbol{A}(\tilde{\boldsymbol{\beta}})'\right) \boldsymbol{W}_n(\tilde{\boldsymbol{\beta}}). \tag{4.8}$$

And, under H_0, $X^2_{RR} \xrightarrow{d} \chi^2_{k-1}$. This result was first given by Rao and Robson (1974) and will be referred as the Rao-Robson test. At that time the test was to be seen as an improvement over the Chernoff-Lehmann (CL) test (Chernoff and Lehmann, 1954), which was simply defined by $X^2_{CL} = \boldsymbol{W}_n(\tilde{\boldsymbol{\beta}})'\boldsymbol{W}_n(\tilde{\boldsymbol{\beta}})$, i.e. the classical Pearson test statistic with the unknown parameter $\boldsymbol{\beta}$ replaced by it maximum likelihood estimator $\tilde{\boldsymbol{\beta}}$ based on the ungrouped data. Unfortunately, the limiting null distribution of X^2_{CL} depends on the underlying distribution:

$$\chi^2_{k-p-1} + \sum_{i=1}^{p} \lambda_i Z_i^2,$$

where the Z_i are i.i.d. standard normal variates, and the λ_i are the roots of the determinantal equation $|(\boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}\boldsymbol{A}) - (1-\lambda)\boldsymbol{J}| = 0$ (note that $\boldsymbol{A}'\boldsymbol{D}_{\pi_0}^{-1}\boldsymbol{A}$ actually is the information matrix of the multinomial obtained by the grouping). Since the information matrix $\boldsymbol{J}$ explicitly depends on the unknown parameter $\boldsymbol{\beta}_0$ and on the null model, the Chernoff-Lehmann statistic is not straightforward to use in practice, and, furthermore, this result also implies that a different null distribution may hold for the statistic when applied to random cells. In a series of papers Roy (1956) and

Watson (1957, 1958, 1959) studied the Chernoff-Lehmann statistic under random cells. They observed that for a scale-location family, and the normal distribution in particular, the cells can be chosen such that the asymptotic null distribution of $\boldsymbol{X}^2_{CL}$ looses its dependence on $\boldsymbol{\beta}_0$, though the distribution is not a simple distribution like e.g. the χ^2 distribution. Therefore, among other reasons, the Roy-Watson test is also not popular for practical use. Tables with percentage points for this distribution can be found in Dahiya and Gurland (1972, 1973). Yet a more important reason why the Rao-Robson test is to be preferred over the Chernoff-Lehmann or the Roy-Watson test is that on average the power of the Rao-Robson test is the highest. Also as compared to Pearson-Fisher the Rao-Robson has generally better power characteristics. These conclusions have many times been confirmed by means of simulation studies in literature (for an overview, see e.g. Moore (1986) and references therein). Intuitively, this is easy to understand: first, Rao-Robson uses the continuous data to get an efficient estimator of $\boldsymbol{\beta}$, and second, independent of the number of parameters in $\boldsymbol{\beta}$, the maximal number of degrees of freedom $(k-1)$ is retained.

4.1.5 A Generalization: the Power Divergence Statistics

Pearson's χ^2-statistic is not the only statistic that is appropriate for testing $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$, i.e. the one-sample GOF problem for a multinomial distribution. Other well known statistics are the likelihood ratio (LR) statistic, the Freeman-Tukey statistic, Neyman's modified X^2 statistic and the modified log-likelihood statistic. All these statistics have the same limiting null distributions and are therefore often referred to as χ^2 statistics in general.

Cressie and Read (1984) introduced a generalization of the above mentioned statistics. They found a family of statistics, indexed by a real valued parameter λ. The family is called the family of power divergence statistics and it is given by

$$2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} X_i \left\{ \left(\frac{X_i}{n\pi_{0i}} \right)^{\lambda} - 1 \right\},$$

where, as before, X_i $(i = 1, \ldots, k)$ are the observed frequencies. For $\lambda = 0$ and $\lambda = -1$, the corresponding statistics are defined by continuity. Then $\lambda = 0$ and $\lambda = 1$ give the LR and Pearson's statistic, respectively.

Cressie and Read (1984) showed that under a simple H_0 and for each $\lambda \in \mathbb{R}$,

$$2nI^{\lambda} - 2nI^{1} \xrightarrow{p} 0,$$

and thus

$$2nI^{\lambda} \xrightarrow{d} \chi^2_{k-1}.$$

When the null hypothesis is composite, say $\boldsymbol{\pi}_0(\boldsymbol{\beta})$ is known up to a p-dimensional nuisance parameter $\boldsymbol{\beta}$, then they defined the power divergence statistic $2n\hat{I}^\lambda$ as the plug-in estimator for which under H_0

$$2n\hat{I}^\lambda \xrightarrow{d} \chi^2_{k-p-1},$$

whenever the estimator $\hat{\boldsymbol{\beta}}$ is satisfies the same conditions as those stated at the construction of the Pearson-Fisher statistic (Section 4.1.4).

A power comparison of some interesting members of the family of power divergence statistics was given by Read (1984)

4.1.6 Testing Independence between two Discrete Variables

In Sections 4.1.1 and 4.1.3 Pearson's χ^2-test was introduced for GOF of a discrete rv. Only later it was shown how these techniques could be applied to continuous data as well. Pearson (1900, 1922) showed that a similar statistic can be constructed for testing independence between 2 discrete rvs. This test will be briefly discussed here.

Let $\boldsymbol{X} = (X_{11}, \ldots, X_{1c}, X_{21}, \ldots, X_{r1}, \ldots, X_{rc})$ denote an observation from a multinomial distribution of two cross classified discrete rvs, U and V, which are defined over a sample space with r and c different elements, respectively. The probabilities of the multinomial distribution are denoted by π_{ij} ($i = 1, \ldots, r; j = 1, \ldots, c$). The null hypothesis of independence between the two discrete rvs is

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \text{ for all } i = 1, \ldots, r; j = 1, \ldots, c, \tag{4.9}$$

where $\pi_{i.}$ and $\pi_{.j}$ are the marginal probabilities of the discrete rvs U and V, respectively. As the alternative hypothesis, the negation of H_0 is considered.

The null hypothesis in Equation 4.9 is composite as both marginal probabilities are unknown, and may be considered as finite dimensional nuisance parameters. They may be estimated from the data by $\hat{\pi}_{i.} = \frac{1}{n}\sum_{j=1}^{c} X_{ij}$ and $\hat{\pi}_{.j} = \frac{1}{n}\sum_{i=1}^{r} X_{ij}$. The following Pearson statistic is easily recognized to be of the same form as Equation 4.2,

$$X^2 = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(X_{ij}/n - \hat{\pi}_{i.}\hat{\pi}_{.j}\right)^2}{\hat{\pi}_{i.}\hat{\pi}_{.j}}.$$

Under H_0, $X^2 \xrightarrow{d} \chi^2_{(r-1)(c-1)}$ (Fisher, 1922).

Roy and Mitra (1956) showed that the above results hold in general under Poisson, multinomial and independent multinomial sampling.
Again, Pearson's statistic may be generalized to the power divergence family of Cressie and Read (1984).

4.1.7 Some Remarks on Exact Pearson-type GOF-Tests

One way to assess the practical value of asymptotic tests is to compare these tests with exact tests for finite sample sizes for the same hypotheses. The asymptotic null distribution is thus seen as an approximation to the null distribution of the exact test. Tate and Hyer (1973) compared Pearson's test with an exact multinomial test, which is based on the ordering of the exact multinomial probabilities

$$\mathrm{P}\left[\boldsymbol{X}=\boldsymbol{y}\right]=n!\Pi_{i=1}^{k}\frac{\pi_{0i}^{y_i}}{y_i!}$$

of observing all possible sample configurations $\boldsymbol{y}$. The p-value is given by the sum of probabilities $\mathrm{P}\left[\boldsymbol{X}=\boldsymbol{y}\right]$ which are smaller than the probability $\mathrm{P}\left[\boldsymbol{X}=\boldsymbol{x}\right]$ of observing the observed sample $\boldsymbol{x}$. He concluded that the χ^2 approximation is rather poor. Radlow and Alf (1975) and Horn (1977) argue that this comparison makes no sense because, although both tests address to the same GOF null hypothesis, they are constructed on a different criterion for measuring the deviation between the observed and the hypothesized frequencies.

Radlow and Alf (1975) proposed an alternative exact test for which it is appropriate to be compared with Pearson's test. Instead of ordering multinomial probabilities for all possible sample configurations $\boldsymbol{y}$, the corresponding X^2 statistics are calculated and ordered and thus both tests are based on the same measure. For this exact test, the χ^2 approximation performs better.

Many *corrected* versions of Pearson's test statistic are proposed in order to obtain a better χ^2 approximation. For a brief review, we refer to Cressie and Read (1989).

Also for the χ^2-test for independence between two discrete rvs exact tests are available. The best known is probably Fisher's exact test (Fisher, 1934). For a discussion on this test, we refer to Agresti (1990).

4.2 EDF-Based GOF-tests

A wide and diverse family of GOF-tests is based on the empirical distribution function (EDF). First, some properties of the EDF are summarized. Then, in Section 4.2.2 a general framework for this type of statistical tests is described. Then, in the subsequent sections the most frequently used and some historically important tests are discussed.
In general it is known that the EDF-based tests are more powerful than the Pearson χ^2-type tests.

4.2.1 The Empirical Distribution Function

An intuitively straightforward approach for testing the GOF null hypothesis $H_0 : F_x \in \mathcal{G}_0$ is to find an appropriate estimator for the CDF F_x and to compare this estimator with the family $\mathcal{G}_0$ based on some appropriate measure. In this section probably the most widely used estimator of F_x is discussed. We will restrict the discussion to univariate distributions. The extension to multivariate CDFs is straightforward.

Suppose that the sample $\mathcal{S}_n$ contains the n observations $X_1, \ldots, X_n$. The CDF $F = F_x$ may be estimated by the empirical distribution function (EDF) $\hat{F}_{x,n} = \hat{F}_n$, which is given by

$$\begin{aligned} \hat{F}_n(x) &= \frac{\text{number of observations } \leq x}{n} \\ &= \frac{\sum_{i=1}^{n} \mathrm{I}\left[X_i \leq x\right]}{n}, \end{aligned} \tag{4.10}$$

for $x \in \mathcal{S}_x$. From Equation 4.10 it may be seen directly that $n\hat{F}_n(x)$ is distributed as a binomial with parameters $F(x)$ and n. Then, of course, $\hat{F}_n(x)$ is an unbiased estimator of $F(x)$. Furthermore, when $n \to \infty$, the central limit theorem implies that

$$\sqrt{n}\left(\hat{F}_n(x) - F(x)\right) \xrightarrow{d} N\left(0, F(x)\left(1 - F(x)\right)\right).$$

Thus the variance of $\hat{F}_n(x)$ is of order $\frac{1}{n}$, and thus the estimator is consistent. Even a stronger convergence holds:

$$\sup_x \left|\hat{F}_n(x) - F(x)\right| \xrightarrow{p} 0 \text{ as } n \to \infty.$$

The definition of the EDF is easily extended to multivariate distributions. E.g. The bivariate CDF F_{xy} is estimated by

$$\hat{F}_{xy,n}(x, y) = \frac{\sum_{i=1}^{n} \mathrm{I}\left[X_i \leq x, Y_i \leq y\right]}{n},$$

where (X_i, Y_i) $(i = 1, \ldots, n)$ constitute the sample $\mathcal{S}_n$.

4.2.2 Statistical Tests: a Framework

Many of the EDF tests fit into a framework for statistical tests for GOF as it was introduced by Romano (1988, 1989).

Recall the hypotheses $H_0 : F \in \mathcal{G}_0$ and $H_1 : F \in \mathcal{G}_1$. Let δ be a measure on $\mathcal{F}_\infty$. δ must not necessarily be a metric, it may even be non-symmetric. Let ρ be a metric or a pseudometric on $\mathcal{F}_\infty$. Then, τ is defined as a mapping from $\mathcal{F}_\infty$ on $\mathcal{G}_0$ such that for all

$F \in \mathcal{F}_\infty$ $\tau(F)$ is smooth in F and

$$\rho\left(F, \tau(F)\right) = \inf_{Q \in \mathcal{G}_0} \rho\left(F, Q\right).$$

Thus, $\tau(F)$ is considered as a projection of F onto $\mathcal{G}_0$ (Figure 4.1). Also, for all $F \in \mathcal{G}_0$, $\tau(F) = F$. Also note that when the null hypothesis is simple, i.e. $\mathcal{G}_0 = \{G\}$, then for all F, $\tau(F) = G$.

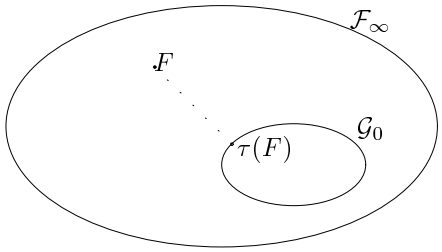


Figure 4.1: The mapping $\tau(F)$ on $\mathcal{G}_0$ can be seen as a projection. The dashed line represents the shortest "distance" between F and $\tau(F) \in \mathcal{G}_0$ w.r.t. the pseudometric ρ.

Then, in general, GOF-test statistics are of the form

$$T_n = \tau_n \delta\left(\hat{F}_n, \tau(\hat{F}_n)\right),$$

where τ_n is a normalizing factor such that T_n has a a non-degenerate null distribution, and where $\hat{F}_n$ is the EDF.

Actually Romano used this framework at first only with δ being the supremum distance over an appropriate Vapnik-Cernovenkis class of sets (cfr. Kolmogorov-Smirnov statistic). With this particular choice, which still results in a wide class of statistics, he showed the (asymptotic) validity of the bootstrap test (Romano, 1988), and under an invariance implying null hypothesis he compared this bootstrap test to the corresponding randomization test (Romano, 1989). The bootstrap test has the advantage of being asymptotically valid under very weak assumptions on the true, underlying distribution F (e.g. F must not be continuous).
Later, Politis, Romano and Wolf (1999) showed that for much weaker conditions on δ, the bootstrap is still a valid method of testing.
Although we will not further apply the bootstrap in this thesis, it is interesting to know that the EDF-based tests, as well as the new tests that are proposed in this work, fit into Romano's framework.

4.2.3 Tests of the Kolmogorov-Smirnov Type

A first class of tests origins from the ideas of Kolmogorov (1933) and Smirnov (1939). Two forms are discussed here: one for the one-sample and one for the k-sample problem. In general tests of this type are based on a supremum distance measure between the EDF of the true distribution and the CDF of the hypothesized distribution.

The One-Sample Problem

First, consider the simple null hypothesis $H_0 : F(x) = G(x)$, for all $x \in \mathcal{S}_x$. The Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933; Smirnov, 1939) is then given by

$$D_n = \sqrt{n} \sup_{x \in \mathcal{S}_x} \left| \hat{F}_n(x) - G(x) \right|,$$

which may be described within Romano's original framework with $\tau(\hat{F}_n(x)) = G(x)$, $\tau_n = \sqrt{n}$ and δ is the supremum operator on the absolute difference.

Kolmogorov (1933) has proven that the limiting null distribution of D_n is given by

$$F_d(d) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1} \exp\left(-2j^2 d^2\right),$$

which clearly does not depend on the distribution G, nor on the true distribution F. Thus the Kolmogorov-Smirnov test is nonparametric. Critical values are however not easily obtained from the limiting null distribution; tables are provided by Massey (1951, 1952), Owen (1962). For moderate sample sizes, Stephens (1970) proposed a modified statistic

$$D_n^m = D_n \left(1 + \frac{0.12}{\sqrt{n}} + \frac{0.24}{n}\right),$$

for which he gave the upper tail critical values for several values for α.
The Kolmogorov-Smirnov test is consistent against essentially any alternative.

An overview of other related statistics are given by D'Agostino and Stephens (1986).

When the null hypothesis is composite, a similar test statistic may be used:

$$D_n = \sqrt{n} \sup_{x \in \mathcal{S}_x} \left| \hat{F}_n(x) - \hat{G}(x) \right|,$$

where $\hat{G} = G_{\hat{\theta}}$. Although this seems to be a natural extension of the statistic in the simple null hypothesis case, there is however no general asymptotic theory available. Critical values must be estimated by e.g. simulation. For the case where G is a normal distribution with unknown mean and variance, Lilliefors (1967) obtained critical values. Stephens

(1974) proposed a modified statistic for testing normality with unknown parameters

$$D_n^m = D_n \left(1 - \frac{0.01}{\sqrt{n}} + \frac{0.85}{n}\right),$$

for which critical values are given by Stephens (1974, 1986).

Although the tests discussed in this section are widely known and very frequently applied in daily statistical practice, it is often reported (e.g. Stephens, 1986) that its power is inferior as compared to other EDF-based GOF tests.

The k-Sample Problem

Smirnov (1939) was the first to introduced the KS test for the 2-sample problem. Let $\hat{F}_{j,n_j}$ ($j = 1, 2$) denote the EDF of sample j. The 2-sample KS statistic is given by

$$D_n = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max_{x \in \mathcal{S}_x} \left| \hat{F}_{2,n_2}(x) - \hat{F}_{1,n_1}(x) \right|.$$

Its limiting null distribution is given by (Kolmogorov, 1941; Smirnov, 1939, 1948).

$$F_d(d) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} \exp\left(-2j^2 d^2\right).$$

Critical values for small sample sizes have been tabulated by Massey (1951, 1952), Owen (1962).

When $k > 2$, several extension have been proposed. An overview is given by Hájek and Šidák (1967). Some generalization have been studied by Romano (1988), for which he used the bootstrap to obtain critical values. We will only mention one extension here. Let $X_1, \ldots, X_n$ denote the combined sample ($n = \sum_{j=1}^{k} n_j$). This statistic is given by (Chang and Fisz, 1957; Fisz, 1960; Dwass, 1960)

$$D_{k,n} = \max_{j=2,\ldots,k} \max_{i=1,\ldots,n} |\hat{F}_j(X_i) - \frac{1}{\sum_{l=1}^{j-1} n_l} \sum_{l=1}^{j-1} n_l \hat{F}_l(X_i)|,$$

which is a maximum of $k - 1$ independent KS statistics. Note that the original 2-sample KS test of Smirnov (1939) is not a special case of the k-sample KS test that is given here.

4.2.4 Tests of the Cramér-von Mises and Anderson-Darling Type

A wide class of tests is based on the Cramér-von Mises statistic, which has its origin in the work of Cramér (1928), von Mises (1931, 1947). Later this test was generalized

by Anderson and Darling (1952, 1954). Tests will be subsequently discussed for the three GOF-problems. For the one-sample and the k-sample problems, Anderson-Darling tests will be given. For the independence problem, a Cramér-von Mises like test will be discussed.

The One-Sample Problem

First the simple null hypothesis is considered.
Anderson and Darling (1952) introduced the family of statistics

$$A_n = n \int_{\mathcal{S}_x} \left(\hat{F}_n(x) - G(x) \right)^2 \Psi\left(G(x)\right) dG(x), \tag{4.11}$$

where $\Psi \geq 0$ is a weight function. If $\Psi = 1$, then A_n reduces to the original Cramér-von Mises statistic. Although many choices for Ψ are allowed, in literature mainly the choice

$$\Psi(u) = \frac{1}{u(1-u)} \tag{4.12}$$

is taken. Moreover, the test with this particular weight function is often called **the** Anderson-Darling (AD) test. It seems as if almost no other choices appear in literature. In the remainder of this thesis, we will restrict the attention to the weight function in Equation 4.12, and the resulting statistic will be referred as the Anderson-Darling statistic. For general Ψ it will be called the Anderson-Darling family of statistics.

It is again straightforward to see that the AD statistic fits into the Romano's framework. Intuitively, the difference between the AD and the KS measure for the deviation between the EDF and G is that the latter looks over the complete sample space for the maximal difference in absolute value, whereas the former integrates the differences over the complete sample space, taking the weight function into account. The AD weight function $\frac{1}{G(x)(1-G(x))}$ upweighs the differences between the EDF and G in the tails of the distribution G. This is generally considered to be the reason why the AD test is overall more powerful than the Cramér-von Mises test.

The formula in Equation 4.11 is however not convenient for computation. The equivalent computation formula is

$$A_n = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \left(\ln(Z_{(i)}) + \ln(1 - Z_{(n+1-i)}) \right), \tag{4.13}$$

where $Z_{(1)} \leq \ldots \leq Z_{(n)}$ are the order statistics of the variable $Z = G(X)$.

The limiting null distribution may be found by considering the empirical process

$$S_n(t) = \frac{\sqrt{n}(\hat{F}_n(t) - t)}{\sqrt{t(1-t)}},$$

which has covariance function $\rho_0(s,t) = (s(1-s)t(1-t))^{-1/2}(s \wedge t - st)$ (Anderson and Darling, 1952). The solution of the corresponding integral equation gives immediately the limiting null distribution. In particular, under H_0, the AD statistic has asymptotically the same distribution as

$$\sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2, \tag{4.14}$$

where the Z_j ($j = 1, \dots$) are i.i.d. standard normal.
Critical values from this distribution have been tabulated by Lewis (1961). A striking observation is that for $n > 3$, the distribution of A_n is accurately approximated by its asymptotic distribution. Percentiles from the asymptotic null distribution may be approximately obtained by fitting Pearson curves (Solomon and Stephens, 1978).

The AD test is consistent against essentially any alternative. Power studies have indicated that the power of the AD test is frequently superior to the power of the Cramér-von Mises test. Both are generally more powerful than the KS test.

When the null hypothesis is composite, the AD statistic becomes

$$A_n = n \int_{\mathcal{S}} \frac{\left(\hat{F}_n(x) - \hat{G}(x)\right)^2}{\hat{G}(x)\left(1 - \hat{G}(x)\right)} d\hat{G}(x), \tag{4.15}$$

where $\hat{G}(x) = G_{\hat{\theta}}$. As with the KS test, the asymptotic theory becomes much more complicated as compared to the simple null hypothesis case. In general the asymptotic null distribution is equivalent to the distribution of

$$\sum_{j=1}^{n} \lambda_j Z_j^2,$$

where the Z_j ($j = 1, \dots$) are i.i.d. standard normal, and where the coefficients λ_j are the eigenvalues of an integral equation in which the covariance function depends on the distribution G, the parameters $\boldsymbol{\theta}$ that must be estimated from the data, and the method of estimation of these nuisance parameters (Darling, 1955; Stephens, 1971, 1976; Sukhatme, 1972). Thus the λ-coefficients must be determined for each G.
When the unknown parameters are those of location (e.g. mean) and of scale (e.g. variance), and when an appropriate method of estimation is used (e.g. maximum likelihood), then the null distribution of A_n does not depend on the unknown parameters. Thus, in these situations the null distributions only depend on the family $\mathcal{G}_{\Theta_0}$ tested and the sample size n. The normal and the exponential distribution are well known examples, and asymptotic results are available for them, as well as modifications to the test statistic such that approximate percentage points from the exact null distributions for finite n can be calculated easily.

The next theorem is a summary of some results of Darling (1955), Durbin (1973), Stephens

(1971), Sukhatme (1972).

Theorem 4.1 *Under the composite null hypothesis $H_0 : F \in \mathcal{G}_{\Theta_0}$, where $\mathcal{G}_{\Theta_0}$ is a location-scale family, the A_n statistic converges weakly to*

$$\sum_{j=1}^{\infty} \lambda_j Z_j^2,$$

where the Z_j^2 are i.i.d. χ_1^2 variates, and where the λ_j are the eigenvalues of the integral equation

$$\lambda h(x) = \int_0^1 \rho(x,y)h(y)dy,$$

where ρ is the covariance function given by

$$\rho(s,t) = \rho_0(s,t) - \phi_1(s)\phi_1(t) - \phi_2(s)\phi_2(t),$$

where $\rho_0(s,t)$ denotes the covariance function associated with the limiting null distribution of the AD_n statistic under a simple null hypothesis, and where the functions ϕ_i ($i = 1, 2$) depend on the family $\mathcal{G}_{\Theta_0}$ but not on the unknown location and scale parameters.

When G is the normal distribution where the mean and the variance must be estimated, Stephens (1976) gives the first ten largest λ-coefficients as well as the upper tail percentage points for some values of α. For moderate sample sizes Stephens (1986) gives a modified statistic

$$A_n^m = A_n \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right),$$

and the corresponding percentage points. Also an approximation formula for the calculation of the p-value is available (Pettit, 1977).

When the unknown parameters are not restricted to those of location and scale, e.g. when a shape parameter is involved as for the gamma distribution, then the null distribution, even the limiting null distribution, will still depend on the unknown parameter(s). A solution to this problem is the half-sample method (Stephens, 1978), but since half of the sample is lost for testing, a severe reduction in power may be expected.

Finally, note that the AD statistic may be interpreted as a plug-in estimator of the statistical functional

$$\begin{aligned} A(F;G_\theta) &= \int_{\mathcal{S}_x} \left(F(x) - G_\theta(x)\right)^2 \Psi\left(G_\theta(x)\right) dG_\theta(x) \\ &= \mathrm{E}_g\left[\left(F(X) - G_\theta(X)\right)^2 \Psi\left(G_\theta(X)\right)\right]. \end{aligned}$$

Thus, the Anderson-Darling statistics are closely related to (degenerate) V-statistics (see e.g. Lee, 1990).

The k-Sample Problem

Pettit (1976) proposed an AD-type test for the 2-sample problem. The statistic is defined as

$$A_{n_1 n_2} = \frac{n_1 n_2}{n} \int_{\mathcal{S}} \frac{\left(\hat{F}_{1,n_1}(x) - \hat{F}_{2,n_2}(x)\right)^2}{\hat{F}_n(x)\left(1 - \hat{F}_n(x)\right)} d\hat{F}_n(x), \tag{4.16}$$

where $\hat{F}_n$ is the EDF of the combined sample, i.e.

$$\hat{F}_n(x) = \frac{1}{n}\left(n_1 \hat{F}_{1,n_1}(x) + n_2 \hat{F}_{2,n_2}(x)\right).$$

$\hat{F}_n$ may be considered as the estimator of the common CDF $F = G$ under the null hypothesis.

The AD statistic in Equation 4.16 simplifies to

$$A_{n_1 n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n-1} \frac{(M_i n - n_1 i)^2}{i(n-i)},$$

where M_i is the number of observations of the first sample that are less than or equal to the ith smallest observation in the combined sample. Thus, the 2-sample AD test is a rank test. Furthermore, the null hypothesis implies an invariance property under the group of $\frac{n!}{n_1! n_2!}$ permutations of the observations across two samples with n_1 and n_2 observations. Therefore the exact null distribution for finite sample sizes may be obtained by enumerating all possible values of the test statistic under the group of permutations. Pettit (1976) gives exact upper percentage points for small sample sizes. He has also proven that the limiting null distribution is exactly the same as for the one-sample (simple null hypothesis) AD test which is given in Equation 4.14.

Since for the one-sample AD test (simple null hypothesis) it is observed that the convergence to the asymptotic null distribution is very fast, it is more or less expected to hold in the present situation as well. Pettit (1976) proposed to use a kind of standardized statistic

$$A^s_{n_1 n_2} = (A_{n_1 n_2} - 1)(1 + \frac{1.55}{n}) + 1,$$

for which the exact expectation and variance are equal to the asymptotic mean and variance. This statistic may be used to enter the table of critical values of the standardized asymptotic null distribution.

A generalization of the AD test to the k-sample case is given by Scholz and Stephens (1987). Their statistic is

$$A_{k,n} = \sum_{j=1}^{k} n_j \int_{\mathcal{S}} \frac{\left(\hat{F}_{j,n_j}(x) - \hat{F}_n(x)\right)^2}{\hat{F}_n(x)\left(1 - \hat{F}_n(x)\right)} d\hat{F}_n(x).$$

Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the n order statistics of the combined sample. The test statistic simplifies to (supposing that there are no ties)

$$A_{k,n} = \frac{1}{n} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{i=1}^{n-1} \frac{(nM_{ij} - in_j)^2}{i(n-i)},$$

where M_{ij} is the number of observations in the jth sample that less than or equal to $X_{(i)}$. Note that when $k = 2$ the the k-sample statistic reduces to the 2-sample statistic of Pettit (1976). Again the statistic is a rank statistic, and the null hypothesis implies a similar group invariance property as for the 2-sample problem. The asymptotic null distribution of $A_{k,n}$ is the same as the distribution of

$$\sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2,$$

where the Z_j^2 ($j = 1, \ldots$) are i.i.d. χ^2 random variables with $k - 1$ degrees of freedom. Scholz and Stephens (1987) suggest that approximate percentiles may be calculated using Pearson curves (Solomon and Stephens, 1978). A good approximation to the asymptotic null distribution is obtained by using a standardization correction for the exact mean and variance, as it was done for the 2-sample case.

The test is consistent against essentially any alternative.

The Independence Problem

The test presented in this section was originally studied by Hoeffding (1948). Later Blum et al. (1961) developed a more general test for testing independence between multivariate variables which reduces to Hoeffding's test when applied to independence between univariate variables. Moreover, in the latter approach the test statistic is specified "in the spirit of" the Cramér-von Mises statistic. There is also a true Cramér-von Mises statistic for testing for independence (De Wet, 1980), but this is not included here. Hoeffding's test statistic is

$$B_n = n \int_{\mathcal{S}} \left(\hat{F}_{xy,n}(x,y) - \hat{F}_{x,n}(x)\hat{F}_{y,n}(y)\right)^2 d\hat{F}_{xy,n}(x,y),$$

where $\hat{F}_{xy,n}$ is the EDF of the joint distribution of X and Y. Note that the difference with the true Cramér-von Mises statistic is that here the integral is taken w.r.t. the true distribution in stead of the hypothesized distribution reflecting independence. Blum et al. (1961) argue that the true Cramér-von Mises statistic would be asymptotically equivalent under the null hypothesis.

The statistic $\frac{B_n}{n}$ is actually the plug-in estimator of the functional

$$\begin{aligned}\Delta(F_{xy}; F_x F_y) &= \int_{\mathcal{S}} \left(F_{xy}(x,y) - F_x(x)F_y(y)\right)^2 dF_{xy}(x,y) \\ &= \mathrm{E}_{f_{xy}}\left[\left(F_{xy}(x,y) - F_x(x)F_y(y)\right)^2\right],\end{aligned}$$

which measures the discrepancy between the true and the hypothesized distribution (cfr. Romano's framework). Note that the expectation in the latter equation is taken w.r.t. the true distribution, whereas a typical Cramér-von Mises functional would be an expectation w.r.t. the hypothesized distribution.

Hoeffding (1948) provided a computational formula for B_n (actually a slight modified form) by recognizing that it actually is a V-statistic based on a kernel of degree 5. Blum et al. (1961) proposed another computational formula,

$$B_n = \frac{1}{n^4}\sum_{i=1}^{n}\left(N_1(i)N_4(i) - N_2(i)N_3(i)\right)^2,$$

where $N_1(i)$, $N_2(i)$, $N_3(i)$ and $N_4(i)$ are the number of observations in the intersection of $\mathcal{S}_n$ with $\{(x,y) : x \leq X_i, y \leq Y_i\}$, $\{(x,y) : x > X_i, y \leq Y_i\}$, $\{(x,y) : x \leq X_i, y > Y_i\}$ and $\{(x,y) : x > X_i, y > Y_i\}$, respectively (see Figure 4.2). Thus, the statistic can be interpreted as a statistic proportional to the average, over all observations, of a measure of dependence in a 2×2 contingency table "centred" at an observation.
From the computational formula it may also be seen that B_n is a rank statistic. Since the null hypothesis of independence implies a group invariance property under the group of all $n!$ permutations of the observations on X (or, equivalently, on Y), the exact null distribution can be enumerated.

Intuitively, it may be seen that the null distribution of B_n will be the same as a true Cramér-von Mises statistic, because f_{xy} becomes $f_x f_y$ under H_0. Under H_0, B_n converges weakly to the random variable

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\frac{1}{\pi^4 j^2 k^2} Z_{jk}^2,$$

where the Z_{jk} $(j,k = 1,\dots)$ are i.i.d. standard normal.

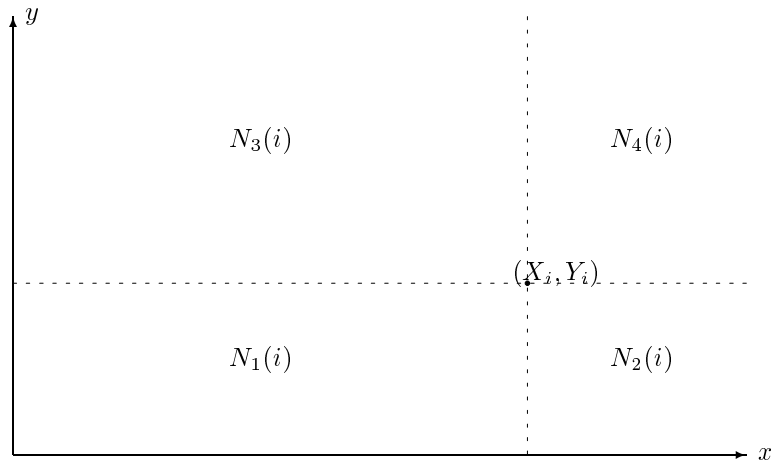


Figure 4.2: Illustration of an observation (X_i, Y_i) implying a 2×2 contingency table.

4.3 Smooth GOF-tests

Neyman (1937) introduced the smooth GOF test, which in its original formulation could only test simple null hypotheses $H_0 : F(x) = G(x)$, for all $x \in \mathcal{S}_x$. In particular Neyman argued that the integral transformation G always can be applied. Therefore he only constructed the smooth test for testing uniformity. From a theoretical point of view, an interesting property of Neyman's smooth test is asymptotically locally uniformly most powerful symmetric, unbiased and of size α.

Since 1937 many changes have been suggested to Neyman's smooth test. In this section a brief overview of the modern interpretation of this class of tests is given.

4.3.1 A General Form of Smooth Tests

The construction of smooth tests that is given in this section is mainly taken from Rayner and Best (1989).

The term "smooth" refers to the characteristic that the distribution G, which is specified in the null hypothesis and which may depend on a nuisance parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$, is imbedded in a family $\mathcal{F}_\Theta$ of alternatives F_θ which varies smoothly with the parameters $\theta_1, \ldots, \theta_s$. The density of the sth order alternative is given by

$$f_s(x; \boldsymbol{\theta}, \boldsymbol{\beta}) = c(\boldsymbol{\theta}, \boldsymbol{\beta}) \exp\left(\sum_{i=1}^{s} \theta_i h_i(x, \boldsymbol{\beta})\right) g(x; \boldsymbol{\beta}), \tag{4.17}$$

where $c(\boldsymbol{\theta}, \boldsymbol{\beta})$ is a normalization constant, and where $\{h_i(., \boldsymbol{\beta})\}$ $(i = 1, \ldots, s)$ is a complete orthonormal system (CONS) of functions on g. Thus, $(i, j = 1, \ldots, s)$

$$\int_{\mathcal{S}_x} h_i(x, \boldsymbol{\beta}) h_j(x, \boldsymbol{\beta}) g(x, \boldsymbol{\beta}) dx = \delta_{ij},$$

where δ_{ij} is Kronecker's delta. The latter condition of complete orthonormality has not always been imposed. The first contributions to smooth tests for composite null hypotheses (e.g Kopecky and Pierce, 1979; Thomas and Pierce, 1979) were all based on moment-like functions $h_i(x) = x^i$ which clearly do not possess the orthonormality restriction. The main advantages of orthonormal functions are that (1) the asymptotic null distribution of the test statistic turns out to be a simple χ^2-distribution, whereas for the other functions the null distributions are far more complicated, that (2) the components often are easily interpretable, that (3) they have a standard normal limit distribution, and that (4) they are asymptotically independent.
Let $\boldsymbol{h}(X; \boldsymbol{\beta}) = (h_1(X; \boldsymbol{\beta}), \ldots, h_s(X; \boldsymbol{\beta}))'$.

From Equation 4.17 it is immediately seen that the null hypothesis is equivalent to

$$H_0 : \boldsymbol{\theta} = \mathbf{0}.$$

Since the df of the family of alternatives is explicitly given, likelihood inference may be applied directly. In particular, the score test is constructed. Let $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimator of the nuisance parameter $\boldsymbol{\beta}$. The maximum likelihood estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{h}(X_i; \hat{\boldsymbol{\beta}}).$$

Note that the score is $\sum_{i=1}^{n} \boldsymbol{h}(X_i; \boldsymbol{\beta}) = n\hat{\boldsymbol{\theta}}$. The score test statistic is then given by

$$S_{s;n} = \hat{\boldsymbol{\theta}}' \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\theta}},$$

where $\hat{\boldsymbol{\Sigma}}$ is the estimator of the variance-covariance matrix $\boldsymbol{\Sigma}$ of the $\hat{\boldsymbol{\theta}}$ with $\boldsymbol{\beta}$ replaced by its maximum likelihood estimator $\hat{\boldsymbol{\beta}}$.

$$\boldsymbol{\Sigma} = \frac{1}{n} \left(\boldsymbol{I} - \mathrm{Cov}_g \left[\boldsymbol{h}, \frac{\partial \ln g}{\partial \boldsymbol{\beta}} \right]' \mathrm{Var}_g \left[\frac{\partial \ln g}{\partial \boldsymbol{\beta}} \right]^{-1} \mathrm{Cov}_g \left[\boldsymbol{h}, \frac{\partial \ln g}{\partial \boldsymbol{\beta}} \right] \right),$$

where $\boldsymbol{I}$ is the $s \times s$ identity matrix. The score statistic is however only defined when $\hat{\boldsymbol{\Sigma}}$ is nonsingular. In this case, under H_0, $S_{s;n} \xrightarrow{d} \chi^2_{s-q}$. For further purposes it is more convenient to write the test statistic as

$$S_{s;n} = \hat{\boldsymbol{U}}' \hat{\boldsymbol{M}}^{-1} \hat{\boldsymbol{U}},$$

where $\hat{\boldsymbol{M}} = n\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{U}} = n^{1/2}\hat{\boldsymbol{\theta}}$.

One way to avoid the problem of $\hat{\boldsymbol{M}}$ being singular is to define a $s \times p$ matrix $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{\beta})$, which depends on $\boldsymbol{\beta}$, and which is chosen such that $\hat{\boldsymbol{B}}'\hat{\boldsymbol{M}}\hat{\boldsymbol{B}}$ is nonsingular. Moreover, it is always possible to construct a $\hat{\boldsymbol{B}}$ such that $\hat{\boldsymbol{B}}'\hat{\boldsymbol{M}}\hat{\boldsymbol{B}} = I$ for any $p \leq s - q$. Under these conditions the score statistic becomes

$$\begin{aligned} S_{s;n} &= \hat{\boldsymbol{U}}'\hat{\boldsymbol{B}}\left(\hat{\boldsymbol{B}}'\hat{\boldsymbol{M}}\hat{\boldsymbol{B}}\right)^{-1}\hat{\boldsymbol{B}}'\hat{\boldsymbol{U}} \\ &= \left(\hat{\boldsymbol{B}}'\hat{\boldsymbol{U}}\right)'\left(\hat{\boldsymbol{B}}'\hat{\boldsymbol{U}}\right) \\ &= \hat{\boldsymbol{V}}'\hat{\boldsymbol{V}}, \end{aligned}$$

where $\hat{\boldsymbol{V}} = \hat{\boldsymbol{B}}'\hat{\boldsymbol{U}}$. Thus the score statistic simply reduces to a sum of p squared components,

$$S_{s;n} = \sum_{i=1}^{p} \hat{V}_i^2,$$

where $\hat{V} = n^{-1/2}\hat{\boldsymbol{B}}'\sum_{i=1}^{n} h(X_i;\hat{\boldsymbol{\beta}})$. Furthermore, $\hat{\boldsymbol{V}}$ is asymptotically multivariate normally distributed with as variance-covariance matrix the $p \times p$ identity matrix, implying that the p components are asymptotically independent.

By choosing an appropriate $\hat{\boldsymbol{B}}$ and an appropriate system of complete orthonormal functions, the components may have a desirable interpretation. Later examples will be given.

4.3.2 An Alternative Formulation of the Smooth Test

Some authors prefer to use another parameterization of a family of "smooth" alternatives to the null hypothesis (e.g Barton, 1955, 1956; Hamdan, 1962, 1964; Kendall and Stuart, 1973):

$$f_s(x;\boldsymbol{\theta},\boldsymbol{\beta}) = \left(1 + \sum_{i=1}^{s} \theta_i h_i(x;\boldsymbol{\beta})\right) g(x;\boldsymbol{\beta}), \tag{4.18}$$

where $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\{h(.;\boldsymbol{\beta})\}$ are as before. The null hypothesis is still $H_0 : \boldsymbol{\theta} = \mathbf{0}$. This family of alternatives does not alter the form of the test statistic.

Inserting f_s in Pearson's functional (Lancaster, 1969)

$$\phi_s^2 = \int_{\mathcal{S}_\S} \left(\frac{f_s(x;\boldsymbol{\beta})}{g(x;\boldsymbol{\beta})}\right)^2 g(x;\boldsymbol{\beta})dx - 1,$$

results in

$$\begin{aligned}\phi_s^2 &= \int_{\mathcal{S}_\S} \left(1 + \boldsymbol{\theta}'\boldsymbol{h}(x;\boldsymbol{\beta})\right)^2 g(x;\boldsymbol{\beta})dx - 1 \\ &= \sum_{j=1}^{s} \theta_j^2.\end{aligned}$$

The parameters $\boldsymbol{\theta}$ can be estimated as (similar as in Eubank et al., 1987)

$$\hat{\boldsymbol{\theta}}_j = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(X_i;\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$. The test statistic becomes

$$\begin{aligned}T_{s;n} &= n\sum_{j=1}^{s} \hat{\theta}_j^2 \\ &= \sum_{j=1}^{s} \left(\sqrt{n}\hat{\theta}_j\right)^2 \\ &= \sum_{j=1}^{s} \hat{V}_j^2.\end{aligned}$$

4.3.3 Smooth Tests for the Three GOF Problems

In this section the above theory is applied to the three GOF problems that are discussed in this thesis.

One Sample GOF Problem

For the one sample problem the methods described above can be applied directly. Two approaches for testing normality with unknown mean and variance ($\boldsymbol{\beta} = (\mu, \sigma)'$) are described briefly next.

The first one is actually due to Lancaster (1969). Within the general framework of smooth tests of Section 4.3.1 the Hermite polynomials may be used, after the observations are standardized as $Z_i = \frac{X_i - \bar{X}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}}$. With these polynomials it is seen immediately that $\sum_{i=1}^{n} h_j(Z_i;\hat{\boldsymbol{\beta}}) = 0$ ($j = 1, 2$) is implied by $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. An evident choice for $\hat{\boldsymbol{B}}$ is thus a matrix with (r, s)th element equal to $\delta_{u-2,v}$ ($u =$

$3, \ldots, s; v = 1, \ldots, s$). With this choice, $\hat{\boldsymbol{B}}'\hat{\boldsymbol{M}}\hat{\boldsymbol{B}} = \boldsymbol{I}$, and the test statistic becomes

$$S_{s;n} = \sum_{j=3}^{s} \hat{V}_j^2,$$

where $\hat{V}$ is defined as before. The first component, $\hat{V}_3$, turns out to be the classical third moment test for testing symmetry (Gupta, 1967).

The second approach resembles more the original formulation of the smooth test. First the observations are transformed according to $Z_i = G(X_i; \hat{\boldsymbol{\beta}})$. with this choice, the Legendre polynomials are appropriate. This method is general in the sense that it can applied to all G. This is exactly what is described by Eubank et al. (1987).

Finally, we like to mention that Thomas and Pierce (1979) proposed another smooth test for the composite null hypothesis which is based on a modified score statistic. It consists mainly in using a rank $s - 1$ matrix as an estimator for the singular matrix $\boldsymbol{M}^{-1}$. This procedure is very similar to the construction of the Rao-Robson statistic (Section 4.1.4). Thomas and Pierce (1979) did hover not use orthonormal polynomials, but rather the moment-like functions.

The k-Sample Problem

Let $H(x) = \sum_{j=1}^{k} \lambda_j F_j(x)$, where $\lambda_j = \lim_{n\to\infty} \frac{n_j}{n}$. Then the original k-sample null hypothesis is equivalent to $H_0 : F_1(x) = H(x); \ldots; F_k(x) = H(x)$, for all $x \in \mathcal{S}_x$, which may be considered as k null hypotheses $H_{0;j} : F_j(x) = H(x)$, for all $x \in \mathcal{S}_x$ ($j = 1, \ldots, k$). As in Eubank et al. (1987), each distribution f_j may be parameterized as in Equation 4.18, where g is replaced with the df corresponding to the CDF H, and with parameters $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{js})'$. For each corresponding null hypothesis $H_{0;j}$ a set of components

$$\hat{V}_{jl} = n_j^{-1/2} \sum_{i=1}^{n_j} h_l(\hat{H}(X_{ji})),$$

is calculated, where $\hat{H} = \sum_{j=1}^{k} \frac{n_j}{n} \hat{F}_j$, and where $\{h_l\}$ are the Legendre polynomials ($l = 1, \ldots, s$). The ks components $\hat{V}_{jl}$ can be combined into s components $\hat{V}_l^2 = \sum_{j=1}^{k} \frac{n_j}{n} \hat{V}_{jl}^2$.

Eubank et al. (1987) showed that $\hat{V}_1$ and $\hat{V}_2$ are the Kruskal-Wallis statistic and the k-sample generalization of the Mood statistic, respectively. More generally, the test statistic is a rank statistic.

The Independence Problem

We consider here only the case of bivariate independence. A small change to the model in Equation 4.17 is needed. When the marginal distributions F_x and F_y are normal distributions, Kallenberg et al. (1997) proposed to consider the alternatives (based on a similar family introduced by Koziol (1979))

$$f_s(x) = \frac{1}{\sigma_x}\frac{1}{\sigma_y} f_x(x) f_y(y) \exp\left(\sum_{j=1}^{s} \theta_j h_j\left(\frac{x-\mu_x}{\sigma_x}\right) h_j\left(\frac{y-\mu_y}{\sigma_y}\right)\right), \qquad (4.19)$$

where the polynomials $\{h_j\}$ are the Hermite polynomials, and where (μ_x, σ_x) and (μ_y, σ_y) are the mean and the standard deviation of f_x and f_y, respectively. The corresponding estimators are denoted by $(\hat{\mu}_x, \hat{\sigma}_x)$ and $(\hat{\mu}_y, \hat{\sigma}_y)$. The resulting score statistic is (originally proposed by Koziol (1979))

$$\begin{aligned} S_{s;n} &= \sum_{j=1}^{s}\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} h_j\left(\frac{X_i-\hat{\mu}_x}{\hat{\sigma}_x}\right) h_j\left(\frac{Y_i-\hat{\mu}_y}{\hat{\sigma}_y}\right)\right)\right)^2 \\ &= \sum_{j=1}^{s}\left(\sqrt{n}\hat{\theta}_j\right)^2. \end{aligned}$$

The first component of $S_{s;n}$ turns out to be n times the square of Pearson's correlation coefficient (Kallenberg et al., 1997).

A more general method consists in taking first the marginal integral transformations $F_x(X)$ and $F_y(Y)$. The family of alternatives is then similar to Equation 4.19,

$$f_s(\boldsymbol{x}) = c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{s} \theta_j h_j(F_x(x)) h_j(F_y(y))\right), \qquad (4.20)$$

(Kallenberg and Ledwina, 1999) where the $\{h_j\}$ are now Legendre polynomials, and where $c(\boldsymbol{\theta})$ is a normalization constant. A further generalization is obtained by dropping the restriction that the cross-product terms in Equation 4.20 are symmetric. Then, the family of alternatives becomes

$$f_{st}(\boldsymbol{x}) = c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{s}\sum_{k=1}^{t} \theta_{jk} h_j(F_x(x)) h_k(F_y(y))\right). \qquad (4.21)$$

Writing R_i and S_i denote the rank of X_i among $X_1, \ldots, X_n$, and of Y_i among $Y_1, \ldots, Y_n$, respectively, the test statistic becomes (Kallenberg and Ledwina, 1999)

$$S_{(s,t);n} = \sum_{j=1}^{s} \sum_{k=1}^{t} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_j \left(\frac{R_i - \frac{1}{2}}{n} \right) h_k \left(\frac{S_i - \frac{1}{2}}{n} \right) \right)^2 .$$

Now, the first component $(j,k) = (1,1)$ is up to a multiplicative constant the square of Spearman's rank correlation coefficient. More generally, the components $S_{(s,t);n}$ may be interpreted as estimators of the grade correlations between $(F_x(X))^s$ and $(F_y(Y))^t$. The components that are derived in this way are identical to those that are obtained by Eubank et al. (1987), though the latter are based on another family of alternatives

$$f_s(\boldsymbol{x}) = \sum_{j=0}^{s} \sum_{k=0}^{s} \theta_{jk} h_j(F_x(x)) h_k(F_y(y)),$$

with $h_0(.) = 1$. As for the k-sample problem, also this statistic is a rank statistic.

4.3.4 Data-Driven Smooth Tests

In the introduction to the smooth tests it was mentioned that Neyman (1937) showed that his smooth test is asymptotically locally uniformly most powerful symmetric, unbiased and of size α. Although this may look like a desirable property, there is however an important critique which is of practical importance. The optimality property does however only hold when the true distribution F belongs to the family of alternatives which is constructed in the previous sections. Moreover, the test is only consistent when F belongs to the specified family of order s. Only in the limiting case, when $s \to \infty$, the family equals $\mathcal{F}_\infty$, the set of all proper distributions. Indeed, in this case the family of alternatives is strongly related to the orthogonal series estimator of the true df f. This limiting situation may be of theoretical interest, but since in practice the sample size n is always finite, the maximal order s of the alternatives is restricted in some sense by n.

Apart from the problem that a finite order s may imply that the true distribution is not captured by the family, there is the problem of dilution. E.g. suppose that in the 2-sample problem the two distributions F_1 and F_2 have the same form but different means (cfr. the location shift problem). When the smooth test of Section 4.3.3 would be used with $s = 1$, then this test is equivalent to the Kruskal-Wallis test, which is originally especially designed for the location shift problem. Among all GOF tests a high power may be expected here. But, on the other hand, when s is taken considerably larger than 1, then the effect of the location shift becomes "diluted" in the test statistic. More specifically, the second component is Mood's test for comparing variances, which in the present example behaves the same as under the 2-sample null hypothesis and thus does not enlarge the test statistic in probability, but only increases its variance. Therefore, in this example, the power of the smooth test with $s > 1$ is smaller as compared to $s = 1$. In practice, when

one wants to apply an omnibus test, one does not know a priori in what sense the true distribution will deviate from the distribution specified in the null hypothesis, and thus the order s of a smooth test may be very hard to be chosen appropriately, and it definitely holds a certain risk w.r.t. the power.

A solution to the problem of choosing an appropriate order s is given by Ledwina (1994): the order s is estimated from the data, i.e. the order is data-driven. In short the reasoning is as follows. Let $\mathcal{F}_{\Theta_j}$ denote the family of alternatives of order j. Then the set

$$\{\mathcal{F}_{\Theta_1}, \mathcal{F}_{\Theta_2}, \dots\} \tag{4.22}$$

may be considered as a sequence of families. If the true df f belongs to $\mathcal{F}_{\Theta_j}$, then of course also $f \in \mathcal{F}_{\Theta_k}$, $k > j$. From the discussion in the previous paragraph, it is clear that in this situation it is best to choose $s = j$, i.e. the family with the smallest order that contains the true df f. The problem of choosing the appropriate order s may be considered as a model selection problem where s is the dimension of the model. Since the families $\mathcal{F}_{\Theta_j}$ are by construction exponential distributions, an appropriate selection rule is e.g. the Bayesian Information Criterion (BIC) of Schwartz (1978). Once the order is estimated from the data, say the order S is selected, then the original smooth test of order S is applied to the data.

Originally, the data-driven smooth tests were exactly based on Schwartz' selection rule (Ledwina, 1994), which is computationally inconvenient because the maximized likelihoods must be calculated for each candidate dimension s. Later a modified selection rule was proposed (Kallenberg and Ledwina, 1997), which is computationally simpler. All data-driven smooth tests that are described in the remainder of this section will be based on this modified selection rule, though it will still be referred to as Schwartz BIC.

In general the modified selection rule selects the order

$$S = \min\{j; 1 \le j \le d : S_{j;n} - j\ln(n) \ge S_{r;n} - r\ln(n), r = 1, \dots, d\}, \tag{4.23}$$

where d is the maximal order considered. (Note: in most of the literature, the order S selected by the modified selection rule is denoted by $S2$.) Since on the one hand the maximal order d is restricted by the sample size, and on the other hand consistency against essentially any alternative (i.e. omnibus consistency) is only guaranteed when $d \to \infty$, the maximal order is made sample size dependent such that $d = d_n \to \infty$ as $n \to \infty$ in some sense.

Once the order S is selected, the test statistic becomes $S_{S;n}$ (test in Section 4.3.1) or $T_{S;n}$ (test in Section 4.3.2). For all data-driven smooth tests that have been defined similarly, the following theorem has been proven (Inglot et al., 1997; Janic-Wróblewska and Ledwina, 2000; Kallenberg and Ledwina, 1999; Ledwina, 1994).

Theorem 4.2 *Under H_0,*

$$S \xrightarrow{p} 1 \text{ and } S_{S;n} \xrightarrow{d} \chi^2_1.$$

The importance of this property is that the limiting null distribution does not depend on d and F.

In the following sections data-driven tests for the three GOF-problems will be discussed briefly.

The One-Sample Problem

The data-driven Neyman smooth test for simple null hypotheses was introduced by Ledwina (1994), but will not be further discussed here. From a practical point of view, the composite null hypothesis is of more importance. The data-driven smooth test for this problem was given by Inglot et al. (1997). In Section 4.3.3 it was seen that the estimation of a q-dimensional nuisance parameter results in a loss of at least q degrees of freedom in the limiting chi-squared distribution when the smooth tests of Rayner and Best (1989) are used. Inglot et al. (1997), however, use the modified score test of Thomas and Pierce (1979) as the core of their smooth test, which its asymptotic null distribution does not depend on the estimation of the nuisance parameters. This approach has some parallelism with the GQF that eventually led to the RR statistic. For this data-driven smooth test Theorem 4.2 holds. Furthermore, consistency is proven for essentially any alternative.

Simulation studies suggested that the convergence of the null distribution towards its limiting distribution is rather slow. In particular, the selection rule does not always select $S = 1$ under H_0, as is should asymptotically according to Theorem 4.2. Therefore a second-order approximation is developed (Kallenberg and Ledwina, 1995), which gives a good approximation for moderate sample sizes.

Inglot et al. (1997) report that their data-driven smooth test based on the Legendre polynomials has good power characteristics. When testing for normality, the data-driven test is even competitive with tests which are specifically designed for testing for normality, e.g. the Shapiro-Wilk test (Shapiro and Wilk, 1965). Also for testing for exponentiality the same conclusion holds w.r.t. Gini's test for exponentiality (Gail and Gastwirth, 1975).

The k-Sample Problem

In Section 4.3.3 a smooth test for the k-sample problem was described (Eubank et al., 1987). Until today, to our knowledge, there is no data-driven version of this test available for $k > 2$. For $k = 2$, on the other hand, recently Janic-Wróblewska and Ledwina (2000) proposed a data-driven rank statistic. Their statistic is exactly the data-driven version, based on the modified Schwartz selection rule, of the k-sample smooth statistic of Eubank et al. (1987) (Legendre polynomials). Theorem 4.2 is proven as well, and a second-order approximation of the asymptotic null distribution is obtained in a similar way as in Kallenberg and Ledwina (1995), resulting in a good approximation for moderate sample sizes.

Suppose that the samples sizes of the two samples are n_1 and n_2, with $n_1 \leq n_2$, then, if

$n_1 \to \infty$ as $n \to \infty$ and $d_n = o\left(\left(\frac{n_1}{\ln(n_1)}\right)^{1/9}\right)$, then their data-driven test is consistent against essentially any alternative.

Janic-Wróblewska and Ledwina (2000) compared their test in a simulation study ($n_1 = n_2 = 50$) with some other tests for the 2-sample problem (among which the Wilcoxon's rank sum test (Wilcoxon, 1945) and Neuhaus' omnibus test (Neuhaus, 1987)). They concluded that their test has overall good power, also under alternatives where linear rank tests completely break down.

The Independence Problem

The data-driven rank test for the independence problem is a straightforward application of the general methodology presented in the previous sections. Only for the asymmetric family (Equation 4.21) a small adaptation is needed. The two tests are discussed briefly here. Both tests are based on the smooth tests described in Section 4.3.3 with Legendre polynomials.

First, for the symmetric family of alternatives (Equation 4.20) the modified selection may be applied exactly as explained above. This test is referred to as the TS2 test ("T" referring to the test statistic, and "S2" to the modified selection rule).
When the asymmetric family is adopted, the sequence of families of Equation 4.22 must be indexed by two indices: one referring to the order of the Legendre polynomials for the X variable, and the other referring to the order of the polynomials for the Y variable. Thus,

$$\{\mathcal{F}_{\Theta_{11}}, \mathcal{F}_{\Theta_{12}}, \mathcal{F}_{\Theta_{21}}, \mathcal{F}_{\Theta_{22}}, \dots\},$$

where $\mathcal{F}_{\Theta_{st}}$ is the family of alternatives given in Equation 4.21 ($s, t = 1, \dots$). The order is now given by the pair (s, t), but the dimension of the corresponding model is still a scalar, $s + t$, i.e. the total number of parameters. Again the maximal order (d_n, d_n) depends on n. The selection rule now becomes

$$\begin{aligned}(S,T) \quad = \quad & \min\big\{1 \le s,t \le d_n, S_{(s,t);n} - (s+t)\ln(n) \ge S_{(i,j);n} - (i+j)\ln(n); \\ & i,j = 1,\dots,d_n\big\}.\end{aligned}$$

The test statistic is $S_{(S,T);n}$. The test will be referred to as the V test.

Kallenberg and Ledwina (1999) have proven a slightly adapted version of Theorem 4.2, in the sense that now under H_0, $(S,T) \xrightarrow{p} (1,1)$, and consequently $S_{(S,T);n} \xrightarrow{d} \chi^2_1$ whenever d_n converges appropriately to infinity. A simulation study revealed that the convergence is again rather slow. Therefore, a second-order approximation using the arguments of Kallenberg and Ledwina (1995) was obtained, resulting in good approximations. Under certain restrictions on the convergence rate of d_n, both tests were proven to be consistent against essentially any alternative. In an extensive simulation study the TS2 and the V test were compared to some other tests for independence (among them

Hoeffding's test (Blum et al., 1961; Hoeffding, 1948)). They used $d_n = 10$ for $n = 50$ with the TS2 test, and $d_n = 2$ for $n = 50$ with the V test. Their general conclusion is that both tests performed well for a wide range of alternatives. In particular the data-driven tests did never show a complete break down of the power, whereas the other tests did for at least some of the alternatives considered. Under alternatives with lacking symmetry the V tests gave better results than the TS2 test.

Some Final Remarks on Data-Driven Smooth Tests

One practical problem seems still to remain for finite sample sizes: the choice of d_n. Though, e.g. Janic-Wróblewska and Ledwina (2000) investigated the effect of the choice of d_n on the power for moderate sample sizes in a simulation studies, and it turned out that for the alternatives that they considered, the power remained very stable even for small choices for d_n.
On the other hand it is expected that the power of the data-driven tests may be small when d_n is chosen too small when one aims to detect "high-frequency" departures from the null hypothesis (with "high-frequency" it is meant that only the coefficients of the higher order polynomials differ substantially from zero). A possible solution to this problem is to construct an alternative sequence of exponential families as compared to the sequence given in Equation 4.22. E.g. sth order alternatives may be constructed as a subfamily of a tth ($t > s$) order family $\mathcal{F}_{\Theta_t}$ with the $t - s$ first coefficients put to zero. A critique to this approach is that the user must specify the sequence beforehand, whereas typically when facing a omnibus problem the user does not have a clue about the true distribution.

4.4 A Link between Smooth Tests and the Anderson-Darling Statistics

Durbin and Knott (1972) showed an interesting relation between statistics of the Cramér - von Mises type and the smooth tests. In particular they showed that the A_n statistic (Section 4.2.4) for the one-sample problem for simple null hypothesis, may be written in an alternative form

$$A_n = \sum_{j=1}^{\infty} \frac{Z_{nj}^2}{j(j+1)},$$

where

$$Z_{nj} = -\left(\frac{1}{n^{1/2}}\right) \sum_{i=1}^{n} P_j(G(x_i)),$$

where $\{P_j\}$ are the Legendre polynomials on $L_2[0,1]$. Thus, the A_n statistic may now be seen as a weighted Neyman's smooth statistic of infinite order, which is with the present

notation $S_{\infty;n} = \sum_{j=1}^{\infty} Z_{nj}^2$. The weights, $\frac{1}{j(j+1)}$ are such that the higher the order of the term, the lower the weight is.
A similar property of the AD statistic in the composite null hypothesis case is studied by Durbin, Knott and Taylor (1975).

For the 2-sample AD statistic, Pettit (1976) gave a similar expansion of the $A_{n_1 n_2}$ statistic,

$$A_{n_1 n_2} = \frac{n}{n_2} \sum_{j=1}^{\infty} \left(\frac{B_j^2}{j(j+1)} \right),$$

where

$$B_j = -\left(\frac{2j+1}{n_1} \right)^{\frac{1}{2}} \sum_{i=1}^{n} P_j \left(\frac{2i}{n+1} - 1 \right) Z_i,$$

where the Z_i ($i = 1, \ldots, n$) are indicator functions, being one when the ith observation belongs to the 1st sample. Thus, the B_j are just linear rank statistics. Furthermore, $\left(\frac{n+1}{n_2}\right)^{1/2} B_1$ and $\left(\frac{n+1}{n_2}\right)^{1/2} B_2$ are the standardized Wilcoxon and Mood statistic, respectively. Again the statistic may be interpreted as a weighted smooth test for the 2-sample problem (Section 4.3.3).

CHAPTER 5

The Sample Space Partition Test for Goodness-of-Fit

The first type of Sample Space Partition test that is proposed in this work is specifically constructed for the one-sample GOF problem. The directed divergences, which are introduced in Section 5.1), however, will also be needed in the subsequent chapters. In Sections 5.2 and 5.3 the tests are developed for the simple and the composite null hypothesis, respectively. A generalization is given in Section 5.4. In Section 5.5 a data-driven test is constructed, and finally, in Section 5.6 an extension to multivariate observations is discussed.

First we repeat the assumptions that were given in Chapter 3.

Assumption (A1).
The CDF of $\boldsymbol{X}$ is differentiable, and the corresponding df is stictly positive over the sample space.

Assumption (A2).
The distributions F and G are defined on the same sample space $\mathcal{S}$.

Later Assumption A2 will be loosened (Section 5.1.2).

5.1 Directed Divergence

In Section 4.2.2 a broad class of EDF GOF tests was fitted into a framework which is essentially characterized by a measure δ for the discrepancy between the true distribution F, estimated by $\hat{F}_n$, and the hypothesized distribution G. This measure must not necessarily be a metric, nor must it even be symmetric in its arguments. One interesting choice for δ is a functional based on a *directed information divergence*. In Section 4.1.5 a related family of power divergence statistics was discussed. These divergences, however, were originally defined for discrete distributions (Cressie and Read, 1984). In this section we will show how this family of divergences can be used for the GOF setting for continuous variables, eventually leading to a new statistical test for GOF.

5.1.1 Directed Divergence for Continuous Distributions

Although most of the discussion in Cressie and Read (1984) is about the divergence for discrete distributions, they give briefly an indication on how a continuous analogue could be constructed. Here, however, we prefer to follow another line of thinking in which the divergence between continuous distributions is defined starting from the definition of the divergence between discrete distributions. To make the link between both more direct we will state the results for the latter divergence for discretized continuous distributions, which, as in Section 4.1.4, imply multinomial distributions.

Let

$$[A] = \{A_1, \ldots, A_c\}$$

denote a partition of size c of the sample space $\mathcal{S}$. As before, F and G denote the true and the postulated distribution function of X, respectively, for which Assumptions A1 and A2 are supposed to hold. Then, the fixed sample space partition (SSP) $[A]$ implies on both F and G a multinomial distribution with probabilities $\mathrm{P}_f[A_i] = \int_{A_i} dF$ and $\mathrm{P}_g[A_i] = \int_{A_i} dG$ $(i = 1, \ldots, c)$, respectively. This is illustrated in Figure 5.1.

Definition 5.1 *The directed divergence of order $\lambda \in \mathbb{R}$ between F and G, based on the size c SSP $[A] = \{A_1, \ldots, A_c\}$ for which for all $i = 1, \ldots, c : P_g[A_i] \neq 0$, is defined as*

$$I^{\lambda}(F; G\|[A]) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{c} P_f[A_i] \left[\left(\frac{P_f[A_i]}{P_g[A_i]} \right)^{\lambda} - 1 \right], \tag{5.1}$$

where the function at $\lambda = 0, -1$ is defined by continuity.

This definition is a straightforward adaptation of the definition of the divergence between two discrete (multinomial) distributions (Cressie and Read, 1984). Sometimes, for short, this divergence will be referred to as the *discrete divergence*.

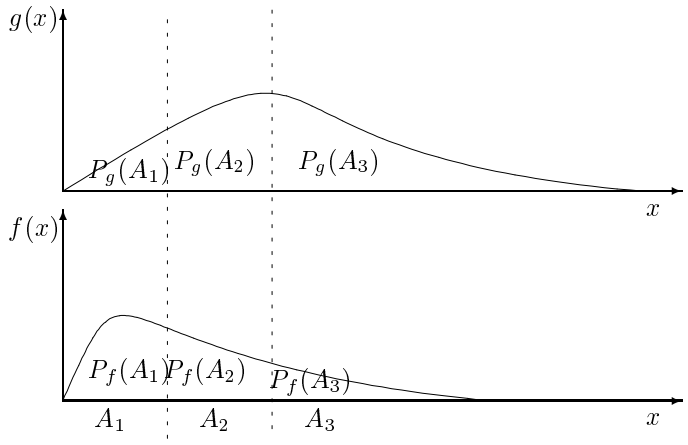


Figure 5.1: A partition $[A] = \{A_1, A_2, A_3\}$ of a sample space of a univariate variable, and the probabilities w.r.t. the true (f) and the postulated (g) distributions.

The condition that $\mathrm{P}_g[A] \neq 0$ for all $A \in [A]$ is assured by Assumption A1 that g is strictly possitive on $\mathcal{S}$.

A few important notes are here in place.

1. First, for $\lambda = 0$ the directed divergence becomes

$$\mathrm{I}^0(F; G\|[A]) = \sum_{i=1}^{c} \mathrm{P}_f[A_i] \ln \frac{\mathrm{P}_f[A_i]}{\mathrm{P}_g[A_i]},$$

   which is a constant multiple of Kullback's information divergence (Kullback, 1959). Another interesting special member of the family is given for $\lambda = 1$,

$$\mathrm{I}^1(F; G\|[A]) = \frac{1}{2} \sum_{i=1}^{c} \frac{(\mathrm{P}_f[A_i] - \mathrm{P}_g[A_i])^2}{\mathrm{P}_g[A_i]},$$

   which is clearly a Pearson-like functional.

2. A second important remark is concerning the asymmetry of the divergence in its arguments, which makes the divergence definitely not a metric. Only for $\lambda = \frac{1}{2}$ the divergence becomes a metric, which is known as the Hellinger or Matusita distance.

Cressie and Read (1984) showed that $\mathrm{I}^\lambda(F; G\|[A])$ is indeed a true information measure according to the criteria of Rathie and Kannappan (1972). They also showed that it satisfies some important properties, of which the most important to the present text are summarized in the following Corollary.

Corollary 5.1 *(Cressie and Read, 1984)*

- *The directed divergence based on SSP* $[A]$ *is* **non-negative**, *i.e.* $I^{\lambda}\left(F;G\|[A]\right) \geq 0$, *with equality if and only if* $P_f\left[A_i\right] = P_g\left[A_i\right]$, $i = 1, \ldots, c$.
- *The directed divergence based on SSP* $[A]$ *satisfies a mild* **symmetry condition** *implying that the divergence is unchanged if the* c *partition elements are taken in any different order.*
- *The directed divergence based on SSP* $[A]$ *satisfies the following* **grouping property***: for any partition* $[A] = \{A_1, \ldots, A_c\}$ *and any derived size* $c-1$ *partition* $[B] = \{A_1, \ldots, A_{c-1} \cup A_c\}$, $I^{\lambda}\left(F;G\|[B]\right) \leq I^{\lambda}\left(F;G\|[A]\right)$, *i.e. the divergence cannot increase if any classes are grouped. The partition* $[A]$ *is sometimes called a* refinement *of partition* $[B]$, *and this property is also known as the* **Refinement Lemma** *(Whittaker, 1990).*

The directed divergence between two continuous distributions is now defined.

Definition 5.2 *The directed divergence of order* $\lambda \in \mathbb{R}$ *between* F *and* G *is defined as*

$$I^{\lambda}\left(F;G\right) = \sup_{[A]} I^{\lambda}\left(F;G\|[A]\right), \tag{5.2}$$

where the supremum is taken over all possible sample space partitions of $\mathcal{S}$.

This divergence will sometimes be referred to as the *continuous divergence*.
The existance of the supremum follows from the extended Dobrushin theorem (cfr. Whittaker, 1990), which states that

$$\sup_{[A]} \mathrm{I}^{\lambda}\left(F;G\|[A]\right) = \frac{1}{\lambda(\lambda+1)} \int_{\mathcal{S}} \left[\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \right)^{\lambda} - 1 \right] f(\boldsymbol{x}) d\boldsymbol{x}. \tag{5.3}$$

The right hand side of Equation 5.3 may be taken as the practical definition of the directed divergence of order λ, $\mathrm{I}^{\lambda}\left(F;G\right)$. Intuitively, Equation 5.3 may be understood as follows: from the Refinement Lemma (Corollary 5.1) it is known that the directed divergence increases as the partition becomes more refined. Thus, the supremum can be obtained by taking the limit for infinitesimal small partition elements, which eventually leads to the integral.

Defining the divergence between continuous distributions as above, has in general two important advantages.

- The first important advantage is that the original definition, which actually says that the divergence between continuous distributions is a special case (limiting case) of

the divergence between discrete distributions, directly implies that all properties for the discrete divergence are still valid for the continuous divergence. More specifically, the first part of Corollary 5.1 still applies, but can now be formulated as

Corollary 5.2 *The directed divergence is* **non-negative**, *i.e.* $I^{\lambda}(F;G) \geq 0$ *with equality if and only if* $f(\boldsymbol{x}) = g(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \mathcal{S}$.

- Finally, since the continuous divergence is according to its definition a discrete divergence w.r.t the most extreme refined partition, the Refinement Lemma now implies immediately the following corollary.

  Corollary 5.3 *For all partitions* $[A]$ *of any size* $c > 1$,

  $$I^{\lambda}(F;G\|[A]) \leq I^{\lambda}(F;G)\,.$$

5.1.2 Directed Divergence Based on SSP in the GOF Setting

Corollary 5.2 actually says that the continuous divergence between F and G is zero if and only if the GOF null hypothesis is true, otherwise the divergence is positive. And Corollary 5.3 states that the continuous divergence is at least as large as the divergence based on a SSP. Thus, for some a priori specified λ, it is sufficient to find only one partition $[A]$ for which $\mathrm{I}^{\lambda}(F;G\|[A]) > 0$ to conclude that G is not the true distribution of $\boldsymbol{X}$. We summarize this straightforward result in a Corollary for it is a central argument in the construction of the proposed statistical test in this thesis.

Corollary 5.4 *If, for some* λ, *there exists a SSP* $[A]$ *of any size* $c > 1$ *for which* $I^{\lambda}(F;G\|[A]) > 0$, *then* G *and* F *are different.*

In the previous section is was assumed that both f and g are defined on a common sample space (Assumption A2) on which they both are strictly positive (Assumption A1). This assumption was an important element to guarantee that the discrete and the continuous divergences exist (division by zero was avoided). A few comments on this assumption in the GOF setting may be in place here.

- The discrete and the continuous divergence are often seen as expectations w.r.t. P_f and f, respectively. Obviously, the summation and integration in their definitions are over the partition elements of the sample space of f, and the sample space of f, respectively. In a typical GOF setting, however, one does not know a priori the true distribution f, and, therefore, one also often does not know exactly the sample space on which f is defined. This is clearly a problem. The distribution g, on the other hand, is completely specified (or at least up to some estimable parameters), and its sample space is known as well.

- Suppose that the sample space f is a subset of the sample space of g, then $\mathrm{I}^{\lambda}(F;G)$ would remain unchanged if the integral is calculated over the sample space of g, and also $\mathrm{I}^{\lambda}(F;G\|[A])$ would remain unchanged by constructing the partition $[A]$ on the larger sample space. (For $\lambda = 0$, this is guaranteed by defining $0 \ln 0 = 0$ as in Whittaker (1990).)

- Suppose that it would be allowed that g is zero on some subset of the sample space where f is strictly positive. Then $\mathrm{I}^{\lambda}(F;G) = +\infty$ and SSPs $[A]$ may be constructed for which $\mathrm{I}^{\lambda}(F;G\|[A]) = +\infty$ as well. The infinitely large values of the divergences only reflects (correcly) the fact that the sample spaces do not coincide and that therefore the distributions f and g are definitely different. Thus, from the point of view of the GOF problem, this situation has a clear interpretation.

Although the last remark does not seem to introduce any problems in a GOF setting, we will not allow this situation since it will invalidate the sample distributions of the statistics that are constructed in the next sections. The situation described in the second remark, on the other hand, will not change the results that are given in the next sections. Therefore, in the remainder of this chapter, we will loosen the assumption on the common sample space:

Assumption (A2b).
The sample space on which F is defined must be equal to the sample space on which G is defined, or at least it must be a subset of the latter.

In the remainder of this chapter the sample space $\mathcal{S}$ refers to the extended sample space on which g is defined.

5.1.3 The Power Divergence Statistics

The major reason why the information divergences, as formulated in the previous section, cannot be used directly in practice is of course that they generally cannot be calculated, because the true distribution F is not known, otherwise the GOF-problem would not exist in the first place. Moreover, in the case of a composite GOF null hypothesis, the exact member G_{θ_0} is not known. Therefore a sample-analogue to both the discrete and the continuous divergence can be looked for as the corresponding estimator, which subsequently might be used as a test statistic, taking the sampling variability into account.

The most straightforward solution is to replace the unknown distributions F and G (if the null hypothesis is composite) by their corresponding estimators. As an estimator of F we propose to use the empirical distribution function (EDF) $\hat{F} = \hat{F}_n$, which was already introduced in Section 4.2.1 as a consistent estimator of F. When G_θ belongs to some parametric family, there is also need for an estimator. As frequently used in Chapter 3, this estimator is given by $\hat{G} = G_{\hat{\theta}}$, where $\hat{\theta}$ denotes some consistent estimator of $\boldsymbol{\theta}$.

The divergences which are defined in Section 5.1, are basically real-valued functions of distributions. Such functions are often called *functionals*. By replacing the distributions

by their estimators, as described in the previous paragraph, a so-called *plug-in* estimator is obtained. In this way the functional with estimators plugged-in for the unknown distributions becomes a function of the sample, and thus it becomes a statistic.

Discrete Divergence

The statistic based on the discrete directed divergence was the main topic of the Cressie and Read (1984) paper. These statistics were treated in detail in Section 4.1.5; we will only repeat the main results here, applied to the grouped (by means of SSP) continuous data setting.

Definition 5.3 *The power divergence statistic of order $\lambda \in \mathbb{R}$ between F and G, based on the size c SSP $[A]$ is defined as*

$$\hat{I}^{\lambda}\left(F;G\|[A]\right)=\frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{c}\hat{P}_f\left[A_i\right]\left[\left(\frac{\hat{P}_f\left[A_i\right]}{\hat{P}_g\left[A_i\right]}\right)^{\lambda}-1\right],$$

where $\hat{P}_f\left[A_i\right]=n^{-1}\#(A_i\cap\mathcal{S}_n)$, and $\hat{P}_g\left[A_i\right]=P_g\left[A_i\right]$ for a simple null hypothesis and $\hat{P}_g\left[A_i\right]=P_{g_{\hat{\theta}}}\left[A_i\right]$ for a composite null hypothesis.

Note that $\hat{\mathrm{I}}^{\lambda}\left(F;G\|[A]\right)$ could just as well be written as $\mathrm{I}^{\lambda}\left(\hat{F}_n;G_{\hat{\theta}}\|[A]\right)$. Note also that by Assumptions A1 and A2b the estimated probabilities $\mathrm{P}_g\left[A_i\right]$ $(i=1,\ldots,c)$ are always greater than zero.

It is shown that for all λ, conditional of $[A]$,

$$2n\hat{\mathrm{I}}^{\lambda}\left(F;G\|[A]\right)-2n\hat{\mathrm{I}}^{1}\left(F;G\|[A]\right)\xrightarrow{p}0,$$

which is Pearson's χ^2-statistic. Further, under a simple H_0, for all λ,

$$2n\hat{\mathrm{I}}^{\lambda}\left(F;G\|[A]\right)\xrightarrow{d}\chi^2_{c-1}.$$

For a composite null hypothesis, and p-vector parameter $\boldsymbol{\theta}$ replaced by its BAN-estimator $\hat{\boldsymbol{\theta}}$ based on the grouped data (see Section 4.1.3 for more details on BAN-estimators), the asymptotic null distribution is given as

$$2n\hat{\mathrm{I}}^{\lambda}\left(F;G\|[A]\right)\xrightarrow{d}\chi^2_{c-p-1}.$$

In the beginning of this chapter we suggested that a statistic based on an information divergence may be a good choice for the measure δ within Romano's framework. One such possibility is the discrete power divergence statistic.

Continuous Divergence

Also in the continuous directed divergence $\mathrm{I}^{\lambda}(F;G)$ the unknown distributions can be replaced by estimators, resulting in a divergence statistic. When the distributions are continuous, then the use of the plug-in estimator seems the most natural choice to use in a GOF test based on the information divergence. Indeed the use of the discrete power divergence statistics described above only seems a generalization of the Pearson χ^2-test for grouped data for which in Section 4.1.4 it was explained that this is definitely not an optimal choice. An important difference between the continuous and the discrete divergence statistic is that for the continuous case not the CDF F, which is basically a probability, must be replaced by its estimator, but rather that the density f must be replaced. A good choice is a kernel estimator $\hat{f}_{h;n}$ with bandwidth h. In the case $\lambda = 1$ (Pearson's functional) Bickel and Rosenblatt (1973) showed that such a divergence statistic has an asymptotic normal null distribution. More recently, similar statistics (Rosenblatt and Wahlen, 1992; Zheng, 1997) were studied in the context of testing independence between two continuous random variables. The major problem with this approach is that it necessarily brings the problem of bandwidth selection along; see e.g. Hall (1987) for a detailed discussion. We will not continue in this line.

At this point it may be interesting to remark that smooth tests may also be constructed from Pearson's functional (Eubank et al., 1987). In Section 4.3.2 it was shown that the corresponding smooth statistic is obtained by substituting f with an s-dimensional exponential family. In this way the densities must not be estimated by means of a kernel density estimator, but rather by replacing the s parameters by their maximum likelihood estimators. Omnibus consistency is however lost, but may be obtained again by making the test data-driven, and allowing s to grow unboundedly with n as $n \to \infty$.

5.2 The SSP Test: Simple Null Hypothesis

In this section a new SSP-based GOF test is constructed for the easiest case: the simple null hypothesis $H_0 : F(x) = G(x)$ for all $x \in \mathcal{S}$, where G is thus a uniquely specified CDF, and where x is univariate. Further, we will restrict the power divergences to $\lambda = 1$, i.e. the Pearson-type divergence, which will be referred to as the *(discrete) Pearson divergence*. The corresponding statistic is the (discrete) Pearson divergence statistic, or the traditional Pearson statistic. Later the λ-restriction will be dropped and the test proposed in this section will be generalized.

First, in Section 5.2.1, an appropriate δ-measure is proposed. Its plug-in estimator, which will be used to construct the test statistic, is presented in Section 5.2.2. The next section is devoted to a discussion on the test statistic and its limiting null distribution, followed by a study about approximation methods for both the asymptotic and the exact (finite sample) null distribution. We end with a small simulation study to show the new test its power characteristics.

5.2.1 The Averaged Pearson Divergence

Starting from Corollary 5.4 one could think of a GOF test by constructing some algorithm to search a SSP $[A]$ for which

$$\mathrm{I}^1\left(F; G \| [A]\right) > 0. \tag{5.4}$$

For each SSP $[A]$ the inequality can be assessed by means of a statistical α-level test based on the Pearson statistic $2n\hat{\mathrm{I}}^1\left(F; G \| [A]\right)$. Once such a SSP is found, by Corollary 5.4 it is concluded that the null hypothesis is rejected. This approach introduces however some difficulties. Since the decision rule is based on performing several tests, this methodology clearly introduces multiplicity. Hence, the overall probability of making a type I error will be larger than α. Furthermore, in general, the partitions for which Equation 5.4 is to be assessed will lead to Pearson test statistics that may be dependent, which makes a correction for multiplicity possibly rather difficult (see e.g. Hsu, 1996). We want to note here, however, that this point of criticism does not mean that such a method is not feasible, but in this thesis we will not continue in that line.

One way to overcome the multiple testing problem is of course to construct a test (decision rule) which is based on only a single test statistic. This test statistic will again be an estimator of some measure for the deviation from the null hypothesis, but now this measure will combine the information against the null hypothesis among a *representative* set of SSPs. In the remainder of this section this measure is introduced, but first some definitions are given.

Definition 5.4 *Consider $\mathcal{P}_{q,n} \subset \mathcal{S}_n$, where $q = \#\mathcal{P}_{q,n}$. Then, a* **partition construction rule** *is an algorithm that uniquely determines a SSP $[A]$ as a function of $\mathcal{P}_{q,n}$, i.e. $[A](\mathcal{P}_{q,n})$. Further, define $\mathbb{P}_{q,n}$ as the set of all q-sized subsamples $\mathcal{P}_{q,n}$, and $\mathcal{A}_{q,n}$ as the set of all corresponding SSPs. Then, $\mathbb{P}_{q,n}$ is a* **partition determining subsample set**.

Definition 5.4 serves as the general definition, and it allows for different types of partitions. In the present chapter, however, X is supposed to be univariate, such that it is straightforward to find a simple and appropriate construction rule and a corresponding partition determining subsample set. The following partition construction rule is the one that is used in the remainder of this chapter. The sample $\mathcal{S}_n$ may be written as $\mathcal{S}_n = \{X_{(1)}, \ldots X_{(n)}\}$, where $X_{(i)}$ denotes the ith order statistic. Let b_l and b_u denote the lower and upper bound of the sample space $\mathcal{S}$, respectively.

Definition 5.5 *The* **partition construction rule** *of a c-sized SSP of the sample space $\mathcal{S}$ of a univariate rv determines partitions of the form*

$$[A] = \{]b_l, X_{(i_1)}] \cap \mathcal{S},]X_{(i_1)}, X_{(i_2)}] \cap \mathcal{S}, \ldots]X_{(i_{c-1})}, b_u] \cap \mathcal{S}\}, \tag{5.5}$$

for any $1 \leq i_1 < i_2 < \ldots < i_{c-1} \leq n$.

Hence, $\mathcal{P}_{q,n} = \{X_{(i_1)}, \ldots, X_{(i_{c-1})}\}$, $q = c-1$, and $\mathbb{P}_{c-1,n}$ is the set of all such $\mathcal{P}_{c-1,n}$. Note that $\#\mathbb{P}_{c-1,n} = \binom{n}{c-1}$. Also, for short, $[A]_{\mathcal{P}} = [A](\mathcal{P}_{q,n})$, and $\mathcal{A} = \mathcal{A}_{q,n}$ will be used. Further $a_{q,n} = \#\mathcal{A} = \#\mathbb{P}_{c-1,n}$. Let $\mathcal{S}_{\mathcal{P}}$ denote the sample space of $\mathcal{P} = \mathcal{P}_{c-1,n}$, i.e. $\mathcal{S}_{\mathcal{P}} = \{(x_1, \ldots, x_{c-1}) \in \mathcal{S}^{c-1} : \text{ for all } i \neq j (i, j = 1, \ldots, c-1), x_i \neq x_j\}$.

Consider any random partition $[A]_{\mathcal{P}}$, where $\mathcal{P} \in \mathbb{P}_{c-1,n}$. Then also the discrete divergence $\mathrm{I}^1(F; G \| [A]_{\mathcal{P}})$ is a random variable, whose distribution is determined by $\mathcal{P}$, which in turn is determined by the distribution F of its $c-1$ components. Averaging these divergences over all SSPs in $\mathcal{A}$ gives the new measure proposed here.

Definition 5.6 *The* **averaged Pearson divergence** *is defined as*

$$\Delta_c(F; G) = E_f\left[I^1(F; G \| [A](\mathcal{P}_{c-1,n}))\right]. \tag{5.6}$$

This measure $\Delta_c(F; G)$ is of course only useful if it measures the deviation of F from G in some sense. This is shown in the following lemma.

Lemma 5.1 *For all* $c > 1$

$$\Delta_c(F; G) = 0 \Leftrightarrow H_0 \textit{ is true}.$$

Proof. If H_0 is true, then, following Corollary 5.1, for all $[A]$ $\mathrm{I}^1(F; G \| [A]) = 0$, and hence $\Delta_c(F; G) = 0$.

If $\Delta_c(F; G) = 0$, i.e.

$$\frac{1}{2}\int_{\mathcal{S}_{\mathcal{P}}} \sum_{i=1}^{c} \frac{\left(\mathrm{P}_f[A_i(\mathcal{P})] - \mathrm{P}_g[A_i(\mathcal{P})]\right)^2}{\mathrm{P}_g[A_i(\mathcal{P})]} dF_{\mathcal{P}_{c-1}}(\mathcal{P}) = 0,$$

where $F_{\mathcal{P}_{c-1}}(\mathcal{P}) = \prod_{i=1}^{c-1} F(x_i)$. Then,

$$\mathrm{P}_f[A_i(\mathcal{P})] = \mathrm{P}_g[A_i(\mathcal{P})] \tag{5.7}$$

almost everywhere. Since $\mathrm{P}_f[.]$ and $\mathrm{P}_g[.]$ are differences of the form $F(b) - F(a)$ and $G(b) - G(a)$, respectively, Equation 5.7 is equivalent to $F(x) = G(x)$ almost everywhere. By the continuity of F and G this means that $F(x) = G(x)$ for all $x \in \mathcal{S}$. □

Thus, $\Delta_c(F; G)$ seems a valid choice for the δ-measure. In the next section its plug-in estimator is discussed. It will be used later to construct the test statistic.

5.2.2 The Averaged Pearson Divergence Statistic

Since in Section 5.1.3 an estimator of $\mathrm{I}^1\left(F;G\|[A]\right)$ was given, and since $\Delta_c(F;G)$ is only an expectation of such divergences, an estimator for $\Delta_c(F;G)$ is easily found as

$$\begin{aligned}\hat{\Delta}_c(F;G) &= \Delta_c(\hat{F};G)\\ &= \mathrm{E}_{\hat{F}}\left[\hat{\mathrm{I}}^1\left(F;G\|[A]_{\mathcal{P}}\right)\right]\\ &= \int_{\mathcal{S}_{\mathcal{P}}} \hat{\mathrm{I}}^1\left(F;G\|[A]_{\mathcal{P}}\right) d\hat{F}_{\mathcal{P}}(\mathcal{P}),\end{aligned}$$

where $F_{\mathcal{P}}(\mathcal{P})$ is further simplified by the assumption that all $c-1$ observations in $\mathcal{P}$ are i.i.d., but are restricted to be different. Thus,

$$\begin{aligned}\hat{\Delta}_c(F;G) &= \int_{\mathcal{S}}\dots\int_{\mathcal{S}\,(x_i\neq x_j, i\neq j=1,\dots,c-1)} \hat{\mathrm{I}}^1\left(F;G\|[A]_{\mathcal{P}}\right)\prod_{i=1}^{c-1} d\hat{F}(x_i|\boldsymbol{h}_i)\\ &= \frac{1}{\binom{n}{c-1}}\sum_{\mathcal{P}\in\mathbb{P}_{c-1,n}} \hat{\mathrm{I}}^1\left(F;G\|[A]_{\mathcal{P}}\right), \qquad (5.8)\end{aligned}$$

where $\boldsymbol{h}_i = (x_1,\dots,x_{i-1})$ ($\boldsymbol{h}_1 \equiv 0$), and where $\mathcal{P} = \{x_1,\dots,x_{c-1}\} \subset \mathcal{S}$. $\hat{\Delta}$ will be referred to as the *Averaged Pearson Divergence Statistic*, or simply the *Averaged Pearson Statistic*.

5.2.3 The Test Statistic and its Asymptotic Null Distribution

As a test statistic we propose

$$T_{c,n} = 2n\hat{\Delta}_c(F;G) = 2n\mathrm{E}_{\hat{F}}\left[\mathrm{I}^1\left(\hat{F};G\|[A](\mathcal{P}_{c-1,n}\right)\right]. \qquad (5.9)$$

An important result for the construction of the statistical test is the limiting null distribution of the test statistic. First a special case ($c=2$) is considered, next the general result is given.

Special Case: $c = 2$

If $c=2$, then the partition determining subsample set $\mathbb{P}_{1;n}$ is just the sample $\mathcal{S}_n$, and the sets $\mathcal{P}^i_{1;n}$ may be replaced by the corresponding observations X_i ($i=1,\dots,n$). Thus, the partitions $[A](\mathcal{P}^i_{1;n})$ may be replaced by $[A](X_i)$ or simply $[A]_i$. For a given sample $a_{1;n} = n$ such partitions can be constructed. Each SSP $[A]_i$ consists of two elements, denoted by $A_{i1} =]b_l, X_i]$ and $A_{i2} =]X_i, b_u]$.

The test statistic is then given by

$$\begin{aligned} T_{2,n} &= 2n\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\mathrm{I}}^1\left(F;G\|[A]_i\right)\right] \\ &= 2n\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(\frac{(\hat{\mathrm{P}}_f\left[A_{i1}\right]-\mathrm{P}_g\left[A_{i1}\right])^2}{\mathrm{P}_g\left[A_{i1}\right]}+\frac{(\hat{\mathrm{P}}_f\left[A_{i2}\right]-\mathrm{P}_g\left[A_{i2}\right])^2}{\mathrm{P}_g\left[A_{i2}\right]}\right)\right] \\ &= n\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(\hat{F}(X_i)-G(X_i))^2}{G(X_i)}+\frac{\left((1-\hat{F}(X_i))-(1-G(X_i))\right)^2}{1-G(X_i)}\right) \\ &= n\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{F}(X_i)-G(X_i))^2}{G(X_i)(1-G(X_i))}. \end{aligned} \tag{5.10}$$

It is interesting to note here that there exists a very related statistic. Suppose that in Section 5.2.1 the Average Pearson Divergence would have been defined as

$$\Delta_{c,a}(F;G)=\mathrm{E}_g\left[\mathrm{I}^1\left(F;G\|[A](\mathcal{P}_{c-1,n})\right)\right],$$

i.e. the expectation is w.r.t. the hypothesised distribution G under H_0 instead of the true distribution F. Then, the corresponding statistic $2n\hat{\Delta}_{c,a}$ would have been

$$2n\hat{\Delta}_{c,a}(F;G)=2n\Delta_{c,a}(\hat{F};G)=2n\int_{\mathcal{S}_{\mathcal{P}}}\hat{\mathrm{I}}^1\left(F;G\|[A]_{\mathcal{P}}\right)dG_{\mathcal{P}}(\mathcal{P}). \tag{5.11}$$

For the special case $c=2$ the test statistic $2n\hat{\Delta}_{2,a}$ based on this alternative definition reduces to

$$A_n=n\int_{\mathcal{S}}\frac{(\hat{F}_n(x)-G(x))^2}{G(x)(1-G(x))}dG(x),$$

which is identical to the Anderson-Darling (AD) test statistic (Anderson and Darling, 1952) which was discussed in Section 4.2.4. The null distribution of $T_{2,n}$ may readily be found by proving that under H_0 it is asymptotically equivalent to A_n, for which Anderson and Darling (1952) presented the asymptotic null distribution.

Theorem 5.1 *Under the simple GOF null hypothesis, for* $c=2$*,*

$$T_{2,n}\xrightarrow{d}\sum_{j=1}^{\infty}\frac{1}{j(j+1)}Z_j^2,$$

where the Z_j^2 *are i.i.d.* χ_1^2*-variates.*

Proof. Without loss of generallity we will suppose that G is the CDF of a uniform

distribution on $[0,1]$ (i.e. the integral probability transformation is applied). Let $U_n(t) = \sqrt{n}(\hat{F}_n(t) - t)$. The test statistic may now be written as (McCabe and Tremayne, 1993; Shorack and Wellner, 1986, cfr. Ito stochastic integral)

$$T_{2,n} = \int_0^1 \frac{U_n^2(t)}{t(1-t)} d\hat{F}_n(t). \tag{5.12}$$

The proof consists in showing that this stochastic integral converges in distribution to

$$A = \int_0^1 \frac{U^2(t)}{t(1-t)} dt,$$

where $U(t)$ is a Brownian bridge. The distribution of A is exactly the limiting null distribution of the Anderson-Darling statistic (Anderson and Darling, 1952).

We provide here a sketch of the proof along the same line as e.g. Pettit (1976), Scholz and Stephens (1987). For processes $X(t)$ and $Y(t)$, let

$$||X - Y|| = \sup_{0 \le t \le 1} |X(t) - Y(t)|$$

denote the supremum metric.
According to the Pyke-Shorack theorem (Shorack, 2000, Theorem 10.1), $||\frac{U_n - U}{\sqrt{t(1-t)}}|| \xrightarrow{p} 0$. Also $||\hat{F}_n - t|| \xrightarrow{p} 0$. Since $\frac{U_n^2(t)}{t(1-t)}$ and $\hat{F}_n(t)$ are independent, by Donsker's theorem and the Continuous Mapping Theorem, for any $0 < \delta < 1$,

$$\int_\delta^{1-\delta} \frac{U_n^2(t)}{t(1-t)} d\hat{F}_n(t) \xrightarrow{d} \int_\delta^{1-\delta} \frac{U^2(t)}{t(1-t)} dt$$

(this is the "central" portion of the integral in Equation 5.12). The proof is completed if we can show that the remainder of the integral in Equation 5.12 is negligible. As in Scholz and Stephens (1987) this may be accomplished by invoking Theorem 4.2 of Billingsley (1968) and Markov's inequality. □

In Section 4.2.4 it was mentioned that the AD statistic belongs to the more widely defined family

$$n \int_{\mathcal{S}} \left(\hat{F}(x) - G(x)\right)^2 \Psi(G(x)) dG(x), \tag{5.13}$$

which is known as the Anderson-Darling family. We show now that the $T_{2,n}$ statistic is also a member of this family.

Lemma 5.2 *The statistic $T_{2,n}$ is a statistic of the Anderson-Darling Family.*

Proof. First note that $T_{2,n}$ has asymptotically the same distribution as (cfr. proof of

Theorem 5.1)

$$2n \int_{\mathcal{S}} \hat{\mathrm{I}}^1 \left(F; G \| [A](x)\right) dF(x).$$

The statistic $T_{2,n}$ has the same limiting distribution as

$$\begin{aligned} & n \int_{\mathcal{S}} \frac{\left(\hat{F}(x) - G(x)\right)^2}{G(x)(1-G(x))} dF(x) \\ = \quad & n \int_{\mathcal{S}} \frac{\left(\hat{F}(x) - G(x)\right)^2}{G(x)(1-G(x))} f(x) dx \\ = \quad & n \int_{\mathcal{S}} \frac{\left(\hat{F}(x) - G(x)\right)^2}{G(x)(1-G(x))} \frac{f(x)}{g(x)} dG(x) \\ = \quad & n \int_{\mathcal{S}} \left(\hat{F}(x) - G(x)\right)^2 \Psi(G(x)) dG(x), \end{aligned}$$

which is a statistic of the Anderson-Darling family with $\Psi(G(x)) = \frac{f(x)}{g(x)} \frac{1}{G(x)(1-G(x))}$. □

From the Anderson-Darling weight function that is associated with the $T_{2,n}$ statistic, one may get some insight in the behaviour of the statistic. In comparisons made between the AD ($\Psi(G(x)) = \frac{1}{G(x)(1-G(x))}$ and the Cramér - von Mises statistic ($\Psi = 1$) it is often argued that the AD statistic is more sensitive to deviations in de tail of the distribution G for its weight function is large in the tails. For the $T_{2,n}$ statistic this argument remains valid; two new features come to it. First, in the tails $g(x)$ is small as well, resulting in an additional increase in the weight function. Second, the weight function is corrected by a factor $f(x)$, being large in regions where many observations are expected, and small in regions where only a few observations are expected. In particular, the latter region corresponds to the tails of the true distribution. Thus, globally, $\frac{f(x)}{g(x)}$ serves as a correction factor, upweighing in regions of $\mathcal{S}$ where more "information" is expected under F than under G.

General Case

The asymptotic null distribution for general $c > 1$ is more complicated than in the $c = 2$ case, and will be stated here only as a proposition with an heuristic proof. In a later section it will be empirically shown that null distribution of $T_{c,n}$ indeed seems to converge to the distribution that is proposed here.

Before we give the Proposition, first the test statistic $T_{c,n}$ is written here explicetly in its computational form. The elements of the partition $[A](\mathcal{P})$ are denoted by $A_i(\mathcal{P})$ ($i =$

$1, \ldots, c$) (cfr. Equation 5.5).

$$
\begin{aligned}
T_{c,n} &= 2n \frac{1}{\binom{n}{c-1}} \sum_{\mathcal{P} \in \mathbb{P}_{c-1,n}} \hat{\mathrm{I}}^1 \left(F; G \| [A]_{\mathcal{P}}\right) \\
&= \frac{n}{\binom{n}{c-1}} \sum_{\mathcal{P} \in \mathbb{P}_{c-1,n}} \sum_{i=1}^{c} \frac{\left(\hat{\mathrm{P}}_f \left[A_i(\mathcal{P})\right] - \mathrm{P}_g \left[A_i(\mathcal{P})\right]\right)^2}{\mathrm{P}_g \left[A_i(\mathcal{P})\right]} \\
&= \frac{n}{\binom{n}{c-1}} \sum_{1 \le i_1 < \ldots < i_{c-1} \le n} \left[\frac{\left(\hat{F}(X_{(i_1)}) - G(X_{(i_1)})\right)^2}{G(X_{(i_1)})} + \right. \\
&\quad \frac{\left(\left(\hat{F}(X_{(i_2)}) - \hat{F}(X_{(i_1)})\right) - \left(G(X_{(i_2)}) - G(X_{(i_1)})\right)\right)^2}{\left(G(X_{(i_2)}) - G(X_{(i_1)})\right)} + \ldots \\
&\quad \frac{\left(\left(\hat{F}(X_{(i_{c-1})}) - \hat{F}(X_{(i_{c-2})})\right) - \left(G(X_{(i_{c-1})}) - G(X_{(i_{c-2})})\right)\right)^2}{\left(G(X_{(i_{c-1})}) - G(X_{(i_{c-2})})\right)} + \\
&\quad \left. \frac{\left(\left(1 - \hat{F}(X_{(i_{c-1})})\right) - \left(1 - G(X_{(i_{c-1})})\right)\right)^2}{\left(1 - G(X_{(i_{c-1})})\right)} \right]
\end{aligned}
$$

Proposition 5.1 *For any $c > 1$, under the simple GOF null hypothesis,*

$$T_{c,n} \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2,$$

where the Z_j^2 are i.i.d. χ^2_{c-1} variates.

Heuristic Proof. The reasoning is based on complete induction. First, suppose the proposition holds for $T_{c-1,n}$ ($c - 1 > 1$), then it will be shown that it also holds for $T_{c,n}$.

Any c-sized partition can be written as (cfr. Equation 5.5)

$$[A]_{\mathcal{P}} = [A_s^{c-1}]_{\mathcal{P}} \cup \{]x_{(i_{c-1})}, b_u]\},$$

where $[A_s^{c-1}]_{\mathcal{P}}$ is a $(c-1)$-sized partition of the subspace $\mathcal{S}_s^{c-1} =]b_l, x_{(i_{c-1})}]$ (see Figure 5.2), and the indices of x are as before (Equation 5.5). Let $n_{\mathcal{P}}$ be the number of observations in $[A_s^{c-1}]_{\mathcal{P}} \cap \mathcal{S}_n$. For any partition $[A]_{\mathcal{P}}$ another associated partition $[A^2]_{\mathcal{P}}$ is defined as

$$[A^2]_{\mathcal{P}} = \{A_1^2, A_2^2\} = \{]b_l, x_{(i_{c-1})}],]x_{(i_{c-1})}, b_u]\},$$

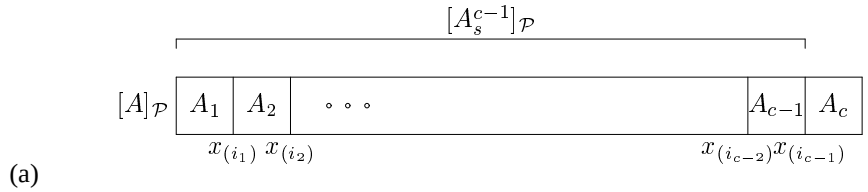
which is a 2-sized partition of $\mathcal{S}$ (see Figure 5.2). According to Lancaster's partitioning

rule (Lancaster, 1969) the test statistic has the same limiting distribution as

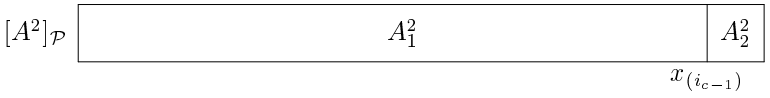
$$\frac{1}{n^{c-1}} \sum_{\mathcal{P} \in \mathbb{P}_{c-1,n}} \left(2n\hat{\mathrm{I}}^1 \left(F; G \| [A^2]_{\mathcal{P}} \right) + 2n_{\mathcal{P}} \hat{\mathrm{I}}^1 \left(F_s; G_s \| [A_s^{c-1}]_{\mathcal{P}} \right) \right),$$

where $\hat{\mathrm{I}}^1 \left(F; G \| [A_s^{c-1}]_{\mathcal{P}} \right)$ is only defined over $\mathcal{S}_s^{c-1} \subset \mathcal{S}$, i.e.

$$\hat{\mathrm{I}}^1 \left(F; G \| [A_s^{c-1}]_{\mathcal{P}} \right) = \frac{1}{2} \sum_{A_i \in [A_s^{c-1}]_{\mathcal{P}}} \frac{\left(\frac{\hat{\mathrm{P}}_f[A_i]}{\hat{\mathrm{P}}_f\left[\mathcal{S}_s^{c-1}\right]} - \frac{\mathrm{P}_g[A_i]}{\mathrm{P}_g\left[\mathcal{S}_s^{c-1}\right]} \right)^2}{\frac{\mathrm{P}_g[A_i]}{\mathrm{P}_g\left[\mathcal{S}_s^{c-1}\right]}}.$$



(a)



(b)

Figure 5.2: Decomposition of $[A]_{\mathcal{P}}$.

Also, still according to Lancaster, for each $\mathcal{P}$ and conditionally on the corresponding SSP, as $n \to \infty$ both terms converge to independent χ^2-distributed random variables. The first having 1, and the second having $c-2$ degrees of freedom. Further, as $n \to \infty$ the test statistic has the same limiting distribution as (proven by a similar technique as in the proof of Theorem 5.1)

$$\begin{aligned}
& \int_{\mathcal{S}^{c-1}} \left(2n\hat{\mathrm{I}}^1 \left(F; G \| [A^2]_{\mathcal{P}} \right) + 2n_{\mathcal{P}} \hat{\mathrm{I}}^1 \left(F; G \| [A_s^{c-1}]_{\mathcal{P}} \right) \right) dF(x_{(i_1)}) \dots dF(x_{(i_{c-1})}) \\
= & \int_{\mathcal{S}} 2n\hat{\mathrm{I}}^1 \left(F; G \| [A^2]_{\mathcal{P}} \right) dF(x_{(i_{c-1})}) \\
& + \int_{\mathcal{S}^{c-2}} 2n_{\mathcal{P}} \hat{\mathrm{I}}^1 \left(F; G \| [A_s^{c-1}]_{\mathcal{P}} \right) dF(x_{(i_1)}) \dots dF(x_{(i_{c-1})}) \\
= & \int_{\mathcal{S}} \left(2n\hat{\mathrm{I}}^1 \left(F; G \| [A^2]_{\mathcal{P}} \right) + S_{i_{c-1}} \right) dF(x_{(i_{c-1})}), \qquad (5.14)
\end{aligned}$$

where

$$S_{i_{c-1}} = \int_{\mathcal{S}} \dots \int_{\mathcal{S}} 2n_{\mathcal{P}} \hat{\mathrm{I}}^1 \left(F; G \| [A_s^{c-1}]_{\mathcal{P}}\right) dF(x_{(i_1)}) \dots dF(x_{(i_{c-2})}).$$

Conditionally on any $x_{(c-1)}$, $n_{\mathcal{P}} \to \infty$ as $n \to \infty$, and $S_{i_{c-1}}$ is asymptotically equivalent to the test statistic based on a $(c-1)$-sized partition of which it is assumed that the proposition applies to it. Thus, under H_0, $T_{c-1,n_{\mathcal{P}}} \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)} V_j^2$, where the V_j^2 are i.i.d. χ^2_{c-2} variates. The first term in the integrandum of Equation 5.14 is a Pearson statistic applied to a 2-sized partition. Thus, $T_{c,n}$ is under H_0 asymptotically equivalent to

$$\int_{\mathcal{S}} 2n \hat{\mathrm{I}}^1 \left(F; G \| [A^2]_{\mathcal{P}}\right) dF(x_{(i_{c-1})}) + \sum_{j=1}^{\infty} \frac{1}{j(j+1)} V_j^2, \tag{5.15}$$

where the first term is the AD-statistic, which is, by Theorem 5.1, under H_0 asymptotically distributed as $\sum_{j=1}^{\infty} \frac{1}{j(j+1)} W_j^2$, where the W_j^2 are i.i.d. χ^2_1-variates. Although Lancaster's Partitioning rule states that for each $x_{(i_{c-1})}$ the terms in the integrandum of Equation 5.14 are independent, it does not imply that the two terms resulting from the integral are still independent, we conjecture here that in this particular case the independence holds, at least in a first order approximation. Therefore, we propose that the limiting null distribution of $T_{c,n}$ becomes

$$\sum_{j=1}^{\infty} \frac{1}{j(j+1)} V_j^2 + \sum_{j=1}^{\infty} \frac{1}{j(j+1)} W_j^2 = \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2,$$

where the Z_j^2 are i.i.d. χ^2_{c-1}-variates.

According to the complete induction reasoning, it now only has to be shown that the theorem holds for $c - 1 = 2$, which was already proven in Theorem 5.1. □

5.2.4 The SSPc Test

In Section 5.2.3 the asymptotic null distribution of the test statistic $T_{c,n}$ was found. The next step is to construct the statistical test. The test will be called the SSPc test in general, where the "c" may be replaced by a number referring to the size of the SSP.

Let $t_{\alpha,c,n}$ denote the $(1-\alpha)$-th percentile of the null distribution of $T_{c,n}$. Then the α-level SSPc test is defined as

$$\begin{aligned} \phi_{\alpha,c,n}(\mathcal{S}_n) &= 1 \text{ if } T_{c,n} > t_{\alpha,c,n} \\ \phi_{\alpha,c,n}(\mathcal{S}_n) &= 0 \text{ if } T_{c,n} \leq t_{\alpha,c,n}. \end{aligned}$$

In the previous section, however, not the exact null distribution for any finite n was given,

but only the asymptotic null distribution. In this limiting case, $t_{\alpha,c}$ is the $(1-\alpha)$-th percentile of the null distribution of $T_{c,n}$ as $n \to \infty$, and the corresponding asymptotic α-level test is denoted by $\phi_{\alpha,c}$. The power function is given by $\beta_{\alpha,c,n}(F_1) = \mathrm{E}_{f_1}\left[\phi_{\alpha,c,n}(\mathcal{S}_n)\right]$ for alternatives $F_1 \in \mathcal{G}_1$.

An important property for a statistical test is consistency against some $F_1 \neq G$, i.e. $\beta_{\alpha,c,n}(F_1) \to 1$ as $n \to \infty$. In the case of GOF tests it is from a practical point of view not sufficient to have consistency against only one or a few F_1 from the infinitely large set $\mathcal{G}_1$. More specifically, a general-purpose GOF tests needs to be consistent against essentially any alternative $F_1 \in \mathcal{G}_1$. Such tests are called omnibus tests. The following theorem states that for any $c \geq 2$ the SSPc test is omnibus consistent.

Theorem 5.2 *For any finite $c \geq 2$ the SSPc test is consistent against essentially any alternative $F_1 \in \mathcal{G}_1$.*

Proof. For $c = 2$ consistency is easily shown by considering any fixed alternative $F \in \mathcal{G}_1$. Then, at least in some subset $R \in \mathcal{S}$ $F(x) \neq G(x), x \in R$. Hence

$$\begin{aligned} n^{-1}T_{2,n} &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{F}(X_i)-G(X_i))^2}{G(X_i)(1-G(X_i))} \\ &\xrightarrow{a.s.} \int_{\mathcal{S}}\frac{(F(x)-G(x))^2}{G(x)(1-G(x))}dF(x) \\ &> 0, \end{aligned}$$

and thus, $T_{2,n}$ grows almost surely unboundedly with n. This shows the consistency of the SSP2-test.

For the decomposition in Equation 5.15 it can be seen that $T_{c,n}$ consists of 2 positive terms; the first is a $T_{2,n}$ term, which according to the first part of the proof grows almost surely unboundedly with n. Thus, the same holds for $T_{c,n}$. Under H_0, on the other hand, $T_{c,n}$ has a non-degenerate asymptotic null distribution from which a finite critical value $t_{\alpha,c}$ is found. Thus, for any $F \in \mathcal{G}_1$. $\beta_{\alpha,c,n}(F) \to 1$ as $n \to \infty$. □

5.2.5 Some Approximations of the Asymptotic Null Distribution

The asymptotic null distribution of $T_{c,n}$, which is given in Section 5.2.3, is not easy to use in practice. The null distributions are weighted sums of an infinite number of independent χ^2-distributed random variables, which are special cases of the more general *quadratic forms*, for which e.g. Solomon and Stephens (1978) have shown that their tails may be excellently approximated by means of *Pearson Curves* (note that especially the right tail of the null distribution is of interest in the present testing situation), which can be considered as a flexible family of distributions that are fitted to the first four moments

of the true asymptotic null distribution. Also, recently, Lindsay and Basak (2000) argue that even a small set of (low order) moments may give a rather accurate approximation in the tails. Yet another approximation method, which is used before by Stephens (1976), is to find the parameters a, b, and p in $V(a,b,p) = a + bZ_p^2$, where $Z_p^2 \sim \chi_p^2$, such that the first three cumulants of both distributions are equal. For $V(a,b,p)$ these are $\mu_1 = \mathrm{E}\left[V(a,b,p)\right] = a + bp$, $\mu_2 = \mathrm{Var}\left[V(a,b,p)\right] = 2b^2p$ and $\mu_3 = 8b^3p$.
Thus, for both methods the cumulants of the asymptotic null distribution must be known. These may easily be calculated from the following corollary, which is a straightforward extension from the results of Anderson and Darling (1952), Solomon and Stephens (1978) and from the null distributions that are given in Theorem 5.1.

Corollary 5.5 *The ith cumulant μ_i of the asymptotic null distribution of $T_{c,n}$ is given by*

$$\mu_i = (c-1)2^{i-1}(i-1)!\sum_{j=1}^{\infty}\left(\frac{1}{j(j+1)}\right)^i,$$

which is $c-1$ times the corresponding cumulant of the asymptotic null distribution of $T_{2,n}$.

Proof. For $c = 2$ the expression for μ_i is given by Anderson and Darling (1952). When $c > 2$ the asymptotic null distribution of $T_{c,n}$ is given by (Proposition 5.1)

$$\sum_{j=1}^{\infty}\frac{1}{j(j+1)}Y_j^2,$$

where the Y_j^2 are i.i.d. χ_{c-1}^2-variates. By replacing each Y_j^2 by a sum of $c-1$ i.i.d. χ_1^2-distributed random variables Z_{jk}^2 ($k = 1, \ldots, c-1$), the null distribution becomes

$$\sum_{j=1}^{\infty}\sum_{k=1}^{c-1}\frac{1}{j(j+1)}Z_{jk}^2.$$

Applying the expression of the ith cumulant of a general quadratic form (Solomon and Stephens, 1978) completes this proof. □

Since for $c > 2$ the cumulants may be calculated from those for $c = 2$ by simply multiplying by $c-1$, it is sufficient to give the cumulants for $c = 2$. These are shown in Table 5.1.

Table 5.1: The first four cumulants for $c = 2$.

μ_1	μ_2	μ_3	μ_4
1	0.5797	1.043	3.040

Since for $c = 2$ $\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = 2.363 > 2$ all $\sqrt{\beta_1}$ are out of the range of the published tables (Pearson and Hartley, 1972), Pearson curves cannot be used. The approach

of Stephens (1976) may still be a way out. Using the cumulants of the limiting null distribution of $T_{c,n}$ (Corollary 5.5), the solution is easily obtained:

$$\begin{aligned} a &= c - 1.6444 \\ b &= 0.4498 \\ p &= 1.4326(c-1). \end{aligned}$$

In table 5.2 some comparisons are made between the exact and the approximate critical values of the asymptotic null distribution of $T_{c,n}$, for $c = 2$ and $c = 3$. The exact critical values are obtained from Scholz and Stephens (1987). It seems that the approximations work rather good for $c = 2$, but when $c = 3$, the approximation is already bad. For larger values of c, the approximation becomes even worse (results not shown).

Table 5.2: Some critical values of the null distribution of $T_{c,n}$, obtained from the exact limiting null distribution and from the approximate limiting null distributions based on 3 and on 4 cumulants. The latter are based on 50000 simulation runs.

	$t_{0.05,2}$	$t_{0.10,2}$	$t_{0.05,3}$	$t_{0.10,3}$
Exact	2.492	1.933	4.030	3.377
Approximate (3)	2.531	1.964	4.965	4.071
Approximate (4)	2.495	1.929	4.085	3.395

Of the two approximation methods described above, the Pearson curves have the advantage that they mimic the first 4 cumulants of the true limiting distribution, whereas the other methods is only based on 3. On the other hand, the latter method has the advantage that the $V(a, b, p)$ variate can be used directly to obtain p-values as well.
Here we propose another approximation method which combines both advantages: the first 4 cumulants are matched to those of the true limiting null distribution, and p-values may be calculated. The disadvantage, however, is that the method basically uses a simulated sample from a distribution. The distribution is of the form

$$W(a, b, p, f) = a + bZ_p^2 + \frac{1}{2}Y_f^2,$$

where $Z_p^2 \sim \chi_p^2$ and $Y_f^2 \sim \chi_f^2$, and where a, b, p and f are determined such that the first 4 cumulants of $W(a, b, p, f)$ match those of the asymptotic null distribution of $T_{c,n}$. The solution is

$$\begin{aligned} a &= 0.1919467416(c-1) \\ b &= 0.1274341352 \\ p &= 2.404852143(c-1) \\ f &= 1.003186011(c-1). \end{aligned}$$

In Table 5.2 this approximation method is compared to the exact solution. The approximated critical values are rather close to the exact values.

One might argue that if one is prepared to use a stochastic approximation as the one that is suggested here, one could just as well simulate the first dominating terms of $\sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2$ (Z_j i.i.d. χ^2_{c-1}), say $j = 1, \ldots 4$. This has been done by Kac, Kiefer and Wolfowitz (1955), but did not result in satisfying approximations.

5.2.6 Null Distributions for Finite n

In the previous sections much attention was given to the asymptotic behaviour of the SSPc test. This may of course be of interest from a theoretical point of view, whereas in a practical situation one is always confronted with finite sample sizes. For a detailed theoretical study of the properties of the SSPc test for finite n, statistical theory is nowadays however still not sufficiently developed. This also applies to the calculation of the exact null distribution of $T_{c,n}$ for finite n. Often the convergence of the exact critical values to those calculated from the asymptotic null distribution is rather fast. Under these circumstances the asymptotic results may still be used as an approximation.

The convergence rate for the SSPc statistic is empirically studied later in this section, but first the methodology of simulating the exact null distribution is explained briefly. Finally an approximation for finite sample sizes will be considered.

Simulating the Exact Null Distribution

If there is no theoretical valuable solution for the approximation of the exact null distribution or the critical values, then it is often still possible to estimate the exact null distribution by means of a Monte Carlo simulation. In general this procedure consists in generating a large number, say $N = 10000$, of independent samples $\mathcal{S}_{n,i}$ ($i = 1, \ldots, N$) from the specified distribution G (by means of a pseudo random generator). Next, the test statistic is calculated for each sample, leading to N realizations $t_{c,n,i}$ from the (simulated) exact null distribution of $T_{c,n}$. An estimate can then be obtained from this large sample from the null distribution. Also, an estimate $\hat{t}_{\alpha,c,n}$ of $t_{\alpha,c,n}$ is calculated as the $(1-\alpha)$-percentile from the EDF of $\{t_{c,n,i}\}$, i.e. the $\lfloor n(1-\alpha) \rfloor$ th order statistic of the sample $\{t_{c,n,i}\}$. It is obvious that this critical value is more accurately estimated as N becomes larger. An approximate $(1-\gamma)$-confidence interval may be calculated as

$$\left[\hat{t}_{\alpha_l,c,n}, \hat{t}_{\alpha_u,c,n}\right],$$

where

$$\alpha_l = \left\lceil N(1-\alpha) - z_{1-\gamma}\sqrt{N\alpha(1-\alpha)} \right\rceil \quad \text{and} \quad \alpha_u = \left\lceil N(1-\alpha) + z_{1-\gamma}\sqrt{N\alpha(1-\alpha)} \right\rceil,$$

where $z_{1-\gamma}$ is the $(1-\gamma)$ percentile of a standard normal distribution.
In this thesis $N = 10000$ for all simulated null distributions.

Figure 5.3 shows a histogram of the simulated exact null distribution of $T_{2,20}$ with the estimated 5%-level critical value indicated.

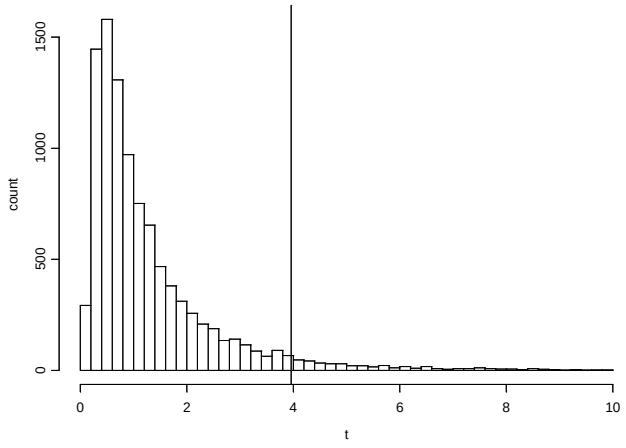


Figure 5.3: A histogram of the simulated exact null distribution of $T_{2,20}$, based on $N = 10000$ simulations. The vertical line indicates the estimate of $t_{0.05,2,20}$.

The simulated exact null distribution may also be used to obtain an estimate $\hat{\alpha}$ of the size of an α-level test which is based on a critical value, say t_{crit}. This value may be e.g. the asymptotic or an approximate critical value. The estimated size is given by $\hat{\alpha} = n^{-1}\#\{t_{c,n,i} : t_{c,n,i} > t_{\text{crit}}\}$. Note that $\hat{\alpha} = \hat{\beta}_{\alpha,c,n}(G)$, i.e. the estimated power of the α-level test under the null hypothesis.
An approximate 95% confidence interval of α may be calculated as

$$\left[\hat{\alpha} - z_{0.975}\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{N}}, \hat{\alpha} + z_{0.975}\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{N}}\right],$$

where $z_{0.975}$ is the 0.975 percentile of a standard normal distribution.

Convergence to the Limiting Null Distribution

When only the limiting null distribution of a test statistic is available, one often hopes that the convergence is rather fast, i.e. the critical values $t_{\alpha,c,n}$ converges quickly to $t_{\alpha,c}$. In this way one can for intermediate large data sets (e.g. $n = 25$) already apply the asymptotic test, and thus one only needs one table of critical values. Moreover, in that situation one can use the approximation proposed in Section 5.2.5 to compute the approximate p-values. This is the first reason why the convergence is studied.
The second reason is to assess the validity of Proposition 5.1 which states the asymptotic null distribution of the SSPc statistic with arbitrary SSP size c.

2-sized sample space partitions:
For a wide range of sample sizes (from 20 to 1000) the critical values $t_{0.05,2,n}$ are estimated and plotted in Figure 5.4, panel (a). The figure suggests clearly that the convergence is very slow. Only for large sample sizes the critical value has sufficiently approached the asymptotic value (the critical value is the same as for the AD-test, and it is obtained from e.g. Stephens (1986, p.105)). Thus, for $c = 2$ and moderate sample sizes one cannot use the asymptotic null distribution, but one may use the simulated exact null distribution instead.

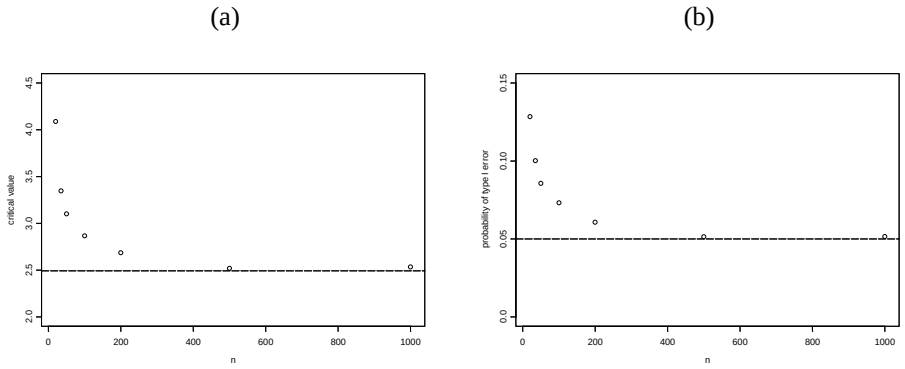


Figure 5.4: Estimated critical values $\hat{t}_{0.05,2,n}$ (a) and the corresponding estimated probabilities of a type I error when the asymptotic critical value $t_{0.05,2}$ is used (b). The horizontal line indicates the asymptotic critical value (a) and the nominal level of 5% (b).

It may be surprising at first sight that the convergence is slow. Especially since in the literature on the Anderson-Darling test, which has the same limiting null distribution, it has many times been mentioned that the AD-statistic has a fast converging null distribution (e.g. Lewis, 1961; Stephens, 1974; Pettit, 1976; Stephens, 1986). It may however be understood by recognizing that the asymptotic null distribution of $T_{c,n}$ is based on the convergence of $\hat{F}_n$, which is the estimator of the distribution with respect to which the expectation in Equation 5.6 is taken, to G under H_0, whereas the Anderson-Darling statistic can be seen as the plug-in estimator of

$$A_n = 2n\mathrm{E}_g \left[\mathrm{I}^1 \left(F; G \| [A](X) \right] \right),$$

which is an expectation w.r.t. G for which there is no need to be replaced by an estimator. Thus, the limiting null distribution of the SSP2 statistic is based on more convergences as the AD statistic. This may be a reason why the convergence of the AD statistic is much faster.

The convergence can also be assessed by using the simulated exact distributions to estimate the probability of making a type I error when the asymptotic critical value would have been used. With increasing sample size this probability should converge to α. Figure 5.4, panel (b), shows the results for $c = 2$ ($\alpha = 0.05$). This figure shows the same

slow convergence, but now the consequence of using the asymptotic critical values may be more realistically quantified. It seems that even for $n = 100$ the actual α is still about 2% above the nominal level.

3-sized sample space partitions:
For $c = 3$, the results are presented in Figure 5.5. They show generally the same pattern as for $c = 2$. Nevertheless, the figure clearly shows that the estimated exact critical value converge, though slowly, to the value calculated from the proposed limiting null distribution (similar convergence behaviour was obtained for other α-levels and for $c = 4$ and $c = 5$ (results not shown)). Thus, the results at least confirm Proposition 5.1.

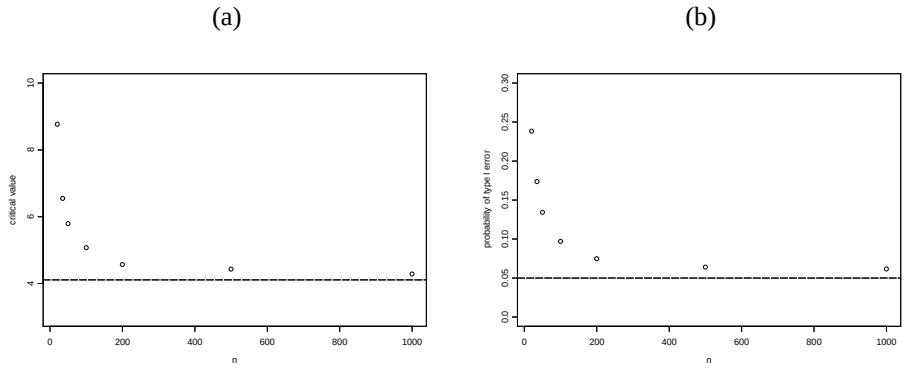


Figure 5.5: Estimated critical values $\hat{t}_{0.05,3,n}$ (a) and the corresponding estimated probabilities of a type I error when the asymptotic critical value $t_{0.05,3}$ is used (b). The horizontal line indicates the asymptotic critical value (a) and the nominal level of 5% (b).

An Interpolation Formula for Critical Values for Finite n

Although the use of the simulated exact null distribution is a good and correct methodology, it also has some disadvantages from a practical point of view. The most important is that for each n a Monte Carlo simulation must be run in order to estimate the critical value. In this paragraph an easy-to-use approximation is given. It is in the line of the work of e.g. Scholz and Stephens (1987).

For each α we suggest to use an interpolation formula of the form

$$\hat{t}_{\alpha,c,n} = t_{\alpha,c} + \frac{b_{1,\alpha,c}}{\sqrt{n}} + \frac{b_{2,\alpha,c}}{n} + \frac{b_{3,\alpha,c}}{n^2} + o(n^{-2}),$$

where the parameters $b_{i,\alpha,c}$ ($i = 1, 2, 3$) are estimated from a regression of the estimated critical values $\hat{t}_{\alpha,c,n}$ from the simulated null distributions from the previous section, on $\frac{1}{\sqrt{n}}$, $\frac{1}{n}$ and $\frac{1}{n^2}$.

For $\alpha = 0.05$ and $c = 2, 3, 4$ the estimated parameters are presented in Table 5.3.

Table 5.3: The estimated parameters of the interpolation formula for $\hat{t}_{\alpha,c,n}$ ($c = 2, 3, 4$).

c	$b_{1,\alpha,c}$	$b_{2,\alpha,c}$	$b_{3,\alpha,c}$
2	0.9045	22.18	111.4
3	4.335	41.06	653.4
4	1.892	129.4	1792

As a small validation experiment, the exact null distributions of SSP2 and SSP3 are simulated for $n = 25$ and $n = 75$. From these simulated null distributions the size of the test, when based on the critical value obtained from the interpolation formula, is estimated. For $c = 2$, the estimated sizes are 0.052 and 0.050 for $n = 25$ and $n = 75$, respectively. When $c = 3$, the estimated sizes for $n = 25$ and $n = 75$ are 0.056 and 0.050, respectively, and when $c = 4$ the estimated sizes are 0.046 and 0.049 , respectively. Taking into account that the estimations are only based on a regression of 7 simulated critical values, which are each based on only 10000 simulation runs, the validation suggest that the proposed interpolation formulae may result in rather good approximations.

5.2.7 Power Characteristics

This section gives a discussion on the power of the SSPc test. First a theoretical property is given. Then the methodology for simulation studies is explained, and finally the results of a simulation study are presented in which the powers of the SSPc test and some other existing GOF tests are compared.

Some Asymptotic Power Characteristics

In Section 5.2.4 it was shown that the SSPc test is consistent, i.e. for essentially all alternatives the SSPc test has a power tending to 1 as the sample size n tends to infinity. Since many GOF tests possess this property, it is not possible to compare tests on this basis. Neyman (1937) proposed to consider as alternative hypotheses a sequence of distributions F_n that converges to G as n tends to infinity. In this way it gets more and more difficult to discriminate the alternative from the null hypothesis as n increases, and as a consequence, for a suitable convergence rate of F_n, the asymptotic power is kept away from 1. Thus under such a sequence of alternatives different tests may have different asymptotic powers and can therefore be compared.

Since the SSPc test is clearly constructed starting from Pearson's statistics, the sequence that is used in the present context is related to sequences used to assess the asymptotic power of Pearson's χ^2 test, in the sense that the convergence rate is the same. Consider the sequence of alternatives

$$f_n(x) = g(x) + n^{-1/2}c(x), \tag{5.16}$$

where $c(.)$ is a continuous function such that for each $n \geq 1$ the df f_n is a proper df (Assumption A1). (Thus, $\int_{\mathcal{S}} c(x)dx = 0$ (normalization condition); for all $x \in \mathcal{S}$, $c(x) > -g(x)$ (f_n must be positive)) Let F_n denote the corresponding CDF.

Theorem 5.3 *The SSP2 test has the same asymptotic power as the AD test under the sequence given in Equation 5.16.*

Proof. We give a sketch of a proof.
Consider the following double array.

$$\begin{aligned} X_{11} &\sim F_1, \text{ which is estimated by } \hat{F}_{11} \\ X_{21}, X_{22} &\sim F_2, \text{ which is estimated by } \hat{F}_{22} \\ \vdots &\quad \vdots \\ X_{n1}, \ldots, X_{nn} &\sim F_n, \text{ which is estimated by } \hat{F}_{nn} \end{aligned}$$

(Thus, $\hat{F}_n$ is here equal to $\hat{F}_{nn}$, which stresses that the distribution F depends on n as well.) Note that for all $x \in \mathcal{S}$,

$$\mathrm{E}\left[\hat{F}_{nn}(x)\right] = F_n(x) \longrightarrow G(x),$$

and

$$\mathrm{Var}\left[\sqrt{n}(\hat{F}_{nn}(x) - F_n(x))\right] = F_n(x)(1 - F_n(x)) \longrightarrow G(x)(1 - G(x))$$

as $n \to \infty$. Moreover, for each $x \in \mathcal{S}$

$$V_n(x) = \frac{\sqrt{n}\left(\hat{F}_{nn}(x) - F_n(x)\right)}{\sqrt{F_n(x)(1 - F_n(x))}} \xrightarrow{d} N(0,1)$$

according to a consequence of the Berry-Esseen theorem (see e.g. Corollary 2.4.1. in Lehmann, 1999).

Lemma 5.2 states that the SSP2 statistic is asymptotically equivalent to the member of the Anderson-Darling Family:

$$U_n = n \int_{\mathcal{S}} \frac{\left(\hat{F}_{nn}(x) - G(x)\right)^2}{G(x)(1 - G(x))} \frac{f(x)}{g(x)} dG(x),$$

where f denotes the df of the true distribution, which is here replaced by the sequence

$f_n(x) = g(x) + n^{-1/2}c(x)$. Hence,

$$U_n = n\int_{\mathcal{S}} \frac{\left(\hat{F}_{nn}(x) - G(x)\right)^2}{G(x)(1-G(x))} dG(x) + n^{-1/2} n \int_{\mathcal{S}} \frac{\left(\hat{F}_{nn}(x) - G(x)\right)^2}{G(x)(1-G(x))} \frac{c(x)}{g(x)} dG(x),$$

where the second term may be written as

$$n^{-1/2}\int_{\mathcal{S}} \frac{\left(\sqrt{n}\left(\hat{F}_{nn}(x) - F_n(x)\right)\right)^2}{G(x)(1-G(x))} \frac{c(x)}{g(x)} dG(x) \tag{5.17}$$

$$+ \quad 2n^{-1/2}\int_{\mathcal{S}} \frac{\sqrt{n}\left(\hat{F}_{nn}(x) - F_n(x)\right) C(x)}{G(x)(1-G(x))} \frac{c(x)}{g(x)} dG(x) \tag{5.18}$$

$$+ \quad n^{-1/2}\int_{\mathcal{S}} \frac{C^2(x)}{G(x)(1-G(x))} \frac{c(x)}{g(x)} dG(x). \tag{5.19}$$

By recognizing that $V_n(x)$ is an empirical process that is closely related to a Brownian bridge, one can show (Donsker's theorem and the Continuous Mapping Theorem, (see e.g. McCabe and Tremayne, 1993)) that the stochastic integrals of the first two terms (5.17 and 5.18) have asymptotically proper distributions, and thus, $n^{-1/2}$ times the integral converges in probabiliy to zero as $n \to \infty$. The last term (5.19), which is not stochastic, also converges to zero as $n \to \infty$.

Thus,

$$T_{2,n} - n\int_{\mathcal{S}} \frac{\left(\hat{F}_{nn}(x) - G(x)\right)^2}{G(x)(1-G(x))} dG(x) \xrightarrow{p} 0,$$

which is the AD-statistic. □

An immediate consequence is given in the following corollary.

Corollary 5.6 *The SSP2 test and the AD test have the same Pitman efficiencies w.r.t. all other tests.*

Methodology for Small Sample Power Studies

Since the SSPc test is consistent the power of the test is 1 for essentially all alternatives whenever n is infinitely large. This important property is shared with many other omnibus GOF tests (see Chapter 4). Two points of criticism are in place here: first, infinitely large sample sizes are not of practical importance, and second, on this basis consistent tests cannot be compared. One way out is to calculate the Bahadur or Pitman efficiencies for tests that have to be compared, but again this is a comparison for infinitely large sample

sizes. For sample sizes that are of practical relevance, power simulation studies may be performed.

In a power study Monte Carlo simulations are run in which data of any sample size n are simulated according to a specified alternative distributions. The general algorithm is briefly stated below. Suppose that the power $\beta_{\alpha,c,n}(F_1)$ of the SSPc test is to be estimated for sample size n, alternative distribution F_1 and level α.

Algorithm:
First, set a counter variable CNT to zero. The following steps are repeated N times.

1. Generate a sample of size n from distribution F_1 by means of a pseudo random generator.
2. Apply the SSPc test to the sample at the α-level.
3. If the SSPc test results in a significant rejection of H_0 at the α-level, then increase the counter CNT by one.

The estimated power is then given by $\hat{\beta}_{\alpha,c,n}(F_1) = \frac{\text{CNT}}{N}$. Note that this is actually an estimate of the probability $\beta_{\alpha,c,n}$ of a binomial variable. Thus, for N Monte Carlo runs, an approximate 95% confidence interval for the power can be constructed as

$$\left[\hat{\beta}_{\alpha,c,n} - z_{0.975}\sqrt{\frac{\hat{\beta}_{\alpha,c,n}(1-\hat{\beta}_{\alpha,c,n})}{N}}, \hat{\beta}_{\alpha,c,n} + z_{0.975}\sqrt{\frac{\hat{\beta}_{\alpha,c,n}(1-\hat{\beta}_{\alpha,c,n})}{N}}\right],$$

where $z_{0.975}$ is the 0.975 percentile of a standard normal distribution.

The choice of alternatives F_1 is often not obvious when an overall view of the performance of an omnibus test is to be obtained (see e.g. Hájek and Šidák, 1967; Lehmann, 1975; Kallenberg and Ledwina, 1999). Indeed, omnibus tests are sensitive to essentially all alternatives, but none of the omnibus GOF-tests is optimal. This means that some tests will have higher power for some alternative and others may have higher powers for other alternatives. Thus, in order to get a good overview a wide range of alternatives should be included in a simulation study. Of course, the alternatives that are included should also have some practical relevance.
The power estimation of other statistical tests is similar. An immediate consequence of the definition of the power function $\beta_{\alpha,n}$ of a test of size α is that $\beta_{\alpha,n}(G) = \alpha$, i.e. $\beta_{\alpha,n}$ is the size of the α-level test. In practice, however, many tests are performed with approximate α-level critical values, i.e. in the definition of the statistical test the exact critical value $t_{\alpha,n}$ is replaced with an approximation, say $\tilde{t}_{\alpha,n}$ or even $\tilde{t}_{\alpha}$. As a consequence the power function, evaluated in G, may be different from α for some sample sizes n. A study in which powers of several tests are compared to each other is only meaningful if all tests are indeed of size α. For all tests included in the power studies in this thesis, the size $\beta_{\alpha,n}(G)$ was estimated, and if it was observed that the size differed from α, the critical values of the test were replaced by the corresponding approximate exact critical

values $\hat{t}_{\alpha,n}$ as it was described in Section 5.2.6. By construction, these tests are of size α. The resulting estimated powers are known as *size-adjusted* powers.

Yet another approach for comparing the performance of GOF tests was proposed by Mudholkar, Kollia, Lin and Patel (1991). The main advantages of this method are that it is fast as compared to the traditional Monte Carlo simulations, and that in one figure the results for a wide range of alternatives can be presented which facilitates the interpretation. In the next paragraph this technique is explained.

Instead of applying the GOF test to N realizations of the alternative F_1, it is only applied once to one specifically constructed sample, which is called the *ideal sample* (e.g. Andrews et al., 1972) or the *profile* of F_1.

Definition 5.7 *A* **profile** P_n *of* F_1 *is a sample of size* n *constructed as*

$$P_n = \{F_1^{-1}(p_1), \ldots, F_1^{-1}(p_n)\},$$

where $p_i = \frac{i-0.5}{n}$ *(*$i = 1, \ldots, n$*).*

The central idea is that a large value of the test statistic calculated for the profile corresponds to a high power. In the original paper Mudholkar et al. (1991) suggest to calculate the test statistic for the profiles of a family of alternatives F_θ, indexed by two parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$, in which the distribution G is embedded. The corresponding profile is denoted by $P_{n,\theta}$. Let $t(P_{n,\theta})$ denote this test statistic. Two types of graphs can be constructed.

- The 3-dimensional surface $(\theta_1, \theta_2, t(P_{n,\theta}))$ may give a good general overview of the relative power in the family indexed by $\boldsymbol{\theta}$.
- The 3-dimensional surface may be reduced to a 2-dimensional graph in the plane (θ_1, θ_2) by plotting the *isotone*, which is the set of points (θ_1, θ_2) for which $t(P_{n,\theta})$ equals some specified constant. The authors suggest to take for this constant the critical value for the test at the α-level. Furthermore, in the same plane isotones for more than one GOF test may be plotted. Since the distribution G is embedded in the family of alternatives, the plot of the isotones may be interpreted easily be starting from the point representing G and moving into a certain direction. The isotone that is crossed first corresponds most likely with the test having the largest power. Of course, the order of isotones crossed may differ from one direction to another. This produces a concise comparison of these tests.

The Power Study for the SSPc-test

Simple null hypotheses like those studied in this section are actually not much of practical interest because it does not happen too often that one knows G completely. Most frequently only the form of G is known, i.e. G is a family of distributions, resulting in a

composite null hypothesis. This interesting case is studied later in Section 5.3. Of course many of the results will be related to those for the simple null hypothesis.
Thus, since the simple null hypothesis situation is not very important from a practical point of view, only a small power study is presented here.

Isotones are constructed for two families of alternatives to the normal distribution G with known mean and variance.

- **Tukey-Lambda** family indexed by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$.
  The quantile function F_λ^{-1} is given by

  $$F_\lambda^{-1}(p) = \frac{p^{\lambda_1}}{\lambda_1} - \frac{(1-p)^{\lambda_2} - 1}{\lambda_2}.$$

  For a description of the Tukey-Lambda family we refer to Appendix A. We only mention here briefly that for $\boldsymbol{\lambda} = (0.1349, 0.1349)$ the Tukey-Lambda distribution is almost indistinguishable from a normal distribution with mean 0 and variance 2.142 . Also for $\boldsymbol{\lambda} = (5.2, 5.2)$ the distribution looks closely to a normal distribution ($\mu = 0$, $\sigma^2 = 0.00647$).
  For this family of alternatives, the distribution G specified under the simple null hypothesis is the normal distribution $N(0, 2.142)$.

- **Mixture of normal distributions** indexed by $\boldsymbol{\theta} = (\delta, \gamma)$ (also referred to as a **contaminated normal distribution**).
  Mixtures of normal distributions can be constructed in many ways. Here, we consider mixtures with densities

  $$f_\theta(x) = (1-\gamma)g(x) + \gamma g\left(\frac{x-\delta}{0.001}\right),$$

  where g is the density of a standard normal distribution as specified in H_0. $\delta = 0$ and $\gamma = 0$ gives the distribution G.
  The interpretation of the parameters γ and δ is clear. γ is the fraction or probability mass of the distribution that is contaminated with a normal distribution with mean δ and standard deviation 0.001. Figure 5.6 shows two examples of a contaminated normal distribution.

Table 5.4: Critical values at the 5%-level for $n = 20$ and $n = 50$ for the GOF tests for (simple) normality used in this thesis.

n	SSP2	SSP3	SSP4	AD	KS
20	3.960	9.102	16.948	2.492	1.358
50	3.253	5.943	9.023	2.492	1.358

The isotones are constructed for $n = 20$ and $n = 50$. All tests are performed at $\alpha = 0.05$. Three SSPc tests are considered: SSP2, SSP3 and SSP4. The SSPc tests are compared to some other GOF tests for simple null hypotheses: Anderson-Darling (AD) (Anderson

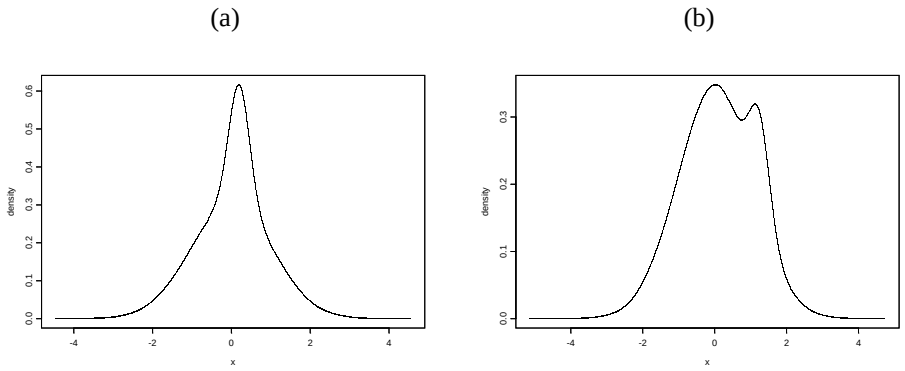


Figure 5.6: Two examples of a contaminated normal distribution with $\delta = 0.2$ and $\gamma = 0.2$ (a), and $\delta = 1.25$ and $\gamma = 0.1$ (b).

and Darling, 1952) and Kolmogorov-Smirnov (KS) (Kolmogorov, 1933; Smirnov, 1939) (here the modified statistic of Stephens (1970) is used). These tests are generally well known and can therefore serve as benchmarks in the comparison. They are all discussed in Sections 4.2 and 4.3. The critical values are given in Table 5.4. Later, in the simulation study for a composite null hypothesis, more tests will be considered. The critical values of the SSPc tests are obtained from the simulated null distributions, based on 10000 simulation runs.

Briefly, the AD test is in general the EDF test with the best overall power characteristics, and the KS test is probably one of the most frequently used GOF tests, although it is frequently shown that it has rather low power to many alternatives (e.g. Stephens, 1986).

The results for the Tukey-Lambda family of alternatives and the family of mixtures are shown in Figure 5.7 and Figure 5.8, respectively. To give an impression of the magnitude of the powers, for a selection of alternatives, the powers are estimated by means of a simulation study based on 10000 runs. The results are shown in Tables 5.5 and 5.6 for the Tukey-Lambda family and the contaminated normal family, respectively.

Discussion

Tukey-Lambda: First, note that all curves are closed around $\boldsymbol{\lambda}_0 = (0.1349, 0.1349)$, and that the curves are closer to $\boldsymbol{\lambda}_0$ as n increases from 20 to 50. This reflects the consistency property of all tests. The best way to discuss the isotones is to start from the point $\boldsymbol{\lambda}_0$ representing the null hypothesis, and to move along certain directions. The order in which the isotones of the different tests are crossed is an indication of the order of the power of these tests.

- The direction along the diagonal towards higher values of λ_1 and λ_2 represents all symmetric distributions, starting from unimodal distributions with continuous tails ($\lambda_1 = \lambda_2 < 1$), over U-shaped distributions with truncated tails ($1 < \lambda_1 = \lambda_2 < 2$), to unimodal distributions with truncated tails ($\lambda_1 = \lambda_2 > 2$). In this direction

Table 5.5: The estimated powers for some alternatives of the Tukey-Lambda family, based on 10000 simulation runs. The mean and the variance of the distributions are also given.

n	λ_1	λ_2	mean	variance	power				
					SSP2	SSP3	SSP4	AD	KS
20	-0.2	-0.2	0	7.42	0.684	0.636	0.592	0.543	0.116
20	1	1	0	0.333	0.001	0.333	0.268	0.698	0.298
20	1.5	1.5	0	0.157	0.003	1.000	0.948	1.000	1.000
20	0.9	2	-0.193	0.213	0.010	0.889	0.721	1.000	1.000
20	2	0	0.666	1.30	0.098	0.258	0.230	0.885	1.000
50	-0.2	-0.2	0	7.42	0.926	0.931	0.928	0.828	0.201
50	0.7	0.7	0	0.57	0.271	0.995	0.974	0.996	0.714
50	-0.2	0	0.25	6.29	0.820	0.826	0.830	0.590	0.147
50	0.5	0	0.333	1.97	0.337	0.590	0.635	0.929	1.000
50	0.3	1	-0.296	0.788	0.321	0.895	0.884	0.999	1.000

Table 5.6: The estimated powers for some alternatives of the contaminated normal family, based on 10000 simulation runs. The mean and the variance of the distributions are also given.

n	δ	γ	mean	variance	power				
					SSP2	SSP3	SSP4	AD	KS
20	0	0.2	0	0.79	0.024	0.716	0.767	0.044	0.148
20	0	0.6	0	0.40	0.142	1.000	1.000	0.737	0.919
20	2	0.2	0.40	1.45	0.477	0.770	0.795	0.546	0.334
50	0	0.2	0	0.79	0.047	0.950	0.981	0.156	0.315
50	0	0.35	0	0.65	0.370	1.000	1.000	0.669	0.836
50	0.18	0.18	0.03	0.83	0.039	0.915	0.967	0.129	0.279

the AD test is the most powerful for both $n = 20$ and $n = 50$. The second highest power is obtained by the KS test when $n = 20$, and by SSP3 when $n = 50$. This feature is clearly seen when comparing Figures 5.7 (a) and (b): the contour of the KS test does not shrimp as much as those of the other tests when n increases from 20 to 50. Especially the contours of the SSPc tests seem to shrimp extensively by the sample size increase. The order in power among the SSPc tests in ascending order is SSP2-SSP4-SSP3.
Note that for these symmetric distributions the mean is always equal to 0, which is also the mean under the null hypothesis; the variance, on the other hand, varies.

- The direction along the diagonal towards negative values of λ_1 and λ_2 also represents symmetric unimodal distributions with continuous tails which become heavier as $\lambda_1 = \lambda_2$ decreases. The isotones show in this direction that now the SSPc tests are more powerful. This is also clearly seen from the estimated powers in Table 5.5. The KS test is least sensitive.

- The isotones of the SSP3 and SSP4 test show for $n = 20$ a rather insensitive behaviour in the directions of $\lambda_1 \approx 0$ and λ_2 positive, and $\lambda_2 \approx 0$ and λ_1 positive.

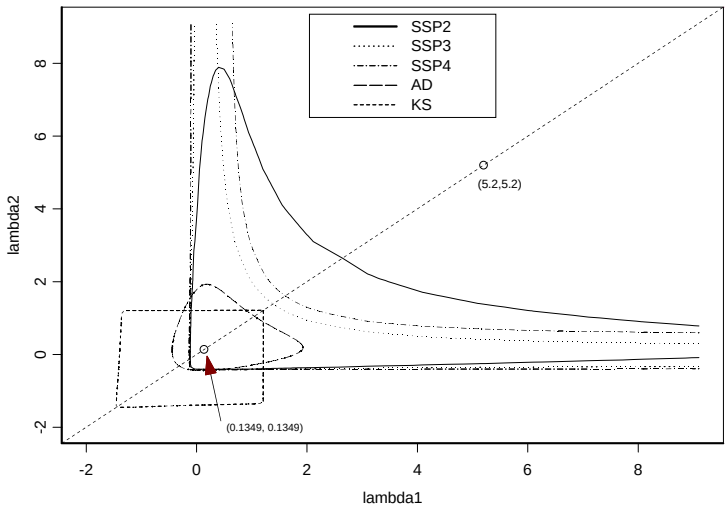
These former direction represents distributions that are similar to the exponential and χ^2 distributions. The latter direction represents distributions with the same shape, but with the tails to the left. This insensitiveness quickly reduces substantivally as n increases.

Contaminated Normal: Figure 5.8 shows mainly the half-plane determined by positive values of δ, since the plot is symmetric around $\delta = 0$. For negative values of γ the distribution is not defined. The null hypothesis is given by $(\delta, \gamma) = (0, 0)$ (standard normal distribution). All contours in both Figure 5.8 (a) and (b) show the same shape, which makes the discussion easier. The only difference between $n = 20$ and $n = 50$ is that the isotones are closer to the point representing H_0 for the higher sample size. Two directions of alternatives are discussed next (note that actually all points for which $\gamma = 0$ represent H_0, so that the corresponding direction must not be discussed).

- The direction from $(0, 0)$ to $(0, \infty)$ represents mixtures of two normal distributions with mean 0; the higher the value of γ the more the standard normal distribution is contaminated with a normal distribution with standard deviation 0.001, resulting in a zero-mean normal distribution with reduced variance. The isotones that are crossed first when moving in this direction, are those of SSP4 and SSP3, indicating that they give the highest powers. Next, the isotones of KS, AD are crossed. The lowest power is obtained with the SSP2 test.
- Along the direction from $(0, 0)$ to $(4, 0.4)$ the mean increases and the variance decreases. Again first the isotones of SSP4 and SSP3 are crossed, next those of SSP2 and AD (their order switched somewhere in around this direction, but they stay close to each other). The KS test gives the lowest power.

Some General Conclusions:
Obviously none of the tests have overall the highest power. Almost under all alternatives, the SSP2 test has the lowest power among the SSPc tests. Furthermore, its power is also often smaller than the power of the AD test, except for $\lambda_1 = \lambda_2 < 0$ in the TL family. All SSPc tests are more powerful than the KS test, except under a few specific alternatives. Under the mixture alternatives, the SSP3 and the SSP4 tests have by far the highest power among all tests considered.

(a)



(b)

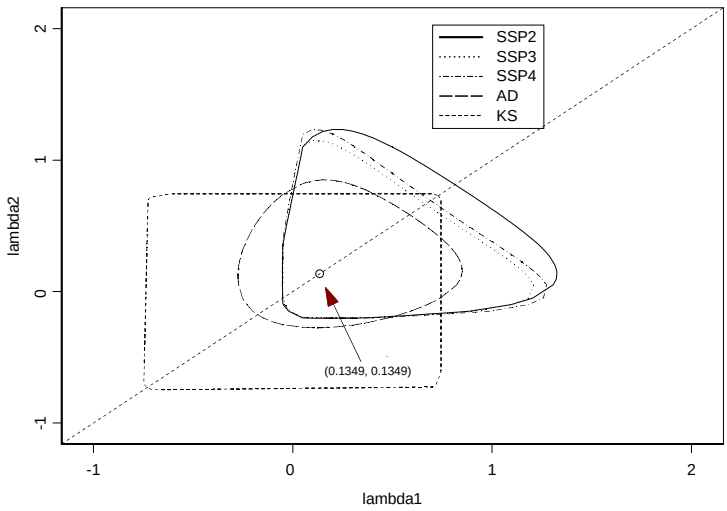
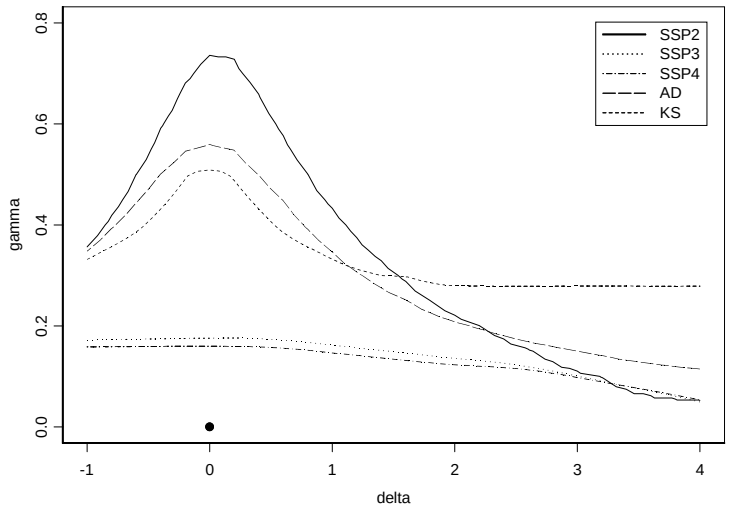


Figure 5.7: The isotones for the Tukey-Lambda family of alternatives, for $n = 20$ (a) and $n = 50$ (b). The diagonal dashed line indicates the symmetric distributions ($\lambda_1 = \lambda_2$).

(a)



(b)

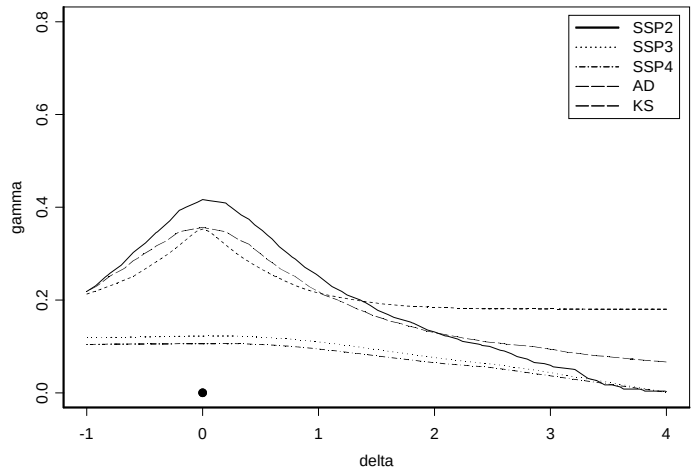


Figure 5.8: The isotones for the family of mixture distributions, for $n = 20$ (a) and $n = 50$ (b).

5.3 The SSP-Test: Composite Null Hypothesis

In the previous section the SSPc test was constructed for a simple null hypothesis in which the distribution G was completely specified. In practice, however, this situation does not occur often. Of more interest is the situation in which a researcher has to test whether a sample comes from a particular family of distributions, denoted by $\mathcal{G}_0 = \mathcal{G}_{\Theta_0}$ with members G_θ, i.e. a functional form of G_θ is assumed, but the parameters $\boldsymbol{\theta}$ are still to be determined. The most classical example, which is still of substantial importance, is testing for normality: in this example $\boldsymbol{\theta} = (\mu, \sigma)$ and $G_\theta(x) = \int_\infty^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$.

Again, as for the simple null hypothesis case, we will suppose x is univariate. The test will be constructed based on Pearson-type divergences. Later the test will be generalized to the power divergence family and to a multivariate setting. The composite null hypothesis is

$$H_0 : F_x \in \mathcal{G}_{\Theta_0}.$$

In the previous section the test was constructed starting from the Averaged Pearson Divergence and the corresponding statistic. Since now G_θ is not completely specified this statistic needs to be changed first. Next, the test statistic is proposed. Since even for the Anderson-Darling statistic there is no analytical solution for the null distribution, it is expected that for general c the null distribution of the SSP based statistic is complicated as well. Finally some results from a power study are given.

5.3.1 The Averaged Pearson Divergence Statistic

In Section 5.2.2 the Averaged Pearson Divergence Statistic of order c was defined as

$$\hat{\Delta}_c(F;G) = \frac{1}{a_{c-1,n}} \sum_{\mathcal{P}\in\mathbb{P}_{c-1,n}} \hat{\mathrm{I}}^1\left(F;G\|[A]_\mathcal{P}\right), \tag{5.20}$$

where $\mathcal{P} = \{X_1, \ldots, X_{c-1}\} \subset \mathcal{S}_n$, and where $\hat{\mathrm{I}}^1\left(F;G\|[A]_\mathcal{P}\right)$ is the plug-in estimator of $\mathrm{I}^1\left(F;G\|[A]_\mathcal{P}\right)$, i.e. F is replaced by $\hat{F} = \hat{F}_n$. If the null hypothesis is composite, not only F is unknown, but also the parameter $\boldsymbol{\theta}$ that determines G_θ. The estimation of $\boldsymbol{\theta}$ may be seen as in the general framework of Romano (1988, 1989) (Section 4.2.2), where $\boldsymbol{\theta}$ is chosen such that among all distributions in $\mathcal{G}_{\Theta_0}$, G_θ is the closest to $\hat{F}$ w.r.t. some pseudometric ρ on $\mathcal{F}_\infty$. As in Section 4.1.4 it may be expected that the choice of ρ will determine the statistical properties of the statistic in which estimator of G_θ is plugged-in. Here, we will consider only the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. The corresponding distribution $G_{\hat{\theta}}$ may also be denoted by $\hat{G}$ or $\hat{G}_\theta$. The Averaged Pearson

Divergence Statistic is now still given by Equation 5.20, but with the plug-in estimator

$$\hat{\mathrm{I}}^1\left(F; G \| [A]_{\mathcal{P}}\right) = \mathrm{I}^1\left(\hat{F}; \hat{G} \| [A]_{\mathcal{P}}\right).$$

5.3.2 The Test Statistic and its Asymptotic Null Distribution

The test statistic which was given in Section 5.2.3 (Equation 5.9) essentially remains unchanged, except that now the extended definition of the Averaged Pearson Divergence Statistic is used, resulting in

$$T_{c,n} = 2n\hat{\Delta}_c(F; G) = 2n\mathrm{E}_{\hat{F}}\left[\mathrm{I}^1\left(\hat{F}; \hat{G} \| [A](\mathcal{P}_{c-1,n}\right)\right]. \tag{5.21}$$

In the next section the asymptotic null distribution for $c = 2$ is shown to be the same as for the AD statistic. Even for the latter there is no full analytic solution available. When the distribution G_θ is the normal distribution with unknown mean and variance, Durbin et al. (1975), Stephens (1976) have given approximations. For general c the null distribution will be even more complicated. Only a general discussion is given.

Special Case: $c = 2$

When $c = 2$ we now show that the $T_{2,n}$ and the AD statistic A_n have asymptotically the same null distribution.

Theorem 5.4 *Suppose that the finite dimensional nuisance parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is estimated by a consistent estimator $\hat{\boldsymbol{\theta}}$, then under a composite null hypothesis,*

$$T_{2,n} - A_n \xrightarrow{p} 0,$$

and,

$$T_{2,n} \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j Z_j^2,$$

where the Z_j are i.i.d. standard normal, and where the λ_j are solutions to the same integral equation as for the corresponding AD statistic.

Proof. Let $\hat{G}(x) = G_{\hat{\theta}}(x)$. Similar calculations as in Section 5.2.3 show that (cfr. Equation 5.10)

$$T_{2,n} = \int_{\mathcal{S}} \frac{\left(\sqrt{n}(\hat{F}(x) - \hat{G}(x))\right)^2}{\hat{G}(x)(1 - \hat{G}(x))} d\hat{F}(x).$$

The AD statistic for a composite null hypothesis was given in Equation 4.15. In order to prove that the difference between both statistics converge in probability to zero, a Taylor series expansion of $G_{\hat{\theta}}$ about the true parameter value $\boldsymbol{\theta}$ is performed,

$$G_{\hat{\theta}}(x) = G_{\theta}(x) + \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \frac{\partial G_{\theta}}{\partial \boldsymbol{\theta}}(x) + O_p\left(\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right).$$

Then,

$$\begin{aligned} T_{2,n} - A_n &= \int_{\mathcal{S}} \frac{\left(\sqrt{n}(\hat{F}(x) - \hat{G}(x))\right)^2}{\hat{G}(x)(1-\hat{G}(x))} d\left(\hat{F}(x) - \hat{G}(x)\right) \\ &= \int_{\mathcal{S}} \frac{\left(\sqrt{n}(\hat{F}(x) - \hat{G}(x))\right)^2}{\hat{G}(x)(1-\hat{G}(x))} d\left(\hat{F}(x) - G_{\theta}(x)\right) \\ &\quad + \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \int_{\mathcal{S}} \frac{\left(\sqrt{n}(\hat{F}(x) - \hat{G}(x))\right)^2}{\hat{G}(x)(1-\hat{G}(x))} \frac{\partial g_{\theta}}{\partial \boldsymbol{\theta}}(x) \frac{1}{g_{\theta}(x)} dG_{\theta}(x) \\ &\quad + R_n. \end{aligned} \tag{5.22}$$

Under H_0, the first term in equation 5.22 converges in probability to zero by the same argument as used in the proof of Theorem 5.1. The second term consists of p terms

$$(\hat{\theta}_i - \theta_i) \int_{\mathcal{S}} \frac{\left(\sqrt{n}(\hat{F}(x) - \hat{G}(x))\right)^2}{\hat{G}(x)(1-\hat{G}(x))} \frac{\partial g_{\theta}}{\partial \theta_i}(x) \frac{1}{g_{\theta}(x)} dG_{\theta}(x),$$

($i = 1, \ldots, p$), of which, by the consistency assumption on the estimator $\hat{\theta}_i$, the factor $\hat{\theta}_i - \theta_i \xrightarrow{p} 0$. Under certain regularity conditions on G_{θ} (see Anderson and Darling, 1952, for details) the integral $\int_{\mathcal{S}} \frac{\left(\sqrt{n}(\hat{F}(x) - \hat{G}(x))\right)^2}{\hat{G}(x)(1-\hat{G}(x))} \frac{\partial g_{\theta}}{\partial \theta_i}(x) \frac{1}{g_{\theta}(x)} dG_{\theta}(x)$ is again a member of the Anderson-Darling family of statistics, which have a proper, non-degenerate asymptotic null distribution. Thus, by Slutsky's theorem, all p terms converge in probability to zero. Finally, by the same arguments also $R_n \xrightarrow{p} 0$.
The second part of the theorem follows immediately from the first part and from Theorem 4.1. □

In contrast to the results under a simple null hypothesis, the asymptotic null distribution of A_n is much more complicated under a composite null hypothesis. This was already commented upon in Section 4.2.4. Because of the identical asymptotic null distributions of the AD statistic and the SSP2 statistic, the problems arising from the estimation of the unknown parameters $\boldsymbol{\theta}$ apply to both statistics. We repeat here the most important specific problems (Stephens, 1976, 1986) (see also Section 4.2.4 for a more detailed discussion).

- In general the null distribution of A_n will depend on the distribution G_{θ} tested, the parameters $\boldsymbol{\theta}$ estimated, the method of estimation and the sample size n. Even for

large n (asymptotically) these dependencies apply.

- When the unknown parameters are those of location (e.g. mean) and of scale (e.g. variance), and when an appropriate method of estimation is used (e.g. maximum likelihood), then the null distribution of A_n does not depend on the unknown parameters.

General Case

The asymptotic null distribution of $T_{c,n}$ ($c > 2$) is expected to be far more difficult to find as in the simplest case of $c = 2$. In this section we will only show what kind of complications are involved. The asymptotic null distribution itself is not given. In a later section exact null distributions are simulated, providing a practical solution for applying the test.

One could think of extending the theory for $T_{2,n}$ to $T_{c,n}$ in the same way as it was done under a simple null hypothesis (cfr. Proposition 5.1). There, an heuristic proof by means of complete induction was given, which was later empirically confirmed in a simulation study. Basically, the complete induction argument is equivalent to the decomposition of $T_{c,n}$. The decomposition is achieved by considering the statistic as a $(c-1)$-fold integral, of which the integrandum is a Pearson χ^2 statistic on the observed frequencies that are induced by a c-sized SSP, over the space $\mathcal{S}_\mathcal{P}$ of SSP determining subsample sets. Conditionally on any SSP, Lancaster's partitioning rule applied to the Pearson statistic results in $c - 1$ asymptotically independent Pearson statistics, each with 1 degree of freedom. The asymptotic independence was crucial to arrive at the proposed limiting distribution. In the present case of a composite null hypothesis, this independence is however lost, for all terms depend on the complete sample by the imputation of the estimator $\boldsymbol{\theta}$ into G_θ. Hence, it may be expected that the asymptotic null distribution will generally depend on the method of estimation and on the true parameter value $\boldsymbol{\theta}$ as well.

Note that there is much analogy with the problems of applying a Pearson χ^2 test to grouped continuous data, discussed in Section 4.1.4 (here the grouping is induced by the SSP). There it was seen that using a maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, based on the original ungrouped data, not only led to a loss of p degrees of freedom, but it also stochastically increased the limiting null distribution with a weighted sum of p i.i.d. χ^2_1-variates (cfr. Chernoff-Lehmann statistic). It is expected that these issues also come into play in the asymptotic null distribution of $T_{c,n}$, eventually leading to a complicated limiting distribution.
One might also try to overcome some of the difficulties by changing the Pearson statistics in the $T_{c,n}$ statistic to Rao-Robson statistics, for these still have asymptotically a χ^2_{c-1} distribution as it is the case under a simple null as well.

5.3.3 Approximation of the Null Distribution

Since we only have the asymptotic null distribution of $T_{2,n}$, we well restrict the approximations to this case as well. Furthermore, the limiting distribution depends on G and the nuisance parameters that have to be estimated. We will consider here only the normal distribution with unknown mean and variance. The first 10 λ_j coefficients are approximated and tabulated by Durbin et al. (1975), Stephens (1976).

As in Section 5.2.5 an approximation of the limiting null distribution is needed. Based on the moments given in Stephens (1976), it is seen that Pearson curves cannot be used. The four-moment stochastic approximation is used once more. Thus, $W(a,b,p,f) = a+bZ_p^2+\frac{1}{2}Y_f^2$, where $Z_p^2 \sim \chi_p^2$, $Y_f^2 \sim \chi_f^2$ and the coefficients a, b, p and f are determined to match the first 4 moments of the true limiting null distribution. The solution is

$$\begin{aligned} a &= 0.1485987138 \\ b &= 0.07645034415 \\ p &= 3.081193646 \\ f &= 0.0002859431225. \end{aligned}$$

The exact asymptotic critical values at the 5% and the 10% level are 0.751 and 0.632, respectively. The corresponding approximations are 0.757 and 0.636, which are rather close approximations. In the following section the convergence is studied.

5.3.4 Convergence to the Limiting Null Distribution

In this section the validity of the asymptotic null distribution of $T_{2,n}$, which was stated in Theorem 5.4, is assessed in a similar way as in Section 5.2.6, i.e. for a range of sample sizes the exact null distributions are simulated under the null hypothesis (here: a standard normal distribution; 10000 simulation runs) and the $1-\alpha$ percentile is used as the estimator of the α-level critical value $t_{\alpha,2,n}$. Only $\alpha = 0.05$ is considered here. At the same time, the observed convergence rate will tell us whether or not the asymptotic null distribution may be used in practice. Since in the simple null hypothesis case the convergence was very slow, it may be expected to be even worse when the null hypothesis is composite.

Figure 5.9 show the estimated critical values, and the estimated sizes when the asymptotic critical value would have been used. As with the SSPc test for simple null hypotheses, the convergence is very slow, and thus the asymptotic critical values do not seem to have much practical value. Still, the results confirm Theorem 5.4.

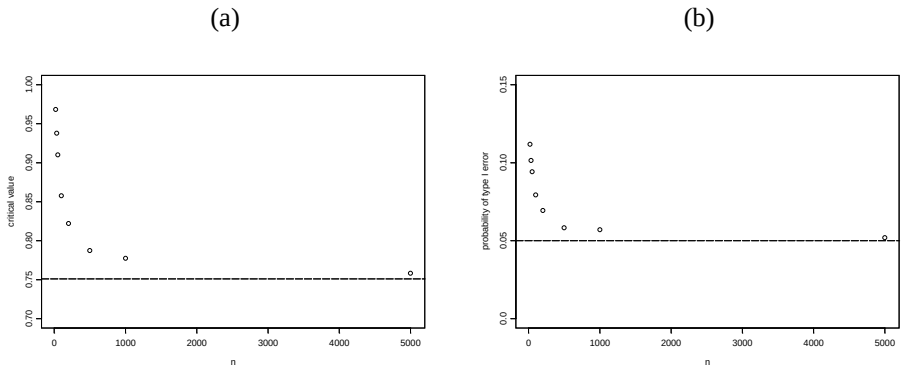


Figure 5.9: Results from the simulated exact null distributions of $T_{2,n}$ under a composite null hypothesis: (a) the estimated critical values, and (b) the estimated sizes when the asymptotic critical value would have been used.

5.3.5 Power Characteristics

In Section 5.2.7 results were presented concerning the power of the SSP2 test for a simple null hypothesis. It was argued that those results were not of much practical value since simple GOF null hypothesis do not occur often. Composite null hypothesis, on the other hand, are frequently encountered in many situations. As for all power studies for GOF tests the main problem is to make a choice among the infinite number of alternatives that can be considered. Here, I restricted the power study to the composite null hypothesis

$$H_0 : F \text{ is the CDF of a normal distribution,}$$

and the alternatives considered are the same as those used to assess the SSPc test for a simple null hypothesis (Section 5.2.7): the Tukey-Lambda family and the contaminated normal family of alternatives. Again a qualitative comparison among a set of selected tests is based on the isotones, and for a few alternatives powers are estimated by means of Monte Carlo simulations. Before the results are given, the next paragraph briefly describes the other tests that are included in this power study.

Some Other GOF Tests for Normality

For a more detailed discussion on the tests mentioned in this paragraph, we refer to Chapter 4.

Probably the best known test for normality is the Shapiro-Wilk (SW) test (originally developed by Shapiro and Wilk, 1965), which belongs to a class of tests called *regression tests*. Many extensions and modifications have been proposed over the years; a detailed overview is given by D'Agostino (1986). The SW test used in this thesis is a modification proposed by Weisberg and Binham (1975), which is as compared to the original

formulation of Shapiro-Wilk easier to calculate. The SW tests are generally recognized as the *overall most powerful* tests for normality. Critical values are obtained from Stephens (1986) (see Table 5.7).

Table 5.7: Critical values at the 5%-level for $n = 20$ and $n = 50$ for the GOF tests for (composite) normality used in this thesis. KL3 up to KL10 denote the data-driven smooth test of Kallenberg and Ledwina of order 3 up to 10.

n	SSP2	SSP3	SSP4	AD	KS	RR	SW	K
20	0.989	4.759	11.037	0.752	0.895	10.628	1.924	6.453
50	0.915	2.813	5.604	0.752	0.895	15.206	2.297	6.424
n	KL3	KL4	KL5	KL6	KL7	KL8	KL9	KL10
20	2.998	3.598	5.093	6.183	6.992	7.430	8.222	8.780
50	3.521	4.926	6.607	7.126	8.244	9.154	9.859	10.805

Among the EDF tests (see Section 4.2) the Anderson-Darling test is generally accepted to be the most powerful. In the present situation of a composite null hypothesis, the modified statistic (Stephens, 1976, 1986)

$$A_n^m = A_n(1 + \frac{0.75}{n} + \frac{2.25}{n^2})$$

is used. Critical values are obtained from Stephens (1976, 1986) and given in Table 5.7.

Although it is argued that χ^2-type tests, which are applied to grouped data, are definitely not as powerful as tests that are specifically constructed for continuous data, we included here the Rao-Robson (RR) (Rao and Robson, 1974) test. This test was described in Section 4.1.4. Since the performance of a test which is based on grouped data may depend to a large extent on the number of (equiprobable) classes, a different number of classes is used for samples of size $n = 20$ ($k = 6$ classes) and $n = 50$ ($k = 9$ classes). These numbers were calculated from a simple rule proposed by Mann and Wald (1942). Critical values were calculated from the simulated exact null distribution for $n = 20$ and $n = 50$ (50000 simulation runs). These values are shown in Table 5.7.

The D'Agostino-Pearson K test (D'Agostino and Pearson, 1973) is actually not an omnibus test because it is only sensitive to deviations from $\sqrt{\beta_1} = 0$ and $\beta_2 = 3$ ($\sqrt{\beta_1}$ and β_2 are the standardized 3rd and 4th moment, respectively). Nevertheless it is experienced as rather sensitive to many alternatives to the normal distribution. The critical values were obtained from the simulated exact null distribution for $n = 20$ and $n = 50$ (50000 simulation runs). They are shown in Table 5.7.

Recently there has been a renewed interest in smooth tests, especially in data-driven smooth tests. Very good power results were obtained by a data-driven smooth test that was proposed by (Inglot et al., 1997; Kallenberg and Ledwina, 1997), This test is in detail discussed in Section 4.3.3 and is included in this power study with a maximal order of $d_n = 10$ for both $n = 20$ and $n = 50$. It will be referred to the KL test. Its critical values were calculated from the simulated exact null distribution for $n = 20$ and $n = 50$ (50000 simulation runs). These values are shown in Table 5.7.

Finally the Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933; Smirnov, 1939; Lilliefors, 1967; Stephens, 1974) is included. Although it is generally known that this test has rather poor power characteristics, it is included here merely as an historical curiosity. The modified statistic of Stephens (1974) is used,

$$D_n^m = D_n(1 - \frac{0.01}{\sqrt{n}} + \frac{0.85}{n}).$$

Critical values are taken from (Stephens, 1974, 1986) and shown in Table 5.7.

The Power Study

Tukey-Lambda: The isotones are presented in Figure 5.10, and some power estimates are shown in Table 5.8.

- The symmetric distributions are situated along the diagonal ($\lambda_1 = \lambda_2$). Starting from the quasi-normal distribution $\lambda_1 = \lambda_2 = 0.1349$ towards negative values for λ_1, λ_2 Figure 5.10 (a) shows clearly that the powers increase very rapidly and that the isotones lay very close together which makes the tests rather similar in this direction. Among the SSP tests the powers increase from SSP4 over SSP3 to SSP2. The SSP2 test has slightly lower powers than the AD test. The SW test has the highest power in this direction. These features show for both $n = 20$ and $n = 50$.

  Since the skewness is exactly zero for symmetric distributions, the power of the K test is rather low, but still increasing because increasing $\lambda_1 = \lambda_2$ corresponds to increasing kurtosis.
  For $n = 50$ the isotones of the K test show an L-shaped region (going from $(1, \infty)$ over $(2.4, 2.4)$ to $(\infty, 1)$) of lower power. This region corresponds to members of the TL family with zero skewness.

- Again starting from the quasi-normal distribution, but now moving along positive values for $\lambda_1 = \lambda_2$, there are more differences observed between the 9 tests. The first isotone that is crossed it the one of the AD test, resulting in the highest powers in this direction.

  The isotones of the KL test in Figure 5.10 (b) ($n = 50$) show an interesting feature in the direction of positive λ_1, λ_2. It seems that the isotones converge only very slowly to the diagonal which represents symmetric distributions with truncated tails. A possible reason is that in theory the KL test is only consistent if the order d_n tends to infinity with increasing n, which is in practice however not possible (here, $d_n = 10$ for both $n = 20$ and $n = 50$). At large values for $\lambda_1 = \lambda_2$ the TL distribution has truncated tails, which is indeed a feature that is only described by high order moments.

  Also for $n = 50$, the K test shows a particular behaviour. In a small range about $\lambda_1 = \lambda_2 = 2$ it gives the highest powers, but on the other hand, for high positive values of λ_1, λ_2 its isotone is the one but last to be crossed.

- In the direction along $\lambda_2 \approx 0$ and λ_1 positive (distributions that are similar to the exponential and χ^2), all tests are rather sensitive. Again, among the SSP tests the SSP2 test is the most powerful, but still less powerful than the AD test and the SW test. Now even the KL test is more powerful than the SSP tests. By the symmetry of the Tukey-Lambda family, the same is seen in the direction of $\lambda_1 \approx 0$ and λ_2 positive.

Table 5.8: The estimated powers for some alternatives of the Tukey-Lambda family, based on 10000 simulation runs.

n	λ_1	λ_2	power								
			SSP2	SSP3	SSP4	AD	KS	RR	SW	K	KL
20	-0.5	-0.5	0.622	0.500	0.409	0.635	0.566	0.556	0.674	0.623	0.588
20	1	1	0.064	0.061	0.064	0.156	0.090	0.094	0.081	0.095	0.006
20	1.5	1.5	0.083	0.081	0.080	0.200	0.105	0.128	0.096	0.156	0.007
20	1.5	2	0.061	0.065	0.067	0.193	0.092	0.117	0.094	0.141	0.009
20	2	0	0.565	0.409	0.299	0.739	0.532	0.566	0.754	0.518	0.627
20	5	10	0.538	0.442	0.346	0.691	0.647	0.563	0.621	0.338	0.413
20	6	6	0.225	0.186	0.140	0.290	0.295	0.251	0.210	0.084	0.079
20	9	9	0.749	0.787	0.720	0.829	0.803	0.784	0.747	0.382	0.347
50	-0.5	-0.5	0.939	0.935	0.909	0.944	0.893	0.909	0.954	0.915	0.880
50	1.5	1.5	0.477	0.440	0.384	0.694	0.324	0.425	0.613	0.846	0.008
50	2	0	0.911	0.876	0.806	0.960	0.808	0.874	0.979	0.830	0.934
50	3	3	0.083	0.092	0.090	0.185	0.087	0.130	0.099	0.316	0.002
50	5	5	0.173	0.239	0.213	0.250	0.231	0.186	0.110	0.002	0.007
50	5	9	0.899	0.917	0.901	0.939	0.923	0.886	0.887	0.517	0.684
50	1	8	0.219	0.203	0.165	0.410	0.178	0.216	0.303	0.549	0.008
50	4	8.5	0.834	0.825	0.759	0.903	0.878	0.722	0.821	0.378	0.646

Contaminated Normal: The isotones are presented in Figure 5.11, and some power estimates are shown in Table 5.9.

- In the direction from $(0, 0)$ to $(0, \infty)$ first the isotones of SSP4 and SSP3 are crossed, indicating the highest powers for these two tests, both when $n = 20$ and $n = 50$. This was also observed in the case of a simple null hypothesis.
- The SSP4 and SSP3 tests remain the most powerful tests when moving in the direction from $(0, 0)$ to $(4, 0.4)$, both when $n = 20$ and $n = 50$.

Some Conclusions

The isotones as well as the estimated powers resulting from the small power studies confirm the findings and conclusions that are found in the literature: the AD test and the SW test have good overall power. The moment based K test does under many of the alternatives in the Tukey-Lambda family rather well, which is because both the 3rd and the 4th moment are highly variable with the λ_1 and λ_2 parameters indexing the family. On the other hand, many authors (e.g. Stephens, 1974) have reported that the KS test often

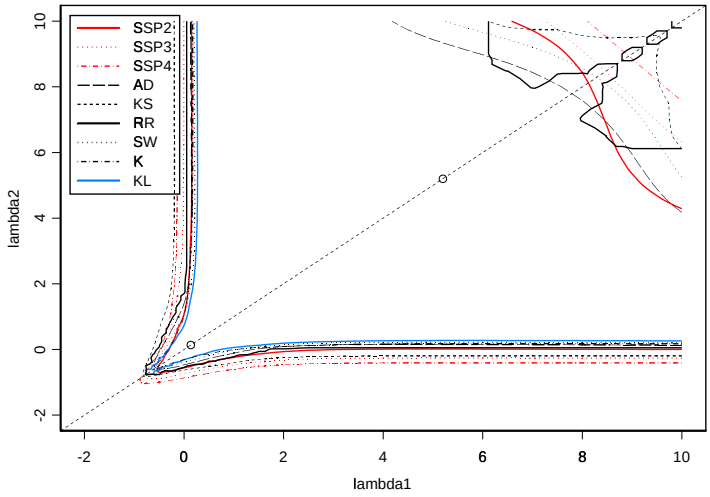
Table 5.9: The estimated powers for some alternatives of the contaminated normal family, based on 10000 simulation runs.

n	δ	γ	power								
			SSP2	SSP3	SSP4	AD	KS	RR	SW	K	KL
20	0	0.2	0.215	0.769	0.781	0.261	0.325	0.270	0.223	0.133	0.114
20	0	0.7	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.823	0.780
20	2	0.4	0.683	0.991	0.991	0.857	0.810	0.845	0.797	0.230	0.132
20	4	0.4	0.948	0.996	0.996	0.990	0.949	0.881	0.972	0.624	0.021
50	0	0.2	0.513	0.986	0.990	0.580	0.706	0.757	0.415	0.173	0.130
50	0	0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.875	0.820
50	1.25	0.38	0.993	1.000	1.000	0.997	0.993	0.999	0.990	0.541	0.632
50	3	0.4	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.938	0.054
50	3.5	0.1	0.754	0.821	0.804	0.784	0.550	0.470	0.844	0.431	0.663

results in low powers as compared to other EDF tests as e.g. the AD test. This is also seen from the present study. A similar conclusion holds for the RR test in the Tukey-Lambda family. In the contaminated normal family, however, the RR test performs rather good on average. The data-driven KL test, which is nowadays often considered to be a very sensible choice, did not perform well, especially not under symmetric alternatives with both tails truncated. Possibly the solution may be to increase the maximal order d_n.

In both families of alternatives studied, the SSP2 test, which is though rather similar to the AD test, did perform always less than the AD test. Also as compared to the other tests it is one of the worst. In the Tukey-Lambda family, the higher order SSP tests (SSP3 and SSP4) have in general even lower powers than the SSP2 test. In the contaminated normal family, on the other hand, both the SSP3 and the SSP4 test have by far the highest powers, and in many situations the difference with the best non-SSP test is extremely large. Thus the choice of the SSP size c seems very important.

(a)



(b)

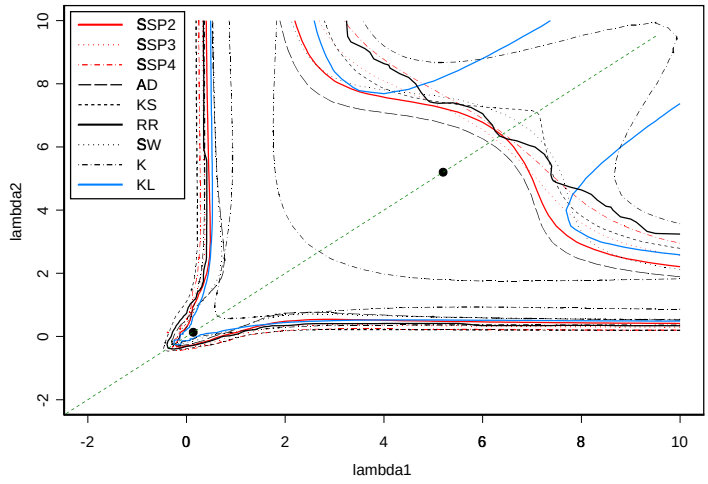
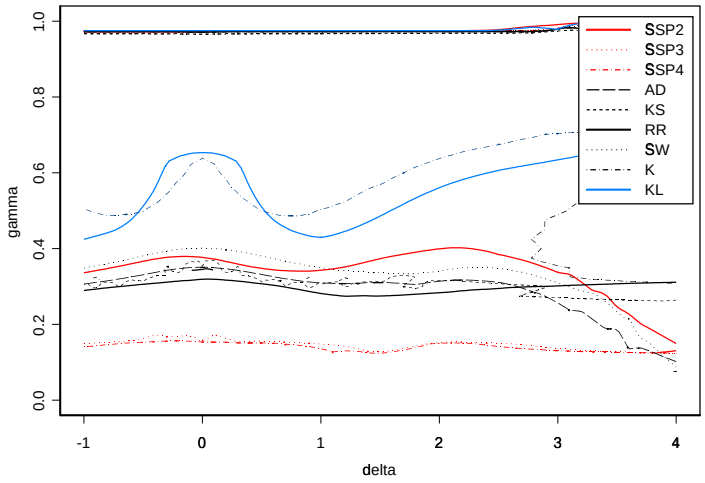


Figure 5.10: The isotones for the Tukey-Lambda family of alternatives, for $n = 20$ (a) and $n = 50$ (b). The diagonal dashed line indicates the symmetric distributions ($\lambda_1 = \lambda_2$).

(a)



(b)

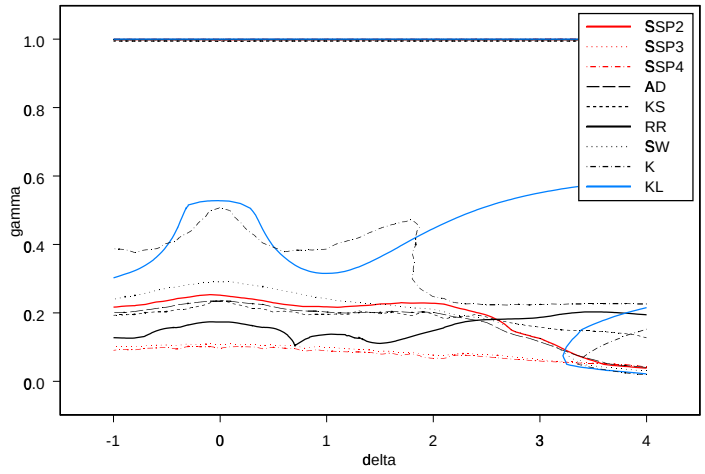


Figure 5.11: The isotones for the family of mixture distributions, for $n = 20$ (a) and $n = 50$ (b).

5.3.6 Examples

In chapter 2 some datasets were introduced. In some of them one was interested to know whether the data are normally distributed or not (composite null hypothesis). These analysis are now performed. All critical values and p-values are derived from simulated null distributions (10000 simulation runs).
Whenever tied observations occurred, they were untied by randomization (Hájek and Šidák, 1967). This is definitely not the best way to deal with ties, for different persons may end up with different values for the test statistic. However, from a theoretical point of view, all properties remain valid.

EXAMPLE 5.1. **Gravity Data**
In Chapter 2 it was seen that with the Kolmogorov-Smirnov test only the 7nd series was determined as not normally distributed ($p = 0.0451$). Here, we perform the SSPc test, with $c = 2, 3$ and $c = 4$, on the 8 series. The results are presented in Table 5.10. As with the KS test, only for the 7nd series the null hypothesis of normality is rejected, with all three partition sizes. The p-values are all smaller than 0.0451, which is the p-value of the KS test. Note also the the p-value of the SSPc tests increase as c increases. This suggests that there are cases (e.g. the same distribution as for the 7nd series, but with a smaller sample) where for a small partition size a significant result may be found, but for larger c the significance disappears. Thus, this stresses the importance of choosing an appropriate c when the sample size is moderate.

Table 5.10: Results of the SSPc test ($c = 2, 3$ and $c = 4$) on the Gravity data. The estimated p-values are presented between brackets.

c	series			
	1	2	3	4
2	0.382 (0.721)	0.412 (0.569)	0.511 (0.377)	0.584 (0.277)
3	1.150 (0.669)	1.203 (0.656)	3.455 (0.159)	1.694 (0.450)
4	2.224 (0.640)	2.280 (0.661)	7.244 (0.171)	3.029 (0.493)
c	series			
	5	6	7	8
2	0.693 (0.264)	0.384 (0.640)	3.525 (0.001)	0.544 (0.291)
3	1.715 (0.444)	0.843 (0.846)	9.061 (0.020)	1.248 (0.638)
4	2.977 (0.501)	1.354 (0.885)	15.054 (0.044)	1.948 (0.760)

□

EXAMPLE 5.2. **The Chemical Concentration Data**
The results of the SSPc test, with $c = 2, 3$ and $c = 4$ are shown in Table 5.11.

None of the SSP tests resulted in a convincing rejection of the null hypothesis. In Section 2.1 the results of some other GOF tests were mentioned. Most of them indicated a non-

Table 5.11: Results of the SSP2, SSP3 and the SSP4 test on the Chemical Concentration Data. .

c	$T_{c,n}$	p-value
2	0.710	0.110
3	1.850	0.178
4	3.056	0.290

significant deviation from normality as well, though with p-values only slightly greater than 0.05. Only a smooth test of Best and Rayner (1985) gave a significant result, in the meanwhile suggesting that most probably the true distribution is too skew to be normal. From the simulation study in Section 5.3.5 it was indeed concluded that the SSP tests performed rather bad in the Tukey-Lambda family of alternatives, which includes skew alternatives to the normal distribution as well. □

EXAMPLE 5.3. **The Singer Data**
In Section 2.2 a Kolmogorov-Smirnov, a Shapiro-Wilk and a data-driven smooth test (KL) were performed on the Singer data. All tests had p-values around 0.3 - 0.4, and thus normality was concluded.

Table 5.12 shows the results of the SSPc test ($c = 2, 3$ and $c = 4$). The results of the SSP2 test ($p = 0.375$) is in the line of the classical tests, but when larger partition sizes are used ($c = 3, 4$), one immediately notices the extreme large values for the test statistic which correspond to p-values smaller than 0.0001. Thus, both the SSP3 and the SSP4 test reject the null hypothesis.
A reason for this tremendous difference between the SSP2 test (and the classical tests) on the one hand, and the SSP3 and the SSP4 test on the other hand, might be that the data actually shows some bimodality. This feature was however not recognized in the exploratory analysis that was presented in Section 2.2 (Figure 2.2, panel (a)), even not in the kernel density estimate. The reason for this might be that the method used (Gaussian kernel with a bandwidth determined by means of the Unbiased Cross Validation method (Silverman, 1986; Venables and Ripley, 1997)) did not select a good bandwidth in the sense that the data was too much smoothed. Of course, since the dataset contains only 35 observations, bumps in the density are often flattened by smoothing. In Figure 5.12 another kernel density estimator is used: a Gaussian kernel with bandwidth manually set to 2.75. Now the density suggest some bimodallity. (Note that this method is definitely not to be recommended, since by setting the bandwidth sufficiently small, bimodallity will show in all datasets.)
Since from the simulation study in Section 5.3.5 it was concluded that the SSP3 and the SSP4 test are especially highly sensitive to contaminated normal distribution, i.e. mixture distributions showing in extreme cases bimodallity, this may the reason here for their extreme significant result.

□

Table 5.12: Results of the SSPc test ($c = 2, 3$ and $c = 4$) on the Singer data.

c	$T_{c,n}$	p-value
2	0.452	0.374
3	306.883	$<$0.0001
4	951.133	$<$0.0001

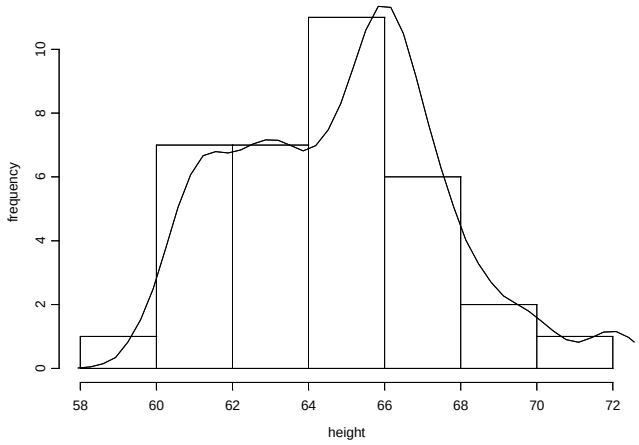


Figure 5.12: A histogram and a Gaussian kernel density estimate (bandwidth = 2.75) of the Singer Data .

5.4 Generalization of the SSPc Test to Divergences of Order λ

In Section 5.1 the Directed Divergence of order λ, and the corresponding Power Divergence Statistic of Cressie and Read (1984) were defined and discussed. Later, in Section 5.2, when first the rationale for the SSPc test was given, the construction of the SSPc test was completely restricted to the Pearson-type divergence for which $\lambda = 1$. For this specific case, the Averaged Pearson Divergence and the related Averaged Pearson Divergence Statistic (Sections 5.2.1 and 5.2.2) were defined. In this section the SSPc test is generalized to the complete family of directed divergences. First, some definitions are given. Then, the order λ SSP2 test is constructed and its null distribution is given, and finally the SSP2 test is extended to the SSPc test for arbitrary λ.

5.4.1 Some Definitions

Definition 5.8 *The* **averaged divergence of order** λ *is defined as*

$$\Delta_c^\lambda(F;G) = E_f\left[I^\lambda\left(F;G\|[A](\mathcal{P}_{c-1,n})\right]\right.\,.$$

This measure $\Delta_c^\lambda(F;G)$ is again only useful if it measures the deviation of F from G in some sense. This is shown in the following lemma, which is the analogue of Lemma 5.1.

Lemma 5.3 *For all* $c > 1$, *and* $\lambda \in \mathbb{R}$

$$\Delta_c^\lambda(F;G) = 0 \Leftrightarrow H_0 \textit{ is true}.$$

Proof. If H_0 is true, then, following Corollary 5.1, for all $[A]$ $\mathrm{I}^\lambda\left(F;G\|[A]\right) = 0$, and hence $\Delta_c^\lambda(F;G) = 0$.
The proof of "$\Rightarrow$" is similar to the proof of Lemma 5.1; only here $\Delta_c^\lambda(F;G) = 0$ implies

$$\frac{1}{\lambda(\lambda+1)}\int_{\mathcal{S}_\mathcal{P}}\sum_{i=1}^{c}\mathrm{P}_f\left[A_i(\mathcal{P})\right]\left[\left(\frac{\mathrm{P}_f\left[A_i(\mathcal{P})\right]}{\mathrm{P}_g\left[A_i(\mathcal{P})\right]}\right)^\lambda - 1\right]dF_{\mathcal{P}_{c-1}}(\mathcal{P}) = 0,$$

which for every λ implies, as in the proof of Lemma 5.1,

$$\mathrm{P}_f\left[A_i(\mathcal{P})\right] = \mathrm{P}_g\left[A_i(\mathcal{P})\right]$$

almost everywhere. The remainder of the proof remains unchanged. □

The usefulness of the extension to arbitrary λ does not only arise from Lemma 5.3, but also from the general philosophy of Romano's framework for GOF tests (Section 4.2.2). There it is suggested that the choice for a certain GOF test is partly determined by the choice for a measure (pseudo-metric) δ that measures the deviation of F from G. Δ_c^λ represents a family of such measures $\delta = \delta_{\gamma,c}$, indexed by both c and λ. By changing λ from 1 to e.g. 0, the pseudo-metric is changed from typical Pearsonian to a measure based on the Kullback-Leibler divergence.

With the same notation as in Section 5.2.2, the definition of the Averaged Pearson Divergence Statistic (Equation 5.8) is now extended to the *Averaged Divergence Statistic of order* λ, given by

$$\hat{\Delta}_c^\lambda(F;G) = \frac{1}{a_{c-1,n}}\sum_{\mathcal{P}\in\mathbb{P}_{c-1,n}}\hat{\mathrm{I}}^1\left(F;G\|[A]_\mathcal{P}\right), \tag{5.23}$$

which is just the plug-in estimator of $\Delta_c^\lambda(F;G)$. Note that the statistic can be both interpreted in terms of a simple null hypothesis (as its definition in Section 5.2.2) as in terms of a composite null hypothesis (as its definition in Section 5.3.1).

5.4.2 The SSPc-test of order λ and its Distribution under a Simple Null

The test statistic is obviously generalized to

$$T_{c,n}^{\lambda} = 2n\hat{\Delta}_c^{\lambda}(F;G) = 2n\mathrm{E}_{\hat{F}}\left[\mathrm{I}^{\lambda}\left(\hat{F};G\|[A](\mathcal{P}_{c-1,n}\right)\right]. \tag{5.24}$$

Its asymptotic null distribution is first derived for the special case $c = 2$. Only simple null hypotheses are considered here, but it will be apparent that it may be extended to the composite null hypothesis case.

Special Case: $c = 2$

For $c = 2$, recall that, as in Section 5.2.3, the partition determining subsample set $\mathbb{P}_{1;n}$ is just the sample $\mathcal{S}_n$, and the sets $\mathcal{P}_{1;n}^i$ may be replaced by the corresponding observations X_i ($i = 1, \ldots, n$). Thus, the partitions $[A](\mathcal{P}_{1;n}^i)$ may be replaced by $[A](X_i)$ or simply $[A]_i$. For a given sample $a_{1;n} = n$ such partitions can be constructed. Each SSP $[A]_i$ consists of two elements, denoted by $A_{i1} = [b_l, X_i]$ and $A_{i2} =]X_i, b_u]$. Then, when the null hypothesis is simple, the test statistic reduces to

$$\begin{aligned} T_{2,n}^{\lambda} &= 2n\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\mathrm{I}}^{\lambda}\left(F;G\|[A]_i\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} n\frac{1}{\lambda(\lambda+1)}\left\{\hat{F}(X_i)\left[\left(\frac{\hat{F}(X_i)}{G(X_i)}\right)^{\lambda}-1\right]\right. \\ &\quad \left.+(1-\hat{F}(X_i))\left[\left(\frac{1-\hat{F}(X_i)}{1-G(X_i)}\right)^{\lambda}-1\right]\right\}. \end{aligned}$$

When $\lambda = 0$ or $\lambda = -1$, $T_{2,n}^{\lambda}$ is defined by continuity. Since $\lambda = 0$ is an interesting special case (based on the Kullback-Leibler divergence and the related likelihood ratio statistic), we give the statistic here explicitly:

$$\begin{aligned} T_{2,n}^{0} &= \frac{1}{n}\sum_{i=1}^{n} 2n\left(\hat{F}(X_i)\ln\frac{\hat{F}(X_i)}{G(X_i)} + (1-\hat{F}(X_i))\ln\frac{1-\hat{F}(X_i)}{1-G(X_i)}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} 2n\ln\left[\left(\frac{1-\hat{F}(X_i)}{1-G(X_i)}\right)^{1-\hat{F}(X_i)}\left(\frac{\hat{F}(X_i)}{G(X_i)}\right)^{\hat{F}(X_i)}\right], \end{aligned}$$

which clearly is based on the log likelihood ratio statistics applied to the multinomial distributions that arise from the SSPs $[A](X_i)$ ($i = 1, \ldots, n$).

An important and very interesting property of the statistics $T_{2,n}^{\lambda}$ is that for each λ their asymptotic null distribution is the same as for the AD statistic, and thus independent of λ.

Theorem 5.5 *Under a simple H_0, for each $\lambda \in \mathbb{R}$,*

$$T_{2,n}^{\lambda} - T_{2,n}^{1} \xrightarrow{p} 0,$$

and

$$T_{2,n}^{\lambda} \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2,$$

where the Z_j^2 are i.i.d. χ_1^2 variates.

Proof. The proof is based on von Mises calculus (von Mises, 1947) and influence functions (for a recent account, see Serfling, 1980; van der Vaart, 1998).

The test statistic $T_{2,n}^{\lambda}$ may be looked upon as a plug-in estimator of the functional $2n\Delta_2^{\lambda}(F;G)$, i.e.

$$T_{2,n}^{\lambda} = 2n\Delta_2^{\lambda}(\hat{F}_n;G) = 2n\int_{\mathcal{S}} \mathrm{I}^{\lambda}\left(\hat{F}_n;G\|[A](x)\right) d\hat{F}_n(x).$$

A Taylor-like expansion of $\Delta_c^{\lambda}(\hat{F}_n;G)$, based on Hadamard derivatives is

$$\begin{aligned}\Delta_2^{\lambda}(\hat{F}_n;G) &= \Delta_2^{\lambda}(F;G) + \int_{\mathcal{S}} IF_1^{\lambda}(y;F;G)\left(d\hat{F}_n(y) - dF(y)\right) dy \\ &\quad + \frac{1}{2}\int_{\mathcal{S}}\int_{\mathcal{S}} IF_2^{\lambda}(y,z;F;G)\left(d\hat{F}_n(y) - dF(y)\right)\left(d\hat{F}_n(z) - dF(z)\right) dydz \\ &\quad + R_n, \end{aligned} \tag{5.25}$$

where the influence functions IF_1^{λ} and IF_2^{λ} are given by

$$\begin{aligned} IF_1^{\lambda}(y;F;G) &= \frac{d}{d\epsilon}\Delta_2^{\lambda}((1-\epsilon)F + \epsilon\delta_y;G)\Big|_{\epsilon=0} \\ IF_2^{\lambda}(y,z;F;G) &= \frac{d}{d\gamma} IF_1^{\lambda}(y;(1-\gamma)F + \gamma\delta_z;G)\Big|_{\gamma=0}, \end{aligned}$$

where $\delta_y = \delta_y(x) = \mathrm{I}[y \le x]$ and $\delta_z = \delta_z(x) = \mathrm{I}[z \le x]$.

It is straightforward to show that under H_0 the first order influence function is zero for any λ, i.e. $IF_1^{\lambda}(y;F;G) = 0$. Also the first term in Equation 5.25 vanishes under H_0 (Lemma 5.3). After some cumbersome algebra it is shown that the second order influence

function becomes, under H_0,

$$\begin{aligned} IF_2^{\lambda}(y,z;F;G) &= \int_{\mathcal{S}} \frac{(\delta_y - G(x))(\delta_z - G(x))}{G(x)(1-G(x))} dG(x) \\ &= -\ln(1-\min(G(y),G(z))) - \ln(\max(G(y),G(z))) - 1, \end{aligned}$$

which is independent of λ. Thus, under H_0, the test statistic $T_{c,n}^{\lambda}$ has under H_0 the same limiting distribution as

$$V_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} IF_2^{\lambda}(X_i,X_j;F;G) \tag{5.26}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(-\ln(1-\min(G(X_i),G(X_j))) - \ln(\max(G(X_i),G(X_j))) - 1\right)$$

$$= -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\ln(X_{(i)}) + \ln(1-X_{(n+1-i)})\right), \tag{5.27}$$

where $X_{(1)} \leq \ldots \leq X_{(n)}$ are the order statistics of $\mathcal{S}_n$.
From Equation 5.27 it is concluded that for any λ the statistic $T_{2,n}^{\lambda}$ has asymptotically the same null distribution as the AD statistic (cfr. Equation 4.13), provided that $nR_n \xrightarrow{p} 0$. The latter condition may be easily checked. This proves the first part of the theorem. The second part follows now immediately from Theorem 5.1. □

For the SSPc test with $\lambda = 1$ it was observed that the convergence of the null distribution to its limiting distribution is very slow (Section 5.2.6). The same behaviour is to be expected here as well. Moreover, for finite n the exact null distributions will still depend on λ. This characteristic is motivated by recognizing that the SSPc statistic is the average of the power divergence statistics applied to the grouped data resulting from all c-sized partitions in $\mathcal{A}$. Conditional on a partition the corresponding power divergence statistic converges weakly to a χ^2-distribution, irrespective of λ (Cressie and Read, 1984), but for finite n the exact null distributions may be quiet different for various values of λ (Read, 1984). This property will be propagated to the SSPc statistic. The practical solution is thus to use the λ-specific simulated exact null distribution.

Before proceeding to the general case, a few notes, which follow from the proof of Theorem 5.5, may be in place here. From equation 5.24 it may be seen that the statistic $T_{2,n}^{\lambda}$ is an expectation functional. Plug-in estimators of such estimators may be considered as V-statistics (von Mises, 1947; Lee, 1990). Here, however, the theory of V-statistics was not applied because of the complexity of the (asymmetric) kernel that follows from Equation 5.24. Instead, we made immediate use of the von Mises calculus, which eventually has led to an asymptotic equivalence with Equation 5.26, which now shows a V-statistic with a proper symmetric kernel of degree 2. Since the statistic was directly recognized as the AD statistic, no further steps had to be made to solve the limiting null distribution. Though, if proceeded from Equation 5.24, it would have turned out that the V-statistic shows a first order degeneracy. By the equivalence relation between V-statistics with

first order degeneracy and stochastic processes (see e.g. De Wet and Randles, 1984), one would immediately end up with the same integral equation of Anderson and Darling (1952).
Yet another consequence is that the asymptotic distribution of $T_{2,n}^{\lambda}$ under any fixed alternative, for which the degeneracy vanishes, is normal.

General Case

The generalization to arbitrary, but finite c involves the same conjecture as in Proposition 5.5. Therefore we state the resulting limiting distribution here also only as a proposition. In the next section its validity is empirically assessed.

Proposition 5.2 *For any $c > 1$ and any λ, under the simple GOF null hypothesis,*

$$T_{c,n}^{\lambda} \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2,$$

where the Z_j^2 are i.i.d. χ_{c-1}^2 variates.

Heuristic Proof. By Theorem 5.5 the asymptotic null distribution of $T_{2,n}^{\lambda}$ does not depend on λ and is thus exactly the same as the limiting null distribution of $T_{2,n}$. We now proceed as in Proposition 5.1, where $\lambda = 1$. The complete induction reasoning that was invoked there, implied that, conditionally on any SSP, the statistic may be written as an integral of $c - 1$ asymptotically independent Pearson statistics. From an asymptotic expansion of the power divergence statistic used by Cressie and Read (1984, p. 442), it follows that the asymptotic independence continuous to hold for arbitrary λ. This would leave the remaining steps unchanged. □

5.4.3 Convergence to the Limiting Distribution

In this section the convergence of the null distribution of $T_{c,n}^{\lambda}$ is empirically investigated. When $\lambda = 1$ this is done already in Sections 5.2.6 and 5.3.4 for the simple and the composite null hypothesis case, respectively.
Here we will consider only the distribution of $T_{2,n}^{0}$ and $T_{3,n}^{0}$ for a wide range of sample sizes. The exact null distributions are estimated by means of 10000 simulation runs. The results are presented in Figures 5.13 and 5.14 for $c = 2$ and $c = 3$, respectively.

From the figures it is concluded that the convergence to the asymptotic critical value is much faster as compared to the Pearsonian statistic, for both $c = 2$ and for $c = 3$. An explanation for the SSP sizes greater than 2 might be found in the following. In the outline of the proof of Propositions 5.1 and 5.2 the limiting null distribution results from a decomposition of power divergence statistics of order 1 and arbitrary λ, respectively. Although asymptotically the decomposition is indeed correct for arbitrary λ, an exact

decomposition for finite sample sizes only occurs if $\lambda = 0$ (the likelihood ratio statistic). In this sense one may expect that the $T^0_{c,n}$ statistic converges faster than $T^1_{c,n}$.

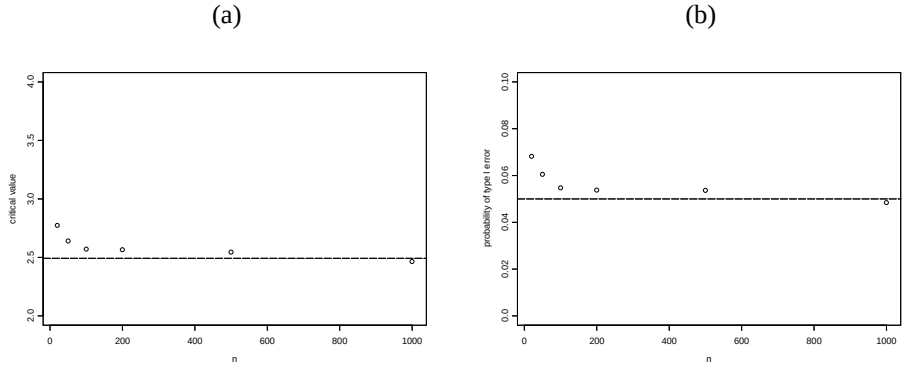


Figure 5.13: Results from the simulated exact null distributions of $T^0_{2,n}$ under a simple null hypothesis: (a) the estimated critical values, and (b) the estimated sizes when the asymptotic critical value would have been used.

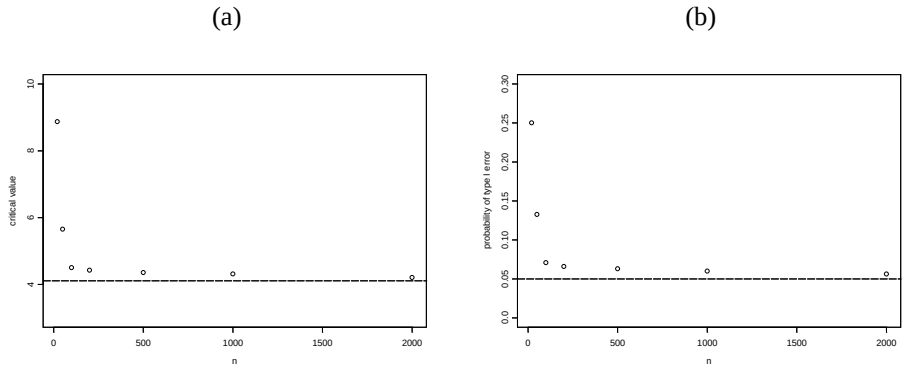


Figure 5.14: Results from the simulated exact null distributions of $T^0_{3,n}$ under a simple null hypothesis: (a) the estimated critical values, and (b) the estimated sizes when the asymptotic critical value would have been used.

5.4.4 A Small Simulation Study

When $\lambda = 0$, the power divergence statistic $\hat{\mathrm{I}}^1(F; G\|[A])$ is the likelihood-ratio (LR) GOF test for a multinomial distribution. A small simulation study is performed to compare the powers of the SSPc tests of orders $\lambda = 1$ (Pearsonian or P-type) and $\lambda = 0$ (LR-type). The critical values of the latter test are approximated by simulating the null distributions (10000 simulation runs). For estimating the powers, the simulations are run under exactly the same conditions as in Section 5.3.5.

Results

The results are shown in Tables 5.13 and 5.14 for the Tukey-Lambda and the contaminated normal families, respectively.

Table 5.13: The estimated powers of SSPc tests of order $\lambda = 1$ and $\lambda = 0$ for some alternatives of the Tukey-Lambda family, based on 10000 simulation runs.

n	λ_1	λ_2	$\lambda = 1$			$\lambda = 0$		
			SSP2	SSP3	SSP4	SSP2	SSP3	SSP4
20	-0.5	-0.5	0.622	0.500	0.409	0.667	0.665	0.645
20	1	1	0.064	0.061	0.064	0.129	0.117	0.117
20	1.5	1.5	0.083	0.081	0.080	0.147	0.143	0.148
20	1.5	2	0.061	0.065	0.067	0.142	0.136	0.136
20	2	0	0.565	0.409	0.299	0.738	0.686	0.649
20	5	10	0.538	0.442	0.346	0.685	0.694	0.686
20	6	6	0.225	0.186	0.140	0.309	0.356	0.363
20	9	9	0.749	0.787	0.720	0.851	0.910	0.915
50	-0.5	-0.5	0.939	0.935	0.909	1.000	1.000	1.000
50	1.5	1.5	0.477	0.440	0.384	0.569	0.549	0.506
50	2	0	0.911	0.876	0.806	0.997	0.996	0.993
50	3	3	0.083	0.092	0.090	0.131	0.121	0.104
50	5	5	0.173	0.239	0.213	0.283	0.347	0.342
50	5	9	0.899	0.917	0.901	0.942	0.948	0.944
50	1	8	0.219	0.203	0.165	0.368	0.354	0.317
50	4	8.5	0.834	0.825	0.759	0.876	0.862	0.845

Table 5.14: The estimated powers of SSPc tests of order $\lambda = 1$ and $\lambda = 0$ for some alternatives of the contaminated normal family, based on 10000 simulation runs.

n	δ	γ	$\lambda = 1$			$\lambda = 0$		
			SSP2	SSP3	SSP4	SSP2	SSP3	SSP4
20	0	0.2	0.215	0.769	0.781	0.276	0.520	0.582
20	0	0.7	1.000	1.000	1.000	1.000	1.000	1.000
20	2	0.4	0.683	0.991	0.991	0.745	0.917	0.940
50	0	0.2	0.513	0.986	0.990	0.593	0.877	0.911
50	0	0.6	1.000	1.000	1.000	1.000	1.000	1.000
50	3.5	0.1	0.754	0.821	0.804	0.837	0.782	0.737

The results obtained under the Tukey-Lambda family of alternatives are very clear in suggesting that the tests of the LR-type are consistently more powerful than the tests of the P-type. In some occasions the power advantage may be as large as 10%. This conclusion holds for all three partition sizes $c = 2, 3$ and $c = 4$. Moreover, when comparing the results of this section with the estimated powers of other GOF tests that were computed under the same TL-alternatives (Table 5.8) and of which many outperformed the SSPc

(P-type) test, it may now be concluded that the powers of the LR-type SSPc test are much closer to the powers of the AD and SW tests, which were overall the most powerful tests. Even in some circumstances the LR-type SSPc tests did better (e.g. large values of $\lambda_1 = \lambda_2$, i.e. unimodal densities with both tails truncated).

Under the contaminated normal alternatives, only the SSP2 LR-type test has a higher power than the corresponding P-type. When the partition size is larger, the LR-type seem to have a slightly smaller power. These larger partition sizes resulted under the contaminated normal alternatives to extremely large powers as compared to other tests (Section 5.3.5). Even the powers of the LR-type are still much higher than those of the other tests.

Conclusion

In general the powers of the SSPc tests depend on the order λ. This was expected since the core of the statistic, the power divergence statistic of order λ, also has as finite sample power function that depends on λ (Cressie and Read, 1984; Read, 1984). It is expected that other choices of λ might result in still higher powers and others in lower powers under different alternatives.
For the TL-family with the LR-type SSPc tests powers comparable to the AD and SW tests were found.

5.5 Data-Driven SSP Tests

In the previous sections the SSPc test was introduced. Actually a whole family of tests was constructed, indexed by the *order* or the *partition size* c, which had to be chosen by the user. This may be considered as a property that makes the SSP methodology flexible in the sense that by choosing an appropriate partition size c the researcher can make the test more sensitive towards alternatives which he is particularly interested in. Indeed, in Section 5.3.5 it was clearly illustrated that the power of the SSPc test did depend to a sometimes large extent on the partition size c. E.g. when a researcher is especially interested in detecting a small contamination in a sample from a normal distribution, he might want to take $c = 3$ or $c = 4$, but when he is merely interested in detecting asymmetry in the distribution then a better choice would be $c = 2$. From another point of view, though, this degree of freedom may be considered as a weakness. Indeed, many researchers are ignorant w.r.t. possible alternatives and they will apply a GOF test as a true omnibus test, i.e. when the distribution is not equal to the specified distribution G then they want to detect it, whatever the true distribution is. In this situation it would be extremely uncomfortably and undesirable if one still had to make a choice for c.

From a theoretical point of view, however, any c results in a consistent test. This means that if the sample size would be infinitely large, any SSPc test would have a power equal to 1 against essentially any fixed alternative, and thus there is no difference between the choices for c anymore. Of course, this property has no practical meaning since all samples

are finite.

In this section a solution to the problem of choosing an appropriate c will be given. The idea is to estimate c from the data itself, by means of some Schwartz-like *selection rule* (Schwartz, 1978). Such a procedure is called **data-driven**: the data suggests the right choice for c according to some criterion. Ledwina (1994) introduced this methodology for selecting the number of components in Neyman's smooth GOF test (see Section 4.3 for an overview). Note that such a test (KL test of Kallenberg and Ledwina (1997)) was already included in the Monte Carlo experiments in Section 5.3.5. Before continuing two important differences with the data-driven Neyman's smooth test are briefly discussed.

- A first important difference is the formal framework in which the selection rule is defined. Ledwina (1994) proposed to use a Schwartz (1978) selection rule, which is a rule to find the right member in (or, dimension of) an exponential family. When the exponential family is parameterized with a s dimensional parameter $\boldsymbol{\theta}$, the setting of some of the components of $\boldsymbol{\theta}$ to zero is equivalent to constructing an embedded lower-dimensional exponential family of distributions or models. When all components are set to zero, the resulting model is the model G specified under the GOF null hypothesis. The selection rule itself is essentially selecting the smallest dimension S that maximizes the maximum log-likelihood reduced by a penalty which is proportional to the dimension and the logarithm of the sample size n, over all models of dimension not greater than k, i.e.

$$S = \min \left\{ j, 1 \leq j \leq k : \mathcal{L}_j = \max_{1 \leq t \leq k} \mathcal{L}_t \right\},$$

  where $\mathcal{L}_t = \sup_{\theta^t} \{\sum_{i=1}^n \ln f_{\theta^t}(X_i)\} - \frac{1}{2} t \ln n$, where $f_{\theta^t}(X_i)$ is the density, evaluated in observation X_i, of the t-dimensional exponential family. The data-driven test is then basically a score test within the S-dimensional exponential family. For practical purposes Kallenberg and Ledwina (1997) note that the maximized log-likelihood in the selection rule may be replaced by the likelihood ratio statistic of $H_0 : \boldsymbol{\theta}^s = \mathbf{0}$ against $H_1 : \boldsymbol{\theta}^s \neq \mathbf{0}$ in the exponential family.

  The SSPc test, on the other hand, is not constructed within the likelihood framework of an exponential family. There was even no need at all to assume or parameterize a family f_θ of distributions. Therefore it would not be straightforward to apply Schwartz's selection rule, which is explicitly based on the likelihood of the exponential family.

- A central idea in Neyman's framework is that his exponential family essentially includes all continuous distributions as the dimensionality goes to infinity. If, however, the dimension is kept finite, then it is not guaranteed that the true distribution F is embedded in the family, and then the score test, and thus also the data-driven test, is not necessarily consistent. Moreover, in practice it is infeasible to use the infinite dimensional exponential model. To overcome this problem Ledwina proposed to make the maximal dimension k sample size dependent: $k = d_n$, where d_n increases with n according to some specified rate to guarantee consistency.

As will become clear in the following sections, the data-driven SSP test that will be constructed here, does not suffer from this drawback. Since for every $c \geq 2$ the corresponding SSPc test is consistent it will be easy to show that also the data-driven SSP test is consistent, whatever the choice for the maximal partition size. Thus, by making the SSPc test data-driven it is only hoped to improve the power of the test and to make it simpler for users who are ignorant to specific alternatives.

5.5.1 The Selection Rule

First we will introduce the selection rule that will be used to build the data-driven SSP test. Since the rule is not clearly related to a likelihood of an exponential model, one might argue that the scale on which the penalty is applied is incorrect to give meaningful results. In an heuristic explanation it will be shown that this scale is relevant. Moreover, it will turn out that for the validity of the data-driven SSP test, one is very free in choosing a penalty, even on completely different scales as Schwartz' BIC penalty.

The Selection Rule

Definition 5.9 *The set Γ of* **Permissible Orders** *or* **Permissible Partition Sizes** *is a set of values for the partition size c that may be selected by the selection rule. It is further supposed that $\#\Gamma$ is finite and that all partition sizes $c \in \Gamma$ are finite as well.*

Definition 5.10 *The* **Minimal Order** c_m *or* **Minimal Partition Size** c_m *is the smallest partition size that may be selected, i.e.* $c_m = \min \Gamma$.

Definition 5.11 *The* **SSP Selection Rule***, which selects the "right" partition size C_n, is given by*

$$C_n = \mathit{ArgMax}_{c \in \Gamma} \left[T_{c,n} - 2(c-1) \ln a_n \right], \tag{5.28}$$

where $2(c-1)\ln a_n$ is called the **penalty**.

Note that the selected order C_n is a random variable. Further, special cases include the Bayesian Information Criterion (BIC) (Schwartz, 1978) when $a_n = n^{1/2}$, and Akaike's Information Criterion (AIC) (Akaike, 1973, 1974) when $a_n = e$. Many other choices for a_n are possible, see e.g. Hannan and Quinn (1979), Haughton, Haughton and Izenman (1990).

An Heuristic Construction of the Selection Rule

To see why the penalty term in the selection rule is on the right scale w.r.t. $T_{c,n}$ when e.g. the BIC is used, we give a brief informal overview of the construction of the selection

rule.

Instead of considering an exponential model for the true continuous distribution f, as it is done for the construction of Neyman's test, we will construct an exponential *working model* for each partition $[A]_{\mathcal{P}} = [A](\mathcal{P}_{c-1,n})$. For a given partition $[A]_{\mathcal{P}} = \{A_1, \ldots, A_c\}$, the working model is given by (see e.g. Kopecky and Pierce, 1979)

$$f_{\theta,\phi}(x) = g_\theta(x) \exp\left\{ \sum_{j=1}^{c-1} \phi_j \mathrm{I}\left[x \in A_j\right] - K(\boldsymbol{\phi}) \right\}, \tag{5.29}$$

where $\boldsymbol{\theta}$ is the vector of nuisance parameters, which are to be estimated in case of a composite null hypothesis, and where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{c-1})'$ is the vector of parameters that model the deviation between the multinomial distribution implied by $[A]_{\mathcal{P}}$ under H_0 and the multinomial under the true distribution f. $K(\boldsymbol{\phi})$ is the normalization constant. It is important to note that $\boldsymbol{\phi} = 0$ corresponds to the original GOF null hypothesis, and that under this hypothesis $K(\boldsymbol{\phi}) = 0$.
Further, the parameterization of the working model is such that setting $d < c - 1$ components of $\boldsymbol{\phi}$ equal to zero, the resulting embedded model is the working model that corresponds to a $c - d$ sized SSP.

When $\boldsymbol{\theta}$ is known, the score test for testing $H_0 : \boldsymbol{\phi} = \mathbf{0}$ is simply Pearson's χ^2 test (Kopecky and Pierce, 1979, used this model to illustrate that Pearson's χ^2 test is actually also a smooth test). Moreover, even when the null hypothesis is composite, the SSPc test statistic is just the average of all these Pearson χ^2 statistics, with the nuisance parameters $\boldsymbol{\theta}$ replaced by their restricted maximum likelihood estimators $\hat{\boldsymbol{\theta}}$. The log-likelihood of the working model in Equation 5.29 is given by

$$l(\theta, \phi | \mathcal{S}_n) = \sum_{i=1}^{n} \ln g_\theta(x_i) + \sum_{i=1}^{n} \left(\sum_{j=1}^{c-1} \phi_j \mathrm{I}\left[x_i \in A_j\right] - K(\boldsymbol{\phi}) \right). \tag{5.30}$$

It is easy to see now that there is an estimation orthogonality between the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ in the working model, which implies here that for any partition $[A]_{\mathcal{P}}$ the estimators $\hat{\boldsymbol{\theta}}$ are the same and, furthermore, that these estimators are also the same as the restricted maximum likelihood estimators in the distribution g (cfr. the first term in Equation 5.30).

In the present context a selection rule will be applied to the working model, still for a given partition $[A]_{\mathcal{P}}$, in order to determine how many of the components of $\boldsymbol{\phi}$ may be set equal to zero, say d components are to be set to zero. Then, the selection dimension is $c - d$, which corresponds exactly to a multinomial distribution implied by a $c - d$ sized SSP.

Typically Schwartz's selection rule discriminates models based on the ordering of ($d =$

$1, \ldots, c-1$)

$$\mathcal{L}_d = \sup_{\theta, \phi^d} \left\{ \sum_{i=1}^{n} \ln f_{\theta,\phi^d}(X_i) \right\} - \frac{1}{2}(c-d)\ln n, \tag{5.31}$$

where ϕ^d denotes the ϕ vector with d components restricted to zero. The supremum in Equation 5.31 is the log-likelihood of the working model, evaluated at the maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\phi}^d$, i.e.

$$\sup_{\theta, \phi^d} \left\{ \sum_{i=1}^{n} \ln f_{\theta,\phi^d}(X_i) \right\} = l(\hat{\boldsymbol{\theta}}, \hat{\phi}^d | \mathcal{S}_n).$$

Since the log-likelihood $l(\hat{\boldsymbol{\theta}}, \mathbf{0}|\mathcal{S}_n)$, which corresponds to $\phi^d = \mathbf{0}$ (and thus also $\phi = \mathbf{0}$), is the same for every d, it may be subtracted from $\mathcal{L}_d$ without changing the order. Also multiplying the log-likelihood terms and the penalty in $\mathcal{L}_d$ by 2 will not alter the order. The new $\mathcal{L}_d$ then becomes

$$\mathcal{L}_d = 2\left[l(\hat{\boldsymbol{\theta}}, \hat{\phi}^d | \mathcal{S}_n) - l(\hat{\boldsymbol{\theta}}, \mathbf{0}|\mathcal{S}_n) \right] - (c-d)\ln n,$$

where $2\left[l(\hat{\boldsymbol{\theta}}, \hat{\phi}^d | \mathcal{S}_n) - l(\hat{\boldsymbol{\theta}}, \mathbf{0}|\mathcal{S}_n) \right]$ is the log likelihood ratio statistic for testing $H_0 : \phi^d = \mathbf{0}$ against $H_1 : \phi^d \neq \mathbf{0}$, in the working model. This test statistic is in the notation used throughout this thesis denoted as $\hat{\mathrm{I}}^0\left(f; g \| [A]_{\mathcal{P}(c-d-1,n)}\right)$. Thus,

$$\mathcal{L}_d = \hat{\mathrm{I}}^0\left(f; g \| [A]_{\mathcal{P}(c-d-1,n)}\right) - (c-d)\ln n.$$

All what is presented yet in this section is based on one SSP $[A]_{\mathcal{P}}$. The SSPc statistic $T_{c,n}$, however, is based on all such c-sized partitions $[A]_{\mathcal{P}} \in \mathcal{A}_{c-1,n}$ ($a_{c-1,n} = \#\mathcal{A}_{c-1,n}$ such partitions). A dimension reduction from c to $c-d$ will imply an change of $a_{c-1,n}$ to $a_{c-d-1,n}$ sample space partitions. Note that this change corresponds to the multiplicity of choosing d out of c components of ϕ to be set to zero.

For the construction of the data-driven test, we will suppose a common *right* dimension for all working models that can be constructed. Therefore it seems appropriate to use the average $\bar{\mathcal{L}}_d$ of all $a_{c-d-1,n}$ statistics $\mathcal{L}_d$, resulting in

$$\bar{\mathcal{L}}_d = T^0_{c-d,n} - (c-d)\ln n.$$

Since under H_0 all power divergence statistics of order λ_1 and λ_2 converge to each other in probability (Cressie and Read, 1984), or, as a consequence, since all $T^{\lambda}_{c-d,n}$ have the same limiting null distribution, it is argued here that all members of the family of SSPc statistics of order λ actually measure deviations from H_0 on a similar scale. Therefore, it seems reasonable to allow the application of a BIC-like penalty to all these members.

5.5.2 The Test Statistic and its Asymptotic Null Distribution

The test statistic of the data-driven SSP test is given by

$$T_{C_n,n}.$$

The asymptotic distribution of the statistic $T_{C_n,n}$ under a simple null hypothesis is now derived in two steps.

Theorem 5.6 *Let Γ be a set of permissible orders. Let c_m denote the Minimal Partition Size. Suppose that $a_n \to \infty$ as $n \to \infty$. Then, under H_0*

$$C_n \xrightarrow{p} c_m$$

as $n \to \infty$.

Proof.

$$\mathrm{P}\left[C_n \neq c_m\right] \quad = \quad \sum_{c \in \Gamma \setminus \{c_m\}} \mathrm{P}\left[C_n = c\right].$$

Next we make use of the characteristic that a selected order equal to c implies that the order c *beats* the minimal order c_m, and hence $T_{c,n} - 2(c-1)\ln a_n > T_{c_m,n} - 2(c_m - 1)\ln a_n$. Let $d = (c-1) - (c_m - 1)$. Then,

$$\begin{aligned}
\mathrm{P}\left[C_n \neq c_m\right] &\leq \sum_{c \in \Gamma \setminus \{c_m\}} P\left[T_{c,n} - 2(c-1)\ln a_n \right. \\
&\quad \left. > T_{c_m,n} - 2(c_m - 1)\ln a_n\right] \\
&\leq \sum_{c \in \Gamma \setminus \{c_m\}} \mathrm{P}\left[T_{c,n} - T_{c_m,n} > 2d \ln a_n\right] \\
&\leq \sum_{c \in \Gamma \setminus \{c_m\}} \mathrm{P}\left[T_{c,n} > 2d \ln a_n\right],
\end{aligned}$$

where in the last step we made use of the fact that $T_{c_m,n} \geq 0$. Let $\mu_{c,n} = \mathrm{E}_g\left[T_{c,n}\right]$ and $\nu_{c,n} = \mathrm{Var}_g\left[T_{c,n}\right]$. Then, we continue by subtracting $\mu_{c,n}$, taking the absolute value and applying Chebychev's inequality.

$$\begin{aligned} \mathrm{P}\left[C_n \neq c_m\right] &\leq \sum_{c \in \Gamma \setminus \{c_m\}} \mathrm{P}\left[T_{c,n} - \mu_{c,n} > 2d \ln a_n - \mu_{c,n}\right] \\ &\leq \sum_{c \in \Gamma \setminus \{c_m\}} \mathrm{P}\left[|T_{c,n} - \mu_{c,n}| > 2d \ln a_n - \mu_{c,n}\right] \\ &\leq \sum_{c \in \Gamma \setminus \{c_m\}} \frac{\nu_{c,n}}{\left(2d \ln a_n - \mu_{c,n}\right)^2}. \end{aligned} \tag{5.32}$$

as $n \to \infty$ we know from Corollary 5.5 that $\mu_{c,n} \to (c-1)$ and that $\nu_{c,n} \to (c-1)\sigma$ ($\mathrm{Var}_g\left[T_{2,n}\right] \to \sigma$ as $n \to \infty$). σ is finite and c is assumed to be finite. Furthermore is was assumed that $a_n \to \infty$ as $n \to \infty$. Hence each term in Equation 5.32 converges to zero. Since $\#\Gamma$ is finite, only a finite number of terms appear in Equation 5.32. Thus, we have

$$\mathrm{P}\left[C_n \neq c_m\right] \to 0$$

as $n \to \infty$, which proves the theorem. □

Next the asymptotic null distribution is given. In will be clear from the "proof" that will be given, that its validity depends on the validity of Proposition 5.1. For this reason we state the asymptotic null distribution of $T_{C_n,n}$ in the format of a proposition as well.

Proposition 5.3 *Let Γ be a set of permissible orders. Let c_m denote the Minimal Partition Size. Suppose that $a_n \to \infty$ as $n \to \infty$. Under H_0,*

$$T_{C_n,n} \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)}\, Z_j^2,$$

where the Z_j^2 are i.i.d. $\chi^2_{c_m-1}$ variates.

Proof. We proceed as in Inglot et al. (1997). Under H_0, for all $x \in \mathcal{S}$,

$$\begin{aligned} \mathrm{P}\left[T_{C_n,n} \leq x\right] &= \mathrm{P}\left[T_{C_n,n} \leq x, C_n = c_m\right] + \mathrm{P}\left[T_{C_n,n} \leq x, C_n \neq c_m\right] \\ &= \mathrm{P}\left[T_{c_m,n} \leq x\right] \mathrm{P}\left[C_n = c_m\right] + \mathrm{P}\left[T_{C_n,n} \leq x, C_n \neq c_m\right], \end{aligned}$$

where, by Theorem 5.6, the last term tends to zero, and $\mathrm{P}\left[C_n = c_m\right] \to 1$ as $n \to \infty$. Thus, by Proposition 5.1, the proposition follows. □

5.5.3 The Data-Driven SSP test

The data-driven SSPc test will be referred to as the SSPdd test. The notation that is used here is adopted from Section 5.2.4.

The α-level SSPdd test $\phi_{\alpha,\Gamma,n}(\mathcal{S}_n)$, based on the set Γ of permissible orders, is defined as

$$\begin{aligned}\phi_{\alpha,\Gamma,n}(\mathcal{S}_n) &= 1 \text{ if } T_{C_n,n} > t_{\alpha,C_n,n} \\ \phi_{\alpha,\Gamma,n}(\mathcal{S}_n) &= 0 \text{ if } T_{C_n,n} \leq t_{\alpha,C_n,n},\end{aligned}$$

where C_n is the selected order based on a SSP selection rule. Note that the critical value $t_{\alpha,C_n,n}$ is also a random variable. Its distribution depends on the set Γ, on the null distributions of the $T_{c,n}$ with $c \in \Gamma$ and on the probabilities of selecting any c from Γ. Thus the test is clearly depending on the selection rule as well, although this is not explicitly shown in the notation.

The power function is now defined as $\beta_{\alpha,\Gamma,n}(F_1) = \mathrm{E}_{f_1}\left[\phi_{\alpha,\Gamma,n}(\mathcal{S}_n)\right]$. Then, $\beta_{\alpha,\Gamma,n}(G)$ gives the size of the SSPdd test.

Since asymptotically $\mathrm{P}\left[C_n = c_m\right] \to 1$, the critical value of the asymptotic SSPdd test is simply the constant t_{α,c_m}.

The following theorem is a more general result.

Theorem 5.7 *The α-level SSPdd test $\phi_{\alpha,\Gamma,n}$ based on the set Γ of permissible orders, with minimal order c_m, is similar and of size α.*

Proof. For every proper $G \in \mathcal{G}_0$,

$$\begin{aligned}\beta_{\alpha,\Gamma,n}(G) &= \mathrm{E}_g\left[\phi_{\alpha,\Gamma,n}(\mathcal{S}_n)\right] \\ &= \sum_{c\in\Gamma} \mathrm{E}_g\left[\phi_{\alpha,\Gamma,n}(\mathcal{S}_n)|C_n = c\right] \mathrm{P}_g\left[C_n = c\right] \\ &= \sum_{c\in\Gamma} \mathrm{P}_g\left[T_{c,n} > t_{\alpha,c,n}\right] \mathrm{P}_g\left[C_n = c\right] \\ &= \sum_{c\in\Gamma} \alpha \mathrm{P}_g\left[C_n = c\right] \\ &= \alpha \sum_{c\in\Gamma} \mathrm{P}_g\left[C_n = c\right] \\ &= \alpha\end{aligned}$$

In these calculations the definition of $t_{\alpha,c,n}$,

$$\mathrm{P}_g\left[T_{c,n} > t_{\alpha,c,n}\right] = \alpha,$$

is used. □

Since for any $c > 2$ the SSPc test was proven to be consistent, it is easy to extend the result to the SSPdd test.

Theorem 5.8 *For any set Γ of permissible partition sizes, and any selection rule satis-*

fying the conditions that are given in Theorem 5.6, the SSPdd test is consistent against essentially any alternative $F_1 \in \mathcal{G}_1$.

Proof. As $n \to \infty$, Theorem 5.6 gives that under H_0, $C_n \xrightarrow{p} c_m$. Consequently, for the critical value that is used in the SSPdd test holds $t_{\alpha,C_n,n} \xrightarrow{p} t_{\alpha,c_m}$, which is the smallest among all possible asymptotic critical values $t_{\alpha,c}$ ($c \in \Gamma$). Since, furthermore, the SSPc test for $c = c_m$ is consistent, the SSPdd test will be consistent as well. □

5.5.4 Null Distributions for Finite n

In Section 5.2.6 it was clearly illustrated that the convergence of the null distribution of $T_{c,n}$ to its limiting null distributions is very slow. It was concluded that for practical purposes the null distribution could be better simulated or approximated by other means, rather than using the asymptotic null distribution. It is obvious that this arguments will still hold for the SSPdd test.

In this section yet another convergence is empirically studied. In Theorem 5.6 it was proven that, under H_0, the selected order C_n converges to the minimal order c_m as $n \to \infty$. The speed of this convergence is assessed here in a small Monte Carlo simulation experiment in which the probabilities $\mathrm{P}_g\,[C_n = c]$ for all $c \in \Gamma$ are estimated, based on 10000 simulation runs.

Three types of penalty are included in the experiment, two of which are well known penalties: a BIC-like criterion ($a_n = n^{1/2}$) and a double logarithmic criterion (Hannan and Quinn, 1979) ($a_n = (\ln n)^{1/2}$). The former will be referred to as BIC and the latter as LL. The third penalty is introduced as an alternative to the AIC criterion, which does not have the property $a_n \to \infty$ as $n \to \infty$, and which therefore does not apply to the conditions of the theory of the data-driven SSP test which is presented in this work. The alternative to the AIC criterion, which will be referred to as the pseudo-AIC (pAIC) criterion, is constructed such that the mean of $2 \ln a_n$ over the sample sizes $n = 20$ up to $n = 100$ is equal to 2, which is the constant value of the true AIC criterion. The choice for the range $n = 20$ up to $n = 100$ is rather arbitrary, but it is thought of to reflect a range of sample sizes that occur often in practice. Of course other choices are possible as well. This condition results in $a_n = \frac{\ln n}{1.38}$.
From the comparison of the three penalties (Figure 5.15) it can now already be concluded that the selection rule based on the BIC criterion will faster select smaller partition sizes as the sample size grows, because $n^{1/2}$ increases faster with n as compared to $(\ln n)^{1/2}$ and $\frac{\ln n}{1.38}$. The pseudo-AIC criterion will be in between the two others.

Simulations were performed for three sample sizes: $n = 20$, $n = 50$ and $n = 100$. The set of permissible partition sizes was taken to be $\Gamma = \{2, 3, 4\}$ in a first series of simulations, and $\Gamma = \{2, 3, 4, 5\}$ in a second series. All simulations were run under a simple null hypothesis (uniform distribution).

The results are presented in Table 5.15. These results clearly confirm the effect of the

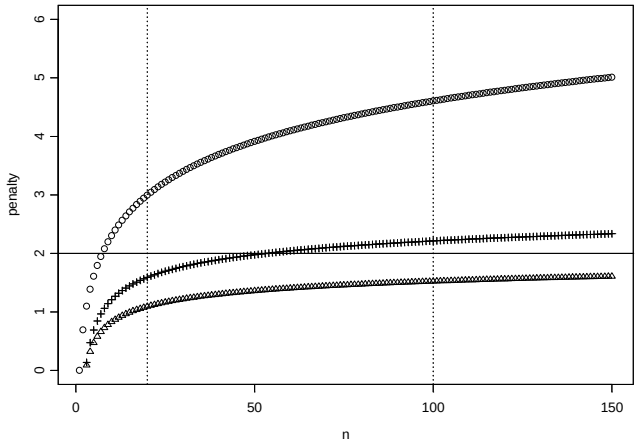


Figure 5.15: The BIC (dots), pAIC (crosses) and LL (triangles) penalties $2a_n$ as a function of the sample size n. The vertical dashed lines indicate the bounds of the range of sample sizes used to determine the pAIC penalty.

penalty on the speed of the convergence of the selected order to c_m. The results also suggest that increasing the set $\Gamma = \{2, 3, 4\}$ to $\Gamma = \{2, 3, 4, 5\}$ slows down the convergence. In general, for moderate sample sizes (say, $n < 100$), it is clear that for all penalties the convergence is too slow to use the asymptotic null distribution which is given in Proposition 5.3. Thus, an important practical consequence is that for applying the SSPdd test one should have approximations (e.g. simulated null distributions, ...) of the exact null distributions of all SSPc test statistics corresponding to the partition sizes in Γ.

5.5.5 Example

EXAMPLE 5.4. **Singer Data**
Of the examples that were analyzed in Section 5.3.6, only the Singer data is analyzed further here by means of the data driven SSP test for it was clear from the results of the SSP2, SSP3 and SSP4 test that the choice of c may be very important. Indeed, with $c = 2$ it would have been concluded that the heights are normally distributed, whereas the other two tests gave extreme small p-values. The set of permissable orders is taken as $\Gamma = \{2, 3, 4\}$, and all three the penalties (BIC, LL and pAIC) are considered. They all selected the largest partition size in Γ, i.e. $c = 4$. Thus also the data-driven test gives $p < 0.0001$ and it is concluded again that the heights of the singers are not normally distributed. □

Table 5.15: Results from a Monte Carlo simulation study to assess the convergence of C_n to the minimal order $c_m = 2$. For each penalty and each sample size, the first and the second line show the estimated probabilities that $C_n = c$ when $\Gamma = \{2, 3, 4\}$, and $\Gamma = \{2, 3, 4, 5\}$, respectively.

		selected partition size c			
penalty	n	2	3	4	5
BIC	20	0.784	0.006	0.210	—
		0.745	0.005	0.004	0.246
	50	0.978	0.002	0.020	—
		0.967	0.001	0.001	0.031
	100	0.996	0.001	0.003	—
pAIC	20	0.447	0.008	0.545	—
		0.384	0.011	0.005	0.600
	50	0.757	0.010	0.233	—
		0.715	0.008	0.004	0.273
	100	0.916	0.020	0.064	—
LL	20	0.220	0.005	0.775	—
		0.186	0.005	0.003	0.806
	50	0.466	0.008	0.526	—
		0.395	0.008	0.004	0.593
	100	0.729	0.018	0.253	—

5.5.6 Power Characteristics

Although, the power of the SSPc test was studied in previous sections under both simple and composite null hypotheses, we will here restrict the results to composite null hypotheses because for practical purposes this is the most important type of hypothesis. To make the results that will be presented in this section easily comparable with the results for the SSPc, Anderson-Darling, Kolmogorov-Smirnov, Shapiro-Wilk, Rao-Robson, moments K, and the data-driven Kallenberg-Ledwina tests, the same null (normal distribution) and alternative distributions (Tukey-Lambda family, and a family of mixtures of 2 normal distributions) will be considered here again. Thus, for more details on these families of alternatives, the construction of the isotones and the simulation conditions, we refer to Section 5.3.5.

The isotones of the SSPdd tests are constructed in a somewhat different way because there is not just one critical value. Indeed, the critical value depends on the partition size that is selected by the selection rule. In the notation of Section 5.3.5, the isotones of a data-driven test are actually those of $t(P_{n,\theta}) - t_{\alpha,c(P_{n,\theta}),n}$, where $c(P_{n,\theta})$ is the order that is selected for profile $P_{n,\theta}$. The constant to which the isotones of the data-driven tests correspond, is zero, which is the threshold for significance at the α-level for whatever selected order. The isotones are presented in Figures 5.16 and 5.17 for the Tukey-Lambda family and the family of mixture distributions, respectively.

Tukey-Lambda: Powers are estimated under the same conditions as the SSPc test (com-

posite null) in Section 5.3.5. The results are shown in Table 5.16 (the powers of the SSPc tests are copied from Table 5.8).

In general the isotones show that the SSPdd-BIC test coincide almost exactly with the isotones of the SSP2 test. The same holds for the SSPdd-LL and the SSP4 tests. Since, for the TL-family, the highest powers among the SSPc test are those of the SSP2 test, and the lowest are those of the SSP4 test, it is expected that the data-driven test with the BIC-like penalty results in higher powers than the test with the LL penalty. It is important to remark here that coinciding isotones does not necessarily mean that the powers are exactly the same. It is only the order of isotones that are crossed, when moving from the point representing the null hypothesis, into a certain direction of alternatives, that approximately corresponds to the order of the powers. For $n = 50$ the results from the simulation experiment are very clear: the estimated powers of the SSPdd-LL and SSPdd-BIC test are almost exactly equal to those of the SSP4 and SSP2 test, respectively. For $n = 20$ the same conclusion still holds for the SSPdd-LL / SSP4 tests, but the estimated powers of the SSPdd-BIC and the SSP2 test sometimes differ little, but still the relative order of the powers is retained.

The pAIC penalty given almost the same results as the LL penalty when $n = 20$. For the larger sample size ($n = 50$), the power of the SSPdd-pAIC test becomes larger as compared to the SSPdd-LL test. In some cases it has even a larger power than the test based on the BIC criterion.

Only in the region $(\lambda_1, \lambda_2) \in [8, 10] \times [4, 7]$ ($n = 20$) there is a distinction between the isotones of the SSP2 and the SSPdd-BIC tests: the isotone of the SSPdd-BIC test follows along the left hand side the isotone of the SSP2 test, but then, about the point $(\lambda_1, \lambda_2) = (8, 7)$, the isotone of the data-driven test starts turning back to the right, almost enclosing an area in the $\boldsymbol{\lambda}$-plane. This area corresponds to a sub-family of members that result in a much lower power of the SSPdd-BIC test as compared to the SSP2 test. This can also be concluded from Table 5.16 (entry $n = 20$, $(\lambda_1, \lambda_2) = (9, 5)$).

With respect to the TL-family of alternatives, the general conclusion is that the data-driven test which is based on the BIC-like criterion has superior power as compared to the SSPdd-LL test. These superior powers are very close to those of the SSP2 test which is for the whole TL-family ($\boldsymbol{\lambda} \in [-2, 10] \times [-2, 10]$) the best among the SSPc tests.

Contaminated Normal: Powers are estimated under the same conditions as the SSPc test (composite null) in Section 5.3.5. The results are shown in Table 5.17 (the powers of the SSPc tests are copied from Table 5.9).

In the contaminated normal family it has been clearly shown that the SSP3 and the SSP4 test outperformed the SSP2 test (Section 5.3.5). Whereas in the TL-family the isotones of the SSPdd-BIC test were almost indistinguishable from those of the SSP2 test, they are now virtually identical to the isotones of the SSP3 and the SSP4 test. Also the isotones of the SSPdd-LL test almost coincide everywhere with those of the SSP3 and the SSP4 tests.

Similar as for the TL-family, the pAIC criterion results in powers comparable to those

Table 5.16: The estimated powers for some alternatives of the Tukey-Lambda family, based on 10000 simulation runs.

n	λ_1	λ_2	power					
			SSP2	SSP3	SSP4	dd - BIC	dd - LL	dd - pAIC
20	-0.5	-0.5	0.622	0.500	0.409	0.521	0.409	0.410
20	1	1	0.064	0.061	0.064	0.092	0.064	0.064
20	1.5	1.5	0.083	0.081	0.080	0.103	0.080	0.080
20	1.5	2	0.061	0.065	0.067	0.092	0.067	0.067
20	2	0	0.565	0.409	0.299	0.418	0.299	0.299
20	5	10	0.538	0.442	0.346	0.383	0.346	0.346
20	6	6	0.225	0.186	0.140	0.204	0.137	0.137
20	9	9	0.749	0.787	0.720	0.774	0.720	0.720
20	9	5	0.567	0.464	0.353	0.436	0.353	0.353
50	-0.5	-0.5	0.939	0.935	0.909	0.939	0.910	0.920
50	1.5	1.5	0.477	0.440	0.384	0.478	0.384	0.433
50	2	0	0.911	0.876	0.806	0.911	0.806	0.866
50	3	3	0.083	0.092	0.090	0.092	0.090	0.111
50	5	5	0.173	0.239	0.213	0.188	0.215	0.240
50	5	9	0.899	0.917	0.901	0.899	0.901	0.916
50	1	8	0.219	0.203	0.165	0.221	0.165	0.225
50	4	8.5	0.834	0.825	0.759	0.840	0.761	0.820

of the SSPdd-LL and SSP4 tests when $n = 20$. When $n = 50$, the SSPdd-pAIC test becomes more powerful.

Some General Conclusions

The isotones and the estimated powers under the Tukey-Lambda and the mixture families of alternatives, suggest some more general conclusions. First, it seems that the BIC-like criterion succeeded rather good in selecting the model resulting in the high powers: in the TL-family it was the *simpler* model with $c = 2$, and in the mixture model it was one of the more *complex* models with $c = 3$ or $c = 4$. The LL-penalty, on the other hand, selected under both families of alternatives the more complex models with $c = 3$ and $c = 4$, which for the TL-family resulted in lower powers.

Under the null hypothesis it was already observed (Section 5.5.4) that the LL-penalty resulted in a slower convergence of the selected order to the minimal partition size, which is $c_m = 2$ here. The simulations presented in this section now also suggest that the LL-penalty has still a tendency to select the higher order models under alternative hypotheses and for moderate sample sizes of $n = 20$ and $n = 50$, even when the smaller sized partitions give better results. Thus, although asymptotically both penalties would select the minimal partition size, it seem that with moderate sample sizes, the BIC-like penalty deserves our favour. Nonetheless, we would like to remark that, possibly, alternatives may

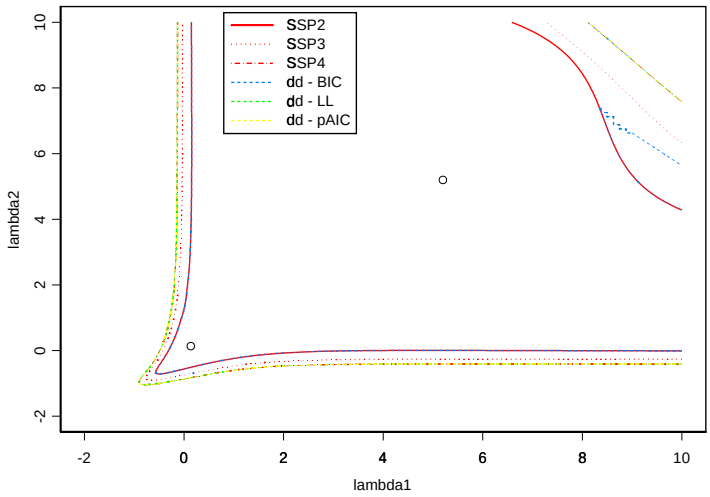
Table 5.17: The estimated powers for some alternatives of the contaminated normal family, based on 10000 simulation runs.

n	λ_1	λ_2	power					
			SSP2	SSP3	SSP4	dd - BIC	dd - LL	dd - pAIC
20	0	0.2	0.215	0.769	0.781	0.800	0.781	0.784
20	0	0.7	1.000	1.000	1.000	1.000	1.000	1.000
20	2	0.4	0.683	0.991	0.991	0.991	0.991	0.991
20	4	0.4	0.948	0.996	0.996	0.997	0.996	0.996
50	0	0.2	0.513	0.986	0.990	0.984	0.990	0.990
50	0	0.6	1.000	1.000	1.000	1.000	1.000	1.000
50	1.25	0.38	0.993	1.000	1.000	1.000	1.000	1.000
50	3	0.4	1.000	1.000	1.000	1.000	1.000	1.000
50	3.5	0.1	0.754	0.821	0.804	0.825	0.810	0.836

be constructed under which the SSPdd-LL test has higher power than the SSPdd-BIC test. Finally, Kallenberg and Ledwina (1997) also conclude that the BIC selection rule gives in general the best results for their data-driven Neyman's smooth test.

In this thesis also a new alternative penalty pAIC, which is constructed on heuristic grounds, is proposed. Since with this penalty the convergence of C_n to c_m under the null hypothesis is also slower as compared to the BIC criterion, it still selects for small sample sizes ($n = 20$) frequently the larger partition sizes, resulting under the TL-family to relative low powers, and under the mixture family to rather high powers. But when $n = 50$, the criterion succeeds to select more frequently the better small partitions as well, which makes is a better choice than the LL penalty. But still better results are obtained with the BIC penalty.

(a)



(b)

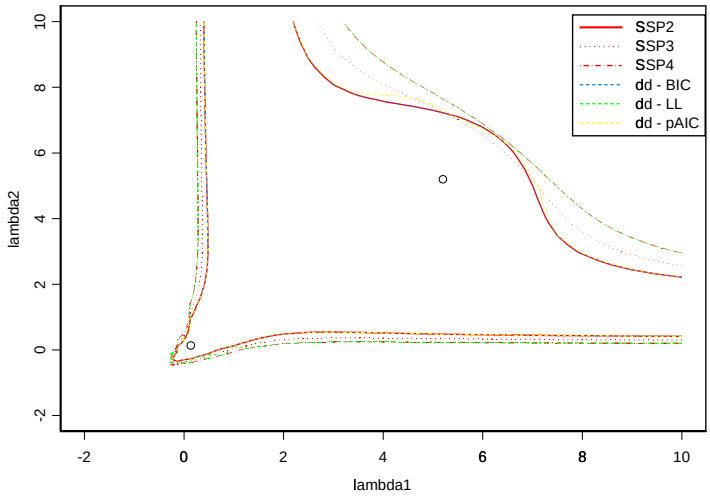
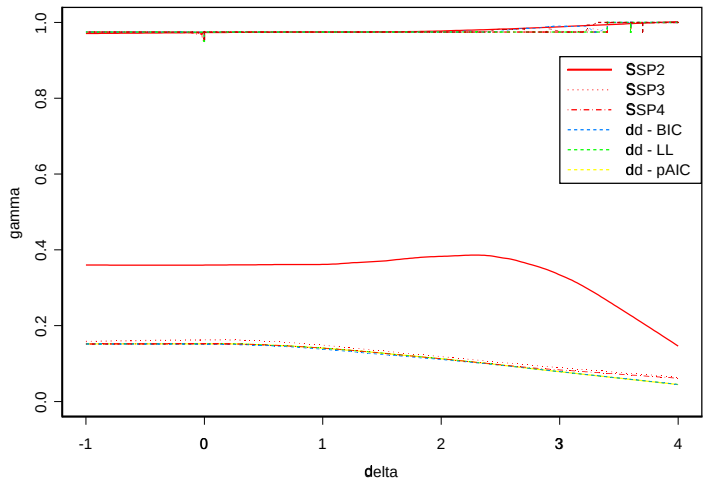


Figure 5.16: The isotones for the Tukey-Lambda family of alternatives, for $n = 20$ (a) and $n = 50$ (b).

(a)



(b)

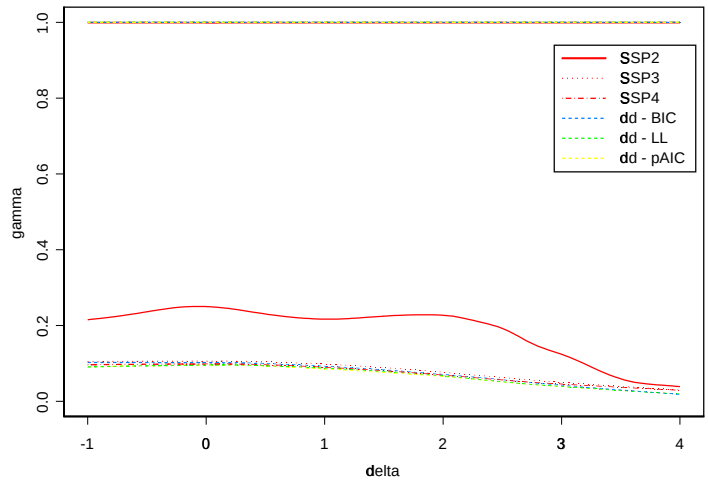


Figure 5.17: The isotones for the family of mixture distributions, for $n = 20$ (a) and $n = 50$ (b).

5.6 An Extension to Multivariate Distributions

Up to now we restricted the discussion to a univariate rv X. In this section we will briefly show how the SSP GOF test may be extended for testing whether or not a sample of multivariate observations come from a distribution G. This section is only meant as an introduction to the topic. We therefore limit us to the simple multivariate null hypothesis. First some general complications are illustrated on the Cramér-von Mises statistic.

5.6.1 The Cramér-von Mises Statistic for Multivariate GOF

In general there is definitely not as much written about the multivariate GOF problem as it is about the univariate situation. One of the most cited researchers in this area is probably Mardia (see Mardia, 1980, for an overview).
Many of the methods for testing univariate GOF that have been discussed in Chapter 4 can be extended to the multivariate setting. Since the Cramér-von Mises statistic is the one which is most closely related to the SSP statistic, we comment here briefly upon the complications that are encountered for at least some of them may be applicable to the SSP setting as well.

In the univariate case the limiting null distribution of the Cramér-von Mises statistic is found by recognizing that the statistic is the integral of a squared empirical process, for which a weak convergence to a Gaussian process (Brownian bridge) holds. The integral equation based on the covariance function of the latter process yields the coefficients that determine the asymptotic null distribution. The p-dimensional multivariate analogue is simply of the form (Csörgö, 1986)

$$\int_{\mathcal{S}} Z_n^2(\boldsymbol{t}) d\boldsymbol{t},$$

where $\boldsymbol{t}$ is a p-dimensional "time" parameter, and Z_n is a p-dimensional empirical process, based on the multivariate EDF. Although this generalization looks rather straightforward and simple, and although there even exists a weak convergence of Z_n to a p-dimensional Gaussian process (Durbin, 1973; Neuhaus, 1976), the resulting integral equation is very complex, such that the statistic is only of theoretical interest. An additional problem, even when the null hypothesis is simple, is that the distribution G cannot be uniquely eliminated. Indeed, since the multivariate integral transformation is not unique, the test is not invariant under the null hypothesis (Koziol, 1986).
Actually the SSP statistics are special cases of the Anderson-Darling family of statistics, which are generalizations of the Cramér-von Mises statistic by introducing a weight function. This clearly even complicates the matter further.

5.6.2 A SSP Test for Multivariate GOF

Despite the expected problems mentioned in the previous section, we will show here that with a certain type of SSPs the resulting statistic may have a simple asymptotic null distribution. Our attention is limited to the bivariate case, but extensions to arbitrary dimensionality will be obvious.

Let F and G denote respectively the true and the hypothesized CDF of the bivariate rv $\boldsymbol{Z} = (X, Y)$.

Translating the Cramér-von Mises statistic of the previous section to a SSP test in the same way as it was done in the beginning of this chapter, would result in

$$T_n = 2n\mathrm{E}_{\hat{F}}\left[\hat{\mathrm{I}}^1\left(F; G \| [A](\boldsymbol{Z})\right)\right],$$

where $[A](\boldsymbol{Z})$ is a SSP determined by $\boldsymbol{Z}$ and a partition construction rule. This rule is simply

$$[A](\boldsymbol{Z}) = \{]b_l^x, X] \times]b_l^y, Y],]b_l^x, X] \times]Y, b_u^y],]X, b_u^x] \times]b_l^y, Y],]X, b_u^x] \times]Y, b_u^y]\},$$

i.e. a 2×2 partition "centred" at $\boldsymbol{Z}$. Both the SSP statistic and the Cramér-von Mises statistic use the same partitions, as it was the case in the univariate setting.
This is however not exactly the way we will proceed.

The problem with the above described procedure is that the resulting empirical process needs a 2-dimensional index parameter, and that generally no unique multivariate integral transformation exists. This problem is overcome by considering only partitions that result in an empirical process that is indexed by only a one dimensional parameter which has a one-to-one relation with $\boldsymbol{Z}$.
One such solution is

$$[A](\boldsymbol{Z}) = \{[b_l^x, X[\times [b_l^y, F_y^{-1}(F_x(X))[, \mathcal{S}_z \setminus [b_l^x, X[\times [b_l^y, F_y^{-1}(F_x(X))[\},$$

(the roles of X and Y may be exchanged). When both univariate integral transformations are applied on X and Y, the partition becomes

$$[A](\boldsymbol{Z}) = [A](U) = \{[0, U]^2, [0, 1]^2 \setminus [0, U]^2\},$$

where $U = F_x(X)$. Let also $V = F_y(Y)$. Thus the partition is now in its copula representation.

Since the partition has only size 2, the test statistic now becomes

$$\begin{aligned} T_n &= \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\sqrt{n}\left(\hat{\mathrm{P}}_f\left[[0,U_i]^2\right] - \mathrm{P}_g\left[[0,U_i]^2\right]\right)\right)^2}{\mathrm{P}_g\left[[0,U_i]^2\right]\left(1-\mathrm{P}_g\left[[0,U_i]^2\right]\right)} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\sqrt{n}\left(\hat{F}_{uv}(U_i,U_i) - G_{uv}(U_i,U_i)\right)\right)^2}{G_{uv}(U_i,U_i)\left(1-G_{uv}(U_i,U_i)\right)}. \end{aligned}$$

The latter representation does essentially not differ from the SSP2 statistic in the univariate setting (cfr. Equation 5.10). Since both arguments (U_i and U_i) in $\hat{F}_{uv}$ and G_{uv} are equal, we may write the distributions as $\hat{F}(U_i)$ and $G(U_i)$, respectively. Moreover, the latter CDF is inversible, i.e. $G^{-1}(s)$ is uniquely determined. It is straightforward now to recognize that the empirical process

$$U_n(s) = \sqrt{n}\left(\hat{F}(G^{-1}(s)) - s\right)$$

is under H_0 a Gaussian process with the covariance function of a Brownian bridge. Hence, from here there is no difference anymore from the SSP2 statistic, and thus is its asymptotic null distribution the same as for the SSP2 statistic in the univariate setting (Theorem 5.1).

5.6.3 Extension of the Multivariate SSP Test to Larger SSPs

An easy way to allow for larger partition sizes is to construct first a 2-sized partition as explained in the previous section, say $[A](\boldsymbol{Z}_i)$, and then to construct again such a 2-sized partition within the subset $[0,U_i]^2$ of the copula sample space, "centred" at $\boldsymbol{Z}_j$ for which $X_j < X_i$. Larger partitions may be constructed by recursively applying this procedure. Let $\mathcal{P} \subset \mathcal{S}$ containing $c-1$ different bivariate observations on $\boldsymbol{Z}$, and let $a_{c-1,n}$ denote the number of such sets $\mathcal{P}$ that can be constructed. As before, $\mathbb{P}_{c-1,n}$ denotes the set of the $a_{c-1,n}$ sets $\mathcal{P}$. Note that it is only the X component, or its integral transformed rv U, that determines the SSP. Thus the c-sized partition is given by

$$[A](\mathcal{P}) = \{[0,U_{(i_1)}], [0,U_{(i_2)}] \setminus [0,U_{(i_1)}], \ldots, [0,1] \setminus [0,U_{(i_{c-1})}]\},$$

where $U_{(i_k)}$ denote the i_kth order statistic of the sample of U observations ($k = 1,\ldots,c-1$). The multivariate SSPc statistic becomes

$$T_{c,n} = \frac{2n}{a_{c-1,n}} \sum_{\mathcal{P}\in\mathbb{P}_{c-1,n}} \hat{\mathrm{I}}^1\left(F;G\|[A](\mathcal{P})\right).$$

By the recursive way in which the partitions are constructed, it is obvious that its asymptotic null distribution will be the same as the one that is proposed for the univariate SSPc statistic (Proposition 5.1).

CHAPTER 6

k-Sample Sample Space Partition Test

In Section 3.3.4 a brief introduction to the k-sample GOF problem was given. The null hypothesis is

$$H_0 : F_1(\boldsymbol{x}) = F_2(\boldsymbol{x}) = \ldots = F_k(\boldsymbol{x}) = G(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{S}_x \text{ and } G \in \mathcal{F}. \tag{6.1}$$

We assume that $\boldsymbol{x}$ is a continuous rv and that Assumption A1 holds for all distributions F_i ($i = 1, \ldots, k$) and distribution G. Furthermore, the sample space $\mathcal{S} = \mathcal{S}_x$ is supposed to be common for all members of $\mathcal{F}$. This replaces Assumption A2 (or A2b) in the previous chapter.

Assumption (A2c).
All distributions F_i ($i = 1, \ldots, k$) and distribution G are defined on a common sample space $\mathcal{S}$.

In the general k-sample problem the alternative hypothesis is given as the negation of the null hypothesis. Thus,

$$H_1 : \exists i \neq j, \boldsymbol{x} : F_i(\boldsymbol{x}) \neq F_j(\boldsymbol{x}), F_i, F_j \in \mathcal{F}.$$

We will restrict this chapter to a univariate rv $\boldsymbol{x} = x$.

Since there are no further restrictions imposed on the set $\mathcal{F}$, the alternative hypothesis may be considered nonparametric. The test must be constructed such that its null distribution does not depend on $G \in \mathcal{F}$. Tests that are constructed for this type of hypothesis are very useful in practice, because the researcher must not have any prior knowledge about the family of distributions to which the distributions belong. Since the alternative hypothesis is just a negation of H_0, the tests must be sensitive to a very wide class of alternatives, i.e. the tests must be an omnibus tests.

The test will be constructed in a similar way as the SSPc GOF test (Section 6.1). Thus, first the appropriate Directed Divergence is proposed, then the Average Pearson Divergence and the related Average Pearson Divergence Statistic are given. The test statistic of the k-sample test that is proposed is based on this statistic. Its null distribution is obtained and its sensitivity towards some families of alternatives is investigated in a small simulation study. In Section 6.2 a data-driven version is proposed. Finally, a decomposition of the test statistic is studied in Section 6.4.

6.1 The k-Sample SSP test

6.1.1 The Directed Divergence

We will show that there is no need to redefine the directed divergence of Cressie and Read (1984), as it was given in Section 5.1.

Consider a discrete random variable Y, defined over a sample space $\mathcal{S}_y = \{1, 2, \ldots, k\}$ and with probabilities $\mathrm{P}[Y = i] = \frac{n_i}{n}$, where n_i is the number of observations in the ith sample $\mathcal{S}_{n_i}$, and $n = \sum_{i=1}^{n} n_i$ is the total number of observations. n_i is considered here as a constant, i.e. the sample sizes n_i are fixed by design. This rv Y may now be considered as an indicator variable, indicating the sample i. Then the joint distribution of X and Y may be constructed as

$$f_{xy}(x, y) = f_{x|y}(x|y) f_y(y),$$

or, in the present context, in a more appropriate notation

$$f_{xy}(x, i) = f_{x|y}(x|i) f_y(i),$$

where $f_y(i) = \mathrm{P}[Y = i] = \frac{n_i}{n}$ and where the conditional df $f_{x|y}(x|i) = f_i(x)$. The latter density is the density of the distribution of the observation of the ith sample. The CDFs may be constructed from these dfs.

Since under the null hypothesis all $f_{x|y}(x|i)$ are equal ($i = 1, \ldots, k$), the joint distribution of X and Y becomes

$$g_{xy}(x, i) = g_{x|y}(x|i) f_y(i) = g_x(x) f_y(i),$$

where $g_x = g$ is the common df of X under H_0.

The directed divergence of order $\lambda \in \mathbb{R}$ between F_{xy} and G_{xy} (Definition 5.2 and Equation 5.3) now becomes

$$\begin{aligned} \mathrm{I}^{\lambda}(F;G) &= \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{k} \int_{\mathcal{S}_x} \left[\left(\frac{f_{xy}(x,i)}{g_{xy}(x,i)} \right)^{\lambda} - 1 \right] f_{xy}(x,i)dx \\ &= \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{k} \int_{\mathcal{S}_x} \left[\left(\frac{f_i(x)}{g_x(x)} \right)^{\lambda} - 1 \right] f_{x|y}(x|i) f_y(i)dx \\ &= \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{k} \frac{n_i}{n} \int_{\mathcal{S}_x} \left[\left(\frac{f_i(x)}{g_x(x)} \right)^{\lambda} - 1 \right] f_i(x)dx \end{aligned} \tag{6.2}$$

In a similar way as in Section 5.1 the directed divergence can also be based on a size c sample space partition $[A] = \{A_1, \ldots, A_c\}$. In particular, the SSP $[A]$ is a partition of the sample space $\mathcal{S}_x$. Therefore, on the sample space of the joint distribution of X and Y, which is given by $\mathcal{S}_{xy} = \mathcal{S}_x \times \mathcal{S}_y$, any partition $[A]$ uniquely implies a ck sized partition $[B]$ which is given by

$$\begin{aligned} [B] &= [A] \times \mathcal{S}_y \\ &= \{A_1 \times \{1\}, A_1 \times \{2\}, \ldots, A_1 \times \{k\}, A_2 \times \{1\}, \ldots, A_c \times \{k\}\}. \end{aligned}$$

let $B_{ji} = A_j \times \{i\}$ ($j = 1, \ldots, c; i = 1, \ldots, k$). Figure 6.1 shows the relation between both types of partitions.

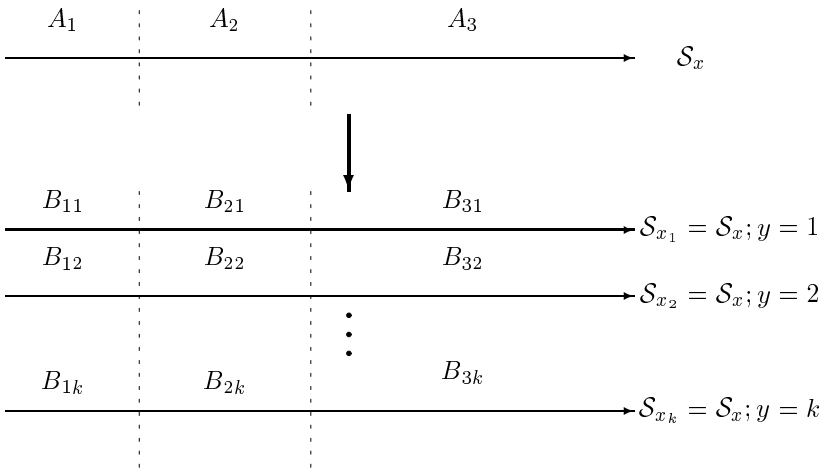


Figure 6.1: Illustration of the relation between the SSP $[A]$ (with $c = 3$) and the SSP $[B]$.

The discrete divergence becomes

$$\begin{aligned}
\mathrm{I}^{\lambda}\left(F;G\|[A]\right) &= \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k}\sum_{j=1}^{c}\mathrm{P}_{f_{xy}}\left[A_j\times\{i\}\right]\left[\left(\frac{\mathrm{P}_{f_{xy}}\left[A_j\times\{i\}\right]}{\mathrm{P}_{g_{xy}}\left[A_j\times\{i\}\right]}\right)^{\lambda}-1\right] \\
&= \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k}\sum_{j=1}^{c}\frac{n_i}{n}\mathrm{P}_{f_i}\left[B_{ji}\right]\left[\left(\frac{\frac{n_i}{n}\mathrm{P}_{f_i}\left[B_{ji}\right]}{\frac{n_i}{n}\mathrm{P}_{g_{x|y}}\left[B_{ji}\right]}\right)^{\lambda}-1\right] \\
&= \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k}\frac{n_i}{n}\sum_{j=1}^{c}\mathrm{P}_{f_i}\left[B_{ji}\right]\left[\left(\frac{\mathrm{P}_{f_i}\left[B_{ji}\right]}{\mathrm{P}_{g}\left[A_j\right]}\right)^{\lambda}-1\right]. \qquad (6.3)
\end{aligned}$$

Since the divergences that are presented above are just a result of the application of the definitions, which are given in Section 5.1, all properties that are given in that section still apply. More specifically, Corollary 5.4 is altered to the present k-sample problem.

Corollary 6.1 *If, for some λ, there exists a SSP $[A]$ of any size $c > 1$ for which $I^{\lambda}\left(F;G\|[A]\right) > 0$, then at least two distributions F_i and F_j $(i \neq j)$ are different.*

In the beginning of this section a random variable Y was introduced. It was basically used to construct the divergences. From the final expressions of the divergences (Equations 6.2 and 6.3), it is seen that the divergences only depend on distributions of the random variable X. The factor $\frac{n_i}{n}$ is to be interpreted as a design related factor.

6.1.2 The Power Divergence Statistic

The directed divergence statistic (Cressie and Read, 1984), based on a c-sized SSP $[A]$, is now given by

$$\hat{\mathrm{I}}^{\lambda}\left(F;G\|[A]\right) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k}\frac{n_i}{n}\sum_{j=1}^{c}\hat{\mathrm{P}}_{f_i}\left[B_{ji}\right]\left[\left(\frac{\hat{\mathrm{P}}_{f_i}\left[B_{ji}\right]}{\hat{\mathrm{P}}_{g}\left[A_j\right]}\right)^{\lambda}-1\right],$$

where $\hat{\mathrm{P}}_{f_i}\left[B_{ji}\right] = n_i^{-1}\#(B_{ji}\cap\mathcal{S}_{n_i})$ and $\hat{\mathrm{P}}_{g}\left[A_j\right] = n^{-1}\#(A_j\cap\mathcal{S}_{n})$. The former probability estimator is computed by making use of only the observations in the ith sample. The latter estimator, on the other hand, is based on the combined sample $\mathcal{S}_n$. If all $n_i\to\infty$, then all the probability estimators are consistent.
Note that a SSP $[A]$ may be constructed such that, for some $A\in[A]$, $\hat{\mathrm{P}}_{g}\left[A\right]=0$. This would make the directed divergence statistic infinitely large. In the remainder of this section such partitions are excluded. In Section 6.1.4 the problem is discussed in more detail.

Cressie and Read (1984) showed that for power divergence statistics of any order λ, and

conditional on any SSP $[A]$, the following convergence holds under H_0:

$$2n\hat{\mathrm{I}}^{\lambda}\left(F;G\|[A]\right)-2n\hat{\mathrm{I}}^{1}\left(F;G\|[A]\right)\xrightarrow{p}0.$$

The power divergence statistic of order 1 is simply a Pearson χ^2 statistic:

$$\begin{aligned}
2n\hat{\mathrm{I}}^{1}\left(F;G\|[A]\right) &= n\sum_{i=1}^{k}\frac{n_i}{n}\sum_{j=1}^{c}\frac{\left(\hat{\mathrm{P}}_{f_i}\left[B_{ji}\right]-\hat{\mathrm{P}}_{g}\left[A_j\right]\right)^2}{\hat{\mathrm{P}}_{g}\left[A_j\right]}\\
&= n\sum_{i=1}^{k}\sum_{j=1}^{c}\frac{\left(\frac{n_i}{n}\hat{\mathrm{P}}_{f_i}\left[B_{ji}\right]-\frac{n_i}{n}\hat{\mathrm{P}}_{g}\left[A_j\right]\right)^2}{\frac{n_i}{n}\hat{\mathrm{P}}_{g}\left[A_j\right]}\\
&= n\sum_{i=1}^{k}\sum_{j=1}^{c}\frac{\left(\hat{\mathrm{P}}_{f_{xy}}\left[B_{ji}\right]-\frac{n_i}{n}\hat{\mathrm{P}}_{g}\left[A_j\right]\right)^2}{\frac{n_i}{n}\hat{\mathrm{P}}_{g}\left[A_j\right]}
\end{aligned}$$

More specifically $2n\hat{\mathrm{I}}^{1}\left(F;G\|[A]\right)$ is the Pearson statistic for testing the null hypothesis that the multinomial distributions, with probabilities $\mathrm{P}_{f_i}\left[B_{ji}\right]$, which are implied by the SSP $[A]$, are the same within each stratum i (independent multinomial sampling). The statistic is however indistinguishable from the statistic for testing the null hypothesis of independence in a $c\times k$ contingency table, which is induced by the SSP $[A]$. In both cases, under the corresponding null hypothesis,

$$2n\hat{\mathrm{I}}^{1}\left(F;G\|[A]\right)\xrightarrow{d}\chi^2_{(c-1)(k-1)},$$

and thus also, for any λ, under H_0,

$$2n\hat{\mathrm{I}}^{\lambda}\left(F;G\|[A]\right)\xrightarrow{d}\chi^2_{(c-1)(k-1)}.$$

In the next sections only $\lambda=1$, which corresponds to the Pearson-like statistic, is further considered. The reason for doing this is mainly that it will give immediate relationships between the test that is developed here and Anderson-Darling type tests.

6.1.3 The Averaged Pearson Divergence

Both the continuous and the discrete divergence are appropriately measuring the deviation from the k-sample null hypothesis, but the latter divergence is only defined for a given SSP. The advantage, though, of the discrete divergence is that it can be easily and non-parametrically estimated by means of its plug-in estimator: the power divergence statistic. In order to retain the advantage of having a straightforward plug-in estimator on the one hand, and to overcome the disadvantage of having to specify a SSP, a new functional (the Averaged Pearson Divergence $\Delta_c(F;G)$) is constructed as the average of all discrete di-

vergences based on SSPs of size c. This was also done in Section 5.2.1 for the one-sample GOF-hypothesis.

The definition (Definition 5.6) of the Averaged Pearson Divergence remains essentially unchanged:

$$\Delta_c(F;G) = \mathrm{E}_f\left[\mathrm{I}^1\left(F;G\|[A](\mathcal{P}_{c-1,n})\right]\right., \tag{6.4}$$

where now the *Partition Construction Rule*, the *Partition Determining Subsample Set* and the set $\mathbb{P}_{c-1,n}$ are most easily defined on the sample space $\mathcal{S}_x$. More specifically, they are defined exactly as in Section 5.2.1 (Definition 5.5). Thus for all $\mathcal{P}_{c-1,n} \in \mathbb{P}_{c-1,n}$, $\mathcal{P}_{c-1,n} \subset \mathcal{S}_n$, the resulting SSP $[A](\mathcal{P}_{c-1,n})$ are partitions of $\mathcal{S}_x$. The reason for defining them like this is that any partition $[A](\mathcal{P}_{c-1,n}) = [A]$, that is constructed in this way, uniquely implies a SSP $[B]$ of the sample space $\mathcal{S}_{xy}$. Furthermore, the expectation in Equation 6.4 is taken w.r.t. the true joint distribution F_{xy} of the rv X and the constructed rv Y.

Lemma 5.1 remains unchanged. Thus

$$\Delta_c(F;G) = 0 \Leftrightarrow H_0 \text{ is true}, \tag{6.5}$$

and the next step is to find an estimator of $\Delta_c(F;G)$ that can be used in the construction of a test statistic.

6.1.4 The Averaged Pearson Divergence Statistic

The Averaged Pearson Divergence Statistic is defined as the plug-in estimator of $\Delta_c(F;G)$. Since in the present context the common distribution G is also unknown, it must be replaced by its plug-in estimator as well. In particular the probabilities $\mathrm{P}_g\left[A_j\right]$ must be estimated, which is easier than estimating G itself.

$$\begin{aligned}\hat{\Delta}_c(F;G) &= \Delta_c(\hat{F};\hat{G}) \\ &= \mathrm{E}_{\hat{F}}\left[\mathrm{I}^1\left(\hat{F};\hat{G}\|[A]_{\mathcal{P}}\right)\right] \\ &= \int_{\mathcal{S}_{\mathcal{P}}} \hat{\mathrm{I}}^1\left(F;G\|[A]_{\mathcal{P}}\right) d\hat{F}_{\mathcal{P}_{c-1}}(\mathcal{P}),\end{aligned}$$

where $F_{\mathcal{P}_{c-1}}$ is the distribution of $\mathcal{P}_{c-1,n}$ w.r.t. the true joint distribution of X and Y. There is however a problem with the above definition of $\hat{\Delta}_c(F;G)$. There are $\binom{n-1}{c-2}$ subsets $\mathcal{P} \in \mathbb{P}_{c-1,n}$ which contain the largest observation from $\mathcal{S}_n$ and which may thus be represented by

$$\mathcal{P} = \{X_{(i_1)}, X_{(i_2)}, \ldots, X_{(i_{c-2})}, X_{(n)}\},$$

where $1 \leq i_1 < i_2 < \ldots i_{c-2} < n$ and where $X_{(1)} < \ldots < X_{(n)}$ are the order statistics of the combined sample $\mathcal{S}_n$. The cth element of the corresponding partition $[A]_{\mathcal{P}}$ is

$A_c =]X_{(n)}, b_u]$ and thus $\hat{\mathrm{P}}_g\left[A_c\right] = 0$. Since this probability estimator is in the nominator of $\mathrm{I}^1\left(\hat{F}; \hat{G} \| [A]_{\mathcal{P}}\right)$, $\hat{\Delta}_c(F;G)$ becomes infinity with probability one. This problem is solved by changing the partition construction rule such that the largest observation is not allowed in $\mathcal{P}$.

Let $\boldsymbol{h}_m = ((x_1, i_1), \ldots, (x_{m-1}, i_{m-1}))$ ($\boldsymbol{h}_1 \equiv 0$), and let $[A](x_1, \ldots, x_{c-1}) = [A]_{\mathcal{P}}$. Since all observations are independent, the statistic becomes

$$\begin{aligned}
\hat{\Delta}_c(F;G) &= \sum_{i_1=1}^{k} \cdots \sum_{i_{c-1}=1}^{k} \int_{\mathcal{S}_{\mathcal{P}}} \hat{\mathrm{I}}^1\left(F;G\|[A]_{\mathcal{P}}\right) \prod_{m=1}^{c-1} d\hat{F}_{x|y}(x_m|i_m, \boldsymbol{h}_m) d\hat{F}_y(i_m|\boldsymbol{h}_m), \\
&= \frac{1}{\binom{n-1}{c-1}} \sum_{1 \le j_1 < \ldots < j_{c-1} < n} \hat{\mathrm{I}}^1\left(F;G\|[A](x_{(j_1)}, \ldots, x_{(j_{c-1})})\right) \\
&= \frac{1}{a_{c-1,n}} \sum_{\mathcal{P} \in \mathbb{P}_{c-1,n}} \hat{\mathrm{I}}^1\left(F;G\|[A]_{\mathcal{P}}\right), \qquad (6.6)
\end{aligned}$$

where $\mathcal{P} = \{x_1, \ldots, x_{c-1}\} \subset \mathcal{S}_x$. It is very interesting to note that there is actually no difference in the calculation of $\hat{\Delta}_c(F;G)$ when the expectation in Equation 6.4 would have been taken w.r.t. the common distribution G, which also must be estimated from the data, and which eventually also would result in Equation 6.6.

6.1.5 The Test Statistic and its Asymptotic Null Distribution

The test statistic based on c-sized SSPs is given by

$$T_{k,c,n} = 2n\hat{\Delta}_c(F;G) = 2n\mathrm{E}_{\hat{F}}\left[\mathrm{I}^1\left(\hat{F}; \hat{G}\|[A](\mathcal{P}_{c-1,n})\right)\right], \qquad (6.7)$$

which is $2n$ times the average of the directed divergence of order 1, over all c sized SSPs that can be constructed according to the construction rule. Now the term SSPkc will be used to refer to this test statistic.

Since the test statistic is very similar to the SSPc test statistic for the one-sample GOF hypothesis, it is expected that the asymptotic null distribution is also similar. First the special case $c = 2$ is discussed. For arbitrary, but finite c, the asymptotic null distribution may be constructed by using similar arguments as those that led to the distribution of the SSPc statistic.

Special Case: $c = 2$

If $c = 2$, then the partition determining subsample set $\mathbb{P}_{1;n}$ is just the sample $\mathcal{S}_n$ with the largest observation deleted, and the sets $\mathcal{P}^i_{1;n}$ may be replaced by the corresponding order statistics $X_{(i)}$ ($i = 1, \ldots, n-1$). Thus, the partitions $[A](\mathcal{P}^i_{1;n})$ may be replaced by $[A](X_{(i)})$ or simply $[A]_i$. For a given sample $a_{1;n} = n - 1$ such partitions can be

constructed. Each SSP $[A]_i$ implies a SSP $[B]_i$ of the sample space $\mathcal{S}_{xy}$ which consists of $2k$ elements:

$$[B]_i = \{B_{i11}, \ldots, B_{i1k}, B_{i21}, \ldots, B_{i2k}\},$$

where $B_{i1j} =]b_l, X_{(i)}] \times \{j\}$ and $B_{i2j} =]X_{(i)}, b_u] \times \{j\}$ $(j = 1, \ldots, k)$.

Before an explicit formula of the test statistic is given, some more notation is introduced. Let R_i denote the rank of X_i among the observation in the combined sample $\mathcal{S}_n$, and let $R_{i(j)}$ denote the rank of X_i among the observation in the jth sample $\mathcal{S}_{n_j}$.

The test statistic is then given by

$$\begin{aligned}
T_{k,2,n} &= 2n \left[\frac{1}{n} \sum_{i=1}^{n-1} \hat{\mathrm{I}}^1 \left(F; G \| [A]_i\right) \right] \\
&= \frac{1}{n-1} \sum_{i=1}^{n-1} n \sum_{m=1}^{k} \sum_{j=1}^{2} \frac{\left(\hat{\mathrm{P}}_{f_{xy}} [B_{ijm}] - \hat{\mathrm{P}}_{g_{xy}} [A_{ij}] \right)^2}{\hat{\mathrm{P}}_{g_{xy}} [A_{ij}]} \\
&= \frac{1}{n-1} \sum_{i=1}^{n-1} n \sum_{m=1}^{k} \frac{n_i}{n} \sum_{j=1}^{2} \frac{\left(\hat{\mathrm{P}}_{f_m} [B_{ijm}] - \hat{\mathrm{P}}_{g} [A_{ij}] \right)^2}{\hat{\mathrm{P}}_{g} [A_{ij}]} \\
&= \frac{1}{n-1} \sum_{i=1}^{n-1} \sum_{m=1}^{k} n_m \left[\frac{\left(\frac{R_{i(m)}}{n_m} - \frac{R_i}{n} \right)^2}{\frac{R_i}{n}} + \frac{\left(\left(1 - \frac{R_{i(m)}}{n_m}\right) - \left(1 - \frac{R_i}{n}\right) \right)^2}{\left(1 - \frac{R_i}{n}\right)} \right] \\
&= \frac{1}{n-1} \sum_{i=1}^{n-1} \sum_{m=1}^{k} \frac{1}{n_m} \frac{\left(n R_{i(m)} - n_m R_i\right)^2}{R_i (n - R_i)}, \qquad (6.8)
\end{aligned}$$

which is up to a multiplicative factor $\frac{n}{n-1}$ exactly the computational formula of the k-sample Anderson-Darling test statistic (Darling, 1957; Pettit, 1976; Scholz and Stephens, 1987). When the distributions are continuous, the probability of the occurrence of ties is zero, and the computational formula reduces to

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \sum_{m=1}^{k} \frac{1}{n_m} \frac{\left(n R_{i(m)} - i n_m\right)^2}{i(n-1)}, \qquad (6.9)$$

which is also given in Scholz and Stephens (1987), up to the factor $\frac{n}{n-1}$. The null distribution of $T_{k,2,n}$ is therefore the same as for the k-sample AD test statistic (see Section 4.2.4), again up the the aforementioned factor. Since $\frac{n}{n-1} \to 1$ as $n \to \infty$, the asymptotic null distributions are exactly the same.

Theorem 6.1 *(Scholz and Stephens, 1987) Under the general k-sample null hypothesis, as $n_i \to \infty$ ($i = 1, \ldots, k$),*

$$T_{k,2,n} \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2,$$

where the Z_j^2 are i.i.d. χ^2_{k-1}-variates.

It is already obvious now that the SSPkc test for the k-sample problem will be a generalization of the Anderson-Darling test. Although the SSP approach that is proposed here for the k-sample problem seems completely analogue as for the one-sample GOF problem, the SSPc test statistic was definitely not a generalization of the AD statistic, because the expectation in the averaged Pearson divergence is taken w.r.t. the true distribution F instead of the distribution G as it is the case with the AD test. In the present context though, as noted before, it does not make a difference whether F or G is taken as soon as these distributions are replaced by their plug-in estimators, which is needed to build the test statistic.

Another interesting property of the SSP2 test statistic is that it is a *rank test statistic*. This is clearly seen from Equation 6.8. The test statistic is only function of the ranking of the observations, but not of the values of the observations themselves.

The General Case

Since an important step that allows the generalization of the distribution of $T_{k,2,n}$ to $T_{k,c,n}$ is only stated as a conjecture, the resulting asymptotic null distribution for arbitrary c is stated only as a proposition, and a sketch of the proof is given. In a later section its validity will be empirically investigated.

Proposition 6.1 *For any finite $c > 1$, under the general k-sample null hypothesis, as $n_i \to \infty$ ($i = 1, \ldots, k$),*

$$T_{k,c,n} \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{\lambda(\lambda+1)} Z_j^2,$$

where the Z_j^2 are i.i.d. $\chi^2_{(c-1)(k-1)}$-variates.

Heuristic Proof. The proof is along the same line as the (heuristic) proof of Proposition 5.1, and is thus accomplished by complete induction.

Suppose the theorem holds for $T_{k,c-1,n}$ ($c - 1 > 1$), then it will be shown that it also holds for $T_{k,c,n}$.

Any c-sized partition $[A]$ of $\mathcal{S}_x$ can be written as

$$[A]_{\mathcal{P}} = [A_s^{c-1}]_{\mathcal{P}} \cup \{]x_{(i_{c-1})}, b_u]\},$$

where $\mathcal{P} = \{x_{(i_1)}, \ldots, x_{(i_{c-1})}\}$ ($1 \leq i_1 < \ldots < i_{c-1} < n$), more specifically, the order statistics are defined on the common sample $\mathcal{S}_n$. Further, as in Theorem 5.1, $[A_s^{c-1}]_{\mathcal{P}}$ is a $(c-1)$-sized partition of the subspace $\mathcal{S}_s^{c-1} =]b_l, x_{(i_{c-1})}] \subset \mathcal{S}_x$. Let $n_{\mathcal{P}}$ be the number of observations in $[A_s^{c-1}]_{\mathcal{P}} \cap \mathcal{S}_n$. For any partition $[A]_{\mathcal{P}}$ another associated partition $[A^2]_{\mathcal{P}}$ is defined as

$$[A^2]_{\mathcal{P}} = \{A_1^2, A_2^2\} = \{]b_l, x_{(i_{c-1})}],]x_{(i_{c-1})}, b_u]\},$$

which is a 2-sized partition of $\mathcal{S}_x$.

Both partitions $[A]_{\mathcal{P}}$ and $[A^2]_{\mathcal{P}}$ uniquely induce partitions of the extended sample space $\mathcal{S}_{xy}$. First, $[A]_{\mathcal{P}}$ induces the $(c-1)k$-sized partition

$$[B]_{\mathcal{P}} = [B_s^{c-1}]_{\mathcal{P}} \cup \{]x_{(i_{c-1})}, b_u]\} \times \mathcal{S}_y.$$

This is schematically shown in Figure 6.2. The partition $[A^2]_{\mathcal{P}}$ induces the $2k$-sized partition

$$[B^2]_{\mathcal{P}} = [A^2]_{\mathcal{P}} \times \mathcal{S}_y,$$

which is illustrated in Figure 6.3.

In Section 6.1.2 it was shown that, for any partition $[A]_{\mathcal{P}}$ the statistic $2n\hat{\mathrm{I}}^1\left(F; G \| [A]_{\mathcal{P}}\right)$ is actually Pearson's statistic for testing independence in a $k \times c$ contingency table. In the same way, it is now easily checked that power divergence statistic based on the partition $[A_s^{c-1}]_{\mathcal{P}}$ is the Pearson statistic for testing independence in a $k \times (c-1)$ table. And, similarly, $2n\hat{\mathrm{I}}^1\left(F; G \| [A^2]_{\mathcal{P}}\right)$ is the Pearson statistic for independence in a $k \times 2$ table. According to Lancaster's partitioning rule (Lancaster, 1969), conditionally on $[A]_{\mathcal{P}}$ these two statistics are asymptotically independent, and distributed as $\chi^2_{(k-1)(c-2)}$ and χ^2_{k-1}, respectively. Applying Lancaster's partitioning rule to all partitions $[A]_{\mathcal{P}}$ ($\mathcal{P} \in \mathbb{P}_{c-1,n}$), the test statistic becomes

$$T_{k,c,n} = \frac{1}{a_{c-1,n}} \sum_{\mathcal{P} \in \mathbb{P}_{c-1,n}} \left(2n\hat{\mathrm{I}}^1\left(F; G \| [A^2]_{\mathcal{P}}\right) + 2n_{\mathcal{P}}\hat{\mathrm{I}}^1\left(F_s; G_s \| [A_s^{c-1}]_{\mathcal{P}}\right)\right),$$

where $n_{\mathcal{P}} = n_{i_{c-1}}$ is still the number of observations in $[A_s^{c-1}]_{\mathcal{P}} \cap \mathcal{S}_n$, and where F_s and G_s indicate that these distributions are restricted to the subspace $\mathcal{S}_s^{c-1} =]b_l, x_{i_{c-1}}]$ (or $\mathcal{S}_s^{c-1} \times \mathcal{S}_y$).

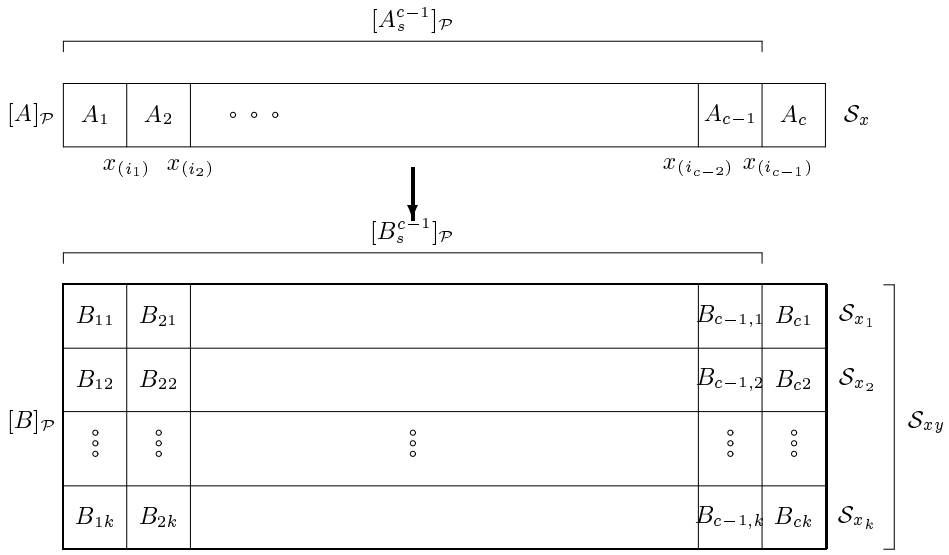


Figure 6.2: The decomposition of $[A]_{\mathcal{P}}$ and the associated decomposition of $[B]_{\mathcal{P}}$. In particular, the relation between $[A_s^{c-1}]_{\mathcal{P}}$ and $[B_s^{c-1}]_{\mathcal{P}}$ is illustrated.

The test statistic, $T_{k,c,n}$, may be written as

$$\frac{1}{n-c}\sum_{i_{c-1}=c-1}^{n-1}\left(2n\hat{\mathrm{I}}^1\left(F;G\|[A^2]_{i_{c-1}}\right)+\frac{1}{a_{c-2,n_{i_{c-1}}}}\sum_{\mathcal{P}\in\mathbb{S}_{i_{c-1}}}2n_{\mathcal{P}}\hat{\mathrm{I}}^1\left(F_s;G_s\|[A_s^{c-1}]_{\mathcal{P}}\right)\right)$$

$$=\frac{1}{n-c}\sum_{i_{c-1}=c-1}^{n-1}\left(2n\hat{\mathrm{I}}^1\left(F;G\|[A^2]_{i_{c-1}}\right)+T_{k,c-1,n_{i_{c-1}}}\right), \tag{6.10}$$

where $T_{k,c-1,n_{i_{c-1}}}$ is the SSPk(c-1) statistic applied to the subspace $\mathcal{S}_S^{c-1}$, and where

$$\mathbb{S}_{i_{c-1}}=\{\mathcal{Q}\in\mathbb{P}_{c-1,n}:\forall X_j\in\mathcal{Q}:X_j\leq X_{i_{c-1}}\}.$$

Since $n_{i_{c-1}}\xrightarrow{p}\infty$ as $n\to\infty$, conditionally on $X_{(i_{c-1})}$, $T_{k,c-1,n_{i_{c-1}}}$ is asymptotically equivalent to the test statistic based on a $(c-1)$-sized partition of which it is assumed that the theorem applies to it. Thus, under H_0, $T_{k,c-1,n_{i_{c-1}}}\xrightarrow{d}\sum_{j=1}^{\infty}\frac{1}{j(j+1)}V_j^2$, where the V_j^2 are i.i.d. $\chi^2_{(k-1)(c-2)}$ variates. Thus, $T_{k,c,n}$ has under H_0 asymptotically the same

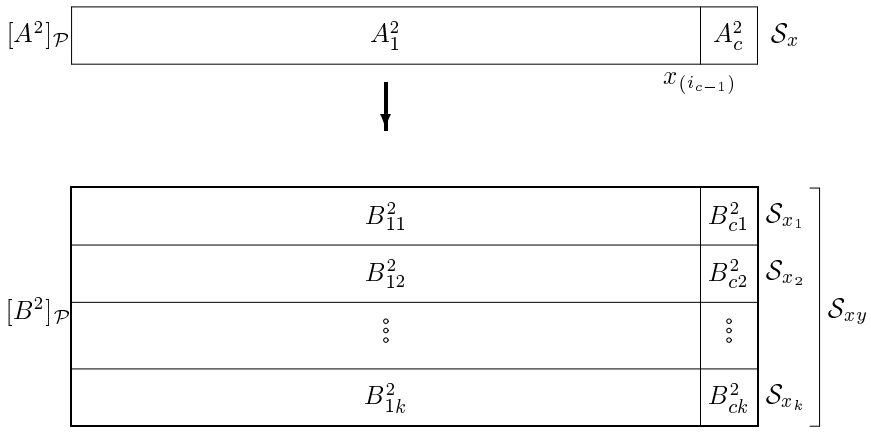


Figure 6.3: The decomposition of $[A]_{\mathcal{P}}$ and the associated decomposition of $[B]_{\mathcal{P}}$. In particular, the relation between $[A^2]_{\mathcal{P}}$ and $[B^2]_{\mathcal{P}}$ is illustrated.

distribution as

$$\frac{1}{n}\sum_{i_{c-1}=1}^{n-1} 2n\hat{\mathrm{I}}^1\left(F;G\|[A^2]_{\mathcal{P}}\right)+\sum_{j=1}^{\infty}\frac{1}{j(j+1)}V_j^2,$$

where the first term is asymptotically equivalent to the k-sample Anderson-Darling statistic (Scholz and Stephens, 1987), which is under H_0, by Theorem 6.1, asymptotically distributed as $\sum_{j=1}^{\infty}\frac{1}{j(j+1)}W_j^2$, where the W_j^2 are i.i.d. χ^2_{k-1}-variates. Although Lancaster's Partitioning rule states that for each $x_{(i_{c-1})}$ the terms between the brackets in Equation 6.10 are independent, it does not imply that the two terms resulting from the sum are still independent, we conjecture here that in this particular case the independence still holds, at least in a first order approximation. Therefore, we propose that the limiting null distribution of $T_{k,c,n}$ becomes

$$\sum_{j=1}^{\infty}\frac{1}{j(j+1)}V_j^2+\sum_{j=1}^{\infty}\frac{1}{j(j+1)}W_j^2=\sum_{j=1}^{\infty}\frac{1}{j(j+1)}Z_j^2,$$

where the Z_j^2 are i.i.d. $\chi^2_{(k-1)(c-1)}$-variates.

According to the complete induction reasoning, it now only has to be shown that the proposition holds for $c-1=2$, which was already proven in Theorem 6.1. □

Since the limiting null distribution of the SSPkc test is exactly the same as for the SSPc test for the one-sample GOF problem, except that the number of degrees of freedom of the

χ^2-variates now also depends on k, the same approximations may be considered. They will be discussed in the next section.

6.1.6 An Approximation to the Asymptotic Null Distribution

Since the asymptotic null distribution is of the same form as in the case of the one-sample GOF test (weighted sum of χ squared variates), the same method of approximation may be considered (Section 5.2.5). The coefficients of the rv $W(a, b, p, f) = a + bZ_p^2 + \frac{1}{2}Y_f^2$, where Z_p^2 and Y_f^2 are i.i.d. χ^2-variates with p and f degrees of freedom, respectively, are determined such that the first four moments are equal to the corresponding moments of the asymptotic null distribution. Hereto the latter moments must be known. These are given in the following corollary, which is a straightforward extension of Corollary 5.5.

Corollary 6.2 *The ith cumulant μ_i of the asymptotic null distribution of $T_{k,c,n}$ is given by*

$$\mu_i = (k-1)(c-1)2^{i-1}(i-1)! \sum_{j=1}^{\infty} \left(\frac{1}{j(j+1)} \right)^i,$$

which is $(k-1)(c-1)$ times the corresponding cumulant of the asymptotic null distribution of $T_{2,2,n}$.

Thus,

$$\begin{aligned} a &= 0.1919467416(c-1)(k-1) \\ b &= 0.1274341352 \\ p &= 2.404852143(c-1)(k-1) \\ f &= 1.003186011(c-1)(k-1). \end{aligned}$$

In Table 6.1 both the (exact) asymptotic and the approximated asymptotic critical values are shown for $k = 2, 3$ and $c = 2, 3, 4$, all at the $\alpha = 0.05$ level.

6.1.7 The SSPkc Test

The α-level SSPkc test $\phi_{\alpha,k,c,n}(\mathcal{S}_n)$ for the k-sample problem is defined as

$$\begin{aligned} \phi_{\alpha,k,c,n}(\mathcal{S}_n) &= 1 \text{ if } T_{k,c,n} > t_{\alpha,k,c,n} \\ \phi_{\alpha,k,c,n}(\mathcal{S}_n) &= 0 \text{ if } T_{k,c,n} \leq t_{\alpha,k,c,n}, \end{aligned}$$

where $t_{\alpha,k,c,n}$ is the $(1-\alpha)$-th percentile of the exact null distribution of $T_{k,c,n}$. The asymptotic null distribution which is given in the previous section only gives asymptotic percentiles $t_{\alpha,k,c}$. Using $t_{\alpha,k,c}$ as a critical value instead of the exact percentiles $t_{\alpha,k,c,n}$,

defines the asymptotic test $\phi_{\alpha,k,c}$. In Section 6.1.8 the convergence of the exact to the asymptotic percentiles will be assessed. In Section 6.1.8 it will be shown how exact critical values may be computed.

The power function for an alternative $f_1 = f_{xy,1}$ is given by $\beta_{\alpha,k,c,n} = \mathrm{E}_{f_1}\left[\phi_{\alpha,k,c,n}(\mathcal{S}_n)\right]$.

Theorem 6.2 *For any finite $c \geq 2$ and any finite k, the SSPkc test is consistent against essentially any alternative, provided that $n_i \to \infty$ for all $i = 1, \ldots, k$.*

Proof. The theorem is proven in exactly the same way as Theorem 5.2. In the proof it is supposed that the the SSP2 test ($c = 2$) is consistent, which is proven by Scholz and Stephens (1987). □

6.1.8 Permutation Test

In Section 5.2.6 a straightforward solution to estimate critical values $t_{\alpha,c,n}$ to use for the SSPc GOF test was given by simulating the exact null distribution. In this section it will be shown that a permutation test can be constructed, but first the construction of permutation tests will be discussed in general.

Permutation and Randomization Tests

The idea behind permutation tests dates back from Fisher (1935). Here, the notation of Hoeffding (1952) is more or less used.

Consider the sample space $\mathcal{S}_x$ of the rv $\boldsymbol{X}$, which may represent a complete sample of observations as well. Let $\mathcal{H}$ be a finite group of transformations h of $\mathcal{S}_x$ onto itself. Let m denote the number of transformations in $\mathcal{H}$. Further, assume that the null hypothesis implies that the distribution of X is invariant under the transformations in $\mathcal{H}$, i.e. for every $h \in \mathcal{H}$, $h(\boldsymbol{X})$ and $\boldsymbol{X}$ have the same distribution under H_0. Thus, conditionally on $\boldsymbol{X}$, each $h(\boldsymbol{X})$ has equal probability $\frac{1}{m}$.
Let $T(\boldsymbol{X})$ be a test statistic for testing H_0. Let

$$T^{(1)}(\boldsymbol{X}) \leq T^{(2)}(\boldsymbol{X}) \leq \ldots \leq T^{(m)}(\boldsymbol{X}) \tag{6.11}$$

be the ordered values of $T(h(\boldsymbol{X}))$ as h varies in $\mathcal{H}$. Conditionally on the sample $\boldsymbol{X}$, the permutation distribution of $T(\boldsymbol{X})$ is completely given by the series in Equation 6.11, in which to each $T^{(i)}(\boldsymbol{X})$ ($i = 1, \ldots, m$) the probability $\frac{1}{m}$ is assigned. Thus, in general the permutation distribution is a conditional distribution, whereas the asymptotic distributions which are typically given in this thesis are all unconditional distributions. Since the SSPkc test statistic is a rank statistic and since the parent distributions of the observations are assumed to be continuous, the conditional and the unconditional null distributions coincide (e.g. Puri and Sen, 1971). For a given nominal significance level α, the integer k

is defined as

$$k = m - \lfloor m\alpha \rfloor,$$

where $\lfloor m\alpha \rfloor$ denotes the largest integer less than or equal to $m\alpha$. Let $M^+(\boldsymbol{X})$ denote the number of values $T^{(i)}(\boldsymbol{X})$ ($i = 1, \ldots, m$) that are greater than $T^{(k)}(\boldsymbol{X})$.

The α-level permutation test is defined as

$$\begin{aligned} \phi_\alpha(\boldsymbol{X}) &= 1 \text{ if } T(\boldsymbol{X}) > T^{(k)}(\boldsymbol{X}) \\ \phi_\alpha(\boldsymbol{X}) &= 0 \text{ if } T(\boldsymbol{X}) \leq T^{(k)}(\boldsymbol{X}). \end{aligned}$$

An essential difference with traditional tests is that the critical values are random variables. More specifically, they are estimated from the sample at hand.

The next corollary shows that the α permutation test has not necessarily size α. In particular, the test is often conservative. The larger m, the less conservative the test will be.

Corollary 6.3 *The size of the α-level permutation test is at most α.*

Proof. First note that

$$\sum_{h \in \mathcal{H}} \phi_\alpha(h(\boldsymbol{X})) = M^+(\boldsymbol{X}) \leq \lfloor m\alpha \rfloor.$$

Then, under H_0,

$$\begin{aligned} \lfloor m\alpha \rfloor &\geq \mathrm{E}_g \left[\sum_{h \in \mathcal{H}} \phi_\alpha(h(\boldsymbol{X})) \right] \\ &= \sum_{h \in \mathcal{H}} \mathrm{E}_g \left[\phi_\alpha(\boldsymbol{X}) \right] \\ &= m \mathrm{E}_g \left[\phi_\alpha(\boldsymbol{X}) \right]. \end{aligned}$$

Hence, the size is $\mathrm{E}_g \left[\phi_\alpha(\boldsymbol{X}) \right] \leq \lfloor m\alpha \rfloor / m \leq \alpha$. □

One solution to make the permutation test of size α is to extend its definition to a *randomization* test. This was originally suggested by Eden and Yates (1933) and then Pitman (1937a, 1937c, 1937b). A recent treatment of the subject can be found in textbooks by e.g. Edgington (1995) and Manly (1997).
Let $M^0(\boldsymbol{X})$ denote the number of values $T^{(i)}(\boldsymbol{X})$ ($i = 1, \ldots, m$) that are equal to $T^{(k)}(\boldsymbol{X})$. The α-level randomization test is defined as

$$\begin{aligned} \phi_\alpha(\boldsymbol{X}) &= 1 \text{ if } T(\boldsymbol{X}) > T^{(k)}(\boldsymbol{X}) \\ \phi_\alpha(\boldsymbol{X}) &= A(\boldsymbol{X}) \text{ if } T(\boldsymbol{X}) = T^{(k)}(\boldsymbol{X}) \\ \phi_\alpha(\boldsymbol{X}) &= 0 \text{ if } T(\boldsymbol{X}) < T^{(k)}(\boldsymbol{X}), \end{aligned}$$

where

$$A(\boldsymbol{X}) = \frac{m\alpha - M^+(\boldsymbol{X})}{M^0(\boldsymbol{X})}.$$

Corollary 6.4 *The α-level randomization test is of size α.*

Proof. First note that

$$\sum_{h\in\mathcal{H}} \phi_\alpha(h(\boldsymbol{X})) = M^+(\boldsymbol{X}) + A(\boldsymbol{X})M^0(\boldsymbol{X}) = m\alpha.$$

Then, under H_0,

$$\begin{aligned} m\alpha &= \mathrm{E}_g\left[\sum_{h\in\mathcal{H}} \phi_\alpha(h(\boldsymbol{X}))\right] \\ &= \sum_{h\in\mathcal{H}} \mathrm{E}_g\left[\phi_\alpha(\boldsymbol{X})\right] \\ &= m\mathrm{E}_g\left[\phi_\alpha(\boldsymbol{X})\right]. \end{aligned}$$

Hence, the size is $\mathrm{E}_g\left[\phi_\alpha(\boldsymbol{X})\right] = \alpha$. □

Thus, whatever the sample size, the size of the randomization test is always equal to the nominal significance level α. Moreover, this result does not depend on the true distribution G under the null hypothesis, and therefore the test is distribution-free. Thus this test is similar and of size α.

In practice, when m is large, it may be computationally (almost) unfeasible to calculate all $h(\boldsymbol{X})$. In these occasions, a Monte Carlo approximation may be performed by considering only a randomly chosen subset of $m' < m$ transformations h from $\mathcal{H}$.

The SSPkc Permutation Test

Basically only one condition must be fulfilled in order to construct a permutation test: the null hypothesis must imply that the distribution G is invariant under a group $\mathcal{H}$ of transformations. Here, for the general k-sample problem, the transformations h are permutations of the observations over the k samples. Indeed, the null hypothesis says that the k distributions are equal, and therefore, observations are interchangeable between samples. This is treated is detail in Hájek and Šidák (1967).

In particular, let the combined sample $\mathcal{S}_n$ be ordered such that the first n_1 elements are those of the 1st sample, the next n_2 elements those of the 2nd sample, etc. Let $h(\mathcal{S}_n)$ be a permutation of the observation in $\mathcal{S}_n$. The first n_1 elements are considered to be the elements of the 1st transformed sample, the next n_2 elements those of the 2nd transformed sample, etc. There are $n!$ such permutations, but since all the observations are assumed

to be independent, the order of the observations within a sample is of no importance, and thus $\mathcal{H}$ must only contain the

$$m = \frac{n!}{\prod_{i=1}^{k} n_i!}$$

transformations that result in different sample configurations.

In Section 6.1.5 it has been shown that the SSPkc test statistic is a rank statistic. Therefore, the computation of the m test statistics $T_{k,c,n}^{(i)}$ ($i = 1, \ldots, m$) must be done only once for each sample size n. But still, when n is rather large, this may be unfeasible, in which case a Monte Carlo approximation may be in place. E.g. when $n = 12$ and $k = 2, 3$ or 4, then $m = 924, 34650$ and 369600, respectively. But when $n = 24$, then $m = 2704156, 9465511770$ and 2308743493056, respectively.

The properties of permutation and randomization tests which are given in the previous section, apply directly to the SSPkc permutation and SSPkc randomization tests.

6.1.9 Convergence to the Limiting Null Distribution

Monte Carlo approximations to some exact critical values for the SSPk2, SSPk3 and SSPk4 tests at the $\alpha = 0.05$ level are given here, for $n = 12, 24$ and $n = 48$, and for $k = 2$ and $k = 3$. The approximation are based on 10000 simulation runs. The results are shown in Table 6.1. To make a comparison with the asymptotic critical values possible, these values are mentioned in the table as well. Moreover, next to each estimated exact critical value, the estimated exact level, when the asymptotic critical value would have been used, is given.

Table 6.1: Approximated exact critical values ($\alpha = 0.05$) of the SSPk2, SSPk3 and SSPk4 tests, for the 2-sample and the 3-sample problem with $n = 12, 24$ and $n = 48$. The entries at $n = \infty$ are the critical values derived from the asymptotic null distributions. Also the approximated asymptotic critical values are shown. Between brackets, the estimated exact level, when the asymptotic critical values would have been used, is shown.

k	n	SSP2	SSP3	SSP4
2	12	2.548 (0.055)	4.099 (0.047)	5.199 (0.040)
	24	2.529 (0.052)	4.118 (0.051)	5.572 (0.048)
	48	2.492 (0.050)	4.184 (0.053)	5.679 (0.054)
	∞	2.492	4.111	5.585
	∞ (approx.)	2.489	4.093	5.526
3	12	3.969 (0.040)	7.033 (0.051)	9.438 (0.041)
	24	3.984 (0.044)	6.985 (0.050)	9.440 (0.041)
	48	4.095 (0.047)	6.931 (0.048)	9.511 (0.047)
	∞	4.111	6.985	9.656
	∞ (approx.)	4.093	6.891	9.491

The results suggest that the convergence to the limiting distribution is rather fast when compared to the convergence rate of the SSPc test (Section 5.2.6). The latter convergence was very slow, even for $c = 2$, despite the fast convergence of the corresponding AD test. Pettit (1976) for the 2-sample Anderson-Darling test, and Scholz and Stephens (1987) for the k-sample AD test, also observed a fast convergence. Both tests are special cases of the present SSPkc test ($c = 2$). An important reason why the convergence is reasonably fast here as compared to the one-sample GOF problem, is that for the GOF problem the calculation of the SSPc test statistic required one additional estimator $\hat{F}$ for the true distribution F, while the AD test made use of the given distribution G, and thus, under H_0, the convergence of the SSPc test rested on one additional $\hat{F} \xrightarrow{d} G$ convergence. Here, on the other hand, both the SSPkc and the AD test statistics rely on exactly the same convergences.

Notice that when $\alpha = 0.05$ the use of the asymptotic critical values gives rather good results; the tests are biased by at most about 1%.
Finally, it is worth noting that, at least under the current simulation conditions, the size of the tests even would have been closer to the nominal level if the approximate asymptotic critical values would have been used.

6.1.10 Ties

It was argued before that the assumption of continues distributions implies that the probability of the occurrence of ties is zero, and that therefore there is no need to discuss the topic, at least from a theoretical point of view. In practice, however, measurements are made with only a finite accuracy often leading to tied observations. In this section we comment briefly on three aspects of the occurrence of ties: the test statistic $T_{k,c,n}$, the asymptotic null distribution and the exact (permutational) null distribution.

When $c = 2$ the test statistic is given in Equation 6.8, which was under the no-ties assumption simplified to Equation 6.9. It is actually the former formula which takes ties into account. Also for general c the proposed test statistic (Equation 6.7) is prepared for tied observations, though there is an additional complication in this general case. When e.g. $c = 3$ then two observations determine a SSP. If, however, these two observations are tied, then actually only a 2-sized partition is constructed. This problem may be overcome by changing the partition construction rule accordingly.
The reason why the statistic accounts naturally for ties is that it is basically a plug-in estimator of a functional. Such statistics handle ties in a way that is directly implied by the definition of the EDF used as plug-in estimator of the true CDF. Since in our case, the EDF evaluated in x is just the number of observations in the sample that are not greater than x, divided by n, the rank of a tied observation X_i equals the number of observations in the sample that are not greater than X_i. Thus the observations in a group of tied observations will all get the same rank, which is the rank of the largest observation in that group. Often the definition of the EDF is changed to handle ties. In particular the average of the left and the right limit of the ordinary EDF might be used. This alternative defini-

tion leads to mid-ranks. Mid-ranks are often encountered and recommended (Lehmann, 1975).

In this work the asymptotic distribution of $T_{k,c,n}$ is derived under the assumption of continuous distribution, and thus with probability one no ties occur. When the assumption is however weakened to allow also for discrete distributions one should take ties into account. A solution for the asymptotic null distribution may be obtained by proceeding in a similar way as Scholz and Stephens (1987). It is however expected that the limiting distribution will depend on the true common distribution G through $G(G^{-1}(u))$ ($u \in [0,1]$).

Also the permutation law for finite sample sizes will change. Since the permutation distribution is essentially a distribution, conditional on the sample, the presence of ties will have an influence on the distribution, especially when the sample size is small.

6.1.11 Examples

All SSPkc tests that are performed here, are based on critical values that have been estimated from the simulated null distributions (10000 simulation runs). When the sample sizes n_j ($j = 1, \ldots, k$) are not all equal (unbalanced), then the null distribution is simulated taking the unbalancedness into account by considering permutations which are restricted to result in samples with n_j ($j = 1, \ldots, k$) observations.

EXAMPLE 6.1. **Gravity Data**
It is of particular interest to test whether the distribution of the observations in the first sample is the same as the distribution of the observations in the last sample. These two samples correspond to the first (in time) and the last series of measurements. It's important to stress that one is not only interested in a difference in mean. Thus a traditional t-test would not suffice, and a test for the 2-sample GOF problem is needed.

The SSP2c test, with $c = 2, 3$ and $c = 4$ is applied to the data. The results are shown in Table 6.2. The table shows the critical values at $\alpha = 0.05$ as well. In the first series there are 8 observations and in the last series there are 13.

For each of the considered SSP sizes c, it is concluded that the distributions are significantly different at the 5% level of significance. The boxplot in Figure 2.5 suggests that the difference may be due to both a difference in mean and in variance.

Table 6.2: The results of the SSPkc test on the Gravity data. Between brackets the p-value is given. Also the critical values are shown.

c	$t_{2,c,21}$	$t_{0.05,2,c,21}$
2	2.634 (0.043)	2.498
3	4.868 (0.029)	4.273
4	6.699 (0.008)	5.457

□

EXAMPLE 6.2. **Sleep data**
The 62 animals are classified into 5 different classes of exposure to danger. The number of observations in the 5 classes are 19, 14, 10, 10, and 9. One of the questions that is of interest is whether or not there is a difference in the distributions of the brain weight between the 5 classes of exposure to danger.
The results are given in Table 6.3.

Table 6.3: The results of the SSPkc test on the Sleep data. Between brackets the p-value is given. Also the critical values are shown.

c	$t_{5,c,62}$	$t_{0.05,5,c,62}$
2	10.332 (0.001)	6.727
3	17.066 (0.002)	12.172
4	22.405 (0.005)	16.980

Thus, for all three SSP sizes one may conclude that the brain weight of animals that are exposed differently to danger, not all come from the same distribution. □

Note that the p-values of the SSPkc tests on the Gravity Data decrease as c increases, whereas for the Sleep Data the opposite trend is detected. This illustrates the importance of a good choice for c.

6.2 The Data-Driven SSP Test

Although for any finite c the SSPkc test is consistent, it may be expected that with finite sample sizes, under certain alternatives the power will be higher for small values of c, but lower under other alternatives (this will be illustrated in a simulation study, later in this chapter). By making the test data-driven, it is hoped that a "good" choice for c w.r.t. the power is made by the data itself.

In general, most of the theory given in Section 5.5 remains valid. In fact, the few changes that must be made, are all imposed by the change in the selection rule, which is a consequence of the difference in the number of degrees of freedom of the χ^2-variates in the limiting null distribution.

Definition 6.1 *The* **SSP Selection Rule**, *which selects the partition size C_n out of the set Γ of permissible orders, is given by*

$$C_n = \mathit{ArgMax}_{c\in\Gamma}\left[T_{k,c,n} - 2(k-1)(c-1)\ln a_n\right], \tag{6.12}$$

where $2(k-1)(c-1)\ln a_n$ is the penalty.

6.2.1 The Test Statistic and its Asymptotic Null Distribution

The test statistic is given by

$$T_{k,C_n,n}.$$

The asymptotic null distribution of the statistic $T_{k,C_n,n}$ is given in the following theorem and proposition.

Theorem 6.3 *Let Γ be a set of permissible orders. Let c_m denote the Minimal Partition Size. Suppose that $a_n \to \infty$ as $n \to \infty$. Then, under H_0*

$$C_n \xrightarrow{p} c_m$$

as $n \to \infty$.

Proof. The proof is almost exactly the same as the proof of Theorem 5.6, with $c-1$ and $c_m - 1$ changed to $(k-1)(c-1)$ and $(k-1)(c_m-1)$, respectively. Further, the asymptotic mean and variance of the test statistic now also depend on k, giving

$$\begin{aligned} \mathrm{E}_g\left[T_{k,c,n}\right] &\to (k-1)(c-1) \\ \mathrm{Var}_g\left[T_{k,c,n}\right] &\to (k-1)(c-1)\sigma \text{ as } n \to \infty, \end{aligned}$$

where σ is the asymptotic standard deviation of $T_{k,2,n}$ under H_0, i.e. $\mathrm{Var}_g\left[T_{k,2,n}\right] \to \sigma < \infty$ as $n \to \infty$.

Since k and all $c \in \Gamma$ are assumed finite, and since σ is also finite, exactly the same arguments as those used in the proof of Theorem 5.6 apply, by which this theorem is also proven. □

Proposition 6.2 *Let Γ be a set of permissible orders. Let c_m denote the Minimal Partition Size. Suppose that $a_n \to \infty$ as $n \to \infty$. Then, under H_0,*

$$T_{k,C_n,n} \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_j^2,$$

where the Z_j^2 are i.i.d. $\chi^2_{(k-1)(c_m-1)}$ variates.

Proof. The proof is exactly the same as the proof of Proposition 5.3, except that now Theorem 6.3 and Proposition 6.1 are needed instead of Theorem 5.6 and Proposition 5.1, respectively. □

6.2.2 The Data-Driven SSP Test

The k-sample data-driven SSPkc test will again be referred to as the SSPdd test.

The α-level SSPdd test $\phi_{\alpha,k,\Gamma,n}(\mathcal{S}_n)$, based on the set Γ of permissible orders, is defined as

$$\begin{aligned}\phi_{\alpha,k,\Gamma,n}(\mathcal{S}_n) &= 1 \text{ if } T_{k,C_n,n} > t_{\alpha,k,C_n,n} \\ \phi_{\alpha,k,\Gamma,n}(\mathcal{S}_n) &= 0 \text{ if } T_{k,C_n,n} \leq t_{\alpha,k,C_n,n},\end{aligned}$$

where C_n is the selected order based on a SSP selection rule.

The following two results are essentially the same as Theorems 5.7 and 5.8. Their proofs must only be adapted to the present notation by introducing k.

Theorem 6.4 *The α-level k-sample SSPdd test based on the set Γ of permissible orders, with minimal order c_m, and based on a selection rule for which $a_n \to \infty$ as $n \to \infty$, is similar and of size α.*

Theorem 6.5 *For any set Γ of permissible partition sizes, and any selection rule, the k-sample SSPdd test is consistent against essentially any alternative.*

6.2.3 The SSPdd Permutation Test

Since the null hypothesis of course still implies the invariance of the distribution of $\mathcal{S}_n$ under the group of transformations $\mathcal{H}$, the SSPdd test may also be implemented as a permutation test. This method gives exact results for finite n, but as noted earlier, the feasibility of the enumeration of all permutations decreases rapidly with increasing sample size n, in which case a Monte Carlo approximation must be considered.

When all permutations would be enumerated, then also the exact probabilities $\mathrm{P}_g\,[C_n = c]$, for finite n, could be computed. A Monte Carlo approximation (10000 runs) to these calculations is done here in order to compare the penalties $a_n = n^{1/2}$ (BIC, Schwartz (1978)), $a_n = (\ln n)^{1/2}$ (LL, Hannan and Quinn (1979)) and $a_n = \frac{\ln(n)}{1.38}$ (pAIC) for the set of permissible partition sizes $\Gamma = \{2, 3, 4\}$, with minimal partition size $c_m = 2$, and for sample sizes $n = 12, 24$ and $n = 48$. The results for $k = 2$ and $k = 3$ are shown in Table 6.4.

When $k = 2$, the results suggest, as it was for the SSPdd test for the one-sample GOF problem, that the convergence to the minimal order is faster with the BIC penalty as with the two other penalties. The slowest convergence is obtained with the LL penalty. When $k = 3$, however, the convergence is with all penalties rather fast, at least for the sample sizes considered here.

Table 6.4: The estimated exact probabilities $\mathrm{P}_g\,[C_n = c]$.

		$k=2$			$k=3$		
penalty	n	$c=2$	$c=3$	$c=4$	$c=2$	$c=3$	$c=4$
BIC	12	0.994	0.006	0.000	0.995	0.005	0.000
	24	0.999	0.001	0.000	0.993	0.007	0.000
	48	1.000	0.000	0.000	1.000	0.000	0.000
LL	12	0.308	0.023	0.669	0.996	0.004	0.000
	24	0.661	0.062	0.277	0.994	0.006	0.000
	48	0.814	0.055	0.131	0.832	0.083	0.085
pAIC	12	0.748	0.097	0.155	0.796	0.128	0.076
	24	0.896	0.071	0.033	0.974	0.022	0.004
	48	0.954	0.034	0.012	0.986	0.012	0.002

6.2.4 Examples

EXAMPLE 6.3. **Gravity data**
Series 1 and 8 are again compared, but this time with the SSPdd test. The set of permissable orders is taken to be $\Gamma = \{2, 3, 4\}$. The BIC criterion resulted in $c_n = 2$, and thus $t_{2,c_n,n} = t_{2,2,n} = 2.634$ ($p = 0.043$). Both the pAIC and the LL based penalty, on the other hand, gave $c_n = 4$ and $t_{2,c_n,n} = t_{2,4,n} = 6.699$ ($p = 0.008$).
In conclusion, the three data-driven tests indicate a significant difference in distribution between series 1 and series 8. □

EXAMPLE 6.4. **Sleep data**
With the penalties BIC and pAIC the 2-sized partition was selected, resulting in $t_{5,c_n,n} = t_{5,2,n} = 10.332$ ($p = 0.001$). The LL criterion led to $c_n = 3$, and thus $t_{5,c_n,n} = t_{5,3,n} = 17.066$ ($p = 0.002$). □

6.3 Power Characteristics

The power of both the SSPkc test and the SSPdd test are compared to some other tests. Isotones will be presented first to given an overall view of the relative sensitivity of the different tests. In a second phase, for some interesting alternatives powers will be estimated in a small simulation study. Alternatives for $k = 2$ and for $k = 3$ are considered.

6.3.1 Other Tests for the k-sample problem

When $k = 2$, probably the best known test is the **Kolmogorov-Smirnov** (KS) test, which was discussed in Section 4.2.3. This test is included in the present study, although it is generally known that this consistent omnibus test does not have very good power characteristics (e.g. Stephens, 1986). For arbitrary k, many extensions to the KS test exist (see Hájek and Šidák, 1967, for an overview). The KS statistic which is considered in this study is given by (Chang and Fisz, 1957; Fisz, 1960; Dwass, 1960)

$$D_{k,n} = \max_{j=2,\ldots,k} \max_{i=1,\ldots,n} |\hat{F}_{j,n}(X_i) - \frac{1}{\sum_{l=1}^{j-1} n_l} \sum_{l=1}^{j-1} n_l \hat{F}_{l,n}(X_i)|.$$

The second test that is included is the **Kruskal-Wallis** (KW) test (Kruskal, 1952; Kruskal and Wallis, 1952), which is equivalent to the Wilcoxon / Mann-Whitney test (Wilcoxon, 1945; Mann and Whitney, 1947) when $k = 2$. This test is however not omnibus consistent, but still it is included here, mainly because of its popularity. The test is most often used for the k-sample location shift problem, for which it is only of size α if the k populations F_i ($i = 1, \ldots, k$) are of the same shape (i.e. all cumulants of order two and higher are equal). If the latter condition does not hold, then the size of the test, under the hypothesis of no location shift, may be larger than the nominal level α. Within the framework of the general k-sample problem this means that the KW test is also sensitive to alternatives different from a location shift. For this reason the test is also considered.

The **Anderson-Darling** (AD) k-sample test (Pettit, 1976; Scholz and Stephens, 1987) is automatically included is this study, since this test is a special case of the SSPkc test ($c = 2$).

Although throughout this thesis there has frequently been made reference to a data-driven Neyman-type test of Janic-Wróblewska and Ledwina (2000), the test is not included in the simulation study because it has been published only after the simulation study that is presented in this section was completed.

The exact (permutation) distributions of all these tests are simulated (10000 runs). The critical values are given in Table 6.5.

Table 6.5: Critical values at the 5%-level for $n = 24$ and $n = 48$ for the k-sample tests used in this thesis.

n	k	SSP2	SSP3	SSP4	KS	KW
24	2	2.529	4.118	5.572	0.500	3.853
	3	3.984	6.985	9.440	0.625	5.780
48	2	2.493	4.184	5.679	0.375	3.757
	3	4.095	6.931	9.511	0.438	5.857

6.3.2 The Alternatives

As with the one-sample GOF problem, there are an infinite number of alternatives of which the small sample properties could be studied. Three families of alternatives are considered for both $k = 2$ and $k = 3$. The first type of alternatives are such that for all F_i ($i = 1, \ldots, k$) the means are equal, but the distributions may differ in all other moments (scale, skewness, kurtosis, ...). The second family consists of contaminated normal distributions, and the third family is a scale-location family of normal distributions.

For the first type, the Tukey-Lambda (TL) family is used in the following way. In contrast to the use of this family in the power study of the SSP test for the one-sample GOF null hypothesis, here k distributions must be specified under the alternative hypothesis. All k distributions are taken to be members of the TL family, but only the kth distribution is different from the first $k-1$ distributions. These first $k-1$ distributions are zero-mean normal distributions ($\lambda_1 = \lambda_2 = 0.1349$). The family of alternatives is then indexed by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ which refers to the kth TL distribution. Since the mean of a TL distributed rv changes with $\boldsymbol{\lambda}$, a correction is needed. Let W_λ a TL distributed rv,

$$W_\lambda = \frac{U^{\lambda_1} - 1}{\lambda_1} - \frac{(1-U)^{\lambda_2} - 1}{\lambda_2},$$

where $U \sim U(0,1)$, then the rv X_λ from the kth distribution F_k is given by

$$X_\lambda = W_\lambda - \mathrm{E}\left[W_\lambda\right] + \mu,$$

where μ is the mean of the first $k-1$ distribution; here $\mu = 0$ for the (quasi) normal distribution $\boldsymbol{\lambda} = (0.1349, 0.1349)$.

The contaminated normal (mixture distribution) alternatives also consist of $k-1$ equal distributions, which are all standard normal, and 1 parameterized distribution, which is supposed to be the kth distribution. The density of the 2-parameter family is given by

$$f_\theta(x) = (1-\gamma)g(x) + \gamma g\left(\frac{x-\delta}{0.001}\right),$$

where g is the density of the first $k-1$ standard normal distributions. The parameter γ correspond to the fraction or probability mass of the distribution that is contaminated with a normal distribution with mean δ and standard deviation 0.001.

As with the two former families, the scale-location family is also used to simulate one of the k samples. The remaining $k-1$ samples are standard normal. As before, let $g(x)$ denote the df of a standard normal distribution, then the distribution $f_\theta(x)$, indexed by $\boldsymbol{\theta} = (\mu, \sigma)$, is the df of a rv $X \sim N(\mu, \sigma^2)$, or equivalently

$$f_\theta(x) = g\left(\frac{x-\mu}{\sigma}\right).$$

6.3.3 Results

Table 6.6: The estimated powers for some alternatives of the Tukey-Lambda family in the 2-sample problem, based on 10000 simulation runs. The first distribution is a normal distribution ($\boldsymbol{\lambda} = (0.1349, 0.1349)$).

n	λ_1	λ_2	power							
			SSP2	SSP3	SSP4	dd - BIC	dd - LL	dd - pAIC	KS	KW
24	1	1	0.153	0.313	0.360	0.154	0.363	0.361	0.100	0.065
24	1.5	1.5	0.309	0.614	0.676	0.318	0.676	0.673	0.196	0.081
24	3	3	0.730	0.952	0.963	0.742	0.963	0.964	0.479	0.108
24	6	6	0.935	0.997	0.999	0.948	0.999	0.999	0.635	0.123
24	2	0	0.080	0.137	0.164	0.081	0.165	0.165	0.063	0.062
24	4	0	0.104	0.169	0.208	0.105	0.209	0.213	0.088	0.079
24	4	1.5	0.533	0.846	0.885	0.544	0.885	0.882	0.340	0.089
24	-1	-1	0.154	0.326	0.379	0.157	0.380	0.377	0.116	0.064
24	1.5	1.75	0.334	0.662	0.717	0.345	0.717	0.715	0.220	0.068
48	1	1	0.506	0.744	0.810	0.506	0.811	0.757	0.206	0.071
48	1.5	1.5	0.870	0.974	0.986	0.780	0.986	0.978	0.565	0.072
48	3	3	1.000	1.000	1.000	1.000	1.000	1.000	0.958	0.093
48	2	0	0.162	0.292	0.373	0.162	0.373	0.340	0.123	0.068
48	4	0	0.215	0.404	0.490	0.215	0.490	0.450	0.149	0.057
48	4	1.5	0.991	1.000	1.000	0.991	1.000	1.000	0.849	0.078
48	-1	-1	0.531	0.750	0.807	0.531	0.807	0.757	0.209	0.054
48	1.5	1.75	0.917	0.984	0.988	0.917	0.988	0.985	0.597	0.084

Tukey-Lambda

The isotones for the TL family of alternatives are shown in Figures 6.4 and 6.5. The estimated powers are shown in Tables 6.6 and 6.7.

In general the estimated powers confirm the relative positions of the isotones. First the $k = 2$ case is discussed.

- Moving away from the point $\lambda_1 = \lambda_2 = 0.1349$, representing the null hypothesis, along the diagonal towards larger λ-values (symmetric distributions with increasing kurtosis), the first isotones that are crossed, are those of the data-driven SSP tests based on the LL and the pAIC criteria, and the SSP24 test. The SSPdd-LL test even has a slightly higher power, especially when $n = 48$. The power of the SSP24 test is almost indistinguishable from the LL based tests. When $n = 24$ also the SSPdd-pAIC test behaves very similarly. With both sample sizes the order of the powers of the SSP2c tests is SSP24 $>$ SSP23 $>$ SSP22, where especially the latter has a far lower power. The power of the BIC based data-driven test was estimated a little bit higher than the power of the SSP22 test, especially when $n = 24$. Recall that the SSP22 test is exactly equivalent to the 2-sample Anderson-Darling test. The reference tests KS and KW resulted clearly in much smaller powers. The former always had larger powers, which was as expected since it is an omnibus test, whereas the KW test is essentially a test for the k-sample location shift problem and in the present simulation study the mean was kept constant.
- In the other direction along the diagonal, into the direction of alternatives with heavier tails, more or less the same behaviour is observed.

Table 6.7: The estimated powers for some alternatives of the Tukey-Lambda family in the 3-sample problem, based on 10000 simulation runs. The first two distributions are normal distributions ($\boldsymbol{\lambda} = (0.1349, 0.1349)$).

n	λ_1	λ_2	power							
			SSP2	SSP3	SSP4	dd - BIC	dd - LL	dd - pAIC	KS	KW
24	1	1	0.103	0.147	0.192	0.103	0.196	0.154	0.032	0.058
24	1.5	1.5	0.163	0.259	0.371	0.163	0.372	0.292	0.040	0.069
24	3	3	0.259	0.579	0.769	0.259	0.769	0.681	0.104	0.085
24	5	5	0.340	0.786	0.935	0.340	0.935	0.896	0.146	0.095
24	2	0	0.078	0.091	0.119	0.078	0.120	0.102	0.026	0.052
24	4	0	0.090	0.117	0.147	0.090	0.150	0.126	0.036	0.061
24	4	1.5	0.223	0.449	0.623	0.223	0.623	0.525	0.076	0.081
24	4	-1	0.956	0.982	0.991	0.956	0.991	0.981	0.897	0.787
24	4	-2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997
24	-1	-1	0.115	0.195	0.242	0.115	0.246	0.199	0.047	0.056
24	1.5	1.75	0.154	0.294	0.413	0.154	0.414	0.333	0.039	0.058
48	1	1	0.193	0.354	0.511	0.193	0.508	0.306	0.078	0.049
48	2	2	0.646	0.935	0.973	0.646	0.973	0.917	0.431	0.066
48	3	3	0.894	0.990	0.998	0.894	0.998	0.990	0.741	0.080
48	6	6	0.999	1.000	1.000	0.999	1.000	1.000	0.968	0.084
48	2	0	0.125	0.170	0.239	0.125	0.237	0.150	0.091	0.066
48	4	0	0.138	0.212	0.310	0.138	0.304	0.187	0.096	0.061
48	6	0	0.140	0.237	0.340	0.140	0.342	0.211	0.094	0.064
48	10	0	0.141	0.234	0.345	0.141	0.341	0.207	0.107	0.065
48	4	1	0.513	0.848	0.924	0.513	0.924	0.820	0.311	0.063
48	2	-1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.969
48	-1	-1	0.327	0.511	0.626	0.327	0.625	0.463	0.186	0.081
48	1.5	1.75	0.475	0.808	0.906	0.475	0.905	0.784	0.294	0.071

- The isotones do not give much information along the the $\lambda_2 = 0$ line towards the point $\boldsymbol{\lambda} = (+\infty, 0)$ (exponential distribution). The corresponding estimated powers, however, indicate that the highest powers are obtained with the data-driven SSP test based on the LL and the pAIC criteria, and the SSP24 test. Further, as in the diagonal direction, among the SSP2c tests the order is SSP24 $>$ SSP23 $>$ SSP22. The latter has shows the same pattern as the SSPdd-BIC test. Again the KW and the KS perform the worst.

Next, the $k = 3$ case is discussed.

- Along the diagonal in the $\boldsymbol{\lambda}$-plane, towards large positive λ-values, the isotones of the LL based data-driven SSP test and the SSP34 test are crossed first. This is confirmed by the estimated powers. When $n = 24$ the power of the SSPdd-pAIC test is always in between those of the SSPdd-LL and the SSP33 tests, but when $n = 48$ the SSP33 test is a bit more powerful than the pAIC based test. The SSP22 test, and this also the AD test, is always less powerful than the other SSP tests. The KS and the KW tests again have the lowest powers.
- In the other direction along the diagonal, again more or less the same behaviour is observed.
- When moving from the null point $\boldsymbol{\lambda} = (0.1349, 0.1349)$ towards the exponential

Table 6.8: The estimated powers for some alternatives of the mixture family in the 2-sample problem, based on 10000 simulation runs. The first distribution is a standard normal distribution ($\delta = \gamma = 0$).

n	δ	γ	power							
			SSP2	SSP3	SSP4	dd - BIC	dd - LL	dd - pAIC	KS	KW
24	0	0.6	0.199	0.473	0.594	0.208	0.594	0.590	0.230	0.067
24	0	0.7	0.328	0.676	0.774	0.360	0.774	0.775	0.311	0.067
24	0	0.85	0.604	0.919	0.953	0.690	0.953	0.953	0.508	0.097
24	0.5	0.6	0.302	0.470	0.563	0.306	0.565	0.565	0.322	0.182
24	0.5	0.8	0.601	0.841	0.914	0.626	0.914	0.912	0.593	0.270
24	1.5	0.5	0.495	0.555	0.578	0.495	0.581	0.578	0.480	0.451
24	2.5	0.6	0.798	0.840	0.838	0.798	0.840	0.851	0.750	0.745
48	0	0.5	0.322	0.618	0.738	0.322	0.738	0.720	0.364	0.072
48	0	0.6	0.553	0.848	0.923	0.553	0.923	0.922	0.563	0.078
48	0	0.8	0.960	0.997	0.998	0.960	0.998	0.998	0.938	0.080
48	0.5	0.5	0.443	0.650	0.790	0.443	0.790	0.760	0.545	0.208
48	1	0.4	0.463	0.550	0.610	0.463	0.610	0.570	0.508	0.372
48	1	0.6	0.853	0.917	0.938	0.853	0.938	0.933	0.907	0.697
48	3	0.4	0.823	0.853	0.882	0.823	0.885	0.865	0.705	0.710
48	3	0.6	0.993	0.997	0.998	0.993	0.998	0.997	0.990	0.977

point $\boldsymbol{\lambda} = (+\infty, 0)$, the powers of all tests remain small, but once more the SSP34 and the LL based data-driven tests performed best. When $n = 48$, the SSPdd-LL test resulted in the highest powers, followed by the SSP34 or the KS test. Then, the SSP33 and the pAIC follow, and finally, the lowest power is obtained with the KW test.

Contaminated Normal

The isotones of the mixture alternatives are presented in Figures 6.6 and 6.7. The estimated powers are shown in Tables 6.8 and 6.9.

First the 2-sample problem is discussed.

- When $\delta = 0$, i.e. a mixture of two equal mean normal distributions, the highest powers are generally obtained with the LL and the pAIC based data-driven tests and the SSP24 test. The next highest power resulted from the SSP23 test, followed by the SSPdd-BIC and the SSP22 tests. The latter two have almost indistinguishable powers when $n = 48$, when $n = 24$, the BIC criterion resulted in slightly higher powers. The power of the KS test is a bit lower than the power of the SSP22 test, except for very small values of γ. Finally, the power of the KW test broke completely down, which is due to the constancy of the mean under the mixture alternative with $\delta = 0$.

- In the situation where $\delta \neq 0$, then the means of the two samples differ, and thus it may be expected that the KW test performs at least better as compared to the $\delta = 0$ case. Indeed this is seen both from the estimated powers as from the isotones, but still it has the lowest powers among the tests compared in this study, except when δ is very large. The best tests are the SSPdd-pAIC, SSPdd-LL and the SSP24 test. The next best test is the SSP23 test, followed by the SSP22 test, which behaves

Table 6.9: The estimated powers for some alternatives of the mixture family in the 3-sample problem, based on 10000 simulation runs. The first two distributions are standard normal distributions ($\delta = \gamma = 0$).

n	δ	γ	power							
			SSP2	SSP3	SSP4	dd - BIC	dd - LL	dd - pAIC	KS	KW
24	0	0.8	0.223	0.492	0.656	0.223	0.656	0.616	0.122	0.073
24	1	0.8	0.643	0.742	0.800	0.643	0.800	0.742	0.637	0.523
24	1.5	0.7	0.694	0.716	0.737	0.694	0.742	0.725	0.616	0.615
24	1.5	0.8	0.837	0.862	0.885	0.837	0.886	0.861	0.799	0.758
24	2.5	0.7	0.810	0.832	0.843	0.810	0.847	0.840	0.655	0.727
24	3	0.4	0.335	0.345	0.363	0.335	0.373	0.353	0.182	0.280
24	3	0.8	0.935	0.943	0.952	0.935	0.955	0.945	0.862	0.887
48	0	0.6	0.298	0.566	0.746	0.298	0.746	0.594	0.329	0.061
48	0	0.7	0.431	0.775	0.909	0.431	0.908	0.800	0.490	0.069
48	0	0.9	0.901	0.995	0.999	0.901	0.999	0.999	0.908	0.085
48	1	0.6	0.676	0.746	0.826	0.676	0.824	0.720	0.745	0.564
48	1	0.8	0.945	0.974	0.986	0.945	0.986	0.966	0.972	0.824
48	2	0.4	0.584	0.615	0.661	0.584	0.668	0.601	0.460	0.474
48	2	0.5	0.789	0.814	0.859	0.789	0.852	0.799	0.726	0.683
48	3	0.6	0.789	0.814	0.859	0.789	0.853	0.799	0.726	0.683

almost exactly the same as the data-driven test based on the BIC criterion when $n = 48$. For the smaller sample size, sometimes the SSPdd-BIC outperforms the SSP22 slightly.
In general, when $\delta \neq 0$, the power differences are smaller then when $\delta = 0$.

Next, the 3-sample problem is discussed.

- When $\delta = 0$, the LL based data-driven SSP test and the SSP34 test resulted in the highest powers, followed by the pAIC based data-driven test and the SSP33 test. A further power decrease occurred between the SSP33 and the SSP32 test. The latter has again a power equal to the power of the SSPdd-BIC test. The KS test performed only slightly less than the SSP32 test. Extreme low powers were again obtained with the KW test.
- In the $\delta \neq 0$ case, the conclusions are more or less the same as in the two-sample case. The LL based data-driven test performs overall the best, immediately followed by the SSP34, SSP33 and the SSPdd-pAIC tests. The SSP32 and the SSPdd-BIC test are equivalent once more, and the lowest powers are obtained with the KS and the KW tests. But all powers are very close to each other, which is also observed from the isotones.

Location-Scale:
Figures 6.8 and 6.9 present the results of the scale-location family of alternatives. The estimated powers are shown in Tables 6.10 and 6.11.

As expected the powers when $k = 2$ are generally higher as compared to $k = 3$.
For most of the $\boldsymbol{\theta} = (\mu, \sigma)$ values the SSPdd-LL test is the most powerful, often almost equivalent to the SSPk4 test. The fact that the SSPdd-LL test mimics the SSP based test

Table 6.10: The estimated powers for some alternatives of the scale-location family in the 2-sample problem, based on 10000 simulation runs. The first distribution is a standard normal distribution ($\mu = 0, \sigma = 1$).

n	μ	σ	power							
			SSP2	SSP3	SSP4	dd - BIC	dd - LL	dd - pAIC	KS	KW
24	0	3	0.236	0.457	0.532	0.243	0.533	0.531	0.144	0.077
24	0	4	0.349	0.659	0.727	0.359	0.727	0.724	0.216	0.077
24	0	6	0.535	0.857	0.904	0.547	0.904	0.902	0.325	0.100
24	1	1	0.622	0.613	0.579	0.622	0.608	0.636	0.452	0.645
24	1	2	0.392	0.452	0.475	0.393	0.482	0.490	0.306	0.328
24	1	4	0.423	0.720	0.764	0.428	0.764	0.765	0.295	0.165
24	1	6	0.571	0.872	0.907	0.585	0.907	0.904	0.369	0.117
24	2	1	0.996	0.997	0.992	0.996	0.994	0.996	0.975	0.995
24	2	4	0.629	0.806	0.834	0.639	0.835	0.834	0.548	0.361
48	0	2	0.325	0.492	0.546	0.325	0.550	0.500	0.152	0.069
48	0	3	0.722	0.891	0.925	0.722	0.925	0.892	0.348	0.072
48	0	4	0.918	0.981	0.987	0.918	0.987	0.982	0.541	0.068
48	0.75	1	0.707	0.673	0.648	0.707	0.684	0.709	0.560	0.727
48	0.75	2	0.597	0.712	0.746	0.597	0.748	0.718	0.475	0.349
48	1	1	0.911	0.897	0.880	0.911	0.903	0.911	0.789	0.919
48	1	2	0.757	0.806	0.823	0.757	0.828	0.810	0.649	0.569
48	1	3	0.853	0.948	0.964	0.853	0.964	0.950	0.647	0.345
48	1.5	3	0.941	0.975	0.982	0.941	0.983	0.975	0.852	0.592

with $c = 4$ has been encountered a few times before, even for the one sample problem. Only under the circumstances of a location shift with a mild change in scale (e.g. $(\mu, \sigma) = (1, 1)$) the SSPk2 test shows the highest power among the SSPkc tests. This behaviour is seen in both the estimated powers and in the isotones. Under these conditions it is the BIC or the pAIC criterion that outperforms the LL penalty. Yet another feature under these conditions that appears from the isotones and that is confirmed by the estimated powers, is that only here the KW test is more powerful than the KS test, which is not surprising for the KW test is indeed specifically sensitive to location shifts when the forms of the k distributions are equal. For all other values of $\boldsymbol{\theta}$ both the KW and the KS test are far less powerful than the SSP based tests.

The pAIC penalty did on average only a bit less than the LL penalty, and as mentioned in the previous paragraph, it did better under a location shift alternative.

In general it may be concluded from this small simulation study that the LL based data-driven SSP test is a good choice. When one expects a location shift but when one is still interested in the general k-sample alternative as well, the pAIC penalty is also a good choice.

6.3.4 Conclusion

Three families of alternatives were empirically investigated, which suggest some general conclusions.

First, we would like to remark that although under the three families very frequently the 2-

Table 6.11: The estimated powers for some alternatives of the scale-location family in the 3-sample problem, based on 10000 simulation runs. The first two distributions are standard normal distributions ($\mu = 0, \sigma = 1$).

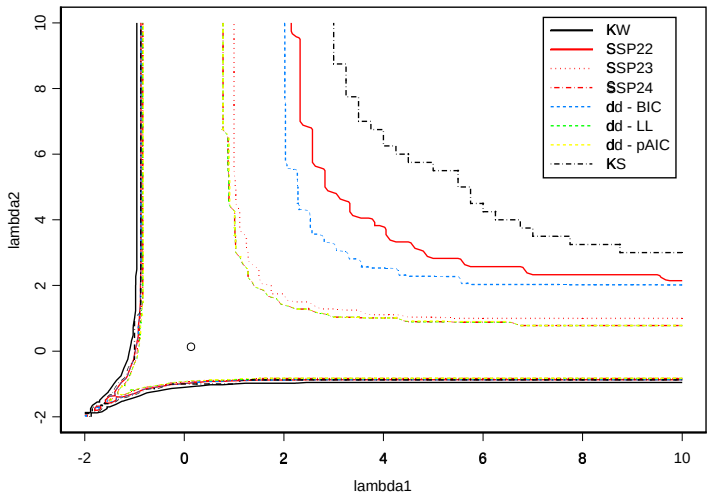
n	μ	σ	power							
			SSP2	SSP3	SSP4	dd - BIC	dd - LL	dd - pAIC	KS	KW
24	0	8	0.346	0.680	0.780	0.346	0.780	0.694	0.156	0.072
24	0	12	0.429	0.800	0.883	0.429	0.883	0.813	0.192	0.080
24	2	1	0.968	0.953	0.944	0.968	0.954	0.970	0.871	0.969
24	2	10	0.458	0.807	0.878	0.458	0.878	0.811	0.246	0.098
24	4	5	0.734	0.827	0.856	0.734	0.858	0.829	0.592	0.517
24	4	10	0.573	0.837	0.908	0.573	0.908	0.842	0.346	0.189
24	4	25	0.555	0.934	0.974	0.555	0.974	0.940	0.281	0.122
48	0	3	0.448	0.668	0.770	0.448	0.768	0.623	0.234	0.059
48	0	5	0.813	0.947	0.975	0.813	0.975	0.929	0.482	0.083
48	0	10	0.991	0.997	0.999	0.991	0.999	0.997	0.785	0.088
48	1	1	0.783	0.746	0.753	0.783	0.777	0.783	0.654	0.793
48	1	3	0.632	0.774	0.844	0.632	0.837	0.735	0.469	0.251
48	1	5	0.868	0.956	0.975	0.868	0.975	0.941	0.580	0.153
48	2	5	0.915	0.977	0.990	0.915	0.990	0.967	0.763	0.343

sized SSP based test performed worst among the SSP tests and the 4-sized test performed best, there does not seem to a theoretically supported argument that allows us to conclude that the SSPk2 test will always results in the lowest powers. This remark makes us a bit precautious w.r.t. to a recommendation of a certain penalty. In particular, most of the simulation results suggest that the LL criterion gives the best powers, and the same simulations also show that often the SSPdd-LL test is very similar to the SSPk4 test. The latter characteristic was also seen in the simulation study of Chapter 5. Moreover, it is also observed that the LL criterion often still selects $C_n = 4$ when a smaller c would be a better choice. Under such circumstances the pAIC criterion, which shows generally a faster convergence of C_n than LL, frequently selects a smaller partition size resulting in a higher power. On the other hand, when the SSPk4 test is the most powerful, the pAIC penalty leads to powers that are only a little bit smaller than those with the LL criterion.

The above discussion brings us to the conclusion that both the LL and the pAIC criteria seem good choices.

Generally, the SSP-based tests result in greater powers than the KS and the KW tests. Only under a location shift alternative the power of the KW test is slightly greater.

(a)



(b)

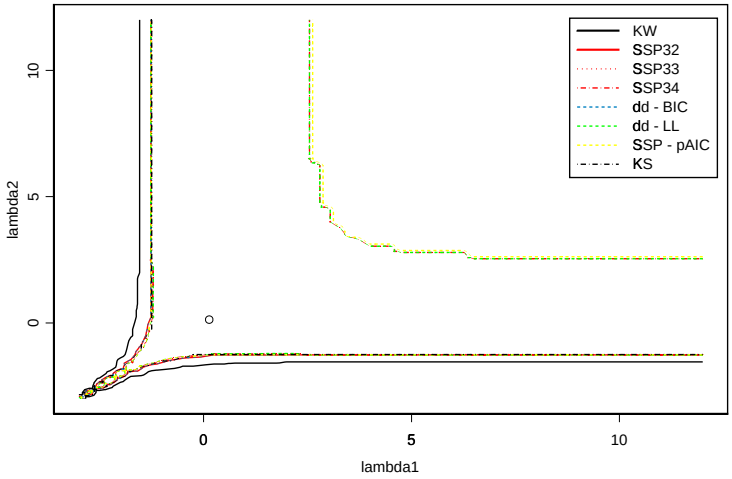
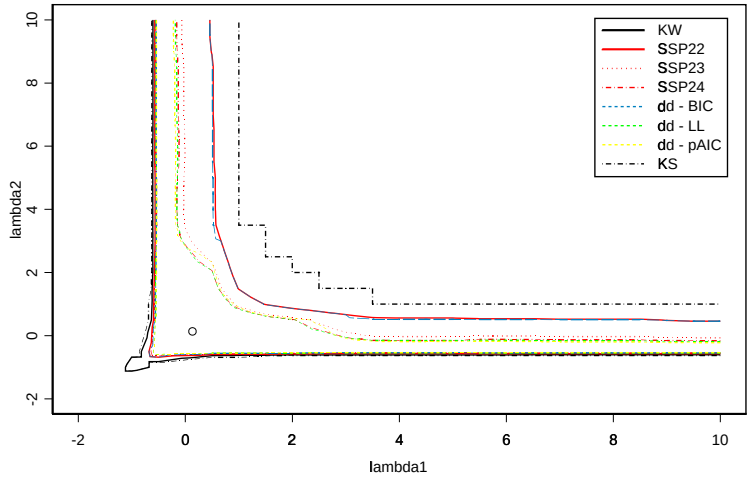


Figure 6.4: The isotones for the Tukey-Lambda family of alternatives, for $k = 2$ (a) and $k = 3$ (b), both with $n = 24$ observations. The first $k - 1$ distributions are normal distributions $\boldsymbol{\lambda} = (0.1349, 0.1349)$.

(a)



(b)

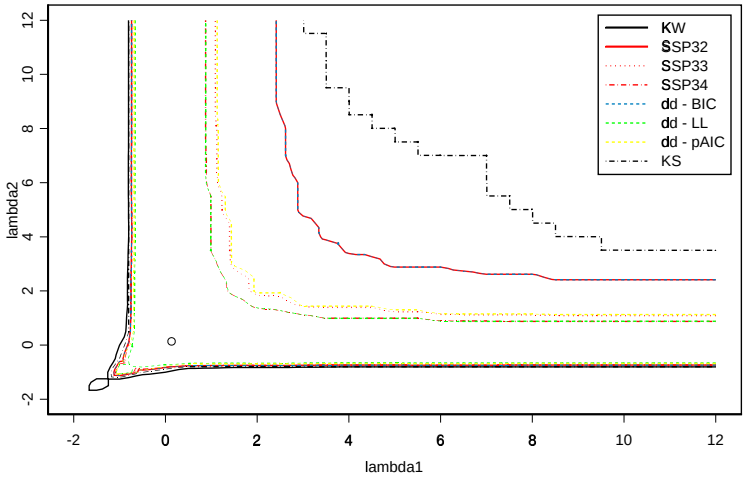
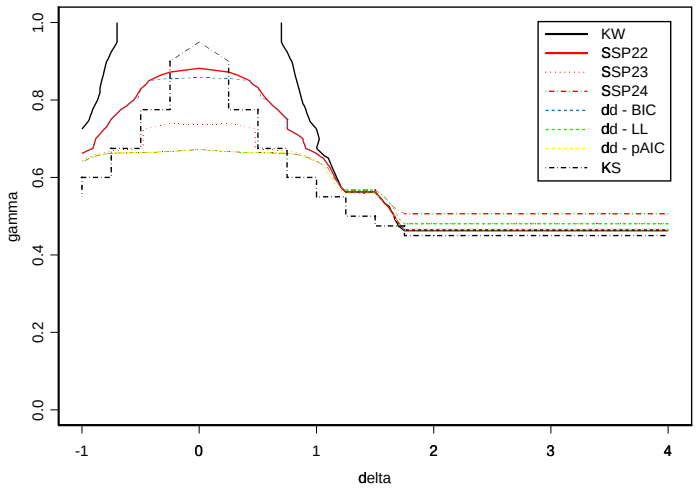


Figure 6.5: The isotones for the Tukey-Lambda family of alternatives, for $k = 2$ (a) and $k = 3$ (b), both with $n = 48$ observations. The first $k - 1$ distributions are normal distributions $\boldsymbol{\lambda} = (0.1349, 0.1349)$.

(a)



(b)

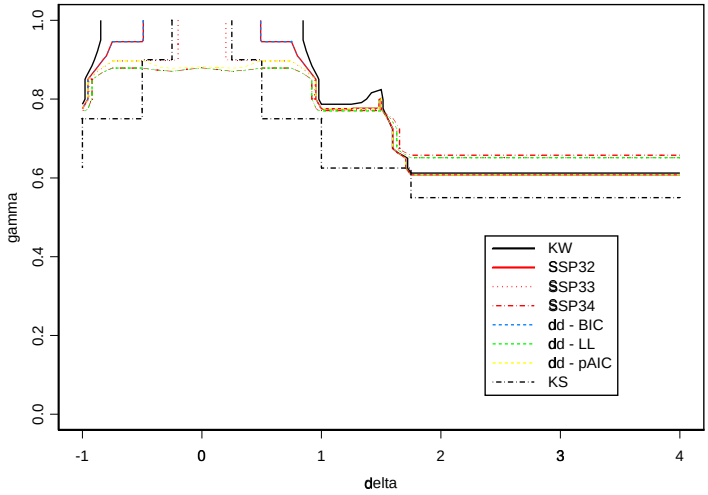
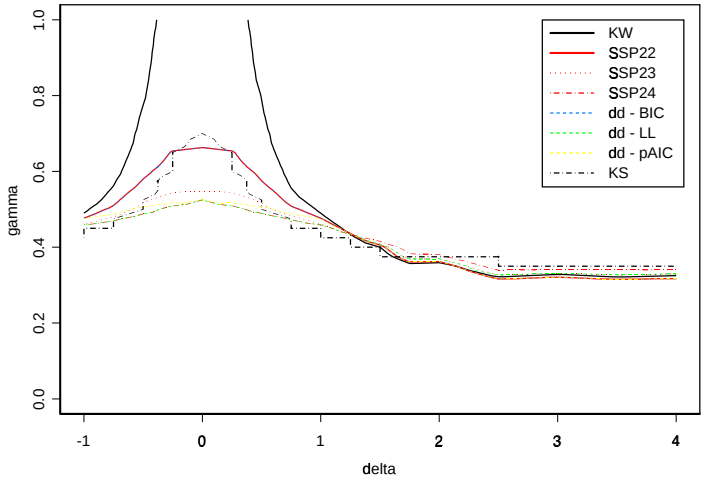


Figure 6.6: The isotones for the mixture family of alternatives, for $k = 2$ (a) and $k = 3$ (b), both with $n = 24$ observations. The first $k - 1$ distributions are standard normal distributions ($\delta = \gamma = 0$).

(a)



(b)

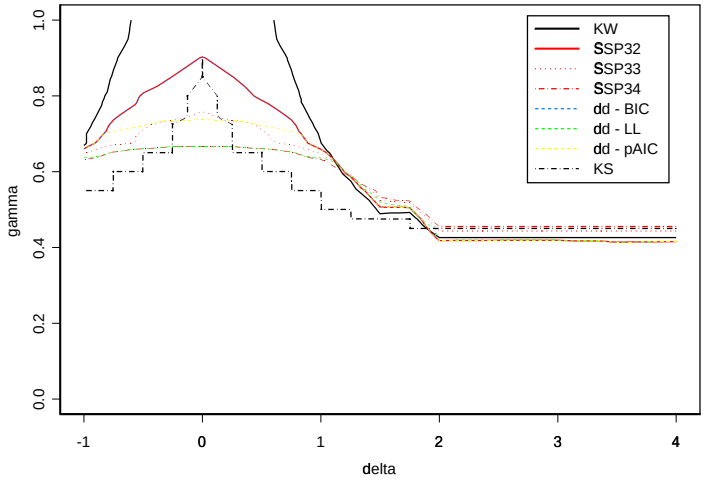
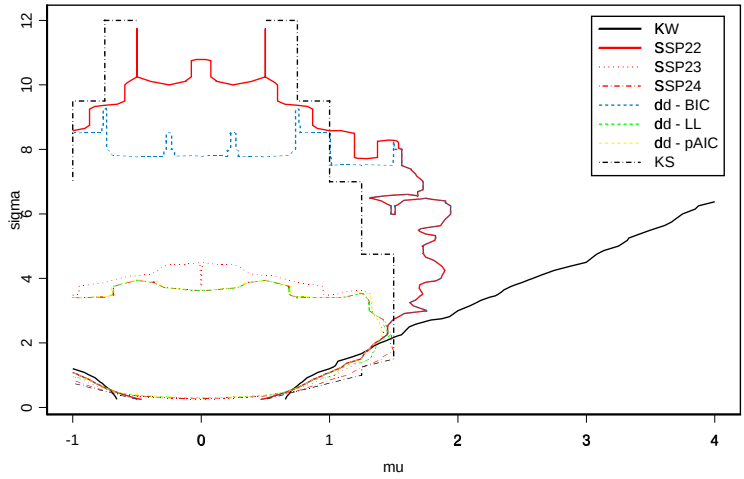


Figure 6.7: The isotones for the mixture family of alternatives, for $k = 2$ (a) and $k = 3$ (b), both with $n = 48$ observations. The first $k - 1$ distributions are standard normal distributions ($\delta = \gamma = 0$).

(a)



(b)

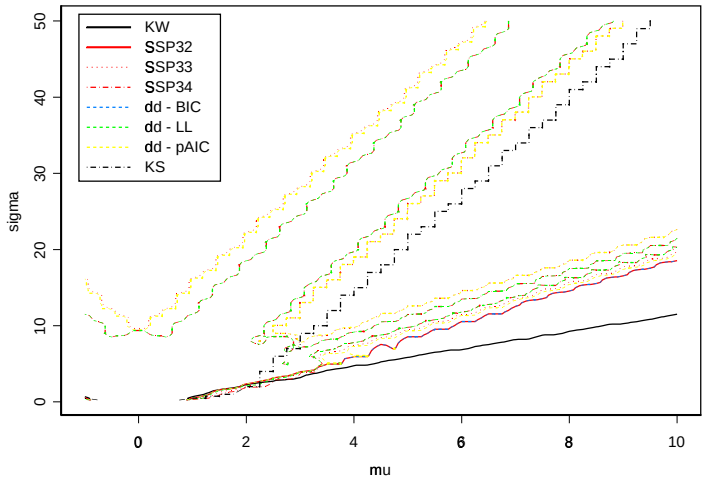
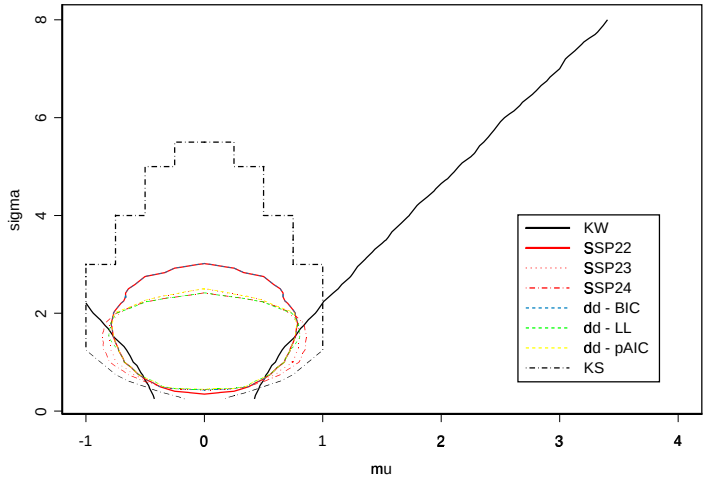


Figure 6.8: The isotones for the scale-location family of alternatives, for $k = 2$ (a) and $k = 3$ (b), both with $n = 24$ observations. The first $k - 1$ distributions are standard normal distributions ($\mu = 0, \sigma = 1$).

(a)



(b)

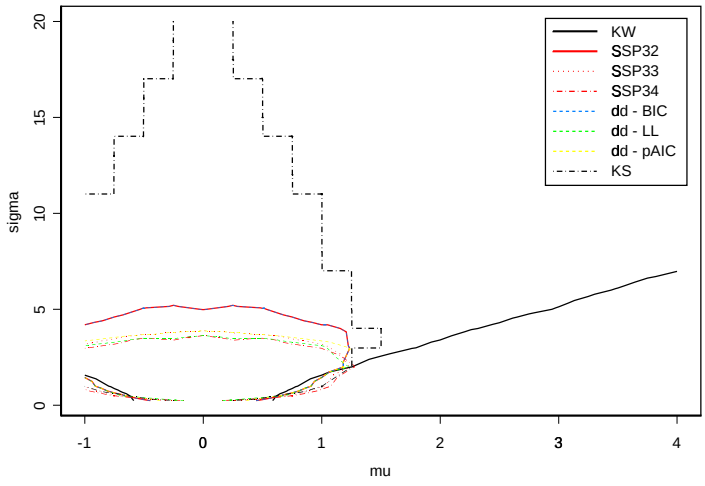


Figure 6.9: The isotones for the scale-location family of alternatives, for $k = 2$ (a) and $k = 3$ (b), both with $n = 48$ observations. The first $k - 1$ distributions are standard normal distributions ($\mu = 0, \sigma = 1$).

6.4 The Decomposition of the SSPkc Test Statistic

The practical advantages of non-parametric omnibus tests are obvious: first, its null distribution does not depend on the true distributions, and second, because of the omnibus sensitivity, the test may be applied to detect whatever deviation from the null hypothesis. Thus, by the latter characteristic, the researcher must not have any prior knowledge on possible deviations. But, on the other hand, the same property has led to a criticism that generally applies to many omnibus tests: when such a test rejects the null hypothesis, the researcher actually only knows that the k samples do not come from the same distribution, but the test does not give an indication in what sense the distributions differ. E.g. all distributions may be of the same form, but may differ in mean or in variance.

For the one-sample GOF problem, Neyman (1937) constructed a smooth test by embedding the hypothesized distribution G in a parametric (exponential) family of continuous alternative distributions F_θ, indexed by an s-dimensional nuisance parameter $\boldsymbol{\theta}$. The test statistic, which is a score statistic within the exponential family, is a sum of s independent components, and each component relates to just one of the s parameters. In this way, when the GOF null hypothesis is rejected by Neyman's test, large components indicate that the corresponding parameters are probably non-zero in the true distribution. This Neyman's smooth test is however only consistent if the true distribution is embedded in the family of alternatives. Furthermore, if this is indeed the case, Neyman (1937) showed that the test is locally uniformly most powerful. Despite this optimality property, the above mentioned embedding-requirement is generally only guaranteed as $s \to \infty$, but of course this is not feasible in practice. A rather recent solution is to make the smooth test data-driven (Ledwina, 1994), but still the dimension $s = s_n$ must grow with n in order to have consistency. For the 2-sample problem, the data-driven version is recently developed by Janic-Wróblewska and Ledwina (2000).

A generalization of the construction of test statistics as sums of components is given by Eubank et al. (1987). They have shown that Pearson's ϕ^2-measure, which is twice the directed (continuous) divergence of order 1 (Equation 5.3),

$$\phi^2 = 2\mathrm{I}^1\left(F;G\right) = \int_{\mathcal{S}} \frac{\left(f(\boldsymbol{x}) - g(\boldsymbol{x})\right)^2}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x},$$

can be decomposed into an infinite number of components a_i^2 ($i = 1, \ldots$), which can easily be estimated from the data. The test statistic is then the corresponding sum of estimated components $\hat{a}_i^2$ (details are given in Section 4.3.2). Of course, only a limited number of components, say s, are computed. The interpretation of the components depends on the choice of a complete orthonormal system (CONS) of polynomials. When, e.g., the system of Legendre polynomials is taken, the same components as those of the Neyman's smooth test statistic are found. With this choice, the first two components for the k-sample problem are the Kruskal-Wallis statistic and the Mood statistic. The former is a rank statistic which is often used to test against a shift in location, and the latter is a statistic used for detecting a change in scale. With this in mind, whenever by means of

such a smooth test, whether data-driven or not, the k-sample null hypothesis is rejected, a close examination of the separate components may reveal some of the features with respect to which the distributions differ. The above mentioned components are those of Janic-Wróblewska and Ledwina (2000) data-driven test statistic as well.

Most of the above mentioned properties hold more generally for other GOF problems, but in the remainder of this section the discussion is restricted to the k-sample problem. Sometimes reference will be made to the data-driven test of Janic-Wróblewska and Ledwina (2000), which is however only constructed for the 2-sample problem.

The aim of this section is to decompose the SSPkc test statistic into interpretable components. Furthermore, it will be shown that the above mentioned statistics may be obtained in a limiting case. Again, as before, we will use the terms discrete and continuous divergence to refer to the divergence based on a SSP and the supremum of such a divergence, respectively.

6.4.1 The Decomposition of the Averaged Pearson Discrete Divergence

The Pearson discrete divergence is the directed divergence of order 1, based on a SSP.

The partition construction rule and the partition determining subsample set are exactly the same as those defined previously for the k-sample problem. Using the same notation as in Section 6.1.3, the averaged Pearson divergence is given by (Equation 6.4)

$$\begin{aligned}\Delta_c(F;G) &= \mathrm{E}_f\left[\mathrm{I}^1\left(F;G\|[A](\mathcal{P}_{c-1,n})\right)\right] \\ &= \frac{1}{2}\mathrm{E}_f\left[\sum_{j=1}^{k}\sum_{i=1}^{c}\frac{\left(\frac{n_j}{n}\mathrm{P}_{f_j}\left[B_{ij}\right]-\frac{n_j}{n}\mathrm{P}_g\left[A_i\right]\right)^2}{\frac{n_j}{n}\mathrm{P}_g\left[A_i\right]}\right], \qquad (6.13)\end{aligned}$$

where for notational comfort the dependence on $\mathcal{P}_{c-1,n}$ of the elements of the partitions $[A](\mathcal{P}_{c-1,n})$ and $[B](\mathcal{P}_{c-1,n})$ is dropped.

Let $\boldsymbol{Q} = \boldsymbol{Q}(\mathcal{P}_{c-1,n})$ be a $c \times k$ matrix with (i,j)th entry q_{ij} equal to

$$n^{1/2}\frac{\frac{n_j}{n}\mathrm{P}_{f_j}\left[B_{ij}\right]-\frac{n_j}{n}\mathrm{P}_g\left[A_i\right]}{\sqrt{\frac{n_j}{n}\mathrm{P}_g\left[A_i\right]}},$$

($i = 1,\ldots,c; j = 1,\ldots,k$). The Pearson divergence $\mathrm{I}^1\left(F;G\|[A](\mathcal{P}_{c-1,n})\right)$ is then equal to $\frac{1}{n}$ times the sum of squared entries of $\boldsymbol{Q}$, which may be written as $\mathrm{tr}\left[\boldsymbol{Q}\boldsymbol{Q}'\right]$. Thus, $\Delta_c(F;G) = \frac{1}{2n}\mathrm{E}_f\left[\mathrm{tr}\left[\boldsymbol{Q}\boldsymbol{Q}'\right]\right]$.

Next, we consider a decomposition of the Pearson divergence, similar to decompositions proposed by Lancaster (1949, 1953). Let $\{v_i\}$ ($i = 1,\ldots,c$) be scores assigned to the c rows of the contingency table induced by the partition $[B](\mathcal{P}_{c-1,n})$. Let $\boldsymbol{R} = \boldsymbol{R}(\mathcal{P}_{c-1,n})$

be a $c \times c$ orthonormal matrix with on the first row the entries

$$v_j = \sqrt{\mathrm{P}_g\left[A_j\right]},$$

($j = 1, \ldots, c$). The other rows are constructed by orthogonality conditions. Let $\{h_{i-1}(v)\}$ ($i = 1, \ldots, c$) be a set of orthonormal polynomials in $v \in \{v_1, \ldots, v_c\}$, associated with the probabilities $\mathrm{P}_g\left[A_j\right]$, and such that $h_0(v_j) = 1$ ($j = 1, \ldots, c$). Then, the entry r_{ij} may be given by ($i, j = 1, \ldots, c$)

$$r_{ij} = h_{i-1}(v_j)\sqrt{\mathrm{P}_g\left[A_j\right]}.$$

Indeed, $\boldsymbol{R}$ constructed in this way is an orthonormal matrix,

$$\sum_{j=1}^{c} r_{lj} r_{mj} = \sum_{j=1}^{c} h_{l-1}(v_j) h_{m-1}(v_j) \mathrm{P}_g\left[A_j\right] = \delta_{lm},$$

($l, m = 1, \ldots, c$) where δ_{lm} is Kronecker's delta.

Note that an important property of the orthonormal matrix $\boldsymbol{R}$ is that

$$\mathrm{tr}\left[(\boldsymbol{RQ})(\boldsymbol{RQ})'\right] = \mathrm{tr}\left[\boldsymbol{R'RQQ'}\right] = \mathrm{tr}\left[\boldsymbol{QQ'}\right],$$

or, equivalently, the sum of squared entries of $\boldsymbol{RQ}$ is equal to the sum of squared entries of $\boldsymbol{Q}$.

Here, we take the same polynomials as those used by Best (1995). The first two are given by ($j = 1, \ldots, c$)

$$\begin{aligned} h_0(v_j) &= 1 \\ h_1(v_j) &= \frac{v_j - \mu}{\sqrt{\nu - \mu^2}}, \end{aligned}$$

where $\mu = \mu(\mathcal{P}_{c-1,n}) = \sum_{j=1}^{c} v_j \mathrm{P}_g\left[A_j\right]$ and $\nu = \nu(\mathcal{P}_{c-1,n}) = \sum_{j=1}^{c} v_j^2 \mathrm{P}_g\left[A_j\right]$. Let $\mu_3 = \mu_3(\mathcal{P}_{c-1,n}) = \sum_{j=1}^{c} v_j^3 \mathrm{P}_g\left[A_j\right]$ and $\mu_4 = \mu_4(\mathcal{P}_{c-1,n}) = \sum_{j=1}^{c} v_j^4 \mathrm{P}_g\left[A_j\right]$. The third polynomial is given by

$$h_2(v_j) = b_4^{-1/2}\left(v_j^2 - b_1^2 b_2 v_j + b_3\right),$$

where $b_1 = (\nu - \mu^2)^{-1/2}$, $b_2 = \mu_3 - \mu\nu$, $b_3 = b_1^2 b_2 \mu - \nu$ and $b_4 = \mu_4 + b_1^4 b_2^2 \nu + b_3^2 - 2b_1^2 b_2 \mu_3 + 2b_3\nu - 2b_1^2 b_2 b_3 \mu$.

Further, we consider for every $\mathcal{P}_{c-1,n} \in \mathbb{P}_{c-1,n}$ the scores $v_j = j$ ($j = 1, \ldots, c$).

$\mathrm{tr}\left[\boldsymbol{QQ'}\right]$ is now decomposed into $c-1$ components. Let $\boldsymbol{r}_i$ be the ith row of $\boldsymbol{R}$. Then, the components are given by the sum of squares of $\boldsymbol{r}_i'\boldsymbol{Q}$ ($i = 1, \ldots, c$). It is easily checked that all entries of $\boldsymbol{r}_1'\boldsymbol{Q}$ are exactly zero, therefore only $c-1$ possibly non-zero components

remain. The corresponding decomposition of $\Delta_c(F;G)$ is

$$\Delta_c(F;G) = \Delta_c^{(1)}(F;G) + \ldots + \Delta_c^{(c-1)}(F;G).$$

The entries of $\boldsymbol{r}_2'\boldsymbol{Q}$ ($j = 1, \ldots, k$) are

$$n^{1/2} \sum_{i=1}^{c} \frac{i-\mu}{\sqrt{\frac{n_j}{n}(\nu-\mu^2)}} \mathrm{P}_{f_{xy}}\left[B_{ij}\right].$$

The first component of the averaged Pearson divergence is thus

$$\Delta_c^{(1)}(F;G) = \frac{1}{2}\mathrm{E}_f\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{c} \frac{i-\mu}{\sqrt{\frac{n_j}{n}(\nu-\mu^2)}} \mathrm{P}_{f_{xy}}\left[B_{ij}\right]\right)^2\right]. \tag{6.14}$$

It may be interesting to note that for this first component, still a property similar to Lemma 5.1 applies.

Lemma 6.1

$$\Delta_c^{(1)}(F;G) = 0 \Leftrightarrow H_0 \textit{ is true}$$

Proof. If H_0 is true, then according to Lemma 6.5, $\Delta_c(F;G) = 0$. Since $\Delta_c(F;G) \geq \Delta_c^{(1)}(F;G)$, also $\Delta_c^{(1)}(F;G) = 0$.

When $\Delta_c^{(1)}(F;G) = 0$, then for all $l = 1, \ldots, k$,

$$\sum_{i=1}^{c} i\mathrm{P}_{f_{xy}}\left[B_{il}\right] = \sum_{i=1}^{c} \mu\mathrm{P}_{f_{xy}}\left[B_{il}\right],$$

almost everywhere. The substitution of $\mu = \sum_{i=1}^{c} i\mathrm{P}_g\left[A_i\right]$ and $\mathrm{P}_{f_{xy}}\left[B_{il}\right] = \frac{n_l}{n}\mathrm{P}_{f_l}\left[B_{il}\right]$ gives for all $l = 1, \ldots, k$,

$$\sum_{i=1}^{c} i\mathrm{P}_{f_l}\left[B_{il}\right] = \sum_{i=1}^{c} i\mathrm{P}_g\left[A_i\right],$$

almost everywhere. Hence, for all $l = 1, \ldots, k$, and for all $i = 1, \ldots, c$,

$$\mathrm{P}_{f_l}\left[B_{il}\right] = \mathrm{P}_g\left[A_i\right],$$

almost everywhere. The remainder of the proof is similar to the proof of Lemma 5.1. □

6.4.2 The Decomposition of the Test Statistic

The test statistic $T_{k,c,n}$ was obtained as the plug-in estimator of $2n\Delta_c(F;G)$. In a similar way the components $T^{(i)}_{k,c,n}$ of the test statistic are the plug-in estimators of the components $2n\Delta^{(i)}_c(F;G)$, resulting in

$$T_{k,c,n} = \sum_{i=1}^{c-1} T^{(i)}_{k,c,n}.$$

The first component is given by (cfr. Equation 6.14)

$$\begin{aligned} T^{(1)}_{k,c,n} &= \frac{1}{a_{c-1,n}} \sum_{\mathcal{P}\in\mathbb{P}_{c-1,n}} \sum_{j=1}^{k} \left(\sqrt{n} \sum_{i=1}^{c} \frac{i-\hat{\mu}(\mathcal{P})}{\sqrt{\frac{n_j}{n}(\hat{\nu}(\mathcal{P})-\hat{\mu}^2(\mathcal{P}))}} \hat{\mathrm{P}}_{f_{xy}}\left[B_{ij}(\mathcal{P})\right] \right)^2 \\ &= \frac{1}{a_{c-1,n}} \sum_{\mathcal{P}\in\mathbb{P}_{c-1,n}} \frac{\sum_{j=1}^{k} n_j \left(\sum_{i=1}^{c} i \hat{\mathrm{P}}_{f_j}\left[B_{ij}(\mathcal{P})\right] - \hat{\mu}(\mathcal{P}) \right)^2}{\hat{\nu}(\mathcal{P})-\hat{\mu}^2(\mathcal{P})}, \end{aligned} \tag{6.15}$$

where $a_{k,c,n} = \#\mathbb{P}_{c-1,n}$, $\hat{\mu}(\mathcal{P}) = \sum_{i=1}^{c} i\hat{\mathrm{P}}_g\left[A_i(\mathcal{P})\right]$, and $\hat{\nu}(\mathcal{P}) = \sum_{i=1}^{c} i^2\hat{\mathrm{P}}_g\left[A_i(\mathcal{P})\right]$.

In the same way as the first component is determined, based on the polynomial $h_2(.)$ the second component is now calculated (for notational comfort the dependence on $\mathcal{P}$ is suppressed),

$$T^{(2)}_{k,c,n} = \frac{1}{a_{c-1,n}} \sum_{\mathcal{P}\in\mathbb{P}_{c-1,n}} \sum_{j=1}^{k} \left(\sqrt{n} \sum_{i=1}^{c} \frac{\left(i^2 - \hat{\nu} - \frac{\hat{\mu}_3-\hat{\nu}\hat{\mu}}{\hat{\nu}-\hat{\mu}^2}(i-\hat{\mu}) \right) \hat{\mathrm{P}}_{f_{xy}}\left[B_{ij}\right]}{\sqrt{\frac{n_j}{n}\left(\hat{\mu}_4 - \hat{\nu}^2 - \frac{(\hat{\mu}_3-\hat{\nu}\hat{\mu})^2}{\hat{\nu}-\hat{\mu}^2} \right)}} \right)^2,$$

where $\hat{\mu}_3 = \hat{\mu}_3(\mathcal{P}) = \sum_{i=1}^{c} i^3\hat{\mathrm{P}}_g\left[A_i(\mathcal{P})\right]$ and $\hat{\mu}_4 = \hat{\mu}_4(\mathcal{P}) = \sum_{i=1}^{c} i^4\hat{\mathrm{P}}_g\left[A_i(\mathcal{P})\right]$.

Both statistics are of the form

$$\frac{1}{n^{c-1}} \sum_{\mathcal{P}\in\mathbb{P}_{c-1,n}} \sum_{j=1}^{k} \left(\sqrt{n} \frac{X_j(\mathcal{P}) - \mathrm{E}\left[X_j(\mathcal{P})|\mathcal{P}\right]}{\sqrt{\mathrm{Var}\left[X_j(\mathcal{P})|\mathcal{P}\right]}} \right)^2.$$

Before we go into the interpretation of the components, a limiting case is discussed. This will help us to understand the components better.

6.4.3 The Decomposition in a Limiting Case

The Decomposition

Consider the limiting case where $\mathbb{P}_{c-1,n} = \mathbb{P}_{n-1,n} = \{\mathcal{S}_n \backslash \{x_{(n)}\}\}$, such that $\#\mathbb{P}_{n-1,n} = 1$, and a partition constructing rule such that the only partition that is formed is equal to

$$[A] = \{]b_l, x_{(1)}],]x_{(1)}, x_{(2)}], \ldots]x_{(n-1)}, b_u]\},$$

where $x_{(1)}, \ldots, x_{(n)}$ are the order statistics (see also Figure 6.10).
Under these conditions we will show that the resulting $T_{k,n,n}$ and its components are well known statistics.

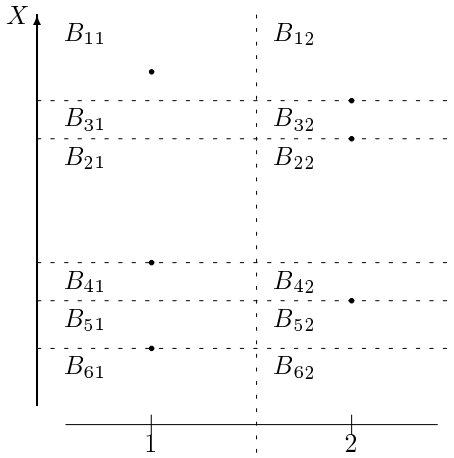


Figure 6.10: The figure illustrates the limiting partition constructed from $\mathcal{P} = \{X_{(1)}, \ldots, X_{(n-1)}\}$ (here $k = 2$, $n_1 = n_2 = 3$).

Lemma 6.2 *If* $\mathcal{P} = \{X_{(1)}, \ldots, X_{(n)}\}$, *then*

$$2I^1\left(F; G \| [A]_{\mathcal{P}}\right) \xrightarrow{p} \phi^2(F; G).$$

Proof. Without loss of generality we suppose that $X \sim U[0,1]$. Consider a triangular array $X_{n1}, \ldots X_{nn}$, and let $[A]_n$ denote the corresponding n-sized SSP with elements A_{ni} $(i = 1, \ldots, n)$. The size of a partition element is given by its norm $||A_{ni}|| = X_{n(i-1)} - X_{n(i)}$,where $X_{n(i)}$ is the ith order statistic of the $X_{n1}, \ldots, X_{nn}$ ($X_{n(0)} \equiv 0$ and $X_{n(n+1)} \equiv 1$). Thus, $||A_{ni}||$ is the ith uniform spacing D_{ni}. According to the definition of $\phi^2(F; G) = 2\mathrm{I}^1\left(F; G\right)$ (Definition 5.2) and the refinement lemma (Corollary 5.1) the proposed convergence would hold if the maximal partition element

size converges in probability to zero.

$$\begin{aligned} \max_{1\leq k\leq n+1} ||A_{nk}|| &= \max_{1\leq k\leq n+1} D_{nk} \\ &\xrightarrow{p} 0, \end{aligned}$$

where the last step is found e.g. in Shorack (2000, p.330). □

The following Corollary now follows immediately.

Corollary 6.5 *If* $\mathcal{P} = \mathcal{S}_n$*, then, as* $n \to \infty$*,*

$$2\Delta_n(F;G) \to \phi^2(F;G).$$

With the choice of $v_j = j$, it is fairly easy to show that the polynomials $h_i(\frac{j}{n})$ $(i = 1, \dots)$ converge in probability to the corresponding Legendre polynomials as $n \to \infty$. Hence, the sum of the squared elements of $n^{-1/2}\boldsymbol{r}_i'\boldsymbol{Q}$ $(i = 2, \dots)$ converge in probability to corresponding a_i^2-terms in the decomposition of $\phi^2(F;G)$ (Eubank et al., 1987, p.822). Thus, the decomposition of $2\Delta_n(F;G)$ is asymptotically equivalent to the decomposition of $\phi^2(F;G)$.
The a_i^2 in the decomposition of $\phi^2(F;G)$ are typically estimated by their plug-in estimators $\hat{a}_i^2$. Also the averaged Pearson divergence for the k-sample problem is estimated by its plug-in estimator in order to construct the test statistic $T_{k,c,n}$. Even for finite n it is seen almost directly that the estimated components $n\hat{a}_i^2$ are exactly equal to the corresponding $T_{k,n,n}^{(i)}$. Indeed, since the partition bounds are the order statistics, the probabilities $\mathrm{P}_g[A_j]$ and $\mathrm{P}_{f_i}[B_{ji}]$ are simply estimated by $\hat{\mathrm{P}}_g[A_j] = \frac{1}{n}$ and $\hat{\mathrm{P}}_{f_i}[B_{ji}] = \frac{1}{n_i}$, respectively, the equivalence follows immediately. Thus, the first term $T_{k,n,n}^{(1)}$ is the Kruskal-Wallis test statistic (Kruskal, 1952; Kruskal and Wallis, 1952) for the k-sample location shift problem,

$$T_{k,n,n}^{(1)} = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{n+1}{2} \right)^2, \tag{6.16}$$

where $\bar{R}_i$ is the average rank of the n_i observations of the ith sample among the n pooled sample observations. There is another way of looking at the Kruskal-Wallis test statistic that might be of interest later in this text. Let S_{ij} be scores assigned to the observations. These scores are defined as $S_{ij} = m$ iff $X_{ij} \in A_m$ $(j = 1, \dots, k; i = 1, \dots, n_j; m = 1, \dots, c$. Since here $c = n$, $S_{ij} = R(X_{ij})$, i.e. the rank of observation X_{ij} among the n observations in the pooled sample. The Kruskal-Wallis statistic can then be written as

$$\frac{\sum_{j=1}^{k} n_j \left(\bar{S}_j - \bar{S}\right)^2}{\hat{\sigma}^2}, \tag{6.17}$$

where $\bar{S}_j$ is the average of the scores in the jth sample, and $\hat{\sigma}^2$ is the sample variance of the scores in the pooled sample. Substituting the scores with the ranks, and thus $\bar{S} = \bar{R} = \frac{n+1}{2}$ and $\hat{\sigma}^2 = \frac{n(n+1)}{12}$, results immediately in the well known form of the statistic (Equation 6.16). In classical Analysis-of-Variance (ANOVA) terms, the denominator of Equation 6.17 is the between-group sum of squares SST, and the nominator is the total sum of squares SSTot divided by n. Thus, with this notation, the statistic may be written as

$$n\frac{\text{SST}}{\text{SSTot}},$$

which is of the form of the F-statistic in ANOVA for testing the hypothesis of equality of means.

The second term becomes the generalized Mood's statistic for the k-sample scale problem,

$$T^{(2)}_{k,n,n} = \frac{180}{(n-1)(n+1)(n-2)(n+2)} \sum_{j=1}^{k} n_j \left(\bar{M}_j - \frac{1}{12}(n-1)(n+1) \right)^2,$$

where

$$\bar{M}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \left(R_i - \frac{n+1}{2} \right)^2.$$

This discussion results in the following Proposition.

Proposition 6.3 *Under H_0, $T_{k,n,n}$ and $n \sum_{i=1}^{n} \hat{a}_i^2$ have the same limiting distribution as $n \to \infty$.*

A Comment on a Data-Driven Version of $T_{k,n,n}$

Let $T_{k,n,n;s}$ be the decomposition of $T_{k,n,n}$, truncated after the s-th term, i.e.

$$T_{k,n,n;s} = \sum_{i=1}^{s} T^{(i)}_{k,n,n}.$$

Then, in a little different way as in Section 6.2, but rather in the sense of the construction of the data-driven Neyman's smooth tests (Janic-Wróblewska and Ledwina, 2000), the order s can be estimated from the data as

$$\begin{aligned} S_n \quad = \quad & \min \left[s : 1 \le s \le d(n_1), T_{k,n,n;s} - 2s(k-1)\ln a_n \ge T_{k,n,n;r} - 2r(k-1)\ln a_n, \right. \\ & \left. 1 \le r \le d(n_1) \right], \end{aligned}$$

where $d(n_1) = o\left((n_1/\ln n_1)^{1/9}\right)$, and a_n is as before. The data-driven test statistic is

$$T_{k,n,n;S_n}.$$

When $k = 2$, this test statistic is exactly the statistic of Janic-Wróblewska and Ledwina (2000). When $k > 2$, their Theorems 1 and 2 may possibly be extended. We state the generalization here only as a conjecture. It remains to be proven for a suitable sequence of maximal orders $d(n_1)$.

Conjecture 6.1 *Under H_0, for a suitable sequence $d(n_1)$ for which $d(n_1) \to \infty$ as $n_1 \to \infty$*

$$\begin{aligned} S_n &\xrightarrow{p} 1 \\ T_{k,n,n;S_n} &\xrightarrow{d} \chi^2_{k-1}. \end{aligned}$$

Conjecture 6.2 *For a suitable sequence $d(n_1)$ for which $d(n_1) \to \infty$ as $n_1 \to \infty$, the data-driven test based on $T_{k,n,n;S_n}$ is consistent against essentially any alternative.*

6.4.4 The Interpretation of the Components of $T_{k,c,n}$

In the previous section it was shown that in the limiting case the components of $T_{k,n,n}$ have the same clear interpretation as for a Neyman's smooth test based on Legendre polynomials. Omnibus consistency of the tests based on the truncated statistic $T_{k,n,n;s}$ or the data-driven version $T_{k,n,n;S_n}$ is only guaranteed when $s \to \infty$ or $d(n_1) \to \infty$, respectively (consistency of the data-driven version is only proven for $k = 2$). The SSPkc test, on the other hand, which has a decomposition into $c - 1$ terms, is consistent for any finite $c \geq 2$. Therefore, the SSPkc test may have the advantage that any departure from the null hypothesis is observed in the $c - 1$ components. The partition size c may be seen as the resolution of the decomposition. E.g., when $c = 2$, the $c - 1 = 1$ component is of course the test statistic itself, and since the SSPk2 test is consistent, any departure from the null hypothesis must be reflected in this one component. Thus, in this case, one cannot gain any information from studying the "component". $c = 2$ is the lowest resolution.
Increasing the partition size with one unit ($c = 3$), results in a decomposition into 2 terms. The first (Equation 6.15) can be interpreted as the average over all partitions $[A](\mathcal{P})$, of statistics that measure a change of location between k samples of c scores $(1, \ldots, c)$ (ties allowed). Each such statistic is of the form

$$\begin{aligned} & n \sum_{j=1}^{k} \left(\sum_{i=1}^{c} \frac{i - \hat{\mu}(\mathcal{P})}{\sqrt{\frac{n_j}{n}(\hat{\nu}(\mathcal{P}) - \hat{\mu}^2(\mathcal{P}))}} \hat{\mathrm{P}}_{f_{xy}}\left[B_{ij}(\mathcal{P})\right] \right)^2 \\ = \quad & \frac{\sum_{j=1}^{k} n_j \left(\bar{S}_j - \bar{S}\right)^2}{\hat{\sigma}^2}, \end{aligned} \tag{6.18}$$

where the scores S_{ij} of the observations X_{ij} are still such that $S_{ij} = m$ iff $X_{ij} \in A_m$ ($j = 1, \ldots, k; i = 1, \ldots, n_j$), but now $m = 1, \ldots, c$ with $c < n$. Hence, the statistic is still $n\frac{\text{SST}}{\text{SSTot}}$, but now it looks as if it is based on tied observations such that only c different values are distinguishable. It is however not the traditional Kruskal-Wallis statistic adjusted for ties. The main difference is that in the latter statistic the tied observations get a score assigned which is their midrank, which would in the present situation result in c midranks between 1 and n, whereas the scores in Equation 6.18 are always $1, 2, \ldots, c$, regardless of the relative ranking of the observations within each SSP element A_m ($m = 1, \ldots, c$). Further, note also that the number of ties is completely determined by c. The higher c, the less ties will occur, and in the limiting case $\mathbb{P}_{c-1,n} = \mathbb{P}_{n-1,n}$, there are no ties anymore (with probability one).
Thus, in conclusion, the first term is especially sensitive to the stochastic ordering of the k samples.

For the second term in the decomposition of $T_{k,c,n}$ a similar argument can be built, such that it may be concluded that the second term has the interpretation of an average measure (averaged over all partitions $[A](\mathcal{P})$) of a difference in dispersion between the k samples.

In general the effect of increasing the partition size / resolution c, is twofold. First, the number of terms in the decomposition increases. When each term can be related to a different feature of the distributions (e.g. location, dispersion, skewness, ...), more components mean that more aspects of the differences among the k distributions may be distinguished. Second, it has been shown that each component is an average of rank statistics, and that the number of ties decreases as c increases. As the number of ties decreases, these rank tests become more and more sensitive to specific deviations from the general k-sample null hypothesis. This is also clearly a resolution effect.

6.4.5 A Small Simulation Study

In order to get a view on how the components relate to each other, on average, under certain alternatives, a small simulation study is included. Actually the interpretation of the components are clear from the theoretical discussion in the previous section, but one question remains. Since each component is basically an average of rank statistics (e.g. Kruskal-Wallis) calculated on highly tied data (only c different values, irrespective of the number of observations), it is not clear to what extend the effect of ties might reduce the resolution or the specificity of the corresponding statistic. It is hoped that by averaging over all possible c-sized partitions some resolution is gained back.

Only the $k = 2$ sample situation is analyzed with 20 and 40 observations (balanced designs). The first sample is taken from a standard normal distribution, and the second from a normal distribution with mean μ and standard deviation one (shift in mean), or from a normal distribution with mean zero and standard deviation σ (change in dispersion). Each analysis is based on 1000 simulation runs. Both the SSP3 and the SSP4 statistic are decomposed; only the first two components are further considered. From the two

components $T^{(1)}_{k,c,n}$ and $T^{(2)}_{k,c,n}$, the quantity

$$F = \frac{T^{(1)}_{k,c,n}}{T^{(2)}_{k,c,n}},$$

is computed. After the 10000 simulation runs the average F value is calculated, which is denoted by $\bar{F}$, as well as the relative fraction f of F values that are greater than one, i.e. an estimate of the probability of obtaining a first component that is greater than the second. The results are presented in Table 6.12.

The results suggest that a relative comparison of the first and the second component may give some insight into the deviation from the null hypothesis. In general, when the difference in mean increases, the probability that the first component is greater than the second increases as well. This is seen for both sample sizes and for both SSP sizes. The opposite is seen when the difference in variances increases. Further, $\bar{F}$ and f change faster with μ and σ when the sample sizes are larger, and when the SSP size is greater. The latter is related to the resolution, i.e. the larger the SSP size c, the less ties, and the more specific the components are. The reason for this phenomenon is possibly that, when $c = 3$, the first component has the desired interpretation, though with a small resolution, whereas the second component is more to be interpreted as a rest term which thus also contains information on e.g. skewness, kurtosis. This is to be expected because the SSP3 test is omnibus consistent, which implies that any deviation from the null hypothesis is to be found in the test statistic. When $c = 4$, it is observed that f becomes more extreme (large when $\mu > 0$ and small when $\sigma > 1$), which, according to our reasoning, is because now the third component mainly takes over the role of rest term. The first two get more specialized.

6.4.6 Examples

EXAMPLE 6.5. **Gravity Data**:
In Sections 6.1.11 and 6.2.4 the SSPkc and its data-driven version were applied to the Gravity data. All SSP tests revealed that the 2 distributions (series 1 and 8) are significantly different at the 5% level.
For $c = 2, 3$ and $c = 4$ the decompositions of the SSPkc statistics are calculated and the results are presented in Table 6.13.

For both $c = 3$ and $c = 4$ the first two components are almost equal. When $c = 4$ the third component is also calculated and seems to be much smaller as compared to the first two. This suggests that the difference in distributions may be attributed to a difference in mean as well as a difference in scale.
An informal guideline that may show useful is to use for each component the critical value of the null distribution of $T_{k,c,n}$ with $c = 1$ as a benchmark value. In the present example with $c = 4$, $T^{(1)}_{k,c,n}$ and $T^{(2)}_{k,c,n}$ are both greater than 2.498 (the estimated exact critical

Table 6.12: The average fraction of the first over the second component and the estimated probability that the former is greater than the latter are presented for both the SSP2 and the SSP3 test in the 2-sample case. Sample sizes of $n = 20$ and $n = 40$ are studied.

n	σ	μ	$c = 3$		$c = 4$	
			f (%)	F	f (%)	F
20	1	0	35	1.33	35	1.77
		0.5	64	2.64	77	3.49
		1	84	5.32	87	8.04
		2	99	11.31		
	2	0	30	1.06	26	1.05
	3		25	1.02		
	4		26	0.91	14	0.61
	5		14	0.64		
40	1	0	35	1.18		
		0.5	71	3.41		
		1	96	9.62		
	2	0	20	0.86	18	0.63
	4		6	0.52	2.6	0.35

Table 6.13: The SSPkc statistics and their components ($c = 2, 3$ and $c = 4$) for the Gravity Data.

c	$T^{(1)}_{k,c,n}$	$T^{(2)}_{k,c,n}$	$T^{(3)}_{k,c,n}$	$T_{k,c,n}$
2	2.634	-	-	2.634
3	2.627	2.241	-	4.868
4	2.656	2.993	1.050	6.699

value). Moreover, the second component increases as c is changed from 3 to 4, which may be due to the increase in resolution (the second term becomes more specifically sensitive to differences in scale). Thus both a difference in location and in scale may be the reason for the rejection of H_0, but since particularly the second component showed a strong increase when c was changed from $c = 3$ to $c = 4$, we suggest that the difference in scale is more important. Recall that in Section 2.4 a difference in scale was also concluded (Mood's test).
Finally, note that the also boxplot in Figure 2.5 confirms a very large difference in variance between the measurements in series 1 and series 8. Also the difference in means is among all pairwise differences but one, the largest. □

EXAMPLE 6.6. **Sleep Data**:
In Sections 6.1.11 and 6.2.4 the SSPkc and the SSPdd tests were performed on the Sleep Data, and all of them showed that at least 2 distributions of the brain weight are different.

In Table 6.14 the components of the SSPkc statistic are presented for $c = 2, 3$ and $c = 4$. The first component is clearly greater than the other components. Moreover, $T^{(1)}_{k,c,n}$

increases with c (cfr. resolution), and it is greater than 6.727 (the estimated exact critical value of $T_{5,2,n}$) whereas the other components do not exceed it. The results suggest that the difference in brain weight between the 5 categories of exposure to danger may be attributed to a difference in average brain weight. This conclusion is supported by the boxplot in Figure 2.8 (b), although the graph also shows big differences in dispersion.

Table 6.14: The SSPkc statistics and their components ($c = 2, 3$ and $c = 4$) for the Sleep Data.

c	$T_{k,c,n}$	$T^{(1)}_{k,c,n}$	$T^{(2)}_{k,c,n}$	$T^{(3)}_{k,c,n}$
2	10.332	-	-	10.332
3	12.383	4.683	-	17.066
4	13.454	4.364	4.587	22.405

□

6.4.7 Other Decompositions

The decomposition that is constructed in Section 6.4.1 started with the construction of the $c \times k$ matrix $\boldsymbol{Q} = \boldsymbol{Q}(\mathcal{P}_{c-1,n})$, for which $\Delta_c(F; G) = \frac{1}{2n}\mathrm{E}\left[\mathrm{tr}\left[\boldsymbol{Q}\boldsymbol{Q}'\right]\right]$. The decomposition into $c - 1$ components was then obtained by introducing a $c \times c$ orthonormal matrix $\boldsymbol{R}$ which post-multiplied with $\boldsymbol{Q}$ results in $\mathrm{tr}\left[(\boldsymbol{R}\boldsymbol{Q})(\boldsymbol{R}\boldsymbol{Q})'\right] = \mathrm{tr}\left[\boldsymbol{Q}\boldsymbol{Q}'\right]$. Similarly a $k \times k$ orthonormal matrix $\boldsymbol{S}$ may be constructed, which pre-multiplied with $\boldsymbol{Q}$ gives again $\mathrm{tr}\left[(\boldsymbol{Q}\boldsymbol{S})'(\boldsymbol{Q}\boldsymbol{S})\right] = \mathrm{tr}\left[\boldsymbol{Q}'\boldsymbol{Q}\right] = \mathrm{tr}\left[\boldsymbol{Q}\boldsymbol{Q}'\right]$, and thus are the k sums of c squared elements of the columns of $\boldsymbol{Q}\boldsymbol{S}$ also components in the same sense as before.
In order to have nice interpretable components, the polynomials in $\boldsymbol{S}$ may be indicator functions arranged in $\boldsymbol{S}$ by the orthonormality condition in such a way that they make contrasts between the k samples. As a result, the components will give information on which samples are the most different.

CHAPTER 7

The Sample Space Partition Test for Independence

A formal definition of the hypothesis of independence has been given in Section 3.3.5. Although most of the methods that will be described in this chapter are easily extendable to general p-variate independence, most of this chapter will be focussed on bivariate independence. In this situation, the null hypothesis of independence between rvs X and Y, for which the notation $X \perp\!\!\!\perp Y$ of Dawid (1979) is used, is given by

$$\begin{aligned} H_0 : F_{xy}(x,y) &= F_x(x)F_y(y) \text{ for all } (x,y) \in \mathcal{S}_{xy}, \text{ or} \\ H_0 : f_{xy}(x,y) &= f_x(x)f_y(y) \text{ for all } (x,y) \in \mathcal{S}_{xy}, \end{aligned}$$

where $\mathcal{S}_{xy}$ is the sample space of (X,Y) and where where F_{xy} is the true joint CDF of rv $\boldsymbol{Z} = (X,Y)$, and $F_x(x)$ and $F_y(y)$ are the univariate marginal CDFs of X and Y, respectively. f_{xy}, f_x and f_y are defined analogously. The corresponding sample spaces are denoted by $\mathcal{S}_z = \mathcal{S}_{xy}$, $\mathcal{S}_x$ and $\mathcal{S}_y$, respectively. The sample of n bivariate observations is denoted by $\mathcal{S}_n$, and the samples of the univariate observations of X and Y are denoted by $\mathcal{S}_n^x$ and $\mathcal{S}_n^y$, respectively. Assumption A1 is assumed to hold for the distribution F_{xy}. The joint distribution under the null hypothesis will be denoted by $G = G_{xy}$ ($g = g_{xy}$). In the two previous chapters there was a need for an assumption on the relation between the sample spaces on which F and G are defined (cfr. Assumptions A2, A2b and A2c). Here, on the other hand, by the construction of the distribution G, it is guaranteed that

both F and G are defined on the same sample space $\mathcal{S}$.
The most general alternative will be considered, i.e. H_1 is just the negation of H_0 (omnibus alternative).

As for the two previous GOF problems, an SSP-based test will be constructed. First, the appropriate Directed Divergence and the corresponding Averaged Pearson Divergence are discussed, then the Averaged Pearson Divergence Statistics, on which the SSP test statistic will be constructed. Its null distribution will be determined, and a data-driven version is proposed. At the end of this chapter, the power is investigated in a small power study. Finally an extension to componentwise independence in a multivariate setting is discussed.

7.1 The SSP Test for Independence

7.1.1 The Directed Divergence

In Section 6.1.1 a directed divergence of order λ was proposed for the k-sample problem. It was constructed by introducing a working variable Y with a multinomial distribution. It was shown that the k-sample null hypothesis reduces to the condition of independence between the working variable Y and the discretized variable of interest, X. Both the continuous and the discrete directed divergence were then constructed accordingly. For the latter divergence, however, a rule had to be defined such that a partition $[A]$ of the sample space $\mathcal{S}_x$ of the univariate variable X uniquely induces a partition $[B]$ of the sample space $\mathcal{S}_{xy}$ of the bivariate variable (X, Y).

In the present situation, the bivariate independence condition is stated directly in the null hypothesis. Thus, the continuous directed divergence of order $\lambda \in \mathbb{R}$ between F_{xy} and G_{xy} becomes

$$\begin{aligned} \mathrm{I}^{\lambda}(F;G) &= \frac{1}{\lambda(\lambda+1)} \int_{\mathcal{S}_x} \int_{\mathcal{S}_y} \left[\left(\frac{f_{xy}(x,y)}{g_{xy}(x,y)} \right)^{\lambda} - 1 \right] f_{xy}(x,y)dxdy \\ &= \frac{1}{\lambda(\lambda+1)} \int_{\mathcal{S}_x} \int_{\mathcal{S}_y} \left[\left(\frac{f_{xy}(x,y)}{f_x(x)f_y(y)} \right)^{\lambda} - 1 \right] f_{xy}(x,y)dxdy. \end{aligned}$$

The corresponding directed divergence of order λ based on a SSP (discrete divergence) can be defined for very general partitions of $\mathcal{S}_{xy}$, but here we will restrict the SSPs to **rectangular** partitions.

Definition 7.1 *The elements $B_i = B_{i_1 i_2}$ of a* **rectangular $r \times c$ partition** *$[B]$ of $\mathcal{S}_{xy}$ are given by $A^x_{i_1} \times A^y_{i_2}$, where $A^x_{i_1}$ and $A^y_{i_2}$ are the elements of a r and c sized partition $[A^x]$ and $[A^y]$ of $\mathcal{S}_x$ and $\mathcal{S}_y$, respectively ($i_1 = 1, \ldots, r; i_2 = 1, \ldots, c$).*

For a rectangular $r \times c$ partition $[B]$ the discrete divergence of order λ is given by

$$\begin{aligned} \mathrm{I}^{\lambda}(F;G\|[B]) &= \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{r}\sum_{j=1}^{c} \mathrm{P}_{f_{xy}}[B_{ij}] \left[\left(\frac{\mathrm{P}_{f_{xy}}[B_{ij}]}{\mathrm{P}_{g_{xy}}[B_{ij}]} \right)^{\lambda} - 1 \right] \\ &= \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{r}\sum_{j=1}^{c} \mathrm{P}_{f_{xy}}[B_{ij}] \left[\left(\frac{\mathrm{P}_{f_{xy}}[B_{ij}]}{\mathrm{P}_{f_x}[A_i^x]\,\mathrm{P}_{f_y}[A_j^y]} \right)^{\lambda} - 1 \right] \end{aligned} \quad (7.1)$$

The importance of the discrete divergence is that it is sufficient to find one partition $[B]$ for which $\mathrm{I}^{\lambda}(F;G\|[B]) > 0$ to conclude that X and Y are dependent. This is stated in the following corollary.

Corollary 7.1 *If, for some λ, there exists a SSP $[B]$ of any size $r \times c$ ($r, c > 1$) for which $I^{\lambda}(F;G\|[B]) > 0$, then X and Y are dependent.*

7.1.2 The Power Divergence Statistic

The power divergence statistic corresponding to the discrete divergence introduced in the previous section is

$$\hat{\mathrm{I}}^{\lambda}(F;G\|[B])\,\frac{1}{\lambda(\lambda+1)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \hat{\mathrm{P}}_{f_{xy}}[B_{ij}] \left[\left(\frac{\hat{\mathrm{P}}_{f_{xy}}[B_{ij}]}{\hat{\mathrm{P}}_{f_x}[A_i^x]\,\hat{\mathrm{P}}_{f_y}[A_j^y]} \right)^{\lambda} - 1 \right], \quad (7.2)$$

where $\hat{\mathrm{P}}_{f_{xy}}[B_{ij}] = n^{-1}\#(B_{ij} \cap \mathcal{S}_n)$, $\hat{\mathrm{P}}_{f_x}[A_i^x] = n^{-1}\#(A_i^x \cap \mathcal{S}_n^x)$ and $\hat{\mathrm{P}}_{f_y}[A_j^y] = n^{-1}\#(A_j^y \cap \mathcal{S}_n^y)$.
As in the previous chapter, a partition may be constructed for which the power divergence statistic becomes infinitely large. This problem will be discussed in detail in Section 7.1.4. The results that are given in this section do not hold for such partitions.

In the present situation, each $r \times c$ SSP $[B]$ induces a $r \times c$ contingency table $\{N_{ij}\}$ ($i = 1, \ldots, r; j = 1, \ldots, c$), where

$$N_{ij} = \#(B_{ij} \cap \mathcal{S}_n).$$

Thus, the power divergence statistic of Equation 7.2 is the power divergence statistic for testing independence in the $r \times c$ contingency table induced by the SSP $[B]$. Special cases of the power divergence statistic include the Pearson ($\lambda = 1$) and the likelihood ratio ($\lambda \to 0$) statistic.

Cressie and Read (1984) showed that for any λ, and conditional on the SSP $[B]$, under H_0,

$$2n\hat{\mathrm{I}}^{\lambda}(F;G\|[B]) - 2n\hat{\mathrm{I}}^{1}(F;G\|[B]) \xrightarrow{p} 0,$$

and thus,

$$2n\hat{\mathrm{I}}^{\lambda}\left(F;G\|[B]\right) \xrightarrow{d} \chi^2_{(r-1)(c-1)}.$$

In the following sections we will consider only $\lambda = 1$, for which the statistic becomes

$$2n\hat{\mathrm{I}}^{1}\left(F;G\|[B]\right) = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\hat{\mathrm{P}}_{f_{xy}}\left[B_{ij}\right] - \hat{\mathrm{P}}_{f_x}\left[A_i^x\right]\hat{\mathrm{P}}_{f_y}\left[A_j^y\right]\right)^2}{\hat{\mathrm{P}}_{f_x}\left[A_i^x\right]\hat{\mathrm{P}}_{f_y}\left[A_j^y\right]}.$$

The power divergence statistic may be interpreted as $2n$ times the plug-in estimator of the discrete divergence. From the asymptotic null distribution, conditional on a SSP $[B]$, it is seen immediately that it is a biased estimator ($\mathrm{E}_g\left[2n\hat{\mathrm{I}}^{\lambda}\left(F;G\|[B]\right)\right] \to (r-1)(c-1)$ as $n \to \infty$). But, on the other hand the α-level test based on the power divergence statistic is consistent. Thus, whenever $\mathrm{I}^{\lambda}\left(F;G\|[B]\right) > 0$, the power of the corresponding test tends to one as n tends to infinity.
The condition $\mathrm{I}^{\lambda}\left(F;G\|[B]\right) > 0$ of Corollary 7.1 can thus be assessed by applying the corresponding statistical test. But since the same corollary implies that one has to keep on looking until a SSP is found for which it can be concluded that the discrete divergence is different from zero, the problem of multiplicity comes into play. One solution to overcome the multiple testing problem it to combine the discrete divergences $\mathrm{I}^{\lambda}\left(F;G\|[B]\right)$ into one single new functional, and then to look for an appropriate test statistic related to this functional. This is discussed in the next sections.

7.1.3 The Averaged Pearson Divergence

As for the one-sample GOF and the k-sample GOF problem, the averaged Pearson divergence is the mean of a set of directed divergences based on representative SSPs. These SSPs are determined by the partition construction rule (Definition 5.5) and the partition determining subsample set (Definition 5.4). In contrast to the previous two GOF problems, there is here not necessarily a one-to-one relation between the number of observations (q) in the sets $\mathcal{P}_{q,n} \in \mathbb{P}_{q,n}$ and the partition size $r \times c$. This is illustrated in Figure 7.1 where two ways of constructing a 2×3 SSP are presented. In panel (a) only $q = 2$ observations are used, and in panel (b) $q = 3$ observations are needed. Both construction rules are valid. Therefore, the notations $\mathbb{P}_{q,n}$, $\mathcal{P}_{q,n}$, $\mathcal{A}_{q,n}$ and $a_{q,n}$ are extended to $\mathbb{P}^{r,c}_{q,n}$, $\mathcal{P}^{r,c}_{q,n}$, $\mathcal{A}^{r,c}_{q,n}$ and $a^{r,c}_{q,n}$, respectively. The exact relation between q and (r,c) is given by the partition construction rule.

Thus, as before, the SSP $[B](\mathcal{P}^{r,c}_{q,n})$ is a random, data-dependent partition, and consequently, its distribution depends on the distribution of the observations.
The averaged Pearson divergence is now defined as the average of all directed divergences of order 1, over all possible $r \times c$ SSPs in the set $\mathcal{A}^{r,c}_{q,n}$, according to the true distribution F_{xy} of the observations.

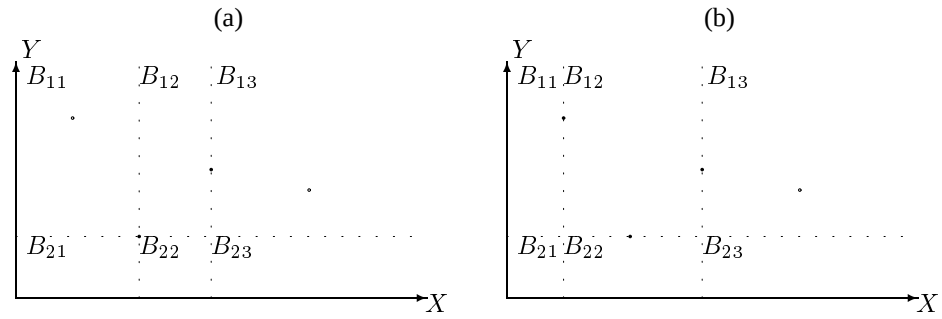


Figure 7.1: Two 2×3 partitions are shown. The dashed lines represent the partition boundaries. The filled and the hollow dots represent the observations that are used or not used for the construction of the partition, respectively.

Definition 7.2 *The* **Averaged Pearson Divergence** *is defined as*

$$\Delta_{r,c}(F;G) = E_f\left[I^1\left(F;G\|[B](\mathcal{P}_{q,n}^{r,c})\right)\right].$$

The relevance of the averaged Pearson divergence for the independence problem is stated in the following lemma.

Lemma 7.1 *For all* r, c *(*$r, c > 1$*)*

$$\Delta_{r,c}(F;G) = 0 \Leftrightarrow X \perp\!\!\!\perp Y.$$

Proof. The proof is completely analogous to the proof of Lemma 5.1. □

In the next section the related statistic is discussed.

7.1.4 The Averaged Pearson Divergence Statistic

The Averaged Pearson Divergence Statistic is defined as the plug-in estimator of the Averaged Pearson Divergence. Let $[B]_{\mathcal{P}}$ denote $[B](\mathcal{P}_{q,n}^{r,c})$. Then,

$$\begin{aligned}\hat{\Delta}_{r,c}(F;G) &= \Delta_{r,c}(\hat{F},\hat{G}) \\ &= \int_{\mathcal{S}_{\mathcal{P}}} \hat{\mathrm{I}}^1\left(F;G\|[B]_{\mathcal{P}}\right) d\hat{F}_{\mathcal{P}}(\mathcal{P}).\end{aligned}$$

A further reduction of the statistic towards a computational formula is possible if the partition construction rule is specified. To illustrate the effect of the construction rule, we give here two different rules for the case $r = c = 2$. The rectangular partitions that are constructed by both rules use one observation for splitting the sample space $\mathcal{S}_x$ and one for $\mathcal{S}_y$. The rules differ in the relation between these two observations.

- A bivariate observation $\boldsymbol{Z}_i = (X_i, Y_i)$ may be used to split both sample spaces in the same way as partitions were constructed for the SSPc one-sample test. Thus $[A^x] = \{]b_l^x, X_i],]X_i, b_u^x]\}$ and $[A^y] = \{]b_l^y, Y_i],]Y_i, b_u^y]\}$, and $\mathbb{P}_{1,n}^{2,2} = \mathcal{S}_n \setminus \{\boldsymbol{Z}_{(n)_x}, \boldsymbol{Z}_{(n)_y}\}$, where $\boldsymbol{Z}_{(i)_x}$ and $\boldsymbol{Z}_{(i)_y}$ are the bivariate component-wise order statistics, according to the X and the Y component, respectively (see e.g Barnett, 1976, for details on mutivariate orderings). The largest observations may not be used for it makes the statistic $\hat{\Delta}$ infinitely large. A similar problem occurred in the k-sample SSPkc statistic as well. The resulting partition may be denoted by $[B](X_i, Y_i)$. Hence,

$$\begin{aligned}\hat{\Delta}_{r,c}(F;G) &= \int_{\mathcal{S}_{xy}} \hat{\mathrm{I}}^1\left(F;G\|[B](x,y)\right) d\hat{F}_{xy}(x,y) \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1} \hat{\mathrm{I}}^1\left(F;G\|[B](X_{(i)},Y_{(i)})\right).\end{aligned}$$

- The second method makes use of two observations, $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$. Of the first observation the X-component determines the partition $[A^x]$ and, similarly, of the second observation the Y component is used to build the $[A^y]$ partition. Now the statistic becomes

$$\begin{aligned}\hat{\Delta}_{r,c}(F;G) &= \int_{\mathcal{S}_{xy}}\int_{\mathcal{S}_{xy}} \hat{\mathrm{I}}^1\left(F;G\|[B](\boldsymbol{z}_x,\boldsymbol{z}_y)\right) d\hat{F}_{xy}(\boldsymbol{z}_x) d\hat{F}_{xy}(\boldsymbol{z}_y) \\ &= \int_{\mathcal{S}_x}\int_{\mathcal{S}_y} \hat{\mathrm{I}}^1\left(F;G\|[B](x,y)\right) d\hat{F}_x(x) d\hat{F}_y(y) \\ &= \frac{1}{(n-1)^2}\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} \hat{\mathrm{I}}^1\left(F;G\|[B](X_{(i)},Y_{(j)})\right).\end{aligned}$$

When r or c are greater than 2, the two methods are readily extended. The first uses the minimal number of observations to construct the partitions $[A^x]$ and $[A^y]$. In particular, $q = r \wedge c - 1$ observations are needed, i.e. suppose that $r \geq c$, then $r-1$ observations are needed to construct the r elements of the $[A^x]$ partition, and of these r elements, only the $c-1$ first observations are used to deliver the Y component for the $[A^y]$ partition. Also, in order to avoid $\hat{\Delta}_{r,c}(F;G)$ to become infinitely large, the observations $\boldsymbol{Z}_{(n)_x}$ and $\boldsymbol{Z}_{(n)_y}$ are not allowed in the partition determining subsample set. In the remainder of this chapter, this partition construction rule will be used.
For the second method, $(r-1)+(c-1)$ observation are needed. This first $r-1$ to construct $[A^x]$ and the $c-1$ remaining observations for the $[A^y]$ partition.

7.1.5 The Test Statistic and its Asymptotic Null Distribution

The test statistic is given by

$$T_{r,c,n} = 2n\hat{\Delta}_{r,c}(F;G).$$

Thus, the test statistic is clearly a rank statistic. Indeed, the statistic depends on the data only through the probability estimators $\hat{\mathrm{P}}_{f_{xy}}[B_{ij}]$, $\hat{\mathrm{P}}_{f_x}[A_i^x]$ and $\hat{\mathrm{P}}_{f_y}[A_j^y]$, which only depend on the relative ranking of the observations because the partition elements are constructed on the observations themselves. Thus, the test that will be constructed from $T_{r,c,n}$ will be a nonparametric test in the sense that its null distribution does not depend on the true distribution F_xF_y.
Furthermore, the null hypothesis of independence implies a group of permutations under which the distribution of (X,Y) is invariant, and thus the (conditional) exact null distribution can be enumerated. This will be discussed later. First, the asymptotic null distribution will be given.

Throughout this chapter it is assumed that the distribution F_{xy} is continuous, which implies that no ties occur with probability one. If, however, in practice a sample should contain ties, due to e.g. rounding off, then some changes might be necessary. Basically these adaptations are similar to those for the SSPkc test, for which a brief discussion was given in Section 6.1.10.

Special Case: $r = c = 2$

Consider the case where $r = c = 2$ and $q = 1$, then the partition construction rule says that each 2×2 partition is *centred* about one observation. The observations with the highest rank among the X- and the Y-variables cannot used as centre. Thus, the sets $\mathcal{P}_m \in \mathbb{P}_{1,n}^{2,2}$ are $\mathcal{P}_m = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n-d}\}$ ($d = 1$ or 2, depending on the bivariate ranking of the X and the Y components of $\boldsymbol{Z}$). Let the elements of $[B]_m = [B]_{\mathcal{P}_m}$ be denoted by B_{mij}. The corresponding elements of the SSPs of $\mathcal{S}_x$ and $\mathcal{S}_y$ are denoted by A_{mi}^x and A_{mj}^y, respectively ($i = 1, 2$; $j = 1, 2$). Then the test statistic becomes

$$\begin{aligned} T_{2,2,n} &= \frac{1}{n-d}\sum_{m=1}^{n-d}\sum_{i,j=1}^{2} n\frac{\left(\hat{\mathrm{P}}_{f_{xy}}[B_{mij}] - \hat{\mathrm{P}}_{f_x}[A_{mi}^x]\,\hat{\mathrm{P}}_{f_y}[A_{mj}^y]\right)^2}{\hat{\mathrm{P}}_{f_x}[A_{mi}^x]\,\hat{\mathrm{P}}_{f_y}[A_{mj}^y]} \\ &= \frac{1}{n-d}\sum_{m=1}^{n-d}\frac{\left\{\sqrt{n}\left(\hat{F}_{xy}(x_m,y_m) - \hat{F}_x(x_m)\hat{F}_y(y_m)\right)\right\}^2}{\hat{F}_x(x_m)\hat{F}_y(y_m)\left(1-\hat{F}_x(x_m)\right)\left(1-\hat{F}_u(y_m)\right)} \end{aligned}$$

Theorem 7.1 *Under H_0, as $n \to \infty$,*

$$T_{2,2,n} \xrightarrow{d} \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{1}{j(j+1)k(k+1)} Z_{jk}^2,$$

where the Z_{jk} ($j,k = 1,\ldots$) are i.i.d. standard normal variates.

Proof. Throughout the whole proof, it is assumed that H_0 is true.

Let $U(s,t)$ ($(s,t) \in [0,1]^2$) be a Brownian bridge (separable Gaussian process) with "time" parameter (s,t). Then, for $(s,t),(u,v) \in [0,1]^2$,

$$\begin{aligned} \mathrm{E}\left[U(s,t)\right] &= 0 \\ \mathrm{Cov}\left[U(s,t),U(u,v)\right] &= (s \wedge u - su)(t \wedge v - tv). \end{aligned}$$

The first part of the proof consists of showing that $T_{2,2,n}$ converges in distribution to

$$T = \int_0^1 \int_0^1 \frac{U^2(s,t)}{s(1-s)t(1-t)} dsdt.$$

We will use techniques that have been used by e.g. Kiefer (1959), Pettit (1976), Rosenblatt (1952), Scholz and Stephens (1987).

The test statistic $T_{2,2,n}$ may be written as

$$T_{2,2,n} = \int_{A_{ns}} \int_{A_{nt}} \frac{V_n^2(s,t)}{\hat{F}_s(s)(1-\hat{F}_s(s))\hat{F}_s(t)(1-\hat{F}_s(t))} d\hat{F}_s(s) d\hat{F}_t(t),$$

where

$$V_n(s,t) = \sqrt{n}\left(\hat{F}_{xy}(F_x^{-1}(s),F_y^{-1}(t)) - \hat{F}_s(s)\hat{F}_t(t)\right),$$

where $\hat{F}_s$ and $\hat{F}_t$ are EDFs of the sample of uniform distributed observations which are obtained by the integral transformation $s = F_x(x)$ and $t = F_y(y)$. Further, the set $A_{ns} = \{s \in [0,1] : 0 < \hat{F}_s(s) < 1\}$.
The process $V_n(s,t)$ is a Gaussian process with asymptotically the same moments as $U(s,t)$. This is seen by (1) noting that

$$\hat{F}_{xy}(F_x^{-1}(s),F_y^{-1}(t)) = \hat{F}_{st}(s,t),$$

where $\hat{F}_{st}$ is the EDF of the copula associated with F_{xy}, and (2) by writing (Blum et al., 1961)

$$\hat{F}_s(s)\hat{F}_t(t) = s\hat{F}_t(t) + t\hat{F}_s(s) - st + O_p(n^{-1}),$$

which is based on the consistency of the estimators. Then, by methods of Pyke and Shorack (1968) and Billingsley (1968, theorem 4.2)

$$T_{2,2,n} \xrightarrow{d} T. \tag{7.3}$$

(The convergence is obtained in two steps: first the convergence of the central part, i.e. over A_{ns} and A_{nt} is established, and then it is shown that the remaining part is negligible.)

The remainder of the proof is a combination of results from Anderson and Darling (1952), Blum et al. (1961). We give a sketch of the proof.

Let $k(s,t,u,v)$ be a continuous function in $[0,1]^4$ (possible not continuous at the corners; details on the exact conditions are given by Anderson and Darling (1952)) which is square integrable in (s,t), (u,v) and $((s,t),(u,v))$. Then it can be expressed as

$$k(s,t,u,v) = \sum_{j=1}^{\infty} \lambda_j f_j(s,t) f_j(u,v),$$

where $\{f_j(s,t)\}$ ($(s,t) \in [0,1]^2$) is a system of orthonormal functions w.r.t. the Lebesgue measure on $[0,1]^2$, i.e.

$$\int_0^1 \int_0^1 \int_0^1 \int_0^1 f_i(s,t) f_j(u,v) ds dt du dv = \delta_{ij},$$

where δ_{ij} is Kronecker's delta ($i,j = 1,\dots$). Let X_j ($j = 1,\dots$) be i.i.d. standard normal variates. We define the process $Z(s,t) = \sum_{j=1}^{\infty} \sqrt{\lambda_j} f_j(s,t) X_j$. It's immediately checked that $Z(s,t)$ is a Gaussian process for which for all $s,t,u,v \in [0,1]$, $\mathrm{E}\left[Z(s,t)\right] = 0$ and $\mathrm{Cov}\left[Z(s,t), Z(u,v)\right] = k(s,t,u,v)$.

When $k(s,t,u,v) = \left(s(s-1)t(t-1)u(u-1)v(v-1)\right)^{-1/2} (s \wedge u - su)(t \wedge v - tv)$, it follows immediately that $Z(s,t)$ is the same stochastic process as

$$\left(s(s-1)t(t-1)\right)^{-1/2} U(s,t).$$

Thus,

$$\begin{aligned} T &= \int_0^1 \int_0^1 \frac{U^2(s,t)}{s(s-1)t(t-1)} ds dt \\ &= \int_0^1 \int_0^1 Z^2(s,t) ds dt \\ &= \sum_{j=1}^{\infty} \lambda_j X_j^2. \end{aligned}$$

Since $k(s,t,u,v)$ is expressed as an infinite series w.r.t. a orthonormal system, the λ_j and the $f_j(s,t)$ can be found as the eigenvalues and the eigenfunctions, respectively, of the

integral equation

$$\int_0^1 \int_0^1 k(s,t,u,v) f(s,t) ds dt = \lambda f(u,v). \tag{7.4}$$

Since $k(.)$ is separable, i.e. $k(s,t,u,v) = k'(s,u)k'(t,v)$, where

$$k'(s,u) = (s(1-s)u(1-u))^{-1/2} (s \wedge u - su),$$

the integral equation reduces to the product of two identical integral equations of the form

$$\int_0^1 k'(s,u) f'(s) ds = \gamma f'(u),$$

which is exactly the integral equation that has been solved by Anderson and Darling (1952), and which gives for the eigenvalues the solutions ($k = 1, \ldots$)

$$\gamma_k = \frac{1}{k(1+k)}.$$

And thus the eigenvalues λ_j of the integral equation of Equation 7.4 are given by

$$\lambda_j = \lambda_{kl} = \gamma_k \gamma_l = \frac{1}{k(1+k)l(1+l)}$$

($k, l = 1, \ldots$). This completes the proof. □

A well known related statistic is Blum et al. (1961), Hoeffding (1948)

$$B_n = \int_0^1 \int_0^1 U_n^2(s,t) d\hat{F}_{xy}(s,t),$$

which is a statistic in the spirit of the Cramér - von Mises statistics.

The General Case

Proposition 7.1 *For any finite $r, c > 1$, under the null hypothesis of independence, as $n \to \infty$,*

$$T_{r,c,n} \xrightarrow{d} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \frac{1}{j(1+j)k(1+k)} X_{jk}^2,$$

where the X_{jk}^2 ($j, k = 1, \ldots$) are i.i.d. χ^2 distributed variates with $(r-1)(c-1)$ degrees of freedom.

Heuristic Proof. The proof for the general case is based on a partitioning rule of Lancaster (1949) for the Pearson statistic for independence in an $r \times c$ contingency table.

Consider $[B]_{\mathcal{P}} \in \mathcal{A}_{q,n}^{r,c}$. Let

$$[B_s^{ij}]_{\mathcal{P}} = \{\{B_{ij}\}, \{\cup_{k<i} B_{kj}\}, \{\cup_{l<j} B_{il}\}, \{\cup_{k<i, l<j} B_{kl}\}\},$$

($i = 2, \ldots, r; j = 2, \ldots, c$) be a sequence of 2×2 partitions of the subsets $\mathcal{S}_s^{ij} \subseteq \mathcal{S}_{xy}$. The corresponding subsamples are $\mathcal{S}_n^{ij} = \mathcal{S}_n \cap \mathcal{S}_s^{ij}$ with sample sizes n_{ij}.

Conditional on the SSP $[B]_{\mathcal{P}}$ the statistics $2n_{ij}\hat{\mathrm{I}}^1\left(F; G \| [B_s^{ij}]_{\mathcal{P}}\right)$ ($i = 2, \ldots, r; j = 2, \ldots, c$) are asymptotically independent (Lancaster, 1949). Moreover, still according to Lancaster (1949) the following decomposition has asymptotically the same distribution as $2n\hat{\mathrm{I}}^1\left(F; G \| [B]_{\mathcal{P}}\right)$,

$$\sum_{i=2}^{r}\sum_{j=2}^{c} 2n_{ij}\hat{\mathrm{I}}^1\left(F; G \| [B_s^{ij}]_{\mathcal{P}}\right).$$

Thus, the test statistic $T_{r,c,n}$ has asymptotically the same distribution as

$$T_n = \frac{1}{a_{q,n}^{r,c}} \sum_{\mathcal{P} \in \mathbb{P}_{q,n}^{r,c}} \sum_{i=2}^{r}\sum_{j=2}^{c} 2n_{ij}\hat{\mathrm{I}}^1\left(F; G \| [B_s^{ij}]_{\mathcal{P}}\right).$$

The same arguments that have led to the convergence in Equation 7.3 may now be applied to each of the $(r-1)(c-1)$ asymptotically independent terms of the decomposition separately, resulting in a rv T with asymptotically the same distribution as $T_{r,c,n}$,

$$T = \int_0^1 \cdots \int_0^1 \sum_{i=2}^{r}\sum_{j=2}^{c} \frac{U_{ij}^2(s_{k_s(ij)}, t_{k_t(ij)})}{s_{k_s(ij)}(1 - s_{k_s(ij)}) t_{k_t(ij)}(1 - t_{k_t(ij)})} ds_1 \ldots ds_{q_s} dt_1 \ldots dt_{q_t}, \tag{7.5}$$

where the $U_{ij}(.,.)$ are independent Brownian bridges indexed by 2-dimensional "time" parameters, and where $k_s(ij)$ and $k_t(ij)$ are functions taking values in $1, \ldots, q_s \leq q$ and $1, \ldots, q_t \leq q$, which determine the indices of s and t, respectively, in accordance with the partition construction rule and the observations in the partition determining subset $\mathcal{P}$. Each term in the integrandum of Equation 7.5 has the same form as the integrandum of Equation 7.3. Thus, as in the proof of Theorem 7.1, each process

$$V_{ij}(s_{k_s(ij)}, t_{k_t(ij)}) = \frac{U_{ij}(s_{k_s(ij)}, t_{k_t(ij)})}{\sqrt{s_{k_s(ij)}(1 - s_{k_s(ij)}) t_{k_t(ij)}(1 - t_{k_t(ij)})}}$$

may be replaced by its expansion

$$V_{ij}(s_{k_s(ij)}, t_{k_t(ij)}) = \sum_{m=1}^{\infty} \sqrt{\lambda_m} f_m(s_{k_s(ij)}, t_{k_t(ij)}) X_{ij,m},$$

where the $X_{ij,m}$ are i.i.d. standard normal variates, and where λ_m and $f_m(.,.)$ are the eigenvalues and eigenfunctions of the integral Equation 7.4. Thus, the expansion becomes (index m expanded to k, l)

$$V_{ij}(s_{k_s(ij)}, t_{k_t(ij)}) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{f_{kl}(s_{k_s(ij)}, t_{k_t(ij)})}{\sqrt{k(k+1)l(l+1)}} X_{ij,kl}.$$

Then,

$$\begin{aligned} T &= \int_0^1 \dots \int_0^1 \sum_{i=2}^{r} \sum_{j=2}^{c} V_{ij}^2(s_{k_s(ij)}, t_{k_t(ij)}) ds_1 \dots ds_{q_s} dt_1 \dots dt_{q_t} \\ &= \sum_{i=2}^{r} \sum_{j=2}^{c} \left(\sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{1}{k(k+1)l(l+1)} X_{ij,kl}^2 \right) \\ &= \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{1}{k(k+1)l(l+1)} X_{kl}^2, \end{aligned}$$

where $X_{kl}^2 = \sum_{i=2}^{r} \sum_{j=2}^{c} X_{ij,kl}^2$, i.e. a sum of $(r-1)(c-1)$ i.i.d. squared standard normal variates. Thus, X_{kl}^2 are i.i.d. $\chi^2_{(r-1)(c-1)}$. This completes the proof. □

The asymptotic null distribution of $T_{r,c,n}$ is again a weighted sum of χ^2-distributed rvs, and thus a general quadratic form in the sense of Solomon and Stephens (1978). Hence, the same type of approximations as before may be considered, for which one needs the cumulants. These are given in the following corollary.

Corollary 7.2 *The ith cumulant of the asymptotic null distribution of $T_{r,c,n}$ is given by*

$$\mu_i = (r-1)(c-1)2^{i-1}(i-1)! \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \left(\frac{1}{j(j+1)k(k+1)} \right)^i.$$

Proof. The proof is analogue to the proof of Corollary 5.5. □

7.1.6 The SSPrc Test

The α-level SSPrc test $\phi_{\alpha,r,c,n}(\mathcal{S}_n)$ for the independence problem is defined as

$$\begin{aligned}\phi_{\alpha,r,c,n}(\mathcal{S}_n) &= 1 \text{ if } T_{r,c,n} > t_{\alpha,r,c,n} \\ \phi_{\alpha,r,c,n}(\mathcal{S}_n) &= 0 \text{ if } T_{r,c,n} \leq t_{\alpha,r,c,n},\end{aligned}$$

where $t_{\alpha,r,c,n}$ is the $(1-\alpha)$th percentile of the exact null distribution of $T_{r,c,n}$. The percentiles $t_{\alpha,r,c}$ denote those of the asymptotic null distribution, and the corresponding asymptotic test is denoted by $\phi_{\alpha,r,c}(\mathcal{S}_n)$.

Theorem 7.2 *For any finite $r, c \geq 2$, the SSPrc test is consistent against essentially any alternative.*

Proof. First consistency of the SSP22 test is proven.

Suppose that the true CDF $F_{xy}(x,y) \neq F_x(x)F_y(y)$ for at least one $(x,y) \in \mathcal{S}_{xy}$, then, by the continuity of the CDF,

$$\frac{1}{n}T_{2,2,n} \xrightarrow{a.s.} \int \frac{\left(F_{xy}(x,y) - F_x(x)F_(y)\right)^2}{F_x(x)(1-F_x(x))F_y(y)(1-F_y(y))} dF_{xy}(x,y) > 0. \tag{7.6}$$

Hence, $T_{2,2,n}$ grows unboundedly with n, and thus the SSP22 test is consistent.

The extension to the SSPrc test is straightforward. Consider the same asymptotic decomposition as in the proof of Proposition 7.1. From that decomposition it is seen that $T_{r,c,n}$ may be written as a sum of $T_{2,2,n}$ and $(r-1)(c-1)-1$ other positive statistics. Since, according to the first part of the proof, $T_{2,2,n}$ grows almost surely unboundedly with n, the $T_{r,c,n}$ statistic will grow unboundedly as well. Hence, the SSPrc test is consistent as well. □

7.1.7 The SSPrc Permutation Test

The statistic $T_{r,c,n}$ is clearly a rank statistic since it only depends on the data through the rankings of the observations: a ranking of the n observations $\{x_i\}$, a ranking of the n observations $\{y_i\}$, and a "ranking" of the observations $\{x_i\}$ relative to the ranking of observations $\{y_i\}$. Concerning the latter ranking, the null hypothesis of independence implies that any ranking of the observations $\{x_i\}$ relative to the ranking of $\{y_i\}$ is equally likely. More specifically, conditional on the observed data $\mathcal{S}_n = \{(x_1,y_1),\ldots(x_n,y_n)\}$, any transformed sample $\mathcal{R}_n = \{(x_1,y_{i_1}),\ldots,(x_n,y_{i_n})\}$, where $i_1,\ldots,i_n$ is a permutation of $1,\ldots n$, has an equal conditional probability under the null hypothesis. More formally, the null hypothesis implies that the distribution $G = F_xF_y$ is invariant under

the group $\mathcal{H}$ of transformations h of $\mathcal{S}_n$ onto itself,

$$\mathcal{H} = \{h : h(\mathcal{S}_n) = \{(x_1, y_{i_1}), \ldots, (x_n, y_{i_n})\} \text{ where } i_1, \ldots, i_n \text{ is a permutation of } 1, \ldots n\},$$

and $\#\mathcal{H} = n!$. Thus, conditional on $\mathcal{S}_n$, and under H_0, for any $h \in \mathcal{H}$,

$$\mathrm{P}_g\left[(X_1, Y_1) = (x_1, y_{i_1}), \ldots, (X_n, Y_n) = (x_n, y_{i_1})\right] = \frac{1}{n!}. \tag{7.7}$$

Note that in Equation 7.7 the observations may be replaced by their componentwise ranks.

Let $T(\mathcal{S}_n)$ denote the SSPrc test statistic based on the bivariate observations in $\mathcal{S}_n$. Let

$$T^{(1)}(h(\mathcal{S}_n)) \leq \ldots \leq T^{(n!)}(h(\mathcal{S}_n))$$

be the ordered values of $T(h(\mathcal{S}_n))$ as h varies in $\mathcal{H}$. Then, conditionally on $\mathcal{S}_n$, the values of $T^{(i)}(.)$ represent the exact conditional distribution of $T_{r,c,n}$. The conditional distribution coincide with the unconditional exact distribution because the test statistic is a rank statistic and because it is assumed that the distribution F_{xy} is continuous. Thus, especially for large but finite sample sizes n, when the null distribution of $T_{r,c,n}$ is estimated by means of Monte Carlo simulation, the resulting estimate of the null distribution is both an estimate of the conditional exact permutation distribution and of the unconditional distribution. Corollary 6.3 is still applicable, indicating that the permutation test is conservative. To overcome this problem, a randomization test may be defined in exactly the same way as explained in Section 6.1.8.

7.1.8 Convergence to the Limiting Distribution

The convergence is studied here for two purposes. First, a fast convergence would mean that that the asymptotic null distribution is an appropriate approximation for moderate sample sizes. Second, the validity of Proposition 7.1 must be assessed empirically.

Exact critical values $t_{\alpha,r,c,n}$ are estimated by means of Monte Carlo simulation at nominal level $\alpha = 0.05$, for partition sizes 2×2, 3×3 and 4×4, and for sample sizes $n = 10, 20, 50, 100, 500$ and 1000. All estimations are based on 10000 simulation runs. The results are presented in Table 7.1. The corresponding approximate asymptotic critical values $t_{\alpha,r,c}$ are also given. For each estimated exact critical value, the estimated exact level if the asymptotic critical value would have been used, is mentioned as well.

The estimations show that the convergence is very slow for all three partition sizes. Thus the asymptotic null distribution is not of much practical value.
The results for the 3×3 and the 4×4 partitions confirm the proposed null distributions of Proposition 7.1.

Table 7.1: Approximated exact critical values ($\alpha = 0.05$) of the SSP22, SSP33 and SSP44 tests with $n = 10, 20, 50, 100, 500$ and $n = 1000$. The entries at $n = \infty$ are the approximated asymptotic critical values. Between brackets, the estimated exact level, when the asymptotic critical value would have been used, is shown.

n	2×2	3×3	4×4
10	2.505 (0.135)	7.820 (0.335)	14.792 (0.527)
20	2.383 (0.127)	7.536 (0.303)	14.664 (0.515)
50	2.242 (0.108)	7.002 (0.230)	14.391 (0.448)
100	2.122 (0.089)		
500	1.920 (0.072)	6.473 (0.127)	
1000	1.800 (0.048)	5.660 (0.055)	
∞	1.823	5.551	11.237

7.1.9 Examples

All critical values and p-values are estimated from the simulated exact null distribution (10000 simulation runs).

EXAMPLE 7.1. **Cultivars data**
A key question for the Cultivars data was whether or not there is a relation between the difference in yields between the two cultivars and the covariate AZ.

SSPrc tests are here performed with $r \times c$ set to 2×2, 3×3 and 4×4. The results, as as well as the critical values for $n = 19$ are presented in Table 7.2. When the partition size is 2×2, the SSPrc test resulted in a significant dependence at the 5% level. For the larger partition sizes, non-significant results are obtained. This example stresses the importance of choosing an appropriate partition size.

Table 7.2: The results of the SSPrc test on the Cultivars data. Between brackets the p-value is given. Also the critical values are shown.

$r \times c$	$t_{r,c,19}$	$t_{0.05,r,c,19}$
2×2	2.894 (0.0251)	2.445
3×3	7.135 (0.0661)	7.391
4×4	13.053 (0.191)	15.267

□

EXAMPLE 7.2. **Sleep data**
One wants to assess the dependence between the variables “brain weight” and “gestation”. Since the scatter plot of the data (Figure 2.8) clearly showed heteroscedasticity such that even a linearity assumption may be questioned, a nonparametric approach may be desirable.

SSPrc tests are here performed with $r \times c$ set to 2×2, 3×3 and 4×4. The results, as as well as the critical values for $n = 58$ are presented in Table 7.3. For all partition sizes considered, the SSPrc test resulted in a highly significant dependence at the 5% level.

Table 7.3: The results of the SSPrc test on the Sleep data. Between brackets the p-value is given. Also the critical values are shown.

$r \times c$	$t_{r,c,58}$	$t_{0.05,r,c,58}$
2×2	16.596 (< 0.0001)	2.212
3×3	31.051 (< 0.0001)	6.959
4×4	44.394 (< 0.0001)	14.328

□

EXAMPLE 7.3. **Ethanol Data**
Visually there is seen a clear quadratic relation between nitric oxide and the equivalence ratio. The results of the SSPrc tests with partition sizes are given in Table 7.4. All tests result in highly significant dependence.

Table 7.4: The results of the SSPrc test on the Ethanol data. Between brackets the p-value is given. Also the critical values are shown.

$r \times c$	$t_{r,c,88}$	$t_{0.05,r,c,88}$
2×2	16.663 (< 0.0001)	2.082
3×3	41.381 (< 0.0001)	6.808
4×4	66.955 (< 0.0001)	14.377

□

7.2 The Data-Driven SSPrc Test

The reasons for constructing a data-driven version of the SSPrc test are very similar to those mentioned for the SSP test for the general GOF problem and for the k-sample problem. First it is important to realize that it is not needed to make the test consistent since Theorem 7.2 guaranteed that for any choice of the partition size $r \times c$ the SSPrc test is consistent. But, if the partition size is chosen too large as compared to the dependence structure between X and Y, then, due to the increase of the variance of the test statistic as a consequence of the large number of degrees of freedom ($(r-1)(c-1)$), the power will be lower than the power that would have been attained when a smaller partition size would have been chosen, i.e. the dilution effect. Also, a too low partition size might result in a test statistic that is not as sensitive to the dependence structure between X and Y as a test based on a larger partition size would have been.

7.2.1 The Selection Rule

Most of the theory on data-driven SSP based tests that is given in Section 5.5 remains valid in the present situation. Only the results that are related to the selection rule must be changed accordingly.

The definition of the permissible partition sizes must be changed slightly, and will depend on the definition of the minimal partition size, which also needs to be altered.

Definition 7.3 *The set Γ of* **Permissible Partition Sizes** *is a set of partition sizes (r, c) that may be selected by the selection rule. It is further supposed that $\#\Gamma$ is finite, all r and c are finite, and that it contains its minimal partition size.*

Definition 7.4 *The* **Minimal Partition Size** *(r_m, c_m) of a set $\Gamma = \{(r_1, c_1), \ldots (r_t, s_t) \in \mathbb{N}^2, t \geq 1\}$ is defined as follows. Let Γ_r and Γ_c denote the sets of values r and c in Γ, respectively. Then, $r_m = \min \Gamma_r$ and $c_m = \min \Gamma_c$. It is also the the smallest partition size that may be selected by the selection rule.*

Definition 7.5 *The* **SSP Selection Rule**, *which selects the right partition size $R_n \times C_n$, is given by*

$$(R_n, C_n) = \mathit{ArgMax}_{(r,c)\in\Gamma} \left[T_{r,c,n} - 2(r-1)(c-1)\ln a_n\right],$$

where $2(r-1)(c-1)\ln a_n$ is called the **penalty**.

7.2.2 The Test Statistic and its Asymptotic Null Distribution

The test statistic of the data-driven SSPrc test is given by

$$T_{R_n,C_n,n}.$$

Thus the test statistic is still a rank statistic.

Theorem 7.3 *Let Γ be a set of permissible partition sizes and let $(r_m, c_m) \in \Gamma$ denote the minimal partition size. Suppose that $a_n \to \infty$ as $n \to \infty$. Then, under H_0,*

$$(R_n, C_n) \xrightarrow{p} (r_m, c_m),$$

as $n \to \infty$.

Proof. The proof is similar to the proof of Theorem 5.6. We only indicate the differences.

The proof is started with the equality

$$\mathrm{P}\left[(R_n, C_n) \neq (r_m, c_m)\right] = \sum_{(r,c)\in\Gamma\setminus\{(r_m,c_m)\}} \mathrm{P}\left[(R_n, C_n) = (r, c)\right].$$

In the remainder of the proof, only the different number of degrees of freedom of the χ^2 variates must be considered. Thus $(c-1)$ and (c_m-1) in the proof of Theorem 5.6 must be changed to $(r-1)(c-1)$ and $(r_m-1)(c_m-1)$, respectively, and d is now defined as $d = (r-1)(c-1) - (r_m-1)(c_m-1)$. When Chebychev's inequality is used, the mean $\mu_{rc,n} = \mathrm{E}_g\left[T_{r,c,n}\right]$ and the variance $\nu_{r,c,n} = \mathrm{Var}_g\left[T_{r,c,n}\right]$ are needed. From Corollary 7.2 it is known that both remain finite as $n \to \infty$, as needed to complete the proof. □

Proposition 7.2 *Let Γ be a set of permissible partition sizes and let $(r_m, c_m) \in \Gamma$ denote the minimal partition size. Suppose that $a_n \to \infty$ as $n \to \infty$. Under H_0,*

$$T_{R_n,C_n,n} \xrightarrow{d} \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{1}{j(j+1)k(k+1)} X_{jk}^2,$$

where the X_{jk}^2 are i.i.d. $\chi^2_{(r_m-1)(c_m-1)}$ variates.

Proof. The proof is exactly the same as the proof of Proposition 5.3, except that now Theorem 7.3 and Proposition 7.1 are needed instead of Theorem 5.6 and Proposition 5.1, respectively. □

7.2.3 The Data-Driven SSP Test

The data-driven version of the SSPrc test will be referred to as the SSPdd test.

The α-level SSPdd test $\phi_{\alpha,\Gamma,n}(\mathcal{S}_n)$ is defined as

$$\begin{aligned} \phi_{\alpha,\Gamma,n}(\mathcal{S}_n) &= 1 \text{ if } T_{R_n,C_n,n} > t_{\alpha,R_n,C_n,n} \\ \phi_{\alpha,\Gamma,n}(\mathcal{S}_n) &= 0 \text{ if } T_{R_n,C_n,n} \leq t_{\alpha,R_n,C_n,n}, \end{aligned}$$

where (R_n, C_n) is the partition size selected by the selection rule.

Theorem 7.4 *The SSPdd test based on the set Γ of permissible partition sizes, including the minimal partition size (r_m, c_m), and based on a selection rule for which $a_n \to \infty$ as $n \to \infty$, is asymptotically similar and of size α.*

Proof. The proof is completely similar to the proof of Theorem 5.7. Only the notation must be changed to the present setting. □

The reason why the SSPdd test is only asymptotically similar and of size α is essentially the same as for the SSPrc test: for finite n the exact null distribution of the rank

statistic is discrete. The definition of quantiles t_α makes that for the test defined above $\mathrm{E}_g\left[\phi_{\alpha,\Gamma,n}(\mathcal{S}_n)\right] \leq \alpha$, i.e. the SSPdd test is conservative. By randomizing the data-driven test its size becomes equal to α for all finite n as well.

Theorem 7.5 *For any set Γ of permissible partition sizes, and any selection rule for which $a_n \to \infty$ as $n \to \infty$, the SSPdd test is consistent against essentially any alternative.*

Proof. The proof is analogue to the proof of Theorem 5.8. □

7.3 Power Characteristics

In this section the power of both the SSPrc, for several partition sizes $r \times c$, and of the SSPdd test will be compared to the power of some other tests.

7.3.1 Other Tests for the Independence Problem

When the SSP22 test for independence was constructed (Section 7.1.5) it was already mentioned that the $T_{2,2,n}$ statistic is closely related to **Hoeffding's** (H) test for independence (Hoeffding, 1948), which is omnibus consistent. The relation is most easily seen when the formulation of Hoeffding's statistic of Blum et al. (1961) is studied. Both the SSP22 and Hoeffding's statistic may be interpreted as the mean of a measure for dependence in a 2×2 contingency table, where the mean is calculated over all observation-centred 2×2 tables. The measures only differ in whether or not a weight function is incorporated. Another way of looking at relation between both tests is that the SSP22 statistic is related to Hoeffding's statistic, like an Anderson-Darling statistic is related to a von Mises (von Mises, 1947) statistic. More details on Hoeffding's test are given in Section 4.2.4.

A test which is a special case of the **Puri and Sen** (Puri and Sen, 1971) rank test for independence will be considered as well. This test is based on the Spearman rank correlation. The main reason to include this test is that it is very easy to apply in practice, i.e. the test statistic is easy to calculate and it has a simple null distribution. Moreover, this test is included in published simulation studies before (Kallenberg and Ledwina, 1999) as well as in papers reporting some ARE's (Ledwina, 1986; Puri and Sen, 1971). This test will be referred to as PS. It is consistent against the class of all continuous distributions with a monotone dependence.

Yet another well known measure for dependence is the quadrant statistic (Blomqvist, 1950), also known as the sample coefficient of medial correlation. Based on this statistic again a Puri and Sen statistic (Puri and Sen, 1971) can be constructed. Here it is however preferred to use the asymptotically equivalent **interdirection quadrant statistic** (IQS)

(Gieser and Randles, 1997), which is also a simple function of the quadrant statistic, because Gieser and Randles (1997) reported that in small samples the IQS test performed better than the Puri and Sen version.

Finally, two recent **data-driven smooth tests** of Kallenberg and Ledwina (1999) are included: **TS2** and **V**. These tests are of particular interest because they are both a smooth and a data-driven tests which is based on Schwartz BIC criterion. Details on the tests are given in Section 4.3.4. In brief, both tests are data-driven score tests within a bivariate exponentially family in which each term is related to a grade correlation between $(F_x(X))^r$ and $(F_y(Y))^s$, where r and s specify the order. The $TS2$ and the V only differ w.r.t the orders (r, s). The former test is restricted to symmetric orders, i.e. $r = s$, while the latter has no such restriction.

7.3.2 The Alternatives

The problem of choosing alternatives to consider in the simulation study is even larger than it was for the general GOF and the k-sample problems. This problem is recognized many times in the literature (e.g. Hájek and Šidák, 1967; Kallenberg and Ledwina, 1999; Lehmann, 1975). One reason is of course that for a simulation study for comparing tests for independence, bivariate distributions must be used. Thus not only two marginal distributions, but also an infinite number of possible types of dependence must be chosen. Since the SSP based tests are nonparametric in the sense that their properties do not depend on the marginal distributions F_x and F_y, it would for instance be meaningful to consider copula's (bivariate distributions with fixed uniform marginal distributions) in which several dependence structures can be studied. An extensive overview of copula's has been given by Joe (1997). Although it seems a very sensible solution, our concern is that, from a practical point of view, these alternatives do not always reflect clearly equivalent situations that occur in statistical practice. E.g. consider a uniform distributed X variable and conditionally on x a normal distributed Y variable with a conditional mean that varies linearly with x (cfr. a classical linear regression). This seems a very simple and meaningful situation, but even this example does not fit easily into a simple copula family, and when the conditional mean is e.g. a quadratic function of X, then it becomes even harder to fit it into a copula. Therefore, we will not use copula's as alternatives.

To our opinion, a more sensible way of constructing bivariate distributions for studying the power of tests for independence, is to build the distribution by factorizing the distribution as $F_{xy}(x, y) = F_x(x)F_{y|x}(y|x)$. Thus, first the marginal distribution of X must be specified, and next the conditional distribution of Y given X. In this way many dependence structures can be constructed in a very flexible manner. In particular the conditional mean and the conditional variance of Y given X may be completely specified as an arbitrary simple or complex function, whatever the form (e.g. normal, lognormal, ...) of the conditional and marginal distribution.

More specifically, in this power study the observations (X_i, Y_i) $(i = 1, \ldots, n)$ of the

alternatives will be constructed using the following regression model:

$$Y_i = \mu(X_i) + \sigma(X_i)\epsilon_i, \tag{7.8}$$

where ϵ_i ($i = 1, \ldots, n$) are i.i.d. distributed random variables with CDF F_ϵ and which are independent of the X_i. The mean and the variance of ϵ are always zero and one, respectively. Further, $\mu(.)$ is the conditional mean of Y given X, and the factor $\sigma(.)$ has in the regression model the interpretation of the conditional standard deviation of Y given X. Although the distribution F_ϵ may be chosen arbitrarily, it is restricted in the present study to the standard normal distribution.

With the regression model of Equation 7.8 there are still an infinite number of alternatives possible. We made a selection of alternatives reflecting several types of dependence that posses practical relevance. Here is an overview of the alternatives.

- **Linear Regression**
  The most simple type of dependence that is studied here is a classical linear regression model with normal residuals and with homoscedastic variances. Thus,

  $$\begin{aligned} \mu(X_i) &= \beta_1 X_i \\ \sigma(X_i) &= 1. \end{aligned}$$

  β_1 values in the range between 0 and 1.5 are considered.

- **Quadratic Regression**
  Similar to the linear regression case, the conditional mean may be a quadratic function of X. The conditional variance is still kept constant. Thus,

  $$\begin{aligned} \mu(X_i) &= \beta_2 X_i^2 \\ \sigma(X_i) &= 1. \end{aligned}$$

  For β_2 the range 0-5 is studied.

- **Cubic Regression**
  The cubic regression that is simulated actually looks a non-linear monotonic regression relation. The conditional variance is still kept constant. Thus,

  $$\begin{aligned} \mu(X_i) &= \beta_3 X_i^3 \\ \sigma(X_i) &= 1. \end{aligned}$$

  For β_3 the range 0-2 is studied.

- **Sinusoidal Regression**
  As an example of a highly non-linear dependence structure a periodic function for

μ is considered.

$$\begin{aligned} \mu(X_i) &= \phi \sin(X_i) \\ \sigma(X_i) &= 1. \end{aligned}$$

- **Constant-mean Heteroscedasticity**
  Dependence is of course not only established by a non-constant conditional mean function. Another type of dependence may be constructed by letting the conditional variance of Y vary with x. In regression terms this is often called heteroscedasticity or heterogeneity. The heteroscedasticity is specified as

$$\begin{aligned} \mu(X_i) &= 0 \\ \sigma(X_i) &= 1 + \gamma \frac{R(X_i)}{n}, \end{aligned}$$

  where $R(X_i)$ is the rank of X_i among the observations in the sample $\mathcal{S}_n^x$, and where n is the sample size. In the simulation study γ is varied between 0 and 10.

- **Linear Regression with Heteroscedasticity**
  A combination of the linear regression model and the heteroscedasticity model is considered.

$$\begin{aligned} \mu(X_i) &= \beta_1 X_i \\ \sigma(X_i) &= 1 + \gamma \frac{R(X_i)}{n}, \end{aligned}$$

  where $\beta_1 = 1$. In the simulation study γ is varied between 0 and 10.

7.3.3 Results

All power estimations are based on 10000 Monte Carlo simulation runs. Powers are estimated at $\alpha = 0.05$ and for sample sizes $n = 20$ and $n = 50$. The critical values are obtained from simulated null distributions, based on 50000 simulation runs. The results are all presented as graphs.

The results for the linear, quadratic and cubic regression models are shown in Figures 7.2, 7.4 and 7.6, respectively, for $n = 20$, and Figures 7.3, 7.5 and 7.7, respectively for $n = 50$. The estimated powers of the sinusoidal regression models are shown in Figure 7.8 ($n = 20$) and Figure 7.9 ($n = 50$). The results of the variance heterogeneity models without and with the linear regression component, are presented in Figures 7.10 and 7.12, respectively, for $n = 20$, and in Figures 7.11 and 7.13, respectively for $n = 50$.

Linear Regression
For the linear regression models the simulations suggest that high powers are obtained with the PS test, which is based on Spearman's rank correlation coefficient. Since Spearman's rank correlation coefficient is known to be sensitive to monotone dependence (see

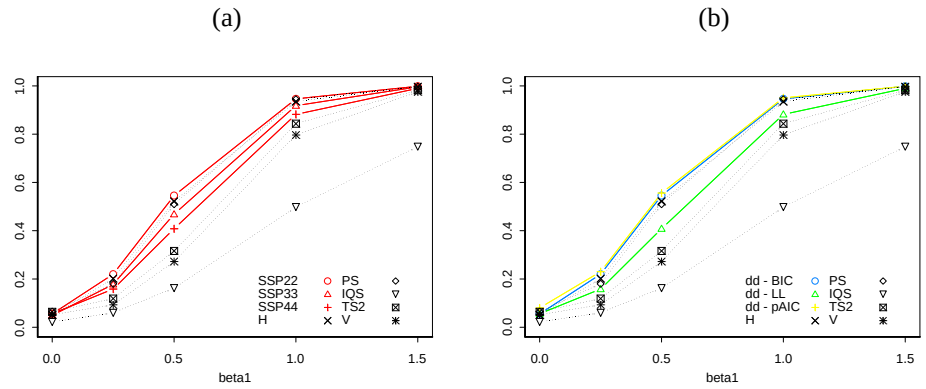


Figure 7.2: Estimated powers for the linear regression model ($n = 20$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

e.g. Joe, 1997, for an overview of its properties), of which the linear regression models form a special case. Also Hoeffding's H test is among the tests with the highest powers. Almost indistinguishable from the latter test, is the SSP22 test, of which it was mentioned already that it is closely related to the H test, except that an Anderson-Darling type weight function is incorporated. Thus, under the present simulation conditions the weight function does not seem to have a substantial influence. Among the SSPrc tests, the SSP22 test is the most powerful, followed by the SSP33 and the SSP44 test, which have lower powers (at most a difference of more or less 10% when $n = 20$). The differences between the SSPrc tests become smaller as the sample size n increases. When $n = 50$ there is virtually almost no difference between all tests but the IQS and the data-driven TS2 and V tests. The latter two tests have a slightly lower power than the SSP44 test. The lowest power is obtained with the IQS test, which is only based on the quadrant statistic of which it is indeed known that in case of a monotone dependence structure it is less sensitive than the Spearman rank correlation coefficient. When $n = 50$ the difference between the TS2, V and IQS tests decreases clearly.
All three SSPdd tests perform rather well, especially those based on the pAIC and the BIC criterion, which have powers that are virtually equal to those of the SSP22 test. The LL-based test has powers near to the powers of the SSP44 test.
The powers of all tests increase as the linear regression parameter β_1 increases.

Quadratic Regression
Under the quadratic regression model alternatives the highest powers are given by the (asymmetric) data-driven rank test, V, which is slightly better than the SSP44 test. Especially when $n = 50$ the difference between both tests become very small. After the SSP44 test, the SSP33 and the SSP22 have the highest powers among the SSPrc tests. It may also be interesting to note that the difference between the SSP44 and the SSP33 test is smaller than the difference between the SSP33 and the SSP22 test. Showing the same pattern as the SSP22 test, but with a lower power (about 10% at most in this study), is

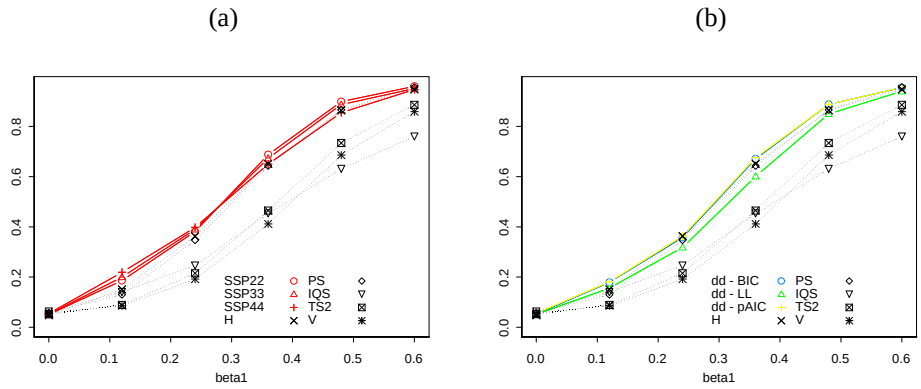


Figure 7.3: Estimated powers for the linear regression model ($n = 50$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

the H test. Thus, in this case, the incorporation of the weight function seems to have a power advantage. The powers of both the PS and the IQS test are extremely low; their powers broke down completely. When $n = 20$, the TS2 test shows the same behaviour. It is remarkable that the data-driven tests of Kallenberg and Ledwina (1999), TS2 and V, which both belong to the same family of score statistics, behave as differently as observed here. The TS2 test is even based on a choice of Legendre polynomials of order 1 up to order 10, whereas the V tests may at most contain Legendre polynomials of order 2. But, on the other hand, the V test may select Legendre polynomials of different order to form a term of the score statistic (i.e. asymmetric), whereas for TS2 test Legendre polynomials used to form a term must be of equal order. In the present situation, this clearly shows that a quadratic regression model alternative is clearly a so-called nonsymmetric alternative (terminology used by Kallenberg and Ledwina, 1999). Note that in their paper, they did not discuss such an extreme nonsymmetric alternative for which the lowest order symmetric grade correlation, i.e. the grade correlation Cov $[F_x(X), F_y(Y)]$, is as low as it is here. It is also this small grade correlation which makes that the power of the PS and IQS test are extremely low.

Among the data-driven SSP tests, again the pAIC penalty corresponds to the best SSPdd test; its power is even higher than that of the SSP44 test, but still a bit lower that the power of the V test. Only a small power difference exists with the LL-based test; when $n = 50$ this difference becomes even negligible. The power of the SSPdd - BIC test is similar to the power of the SSP22 test.

All powers, except those that broke down completely (PS and IQS), increase as the quadratic regression coefficient β_2 increases.

Cubic Regression

Before discussing the results of the cubic regression model, it is worth mentioning that the cubic model basically still results in a monotone dependence between X and Y, but for values of X in the neighbourhood of zero this dependence is virtually absent.

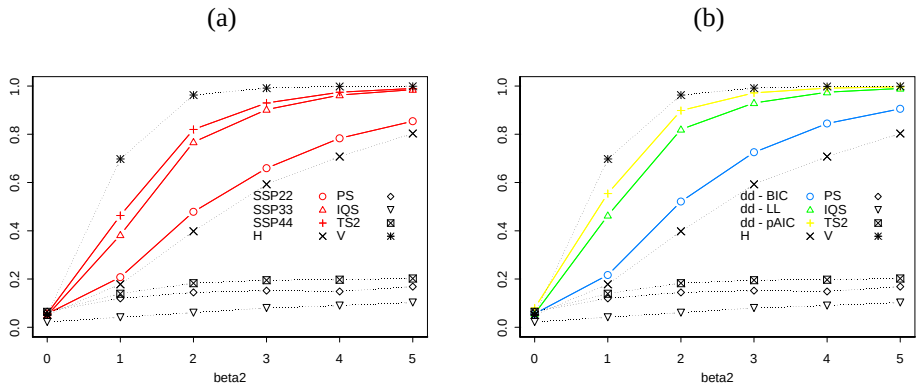


Figure 7.4: Estimated powers for the quadratic regression model ($n = 20$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

In general the same conclusions as for the linear regression model alternatives may be stated, except that now, especially when $n = 50$, the SSPrc and the SSPdd tests clearly have higher powers than the other tests.

Sinusoidal Regression

As with the quadratic regression model, the sinusoidal regression model alternatives have a very low grade correlation $\mathrm{Cov}\,[F_x(X), F_y(Y)]$, resulting in low powers for the PS and IQS tests. When $n = 50$, however, the IQS tests has a power of about 20%, but when $\phi > 2$ it does not seem to increase anymore with ϕ. Also, as with the quadratic regression model alternative, the TS2 test shows a complete power breakdown. And, more surprisingly at first sight, even the power of the V test is extremely low. This result may be explained by the inappropriateness of the low-order Legendre polynomials to capture a dependence structure with a kind of periodicity.

When $n = 20$, also the SSP22 test has very low power. The H test shows again the same pattern as the SSP22 test, both with $n = 20$ and $n = 50$, but now with slightly higher powers as the SSP22 test. With both sample sizes considered, the SSP44 test has definitely the highest power among all tests studied. The behaviour of the SSP33 test is in between the other two SSPrc tests.

Both the pAIC and the LL based SSPdd tests have good power characteristics, but not as good as the SSP44 test. When the BIC penalty is used, the SSPdd test has similar powers as the SSP22 test; when $n = 20$ its power is slightly higher than the power of the SSP22 test.

In general, the powers increase as ϕ increases.

Constant-mean Heteroscedasticity

For the variance heterogeneity models with constant mean, the TS2, PS and the IQS test have almost no power, not increasing with γ. When $n = 20$, also the SSP22 and the H test have very low power. When $n = 50$, their powers are still very poor, by now they show at least a small increase with γ. The power of the H test is generally a bit higher

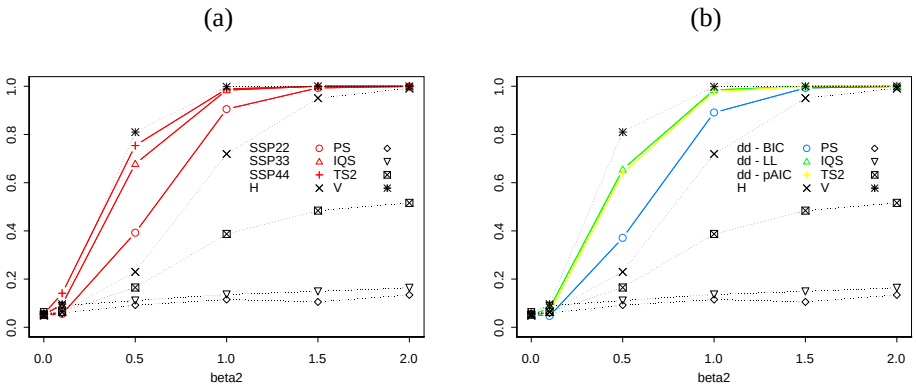


Figure 7.5: Estimated powers for the quadratic regression model ($n = 50$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

than the power of the SSP22 test. The best power is definitely obtained with the V test. Both when $n = 20$ and $n = 50$, the power of the SSP44 test is only a bit smaller (at most about 2% and 10% for $n = 20$ and $n = 50$, respectively). The power of the SSP33 test is somewhere in between those of the SSP44 and the SSP22 test.
The power of the data-driven SSP tests is rather good when the pAIC and the LL penalties are used; the former results in better powers than the latter when $n = 20$, and the other way around when $n = 50$. Their powers are in between those of the SSP44 and the SSP33 tests. When the BIC penalty is used, powers comparable as those of the SSP22 test are obtained.
Except for those tests of which the power breaks down completely, the power increases with increasing γ.

Linear Regression with Heteroscedasticity

Finally the results of the linear regression model with variance heterogeneity is discussed. Note that when $\gamma = 0$, the powers are the same as for the variance homogeneity linear regression model alternatives with $\beta_1 = 1$ and $\beta_1 = 0.5$ for $n = 20$ and $n = 50$, respectively. When $n = 20$, in general all tests show the same behaviour: decreasing power as γ increases. Moreover, the powers show the same ordering as for the variance homogeneity linear regression model. Only the V test shows a deviating behaviour: its power starts increasing again when γ becomes larger than 4. When $n = 50$, the power increase even starts as soon as γ exceeds 1. For this larger sample size, also the powers of the SSP44, SSP33, SSPdd-pAIC and the SSPdd-LL tests show this pattern, but under the conditions used in this simulation study they never exceeded the power of the V test for large γ. When $n = 50$, the powers of the SSP22 and the SSPdd-BIC tests start levelling off for larger values of γ, so that there is at least no complete breakdown of their powers.

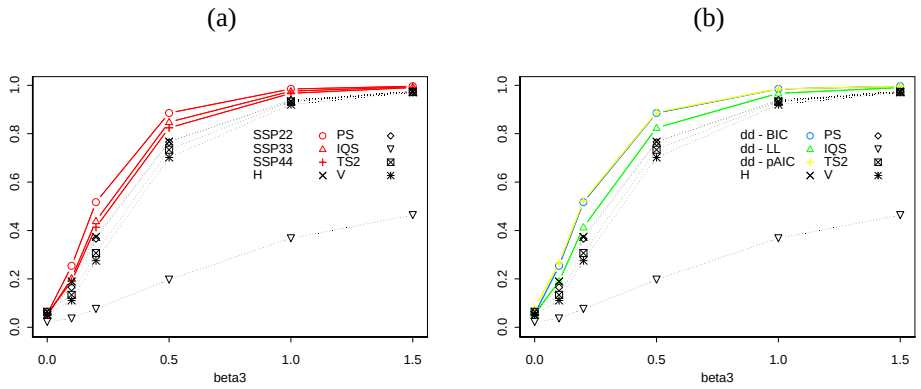


Figure 7.6: Estimated powers for the cubic regression model ($n = 20$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

7.3.4 Conclusions

Under all alternatives considered in the simulation study, none of the SSPrc and SSPdd tests showed a complete power break down, except in small samples for the SSP22 and SSPdd - BIC tests under the constant mean variance heterogeneity model. In general, the SSP22 test is among the SSPrc tests the more powerful when a monotonic dependence relation is present (e.g. linear and cubic regression). The test is even almost indistinguishable from the PS test based on the Spearman's rank correlation coefficient. When the relationship between the variables is more complex (e.g. quadratic, sinusoidal, variance heterogeneity) the SSP44 test is clearly better then those based on smaller partition sizes. In some cases it is outperformed by the V test (e.g. quadratic and variance heterogeneity models), but the difference in power is never large.
Different penalties in the SSPdd statistics clearly result in different behaviour. In all simulations the BIC criterion made the data-driven SSP test look very similar to the SSP22 test, and the use of the LL criterion resulted in a test similar to the SSP44 test. The pAIC penalty, which is proposed in this work, does not seem to have the property of following one specific SSPrc test closely. In particular, it switches more often from partition size than the other two, such that it does not stick to a partition size when the corresponding SSPrc test shows a lower power. This seems to make the SSPdd-pAIC test a good choice, at least under the alternatives studied here.

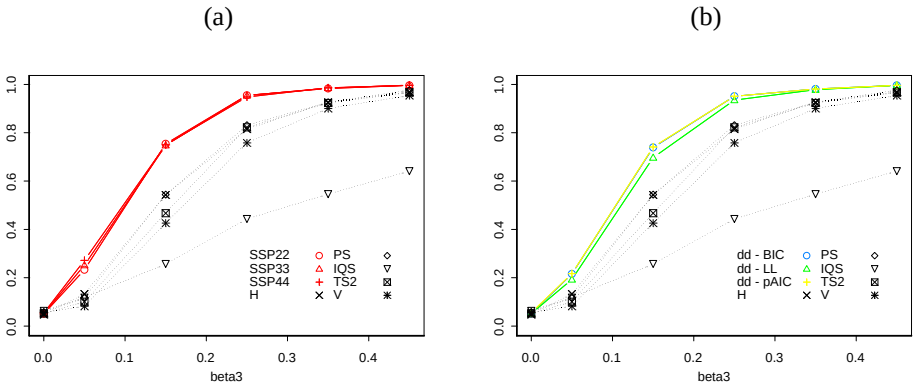


Figure 7.7: Estimated powers for the cubic regression model ($n = 50$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

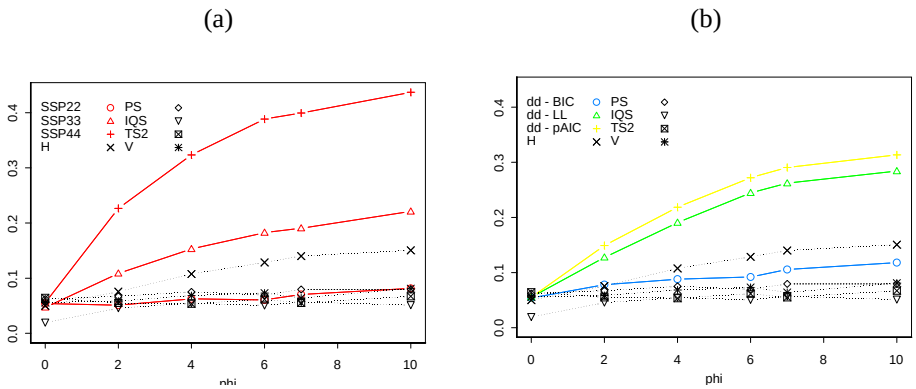


Figure 7.8: Estimated powers for the sinusoidal regression model ($n = 20$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

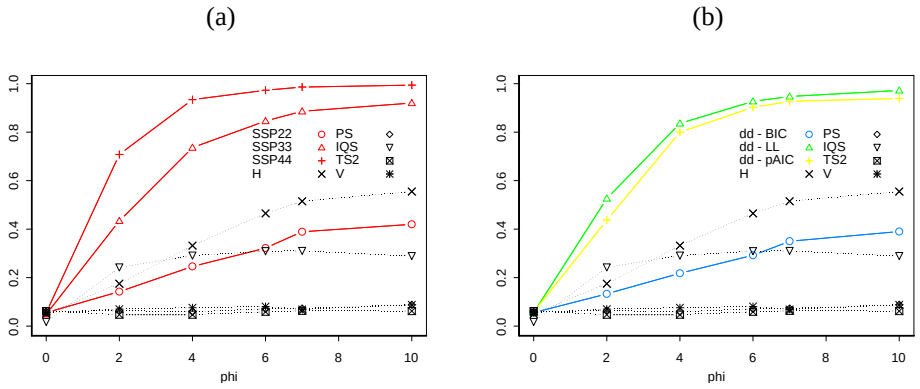


Figure 7.9: Estimated powers for the sinusoidal regression model ($n = 50$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

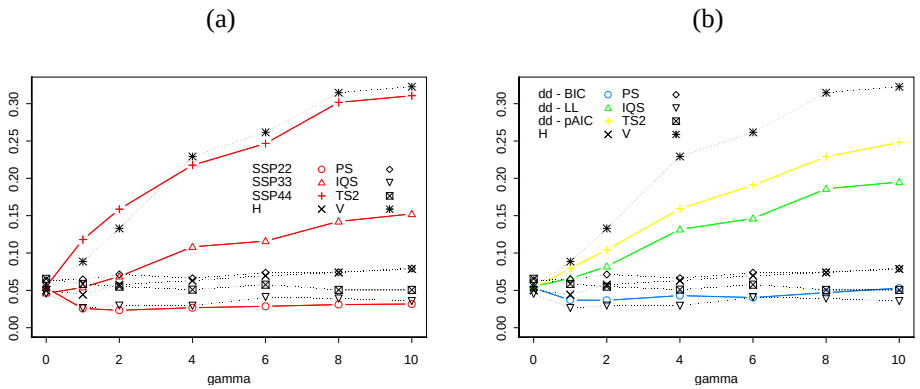


Figure 7.10: Estimated powers for the variance heterogeneity model with constant mean ($n = 20$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

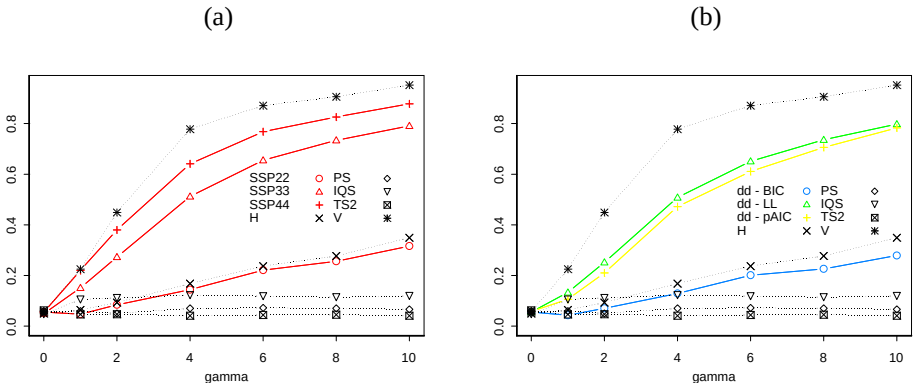


Figure 7.11: Estimated powers for the variance heterogeneity model with constant mean ($n = 50$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

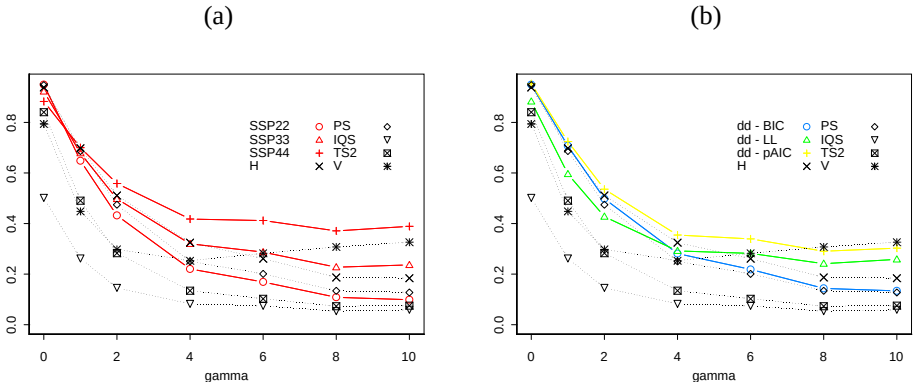


Figure 7.12: Estimated powers for the variance heterogeneity model within the linear regression model ($n = 20$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

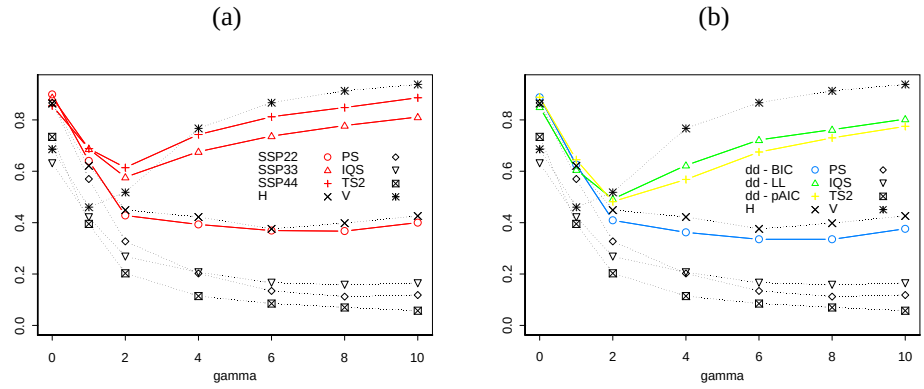


Figure 7.13: Estimated powers for the variance heterogeneity model within the linear regression model ($n = 50$). Panel (a) shows the SSP22, SSP33 and SSP44 tests together with the other tests, and panel (b) shows these other tests with the data-driven SSP tests.

7.4 Examples

For all examples the set of permissable orders is

$$\Gamma = \{(2,2),(2,3),(3,2),(3,3),(3,4),(2,4),(4,3),(4,2),(4,4)\}, \tag{7.9}$$

and all three the penalties BIC, pAIC and LL are used.

EXAMPLE 7.4. **Cultivars data**
In Section 7.1.9 it was seen that only the SSP22 test gave a significant results at the 5% level. Thus a "wrong" choice of $r \times c$ would not have revealed the dependence between the yield difference and the AZ covariate.
Both the BIC and the pAIC criterion selected the 2×2. The LL criterion, on the other hand, selected the largest partition size considered (4×4). Thus with both the former two SSPdd tests, a significant dependence is concluded, but with the latter test (LL criterion) no dependence would have been concluded.

Note that this agrees with the conclusions from the simulation study. There it was seen that for the linear regression model alternatives (from Figure 2.4 it may be seen that there is more or less a linear relation between the two variables) the SSP44 test had the lowest power among the SSPrc tests (in Section 7.1.9 the SSP44 test resulted indeed in the highest p-value) and that the LL-based data-driven test behaves similarly as the SSP44 test. □

EXAMPLE 7.5. **Ethanol data**
In Section 7.1.9 all three SSPrc tests indicated a highly significant dependence between the nitric oxide concentration and the equivalence ratio. All three penalties selected the 4×4 partition size.

It is interesting to see that even the BIC criterion selected the largest partition size. In the simulation study it was observed though, that the power of the BIC-based test was almost the same as the power of the SSP22 test, suggesting that the criterion frequently selects the 2×2 partition size. Here, on the other hand, also with the BIC penalty the 4×4 partition size is selected. This might indicate that the SSP44 statistic is extremely large relatively to the SSP22 statistic. □

7.5 Generalization of the SSPrc Test to Divergences of order λ

In Section 5.4 the one-sample SSPc statistic was generalized to a SSPc statistic of order λ, where λ refers to the order of the power divergence statistic which is the core of all SSP based statistics. Thus, a similar extension may be considered here.

First the averaged divergence of order λ and the corresponding statistic are defined. These will function as the basis for the test statistic, for which its asymptotic null distribution is proposed in the simplest case ($r = c = 2$). Finally a small simulation study is performed to compare the results of the Pearsonian with the LR-type SSPrc test.

7.5.1 Some Definitions

The directed divergence of order λ for the independence problem was defined in Section 7.1.1. By introducing the data-dependent $r \times c$ sized SSP $[B]_{\mathcal{P}} = [B](\mathcal{P}_{q,n}^{r,c})$ and taking the expectation w.r.t. the true distribution F, the averaged divergence of order λ is obtained.

Definition 7.6 *The* **Averaged Divergence of order λ** *is defined as*

$$\Delta_{r,c}^{\lambda}(F;G) = E_f\left[I^{\lambda}\left(F;G\|[B](\mathcal{P}_{q,n}^{r,c})\right)\right].$$

The usefulness of $\Delta_{r,c}^{\lambda}(F;G)$ as a measure of the deviation of F from G follows from the following result, which is an extension of Lemma 7.1, and which is proven in exactly the same way.

Lemma 7.2 *For any $\lambda \in \mathbb{R}$,*

$$\Delta_{r,c}^{\lambda}(F;G) = 0 \Leftrightarrow H_0 \textit{ is true}.$$

The Averaged Divergence of order λ is estimated by its plug-in estimator $\hat{\Delta}_c^{\lambda}(F;G)$, the Averaged Divergence Statistic of order λ,

$$\hat{\Delta}_{r,c}^{\lambda}(F;G) = \frac{1}{a_{q,n}^{r,c}} \sum_{\mathcal{P}\in\mathbb{P}_{q,n}^{r,c}} \hat{\mathrm{I}}^{\lambda}\left(F;G\|[B]_{\mathcal{P}}\right).$$

7.5.2 The SSPrc Test of order λ and its Null Distribution

The test statistic is simply $T_{r,c,n}^{\lambda} = 2n\hat{\Delta}_{r,c}^{\lambda}(F;G)$.

When 2×2 the test statistic reduces to

$$\begin{aligned} T^{\lambda}_{2,2,n} &= 2n\mathrm{E}_{\hat{F}}\left[\hat{\mathrm{I}}^{\lambda}\left(F; G \| [B](\mathcal{P}^{r,c}_{q,n})\right)\right] \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1} n \frac{1}{\lambda(\lambda+1)}\left\{\hat{F}_{xy}(X_i,Y_i)\left[\left(\frac{\hat{F}_{xy}(X_i,Y_i)}{\hat{F}_x(X_i)\hat{F}_y(Y_i)}\right)^{\lambda} - 1\right] + \right. \\ &\quad \left. (1-\hat{F}_{xy}(X_i,Y_i))\left[\left(\frac{1-\hat{F}_{xy}(X_i,Y_i)}{1-\hat{F}_x(X_i)\hat{F}_y(Y_i)}\right)^{\lambda} - 1\right]\right\}. \end{aligned}$$

$T_{2,2,n}$ is a plug-in estimator of a functional. Hence its asymptotic null distribution may be found by von-Mises calculus. However, for two reasons the calculations are more difficult than in Chapter 5. First, bivariate distributions are involved, and second, the independence hypothesis is basically a composite null hypothesis in the sense that G depends on the data by $G(x,y) = \hat{F}_x(x)\hat{F}_y(y)$. The former problem introduces only some more cumbersome algebra, whereas the latter complication invokes some additional convergences of $\hat{F}_x$ and $\hat{F}_y$, which may be handled by the methods of Billingsley (1968), Pyke and Shorack (1968). After rather long calculations it may be shown that for any λ, $T^{\lambda}_{2,2,n}$ has asymptotically the same null distribution as the V-statistic

$$V_n = n\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^1\int_0^1 \frac{\left(\delta_{z_i} - F_x(x)F_y(y)\right)\left(\delta_{z_j} - F_x(x)F_y(y)\right)}{F_x(x)(1-F_x(x))F_y(y)(1-F_y(y))} dF_x(x)dF_y(y),$$

where $\delta_z = \delta_z(x,y) = \mathrm{I}\left[z_x \leq x, z_y \leq y\right]$ where $\boldsymbol{z} = (z_x, z_y)$. V_n is not depending on λ anymore, and, moreover, it has the same kernel as the Pearsonian SSPrc statistic. Thus its limiting null distribution is also the same. This is stated in the following result, which will be referred to rather as a proposition for we did not include the whole proof here.

Proposition 7.3 *Under H_0, for each $\lambda \in \mathbb{R}$,*

$$T^{\lambda}_{2,2,n} \xrightarrow{p} T^{1}_{2,2,n},$$

and

$$T^{\lambda}_{2,2,n} \xrightarrow{d} \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\frac{1}{j(j+1)k(k+1)}Z^2_{jk},$$

where the Z^2_{jk} are i.i.d. χ^2_1 variates.

When r and c are greater than 2, Proposition 7.3 may be extended to arbitrary partition sizes in a similar way as before. Without proof we conjecture the following proposition, which will be empirically assessed in the next section.

Proposition 7.4 *Under H_0, for each $\lambda \in \mathbb{R}$,*

$$T^{\lambda}_{r,c,n} \xrightarrow{p} T^{1}_{r,c,n},$$

and

$$T^{\lambda}_{r,c,n} \xrightarrow{d} \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{1}{j(j+1)k(k+1)} Z^2_{jk},$$

where the Z^2_{jk} are i.i.d. $\chi^2_{(r-1)(c-1)}$ variates.

7.5.3 Convergence to the Limiting Distribution

The limiting null distributions of $T^0_{r,c,n}$ (i.e. $\lambda = 0$), with 2×2 and 3×3 partitions, are empirically assessed by means of Monte Carlo simulations (10000 simulation runs). The results are presented in Table 7.5. For moderate sample sizes, the convergence is too slow to use the (approximate) asymptotic critical values. The convergence is only a little bit faster as compared to the corresponding Pearsonian statistics ($\lambda = 1$).
Note that the estimated critical values of the SSP33 test seem to converge to the asymptotic values which were derived from Proposition 7.4. Thus the results confirm the proposed limiting null distribution.

Table 7.5: Estimated critical values of $T^0_{r,c,n}$ based 10000 simulation runs, for 2×2 and 3×3 partitions. Between brackets the estimated size of the test when the approximated critical value is used, is mentioned. On the line indicated with $n = \infty$, the approximated asymptotic critical value is given.

n	Partition size	
	2×2	3×3
20	2.621 (0.177)	7.361 (0.297)
50	2.181 (0.103)	6.732 (0.188)
100	2.040 (0.078)	6.491 (0.146)
500	1.825 (0.051)	6.255 (0.120)
1000	1.824 (0.052)	5.587 (0.056)
∞	1.823	5.551

7.5.4 A Small Simulation Study

It is again expected that the small sample power characteristics of the SSPrc test of order λ depend to some extent on λ. A rather extensive power study was presented in Section 7.3 for the Pearsonian SSPrc test ($\lambda = 1$). Under the same alternatives for the linear and quadratic regression models the powers of the SSPrc test of order $\lambda = 0$ (LR-type) are estimated. Sample space partitions of sizes 2×2, 3×3 and 4×4 are considered. Also the three data-driven tests are included, with Γ as in Equation 7.9. All powers are

estimated based on 10000 simulation runs. The critical values of the tests were taken from the simulated exact (permutational) null distribution (50000 runs).

Results

The results are presented in Figures 7.14 and 7.15 for the linear and quadratic regression models, respectively.

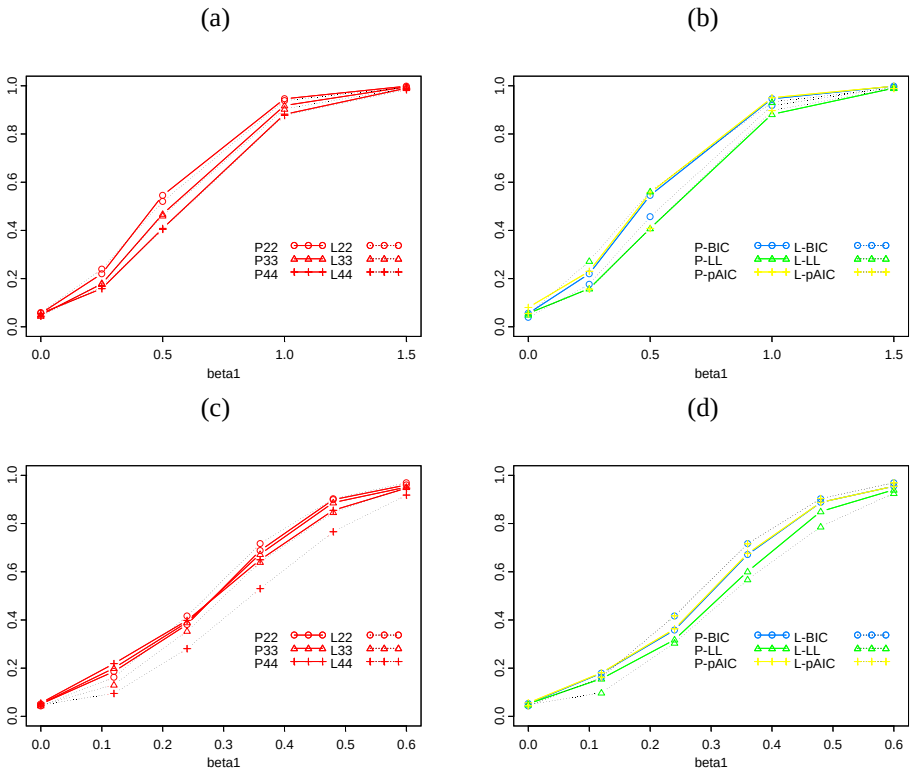


Figure 7.14: Estimated powers of the Pearsonian and the LR-type tests for the linear regression model. Panels (a) and (b) show the SSPrc and the SSPdd tests, respectively, for $n = 20$, and panels (c) and (d) for $n = 50$.

Both figures show that with the non data-driven SSPrc tests, there are not much differences observed between the Pearsonian and the LR-type tests. In particular, the ordering among the SSP22, SSP33 and SSP44 tests is the same. For the data-driven tests, on the other hand, more differences occur, especially in the ordering among the BIC, LL and pAIC based tests. This may look strange at first sight for the SSPrc tests did not show must differences at all between the P- and the LR-type. Nevertheless, the reason might be due to a different behaviour of the the asymmetric SSPrc tests ($r \neq c$), which may be selected by the SSPdd tests.

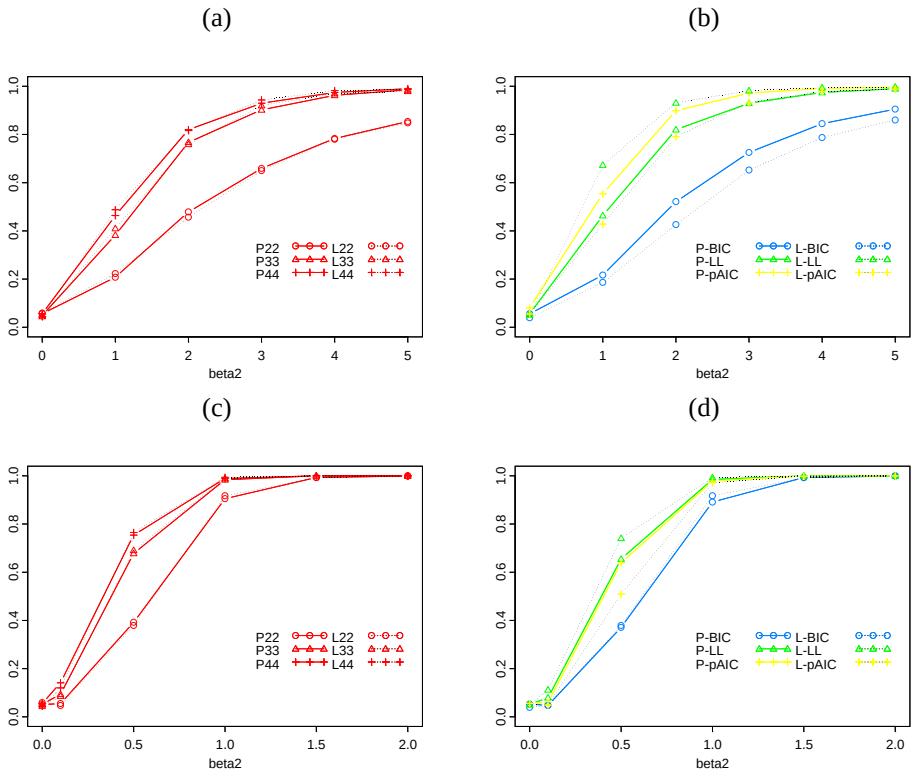


Figure 7.15: Estimated powers of the Pearsonian and the LR-type tests for the quadratic regression model. Panels (a) and (b) show the SSPrc and the SSPdd tests, respectively, for $n = 20$, and panels (c) and (d) for $n = 50$.

In general the difference between the P- and the LR-type SSP tests seems not to be as big as with the one-sample SSP tests, though possibly other values of λ may result in larger differences.

7.6 An Extension to Multivariate Independence

In this section it is indicated how the SSPrc test for bivariate independence is extended to a SSP test for joint or mutual independence between $p \geq 2$ rvs. Let $F_{1\ldots p}$ denote the joint CDF, and let F_j ($j = 1, \ldots, p$) denote the univariate CDFs of the corresponding components in $\boldsymbol{Z} = (X_1, \ldots, X_p)$. The components of an observation $\boldsymbol{Z}_i$ are denoted by $X_{1i} \ldots X_{pi}$. The sample space of the multivariate rv is $\mathcal{S} = \mathcal{S}_z$, and those of the components of $\boldsymbol{Z}$ are denoted by $\mathcal{S}_j$ ($j = 1, \ldots, p$). The null hypothesis is

$$H_0 : F_{1\ldots p}(\boldsymbol{z}) = F_1(x_1)F_2(x_2)\ldots F_p(x_p) \text{ for all } \boldsymbol{Z} \in \mathcal{S}_z.$$

By G the CDF of $\boldsymbol{Z}$ under H_0 is denoted. As the alternative hypothesis we take the complement of H_0.

First it is explained how partitions are constructed in the p dimensional sample space, then the SSP22 statistic is extended to 2^p partitions. Based on these results one could think of a further extension to arbitrary $r_1 \times r_2 \times \ldots \times r_p$ sized partitions.

7.6.1 Partitions in a p-Dimensional Sample Space

In Definition 7.1 rectangular partitions of size $r \times c$ of a 2-dimensional sample space were defined. This definition is readily extended to a $r_1 \times r_2 \times \ldots \times r_p$ partition of $\mathcal{S}_z$.

Definition 7.7 *The elements $B_i = B_{i_1 i_2 \ldots i_p}$ ($i_1 = 1, \ldots, r_1; i_2 = 1, \ldots, r_2; \ldots; i_p = 1, \ldots, r_p$) of a* **rectangular $\boldsymbol{r_1 \times r_2 \times \ldots \times r_p}$ partition** *$[B]$ of $\mathcal{S}_z$ are given by $A^1_{i_1} \times A^2_{i_2} \times \ldots \times A^p_{i_p}$, where the $A^j_{i_j}$ are the elements of an r_j sized partition $[A^j]$ of $\mathcal{S}_j$ ($j = 1, \ldots, p$; $i_1 = 1, \ldots, r_1; \ldots; i_p = 1, \ldots, p$).*

As in the bivariate case, partitions are data-dependent according to a partition construction rule. Basically this rule remains unaltered. Only some notation needs to be extended. The sets $\mathcal{P}^{r,c}_{q,n}$ and $\mathbb{P}^{r,c}_{q,n}$ become $\mathcal{P}^{r_1,\ldots,r_p}_{q,n}$ and $\mathbb{P}^{r_1 \ldots r_p}_{q,n}$, respectively, where q is the number of observations in $\mathcal{P}$ which is determined by the partition size and the partition construction rule. $a^{r_1,\ldots,r_p}_{q,n}$ denotes the cardinallity of $\mathbb{P}^{r_1,\ldots,r_p}_{q,n}$.

The smallest partition size in p dimensions is 2^p, i.e. $r_1 = r_2 = \ldots = r_p = 2$. For such partitions only one observation in $\mathcal{P}$ is needed. The corresponding SSP statistic will be represented by SSP2^p.

7.6.2 The SSP2^p Test for Independence

The test statistic is

$$T_{2^p,n} = 2n \frac{1}{n-d} \sum_{i=1}^{n-d} \hat{\mathrm{I}}^1 \left(F; G \| [A](\boldsymbol{Z}_i)\right),$$

where the sample of observations $\boldsymbol{Z}_i$ are ordered such that $\boldsymbol{Z}_{n-d+1}, \ldots, \boldsymbol{Z}_n$ are the only observations of which at least one component is the component-wise nth order statistic, and thus $d \leq p$. The statistic further reduces to

$$T_{2^p,n} = \frac{1}{n-d} \sum_{i=1}^{n-d} \frac{\left(\sqrt{n}\left(\hat{F}_z(\boldsymbol{Z}_i) - \prod_{j=1}^p \hat{F}_j(X_{ji})\right)\right)^2}{\prod_{j=1}^p \hat{F}_j(X_{ji})\left(1 - \hat{F}_j(X_{ji})\right)}.$$

Its null distribution is a straightforward extension of the bivariate case.

Theorem 7.6 *Under H_0, as $n \to \infty$,*

$$T_{2^p,n} \xrightarrow{d} \sum_{j_1=1}^{\infty} \cdots \sum_{j_p=1}^{\infty} \frac{1}{\prod_{k=1}^{p} j_k(j_k+1)} Z^2_{j_1,\ldots,j_p},$$

where the $Z_{j_1,\ldots,j_p}$ $(j_1,\ldots,j_p = 1,\ldots)$ are i.i.d. standard normal variates.

Proof. We only give a sketch of the proof for it is completely similar to the proof of Theorem 7.1.

Basically the first step is to show that the statistic $T_{2^p,n}$ converges weakly to

$$T = \int_0^1 \cdots \int_0^1 \frac{U^2(s_1,\ldots,s_p)}{\prod_{k=1}^{p} s_k(1-s_k)} ds_1 \ldots ds_p,$$

where $U(s_1,\ldots,s_p)$ is a separable Gaussian process with a p-dimensional "time" parameter $(s_1,\ldots,s_p)$. This part of the proof is a straightforward extension of the method in the proof of Theorem 7.1.

Also the second part, in which the integral equation is solved for which it is recognized that its solution is the p-fold product of solutions of the 1-dimensional Brownian bridge, is completely similar. □

CHAPTER 8

Conclusions and Outlook

8.1 Conclusions

In this work a new flexible methodology for constructing goodness-of-fit tests has been proposed and applied to three specific GOF problems. The method is based on (1) repeatedly partitioning the sample space according to a data-driven partition construction rule, subsequently (2) calculating for each partition a power divergence statistic, and, finally, (3) averaging these statistics into a new statistic, which is called the sample space partition (SSP) test statistic.

The conclusions are presented in three subsections. In the first the **construction** of the tests is briefly discussed. Next, some **links** with other tests are shown, and finally an overview of the conclusions on the **power characteristics** is given.

8.1.1 Construction of a New Family of Tests

In its most general setting a whole family of new test statistics has been proposed. This family is characterized by two parameters: the size of the sample space partitions, and the order λ of the power divergence statistic. Most of the results that are presented in this thesis are however given for $\lambda = 1$, which corresponds to the Pearson statistic. For the one-sample and the independence problem, the extensions to other values of λ have been

discussed (see later). When the random variable, on which the observations are made, is univariate, the SSP size is simply $c \geq 2$. When this random variable is bivariate, partitions of size r and c may be constructed for both components separately. The product of these two partitions results in a rectangular SSP of size $r \times c$. This occurs with the SSP test for (bivariate) independence.
The power of the SSP tests depends on λ as well as on the partition size. In particular the power may be high with a small partition size under one alternative, and low under another alternative. This characteristic has been shown many times throughout this thesis, in both the simulation studies and the examples. Thus the choice of the partition size is very important. Since the SSP test is originally designed as an omnibus test, which is in practice most often applied in situations where the researcher has no prior knowledge on the true alternative to the null hypothesis, it is most difficult to choose the "optimal" partition size. Therefore a data-driven version has been developed. Based on a selection rule, the data-driven SSP test selects a partition size out of a prespecified set of permissible partition sizes. With this approach, only this set must be chosen by the researcher. In contrast to many other data-driven (smooth) tests, there is no need here to make the set large. Indeed, it has been proven that the data-driven SSP test is consistent whatever the content of the set.

The resulting tests are distribution-free and omnibus consistency is established.

Most of the theory and results in the thesis have been obtained for univariate observations. There have been extensions to multivariate observations proposed for the one-sample and the independence problem.

8.1.2 Links with Other Tests

Many interesting links with other tests for GOF exist:

- The SSP statistic ($\lambda = 1, c = 2$) for the one-sample problem belongs to the Anderson-Darling family of statistics with a specific weight function. For SSP sizes $c > 2$, the SSP statistic is a generalization of the Anderson-Darling statistics. The AD test is asymptotically also a smooth test; a similar link of the SSP test with smooth tests is thus to be expected.

- Many smooth tests have the desirable property of being decomposable into interpretable components. Typically for smooth omnibus test statistics is that they expand into an infinite number of terms, or at least in theory this expansion is established asymptotically. With finite sample sizes, on the other hand, the user generally has to choose a finite-length sequence of components for which it is not guaranteed that it makes sense for the data at hand, for the exponential model that corresponds to the selected component may not contain the true distribution. Thus, a wrong choice may possibly result in highly insensitive statistics.
  It has been shown that the SSP statistic is also decomposable into interpretable components. In this thesis this has been worked out in detail for the k-sample problem,

where the SSP statistic based on c-sized SSPs is decomposed into $c-1$ components. Since the test is consistent, whatever the value of c, it is very likely that any departure from H_0 is captured in at least one of the components. It has been shown that the first couple of components are very closely related to the Kruskal-Wallis and the Mood statistic, and that the components may be interpreted accordingly. The usefulness has been illustrated by means of a small simulation study, and some examples.
In a limiting situation where $c = n$ we have shown that the resulting SSP statistic is exactly equivalent to a well known smooth statistic. This sheds a light on a difference between the data-driven versions of the SSP test and the smooth tests of Neyman's kind. The data-driven tests of the latter type results in a truncated series (truncation point is estimated from the data); any information that was contained in the components after the truncation point is lost completely. The data-driven SSP test results, at first sight, also in a series of which its length is determined by the data as well, but actually a decrease in partition size will mainly cause only a rearrangement of the information into a smaller number of terms.

In conclusion, the SSP statistics fill in a gap between the traditional smooth tests and the EDF-based Anderson-Darling tests.

8.1.3 Power Characteristics

For the three GOF problems simulation studies have been performed to estimate the power, under many alternatives, of the SSP tests, their data-driven versions, and some traditional tests. Typical for omnibus tests is that some tests perform better on some alternatives, and other perform better on other alternatives. We give here a brief summary of the conclusions. The conclusions are all for the tests based on the power divergence statistic of order $\lambda = 1$.

- **One-sample Problem**
  The null hypothesis is that F is a normal distribution (unknown mean and variance).
  The results have shown clearly the importance of the choice of c. Under some alternatives the SSP2 test performs the best among the SSP tests (e.g. skew alternatives), whereas under other alternatives, it performs worse (e.g. contaminated normal). It is furthermore concluded that the BIC and the LL criteria select very frequently $c = 2$ and $c = 4$, respectively, even if their choice is not the best one. Thus they resembled closely the SSP2 and the SSP4 test, respectively. Only in very extreme occasions a deviation from this behaviour has been observed. The pAIC criterion, on the other hand, seems more flexible, i.e. it switches faster to the best choice for c, but its power is often only slightly smaller than the best choice.
  Except for the contaminated normal alternatives, where the SSP3 and the SSP4 test perform extremely well, the SSP based tests are worse than the classical tests (Kolmogorov-Smirnov, Shapiro-Wilk, Anderson-Darling, ...).

- **k-sample Problem**
  In the simulation study $k = 2$ and $k = 3$ have been considered. The k-sample Kolmogorov-Smirnov and the Kruskal-Wallis tests have been included as reference. Generally the SSP4 test is the most powerful, closely followed by the SSP3 test. The SSP2 test has in general far lower powers, though still higher as compared to the two classical tests. Again the BIC and the LL criteria consistently select $c = 2$ and $c = 4$, respectively. Under most of the alternatives, the pAIC criterion selects $c = 4$ as well, generally resulting though in a slightly smaller power than the LL based test, but on the other hand, in the few situations where the SSP2 test performs better than the tests based on higher partition sizes, the pAIC criterion selects often $c = 2$, whereas the LL criterion sticks to $c = 4$. Thus, in conclusion, the data-driven test based on the LL criterion is very frequently the best choice, but its power may decrease severely under some alternatives where small partition sizes are the best. When large partition sizes are the best choice, the use of pAIC is only a bit less powerful and has in addition the advantage of retaining high powers under alternatives for which small partition sizes are best.

- **Independence Problem**
  Many alternatives have been investigated: linear, quadratic and cubic regression dependence, a sinusoidal regression dependence, and two variance heterogeneity dependencies.
  Only under monotonic relations the SSP22 test is the most powerful. When the dependence structure is more complex, the larger partition sizes result in higher powers.
  Again the BIC and the LL criterion show the tendency to select small and large partition sizes, respectively, and therefore their powers are very similar to those of the SSP22 and the SSP44 test, respectively. The pAIC criterion does not share this property. Its power is under all alternatives studies almost always only slightly smaller than the optimal SSPrc test.
  In comparison to other tests, the SSP tests are very powerful under all alternatives studied; only the TS2 and the V test of Kallenberg and Ledwina (1999) perform a bit better under a few alternatives.

In general the SSP tests are powerful omnibus tests that are very competitive to classical tests. Moreover, under many alternatives investigated in the simulation study, the SSP tests gives much better results. In addition, the power of the SSP tests does not break down often, which makes the SSP tests a good choice for typical omnibus applications. Overall we recommend the data-driven SSP test based on the pAIC criterion.

For the one-sample and the independence problem, a small simulation study has been performed to study the power of the SSP tests when $\lambda = 0$ is used. Generally it is concluded that the choice of λ has an influence on the small sample power characteristics of the SSP tests. In particular, for the one-sample problem, an overall power increase is observed. For the independence problem the power is rather invariant to the choice between both λ's.

8.2 Outlook

The methodology that has been proposed in this study looks promising for future research. We make here a distinction between further research that follows immediately from this thesis in the sense that they are actually already proposed here, but not worked out in detail. This kind of work basically consists of filling in the gaps. This will be discussed first. Next, more indirect further research is briefly commented upon.

8.2.1 Filling in the Gaps

Multivariate Observations

The three chapters on the three GOF problems handled in this work all start in the same way: first the Directed Divergence of order λ has been defined, next the corresponding statistic has been incorporated into the Averaged Pearson Divergence. In this step we have restricted λ to be 1, i.e. the Pearsonian case. Moreover in the same step the attention has been reduced to univariate variables. Only in the chapter on the one-sample and on the independence problem the multivariate extension has been discussed. From these extensions it is, though, clear that also the k-sample SSP test might be easily extendable to multivariate situations. This is a first gap that may be filled.
Further, the discussion on the multivariate SSP tests has been limited to the construction of the test statistic and its asymptotic null distribution. It would be interesting to perform simulation studies to compare the multivariate SSP tests to other, classical tests.

Arbitrary λ

The k-sample SSP test is also the only one for which no extension to arbitrary λ has been given. We believe, though, that an extension is possible in exactly the same way as has been done for the other two GOF problems.

Decomposition of the Test Statistic

The decomposition of the SSP test statistic has only been mentioned in the chapter on the SSP test for the k-sample problem. There it has been applied to decompose the test statistic into $c-1$ terms, each having an interpretation in terms of how the k distributions differ (e.g. in location, in scale, ...). It has also been briefly mentioned that a similar decomposition is possible to decompose the statistic into $k-1$ components which may serve to indicate which of the k distributions differ. This has however not been treated in detail.
The one-sample SSP test and the SSP test for independence may be decomposed as well.

The former into $c-1$ terms, and the second in $(r-1)(c-1)$ components. Further research is needed to get a more precise insight in the interpretability of the components.

Formal Proof of SSP Statistic with Arbitrary Partition Size

Finally, the validity of the asymptotic null distributions of the SSP tests based on partition sizes greater than 2 (or 2×2) are still constructed on a conjecture. Nevertheless, in simulation studies it has been confirmed that the null distributions indeed converge to the proposed asymptotic null distributions. From a theoretical point of view it would be nice to proof the conjecture. On the other hand, the simulation experiments have indicated that the convergence is very slow, even for small partition sizes. Thus, the asymptotic null distribution is not of much practical value.

8.2.2 Further Research

The SSP Test for Discontinuous Distributions

Throughout the complete work we have assumed that the distributions (F and G) are continuous, and that, consequently, no ties occur. It would definitely be interesting to extend the SSP test to situations where F or G are discontinuous, or even discrete.

An SSP Test more Closely Related to the Anderson-Darling Test

The simulation experiments show that the convergence to the asymptotic null distribution is extremely slow for the one-sample SSP test, even for the smallest partition size. The Anderson-Darling test, on the other hand, shows an extremely fast converging null distribution. We have suggested that a reason might be found in the difference in definition of both statistics. The Anderson-Darling statistic is defined as a plug-in expectation w.r.t. the hypothesized distribution G, which is completely known under a simple null hypothesis,

$$\mathrm{E}_g \left[\hat{\mathrm{I}}^1 \left(F; G \| [A](X) \right) \right],$$

where $[A](X)$ is a random 2-sized sample space partition, centred at X. The Pearsonian SSP statistic is defined as a plug-in expectation of the same statistic, but w.r.t. the true distribution F. Since the latter is unknown, it is replaced by its estimator $\hat{F}$. Hence an additional convergence is involved, and may cause the slow convergence.
The aforementioned difference in definition is a first difference. A second is the generalization to arbitrary sample space partition size c. Thus the SSP statistic is

$$\mathrm{E}_{\hat{F}} \left[\hat{\mathrm{I}}^1 \left(F; G \| [A]_{\mathcal{P}_c} \right) \right],$$

where $[A]_{\mathcal{P}_c}$ is the data-dependent sample space partition, based on $\mathcal{P}_c \subset \mathcal{S}_n$ through a partition construction rule.
It would be interesting to study the class of related statistics, defined as

$$\mathrm{E}_g \left[\hat{\mathrm{I}}^1 \left(F; G \| [A]_{\mathcal{P}_c} \right) \right].$$

It is expected that the convergence may be much faster, such that perhaps the asymptotic null distribution may be used. This would make the test easier to apply in practice. Further, its power characteristics may be substantially different from those of the SSP test constructed in this work.

A Lack-of-Fit Test

In this thesis we formulated the one-sample GOF problem merely in terms of distributions, although in Chapter 3 the relation between a distribution and a statistical model was explained. In case one wants to assess to quality of the fit of a statistical model, the term "lack-of-fit" is often preferred over GOF.
A straightforward example is simple linear regression with i.i.d. normal distributed error terms. This model is clearly equivalent to a specification of a bivariate distribution with 3 estimable parameters. Thus a SSP GOF test may be constructed for this specific problem, though by the complexity of the composite null hypothesis (H_0: the simple linear regression model suits the data) we may expect that the sampling distribution of the test statistic will not be easy to find. Still we think that it is possible to use some of the concepts of repeatedly constructing sample space partitions in the development of a lack-of-fit test, for it has a clear link with GOF problems.

APPENDIX A

The Generalized Tukey-Lambda Family

The generalized Tukey-Lambda (TL) family was studied in detail by Freimer, Mudholkar, Kollia and Lin (1988). Consider the random variables X_{λ_1,λ_2} which are defined as

$$X_{\lambda_1,\lambda_2} = F^{-1}_{\lambda_1,\lambda_2}(U) = \frac{U^{\lambda_1}-1}{\lambda_1} - \frac{(1-U)^{\lambda_2}-1}{\lambda_2}, \tag{A.1}$$

where $U \sim U(0,1)$ and $\lambda_1, \lambda_2 \in \mathbb{R}$, and which are for $\lambda_1 = 0$ and $\lambda_2 = 0$ defined by continuity. The family of distributions that is defined in this way, is the TL-family, indexed by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. The family includes a wide variety of distributions: symmetric and asymmetric, infinite and truncated tails, unimodal, U- and J-shaped distributions. Many commonly used distributions are extremely well approximated by members of this family.

Equation A.1 immediately gives the inverse CDF (quantile function) at a probability p, also indexed by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, i.e.

$$F^{-1}_{\lambda_1,\lambda_2}(p) = \frac{p^{\lambda_1}-1}{\lambda_1} - \frac{(1-p)^{\lambda_2}-1}{\lambda_2}.$$

The family (approximately) includes the following well known distributions as members.

- **Normal distribution**. The TL-distribution with $\lambda_1 = \lambda_2 = 0.1349$ is approximately a normal distribution with mean 0 and standard deviation 1.46357. When standardized, their distributions only differ by a maximum of 0.00108 at about the 27th and the 73rd percentiles.

  The member with $\lambda_1 = \lambda_2 = 5.2$ agrees rather well in 3rd and 4rd moment with a normal distribution, but it has high truncated tails.

- **Exponential distribution**. For $\lambda_1 = +\infty$ and $\lambda_2 = 0$, the TL-distribution is exactly an exponential distribution.

- **Uniform distribution**. The uniform distribution is exactly equivalent to the members with $\lambda_1 = \lambda_2 = 1$ and $\lambda_1 = \lambda_2 = 2$.

- **Logistic distribution**. By continuity the TL-family is also defined at $\lambda_1 = \lambda_2 = 0$. This member corresponds to the logistic distribution.

More generally, the members of the family are classified into 5 categories.

- **Class I**: $\lambda_1 < 1$, $\lambda_2 < 1$
  This is the richest class of distributions that agree well with distributions that occur often in practice. All members have unimodal densities and infinitely long tails. Some examples are shown in Figure A.1.

- **Class II**: $\lambda_1 > 1$, $\lambda_2 < 1$
  This class consists of distributions that are similar in shape to the exponential and χ^2 distributions. The exponential distribution is contained in this class ($\lambda_1 = +\infty$, $\lambda_2 = 0$). Some examples are shown in Figure A.2.

- **Class III**: $1 < \lambda_1 < 2$, $1 < \lambda_2 < 2$
  The members in this class have U-shaped densities and both tails truncated. Some examples are shown in Figure A.3.

- **Class IV**: $\lambda_1 > 2$, $1 < \lambda_2 < 2$
  The members in this class are characterized by bimodal distributions with one mode and one antimode, with both tails truncated. An example is shown in Figure A.4.

- **Class V**: $\lambda_1 > 2$, $\lambda_2 > 2$
  In this class the members have unimodal densities with both tails truncated. Some examples are shown in Figure A.5.

Further, members for which $\lambda_1 = \lambda_2$ are all symmetric distributions. Moving in orthogonal directions away from this diagonal, increasingly asymmetric distributions are found. However, members that are situated on a curve going through $\lambda_1 = \lambda_2 = 2.4$ and converging to the points $\boldsymbol{\lambda} = (\infty, 1)$ and $\boldsymbol{\lambda} = (1, \infty)$ have also zero skewness.
Starting from the quasi-normal member $\boldsymbol{\lambda} = (0.1349, 0.1349)$ and moving along the line of symmetry, the distributions become more and more heavy-tailed as $\lambda_1 = \lambda_2$ decreases, and the tail weights decreases as $\lambda_1 = \lambda_2$ increases. The kurtosis increases as $\lambda_1 = \lambda_2$ increases.

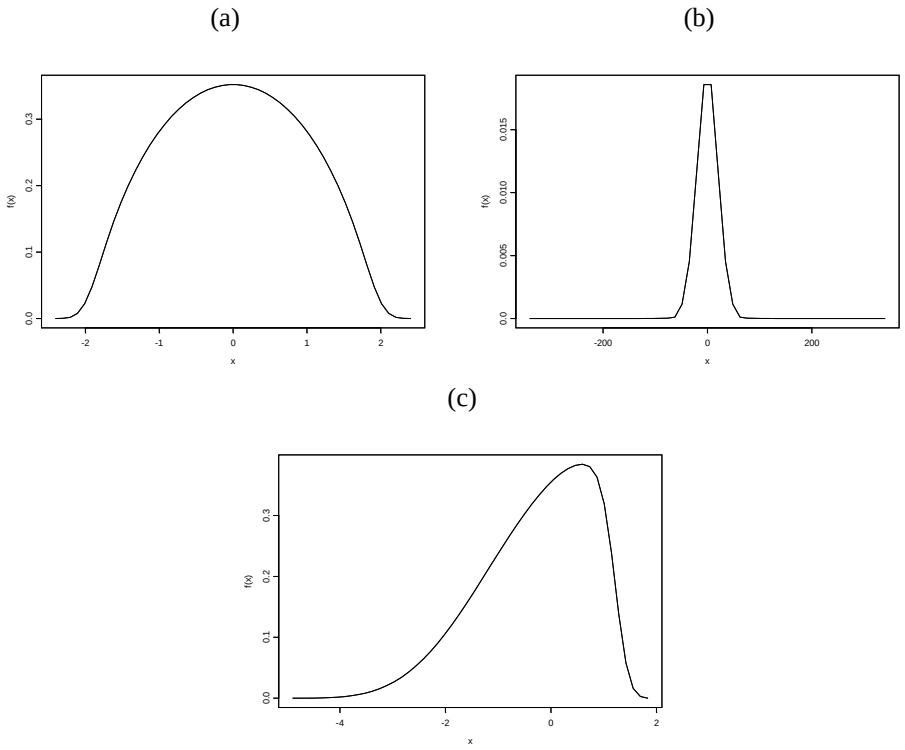


Figure A.1: Three examples of TL class I densities: (a) $\lambda_1 = \lambda_2 = 0.5$, (b) $\lambda_1 = \lambda_2 = -0.5$ and (c) $\lambda_1 = 0.2$, $\lambda_2 = 0.8$.

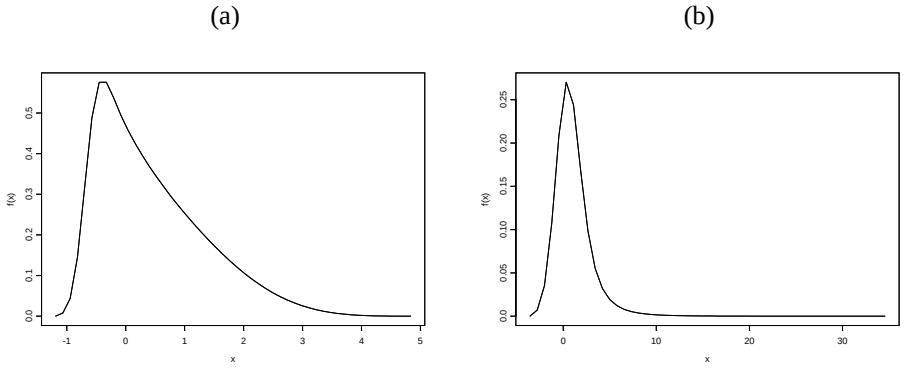


Figure A.2: Two examples of TL class II densities: (a) $\lambda_1 = 1.5$, $\lambda_2 = 0.2$ and (b) $\lambda_1 = 4$, $\lambda_2 = -0.2$.

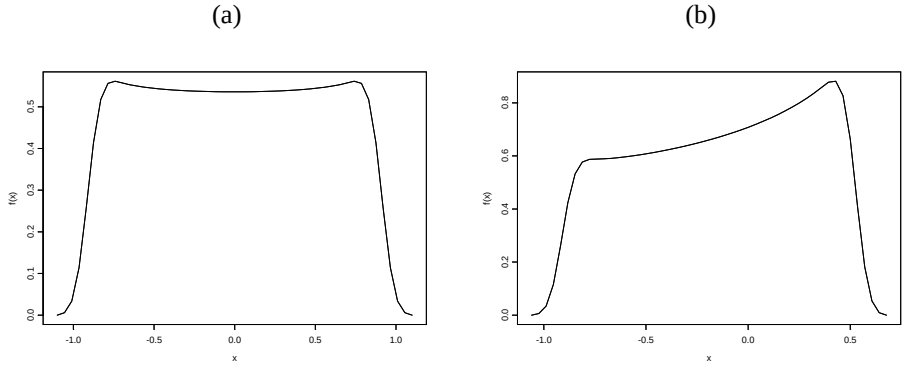


Figure A.3: Two examples of TL class III densities: (a) $\lambda_1 = 1.1$, $\lambda_2 = 1.1$ and (b) $\lambda_1 = 1.1$, $\lambda_2 = 1.9$.

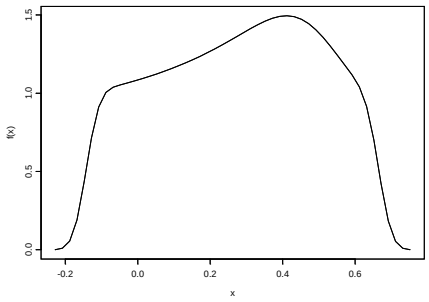


Figure A.4: An examples of a TL class IV density: $\lambda_1 = 7$, $\lambda_2 = 1.5$.

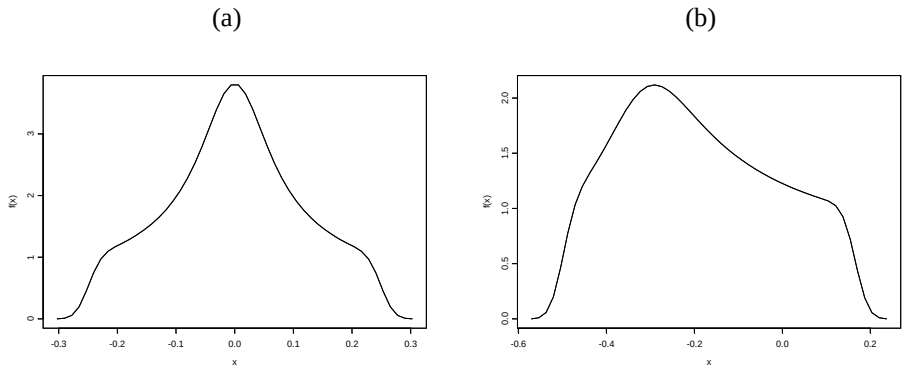


Figure A.5: Two examples of TL class V densities: (a) $\lambda_1 = \lambda_2 = 4$ and (b) $\lambda_1 = 2$, $\lambda_2 = 6$.

APPENDIX B

Computations

All computations that have been done in this thesis are performed with S-plus 2000 Professional Edition for Windows, Release 1. S-plus is an flexible object-oriented programming and analysis environment for statistical purposes. It has many build in commands for data-handling, calculation and statistical data-analysis. Despite the many build-in statistical functions, only a few of the methods that were needed, are directly available in S-plus: Pearson's χ^2 test, Kolmogorov-Smirnov test and Kruskal-Wallis test. For all other tests we wrote S-plus functions ourselves. For validation purposes, each test was simulated under its specific null hypothesis, and the estimated critical values were compared to those obtained from the literature.

The new tests that were proposed in this work are all based on repeatedly constructing sample space partitions and calculating for each partition a Pearson χ^2 statistic. Since the number of partition quickly increases with the number of observations, and since S-plus is generally inefficient with its memory management, the computation time of the SSP based tests would have been rather high when they would have been implemented directly in S-plus. To overcome this practical drawback, the computations of the SSP based test statistics is implemented in the C programming language. The code is compiled to a dll file, which is subsequently linked to S-plus. In this way, the tests are all available in S-plus, but the heavy computations are efficiently done by the linked dll code.

Bibliography

Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley & Sons.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csàki (Eds.), *Second international symposium on inference theory* (p. 267-281). Budapest: Akadémiai Kiadó.

Akaike, H. (1974). A new look at statistical model identification. *I. E. E. E. Trans. Auto. Control*, *19*, 716-723.

Allison, T. and Cicchetti, D. (1976). Sleep in mammals: Ecological and constitutional correlates. *Science*, *194*, 732-734.

Anderson, T. and Darling, D. (1952). Asymptotic theory of certain "goodness of fit"criteria based on stochastic processes. *Annals of Mathematical Statistics*, 193-212.

Anderson, T. and Darling, D. (1954). A test of goodness-of-fit. *Journal of the American Statistical Association*, *49*, 765-769.

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W. and Tukey, J. (1972). *Robust estimators of location.* Princeton, USA: Princeton University Press.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society, Series A*, *139*, 318-354.

Barton, D. (1955). A form of Neyman's ψ_k^2 test of goodness of fit applicable to grouped and discrete data. *Skandinavisk Aktuarietidskrift*, *38*, 1-16.

Barton, D. (1956). Neyman's ψ_k^2 test of goodness of fit when the null hypothesis is composite. *Skandinavisk Aktuarietidskrift*, *39*, 216-245.

Bednarski, T. and Ledwina, T. (1978). A note on biasedness of tests of fit. *Mathematische Operationsforschung und Statistik, Series Statistics*, *9*, 191-193.

Belsey, D., Kuh, E. and Welsch, R. (1980). *Regression diagnostics.* New York, USA: Wiley.

Best, D. (1995). Consumer data - statistical tests for differences in dispersion. *Food Quality and Preference*, *6*, 221-225.

Best, D. and Rayner, J. (1985). Lancaster's test of normality. *Journal of Statistical Planning and Inference*, *12*, 395-400.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics*, *1*, 1071-1095.

Billingsley, P. (1968). *Convergence of probability measures.* New York, USA: Wiley.

Birch, M. (1964). A new proof of the Pearson-Fisher theorem. *Annals of Mathematical Statistics*, *35*, 817-824.

Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA, USA: MIT Press.

Blomqvist, N. (1950). On a measure of dependence between two random variables. *Annals of Mathematical Statistics*, *21*, 593-600.

Blum, J., Kiefer, J. and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics*, *32*, 485-498.

Boos, D. and Hughes-Oliver, J. (2000). How large does n have to be for z and t intervals. *The American Statistician*, *54*, 121-128.

Box, J. (1978). *R. A. Fisher, the life of a scientist.* NY, USA: Wiley.

Chang, L. and Fisz, M. (1957). Exact distributions of the maximal values of some functions of empirical distribution functions. *Science Record, New Series*, *1*, 341-346.

Chernoff, H. and Lehmann, E. (1954). The use of maximum-likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics*, *25*, 579-586.

Cleveland, W. (1993). *Visualizing data.* Summit, NJ, USA: Hobart Press.

Cochran, W. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315-345.

Cohen, A. and Sackrowitz, H. (1975). Unbiasedness of the chi-square, likelihood ratio, and other goodness of fit tests for the equal cell case. *Annals of Statistics*, *3*, 959-964.

Cook, R. and Weisberg, S. (1982). *Residuals and influence in regression.* New York, USA: Chapman and Hall.

Cramér, H. (1928). On the composition of elementary errors. *Skandinavisk Aktuarietidskrift, 11*, 13-74, 141-180.

Cramér, H. (1954). *Mathematical methods of statistics.* Princeton, USA: Princeton University Press.

Cressie, N. (1982). Playing safe with misweighted means. *Journal of the American Statistical Association, 77*, 754-759.

Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B, 46*, 440-464.

Cressie, N. and Read, T. (1989). Pearson's X^2 and the loglikelihood ration statistic G^2: A comparitive review. *International Statistical Review, 57*, 19-43.

Csörgö, S. (1986). Testing for normality in arbirtary dimension. *Annals of Statistics, 14*, 708-723.

D'Agostino, R. (1986). Tests for the normal distribution. In R. D'Agostino and M. Stephens (Eds.), *Goodness-of-fit techniques* (p. 367-419). Marcel Dekker.

D'Agostino, R. and Pearson, E. (1973). Testing for departure from normality. I. Fuller empirical results for the distribution of b_2 and $\sqrt{b_1}$. *Biometrika, 60*, 613-622.

D'Agostino, R. and Stephens, M. (1986). *Goodness-of-fit techniques.* NY, USA: Marcel Dekker.

Dahiya, R. and Gurland, J. (1972). Pearson chi-square test of fit with random intervals. *Biometrika, 59*, 147-153.

Dahiya, R. and Gurland, J. (1973). How many classes in the pearson chi-square test? *Journal of the American Statistical Association, 68*, 707-712.

Darling, D. (1955). The Cramér-Smirnov test in the parametric case. *Annals of Mathematical Statistics, 26*, 1-20.

Darling, D. (1957). The Kolmogorov-Smirnov, Cramer-von Mises tests. *Annals of Mathematical Statistics, 28*, 823-838.

Davison, A. and Hinkley, D. (1997). *Bootstrap methods and their applications.* Cambridge, UK: Cambridge university press.

Dawid, A. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B, 41*, 1-31.

De Wet, T. (1980). Cramér-von Mises tests for independence. *Journal of Multivariate Analysis, 10*, 38-50.

De Wet, T. and Randles, R. (1984). *On limiting chi square u- and v-statistic form with auxiliary estimators* (Tech. Rep. No. 227). Department of Statistics, University of Florida.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics, 1*, 279-290.

Durbin, J. and Knott, M. (1972). Components of Cramér - von Mises statistics. *Journal of the Royal Statistical Society, Series B, 34*, 290-307.

Durbin, J., Knott, M. and Taylor, C. (1975). Components of Cramér - von Mises statistics: II. *Journal of the Royal Statistical Society, Series B, 37*, 216-237.

Dwass, M. (1960). Some k-sample rank-order tests. In *Contributions to probabilitiy and statistics, essays in honor of H. Hotelling* (p. 198-202).

Eden, T. and Yates, F. (1933). On the validity of Fisher's z test when applied to an actual example of non-normal data. *Journal of Agricultural Science, 23*, 6-16.

Edgington, E. (1995). *Randomization tests.* New York, USA: Marcel Dekker.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman and Hall.

Eubank, R., LaRiccia, V. and Rosenstein, R. (1987). Test statistics derived as components of Pearson's phi-squared distance measure. *Journal of the American Statistical Association, 82*, 816-825.

Fisher, R. (1922). On the interpretation of chi square from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society, 85*, 87-94.

Fisher, R. (1924). The conditions under which χ^2 measures the discripancy between observations and hypothesis. *Journal of the Royal Statistical Society, 87*, 442-450.

Fisher, R. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver and Boyd.

Fisher, R. (1934). *Statistical methods for research workers.* Edingburgh, UK: Oliver and Boyd.

Fisher, R. (1935). *The design of experiments.* London, UK: Oliver and Boyd.

Fisz, M. (1960). Some non-parametric tests for the k-sample problem. *Colloquium math., 7*, 289-296.

Freimer, M., Mudholkar, G., Kollia, G. and Lin, C. (1988). A study of the generalized Tukey lambda family. *Communications in Statistics. Theory and Methods, 10*, 3547-3567.

Gail, M. and Gastwirth, J. (1975). A scale free goodness-of-fit for the exponential distribution based on the gini statistic. *Journal of the Royal Statistical Society, 40*, 350-357.

Gieser, W. and Randles, R. (1997). A nonparametric test of independence between two vectors. *Journal of the American Statistical Association, 92*(438), 561-567.

Gupta, M. (1967). An asymptotically nonparametric test for symmetry. *Annals of Mathematical Statistics*, *38*, 849-866.

Hájek, J. and Šidák, Z. (1967). *Theory of rank tests.* New York, USA: Academic Press.

Hall, P. (1987). On kullback-leibler loss and density estimation. *Annals of Statistics*, *15*, 1491-1519.

Hamdan, M. (1962). The powers of certain smooth tests of goodness of fit. *Australian Journal of Statistics*, *4*, 25-40.

Hamdan, M. (1964). A smooth test of goodness of fit based on the Walsh function. *Asustralian Journal of Statistics*, *6*, 130-136.

Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, *41*, 190-195.

Haughton, D., Haughton, J. and Izenman, A. (1990). Information criteria and harmonic models in time series analysis. *Journal of Statistical Computation and Simulation*, *35*, 187-207.

Heiberger, R. and Becker, R. (1992). Design of an S function for robust regression using interatively reweighted least squares. *Journal of Computational and Graphical Statistics*, *1*, 181-196.

Hoeffding, W. (1948). A non-parametric test of independence. *Annals of Mathematical Statistics*, *19*, 546-557.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, *23*, 169-192.

Horn, S. (1977). Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics*, *33*, 237-248.

Hsu, J. (1996). *Multiple comparisons, theory and methods.* London, UK: Chapman and Hall.

Inglot, T., Kallenberg, W. and Ledwina, T. (1997). Data driven smooth tests for composite hypotheses. *Annals of Statistics*, *25*, 1222-1250.

Janic-Wróblewska, A. and Ledwina, T. (2000). Data driven rank test for two-sample problem. *Scandinavian Journal of Statistics*, *27*, 281-297.

Joe, H. (1997). *Multivariate models and dependence concepts.* London: Chapman and Hall.

Kac, M., Kiefer, J. and Wolfowitz, J. (1955). On tests of normality and other tests of goodness-of-fit based on distance methods. *Annals of Mathematical Statistics*, *26*, 189-211.

Kallenberg, W. and Ledwina, T. (1995). On data-driven Neyman's tests. *Probability and Mathematical Statistics, 15*, 409-426.

Kallenberg, W. and Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association, 92*, 1094-1104.

Kallenberg, W. and Ledwina, T. (1999). Data-driven rank tests for independence. *Journal of the American Statistical Association, 94*, 285-301.

Kallenberg, W., Ledwina, T. and Rafajlowicz, E. (1997). Testing bivariate independence and normality. *Sankhya, Series A, 59*, 42-59.

Kallenberg, W., Oosterhoff, J. and Schriever, B. (1985). The number of classes in chi-squared goodness-of-fit tests. *Journal of the American Statistical Association, 80*, 959-968.

Karpenstein-Machan, M. and Maschka, R. (1996). Investigations on yield structure and local adaptibility. *Agrobiological Research, 49*, 130-143.

Karpenstein-Machen, M., Honermeier, B. and Hartmann, F. (1994). *Produktion aktuell, triticale.* Frankfurt, Germany: DLG Verlag.

Kempthorne, O. (1967). The classical problem of inference - goodness of fit. In L. Le Cam and J. Neyman (Eds.), *Proceedings of the 5th berkeley symposium on mathematical statistics and probability* (p. 235-249). Berkeley, CA, USA: University of California Press.

Kendall, M. and Stuart, A. (1973). *The advanced theory of statistics, vol.* New York, USA: Hafner publishing company.

Kiefer, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises tests. *Annals of Mathematical Statistics, 30*, 420-447.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Gior. Ist. Ital. Attuari, 4*, 83-91.

Kolmogorov, A. (1941). Confidence limits for an unknown distribution function. *Annals of Mathematical Statistics, 12*, 461-463.

Kopecky, K. and Pierce, D. (1979). Efficiency of smooth goodness-of-fit tests. *Journal of the American Statistical Association, 74*, 393-397.

Koziol, J. (1979). A smooth test for bivariate independence. *Sankhya, Series B, 41*, 260-269.

Koziol, J. (1986). Assessing multivariate normality: A compendium. *Communications in statistics, theory and methods, 15*, 2763-2783.

Kruskal, W. (1952). A nonparametric test for the several sample problem. *Annals of Mathematical Statistics, 23*, 525-540.

Kruskal, W. and Wallis, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583-621.

Kullback, S. (1959). *Information theory and statistiscs.* New York, USA: Wiley.

Lancaster, H. (1949). The derivation and partition of χ^2 in certain discrete distributions. *Biometrika*, *36*, 117-129.

Lancaster, H. (1953). A reconciliation of χ^2, considered from metrical and enumerative aspects. *Sankhya*, *13*, 1-10.

Lancaster, H. (1969). *The chi-squared distribution.* New York: Wiley.

Lauritzen, S. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, *17*, 31-57.

Ledwina, T. (1986). Large deviations and Bahadur slopes of some rank tests of independence. *Synkhya, series A*, *48*, 188-207.

Ledwina, T. (1994). Data-driven version of neyman's smooth test of fit. *Journal of the American Statistical Association*, *89*, 1000-1005.

Lee, A. (1990). *U-statistics.* New York, USA: Marcel Dekker.

Lehmann, E. (1975). *Nonparametrics. Statistical methods based on ranks.* Oakland, CA, USA: Holden-Day.

Lehmann, E. (1999). *Elements of large-sample theory.* New York: Springer-Verlag.

Lewis, P. (1961). Distribution of the Anderson-Darling statistic. *Annals of Mathematical Statistics*, *32*, 1118-1124.

Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*, 399-402.

Lindsay, B. and Basak, P. (2000). Moments determine the tail of a distribution (but not much else). *The American Statistician*, *54*, 248-251.

Lindsey, J. (1996). *Parametric statistical inference.* Oxford: Oxford University Press.

Manly, B. (1997). *Randomization, bootstrap and Monte Carlo methods in biology.* London, UK: Chapman and Hall.

Mann, H. and Wald, A. (1942). On the choice of the number of class intervals in the application of the chi-square test. *Annals of Mathematical Statistics*, *13*, 306-317.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50-60.

Mardia, K. (1980). Tests of univariate and multivariate normality. In P. Krishnaiah (Ed.), *Handbook of statistics, vol.1* (p. 279-320). North Holland.

Massey, F. (1951). The distribution of the maximum deviation between two sample cumulative step functions. *Annals of Mathematical Statistics*, *22*, 125-128.

Massey, F. (1952). Distribution table for the deviation between two sample cumulatives. *Annals of Mathematical Statistics*, *23*, 435-441.

McCabe, B. and Tremayne, A. (1993). *Elements of modern stochastic theory with statistical applications.* Manchester, UK: Manchester University Press.

Moore, D. (1977). Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association*, *72*, 131-137.

Moore, D. (1986). Tests of chi-squared type. In R. D'Agostino and M. Stephens (Eds.), *Goodness-of-fit techniques* (p. 63-95). Marcel Dekker.

Moore, D. and Spruill, M. (1975). Unified large-sample theory fo general chi-squared statistics for tests of fit. *Annals of Statistics*, *3*, 599-616.

Mudholkar, G., Kollia, G., Lin, C. and Patel, K. (1991). A graphical procedure for comparing goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, *53*, 221-232.

Neuhaus, G. (1976). Weak convergence under contiguous alternatives of the empirical process when parameters are estimated. In P. Gaenssler and P. Révész (Eds.), *Empirical distributions and processes* (p. 68-82). Springer-Verlag.

Neuhaus, G. (1987). Local asymptotics for linear rank statistics with estimated score functions. *Annals of Statistics*, *15*, 491-512.

Neyman, J. (1937). "Smooth"test for goodness of fit. *Skand. Aktuartidskr.*, *20*, 149-199.

Oosterhoff, J. (1985). The choice of cells in chi-square tests. *Statistica Neerlandica*, *39*, 115-128.

Owen, D. (1962). *Handbook of statistical tables.* Reading, MA, USA: Addison-Wesley.

Pearson, E. and Hartley, H. (1972). *Biometrika tables for statisticians, vol. 2.* Cambridge, UK: Cambridge University Press.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157-175.

Pearson, K. (1922). On the χ^2 test of goodness of fit. *Biometrika*, *14*, 186-191.

Pettit, A. (1976). A two-sample Anderson-Darling rank statistic. *Biometrika*, *63*, 161-168.

Pettit, A. (1977). Testing the normality of several independent samples using the anderson-darling statistic. *Journal of the Royal Statistical Society, Series C*, *26*, 156-161.

Piepho, H. (2000). Exact confidence limits for covariate-dependent risk in cultivar trials. *Journal of Agricultural, Biological and Environmental Statistics*, *5*, 202-213.

Pitman, E. (1937a). Significance tests that may be applied to samples from any population. *Journal of the Royal Statistical Society Supplement*, *4*, 119-130.

Pitman, E. (1937b). Significance tests that may be applied to samples from any population III the analysis of variance test. *Biometrika*, *29*, 322-335.

Pitman, E. (1937c). Significance tests that may be applied to samples from any population II the correlation coefficient test. *Journal of the Royal Statistical Society Supplement*, *4*, 225-232.

Politis, D., Romano, J. and Wolf, M. (1999). *Subsampling*. NY, USA: Springer-Verlag.

Pollard, D. (1979). General chi-square goodness-of-fit tests with data-dependent cells. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, *50*, 317-331.

Puri, M. and Sen, P. (1971). *Nonparametric methods in multivariate analysis.* New York: John Wiley & Sons.

Pyke, R. and Shorack, G. (1968). Weak convergences of a two-sample empirical process and a new approach to Chernoff-Savage theorems. *Annals of Mathematical Statistics*, *39*, 755-771.

Quine, M. and Robinson, J. (1985). Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. *Annals of Statistics*, *13*, 727-742.

Radlow, R. and Alf, E. (1975). An alternate multinomial assessment of the accuracy of the χ^2 test of goodness of fit. *Journal of the American Statistical Association*, *70*, 811-813.

Rao, C. (1973). *Linear statistical inference and its applications.* NY, USA: Wiley.

Rao, K. and Robson, D. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. *Communications in Statistics*, *3*, 1139-1153.

Rathie, P. and Kannappan, P. (1972). A directed-divergence function of type β. *Information and Control*, *20*, 38-45.

Rayner, J. and Best, D. (1989). *Smooth tests of goodness-of-fit.* New York, USA: Oxford University Press.

Read, T. (1984). Small sample comparisons for the power divergence goodness-of-fit statistics. *Journal of the American Statistical Association*, *79*, 929-935.

Read, T. and Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data.* NY, USA: Springer-Verlag.

Risebrough, R. (1972). Effects of environmental pollutants upon animal other than man. In *Proceedings of the 6th Berkeley symposium on mathematics and statistics* (p. 443-463). Derkeley, CA, USA: University of California University Press.

Romano, J. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, *83*, 698-708.

Romano, J. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics*, *17*, 141-159.

Rosenblatt, M. (1952). Limit theorems associated with variants of the von-Mises statistic. *Annals of Mathematical Statistics*, *23*, 617-623.

Rosenblatt, M. and Wahlen, B. (1992). A nonparametric measure of independence under a hypothesis of independent components. *Statistics and Probability Letters*, *15*, 245-252.

Rosenkrantz, W. (2000). Confidence bands for quantile functions: A parametric and graphic alternative for testing goodness of fit. *The American Statistician*, *54*, 185-190.

Roy, A. (1956). *On χ^2 statistics with variable intervals* (Tech. Rep. No. 1). Department of Statistics, Stanford University.

Roy, S. and Mitra, S. (1956). An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis. *Biometrika*, *43*, 361-376.

Scholz, F. and Stephens, M. (1987). k-sample Anderson-Darling tests. *Journal of the American Statistical Association*, *82*, 918-924.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Serfling, R. (1980). *Approximation theorems of statistics.* New York, USA: Wiley.

Shapiro, S. and Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591-611.

Shorack, G. (2000). *Probability for statisticians.* NY, USA: Springer-Verlag.

Shorack, G. and Wellner, J. (1986). *Empirical processes with applications to statistics.* New York, USA: Wiley.

Silverman, B. (1986). *Density estimation for statistics and data analysis.* London, UK: Chapman and Hall.

Simonoff, J. (1996). *Smoothing methods in statistics.* New York, USA: Springer.

Smirnov, N. (1939). Sur les ecarts de la courbe de distribution empirique (in Russian). *Rec. Math.*, *6*, 3-26.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, *19*, 279-281.

Solomon, H. and Stephens, M. (1978). Approximations to density functions using Pearson curves. *Journal of the American Statistical Association*, *73*, 153-160.

Stephens, M. (1970). Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society, Series B*, *32*, 115-122.

Stephens, M. (1971). *Asymptotic results for goodness-of-fit statistics when parameters must be estimated* (Tech. Rep. No. 159 and 180). Department of Statistics, Stanford University.

Stephens, M. (1974). EDF statistics for goodness-of-fit and some comparisons. *Journal of the American Statistical Association*, *69*, 730-737.

Stephens, M. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, *4*, 357-369.

Stephens, M. (1978). On the half-sample method for goodness-of-fit. *Journal of the Royal Statistical Society, Series B*, *40*, 64-70.

Stephens, M. (1986). Tests based on EDF statistics. In R. D'Agostino and M. Stephens (Eds.), *Goodness-of-fit techniques* (p. 97-193). New-York, USA: Marcel Dekkers.

Sukhatme, S. (1972). Fredholm determinant of a positive definite kernel of a special type and its application. *Annals of Mathematical Statistics*, *43*, 1914-1926.

Tate, M. and Hyer, L. (1973). Inaccuracy of the X^2 test of goodness of fit when expected frequencies are small. *Journal of the American Statistical Association*, *68*, 836-841.

Thomas, D. and Pierce, D. (1979). Neyman's smooth goodness-of-fit test when the null hypothesis is composite. *Journal of the American Statistical Association*, *74*, 441-445.

van der Vaart, A. (1998). *Asymptotic statistics.* Cambridge, UK: Cambridge University Press.

Venables, W. and Ripley, B. (1997). *Modern applied statistics with S-plus.* Springer.

von Mises, R. (1931). *Wahrscheinlichkeitsrechnung.* Vienna, Austria: Deuticke.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, *18*, 309-348.

Watson, G. (1957). The χ^2 goodness-of-fit test for normal distributions. *Biometrika*, *44*, 336-348.

Watson, G. (1958). On chi-square goodness-of-fit tests for continuous distributions. *Journal of the Royal Statistical Society, Series B, 20*, 44-61.

Watson, G. (1959). Some recent results in chi-square goodness-of-fit tests. *Biometrics*, 440-468.

Weisberg, S. and Binham, C. (1975). An approximation analysis of variance test for non-normality suitable for machine calculation. *Technometrics, 17*, 133-134.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics.* Chichester: Wiley.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80-83.

Zheng, J. (1997). A consistent specification test of independence. *Journal of Nonparametric Statistics, 7*, 297-306.

Samenvatting

Reeds meer dan 100 jaar worden door onderzoekers statistische toetsen voor aanpassing (*Goodness-of-Fit*) toegepast. Waarschijnlijk de meest gekende zijn deze voor het toetsen of een steekproef uit een normale distributie komt. Het belang van deze vraag komt voornamelijk voort uit de ruime beschikbaarheid van statistische technieken die gebaseerd zijn op de normaliteitsveronderstelling, maar ook toetsen voor exponentialiteit en uniformiciteit kennen in bepaalde wetenschappen hun toepassing. Dergelijke vraagstellingen worden het "één-steekproef probleem" genoemd (*one-sample problem*). In het algemeen stelt de nulhypothese dat de werkelijke, maar ongekende distributie F behoort tot een vooropgestelde familie van distributies $G_\theta \in \mathcal{G}_\Theta$, geïndexeerd door de parameter $\boldsymbol{\theta}$. Tegenwoordig worden ook nog andere statistische vraagstellingen als bijzonder geval van de bovenstaande hypothese beschouwd; we vernoemen er hier twee, die eveneens een heel groot toepassingsdomein bestrijken:

- het "meerdere-steekproeven probleem" (k-*sample problem*), waarbij men wenst na te gaan of k steekproeven uit een zelfde distributie afkomstig zijn ;
- het "onafhankelijkheidsprobleem" (*independence problem*), waarbij men wenst na te gaan of de componenten van een multivariate variabele onafhankelijk zijn.

Dikwijls vertonen de statistische toetsen voor de hoger vermelde probleemstellingen vele overeenkomsten.

Niettegenstaande er reeds vele toetsen bestaan voor de drie aanpassingsproblemen, blijft er intensief onderzoek naar nieuwe en "betere" methoden plaatsvinden. Ook het voorliggend doctoraatsonderzoek kadert hierin. Een belangrijke drijfveer voor verder onderzoek wordt gevormd door het ontbreken van een optimale statistische omnibus toets voor het aanpassingsprobleem, waardoor de deur open blijft voor een zoektocht naar een toets die algemeen krachtig is t.o.v. de meest relevante alternatieven, of toch tenminste geen algeheel verlies van kracht vertoont t.o.v. realistische alternatieven.

In dit werk wordt een nieuwe methodologie voorgesteld die toepasbaar is op continu verdeelde data. De methode is gebaseerd op het herhaaldelijk construeren van partities van de steekproefruimte; iedere partitie induceert een discretisatie van de data, waarop vervolgens een Pearson χ^2 toets uitgevoerd wordt. De voorgestelde test statistiek is het gemiddelde van deze Pearson χ^2 statistieken.
De methodologie werd eerst ontwikkeld voor het één-steekproef probleem, vervolgens voor het meerdere-steekproefprobleem en ten slotte voor het onafhankelijkheidsprobleem. Deze drie ontwikkelingsstadia komen overeen met drie belangrijkste hoofdstukken in deze thesis.

De eenvoudigste situatie is het één-steekproefprobleem. Aan de hand hiervan werd de basis gelegd waar ook de andere twee aanpassingsproblemen op steunen.
Eigenlijk werd een familie van test statistieken $T_{c,n}$ geconstrueerd. De familie wordt geïndexeerd door de grootte (c) van de partities. De overeenkomstige testen worden de SSPc testen genoemd. Vooreerst werd er gezocht naar de nuldistributie van de test statistiek. Hiertoe dient er een onderscheid gemaakt te worden tussen een eenvoudige en een samengestelde nulhypothese. Wanneer $c = 2$ werd er aangetoond dat de asymptotische distributie van $T_{c,n}$ onder een eenvoudige nulhypothese dezelfde is als deze van de Anderson-Darling (Anderson and Darling, 1952) statistiek. Deze laatste staat bekend als een omnibus test die algemeen een goede kracht bezit. Ook voor $c > 2$ werd er een asymptotische nuldistributie voorgesteld. In een simulatiestudie werden beide distributies bevestigd, maar tevens werd besloten dat de convergentie te traag verloopt om de asymptotische distributie in de praktijk te kunnen gebruiken. Als alternatief werd voorgesteld om ofwel te werken met de gesimuleerde exacte nuldistributie, ofwel gebruik te maken van benaderingsformules voor de berekening van de kritische waarden. Deze laatste blijken redelijk accurate waarden te genereren.
Zoals voor vele GOF testen is asymptotische nuldistributie onder een samengestelde nulhypothese niet eenvoudig. Enkele in het eenvoudigste geval ($c = 2$) werd deze bekomen voor $T_{2,n}$. Weerom is het dezelfde als voor de Anderson-Darling statistiek, en verloopt de convergentie te traag om bruikbaar te zijn. Voor $c > 2$ werd de asymptotische nuldistributie niet gevonden, maar dit belet uiteraard niet de test te gebruiken a.d.h.v. een Monte Carlo benadering van de exacte nuldistributie. Voor alle SSPc testen werd de omnibus consistentie bewezen.
Er werd een simulatiestudie uitgevoerd om de kracht van de SSPc testen (met $c = 2, 3$ en $c = 4$) te vergelijken met de kracht van andere veel gebruikte klassieke GOF testen (o.a. Kolmogorov-Smirnov, Anderson-Darling. Shapiro-Wilk, en een meer recente data-gedreven geleidelijke test (*smooth test*) van Kallenberg and Ledwina (1997)). De distributie die onder de nulhypothese gespecificeerd werd, was de normale distributie met ongekend gemiddelde en variantie. Uit deze uitgebreide studie werd geconcludeerd dat voor scheve alternatieven (bv. exponentiële distributie) de SSPc testen minder krachtig zijn dan bv. de Anderson-Darling test. Onder dergelijke alternatieven blijkt ook de kracht af te nemen naarmate c groter wordt. Wanneer echter gecontamineerde normale distributies, die de neiging tot een bimodaliteit vertonen, beschouwd worden, werd vastgesteld dat de SSP3 en de SSP4 test heel veel krachtiger zijn dan alle andere testen die in de studie opgenomen werden. Tevens illustreert dit het belang van een goede keuze voor

de partitiegrootte c, wat in de praktijk een probleem stelt daar men meestal niet weet in welke zin de werkelijke distributie afwijkt van de vooropgestelde. Als oplossing hiervoor werd een data-gedreven versie van de SSPc test geconstrueerd.
Bij een data-gedreven SSPc test wordt de partitiegrootte c bepaald door de data zelf, aan de hand van een selectieregel, welke veel overeenkomst vertoont met het BIC (Schwartz, 1978). Naast een BIC-gebaseerde regel, werden ook twee andere criteria voorgesteld: het LL-criterium (Hannan and Quinn, 1979) en een criterium dat veel overeenkomst vertoont met het AIC van Akaike (1973, 1974). Een simulatiestudie toont dat voornamelijk de BIC-gebaseerde SSP test er in slaagt een c te selecteren die aanleiding geeft tot een grote kracht.
Tenslotte werden twee veralgemeningen voorgesteld. In een eerste veralgemening wordt de Pearson χ^2 statistiek vervangen door de power divergentie statistiek van order $\lambda \in \mathbb{R}$ (Cressie and Read, 1984). Voor $\lambda = 1$ wordt terug de Pearson statistiek bekomen en voor $\lambda = 0$ de likelihood-ratio statistiek. We hebben bewezen dan de asymptotische nuldistributie niet wijzigt onder de keuze van λ; voor eindige steekproefgrootte, daarentegen, is het gedrag wel afhankelijk van λ. In een kleine simulatiestudie werden de SSPc testen van order $\lambda = 1$ en orde $\lambda = 0$ met elkaar vergeleken. Er werd geconcludeerd dat de laatste test iets krachtiger is dan de eerste; zelfs in die mate dat het verschil in kracht met klassieke GOF testen onder scheve alternatieven erg klein geworden is.
Een tweede veralgemening is naar het multivariate geval. Hiervoor hebben we beknopt aangetoond hoe deze veralgemening mogelijk is zonder hierdoor extra complicaties te induceren.

Vervolgens werd de SSP statistiek aangepast voor het meerdere-steekproeven probleem (k steekproeven). Voor $c = 2$ blijkt de SSP statistiek identiek te zijn aan de Anderson-Darling statistiek voor het meerdere-steekproeven probleem (Pettit, 1976; Scholz and Stephens, 1987), zodat nuldistributie onmiddellijk volgt. Een belangrijk verschil t.o.v. de SSP statistiek voor het één-steekproef probleem is dat de statistiek nu een rank statistiek is, wat toelaat de test als een permutatietest te beschouwen. De uitbreiding naar grotere partities ($c > 1$) gebeurt zoals voorheen, en kan nu dus beschouwd worden als een veralgemening van de Anderson-Darling statistiek. Consistentie werd voor alle SSPkc testen bewezen. Ook werd een data-gedreven versie geconstrueerd. Uit een simulatiestudie werd besloten dat voor vele alternatieven de SSPkc testen voor $c > 2$ krachtiger zijn dan voor $c = 2$. Tevens zijn de SSPkc testen en de data-gedreven versies in het algemeen krachtiger dan de klassieke Kolmogorov-Smirnov en de Kruskal-Wallis test. In tegenstelling tot de resultaten uit het één-steekproef probleem werd hier waargenomen dat het LL-criterium tot hogere krachten leidt dan het BIC criterium.
Om de interpreteerbaarheid bij het verwerpen van de nulhypothese te bevorderen werd een decompositie van de SSPkc statistiek in $c - 1$ componenten voorgesteld. De eerste twee componenten zijn sterk verwant met respectievelijk de Kruskal-Wallis en de Mood statistiek. Wanneer er slechts twee steekproeven vergelijken moeten worden ($k = 2$), werd er een interessante link gevonden met de geleidelijke test van Janic-Wróblewska and Ledwina (2000) in het limiet geval $c = n$. Steunend op dit verband, denken we dat het mogelijk is de test van Janic-Wróblewska and Ledwina (2000) uit te breiden naar het algemene meerdere-steekproeven probleem ($k \geq 2$).

Tenslotte werd een SSP test ontwikkeld voor het onafhankelijkheidsprobleem. Aanvankelijk werd de methode geconstrueerd voor het toetsen van bivariate onafhankelijkheid. De steekproefruimte heeft nu dus twee dimensies, zodat een partitie ervan opgebouwd kan worden door eerst van iedere dimensie afzonderlijk een partitie te maken (stel met groottes r en c) om vervolgens de partitie van de bivariate steekproefruimte te bekomen als het product van de twee univariate. De grootte van een dergelijke partitie is dus $r \times c$ en de overeenkomstige statistiek wordt als SSPrc genoteerd.
Voor $r = c = 2$ werd de asymptotische nuldistributie bewezen en voor grotere partities werd een distributie voorgesteld die in een simulatie-experiment bevestigd werd. De statistiek is terug een rankstatistiek. Om de keuze van r en c te vereenvoudigen, werd een data-gedreven versie in deze bivariate context geconstrueerd.
In een uitgebreide simulatiestudie werden de SSPrc testen vergeleken met enkele andere testen voor onafhankelijkheid (o.a. een test gebaseerd op de Spearman rank correlatie, Hoeffding's omnibus test (Blum et al., 1961; Hoeffding, 1948) en twee recente data-gedreven geleidelijke testen (Kallenberg and Ledwina, 1999)). Hieruit werd besloten dat de SSPrc testen in het algemeen heel krachtig zijn en dat hun kracht in grote mate afhangt van de keuze van r en c. De data-gedreven versies brengen hier een goede oplossing; voornamelijk het BIC en het pAIC criterium selecteren veelal een goede partitiegrootte.
Ook hier werden twee uitbreidingen besproken: een uitbreiding naar power divergentie statistieken order $\lambda \in \mathbb{R}$ en een uitbreiding naar paarsgewijze onafhankelijkheid tussen de componenten van een multivariate variabele.

Samengevat kunnen de SSP testen op drie manieren ingepast worden in de bestaande literatuur.

- In het veralgemeende geval, is de SSP statistiek die gebaseerd is discretiseren van de data, waarna in de geïnduceerde tabel een power divergentie statistiek van orde λ (bv. een Pearson χ^2 statistiek) berekend wordt. Dit vertoont veel gelijkenis met de oudste manier van GOF testen. Om informatieverlies te voorkomen worden meerdere discretisaties beschouwd en worden de resulterende statistieken gecombineerd tot één nieuwe test statistiek. We hebben vastgesteld dat veel eigenschappen van de toepassing van een Pearson χ^2 test op gediscretiseerde data overdraagbaar zijn op de SSP test.

- Een belangrijk verband is er met de Anderson-Darling statistiek. Door de partitiegrootte niet te beperken tot 2, kan de SSPc statistiek in deze zin als een veralgemening van Anderson-Darling gezien worden.

- In een limietgeval ($c = n$) hebben we aangetoond dat de SSPc test equivalent is met een geleidelijke test. Ook de decompositie van de SSPc statistiek in interpreteerbare componenten is een typische techniek bij geleidelijke testen.

Dus de SSPc test vult een leemte in tussen de klassieke geleidelijke testen en de Anderson-Darling testen. Bij deze laatste twee statistieken wordt de informatie tegen de nulhypothese verdeeld over een oneindig aantal componenten, daar waar bij de SSP statistieken de informatie over een beperkt aantal interpreteerbare componenten verdeeld wordt.

Voornamelijk bij eindige steekproefgroottes lijkt dit een oplossing te zijn die vanuit een praktijk-georiënteerde invalshoek waardevol is.

Curriculum Vitae

Personalia

Naam: Olivier Thas
Geboortedatum en -plaats: 2 oktober 1971 te Lommel
Adres: Rode Leeuwlaan 19, 2640 Mortsel
Tel.: 03/440.28.67 en 0486/57.89.74
E-mail: olivier.thas @ rug.ac.be

Studies

Post-universitaire opleiding
Master of Science in Biostatistics, 1996, LUC

Universitaite opleiding
Bio-ingenieur, scheikunde, toegepaste scheikunde en microbiologie, 1994, UG

Loopbaan

1995-1996
Wetenschappelijk medewerker aan de Vakgroep Toegepaste Wiskunde, Biometrie en Procesregeling (TWBPR), FLTBW, UG; project gefinancierd door de Vlaamse Milieumaatschappij, getiteld: “Doorgedreven statistische interpretatie van de data in de meetdatabank”

1996-1997
Contractueel wetenschappelijk medewerken aan de Vakgroep TWBPR, FLTBW, UG

vanaf 1997
Assistent aan de Vakgroep TWBPR, FLTBW, UG; ondersteuning bij de opleidingsonderdelen Biometrie, Gegevensverwerking en Statistics en interne dienstverlening (statistische consulting)

Wetenschappelijke verwezenlijkingen

Publicaties

1. Boucneau, G., Van Meirvenne, M., Thas, O. and Hofman, G. (1996). Stratified kriging using vague transition zones. Proceedings of the second international symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, p. 251-258, May 21-23, Fort Collins, Colorado.

2. Thas, O., Van Vooren, L. and Ottoy, J.P. (1996). Modellen voor Categorische Data (in Dutch). *Verhandelingen van de Fac. L & TBW, Universiteit Gent, 36*, 5/1-38.

3. Goossens, E., Thielemans, S. and Thas, O. (1996). The Popularity of Painting Sport Scenes on Attic Black and Red Figure Vases: a CVA-based Research - Part A. *BABESCH, 71*, 59-94.

4. Struys, M., Versichelen, L., Thas, O., Herregods, L. and Rolly, G. (1997). Comparison of Computer Controlled Administration of Propofol versus two Manual Controlled Infusion Techniques. *Anaesthesia, 52*, 41-50.

5. Thas, O., Vanrolleghem, P., Kops, B., Van Vooren, L. and Ottoy, J.P. (1997). Extreme Value Statistics: Potential Benefits in Water Quality Management. *Water Science and Technology, 36*, 133-140.

6. Thas, O., Van Vooren, L. and Ottoy, J.P. (1998). Nonparametric Test Performance for Trends in Water Quality with Sampling Design Applications. *Journal of The American Water Resources Association, 34*, 347-357.

7. Boucneau, G., Van Meirvenne, M., Thas, O. and Hofman, G. (1998). Integrating Properties of Soil Map Delineations into Ordinary Kriging. *European Journal of Soil Science, 49*, 213-229.

8. Thas, O. and Ottoy, J.P. (1999). An Approach to Intervention Analysis in Spatio-Temporal Modelling of River Quality. In V. Barnett, A. Stein and K.F. Turkman (eds.), *Statistics for the Environment 4: Statistical Aspects of Health and the Environment*, Wiley, Chichester.

9. Darius, P., Ottoy, J.P., Solomin, A., Thas, O. Raeymaekers, B. and Michiels, S. (2000). A collection of applets for visualising statistical concepts. In: Proceedings of the 14th Conference of the International Association for Statistical Computing (COMPSTAT), 21-25 August, University of Utrecht, The Netherlands.

Voordrachten aan internationale congressen

1. Thas, O., Van Vooren, L. and Ottoy, J.P. (1996). Selection of Nonparametric Methods for Monotonic Trend Detection in Water Quality. Proceeding of the 5th HARMA workshop, February 23, Bari, Italy.

2. Thas, O., Van Vooren, L. and Ottoy, J.P. (1996). Graphical Models: Applicability and Software. Proceedings of the 6th HARMA workshop, p. 33-48, September 20, Berlin, Germany.

3. Thas, O., Vanrolleghem, P., Kops, B., Van Vooren, L. and Ottoy, J.P. (1997). Extreme Value Statistics: Potential Benefits in Water Quality Management. WATERMATEX 97, 4th IAWQ Symposium on Systems Analysis and Computing in Water Quality Management, Quebec City, Quebec, Canada, 17-20 June, 1997.

4. Thas, O. and Ottoy, J.P. (1997). An Approach to Intervention Analysis in Spatio-Temporal Modelling of River Quality. SPRUCE IV, International Conference on Statistical Aspects of Health and the Environment, Enschede, The Netherlands, 7-12 September, 1997.

5. Thas, O. and Ottoy, J.P. (1998). A nonparametric test for independence: a sample space partition resampling approach. IBC98, 19th Internation Biometric Conference, Cape Town, South Africa, 14-18 December 1998.

6. Thas, O. and Ottoy, J.P. (1999). The sample space partition resampling method: a nonparametric test for independence between both continuous and discrete variables. In: Proceedings of the Joint Statistical Meetings, Institute of Mathematical Statistics, Baltimore, USA, 8-12 August 1999, p. 260.

7. Thas, O., Vantieghem, S. and Ottoy, J.P. (1999). Spatio-temporal modelling for intervention analysis in river monitoring networks: a case study. In: Proceedings of the International Biometrics Society and Australian Region Biennal Conference (Biometrics99), Hobart, Australia, 12-16 December 1999, p. 66.

Voordrachten aan nationale symposia

1. Thas, O. and Ottoy, J.P. (2000). A data-driven goodness-of-fit test based on sample space partitions. In: Proceedings of the Anual Meeting of the Belgian Statistical Society (BVS), 5-6 October 2000, Heubermont, Belgium.

Posters

1. Kops, S.G.T., Gernaey, K., O. Thas and P.A. Vanrolleghem (1998). The Modelling of Noise Processes in Stochastics Differential Equations: Applications to Biotechnological Processes. Poster at the 7th IFAC Conference on Computer Applications in Biotechnology, Osaka, Japan, 31 May - 4 June 1998.

Verblijven in het buitenland

September 1996: Department of Statistics, Lancaster University, UK (Prof. Peter Diggle)

Juli-Augustus 2000: Statistics Department, Stanford University, USA (Prof. Joe Romano)

Didactische activiteiten

mei 1997
Cursus “Introduction to Statistics”, in het kader van TEMPUS, (University of Krakow), Krakow, Polen

mei 1999
Cursus “Introduction to Statistics”, in het kader van TEMPUS, (University of Sofia), Sofia, Bulgarije

1997-
Oefeningen van het opleidingsonderdeel “Biometrie”, 1ste proef bio-ingenieur (FLTBW, UG)

1997-
Oefeningen van het opleidingsonderdeel “Gegevensverwerking”, 2de en 3de proef bio-ingenieur (FLTBW, UG)

1999-
Oefeningen van het opleidingsonderdeel “Statistics”, Master of Science in Physical Land Resources, Environmental Sanitation, Food Science and Technology (FLTBW, UG)

1998-1999
Oefeningen van de cursus “Basiscursus Statistiek” in het kader van permanente vorming (IVPV, FTW, UG)

1998-
Cursus “Multivariate Statistiek” in het kader van permanente vorming (IVPV, FTW, UG)

2000-
Cursus “Multivariate Statistiek” in het kader van permanente vorming (CvS, UG)