



Rijksuniversiteit Groningen

**Comparing the Student's  $t$  and the ANOVA contrast procedure with  
five alternative procedures.**

Master Thesis

**Wobbe Zijlstra**

Groningen, December 2004

**Supervisors:**

**Prof. Dr. H.A.L. Kiers**

Department of Statistics and Data Analysis, RuG

**Dr. Ir. L.J.M. Mulder**

Department of Experimental and Work Psychology, RuG

## Abstract

A robustness study to investigate the performance of six procedures for comparing two or more groups, under several circumstances, has been carried out. The study was divided in four parts. The first three were Monte-Carlo studies, and the fourth was a study on empirical data. The six procedures were the ANOVA (or its special case, the Student's  $t$ ) procedure (1), the Satterthwaite (or its special case, the Welch) procedure (2), the bootstrap Satterthwaite (the Welch) procedure (3), the bootstrap percentile procedure for the mean (4) and for the median (5), and the Bonett and Price median procedure (6). The procedures were compared on confidence interval widths and coverage percentages, based on 10.000 replications. In Studies 1 and 2, samples were drawn from simulated population distributions with the same (non-) normality. The conditions were varied with different non-normal distributions, unequal population variances, and different equal or unequal sample sizes. In Study 1 two samples were compared on the mean or median and in Study 2 four samples were compared on the mean or median with a contrast analysis. For Study 3 simulated cardiovascular data obtained by Van Roon's baroreflex model (1998) were used. In this study, the results of logarithmically transformed data were compared to the results of untransformed data. Study 4 aimed at comparing real-world experimental data in a one-way ANOVA design with three independent groups.

The Student's  $t$  and the ANOVA procedure performed well under many conditions but failed in particular when unequal population standard deviations were accompanied with unequal sample sizes. In general, the problems encountered with the Student's  $t$  and the ANOVA procedures were best overcome by the Welch and the Satterthwaite procedures. The Student's  $t$ , the ANOVA, the Welch, and the Satterthwaite procedures, however, appeared more robust to deviations from normality than expected.

It was also found that the choice for logarithmically transformation of the cardiovascular data was not as obvious as expected on theoretical basis. The intended result of transformation was not always met. Furthermore, in particular for the Welch (and the Satterthwaite) procedure transformation of the cardiovascular data was not needed to get reliable results.

For the conditions investigated in this study, the Welch and the Satterthwaite procedure are as good as or much better than the Student's  $t$  and the ANOVA procedure. Therefore, it is concluded that the Welch procedure for comparing two samples and the Satterthwaite procedure for comparing more than two samples are the best procedure to use in experimental research.

## Contents

<b>1. Introduction.....</b>	<b>4</b>
§1.1 This study.....	4
§1.2 The Student's <i>t</i> and the ANOVA procedure.....	6
§1.2.1 Assumptions for the Student's <i>t</i> and the ANOVA procedure .....	6
§1.2.2 Earlier researches to the Student's <i>t</i> and the ANOVA procedure .....	8
§1.3 Alternative procedures .....	10
§1.3.1 The Welch and the Satterthwaite procedure .....	12
§1.3.2 The bootstrap procedures .....	13
§1.3.3 The Bonett and Price procedure.....	14
§1.4 Summarized description of the present study .....	15
§1.5 Questions.....	16
<b>2. Method .....</b>	<b>17</b>
§2.1 Monte-Carlo Simulation study.....	17
§2.1.1 Study 1: Simulation study, comparing two groups .....	20
§2.1.2 Study 2: Simulation study, contrast analysis .....	21
§2.1.3 Study 3: Cardiovascular simulation study, comparing two groups .....	21
§2.2 Study 4: Real-world experimental data.....	23
<b>3. Results .....</b>	<b>24</b>
§3.1 Study 1: Comparison of two samples from simulated distributions.....	24
§3.1.1 Normal distributions .....	25
§3.1.2 Non-normal distributions .....	29
§3.1.2.1 Leptokurtic distributions.....	29
§3.1.2.2 Platykurtic distributions.....	30
§3.1.2.3 Moderately skewed distributions .....	30
§3.1.2.4 Extremely skewed distributions.....	31
§3.1.3 Power properties .....	31
§3.1.4 Study 1 summarized.....	32
§3.2 Study 2: Contrast analysis with four samples from simulated distributions.....	39
§3.3 Study 3: Comparison of two samples from cardiovascular simulated data.....	48
§3.4 Study 4: Using real-world experimental data .....	57

<b>4. Discussion and conclusion .....</b>	<b>64</b>
§4.1 The three simulation studies .....	64
§4.2 Logarithmic transformation of cardiovascular data .....	67
§4.3 Smooth ideal simulated data versus experimental simulated data.....	68
§4.4 To conclude.....	69
<b>Appendices.....</b>	<b>71</b>
Appendix A: The procedures in formulas.....	71
Appendix B: Population distribution properties for Study 1 .....	75
Appendix C: Population distribution properties for Study 2 .....	76
<b>References.....</b>	<b>77</b>

# 1. Introduction

## §1.1 This study

In this study the Student's  $t$  procedure for comparison of means from two samples and the ANOVA contrast analyses for comparison of means from more than two samples are investigated. The Student's  $t$  and the ANOVA procedure are the most used procedures in (experimental) psychological research. The main question here is whether the Student's  $t$  and the ANOVA contrast procedure are still good procedures to use when assumptions of normality and/or equality of variances are violated, and which procedures are the best alternatives in such situations.

In experimental psychological research frequently the assumptions that scores in the population should be normally distributed and the population variances for the different groups should be equal are violated. Group manipulations and treatments often produce changes in means, as well as in variance, skewness, kurtosis and other population parameters (Sawilowsky & Blair, 1992). As a consequence, the Student's  $t$  and the ANOVA procedure could lead to wrong conclusions (e.g. Lix, Keselman & Keselman, 1996). It is expected that the main reason for rejecting the null hypothesis is not inequality of means, but rather inequality of variances, differences in skewness, or differences in other population parameters (Wilcox, 2001). Currently, when confronted with data from a non-normal distribution, many users of statistical methods will routinely use the Student's  $t$  and the ANOVA procedure without questioning the assumption of normality, and others will use it, hoping that the sample size is large enough so that the results are reasonable (Sutton, 1993). In the literature, various studies into the tenability of this hope have been reported, as will be discussed in §1.2.

The use of alternative procedures (e.g. the Welch procedure) or the use of other location parameters (e.g. the median) may under certain circumstances be better than the standard Student's  $t$  and ANOVA procedures. Here, five alternative procedures for comparing two or more groups are considered and discussed in §1.3.

The purpose of the present thesis is to report an extensive comparison of the six procedures, both on simulated and real data. A Monte-Carlo study is performed to investigate the performance of the six procedures for comparing two or more groups,

under different circumstances. Specifically, the performance under different non-normal distributions, equal or unequal population variances, and different equal or unequal sample sizes is investigated. In a Monte-Carlo study, samples are drawn from a known population. The data used for this Monte-Carlo study are drawn from populations with 'smooth or ideal' population distributions generated from a mathematical function. This study compares the procedures on the basis of the confidence intervals they provide, and the associated coverage percentages. Confidence intervals are used rather than actual probability Type I errors, because they are currently considered the best reporting strategy (APA, 2001). The coverage percentage is the percentage of time that the confidence interval includes the population location parameter, out of a large number of replicates, e.g. 10.000 (Lumley et al, 2002).

Micceri (1989) mentioned that conclusions from robustness studies that use smoothed ideal data or mathematical functions cannot be applied in educational and psychological settings. Micceri pleaded that research to real-world data would be more useful to get a better understanding of the use of procedures in psychological experimental research. Therefore, the procedures will also be investigated on realistically simulated data, and on real experimental (cardiovascular) psychological data.

First, a Monte-Carlo study is performed using simulated cardiovascular data based on Van Roon's baroreflex model (1998). Cardiovascular variability data are not normally distributed but Chi-squared distributed. Examples are the spectral functions of the heart rate variability (HRV) and blood pressure variability (BPV). Therefore, in practice cardiovascular variability data are logarithmically transformed to approximate normality. The logarithmic transformation of the spectral functions is theoretically grounded (e.g. Van Roon, 1998, p.96-97). A specific question in the present study is whether transformation is necessary, or whether other procedures without transformation lead to reliable and interpretable conclusions as well.

The procedures are also compared on real-world experimental data. The objective for this comparison is to see whether the procedures will lead to different conclusions, and, whether these findings can be compared to the findings of the Monte-Carlo study.

## **§1.2 The Student's *t* and the ANOVA procedure**

First the assumptions for the Student's *t* and the ANOVA procedure are discussed. After that earlier research on the performance of the Student's *t* and the ANOVA procedure will be briefly summarized.

### **§1.2.1 Assumptions for the Student's *t* and the ANOVA procedure**

In most experimental research the Student's *t* and the ANOVA procedure are used to draw inferences on the differences of means of two or more populations. These techniques employ the assumptions of independent observations with identically and normally distributed random variables with equal variance  $\sigma^2$  in all populations (Glass et al, 1972; Penfield, 1994; Scheffé, 1959; Stevens, 1996). Three distinct assumption violations can be considered: (a) non-independence, (b) unequal variances (heterogeneity), (c) non-normality (Glass et al, 1972).

(a) In case of non-independence, the scores/observations of the subject are influenced by other subjects or previous scores. Failure to satisfy the independence assumption can have serious effects on the validity of probability statements in the Student's *t* or the ANOVA procedure (Glass et al, 1972; Harwell et al, 1992; Scheffé, 1959, Stevens, 1996). For a single variable an intraclass correlation can be used to assess whether this assumption is tenable (Stevens, 1996). However, assessing dependence will not be possible if one does not know what causes dependence. When observations are dependent within subjects one speaks of repeated measures. When observations are dependent between subjects one has to know the factor causing dependence.

(b) The (in)equality of the (unknown) population variances is often investigated by a preliminary significance test of homogeneity (e.g. Moser & Stevens, 1992). However, such a preliminary test is senseless, because it is virtually inconceivable that such population variances would ever be exactly equal (Glass et al, 1972), hence the outcome of the significance test is mainly a reflection of the sample size, and does not indicate whether population variances are nearly equal or not.

(c) For the Student's  $t$  procedure and the ANOVA contrast analysis, the test statistic has to be  $t$ -distributed. For comparing two groups the statistic will be based on the mean difference and for more than two groups the statistic will be based on a contrast. For the test statistic to be  $t$ -distributed, a normal distribution of the population scores is sufficient, but the test statistic still can be (nearly)  $t$ -distributed in case of non-normality.

Two parameters play a central role in the effects of non-normality, the skewness ( $\beta_1$ ) and the kurtosis ( $\beta_2$ ) of the distribution (Miller, 1986; Scheffé, 1959):

$$\beta_1 = \frac{E(x - \mu)^3}{\sigma^3} \qquad \beta_2 = \frac{E(x - \mu)^4}{\sigma^4}$$

The skewness provides information about the symmetry of the distribution. A distribution is symmetric if it is the same to the left and right of the center point. A distribution is skewed if one of its tails is longer than the other. A positive skewness means that it has a long tail in the positive direction (skewed to the right). A negative skewness has a long tail in the negative direction (skewed to the left).

The kurtosis provides information about the size of the distribution's tails (tail weight/mass). Distributions that are highly peaked with thicker/heavier tails are called "leptokurtic" (kurtosis $>$ 3); those that are a flat-topped with thinner/lighter tails are called "platykurtic" (kurtosis $<$ 3). A uniform distribution would be the extreme platykurtic case. The normal distribution is called mesokurtic (kurtosis=3).

The difference of two means may still be approximately  $t$ -distributed if both means have skewed distributions, when the skewnesses, variances and sample sizes are about equal (Miller, 1986; Scheffé, 1959). Another type of violation of satisfying a  $t$ -distribution is due to the presence of outliers (also extreme values or influential points). These are observed values that are substantially different from the rest of the observations, but cannot be labeled as being erroneous measurements, miscalculations, etc. They are judged not to have come from another distribution than the rest of the data (Miller, 1986, p.9-10). The mean and especially the variance are very sensitive to outliers (Miller, 1986; Wilcox, 2001).

### **§1.2.2 Earlier researches to the Student's *t* and the ANOVA procedure**

"The assumptions of most mathematical models are always false to a greater or lesser extent. The relevant question is not whether ANOVA assumptions are met exactly, but rather whether the plausible violations of the assumptions have serious consequences on the validity of probability statements based on the standard assumptions" (Glass et al., 1972, p.237).

A summary of earlier research on the performance of the Student's *t* procedure will be given when assumptions are met and when the assumptions are violated. For a quick overview, see Table 16 of Glass et al (1972, p.273) and Table 7 of Harwell et al (1992, p.333). Earlier research on the performance of the ANOVA contrast analysis will also be described.

Statistical methods are called robust if the inferences are not seriously invalidated by the violations of the assumptions (Miller, 1986; Scheffé, 1959). Robustness is often operationalized as the actual Type I errors being near the nominal  $\alpha$  level. For a discussion about criteria for robustness see Bradley (1978).

When the two samples consist of independent observations from identically and normally distributed populations with equal sample sizes, the Student's *t* procedure is the best procedure in terms of power and control of Type I error (Penfield, 1994; Ramsey, 1980; Wilcox, 1997). Results from simulation studies show that the Student's procedure has the highest power and that the Type I error is very close to  $\alpha$ , with no influence of unequal sample sizes on either Type I error (Gibbons & Chakraborti, 1991; Penfield, 1994; Zimmerman, 1987;) or power (Gibbons & Chakraborti, 1991; Zimmerman, 1987). In terms of confidence intervals, the Student's procedure will be closest to the actual confidence percentage with the interval width being the smallest.

Complete robustness of Student's *t* to unequal variances in normal populations is reached, with  $\alpha=.05$ , when the sample sizes are equal and larger than six in terms of Bradley's (1978) liberal criterion, and larger than 15 in terms of Cochran's limits (1954). When the variances have ratio 1:4, equal sample sizes as small as four will result in acceptable Type I errors (Ramsey, 1980). Other researches have found comparable results (Gibbons & Chakraborti, 1991; Glass et al, 1972, Wilcox, 2001; Zimmerman, 1987) with Stevens (1996) stating these results even will hold when the sample sizes are unequal up to a ratio of 3:2. When the two sample sizes are equal, the two-sample

Student's  $t$  procedure is in every case a safe one to use. Although much is known about the effect of unequal population variances, Harwell et al (1992, p.332/334) conclude that "the conventional conclusion that heterogeneous variances are not important when  $n$ 's are equal seems to have boundary conditions like all other conclusions in this area, and the boundary conditions may not have been sufficiently probed".

According to some researchers (e.g. Miller, 1986; Pearson & Please, 1975; Scheffé, 1959) it is very useful to balance the design of an experiment as closely as possible. In balanced designs the boundaries of confidence intervals are reliable, but in case of few observations the confidence interval width will be large. In heavily unbalanced designs with more observations, the boundaries will be unreliable but the width will be (much) smaller than for a design with all sample sizes equal to the smallest. It can be considered that smaller confidence intervals are more important than highly reliable boundaries. Thus comparing a small patient group with a large control group is more informative than comparing a small patient group with a small control group. Perhaps, alternative procedures give more reliable and smaller confidence intervals when the design is heavily unbalanced.

When the population scores are normally distributed but have unequal variances and unequal sample sizes, the Type I error can differ dramatically from  $\alpha$  with a large 'pairing-effect' (Gans (1981); Gibbons & Chakraborti, 1991; Glass et al, 1972; Penfield, 1994; Ramsey, 1980; Scheffé, 1959; Stevens, 1996; Wilcox, 2001; Zimmerman, 1987;). Positive pairing occurs when the largest sample is associated with the largest variance. Positive pairing will have conservative Type I errors (smaller than  $\alpha$ ). Negative pairing occurs when the largest sample is associated with the smallest variance. Negative pairing will have liberal Type I errors (larger than  $\alpha$ ). The worst situation is negative pairing, where the least information is available on the population with the larger variance (Miller, 1986).

In the presence of non-normality, the Student's  $t$  procedure is "reasonably robust to Type I error when (a) sample sizes are equal, (b) sample sizes are fairly large, and (c) tests are two-tailed rather than one-tailed" (Sawilowsky & Blair, 1992, p. 352). Sample sizes are fairly large with 25 or 30 observations (Boneau, 1960). Skewness and kurtosis slightly affect the actual Type I errors (Gans (1981); Glass et al, 1972; Harwell et al,

1992; Pearson & Please, 1975; Penfield, 1994; Scheffé, 1959) and have more influence on the power of the two sample Student's *t* or multiple sample ANOVA procedures (Penfield, 1994). However, when the non-normal distributions are equally distributed (same skewnesses and variances) with equal sample sizes, the actual Type I error doesn't deviate from the nominal level (Miller, 1986; Scheffé, 1959; Wilcox, 1997, 2001). Harwell et al (1992) conclude that the effects of skewness and kurtosis are negligible, except for the situation when the variances are unequal. When the population variances differ or samples sizes are unequal, the effect of nonzero skewness on the Student's *t* procedure tends to increase (Miller, 1986; Scheffé, 1959) but the effect of nonzero kurtosis tends to increase only a little (Miller, 1986). Glass et al (1972) conclude that skewness has no effect on power while kurtosis can substantially affect the power when sample sizes are small. There is no interaction between the effect of unequal population variances and unequal sample sizes with non-normality on the Type I errors (Glass et al, 1972).

The effects of unequal population variances, unequal samples sizes, and pairing on Type I errors and power are about the same for the ANOVA procedure (Glass et al, 1972; Lix et al, 1996; Scheffé, 1959). There might be an effect of number of groups on the Type I error, with a worsening effect on Type I errors when more groups are compared (Scheffé, 1959). For a one-way layout with more than two equally sized samples, non-normality has a small effect on the ANOVA procedure (Scheffé, 1959; Wilcox, 1997). The ANOVA procedure is relatively insensitive to violations of the normality assumption in terms of Type I errors, it is highly sensitive to differences in population variances (Keselman et al, 1998)

### **§1.3 Alternative procedures**

Are there procedures known to perform well when the assumptions of equal population variances and of normality are questionable? Classically, the Mann-Whitney procedure using rank scores is used in presence of non-normality and the Welch procedure is used in presence of unequal population variances (Zimmerman, 1992). This study the use of rank scores is not investigated, and therefore the Welch procedure using rank scores suggested by Zimmerman (1992) is also not used in this study.

Generally, the solutions for obtaining more valid inferences, when assumption are violated, can be divided in two categories; firstly the use of other location parameters and secondly the use of other procedures for estimating confidence intervals.

As other location parameter, only the median is investigated in this study. Other so called "robust estimators" are not investigated here; for performances of robust estimators see for example Keselman, Wilcox & Lix (2003), Lix & Keselman (1998), Van der Wal (2004), and Wilcox (1997, 2001).

When the population data are symmetrically distributed, the median is equal to the mean. Thus, in this case both the sample mean and median can be used to estimate the population mean. When the population data are normally distributed, the mean is the most accurate estimator: nothing beats the mean under normality (Wilcox, 2001). When the population is non-normally and symmetrically distributed or the sample sizes are small, the median can be a more accurate estimate of the population mean (Bonett and Price, 2002).

When a researcher uses the mean as location statistic and the population data are asymmetrically distributed, the mean can lead to highly inaccurate estimations of the location statistic. On the other hand, median based procedures do not make the assumption of normality, and are robust to almost any type of non-normality (Bonett and Price, 2002).

As alternative procedures for estimating confidence intervals, five alternative procedures are investigated in this study; the Welch (for two groups) and the Satterthwaite procedure (for more than two groups) (1), the bootstrap Welch and the bootstrap Satterthwaite procedure (2), the bootstrap percentile procedure for the mean (3) and for the median (4), and the Bonett and Price median procedure (5). The Satterthwaite's procedure, a generalization of the Welch procedure, will be used for the ANOVA contrast analyses. The procedures are introduced briefly below; for formulas of the procedures see appendix A.

### **§1.3.1 The Welch and the Satterthwaite procedure**

The Welch procedure (in SPSS (2001) denoted by 'equal variances not assumed') is an often-used parametric alternative for comparing two groups. This procedure also assumes normal population distributions, but not equality of population variances. The Welch procedure uses adjusted degrees of freedom and weighted variances instead of pooled variances. When the sample sizes are equal and the *sample* variances are exactly equal, the Welch procedure is the same as the Student's *t* procedure. But when the sample variances are slightly different while the population variances are in fact equal, the Welch procedure will have smaller degrees of freedom which will result in a somewhat lower power (Ramsey, 1980). Zimmerman (1992) concluded that there are no undesirable consequences using the Welch procedure instead of the Student's *t* procedure when the population variances are equal.

The Welch procedure has been generalized for use with more than two groups with unequal population variances. This procedure is called the Satterthwaite procedure, for analyzing contrasts of any number of groups (actually comparing two groups also pertains to a contrast with weights -1 and 1). For the ANOVA procedure, the assumption of equal population variances is most of the times ignored. Hence, the Satterthwaite procedure is seldom used. The Satterthwaite is available in SPSS for one-way ANOVA procedures and for one-way ANOVA contrast analyses as the Welch-option, but it is not available for more complex contrast analyses.

The Welch and Satterthwaite procedures give a good control on Type I error when populations are normally distributed with unequal variances and/or unequal sample sizes (Gans (1981); Gibbons & Chakraborti, 1991; Harwell et al, 1992; Lix et al, 1996; Penfield, 1994; Ramsey, 1980).

The Welch procedure is robust to many types of non-normal distributions with equal sample sizes of ten or more per group (Bonett & Price, 2002). The Welch procedure is influenced by the shape of a population when sample sizes are unequal (Bonett & Price, 2002; Gans (1981); Harwell et al, 1992), when the sample sizes are small and when the two population distributions have very dissimilar shapes (Bonett & Price, 2002). The skewness is more important for Type I errors and the kurtosis is more

important for the power (Penfield, 1994). Lix et al (1996, p.613) give a good summary when to use the Welch procedure.

### **§1.3.2 The bootstrap procedures**

Bootstrap procedures are non-parametric procedures which are easy applicable but are computer intensive. The bootstrap procedure could be a good alternative for determining confidence intervals because no assumptions concerning the population distribution are made (Wasserman & Bockenholt, 1989). The bootstrap procedure is therefore an appropriate procedure for data from non-normal distributions that can be used for every statistic. The use of bootstrap procedures can lead to asymmetric confidence intervals. This asymmetry represents an important part of the improvement in the procedure's coverage percentage (Efron & Tibshirani, 1993). Noreen (1989) mentioned that the bootstrap procedures are easy to use and flexible, but some "appear to be unreliable and should be used with caution" (p.63).

The basic idea of the bootstrap procedure is that B new samples (bootstrap samples) are drawn with replacement from each of the original sample (see for example Wilcox, 2001; Wasserman & Bockenholt, 1989). In the bootstrap percentile procedure, for each of the B bootstrap samples the location statistic (mean  $\bar{x}_j$  or median  $\eta_j$ ) is calculated. In case of comparing two groups the statistics are subtracted from each other resulting in differences of mean or medians. In case of contrast analysis the contrast is calculated. The B differences or contrasts, give an approximation for the distribution of sample differences. The middle 95% of these differences are taken to get a 95% percentile interval which can be considered an approximation to a 95% confidence interval (see Wilcox, 2001). The bootstrap percentile procedure is transformation-respecting: the confidence interval before transformation is the same as after transformation (Efron & Tibshirani, 1993).

For the bootstrap Welch and bootstrap Satterthwaite procedure, for each of the B bootstrap samples the mean ( $\bar{x}_j^*$ ) and standard deviation ( $s_j^*$ ) are used to calculate the Welch value ( $W^*$ ). Next the B  $W^*$  values are ordered. The lower (at 2.5%) and upper (at 97.5%) values are set to be the new critical values for determining the 95% confidence interval. The Welch procedure uses the samples' standard error and the new critical

values to get the 95% confidence interval (see Wilcox, 2001). For contrast analysis the same procedure is carried out except that Welch is substituted by Satterthwaite (see also appendix A). The bootstrap Welch and bootstrap Satterthwaite procedure are not transformation-respecting. It makes a difference which scale is used to construct the confidence interval (Efron & Tibshirani, 1993).

The bootstrap procedures are very useful for both univariate and multivariate analyses of variances when it is doubtful that the normality assumption is (nearly) satisfied, particularly when sample sizes are small (Wasserman & Bockenholt, 1989). The bootstrap procedures perform about as well as, and in some cases much better than, the Student's  $t$  procedure. However, the control of Type I error is still unsatisfactory when using a bootstrap procedure with the mean (Wilcox 2001), especially when the population distribution is skewed (Noreen, 1989). The bootstrap Welch procedure has no advantage with respect to Type I errors over the Student's  $t$  procedure when the groups are equally distributed, but it does have an advantage when the distributions of the groups differ considerably (Wilcox, 2001).

### **§1.3.3 The Bonett and Price procedure**

The Bonett and Price (2002) median procedure is developed to estimate confidence intervals for a linear function of medians. The method does not rely on distribution assumptions, and is computationally simple (even without a computer). For calculating the confidence interval the median and a distribution-free estimate of the variance of the median are determined. For calculating the median's distribution-free estimate of the variance, a value from a table and a lower and an upper score are determined from the ordered scores (see also appendix A).

The Bonett and Price procedure performs well for realistic non-normal distributions with sample sizes as small as five per group. In a Monte-Carlo study performed with non-normal distributions and equal samples ( $n_1=n_2=5$ ), the coverage percentages were between 94.0% and 96.4% and in most cases it exceeds 95%. However, when distributions are highly skewed the small-sample coverage percentage can be slightly below 95% but will quickly approach 95% as the sample size increases (Bonett and Price, 2002).

#### **§1.4 Summarized description of the present study**

To sum up, the present study is divided in four parts (for more details see the Method section):

##### Study 1- Monte-Carlo simulation for comparing two samples.

Six procedures for comparing two samples are compared on coverage percentage under several conditions. The samples are drawn from two simulated population distributions with the same non-normality. The conditions are varied with different non-normal distributions, unequal population variances, and different equal or unequal sample sizes. For each procedure the coverage percentage, based on 10.000 replications, is calculated.

##### Study 2- Monte-Carlo contrast simulation.

The general design for contrast simulation is the same as for comparing two samples. Now, four samples are used to calculate confidence intervals for the contrast with weights -1, -1, 1, and 1. This contrast can be seen as a one-way ANOVA design with four groups or as an interaction effect in a 2 by 2 design.

##### Study 3- Monte-Carlo cardiovascular simulation for comparing two samples.

Six procedures for comparing two samples from two different populations are compared on coverage percentage for nine simulated cardiovascular variables obtained by Van Roon's baroreflex model (1998). Also, for six variables the results with the logarithmically transformed version of the variable are compared to those with the non-transformed variable. Two different equal sample sizes and two unequal sample sizes are used. For each procedure the coverage percentage, based on 10.000 replications, is calculated.

#### Study 4- Real-world experimental data.

The six confidence intervals resulting from the procedures are compared for nine real-world cardiovascular variables and six transformed variables (Althaus et al, 2004). The confidence interval is calculated for the contrast of three groups used in the study.

### **§1.5 Questions**

With this study, it is aimed to answer the following questions:

#### Question 1 (via Studies 1, 2 and 3):

In what circumstances do the Student's  $t$  and the ANOVA procedure fail, while alternative procedures perform well?

#### Question 2 (via Studies 1, 2 and 3):

- a) Which procedures perform best under what circumstances?
- b) Which procedures perform well under all circumstances (robustness)?

#### Question 3 (via Studies 3 and 4):

Is logarithmic transformation of cardiovascular data necessary in cardiovascular research or can equal good results be obtained by alternative procedures without transformation?

#### Question 4 (via Study 4):

Do the procedures lead to different conclusions when employed to real-world data?

## 2. Method

This study consists of four parts. The first two are Monte-Carlo studies with simulated data for comparing two samples and for contrast analysis. The third part is a Monte-Carlo study for comparing two samples with Van Roon's cardiovascular simulated data. The last part aims at comparing real-world experimental data in a one-way ANOVA design with three independent groups.

### §2.1 Monte-Carlo Simulation study

The simulation study is threefold. In the first and third part six procedures for comparing two groups are investigated and in the second part six procedures for contrast analysis with four groups are investigated. The performance of the six procedures is compared on coverage percentage under different non-normal distributions, equal or unequal population variances, and different equal or unequal sample sizes.

For each set of samples (two samples for Studies 1 and 3, and four samples for Study 2) drawn in the simulation studies, a confidence interval for the mean or median is calculated according to all six procedures. In each condition 10.000 samples were drawn, and for each sample confidence intervals were computed. Their widths are recorded, but most attention is paid to the percentage of times that the population parameter is covered by the confidence interval, called the coverage percentage. For the bootstrap procedures, the number of bootstrap resamples is set to 1000.

The population distributions are simulated by using the generalized lambda distribution (GLD) developed by Ramberg and Schmeiser (1974) and Ramberg et al. (1979). The amount of skewness and kurtosis of a distribution can be varied while the mean and standard deviation are set to respectively zero and one. The GLD is given by

$$R(p) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, \quad (0 \leq p \leq 1)$$

where  $p$  is a uniform (0,1) random variable,  $\lambda_1$  is a location parameter,  $\lambda_2$  is a scale parameter and  $\lambda_3$  and  $\lambda_4$  are shape parameters. Ramberg et al (1979) provided a Table

with the parameter choices required for obtaining a wide range of skewness and kurtosis values.

By adding two parameters to the GLD, the mean and standard deviation of the distribution can be set at any value desired. The function then becomes:

$$R(p) = b \left( \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2} \right) + a$$

where  $a$  is the shift in mean and  $b$  is the standard deviation of the distribution.

In this simulation study five population distributions are used that vary in skewness and kurtosis. Based on the literature (Glass et al, 1972; Penfield, 1994; Wilcox & Charlin, 1986) the normal, a leptokurtic, a platykurtic (rectangular), a moderately skewed, and an extremely skewed distribution are used (see Table 2.1 and Figure 2.1 ). A population distribution is constructed by taking randomly a large sample of 1.000.000 from Ramberg's GLD. The large sample population distribution is approximately the same as that given by the density function. Using a finite population was preferred over using a density function for an infinite population, because in this way the median is easily determinable. On the other hand, the finite population distributions differ slightly from the intended skewness, standard deviation and effect size with a few thousandths. The intended kurtosis differs in some conditions a bit more, with a few hundredths to even a few tenths (see appendix B and appendix C).

**Table 2.1**

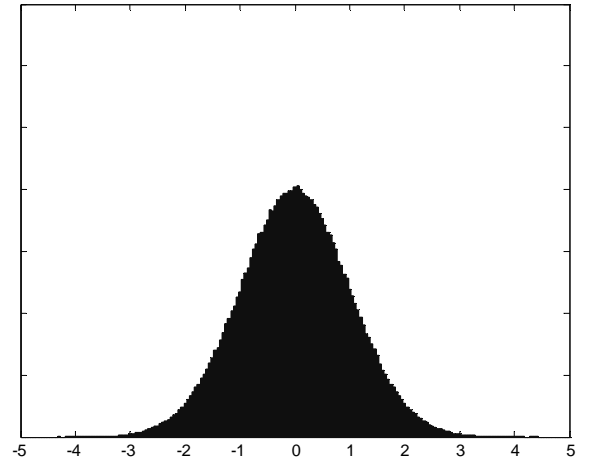
The five distributions used in Studies 1 and 2 with Ramberg's GLD parameters

Distribution	Skewness	Kurtosis	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$
1 Normal	0.0	3.0	0	.1974	.1349	.1349
2 Leptokurtic	0.0	9.0	0	-.3203	-.1359	-.1359
3 Platykurtic	0.0	1.8	0	.5774	1	1
4 Moderately Skewed	0.65	3.8	-.472	.1072	.0377	.0952
5 Extremely Skewed	2.0	9.6	-.865	-.0331	-.002125	-.0298

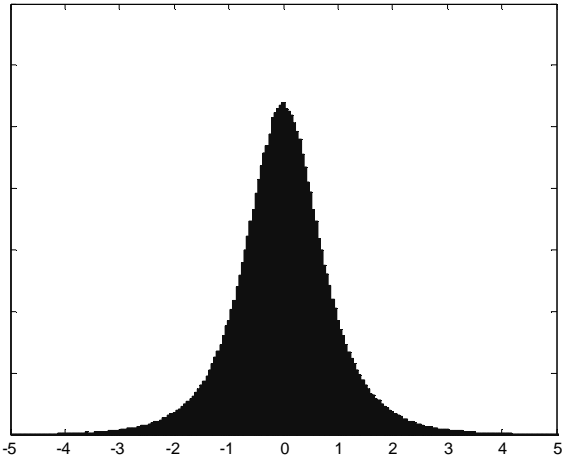
**Figure 2.1**

The five distributions based on Ramberg's GLD used in Studies 1 and 2

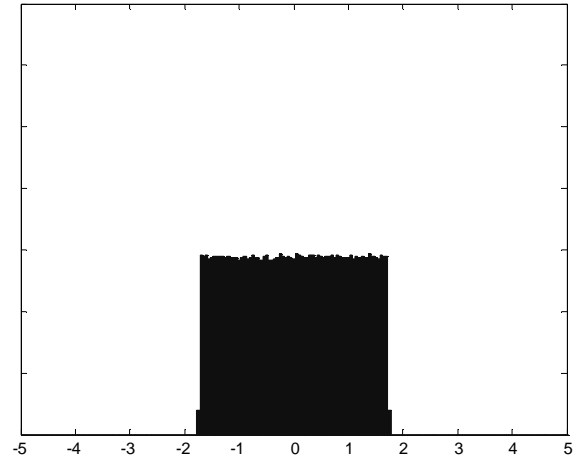
a. Normal



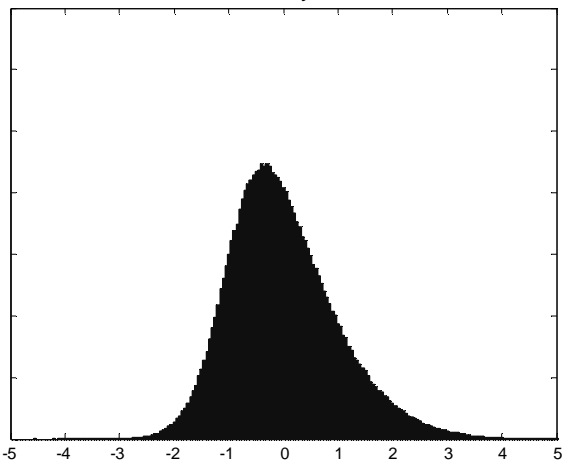
b. Leptokurtic



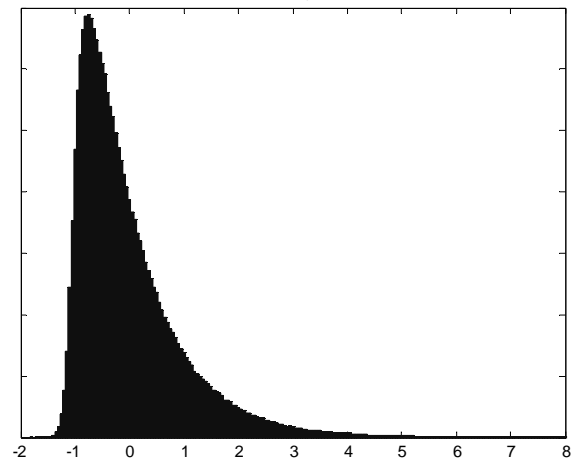
c. Platykurtic



d. Moderately skewed



e. Extremely skewed



### §2.1.1 Study 1: Simulation study, comparing two groups

The design for studying procedures comparing two groups is as follows. For each of the five population distributions 28 conditions were considered. The 28 conditions were combinations of four data type conditions with seven sample size conditions (see Tables 2.2a and 2.2b).

**Table 2.2a**  
The four data type conditions  
for Study 1

	$\mu_1$	$\mu_2$	$\sigma_1$	$\sigma_2$
A	0	0.8	1	1
B	0	0.8	2	1
C	0	0.8	4	1
D	0	0	2	1

**Table 2.2b**  
The seven sample size conditions  
for Study 1

	$n_1$	$n_2$
1	8	8
2	16	16
3	24	24
4	8	16
5	8	24
6	16	32
7	32	16

The effect size ( $\mu_2 - \mu_1$ ) was varied, because in real research the samples will have different means and medians. To investigate the influence of effect size on the coverage percentage, condition D will be compared to condition B. To investigate the effect of different population standard deviations ( $\sigma$ ) conditions A, B and C will be compared. For studying the effect of sample size, different equal sample sizes and different unequal sample sizes with different ratios are used.

Also the effect of pairing is studied. Pairing occurs when unequal group sizes are paired with unequal standard deviations. A negative pairing condition is one in which the smaller group has the greater population variance (data type conditions B, C and D with sample size conditions 4, 5, and 6), and a positive pairing condition is one in which the smaller group has the smaller population variance (data type conditions B, C and D with sample size condition 7).

### §2.1.2 Study 2: Simulation study, contrast analysis

The design for the study of contrast analysis procedures (see Tables 2.3a and 2.3b) is similar to that in Study 1. The means for the four groups are varied but they are the same for all conditions. The choice of variation in standard deviations and sample size conditions 1-3 and 6-7 is based on studies by Wilcox (2001), Glass et al (1972), and Lix & Keselman (1998). Sample size condition 4 is chosen this way because in practical research small variations in the number of subjects may occur as follows. The researcher starts in every group for example with 16 subjects, but the drop out will be differently for the groups. Sample size condition 5 is the same as condition 4 but now with doubled sample sizes.

**Table 2.3a**  
The three data type conditions for Study 2

	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$	$\sigma_1$	$\sigma_2$	$\sigma_3$	$\sigma_4$
A	0	0.3	0.6	0.9	1	1	1	1
B	0	0.3	0.6	0.9	4	1	1	1
C	0	0.3	0.6	0.9	4	3	2	1

**Table 2.3b**  
The seven sample size conditions for Study 2

	$n_1$	$n_2$	$n_3$	$n_4$
1	8	8	8	8
2	16	16	16	16
3	24	24	24	24
4	8	10	12	14
5	16	20	24	28
6	8	8	8	24
7	24	8	8	8

### §2.1.3 Study 3: Cardiovascular simulation study, comparing two groups

Six procedures for comparing two samples from two different populations are compared on coverage percentage for nine cardiovascular variables simulated by Van Roon's baroreflex model (1998, see also Van Roon et al, 2004), as will be explained below. Also, for six variables, the results for logarithmically transformed variables are compared to the results for non-transformed variables. Two different equal sample sizes (8-8 and 24-24) and two unequal sample sizes (8-24 and 24-8) are used.

Van Roon's baroreflex model (1998) estimates the latent vagal (parasympathetic) and latent sympathetic influence on cardiovascular responses especially in mental loading tasks. The vagal and sympathetic nervous systems are part of the autonomic nervous

system. The effect of mental loading is a decreased baroreflex sensitivity which is accomplished by a decreased vagal activation and an increased sympathetic activation. A decrease in vagal activation results (via the nervus vagus) in an increase in heart rate (HR). In the model this is accomplished by a reduction of vagal gain ( $G_v$ ). An increase in sympathetic activation results (via the spinal cord) in an increase in heart rate, systemic resistance, maximum elastance of the heart, and stretched venous volume. In the model this is accomplished by a reduction of sympathetic gain ( $G_s$ ). In theory, the reduction of vagal and sympathetic gain should both increase the blood pressure (BP). However, the model shows that the mean blood pressure (MBP) is almost independent of vagal gain and strongly dependent of sympathetic gain.

The model searches for scores on the latent variables  $G_v$  and  $G_s$  such that the model's cardiovascular output has the best fit with experimentally measured cardiovascular data. For  $G_v=1$  and  $G_s=1$  the model's parameters were set in such a way that the model's cardiovascular output has the best fit with experimental data for the rest condition. Subsequently Van Roon (1998) searched for the  $G_v$  en  $G_s$  values that lead to the best approximation of the cardiovascular response in a mental loading task for a particular sample of subjects.

Van Roon (1998) systematically varied the gains,  $G_v$  and  $G_s$ , to find the best fit. For each gain the values were varied from .5 to 1.2 in steps of .02, which resulted in 1296 combinations of  $G_v$  and  $G_s$ . For each combination of  $G_v$  and  $G_s$  nine variables were computed by the model; inter-beat interval (IBI), heart rate variability (HRV in low, mid, and high frequency band), blood pressure variability (BPV in low, mid, and high frequency band), and Modulus (mid frequency band). The modulus is the strength of the relation between blood pressure and inter-beat interval changes (expressed in ms/mmHg), which may be considered an index of baroreflex sensitivity.

For the HRV and BPV a power spectrum is calculated. The power spectrum can be expressed in a low frequency band (.02-.06 Hz), a mid frequency band (.07-.14 Hz), and a high frequency band (.15-.40 Hz). The low frequency band is associated with homeostatic processes and body temperature regulation, the mid frequency band is associated with regulatory properties of the short-term blood pressure control, and the high frequency band is associated with respiratory activity. Also the modulus function is

frequency dependent. In the present study the mean modulus in the mid frequency band is used as relevant variable.

In this study two populations were constructed. The first population is associated with high sympathetic activation and rather normal vagal activation, and can be labeled as the "hypertension group". The second group is associated with low vagal activation and rather normal sympathetic activation, and can be seen as the "high heart rate group".

Population 1 = hypertension (more sympathetic effect)	Gv=.90, Gs=.74
Population 2 = high heart rate (more vagal effect)	Gv=.68, Gs=.90

For constructing a population, some assumptions about the distributions of Gv and Gs were made; Gv and Gs are normally distributed around the indicated values with a standard deviation of .08. A population is constructed by taking 1.000.000 times two numbers randomly from a standard normal distribution; one will lead to a Gv-value and one to a Gs-value. G-values that end up lower than .50 will be set to .50, and G-values that end up higher than 1.2 will be set to 1.2. Each combination of a Gv-value and a Gs-value will lead to scores on the nine variables, and thus 1.000.000 such pairs lead to population scores for all nine variables.

## **§2.2 Study 4: Real-world experimental data**

The confidence intervals resulting from the six procedures are compared for nine real-world cardiovascular variables. The data are from Althaus et al (2004) where two groups of children ( $n_1=16$ ,  $n_2=16$ ) with autistic-type behavior problems were compared to a group of normal children ( $n_3=16$ ) with respect to their cardiovascular responses. For the nine cardiovascular variables, confidence intervals are calculated for the contrasts of groups 1 and 2 against group 3 (control), using all six procedures. The intervals are compared in terms of width, and in terms of conclusions they imply.

Here also, for the six variables HRV (low, mid, and high frequency band) and BPV (low, mid, and high frequency band), the results for logarithmically transformed variables are compared to the results for non-transformed variables.

### 3. Results

In the present section, the results are reported for the four studies described in Section 2. In all studies the bootstrap procedures used 1000 bootstrap resamples. The results for Studies 1, 2, and 3 are based on 10000 replications. The main focus is on coverage percentages, which ideally should be 95%, but also some information is given on the widths of the confidence intervals. With 10000 replications the estimated 95% margin of error of the coverage percentages of the 95% confidence interval will be approximately  $\pm 0.4\%$ . So if the procedure performs correctly the coverage percentages are expected to be between 94.6% and 95.4% in 95% of the cases.

To investigate the influence of the different conditions on the procedures' performances, the coverage percentages for Study 1 (comparing two samples) are reported in detail, while the main distinction is made between conditions using normal distributions and those using non-normal distributions. The results for Study 2 (contrast analysis) and Study 3 (comparing two samples using cardiovascular data) are reported in less detail because the patterns of coverage percentages are very similar to those found in Study 1. For Study 4 (real experimental data) the confidence intervals of the different procedures are compared. For six cardiovascular variables also the performance after logarithmic transformation is compared to the performance in case of no transformation.

For Study 1 and 2 the results are also summarized by the proportion of "bad" coverage percentages for the different conditions. A bad coverage percentage is defined here as a coverage percentage below 94% or above 96%.

#### **§3.1 Study 1: Comparison of two samples from simulated distributions**

In this study six procedures for comparing two samples are compared. In Table 3.1 the coverage percentages and the mean interval widths for the six procedures are shown for the case of normal population distributions. The coverage percentages are displayed graphically in Figure 3.1. In Figures 3.2 to 3.5 the coverage percentages for the four types of non-normal population distributions are shown.

The influence of effect size (difference between means or medians) on the coverage percentage is negligible for all procedures in all distribution types. The

coverage percentages of the condition with zero effect size are comparable to the coverage percentages of the same condition with nonzero effect size.

### §3.1.1 Normal distributions

The discussion of the results for normal distributions is based on Table 3.1 and Figure 3.1. The Student's  $t$  is the most used procedure for comparing two samples. Therefore the performance of the Student's  $t$  procedure is discussed first. After that the performances of the alternative procedures are given.

The Student's  $t$  procedure performs very well in the case of  $n_1=n_2$ , even when  $\sigma_1\neq\sigma_2$  and also very well in the case of  $\sigma_1=\sigma_2$ , even when  $n_1\neq n_2$ . With sample sizes as small as  $n=8$  when  $n_1=n_2$  and/or  $\sigma_1=\sigma_2$  the Student's  $t$  procedure performs acceptably, the coverage percentages are between 93.7% and 95.2%.

The Student's  $t$  procedure performs very poorly when both  $\sigma_1\neq\sigma_2$  and  $n_1\neq n_2$ , as is striking in Figure 3.1d and 3.1f. With negative pairing (the smaller sample associated with the larger population standard deviation) the coverage percentages are far below 95% (liberal). The worst coverage percentage (75.7%) is seen when sample sizes and population standard deviations differ most. With positive pairing (the larger sample associated with the larger population standard deviation) the coverage percentages are above 95% (conservative).

Furthermore, it can be seen for the Student's  $t$  procedure that, when the group sizes ratio  $n_2/n_1=2$ , the coverage percentages decline with growing differences in population standard deviation. This decline is stronger when the group sizes ratio  $n_2/n_1=3$  ( $n_1=8$  and  $n_2=24$ ).

The Welch and the bootstrap-Welch procedure perform very well in all conditions; the coverage percentages are between 94.0% and 95.4%. The bootstrap percentile mean always has coverage percentages below 94% and is most affected by group size: the coverage percentages decrease with the size of the smallest sample. The bootstrap percentile median and the Bonett and Price (B&P) procedures have conservative coverage percentages when  $\sigma_1=\sigma_2$ . The coverage percentages decrease with increasing difference in population standard deviation and the coverage percentages also

decrease with increasing difference in sample size. None of the five 'alternative procedures' have clear effects of negative or positive pairing.

The confidence interval widths for the six procedures do not differ much. The bootstrap percentile mean gives smaller confidence intervals while the bootstrap percentile median and the B&P procedure give wider confidence intervals compared to the Welch and bootstrap-Welch procedure. When  $\sigma_1 \neq \sigma_2$  the Student's  $t$  confidence interval widths for  $n_1=n_2=16$  and  $n_1=n_2=24$  are about equal to the confidence interval widths of respectively  $n_1=8, n_2=16$  and  $n_1=8, n_2=24$ , while the alternative procedures give wider confidence intervals in the latter cases (see Table 3.1). The relation between coverage percentages and confidence interval widths is clear; in general when a procedure's interval is narrower compared to the other procedures the coverage percentage is lower compared to the other procedures. This relation can, for a great part, explain the under-coverage (<95%) and the over-coverage (>95%) that is seen for some procedures in some conditions.

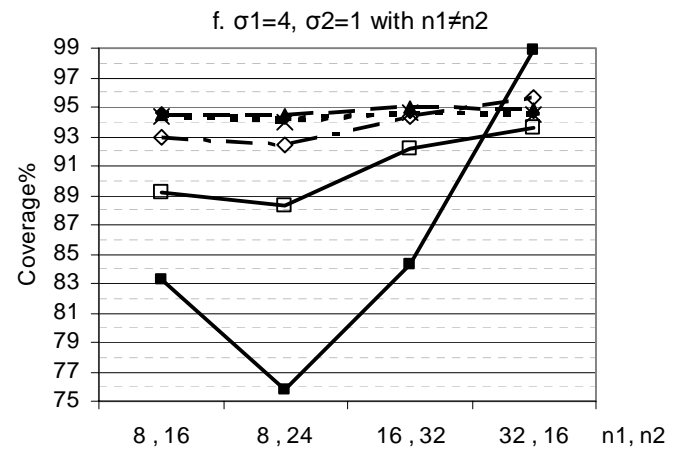
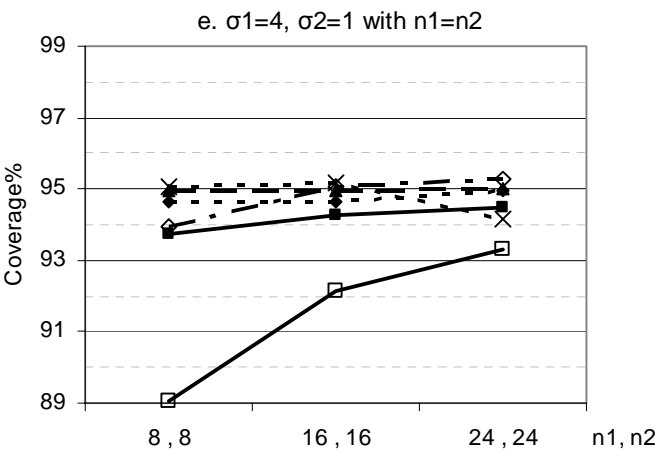
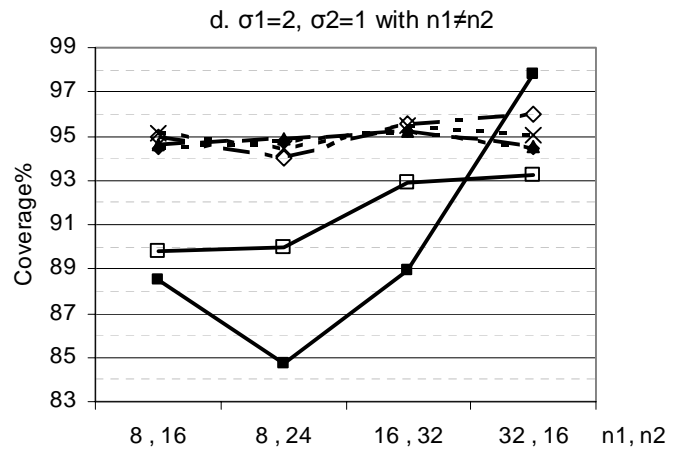
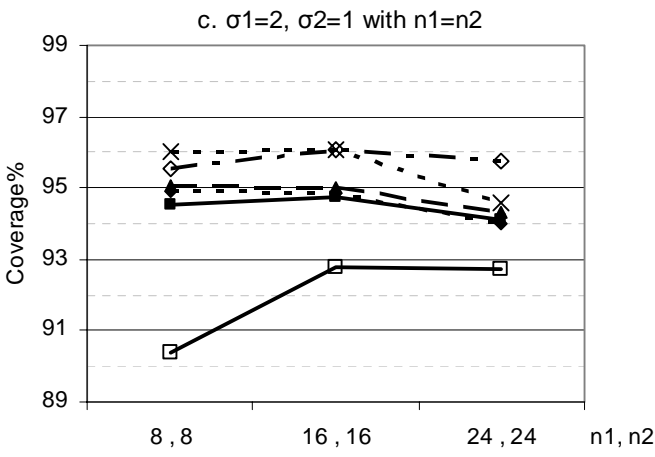
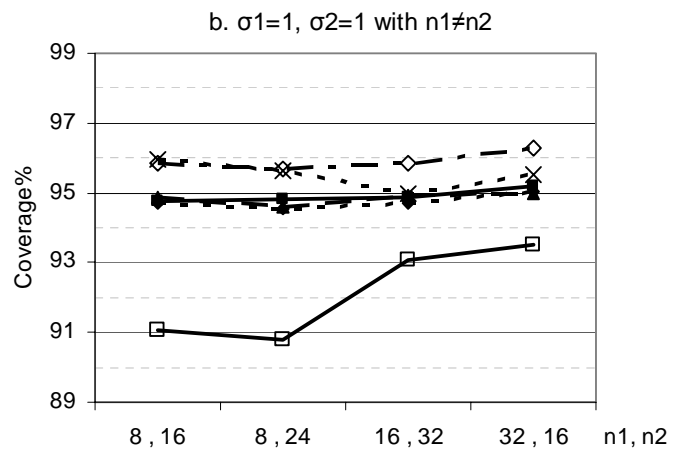
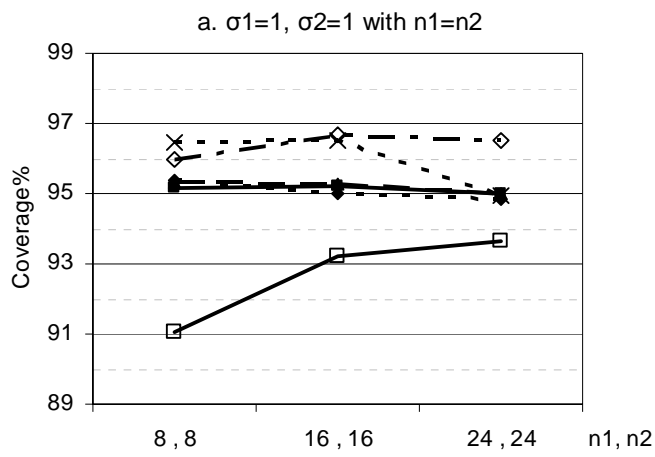
In short, when the populations are normally distributed, the Student's  $t$  procedure performs very badly when unequal population standard deviations are accompanied with unequal sample sizes. The Welch and the bootstrap-Welch procedure are robust with respect to differences in sample sizes and population standard deviations, and generally perform well. The bootstrap percentile mean is not a good alternative, because it has bad coverage percentages in almost all conditions. The bootstrap percentile median and the B&P procedure perform overall rather well, although they both have a conservative bias when the sample sizes and population standard deviations are equal.

**Table 3.1:**

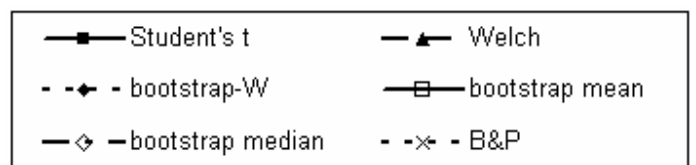
Coverage percentages\* and mean confidence interval widths (in parentheses) for Study 1 with normal distributions

$\sigma_2 = 1$ $\mu_1 = 0.0$	$n_1$	$n_2$	Student's $t$	Welch	bootstrap-W	bootstrap mean	bootstrap median	B&P	
$\sigma_1 = 1$ $\mu_2 = 0.8$	8	8	<b>95.2</b> (2.1)	95.4 (2.1)	95.4 (2.2)	91.1 (1.8)	96.0 (2.4)	96.5 (2.7)	
	16	16	95.2 (1.4)	95.3 (1.4)	<b>95.0</b> (1.4)	93.2 (1.3)	96.7 (1.7)	96.5 (1.9)	
	24	24	<b>95.0</b> (1.2)	<b>95.0</b> (1.2)	94.9 (1.2)	93.7 (1.1)	96.5 (1.4)	94.9 (1.5)	
	8	16	94.7 (1.8)	<b>94.9</b> (1.8)	94.7 (1.9)	91.1 (1.6)	95.9 (2.1)	96.0 (2.3)	
	8	24	<b>94.8</b> (1.7)	94.6 (1.7)	94.6 (1.8)	90.8 (1.5)	95.7 (2.0)	95.7 (2.1)	
	16	32	94.9 (1.2)	94.9 (1.2)	94.7 (1.2)	93.1 (1.2)	95.8 (1.5)	<b>95.0</b> (1.6)	
	32	16	95.2 (1.2)	<b>95.0</b> (1.2)	<b>95.0</b> (1.2)	93.5 (1.2)	96.3 (1.5)	95.5 (1.6)	
	$\sigma_1 = 2$ $\mu_2 = 0.8$	8	8	94.5 (3.3)	<b>95.1</b> (3.4)	<b>94.9</b> (3.6)	90.4 (2.8)	95.6 (3.7)	96.0 (4.2)
		16	16	94.8 (2.3)	<b>95.0</b> (2.3)	94.9 (2.3)	92.8 (2.1)	96.1 (2.7)	96.1 (2.9)
24		24	94.1 (1.8)	94.3 (1.8)	94.0 (1.8)	92.8 (1.7)	95.7 (2.2)	<b>94.6</b> (2.3)	
8		16	88.5 (2.5)	94.7 (3.3)	94.4 (3.5)	89.8 (2.7)	<b>95.0</b> (3.5)	95.1 (3.9)	
8		24	84.7 (2.1)	<b>94.9</b> (3.3)	94.8 (3.5)	89.9 (2.6)	94.0 (3.4)	94.4 (3.8)	
16		32	89.0 (1.7)	<b>95.2</b> (2.2)	<b>95.2</b> (2.2)	92.9 (2.0)	95.6 (2.6)	95.5 (2.8)	
32		16	97.8 (2.1)	94.6 (1.7)	94.5 (1.7)	93.3 (1.7)	96.0 (2.1)	<b>95.0</b> (2.2)	
$\sigma_1 = 4$ $\mu_2 = 0.8$		8	8	93.7 (6.1)	<b>95.0</b> (6.5)	94.6 (7.1)	89.1 (5.2)	94.0 (6.8)	95.1 (7.7)
	16	16	94.3 (4.1)	<b>95.0</b> (4.3)	94.6 (4.4)	92.1 (3.8)	95.1 (4.9)	95.2 (5.4)	
	24	24	94.5 (3.4)	<b>95.0</b> (3.4)	94.9 (3.4)	93.3 (3.2)	95.3 (4.1)	94.2 (4.2)	
	8	16	83.2 (4.2)	94.4 (6.5)	<b>94.6</b> (7.1)	89.2 (5.1)	93.0 (6.6)	94.3 (7.5)	
	8	24	75.7 (3.4)	<b>94.5</b> (6.5)	94.2 (7.1)	88.3 (5.0)	92.4 (6.5)	94.0 (7.4)	
	16	32	84.2 (2.9)	<b>95.0</b> (4.2)	94.6 (4.3)	92.1 (3.8)	94.3 (4.8)	94.6 (5.2)	
	32	16	98.9 (4.1)	<b>94.9</b> (3.0)	94.7 (3.0)	93.6 (2.9)	95.7 (3.7)	94.5 (3.8)	

\* For each condition the best coverage percentages are set in bold.



**Figure 3.1:**  
Coverage percentages of six procedures for comparing two groups with equal and unequal population standard deviations and sample sizes in case of normal distributions



### **§3.1.2 Non-normal distributions**

The six procedures will be discussed for each of the four non-normal distributions as displayed in Figures 3.2 through 3.5. In Table 3.2 the proportion of 'bad' coverage percentages of the six procedures are shown for each distribution, sample size condition, and population standard deviation condition. For each non-normal distribution the procedure's coverage percentages are compared to the coverage percentages in case of normal distributions.

#### **§3.1.2.1 Leptokurtic distributions**

For leptokurtic distributions the Student's  $t$  procedure performs similarly as with normal distributions. In case of  $\sigma_1=\sigma_2$  and/or  $n_1=n_2$  the coverage percentages are still good and the effects of negative and positive pairing are still present. The Welch procedure also performs similarly as with normal distributions; in most conditions the coverage percentages are very good. The Student's  $t$  and the Welch procedure have the same proportion of 'bad' coverage percentages as with normal distributions (see Table 3.2), although on average the coverage percentages are about 0.6% higher (not reported in a table). The bootstrap-Welch coverage percentages in case of leptokurtic distributions differ from those for normal distributions: the coverage percentages are always below 94% (see Table 3.2). The bootstrap percentile mean and the bootstrap percentile median give very similar coverage percentages for leptokurtic distributions and normal distributions. The B&P procedure has on average 0.9% higher coverage percentages (between 94.7%-97.3%) in case of leptokurtic distributions compared to normal distributions, which also is reflected by the higher proportion of 'bad' coverage percentages.

To sum up, when the two samples are from leptokurtic distributions and  $\sigma_1=\sigma_2$  and/or  $n_1=n_2$  the Student's  $t$  procedure is most of the times the best procedure but the Welch procedure performs almost as well. In the cases  $\sigma_1/\sigma_2=2$  with  $n_1\neq n_2$  the Welch procedure is overall the best procedure and when  $\sigma_1/\sigma_2=4$  with  $n_1\neq n_2$  the B&P procedure is overall the best procedure with the Welch procedure performing almost as well.

### §3.1.2.2 Platykurtic distributions

For platykurtic distributions the Student's  $t$  procedure performs similarly as with normal distributions. In case of  $\sigma_1=\sigma_2$  and/or  $n_1=n_2$  the coverage percentages are still good and the effects of negative and positive pairing are still present. The Welch procedure also performs similarly for both distributions; in most conditions the coverage percentages are very good. The Student's  $t$  and the Welch procedure have the same proportion of 'bad' coverage percentages in comparison with normal distributions (see Table 3.2). Here, the coverage percentages for the Student's  $t$  and the Welch procedure are on average only about 0.2% lower compared to those for normal distributions. For the bootstrap-Welch platykurtic distributions lead to higher coverage percentages than normal distributions: the coverage percentages are always above 95.6%. The bootstrap percentile mean and the bootstrap percentile median have very similar coverage percentages for platykurtic and normal distributions. The B&P procedure has on average 1.9% lower coverage percentages (between 91.6%-94.4%) in case of platykurtic distributions, which also is reflected by the higher proportion of 'bad' coverage percentages.

To sum up, when the two samples are from platykurtic distributions and  $\sigma_1=\sigma_2$  the Student's  $t$  procedure is most of the times the best procedure but the Welch procedure is almost as good. In the cases  $\sigma_1\neq\sigma_2$  the Welch procedure is overall the best procedure.

### §3.1.2.3 Moderately skewed distributions

For moderately skewed distributions all six procedures have about the same coverage percentages as for normal distributions, which also can be seen in Table 3.2. When the two samples are from moderately skewed distributions and  $\sigma_1=\sigma_2$ , or  $\sigma_1/\sigma_2=2$  with  $n_1=n_2$ , the Welch procedure is on average the best procedure with the Student's  $t$  procedure performing almost as good. In the cases  $\sigma_1/\sigma_2=2$  with  $n_1\neq n_2$  or  $\sigma_1/\sigma_2=4$  with the smallest  $n=8$ , the B&P procedure is on average the best procedure and the cases  $\sigma_1/\sigma_2=4$  with the smallest  $n\geq 16$ , the bootstrap percentile median is the best procedure, but the Welch procedure also performs well in those conditions. Overall the Welch and the bootstrap-Welch procedures give the fewest bad coverage percentages.

### §3.1.2.4 Extremely skewed distributions

For extremely skewed distributions the Student's  $t$  procedure has rather different coverage percentages compared to those for normal distributions. When  $\sigma_1=\sigma_2$  the coverage percentages are still good, but when  $\sigma_1\neq\sigma_2$  with  $n_1=n_2$  the coverage percentages are below 94% and decrease with increasing difference in population standard deviations and sample sizes. The effects of negative and positive pairing are still present. The Welch procedure has also different coverage percentages in case of extremely skewed distributions compared to normal distributions. When  $\sigma_1=\sigma_2$  the coverage percentages are acceptable, but when  $\sigma_1\neq\sigma_2$  the coverage percentages are below 94% and decrease with increasing difference in population standard deviations and sample sizes. When  $\sigma_1<\sigma_2$  with  $n_1=32$  and  $n_2=16$  the Welch procedure performs well. The Student's  $t$  and the Welch procedure have clearly higher proportions of 'bad' coverage for extremely skewed distributions (see Table 3.2) than for normal distributions. The bootstrap-Welch has here also different coverage percentages: the coverage percentages are always below 94%, with a remarkable result, that the coverage percentages with  $n_1\neq n_2$  are better when  $\sigma_1/\sigma_2=4$  than when  $\sigma_1/\sigma_2=2$  or  $\sigma_1=\sigma_2$ . The bootstrap percentile mean procedure performs worse than with the normal distributions with worsening coverage percentages with increasing difference in population standard deviations. The bootstrap percentile median and the B&P procedure have about similar coverage percentages compared with normal distributions.

To sum up, when the two samples are from extremely skewed distributions and  $\sigma_1=\sigma_2$  the Student's  $t$  procedure is on average the best procedure, but the Welch procedure also performs acceptable. In the cases  $\sigma_1/\sigma_2=2$  with  $n_1=n_2\leq 16$ , the bootstrap percentile median is the best procedure and in the cases  $\sigma_1/\sigma_2=2$  with  $n_1=n_2=24$  or in the cases  $\sigma_1/\sigma_2=4$ , the B&P procedure is on average the best procedure. Overall, the B&P procedure gives fewest bad coverage percentages. Of the procedures comparing means the Welch procedure gives the fewest bad coverage percentages.

### §3.1.3 Power properties

The power of a procedure increases when the confidence interval becomes narrower with the coverage percentage remaining the same for a condition. When for all distributions

the sample size conditions with  $n_1=n_2=8$  are compared to the conditions with  $n_1=8, n_2=16$  and  $n_1=8, n_2=24$ , it can be seen that for all procedures the confidence intervals get narrower with increasing total sample size when  $\sigma_1=\sigma_2$  (e.g. see Table 3.1). For the conditions with  $\sigma_1\neq\sigma_2$  the confidence intervals of all procedures, except of the Student's  $t$  procedure, are about equally wide or only a little smaller for the conditions with  $n_1=8, n_2=16$  and  $n_1=8, n_2=24$ . The Student's  $t$  procedure does get considerably smaller confidence intervals, but also has worse coverage percentages. Only for the Welch procedure the coverage percentages remain close to 95% for all conditions in all distributions, except for extremely skewed distributions. So when  $\sigma_1=\sigma_2$ , the Student's  $t$  and the Welch procedure show the same coverage percentages with smaller confidence intervals (higher power) when total sample size increases, even when sample sizes are unequal. But when  $\sigma_1\neq\sigma_2$  and the distribution is not extremely skewed, only the Welch procedure has about the same power when total sample size increases, even when sample sizes are unequal. The Student's  $t$  has in the conditions with  $\sigma_1\neq\sigma_2$  and  $n_1\neq n_2$  much worse coverage percentages.

### §3.1.4 Study 1 summarized

To summarize Study 1, it has been found that the Student's  $t$  procedure performs very well when  $\sigma_1=\sigma_2$  and  $n_1=n_2$  with all different distributions. When  $\sigma_1\neq\sigma_2$  with  $n_1=n_2=16$  or  $n_1=n_2=24$  the coverage percentages are good for all distributions except for extremely skewed distributions. The Student's  $t$  procedure has very low coverage percentages in case of  $\sigma_1\neq\sigma_2$  in combination with  $n_1\neq n_2$ . This can be seen in Table 3.2 under 'pairing-effect'. In case of  $\sigma_1=\sigma_2$  with  $n_1\neq n_2$  all coverage percentages are good and in case of negative ( $\sigma_1>\sigma_2$  with  $n_1<n_2$ ) or positive pairing ( $\sigma_1>\sigma_2$  with  $n_1>n_2$ ) all coverage percentages are 'bad'. This trend is not shown by the alternative procedures.

In cases where the Student's  $t$  procedure performs well, the Welch procedure performs as well. The Welch procedure performs close to 95% in all conditions for all distributions, except for extremely skewed distributions. Surprisingly, the bootstrap-Welch procedure is not robust to non-normal distributions. The coverage percentages are too liberal for leptokurtic and extremely skewed distributions and too conservative for

platykurtic distributions. The bootstrap percentile mean procedure has for all distributions too liberal coverage percentages.

The bootstrap percentile median procedure has for all distributions coverage percentages that are a little conservative. The B&P procedure is mostly influenced by the kurtosis of the distributions, for leptokurtic distributions the coverage percentages are too conservative and for platykurtic distributions too liberal.

If one would know the population distributions, as was the case in this study, the conclusion would be that in case of  $\sigma_1=\sigma_2$  and  $n_1=n_2$  the Student's  $t$  procedure is the best procedure to use, while the Welch procedure is satisfying as well. In case of  $\sigma_1\neq\sigma_2$  and/or  $n_1\neq n_2$  the Welch procedure is the best alternative, but when the distributions are extremely skewed the Welch procedure also gives 'bad' coverage percentages.

In case of symmetric distributions (zero skewness) the comparison between the mean based and the median based procedures is useful, since the mean and the median are the same. When both populations are from the same symmetric distribution in terms of skewness and kurtosis, the Welch procedure is the best alternative.

In case of asymmetric distributions the mean and median are different. The researcher must make a fundamental choice between using the mean or the median. If the researcher chooses to use the mean, than the Welch procedure is the best choice. If the researcher chooses to use the median, than the B&P procedure is the best choice.

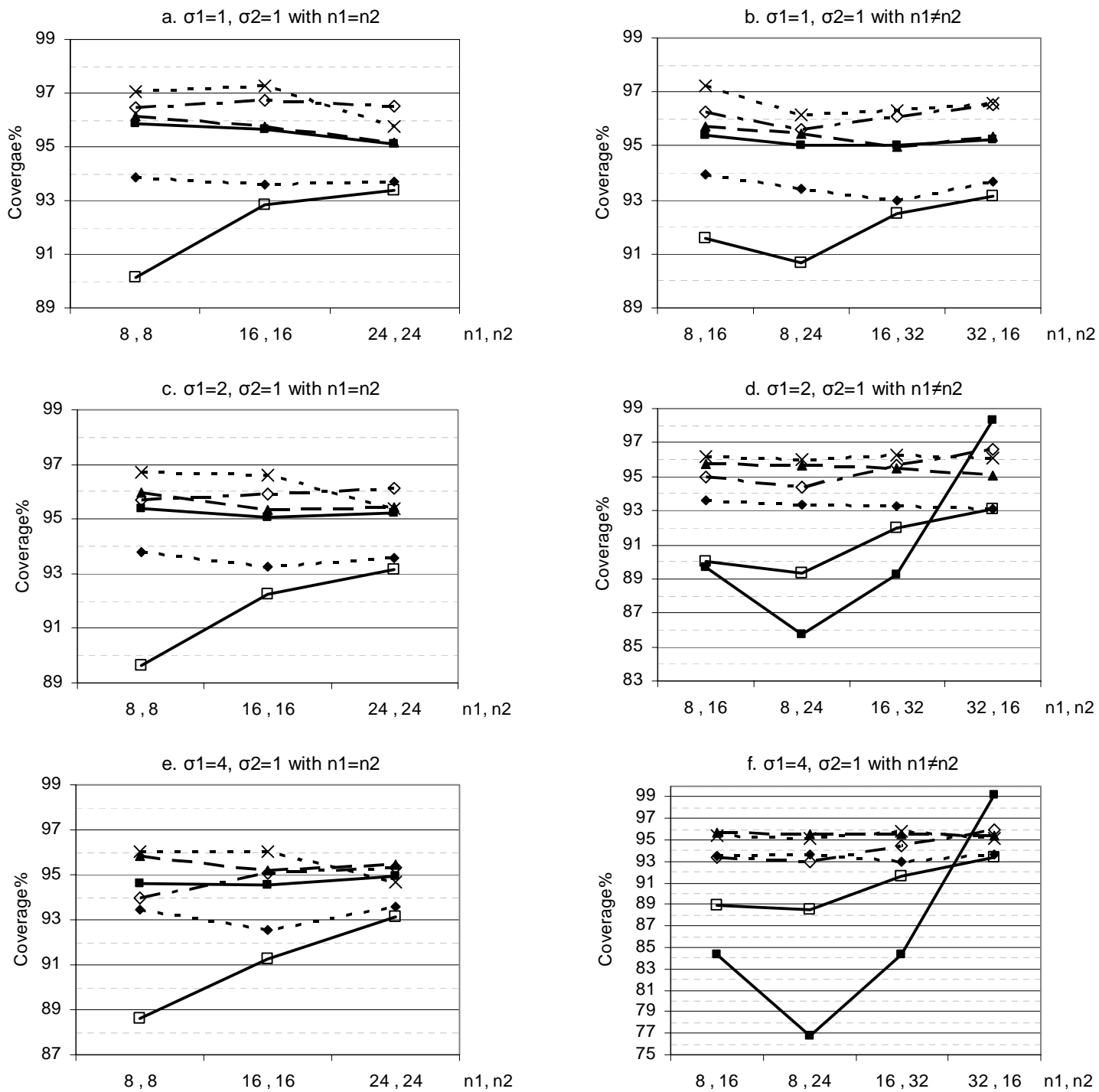
Since in experimental research the distributions of the populations are unknown, a procedure is needed that is not much influenced by unequal population standard deviations or non-normality. The Welch procedure is the best procedure in that case because it is robust to unequal population standard deviations and robust (to a great extent) to non-normality (when the two samples are from the same non-normal distribution in terms of skewness and kurtosis).

**Table 3.2:**

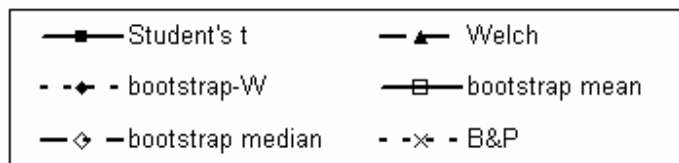
Proportion of coverage percentages smaller than 94% or greater than 96% for Study 1

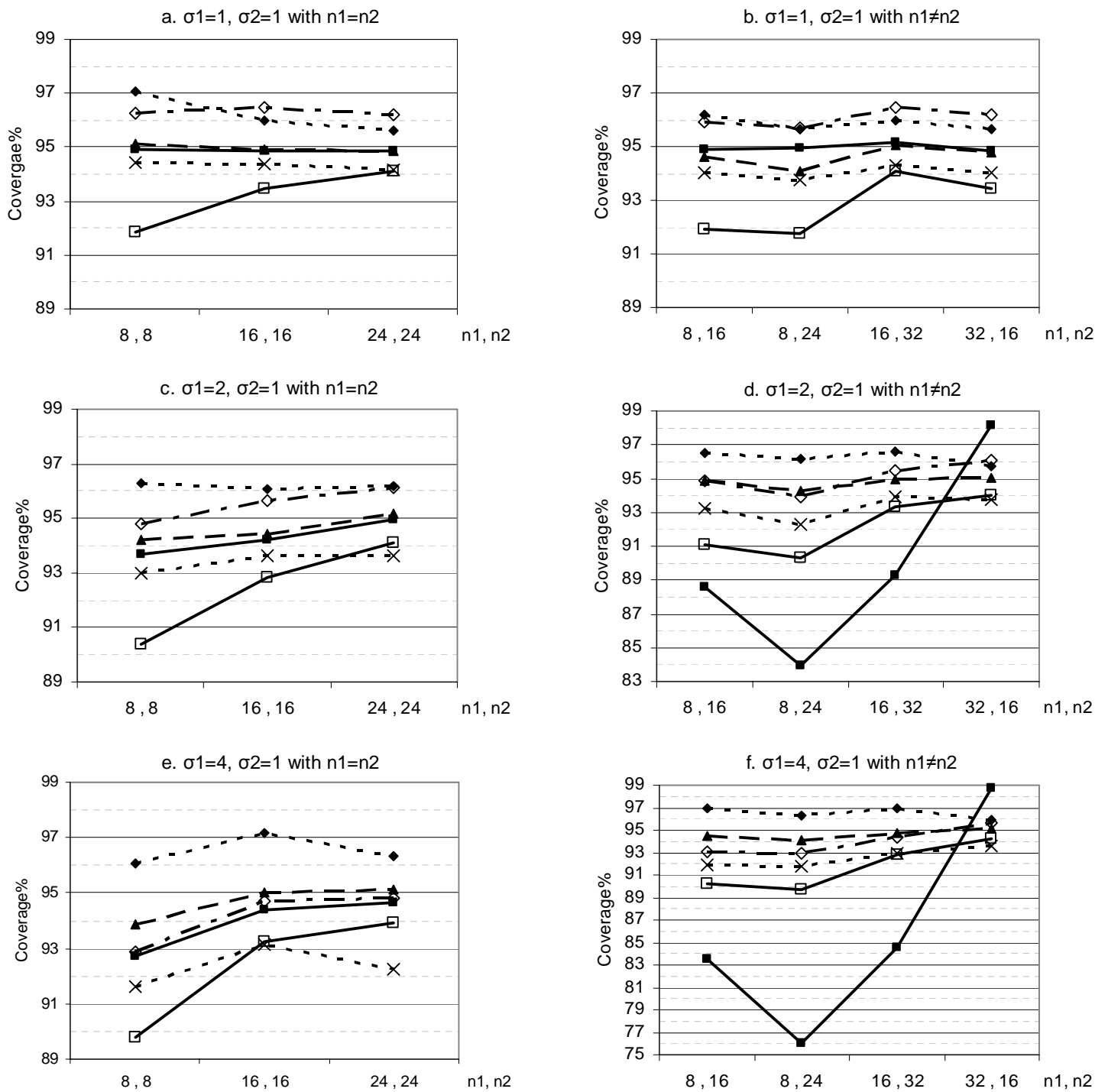
	Student's <i>t</i>	Welch	bootstrap-W	bootstrap mean	bootstrap median	B&P
<b>Distribution</b>						
Normal	0.4	0.0	0.0	1.0	0.4	0.2
Leptokurtic	0.4	0.0	1.0	1.0	0.5	0.6
Platykurtic	0.5	0.0	0.7	0.8	0.5	0.7
Moderate skew	0.4	0.1	0.1	1.0	0.4	0.4
Extreme skew	0.7	0.6	1.0	1.0	0.5	0.3
<b>Sample size</b>						
$n_1=8$ $n_2=8$	0.4	0.3	0.6	1.0	0.5	0.7
$n_1=16$ $n_2=16$	0.1	0.1	0.5	1.0	0.5	0.7
$n_1=24$ $n_2=24$	0.1	0.1	0.5	0.9	0.5	0.2
$n_1=8$ $n_2=16$	0.7	0.2	0.7	1.0	0.4	0.5
$n_1=8$ $n_2=24$	0.7	0.3	0.6	1.0	0.4	0.4
$n_1=16$ $n_2=32$	0.7	0.1	0.5	0.9	0.2	0.3
$n_1=32$ $n_2=16$	0.7	0.0	0.4	0.9	0.7	0.4
<b>Std. Dev.</b>						
$\sigma_1 = 1$ $\sigma_2 = 1$	0.0	0.1	0.5	0.9	0.6	0.5
$\sigma_1 = 2$ $\sigma_2 = 1$	0.7	0.1	0.6	0.9	0.3	0.5
$\sigma_1 = 4$ $\sigma_2 = 1$	0.7	0.3	0.6	1.0	0.4	0.4
<b>Average</b>	<b>0.5</b>	<b>0.2</b>	<b>0.6</b>	<b>1.0</b>	<b>0.5</b>	<b>0.4</b>
<b>Pairing-effect</b>						
$\sigma_1=\sigma_2$ and $n_1 \neq n_2$ *	0.0	0.0	0.4	0.9	0.8	0.5
negative pairing**	1.0	0.2	0.6	1.0	0.0	0.3
positive pairing***	1.0	0.0	0.4	0.8	0.5	0.3

\*  $\sigma_1=1, \sigma_2=1$  with  $n_1=16, n_2=32$  and  $n_1=32, n_2=16$ \*\*  $\sigma_1=2, \sigma_2=1$  and  $\sigma_1=4, \sigma_2=1$  with  $n_1=16, n_2=32$ \*\*\*  $\sigma_1=2, \sigma_2=1$  and  $\sigma_1=4, \sigma_2=1$  with  $n_1=32, n_2=16$

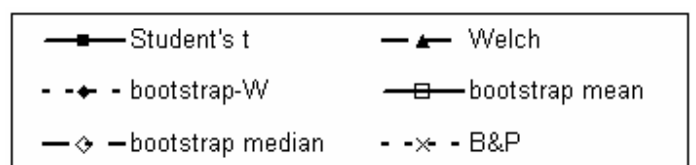


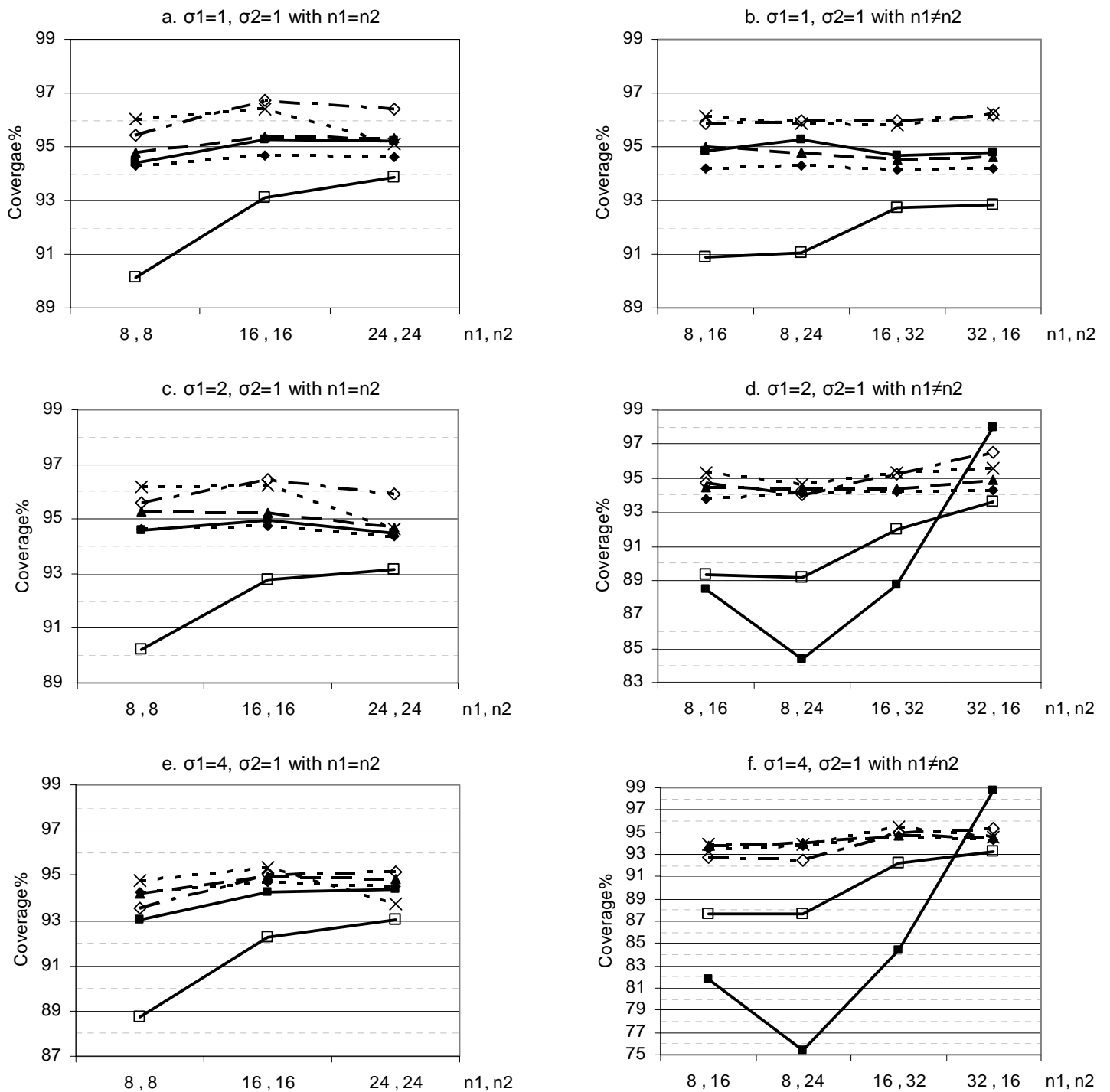
**Figure 3.2:**  
Coverage percentages of six procedures for comparing two groups with equal and unequal population standard deviations and sample sizes in case of leptokurtic distributions



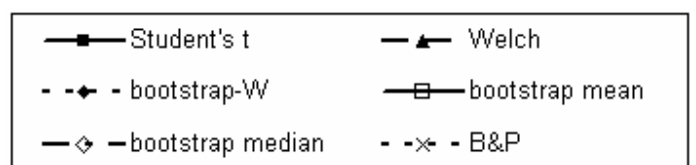


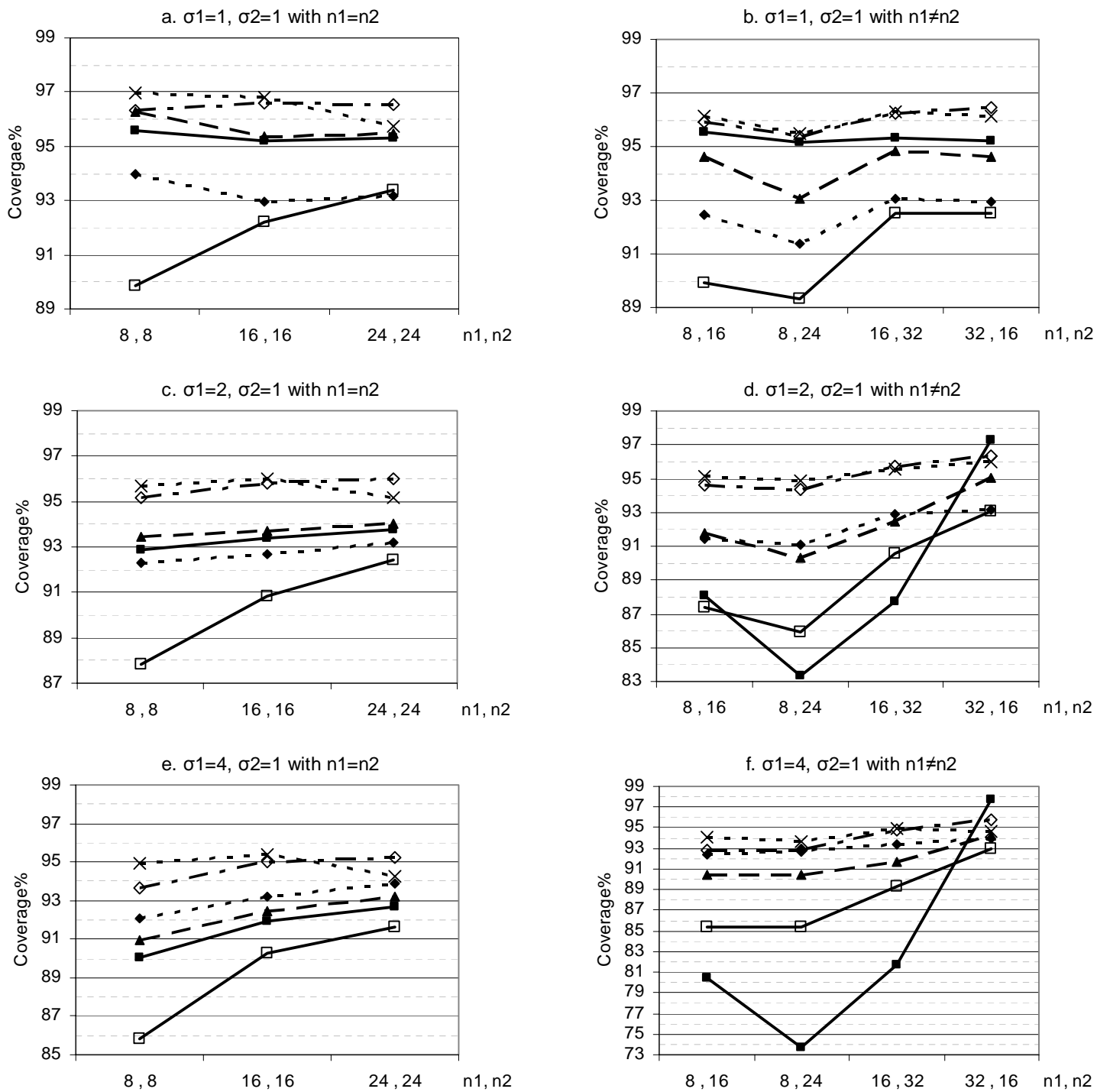
**Figure 3.3:** Coverage percentages of six procedures for comparing two groups with equal and unequal population standard deviations and sample sizes in case of platykurtic distributions



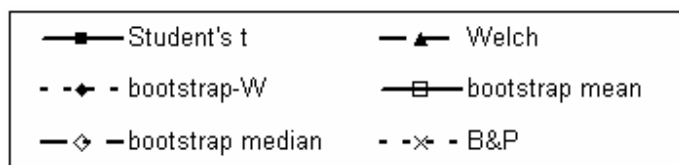


**Figure 3.4:** Coverage percentages of six procedures for comparing two groups with equal and unequal population standard deviations and sample sizes in case of moderately skewed distributions





**Figure 3.5:** Coverage percentages of six procedures for comparing two groups with equal and unequal population standard deviations and sample sizes in case of extremely skewed distributions



### **§3.2 Study 2: Contrast analysis with four samples from simulated distributions**

In this study six procedures for contrast analysis with four groups are compared. The six procedures are generalizations of those in Study 1. The Student's  $t$  procedure and the Welch procedure are actually, respectively, an ANOVA procedure and a Satterthwaite procedure for two samples with contrast weights -1 and 1. The pattern of differences for Study 2 is very similar to Study 1. Therefore the results are discussed only briefly.

In Table 3.3 the coverage percentages and the interval widths for the six procedures are shown for the case of normal population distributions. These coverage percentages are displayed graphically in Figure 3.6. In Figures 3.7 to 3.10 the coverage percentages for the four types of non-normal population distributions are shown. In Table 3.4 the proportion of 'bad' coverage percentages of the procedures are shown for each distribution, sample size condition, and population standard deviation condition.

The ANOVA procedure performs very well in case of equal population standard deviations with equal sample sizes and has poor coverage percentages when unequal population standard deviations are accompanied with unequal sample sizes (see Table 3.3 and Table 3.4). In cases where the ANOVA procedure performs well, the Satterthwaite procedure performs almost as well or sometimes better (see Table 3.3). Moreover, the Satterthwaite procedure performs very well in all conditions for all distributions except for extremely skewed distributions, but even in that case the Satterthwaite procedure is on average closest to 95% compared to the other procedures (see Table 3.4). The bootstrap-Satterthwaite procedure is not robust to non-normal distributions. The coverage percentages are a little liberal for leptokurtic and extremely skewed distributions and a little conservative for platykurtic distributions. The bootstrap-Satterthwaite procedure does perform well, however, when both populations are from a normal or moderately skewed distribution. The bootstrap percentile mean procedure has in almost all conditions coverage percentages that are below 94%. The bootstrap percentile median procedure has for all distributions coverage percentages that are conservative when the population standard deviations are equal and the coverage percentages get closer to 95% with increasing differences in population standard deviations. The B&P procedure is in general too conservative except for platykurtic distributions: the B&P procedure is in that case a little liberal.

**Table 3.3**

Coverage percentages\* and mean confidence interval widths (in parentheses) for Study 2 with normal distributions

	$n_1$	$n_2$	$n_3$	$n_4$	ANOVA	Satterthwaite	bootstrap-S	bootstrap mean	bootstrap median	B&P
$\sigma_1=1$	8	8	8	8	94.8 (2.9)	<b>95.1</b> (2.9)	<b>95.1</b> (2.9)	92.0 (2.6)	97.4 (3.5)	97.5 (3.8)
	16	16	16	16	<b>95.0</b> (2.0)	<b>95.0</b> (2.0)	94.8 (2.0)	93.5 (1.9)	97.4 (2.5)	97.1 (2.7)
	24	24	24	24	94.7 (1.6)	<b>94.8</b> (1.6)	94.6 (1.6)	93.8 (1.6)	96.8 (2.1)	95.7 (2.1)
$\sigma_2=1$	8	10	12	14	95.3 (2.5)	95.3 (2.5)	<b>95.0</b> (2.5)	92.8 (2.3)	97.2 (3.1)	97.2 (3.3)
	16	20	24	28	<b>95.1</b> (1.7)	<b>95.1</b> (1.7)	94.8 (1.7)	94.0 (1.7)	97.4 (2.2)	96.8 (2.3)
	8	8	8	24	95.3 (2.6)	95.4 (2.7)	<b>95.2</b> (2.7)	92.4 (2.3)	97.2 (3.2)	97.2 (3.5)
$\sigma_3=1$	24	8	8	8	<b>95.0</b> (2.6)	95.1 (2.7)	<b>95.0</b> (2.7)	92.3 (2.4)	97.1 (3.2)	96.9 (3.5)
	8	8	8	8	93.4 (6.2)	94.6 (6.7)	94.4 (7.0)	90.4 (5.5)	<b>95.3</b> (7.3)	95.7 (8.2)
	16	16	16	16	94.1 (4.3)	<b>94.8</b> (4.5)	94.6 (4.5)	92.6 (4.1)	96.0 (5.3)	96.0 (5.7)
$\sigma_4=1$	24	24	24	24	94.6 (3.5)	<b>95.1</b> (3.6)	<b>94.9</b> (3.6)	93.6 (3.4)	96.1 (4.4)	94.9 (4.5)
	8	10	12	14	85.8 (4.6)	94.8 (6.6)	94.5 (6.9)	90.0 (5.3)	<b>95.0</b> (7.0)	95.4 (7.9)
	16	20	24	28	87.1 (3.3)	95.3 (4.4)	<b>95.1</b> (4.4)	92.8 (4.0)	95.4 (5.1)	95.6 (5.5)
$\sigma_2=3$	8	8	8	24	86.6 (4.7)	94.6 (6.6)	94.4 (7.0)	89.8 (5.4)	<b>95.0</b> (7.1)	95.4 (8.0)
	24	8	8	8	99.8 (7.6)	<b>94.7</b> (4.1)	94.6 (4.1)	93.0 (3.8)	96.7 (5.0)	96.4 (5.3)
	8	8	8	8	94.7 (7.8)	95.3 (8.1)	<b>95.1</b> (8.2)	91.4 (7.0)	97.0 (9.4)	97.0 (10.5)
$\sigma_3=2$	16	16	16	16	94.6 (5.4)	<b>94.9</b> (5.5)	94.8 (5.5)	93.2 (5.2)	97.1 (6.8)	96.7 (7.3)
	24	24	24	24	94.8 (4.4)	<b>95.0</b> (4.5)	94.9 (4.5)	93.7 (4.3)	96.9 (5.6)	95.5 (5.7)
	8	10	12	14	90.0 (6.2)	94.7 (7.6)	<b>94.8</b> (7.8)	91.5 (6.6)	96.6 (8.9)	96.7 (9.8)
$\sigma_4=1$	16	20	24	28	90.5 (4.3)	<b>94.7</b> (5.2)	94.5 (5.2)	93.4 (4.9)	96.3 (6.4)	96.1 (6.7)
	8	8	8	24	85.8 (5.8)	<b>94.7</b> (8.0)	94.6 (8.2)	91.1 (6.9)	96.6 (9.3)	96.4 (10.3)
	24	8	8	8	99.1 (8.4)	<b>94.9</b> (6.4)	<b>94.9</b> (6.4)	92.2 (5.7)	97.0 (7.7)	96.7 (8.3)

\* For each condition the best coverage percentages are set in bold.

For the ANOVA procedure, the coverage percentages for the conditions with  $\sigma_1=4$ ,  $\sigma_2=3$ ,  $\sigma_3=2$ ,  $\sigma_4=1$  are clearly lower than those for conditions with  $\sigma_1=1$ ,  $\sigma_2=1$ ,  $\sigma_3=1$ ,  $\sigma_4=1$ , while for the bootstrap percentile mean procedures, the bootstrap percentile median, and the B&P procedure differences in coverage percentage are small (see Table 3.3). For the ANOVA, the bootstrap percentile mean procedures, the bootstrap percentile median, and the B&P procedure the coverage percentages for the conditions with  $\sigma_1=4$ ,  $\sigma_2=1$ ,  $\sigma_3=1$ ,  $\sigma_4=1$  are clearly smaller than those for conditions with  $\sigma_1=4$ ,  $\sigma_2=3$ ,  $\sigma_3=2$ ,  $\sigma_4=1$ . For this comparison the Satterthwaite and the bootstrap-Satterthwaite procedure only show considerable differences when comparing results for populations with extremely skewed

distributions to those from normal distributions: the coverage percentages get further away from 95% when  $\sigma_1=4, \sigma_2=1, \sigma_3=1, \sigma_4=1$ .

When for each procedure the conditions with  $n_1=8, n_2=10, n_3=12, n_4=14$  are compared to the conditions with  $n_1=16, n_2=20, n_3=24, n_4=28$ , no clear differences in coverage percentages are seen, except for the bootstrap percentile mean procedure (see Table 3.3). The coverage percentages of the bootstrap percentile mean procedure are worsening with decrease of total sample size (which is also the case when sample sizes are equal). When for each procedure the conditions with  $n_1=8, n_2=10, n_3=12, n_4=14$  are compared to the conditions with  $n_1=8, n_2=8, n_3=8, n_4=24$ , the coverage percentages are roughly equal, except for the ANOVA procedure. With the ANOVA procedure the coverage percentages are clearly worse for  $n_1=8, n_2=8, n_3=8, n_4=24$  compared to  $n_1=8, n_2=10, n_3=12, n_4=14$  in case of  $\sigma_1=4, \sigma_2=3, \sigma_3=2, \sigma_4=1$ , but the coverage percentages are rather similar in case of  $\sigma_1=4, \sigma_2=1, \sigma_3=1, \sigma_4=1$  or in case of  $\sigma_1=1, \sigma_2=1, \sigma_3=1, \sigma_4=1$ .

In general, the confidence intervals get narrower with increasing total sample size for all procedures in all distributions when the population standard deviations or sample sizes are equal (see Table 3.3 for normal distributions). In the conditions with  $\sigma_1=4, \sigma_2=1, \sigma_3=1, \sigma_4=1$  or  $\sigma_1=4, \sigma_2=3, \sigma_3=2, \sigma_4=1$  the confidence intervals for the ANOVA procedure are smaller for  $n_1=8, n_2=8, n_3=8, n_4=24$  (negative pairing) and wider (!) for  $n_1=24, n_2=8, n_3=8, n_4=8$  (positive pairing) than those for conditions with  $n_1=8, n_2=8, n_3=8, n_4=8$ , while the other procedures show a different pattern. The confidence intervals are about equally wide for  $n_1=8, n_2=8, n_3=8, n_4=24$  and smaller for  $n_1=24, n_2=8, n_3=8, n_4=8$  than those for conditions with  $n_1=8, n_2=8, n_3=8, n_4=8$ . For the ANOVA this clearly shows the relation between coverage percentages and confidence interval width: for negative pairing the confidence intervals are smaller with decreased coverage percentages and for positive pairing the confidence intervals are narrower with increased coverage percentages. Of the alternative procedures only the Satterthwaite procedure remains to have good coverage percentages for all conditions. Therefore, the power of the Satterthwaite procedure increases with increasing sample size with equal population standard deviations, with equal sample sizes or with positive pairing. In case of negative pairing the coverage percentages and confidence intervals are about equal, therefore the

power is about the same. For the Satterthwaite procedure in most cases the power improves with increasing sample sizes.

To summarize Study 2, the Satterthwaite procedure is least influenced by unequal population standard deviations, unequal sample sizes and non-normal population distributions. All coverage percentages are very close to 95% (see Table 3.4). However, the Satterthwaite procedure is too liberal when the distributions are extremely skewed and  $\sigma_1=4, \sigma_2=1, \sigma_3=1, \sigma_4=1$ .

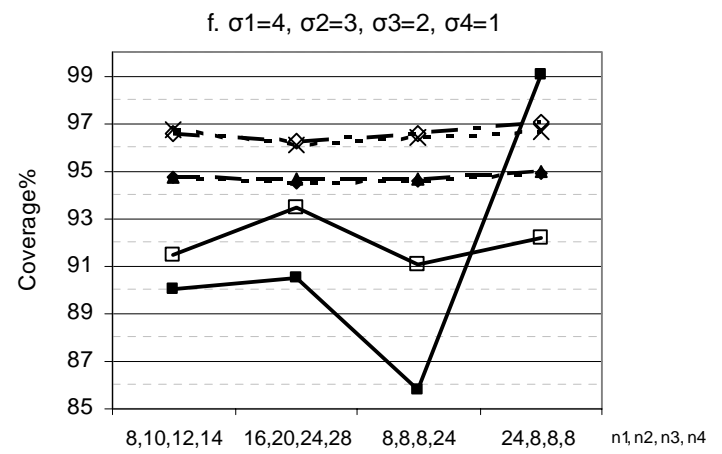
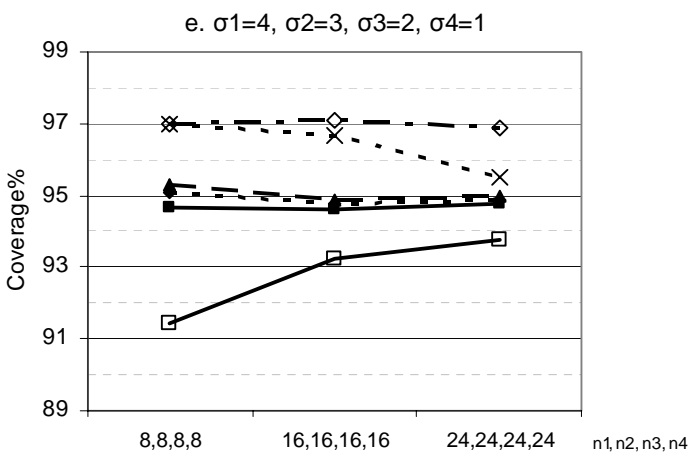
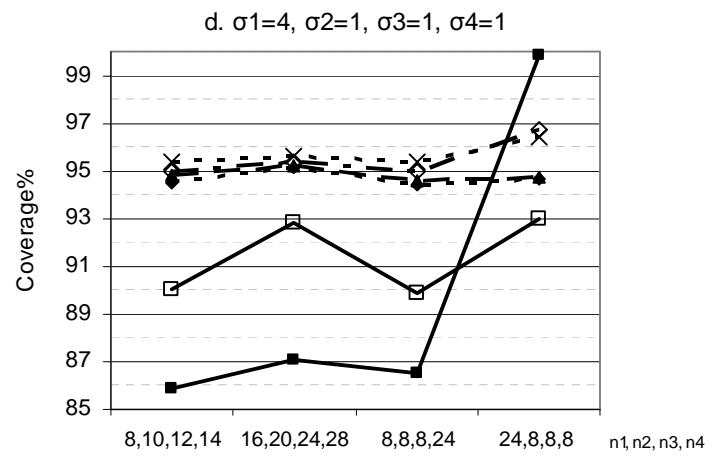
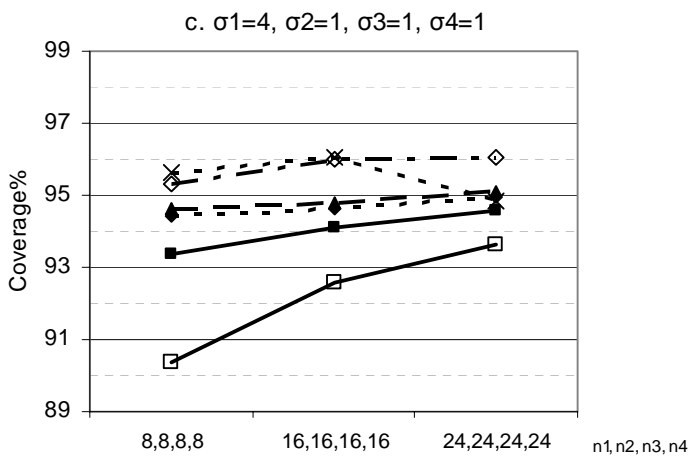
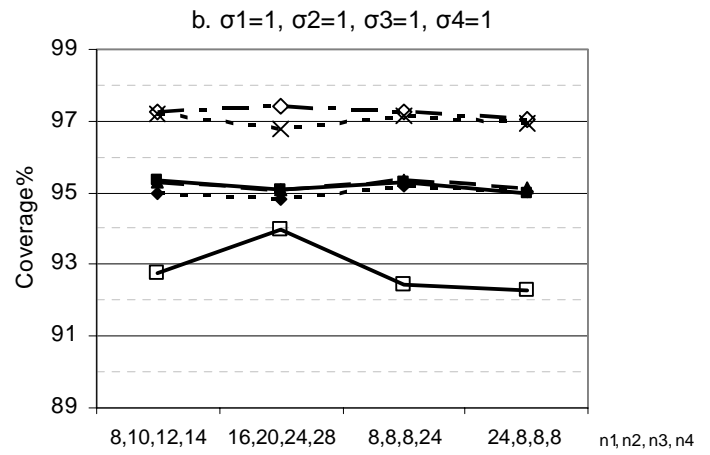
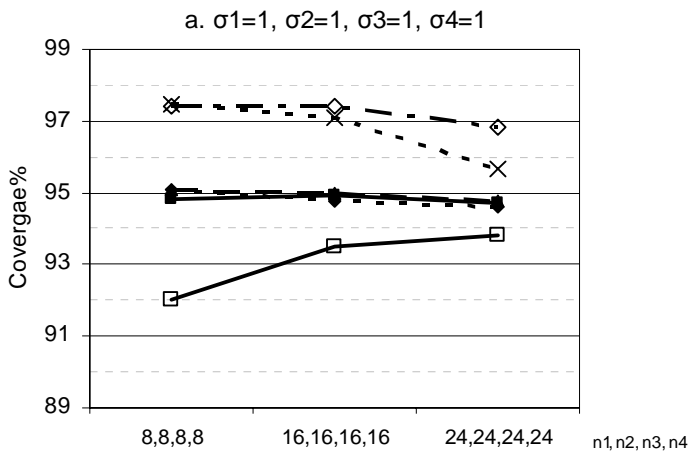
**Table 3.4:**  
Proportion of coverage percentages smaller than 94% or greater than 96% for Study 2

	ANOVA	Satterthwaite	bootstrap-S	bootstrap mean	bootstrap median	B&P
<b>Distribution</b>						
Normal	0.4	0.0	0.0	1.0	0.8	0.7
Leptokurtic	0.4	0.0	0.8	1.0	0.8	1.0
Platykurtic	0.5	0.0	0.5	0.9	0.7	0.3
Moderate skew	0.5	0.0	0.1	1.0	0.8	0.7
Extreme skew	0.5	0.3	1.0	1.0	0.8	0.8
<b>Sample size</b>						
$n_1=8 \quad n_2=8 \quad n_3=8 \quad n_4=8$	0.3	0.1	0.5	1.0	0.7	0.8
$n_1=16 \quad n_2=16 \quad n_3=16 \quad n_4=16$	0.2	0.1	0.4	1.0	0.8	0.8
$n_1=24 \quad n_2=24 \quad n_3=24 \quad n_4=24$	0.1	0.1	0.4	0.8	0.8	0.3
$n_1=8 \quad n_2=10 \quad n_3=12 \quad n_4=14$	0.7	0.1	0.5	1.0	0.7	0.7
$n_1=16 \quad n_2=20 \quad n_3=24 \quad n_4=28$	0.7	0.1	0.5	0.9	0.7	0.7
$n_1=8 \quad n_2=8 \quad n_3=8 \quad n_4=24$	0.7	0.1	0.7	1.0	0.7	0.7
$n_1=24 \quad n_2=8 \quad n_3=8 \quad n_4=8$	0.7	0.0	0.3	1.0	1.0	0.8
<b>Std. Dev.</b>						
$\sigma_1=1 \quad \sigma_2=1 \quad \sigma_3=1 \quad \sigma_4=1$	0.0	0.0	0.4	0.9	1.0	0.8
$\sigma_1=4 \quad \sigma_2=1 \quad \sigma_3=1 \quad \sigma_4=1$	0.8	0.2	0.6	1.0	0.3	0.5
$\sigma_1=4 \quad \sigma_2=3 \quad \sigma_3=2 \quad \sigma_4=1$	0.6	0.0	0.4	1.0	1.0	0.7
<b>Average</b>	0.5	0.1	0.5	1.0	0.8	0.7
<b>'Pairing-effect'</b>						
$\sigma_i=\sigma_j$ and $n_i \neq n_j^*$	0.0	0.0	0.5	1.0	1.0	0.8
negative pairing**	1.0	0.2	0.7	1.0	0.5	0.6
positive pairing***	1.0	0.0	0.3	1.0	1.0	0.8

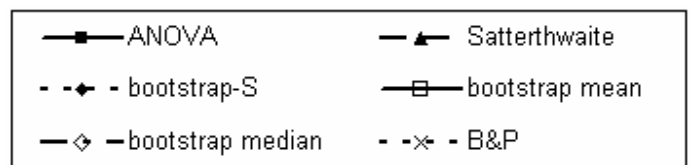
\*  $\sigma_1=1, \sigma_2=1, \sigma_3=1, \sigma_4=1$  with  $n_1=8, n_2=8, n_3=8, n_4=24$  and  $n_1=24, n_2=8, n_3=8, n_4=24$

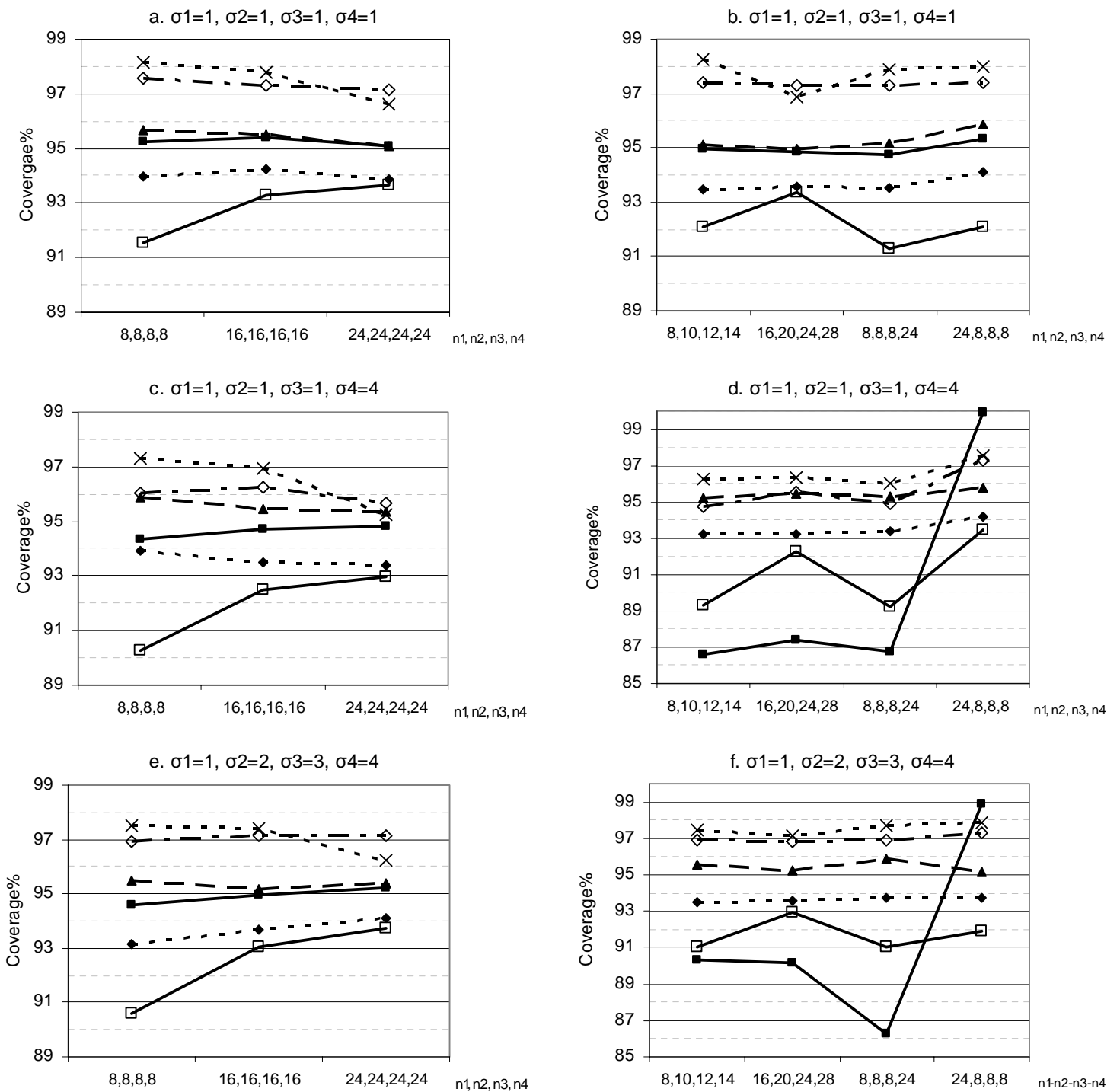
\*\*  $\sigma_1=4, \sigma_2=1, \sigma_3=1, \sigma_4=1$  and  $\sigma_1=4, \sigma_2=3, \sigma_3=2, \sigma_4=1$  with  $n_1=8, n_2=8, n_3=8, n_4=24$

\*\*\*  $\sigma_1=4, \sigma_2=1, \sigma_3=1, \sigma_4=1$  and  $\sigma_1=4, \sigma_2=3, \sigma_3=2, \sigma_4=1$  with  $n_1=24, n_2=8, n_3=8, n_4=24$

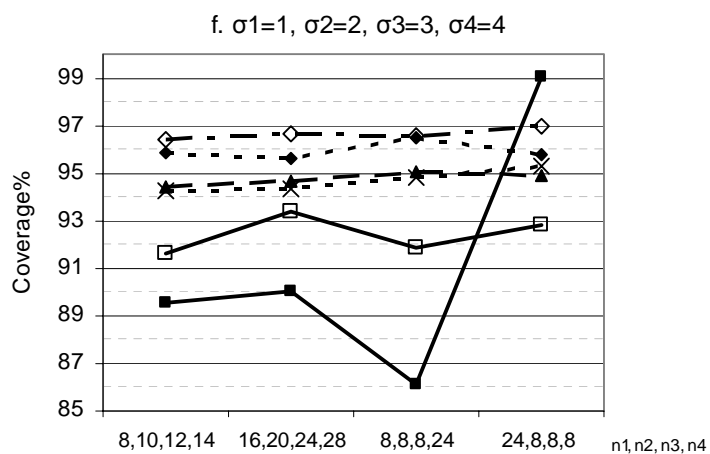
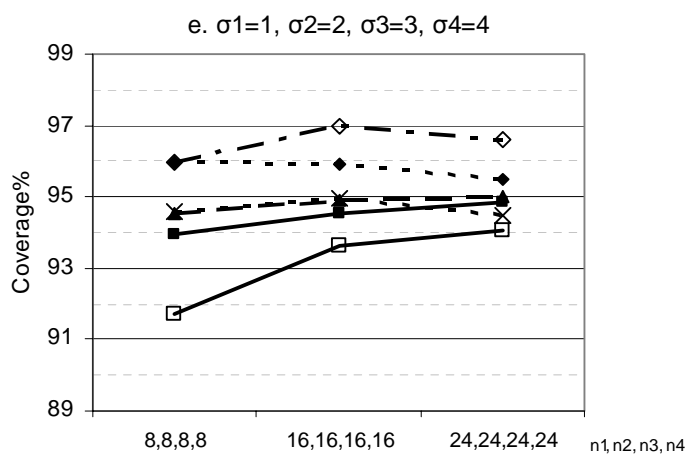
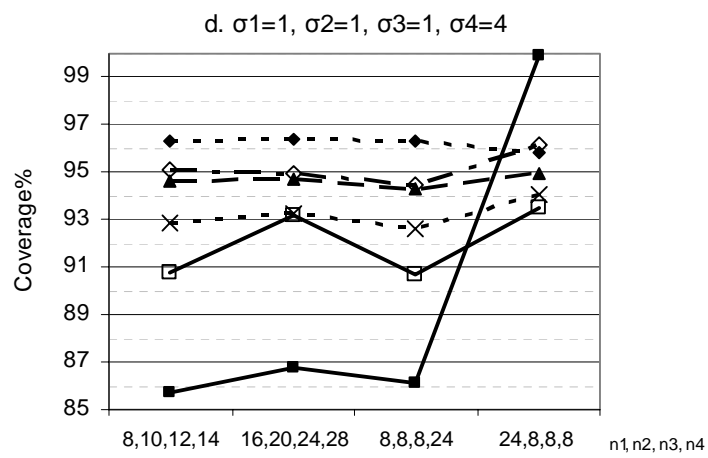
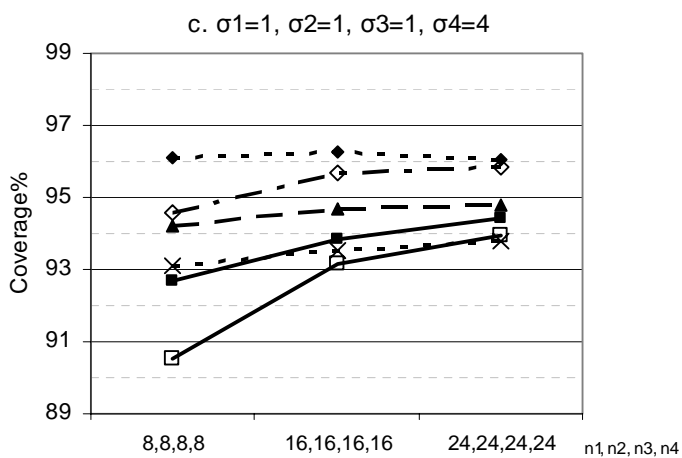
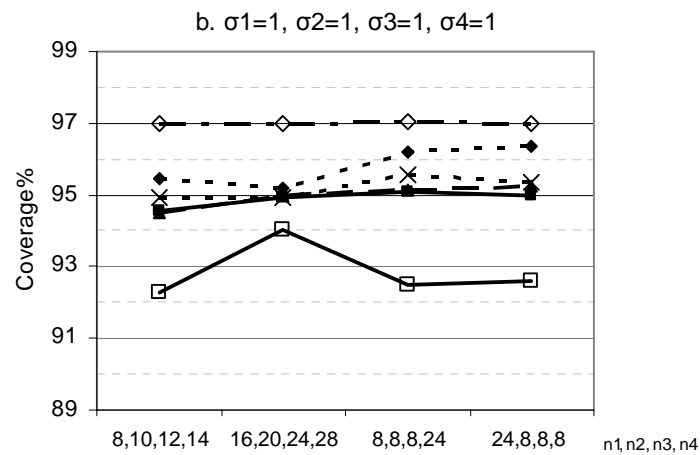
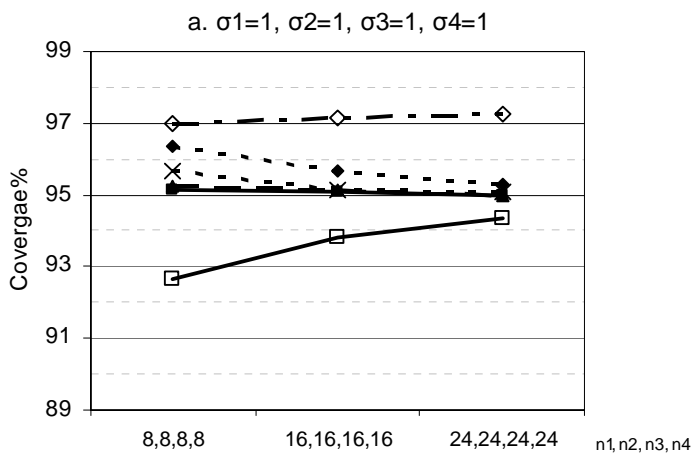


**Figure 3.6:** Coverage percentages of six procedures for contrast analysis with equal and unequal population standard deviations and sample sizes in case of normal distributions

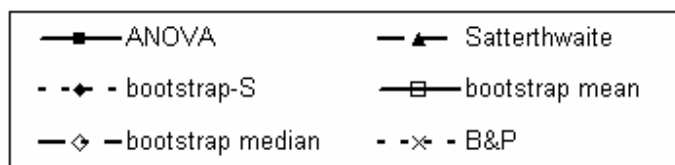


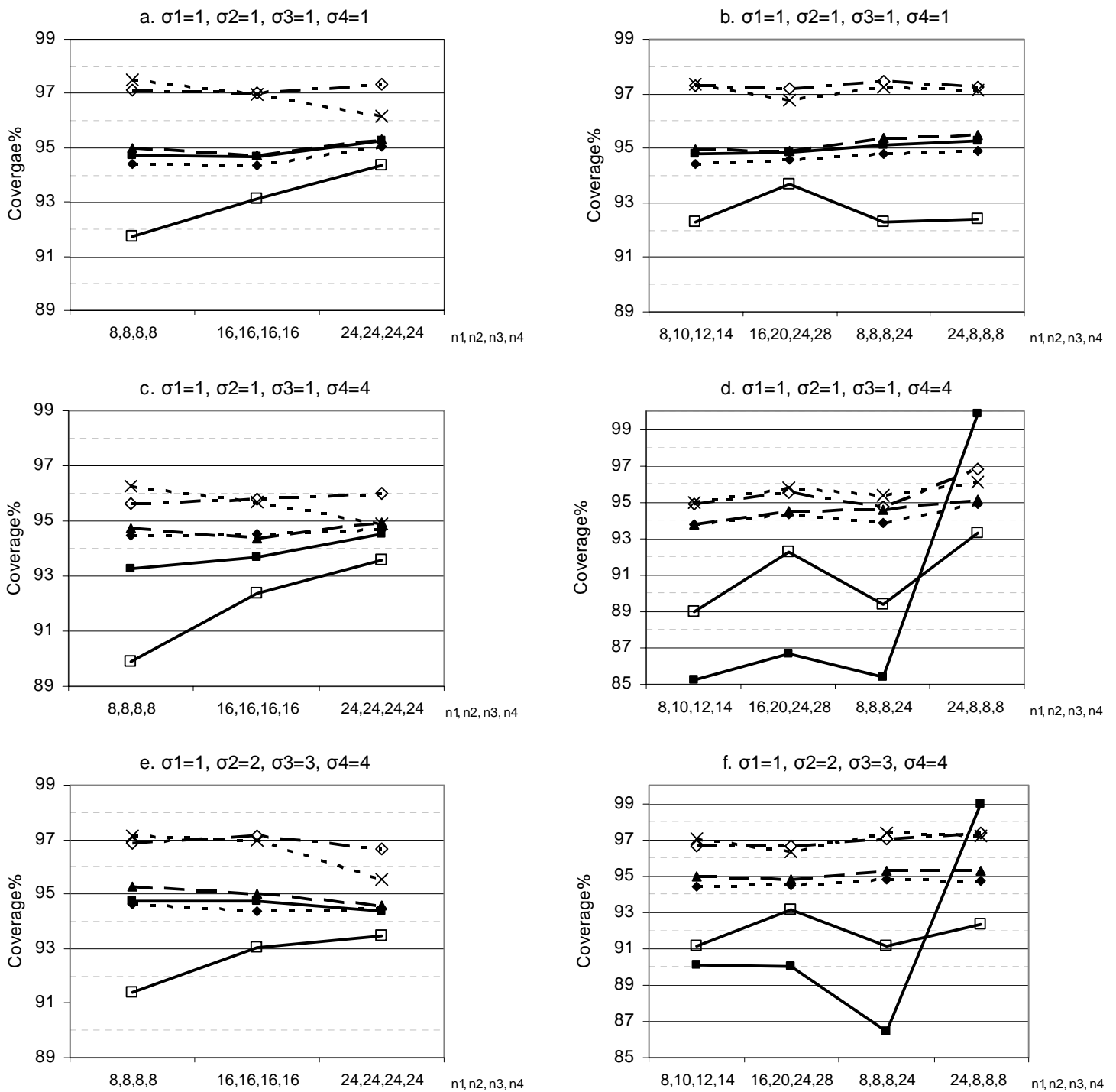


**Figure 3.7:**  
Coverage percentages of six procedures for contrast analysis with equal and unequal population standard deviations and sample sizes in case of leptokurtic distributions

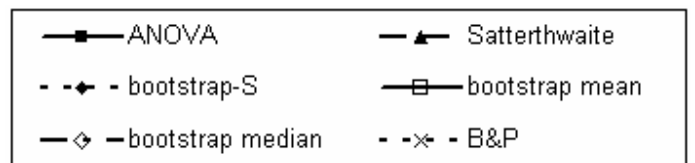


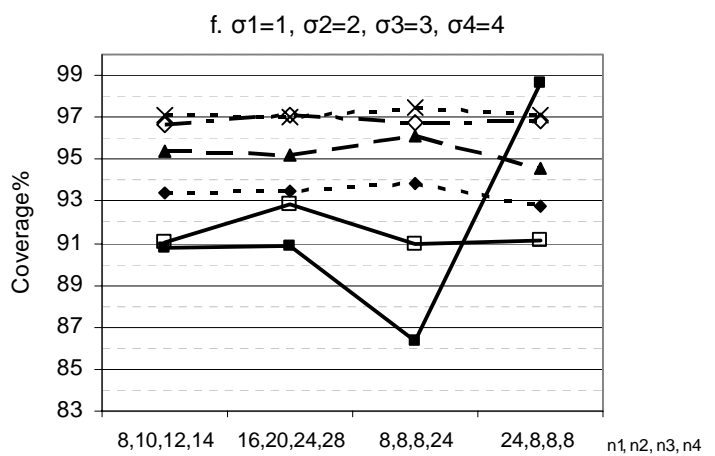
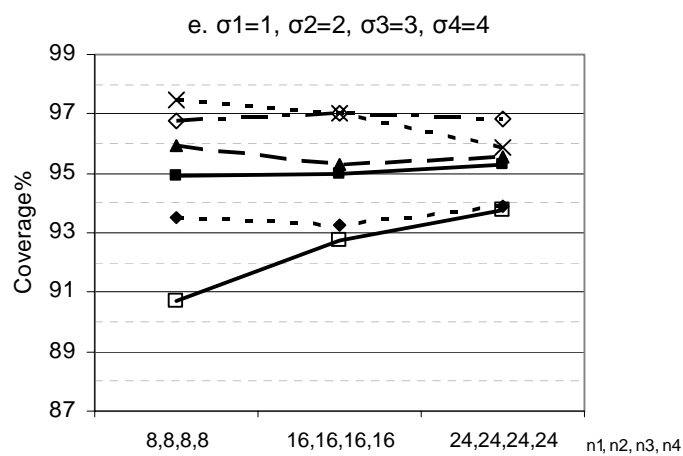
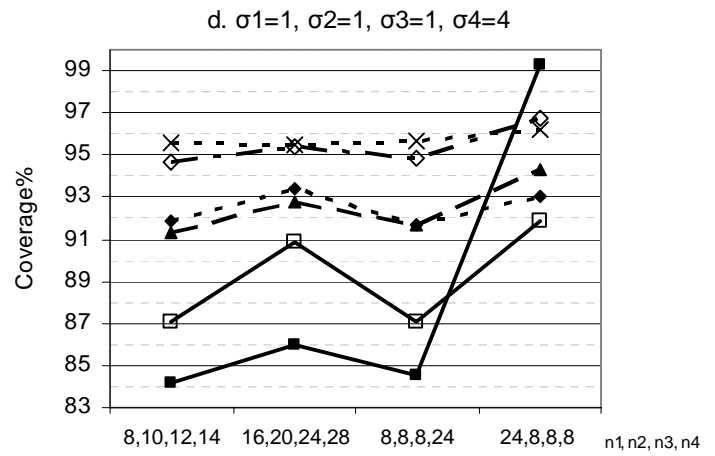
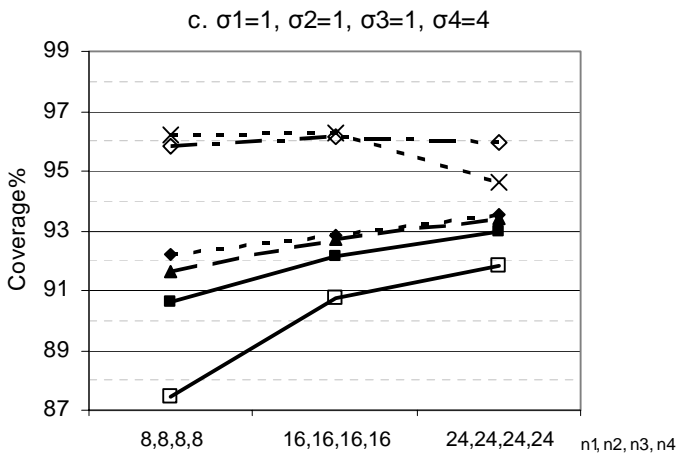
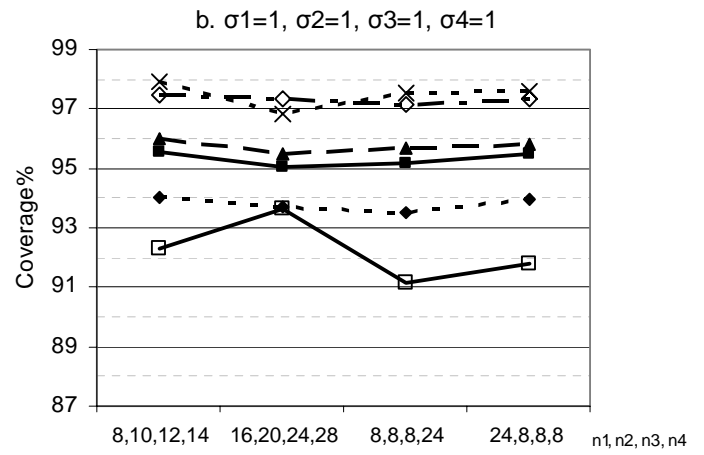
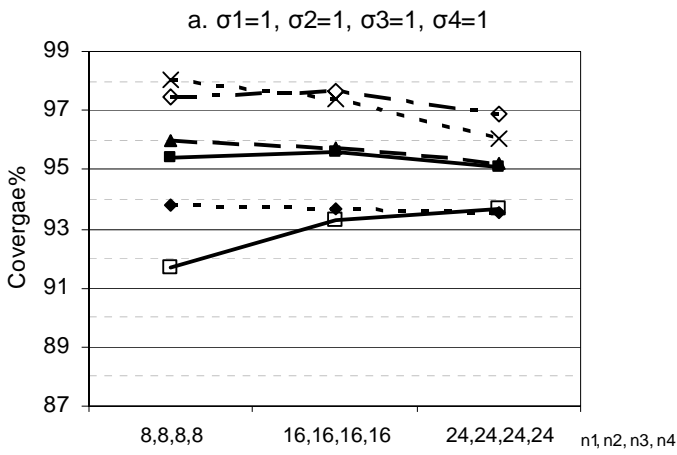
**Figure 3.8:**  
Coverage percentages of six procedures for contrast analysis with equal and unequal population standard deviations and sample sizes in case of platykurtic distributions



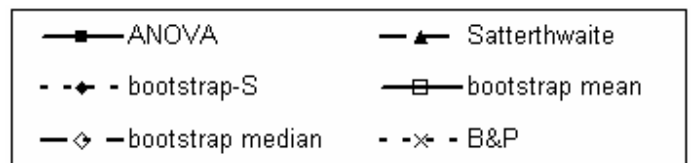


**Figure 3.9:** Coverage percentages of six procedures for contrast analysis with equal and unequal population standard deviations and sample sizes in case of moderately skewed distributions





**Figure 3.10:** Coverage percentages of six procedures for contrast analysis with equal and unequal population standard deviations and sample sizes in case of extremely skewed distributions



### §3.3 Study 3: Comparison of two samples from cardiovascular simulated data

In this study six procedures for comparing two samples are compared again, but now cardiovascular simulated populations are used. In Table 3.5a the coverage percentages of the six procedures are shown for the simulated cardiovascular variables. In Table 3.5b the coverage percentages of the six procedures are shown for the six logarithmically transformed simulated cardiovascular variables. In Figure 3.11 the coverage percentages for the six procedures are shown for the inter-beat interval (IBI), the mean blood pressure (MBP), and the modulus mid frequency band (Modulus). In Figures 3.12 and 3.13 the coverage percentages for the six procedures are shown for respectively heart rate variability (HRV) frequency bands and blood pressure variability (BPV) frequency bands. First the untransformed cardiovascular variables are discussed, succeeded by the logarithmically transformed variables.

For the untransformed cardiovascular variables the coverage percentages found in this study are very much in line with Study 1. The Student's  $t$  procedure has good coverage percentages when  $n_1=n_2$  except for 'HRV low' when  $n_1=n_2=8$  (see Table 3.5a). This was to be expected from Study 1, because the populations both have moderately/extremely skewed distributions and very unequal population standard deviations with ratio  $\sigma_1/\sigma_2=2.8$  (see Table 3.6a). However, for 'BPV low' the populations also are skewed and have unequal standard deviations with ratio  $\sigma_1/\sigma_2=1.9$ , therefore also bad coverage percentages could be expected for 'BPV low', but these are not found. The skewness, the kurtosis and the ratio  $\sigma_1/\sigma_2$  are smaller for 'BPV low' than for 'HRV low'. Probably this would explain that no bad coverage percentages were found at  $n_1=n_2=8$  for 'BPV low'. When  $n_1 \neq n_2$  the coverage percentages for the Student's  $t$  procedure are acceptable for 'IBI' and good for 'HRV mid' where the other variables have bad coverage percentages. The same effect was also to be expected from Study 1 for 'IBI' and 'HRV mid' because the two populations are approximately normally distributed with about equal standard deviations. However, for 'Modulus' and 'BPV high' the populations also have about equal standard deviations, therefore also good coverage percentages would be expected for 'Modulus' and 'BPV high', but are not found. The reason has to be related to the fact that the two populations for 'Modulus' and 'BPV high' differ more in terms of skewness or kurtosis than for 'IBI' and 'HRV mid'.

**Table 3.5a:**

Coverage percentages\* and mean confidence intervals (in parentheses) for Study 3 with simulated cardiovascular data

	$n_1$	$n_2$	Student's $t$	Welch	bootstrap-W	bootstrap mean	bootstrap median	B&P
IBI	8	8	<b>95.3</b> (43.4)	95.6 (43.9)	95.5 (45.4)	91.3 (37.0)	96.3 (49.7)	96.3 (55.4)
	24	24	95.2 (23.8)	95.3 (23.8)	95.2 (23.8)	94.0 (22.7)	96.7 (29.8)	<b>95.1</b> (30.4)
	8	24	93.9 (33.1)	<b>94.8</b> (37.2)	94.6 (38.3)	91.0 (31.2)	95.4 (41.6)	95.5 (45.6)
	24	8	96.2 (35.0)	<b>95.0</b> (34.7)	<b>95.0</b> (35.3)	91.9 (29.8)	95.9 (39.7)	95.5 (43.2)
MBP	8	8	94.6 (71.5)	<b>95.1</b> (73.5)	94.6 (77.7)	90.5 (60.8)	96.0 (79.0)	96.2 (88.1)
	24	24	95.0 (39.4)	95.2 (39.6)	94.9 (40.0)	93.8 (37.6)	96.4 (46.4)	<b>95.0</b> (47.4)
	8	24	88.6 (49.5)	93.8 (67.4)	93.6 (72.8)	89.5 (54.7)	94.7 (69.9)	<b>95.0</b> (77.7)
	24	8	98.3 (62.5)	<b>95.2</b> (51.1)	94.7 (51.8)	92.4 (45.5)	96.5 (58.0)	95.7 (61.9)
Modulus mid	8	8	<b>95.1</b> (3.4)	95.4 (3.4)	95.4 (3.5)	91.2 (2.9)	95.8 (3.8)	96.1 (4.3)
	24	24	<b>95.1</b> (1.8)	95.2 (1.8)	<b>95.1</b> (1.9)	94.0 (1.8)	96.3 (2.3)	94.8 (2.4)
	8	24	97.4 (2.8)	<b>95.0</b> (2.6)	94.7 (2.6)	91.9 (2.2)	96.0 (3.0)	95.5 (3.2)
	24	8	91.5 (2.5)	94.7 (3.0)	94.7 (3.1)	90.6 (2.5)	<b>95.0</b> (3.3)	<b>95.0</b> (3.7)
HRV low	8	8	93.0 (679.5)	93.9 (722.2)	93.7 (810.1)	89.0 (576.2)	94.0 (743.4)	<b>95.0</b> (838.5)
	24	24	94.4 (376.5)	94.7 (382.8)	<b>94.9</b> (393.2)	93.1 (358.3)	95.5 (440.9)	94.0 (455.8)
	8	24	79.1 (409.2)	93.2 (706.2)	<b>94.3</b> (817.0)	88.3 (553.3)	93.1 (706.7)	93.9 (802.5)
	24	8	99.2 (642.0)	95.2 (424.1)	94.5 (431.6)	92.8 (390.7)	96.1 (489.8)	<b>95.1</b> (511.8)
HRV mid	8	8	<b>95.1</b> (340.4)	95.4 (344.3)	95.8 (355.3)	91.3 (290.1)	96.3 (393.7)	96.1 (438.0)
	24	24	95.3 (186.3)	95.3 (186.5)	95.3 (186.9)	94.3 (177.8)	96.7 (239.1)	<b>95.1</b> (243.8)
	8	24	95.2 (268.6)	94.8 (280.2)	94.8 (287.2)	91.3 (238.7)	95.7 (322.6)	<b>95.1</b> (351.6)
	24	8	95.0 (265.3)	<b>95.0</b> (283.5)	<b>95.0</b> (291.0)	91.5 (240.2)	95.8 (324.2)	95.5 (353.3)
HRV high	8	8	94.6 (279.3)	<b>95.0</b> (288.0)	94.9 (302.2)	90.7 (237.8)	95.8 (316.7)	95.8 (353.8)
	24	24	<b>95.3</b> (153.4)	95.4 (154.5)	<b>95.3</b> (155.3)	94.0 (146.3)	96.4 (190.4)	94.6 (194.8)
	8	24	87.1 (188.7)	94.3 (268.3)	94.4 (286.4)	89.9 (216.5)	<b>94.6</b> (284.9)	<b>94.6</b> (318.0)
	24	8	98.6 (246.9)	95.2 (194.3)	<b>94.9</b> (196.1)	92.7 (174.5)	96.5 (230.9)	95.6 (245.3)
BPV low	8	8	94.2 (341.4)	94.9 (355.2)	95.1 (377.7)	90.7 (290.5)	94.7 (396.3)	<b>95.0</b> (443.4)
	24	24	<b>95.2</b> (187.3)	95.5 (189.1)	95.7 (190.1)	94.1 (178.5)	95.8 (247.6)	93.6 (254.4)
	8	24	84.7 (223.8)	94.0 (336.3)	<b>95.1</b> (364.6)	90.1 (269.2)	93.7 (365.2)	93.6 (411.0)
	24	8	99.0 (306.3)	95.3 (229.2)	<b>95.0</b> (232.2)	92.8 (207.2)	96.2 (286.8)	95.1 (300.3)
BPV mid	8	8	94.3 (252.0)	<b>95.0</b> (262.7)	95.2 (277.5)	90.4 (214.4)	94.6 (290.2)	95.1 (325.0)
	24	24	<b>94.9</b> (138.6)	95.2 (140.1)	95.5 (140.8)	93.9 (132.1)	96.0 (180.0)	93.9 (184.9)
	8	24	99.1 (229.4)	<b>94.9</b> (166.8)	<b>94.9</b> (168.7)	92.7 (151.8)	96.1 (208.2)	95.4 (218.5)
	24	8	83.9 (161.7)	<b>94.9</b> (251.5)	95.3 (269.1)	90.1 (200.3)	94.0 (269.0)	94.2 (302.4)
BPV high	8	8	<b>95.1</b> (49.1)	95.2 (49.9)	95.2 (51.9)	91.0 (41.8)	96.2 (55.5)	96.3 (62.1)
	24	24	<b>95.2</b> (26.9)	95.3 (27.0)	95.3 (27.0)	94.2 (25.6)	96.4 (32.3)	95.3 (33.0)
	8	24	92.0 (36.1)	94.7 (43.7)	94.6 (45.4)	90.7 (36.2)	<b>94.9</b> (47.4)	<b>94.9</b> (52.7)
	24	8	97.3 (40.8)	<b>95.2</b> (37.5)	94.6 (38.2)	91.8 (32.7)	96.5 (42.4)	96.5 (45.8)

\* For each condition the best coverage percentages are set in bold.

The Welch and the bootstrap-Welch procedure have coverage percentages very close to 95% for all different cardiovascular variables for all sample size conditions (see Table 3.5a). The bootstrap percentile mean procedure performs always too liberal and the bootstrap percentile median procedure has overall slightly conservative coverage percentages. The B&P procedure performs well for most variables but tends to be a little conservative especially when  $n_1=n_2=8$ . The coverage percentages for these procedures are as expected from Study 1.

**Table 3.5b:**

Coverage percentages\* and mean confidence intervals (in parentheses) for Study 3 with logarithmically transformed simulated cardiovascular data

	$n_1$	$n_2$	Student's $t$	Welch	bootstrap-W	bootstrap mean	bootstrap median	B&P
Ln HRV low	8	8	94.6 (0.44)	95.2 (0.45)	<b>95.1</b> (0.47)	90.9 (0.37)	95.4 (0.50)	95.7 (0.56)
	24	24	<b>95.5</b> (0.24)	95.6 (0.24)	95.4 (0.24)	94.1 (0.23)	95.8 (0.30)	<b>94.5</b> (0.31)
	8	24	87.6 (0.30)	94.2 (0.42)	<b>94.7</b> (0.44)	90.0 (0.34)	94.1 (0.45)	94.3 (0.50)
	24	8	98.5 (0.38)	95.3 (0.31)	94.9 (0.31)	92.8 (0.27)	96.4 (0.37)	95.8 (0.39)
Ln HRV mid	8	8	<b>95.1</b> (0.24)	95.3 (0.24)	95.5 (0.25)	91.2 (0.21)	95.8 (0.28)	96.0 (0.31)
	24	24	<b>95.3</b> (0.13)	<b>95.3</b> (0.13)	<b>95.3</b> (0.13)	94.3 (0.13)	96.1 (0.17)	94.6 (0.17)
	8	24	96.2 (0.20)	94.7 (0.19)	94.7 (0.20)	91.6 (0.17)	95.4 (0.22)	<b>95.0</b> (0.24)
	24	8	93.9 (0.18)	<b>94.9</b> (0.21)	94.7 (0.21)	90.9 (0.17)	95.2 (0.23)	95.4 (0.25)
Ln HRV high	8	8	<b>95.4</b> (0.32)	95.6 (0.32)	95.6 (0.33)	91.4 (0.27)	96.2 (0.37)	96.1 (0.41)
	24	24	<b>95.2</b> (0.17)	95.3 (0.17)	95.4 (0.18)	94.1 (0.17)	96.4 (0.22)	94.5 (0.23)
	8	24	94.3 (0.25)	94.7 (0.27)	94.6 (0.28)	91.0 (0.23)	95.4 (0.31)	<b>95.2</b> (0.34)
	24	8	96.0 (0.26)	<b>94.9</b> (0.26)	<b>94.9</b> (0.26)	91.8 (0.22)	95.8 (0.30)	95.4 (0.32)
Ln BPV low	8	8	94.6 (0.34)	<b>95.0</b> (0.35)	95.7 (0.36)	91.0 (0.29)	95.1 (0.40)	95.1 (0.44)
	24	24	<b>95.2</b> (0.18)	95.4 (0.19)	95.7 (0.19)	94.2 (0.18)	95.9 (0.25)	93.9 (0.25)
	8	24	87.6 (0.23)	94.4 (0.32)	<b>95.2</b> (0.34)	90.2 (0.26)	94.0 (0.36)	93.7 (0.40)
	24	8	98.5 (0.29)	94.8 (0.24)	<b>95.1</b> (0.24)	92.5 (0.21)	96.2 (0.30)	95.2 (0.31)
Ln BPV mid	8	8	94.5 (0.45)	<b>94.7</b> (0.46)	95.5 (0.48)	91.0 (0.38)	95.9 (0.52)	95.8 (0.58)
	24	24	<b>95.3</b> (0.25)	<b>95.3</b> (0.25)	95.6 (0.25)	94.3 (0.24)	96.7 (0.32)	94.6 (0.33)
	8	24	96.8 (0.37)	94.3 (0.35)	<b>94.9</b> (0.36)	91.6 (0.30)	95.9 (0.41)	95.2 (0.45)
	24	8	92.2 (0.33)	94.7 (0.40)	<b>95.0</b> (0.42)	90.7 (0.33)	95.4 (0.45)	94.9 (0.49)
Ln BPV high	8	8	94.7 (0.26)	95.2 (0.26)	<b>94.9</b> (0.28)	90.7 (0.22)	95.3 (0.29)	96.0 (0.32)
	24	24	<b>95.1</b> (0.14)	95.3 (0.14)	<b>95.1</b> (0.14)	94.1 (0.13)	95.8 (0.17)	95.2 (0.17)
	8	24	87.4 (0.18)	<b>94.6</b> (0.24)	94.5 (0.26)	89.8 (0.20)	94.0 (0.26)	<b>94.6</b> (0.29)
	24	8	98.5 (0.23)	95.4 (0.18)	<b>94.7</b> (0.18)	92.6 (0.16)	96.2 (0.21)	96.6 (0.22)

\* For each condition the best coverage percentages are set in bold.

**Table 3.6a:**  
Descriptives for untransformed simulated cardiovascular data

Variable	Pop.	Mean	Median	Std. Dev.	Skewness	Kurtosis
IBI	1	892	891	21.6	0.05	3.00
	2	825	825	19.4	0.12	2.89
MBP	1	1105	1099	41.1	0.90	4.39
	2	1046	1044	25.8	0.69	4.12
Modulus	1	17.2	17.2	1.38	0.21	4.08
	2	12.0	11.9	1.79	0.23	2.76
HRV low	1	1664	1582	438	0.98	3.98
	2	924	896	155	1.66	8.28
HRV mid	1	1508	1504	159	0.02	2.75
	2	1391	1400	163	-0.23	2.95
HRV high	1	1027	1012	162.7	0.50	3.26
	2	639	630	94.8	0.63	3.62
BPV low	1	1056	1029	203	0.40	2.40
	2	836	810	107	1.56	6.60
BPV mid	1	367	352	74.0	1.16	4.63
	2	678	679	152.8	0.09	2.58
BPV high	1	178	178	25.8	0.05	2.65
	2	236	239	20.5	-0.55	3.85

**Table 3.6b:**  
Descriptives for logarithmically transformed simulated cardiovascular data

Variable	Pop.	Mean	Median	Std. Dev.	Skewness	Kurtosis
ln HRV low	1	7.38	7.37	0.25	0.35	2.68
	2	6.82	6.80	0.15	0.89	4.64
ln HRV mid	1	7.31	7.32	0.11	-0.27	2.94
	2	7.23	7.24	0.12	-0.60	3.53
ln HRV high	1	6.92	6.92	0.16	0.05	2.86
	2	6.45	6.45	0.15	0.19	2.93
ln BPV low	1	6.94	6.94	0.19	0.05	2.22
	2	6.72	6.70	0.12	1.08	4.54
ln BPV mid	1	5.89	5.86	0.19	0.65	3.04
	2	6.49	6.52	0.24	-0.48	2.97
ln BPV high	1	5.17	5.18	0.15	-0.32	2.81
	2	5.46	5.47	0.09	-0.91	4.66

The logarithmically transformed variables 'ln HRV low', 'ln HRV high', 'ln BPV low', and 'ln BPV mid' have, compared to their untransformed variable, a more normal distribution in terms of skewness and kurtosis and have less different population standard deviations (see Table 3.6b). The transformed variables 'ln HRV mid' and 'ln BPV high' have a more non-normal distribution in terms of skewness and kurtosis and have more different population standard deviations. The distributions have a considerable negative skewness due to the transformation because the untransformed variables were approximately normally distributed with even a small negative skewness.

The Student's *t* procedure has good coverage percentages for the logarithmically transformed variables when  $n_1=n_2$  (see Table 3.5b). However, the coverage percentages are not better after logarithmic transformation; only for 'HRV low' in case  $n_1=n_2=8$  the coverage percentage is improved. When  $n_1 \neq n_2$  the coverage percentages for 'ln HRV mid' and 'ln HRV high' are acceptable where the other variables have bad coverage percentages. This can be explained from the results of Study 1: the two populations for 'ln HRV mid' and 'ln HRV high' are about equally distributed (the same standard deviation, skew and kurtosis) where for the other variables the two populations have more different distribution shapes (see Table 3.6b). When the transformation leads to a more normal distribution the coverage percentages are a little closer to 95% than with no transformation and when the transformation leads to a more non-normal distribution the coverage percentages are further away from 95% than without transformation.

For the five other procedures the changes in coverage percentages due to transformation are negligible (see Table 3.5b). The Welch and the bootstrap-Welch procedure perform very close to 95% for all transformed cardiovascular variables and for all sample size conditions. The bootstrap percentile mean procedure still performs always too liberal and the bootstrap percentile median has overall slightly conservative coverage percentages. The B&P procedure performs well for most variables but in general the coverage percentages tend to be a little conservative and tend to decrease with increasing total sample size.

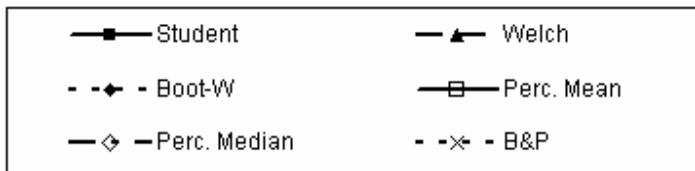
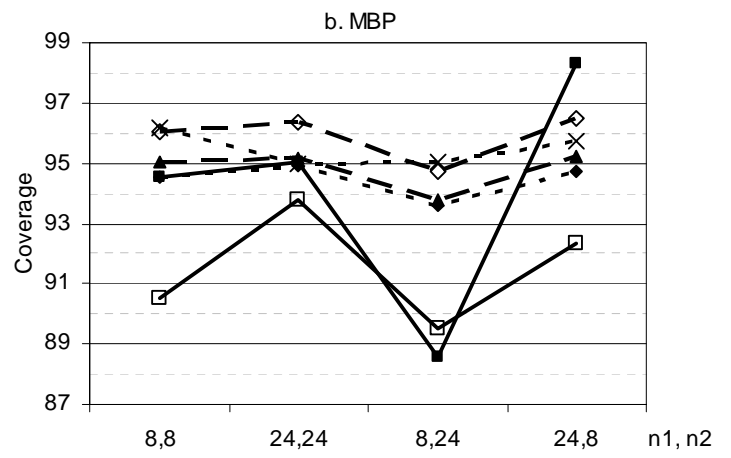
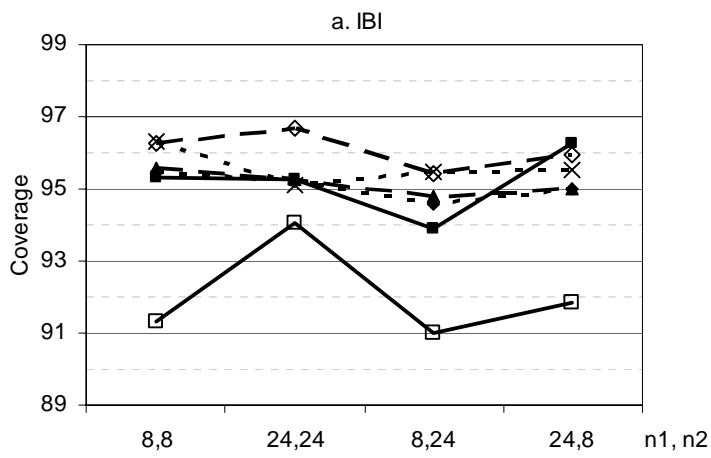
For the untransformed as well as for the transformed situation it can be seen that the confidence intervals get narrower with increasing total sample size for all procedures (see Table 3.5a and 3.5b). When the sample size conditions with  $n_1=n_2=8$  are compared

to the conditions with  $n_1=8$ ,  $n_2=24$  and  $n_1=24$ ,  $n_2=8$  the Welch, the bootstrap-Welch, and to a lesser extent the B&P procedure also remain to have good coverage percentages. Therefore, the power of the Welch, bootstrap-Welch, and the B&P procedures increase with increasing sample size, even with unequal sample sizes.

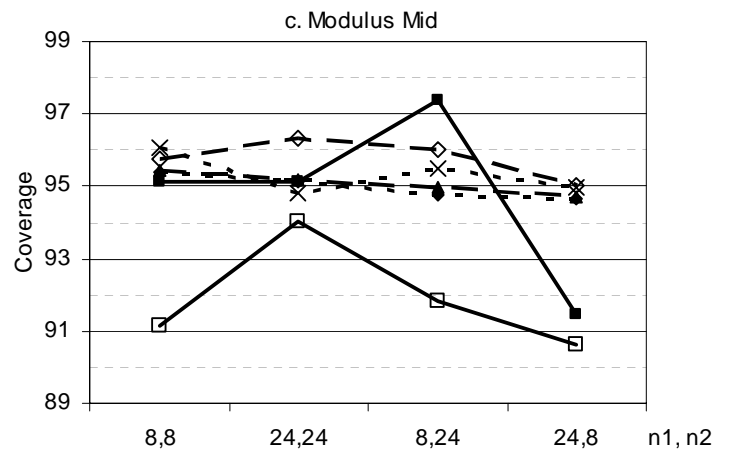
Logarithmic transformation on cardiovascular variables is carried out to reduce a positively skewed distribution to a more normal distribution. The logarithmic transformation of the spectral functions is theoretically grounded (e.g. Van Roon, 1998, p.96-97). The mean, standard deviation, skewness and kurtosis of a transformed variable will be different (when transformed back) to the mean of the untransformed variable, while the median remains the same. The bootstrap percentile mean and median procedures are transformation respecting: variable transformation does not have an influence on the confidence interval, after this has been transformed back, provided that the confidence interval concerns a (single) mean or median.

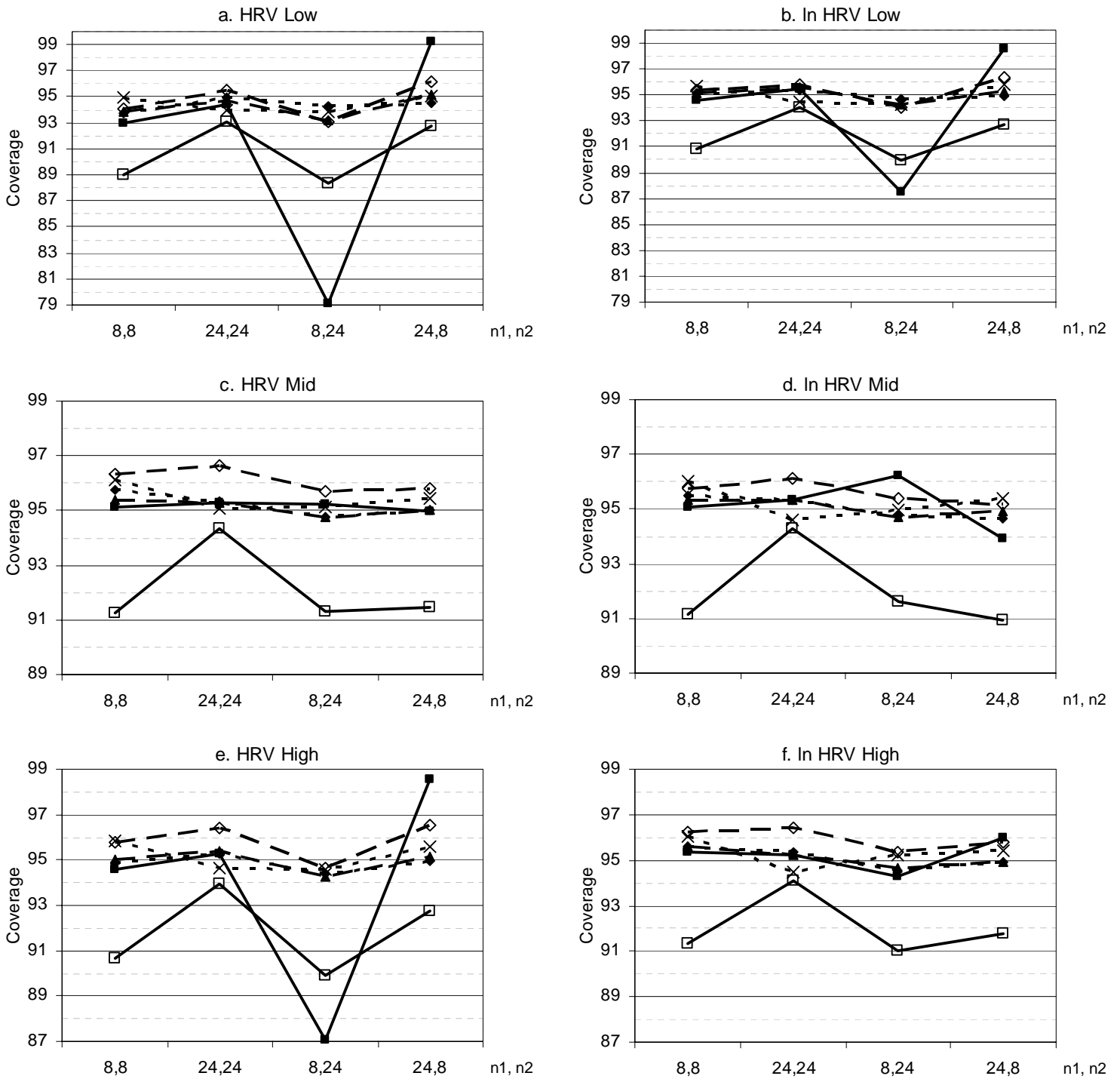
In this study confidence intervals for differences between means or medians are considered. Unfortunately, the difference in population means and medians for the two transformed variables is dissimilar (when transformed back) from the difference in means and medians for the two untransformed variables. Therefore the confidence intervals and the coverage percentages for transformed variables are different from those for untransformed variables for all used procedures.

To summarize Study 3, it has been found that simulated experimental data that are not ideally distributed show about the same results as smooth, ideal simulated data from Study 1. The Welch procedure and the bootstrap-Welch procedure are the best procedures for comparing means: both procedures perform very well for both untransformed and transformed variables. When comparing median values, the B&P procedure performs better than the bootstrap percentile median procedure for both untransformed and transformed variables. Further, it was observed that transformation of cardiovascular data does not always lead to more normal distributions.

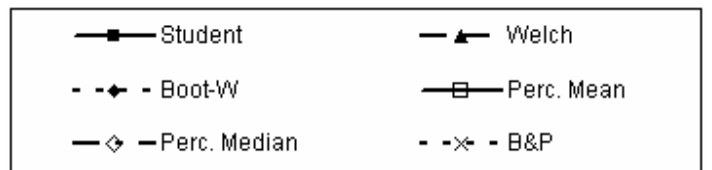


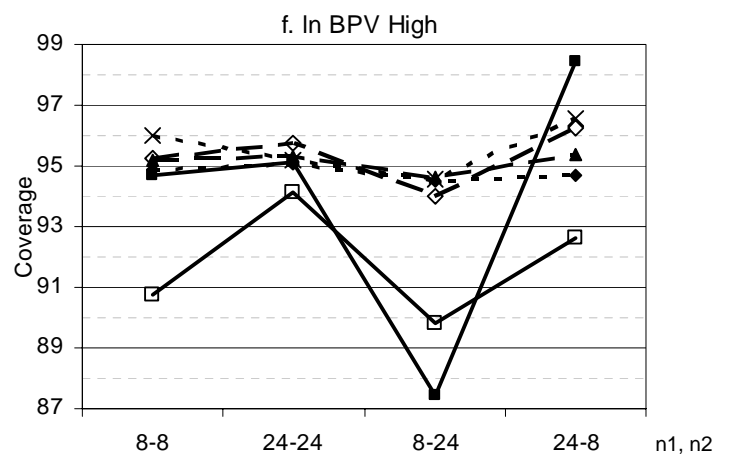
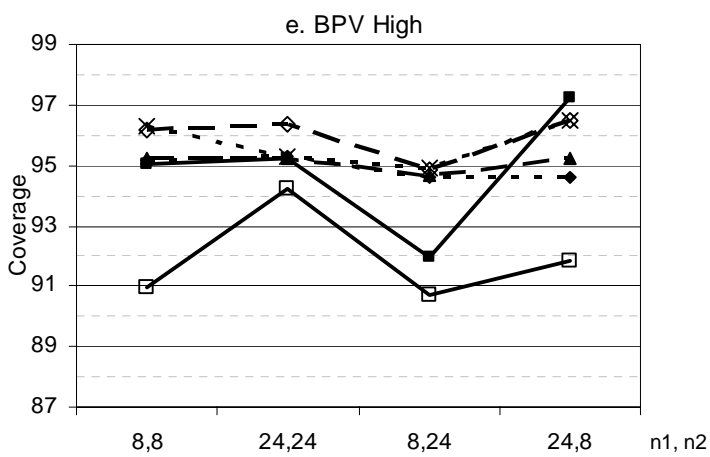
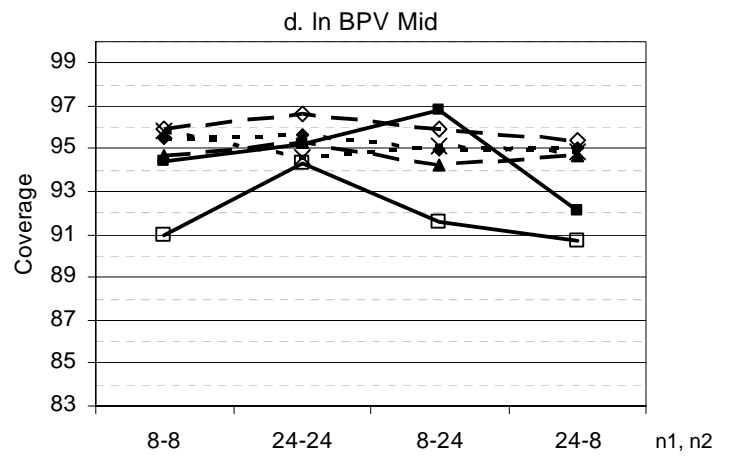
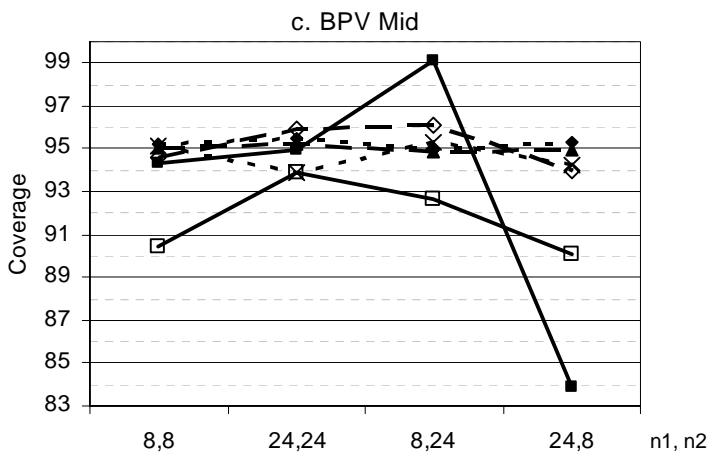
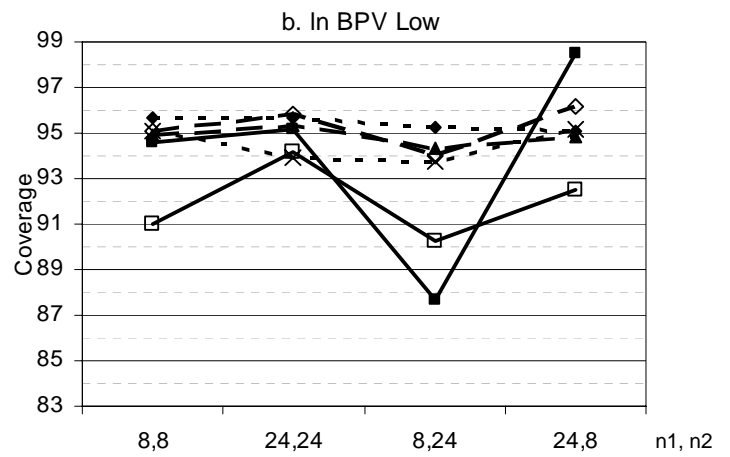
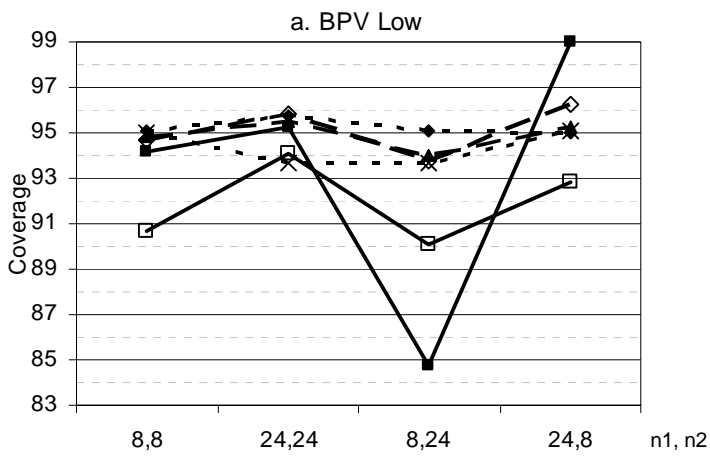
**Figure 3.11**  
Coverage percentages of six procedures with equal and unequal sample sizes for IBI, MBP, and Modulus Mid





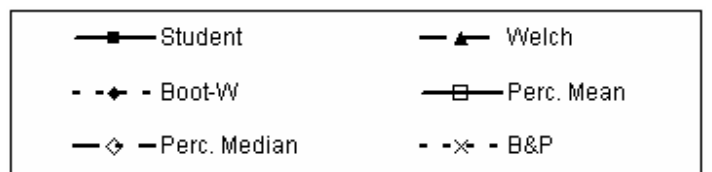
**Figure 3.12**  
Coverage percentages of six procedures with equal and unequal sample sizes for HRV (low, mid, and high) and logarithmic transformed HRV (low, mid, and high)





**Figure 3.13**

Coverage percentages of six procedures with equal and unequal sample sizes for BPV (low, mid, and high) and logarithmic transformed BPV (low, mid, and high)



### **§3.4 Study 4: Using real-world experimental data**

Two groups of children ( $n_1=16$ ,  $n_2=16$ ) with autistic-type behavior problems were compared to a group of normal children ( $n_3=16$ ) with respect to their cardiovascular responses. The cardiovascular responses are measured for a rest condition and a task condition. The difference scores between rest and task condition of nine cardiovascular variables are used. Also for six variables the logarithmic variables are used, so the logarithmic score for the rest condition is subtracted from the logarithmic score of the task condition. Note that the difference variable itself is not a logarithmically transformed variable.

The six procedures for contrast analysis with the weights -0.5, -0.5, and 1 are compared on confidence intervals. The weights are chosen in such a way that the two autistic-type groups are compared to the control group. If there is no difference between the two autistic-type groups with the control group, the confidence interval would include zero.

All procedures give confidence intervals for almost all untransformed variables that cover zero; only for 'HRV mid' the confidence intervals do not include zero (see Table 3.7a and Figure 3.14a). For some variables the different procedures give rather different confidence intervals. For 'HRV low' the mean based procedures give much wider confidence intervals than the median based procedures and for 'IBI' and 'BPV mid' the mean based procedures give much narrower confidence intervals than the median based procedures. For all other variables the procedures give roughly equally wide confidence intervals. For 'HRV mid' the confidence intervals for the mean based procedures are located more to the "right" (further away from zero) compared to the median based procedures and for 'BPV low' the median based procedures give confidence intervals located more to the "right" (lower bound closer to zero) compared to the mean based procedures.

For the six logarithmically transformed ( $\ln$ ) variables, the procedures give similar confidence intervals, but with some exceptions (see Table 3.7a and Figure 3.14b). For ' $\ln$  HRV high' and ' $\ln$  BPV mid' the median based procedures give wider confidence intervals than the mean based procedures and for ' $\ln$  BPV high' only the B&P procedure gives much wider confidence intervals compared to the other procedures. Further, for ' $\ln$

HRV high', 'ln BPV low', and 'ln BPV mid' the procedures give confidence intervals with a lower bound that is nearly zero.

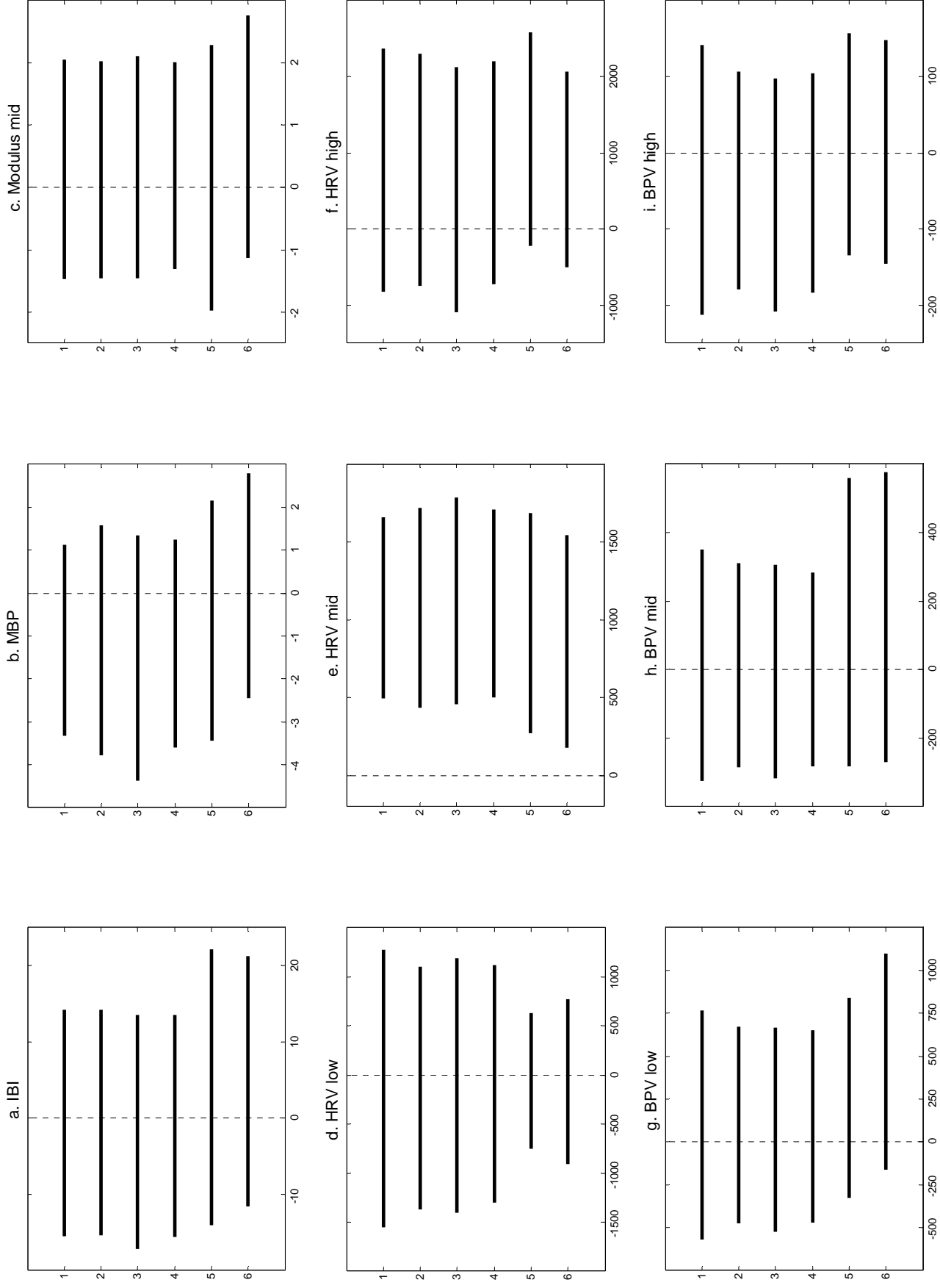
To compare the confidence intervals for transformed variables to those for untransformed, both are standardized by division with the standard error of contrasts of the means or medians. This is a way to compare the confidence intervals with transformed variables to the confidence intervals without transformed variables. In Table 3.7b and Figure 3.15 the standardized confidence intervals for the six logarithmically transformed and untransformed variables are shown. The standardized confidence intervals for the untransformed and transformed variables are by definition equally wide for the ANOVA (twice the  $t$ -value) and the B&P procedure (twice the  $z$ -value), almost equally wide for the Satterthwaite, bootstrap-Satterthwaite, and bootstrap percentile mean procedure and somewhat differently wide for the bootstrap percentile median procedure.

For 'ln HRV low' the standardized confidence interval of all procedures show a shift to the "right" (higher bound further away from zero) compared to the standardized confidence intervals of 'HRV low'. For 'ln HRV mid', 'ln HRV high', and 'ln BPV high' the standardized confidence intervals of the mean based procedures show a little shift to the "right" compared to no transformation, while the median based procedures have about the same standardized confidence intervals. For 'ln BPV low' and 'ln BPV mid' the shift to the "right" of the mean based procedures is larger, resulting in standardized confidence intervals with a lower bound that is almost zero.

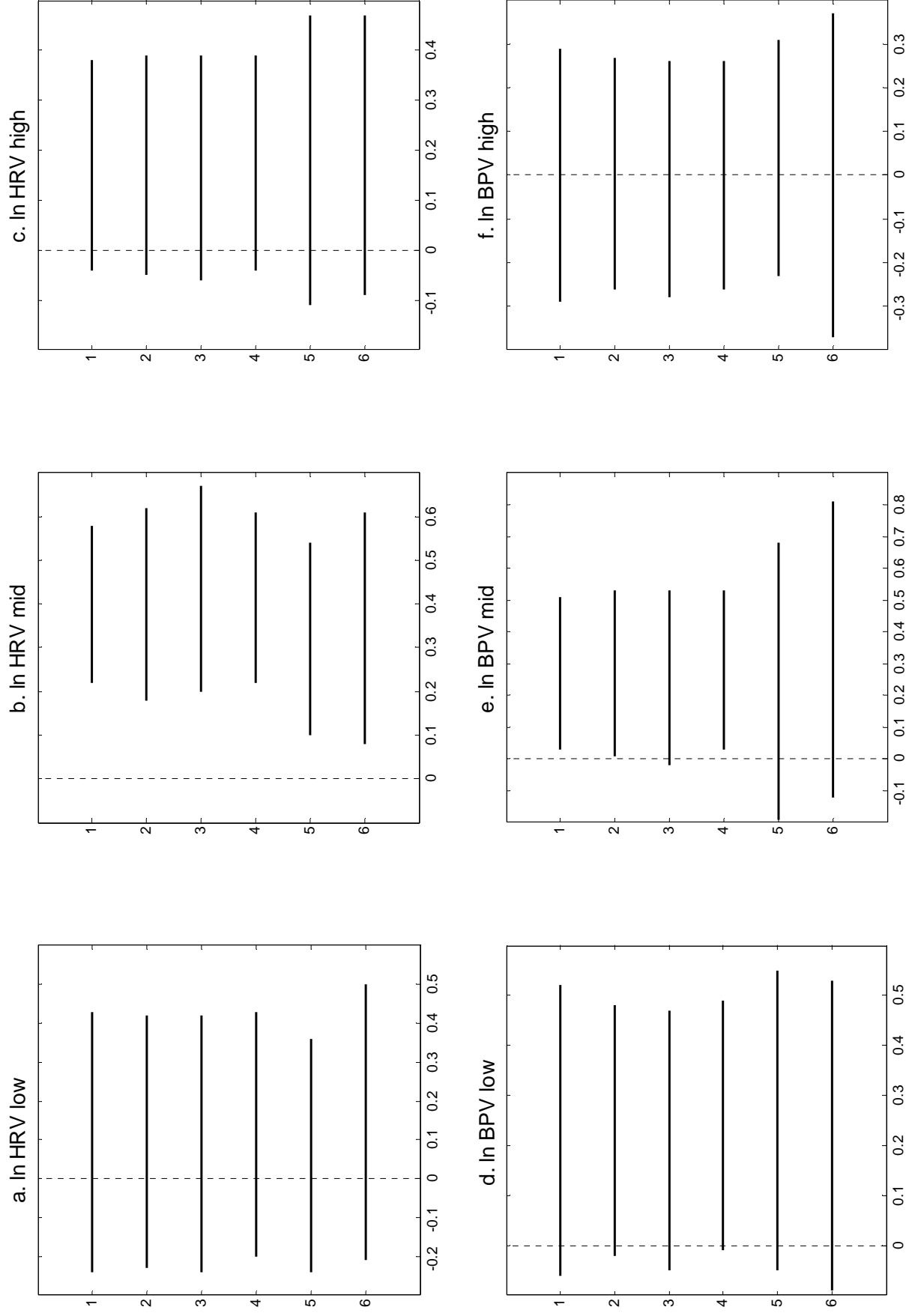
In short, for the untransformed cardiovascular variables some differences in location and width of the confidence intervals are seen, especially between mean based procedures and median based procedures. The mean based procedures give confidence intervals that are more affected by transformation than the median based procedures, as can be seen from the standardized confidence intervals. The confidence intervals of the mean based procedures show a considerable shift to the "right" after transformation, while the confidence intervals of the median based procedures show a smaller shift or hardly any shift at all.

**Table 3.7a:**  
Confidence intervals for Study 4

	ANOVA	Satterthwaite	bootstrap-S	bootstrap mean	bootstrap median	B&P
BI	-15.5 ≤μ≤ 14.2	-15.4 ≤μ≤ 14.2	-17.1 ≤μ≤ 13.6	-15.6 ≤μ≤ 13.5	-14.0 ≤μ≤ 22.1	-11.6 ≤μ≤ 21.3
BP	-3.32 ≤μ≤ 1.13	-3.78 ≤μ≤ 1.59	-4.37 ≤μ≤ 1.34	-3.61 ≤μ≤ 1.24	-3.45 ≤μ≤ 2.15	-2.45 ≤μ≤ 2.79
odulus	-1.47 ≤μ≤ 2.05	-1.45 ≤μ≤ 2.02	-1.46 ≤μ≤ 2.10	-1.31 ≤μ≤ 2.01	-1.97 ≤μ≤ 2.28	-1.13 ≤μ≤ 2.76
RV lo	-1549.0 ≤μ≤ 1281.6	-1372.9 ≤μ≤ 1105.4	-1406.4 ≤μ≤ 1191.5	-1296.7 ≤μ≤ 1119.1	-753.4 ≤μ≤ 636.6	-906.1 ≤μ≤ 770.2
n HRV lo	-0.24 ≤μ≤ 0.43	-0.23 ≤μ≤ 0.42	-0.24 ≤μ≤ 0.42	-0.20 ≤μ≤ 0.43	-0.24 ≤μ≤ 0.36	-0.21 ≤μ≤ 0.50
RV mi	495.9 ≤μ≤ 1656.9	433.9 ≤μ≤ 1719.0	455.9 ≤μ≤ 1786.5	503.6 ≤μ≤ 1706.5	270.7 ≤μ≤ 1686.9	177.6 ≤μ≤ 1546.0
n HRV mi	0.22 ≤μ≤ 0.58	0.18 ≤μ≤ 0.62	0.20 ≤μ≤ 0.67	0.22 ≤μ≤ 0.61	0.10 ≤μ≤ 0.54	0.08 ≤μ≤ 0.61
RV hi	-819.9 ≤μ≤ 2366.0	-750.7 ≤μ≤ 2296.8	-1090.0 ≤μ≤ 2126.0	-724.5 ≤μ≤ 2196.1	-220.0 ≤μ≤ 2578.9	-501.0 ≤μ≤ 2067.4
n HRV hi	-0.04 ≤μ≤ 0.38	-0.05 ≤μ≤ 0.39	-0.06 ≤μ≤ 0.39	-0.04 ≤μ≤ 0.39	-0.11 ≤μ≤ 0.47	-0.09 ≤μ≤ 0.47
PV lo	-570.3 ≤μ≤ 766.5	-475.5 ≤μ≤ 671.6	-525.8 ≤μ≤ 667.5	-471.6 ≤μ≤ 650.5	-327.6 ≤μ≤ 840.6	-162.0 ≤μ≤ 1098.7
n BPV lo	-0.06 ≤μ≤ 0.52	-0.02 ≤μ≤ 0.48	-0.05 ≤μ≤ 0.47	-0.01 ≤μ≤ 0.49	-0.05 ≤μ≤ 0.55	-0.09 ≤μ≤ 0.53
PV mi	-324.0 ≤μ≤ 351.5	-283.9 ≤μ≤ 311.3	-315.9 ≤μ≤ 305.2	-280.9 ≤μ≤ 284.4	-282.8 ≤μ≤ 559.3	-268.7 ≤μ≤ 577.2
n BPV mi	0.03 ≤μ≤ 0.51	0.01 ≤μ≤ 0.53	-0.02 ≤μ≤ 0.53	0.03 ≤μ≤ 0.53	-0.19 ≤μ≤ 0.68	-0.12 ≤μ≤ 0.81
PV hi	-212.8 ≤μ≤ 141.4	-178.6 ≤μ≤ 107.2	-207.7 ≤μ≤ 98.0	-184.0 ≤μ≤ 104.4	-134.7 ≤μ≤ 156.6	-145.6 ≤μ≤ 147.6
n BPV hi	-0.29 ≤μ≤ 0.29	-0.26 ≤μ≤ 0.27	-0.28 ≤μ≤ 0.26	-0.26 ≤μ≤ 0.26	-0.23 ≤μ≤ 0.31	-0.37 ≤μ≤ 0.37



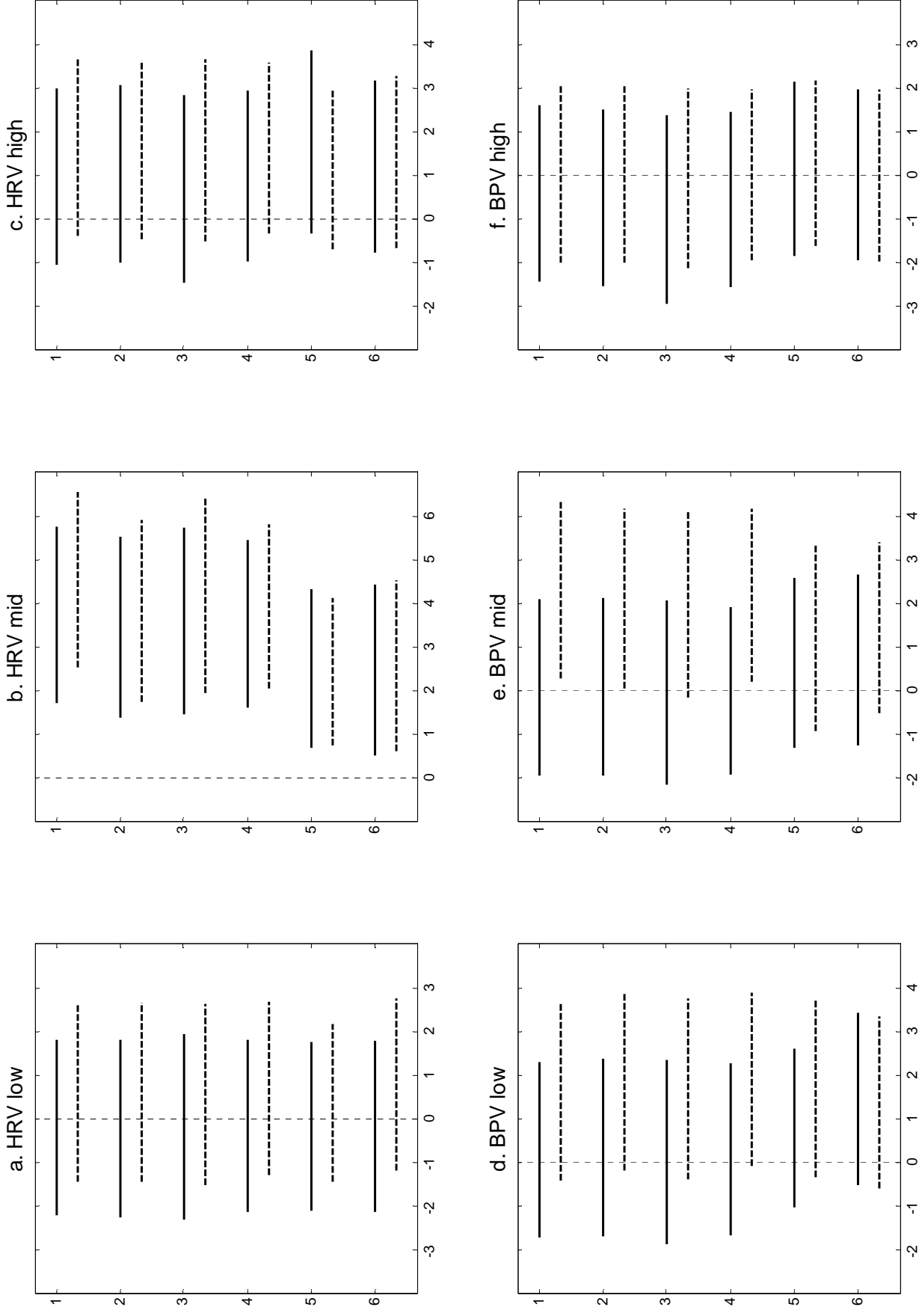
**Figure 3.14a:** Confidence intervals for ANOVA (1), Satterthwaite (2), bootstrap-S (3), bootstrap percentile mean (4) and median (5), and B&P procedure (6)



**Figure 3.14b:** Confidence intervals for Study 4 for transformed variables for ANOVA (1), Satterthwaite (2), bootstrap-S (3), bootstrap percentile mean (4) and median (5), and B&P procedure (6)

**Table 3.7b:**  
Standardized confidence intervals for Study 4 for untransformed and transformed variables

	ANOVA	Satterthwaite	bootstrap-S	bootstrap mean	bootstrap median	B&P
HRV lo	-2.20 $\leq \mu \leq$ 1.82	-2.25 $\leq \mu \leq$ 1.81	-2.30 $\leq \mu \leq$ 1.95	-2.11 $\leq \mu \leq$ 1.82	-2.09 $\leq \mu \leq$ 1.76	-2.12 $\leq \mu \leq$ 1.80
ln HRV lo	-1.43 $\leq \mu \leq$ 2.60	-1.43 $\leq \mu \leq$ 2.65	-1.52 $\leq \mu \leq$ 2.63	-1.27 $\leq \mu \leq$ 2.69	-1.44 $\leq \mu \leq$ 2.17	-1.17 $\leq \mu \leq$ 2.75
HRV mi	1.72 $\leq \mu \leq$ 5.75	1.39 $\leq \mu \leq$ 5.52	1.46 $\leq \mu \leq$ 5.74	1.61 $\leq \mu \leq$ 5.46	0.70 $\leq \mu \leq$ 4.34	0.51 $\leq \mu \leq$ 4.43
ln HRV mi	2.53 $\leq \mu \leq$ 6.56	1.74 $\leq \mu \leq$ 5.91	1.95 $\leq \mu \leq$ 6.41	2.06 $\leq \mu \leq$ 5.82	0.75 $\leq \mu \leq$ 4.12	0.62 $\leq \mu \leq$ 4.54
HRV hi	-1.04 $\leq \mu \leq$ 2.99	-1.00 $\leq \mu \leq$ 3.06	-1.45 $\leq \mu \leq$ 2.83	-0.97 $\leq \mu \leq$ 2.93	-0.33 $\leq \mu \leq$ 3.86	-0.76 $\leq \mu \leq$ 3.16
ln HRV hi	-0.37 $\leq \mu \leq$ 3.66	-0.46 $\leq \mu \leq$ 3.64	-0.51 $\leq \mu \leq$ 3.66	-0.33 $\leq \mu \leq$ 3.59	-0.68 $\leq \mu \leq$ 2.93	-0.65 $\leq \mu \leq$ 3.27
BPV lo	-1.72 $\leq \mu \leq$ 2.31	-1.68 $\leq \mu \leq$ 2.37	-1.85 $\leq \mu \leq$ 2.35	-1.65 $\leq \mu \leq$ 2.27	-1.02 $\leq \mu \leq$ 2.62	-0.50 $\leq \mu \leq$ 3.42
ln BPV lo	-0.40 $\leq \mu \leq$ 3.63	-0.18 $\leq \mu \leq$ 3.86	-0.37 $\leq \mu \leq$ 3.77	-0.08 $\leq \mu \leq$ 3.88	-0.33 $\leq \mu \leq$ 3.75	-0.58 $\leq \mu \leq$ 3.34
BPV mi	-1.93 $\leq \mu \leq$ 2.10	-1.93 $\leq \mu \leq$ 2.12	-2.15 $\leq \mu \leq$ 2.07	-1.91 $\leq \mu \leq$ 1.93	-1.31 $\leq \mu \leq$ 2.58	-1.25 $\leq \mu \leq$ 2.67
ln BPV mi	0.29 $\leq \mu \leq$ 4.31	0.06 $\leq \mu \leq$ 4.18	-0.16 $\leq \mu \leq$ 4.15	0.20 $\leq \mu \leq$ 4.17	-0.92 $\leq \mu \leq$ 3.37	-0.52 $\leq \mu \leq$ 3.40
BPV hi	-2.42 $\leq \mu \leq$ 1.61	-2.52 $\leq \mu \leq$ 1.51	-2.93 $\leq \mu \leq$ 1.38	-2.55 $\leq \mu \leq$ 1.45	-1.84 $\leq \mu \leq$ 2.14	-1.95 $\leq \mu \leq$ 1.97
ln BPV hi	-1.99 $\leq \mu \leq$ 2.04	-2.00 $\leq \mu \leq$ 2.05	-2.13 $\leq \mu \leq$ 1.99	-1.95 $\leq \mu \leq$ 1.98	-1.61 $\leq \mu \leq$ 2.18	-1.96 $\leq \mu \leq$ 1.96



**Figure 3.15:** Standardized confidence intervals for Study 4 for untransformed (solid line) and transformed variables (dashed line) for ANOVA (1), Satterthwaite (2), bootstrap-S (3), bootstrap percentile mean (4) and median (5), and B&P procedure (6)

## 4. Discussion and conclusion

First a summary of the results of the first three studies will be given. That is followed by discussing Question 1: under what circumstances do the Student's  $t$  and the ANOVA procedure fail, while alternative procedures perform well? and Question 2: which procedures perform best under what circumstances and which procedures perform well under all circumstances? Thereafter, the use of logarithmic transformation of cardiovascular data is discussed with Question 3: is logarithmic transformation of cardiovascular data necessary in cardiovascular research or do alternative procedures work as well or better? and Question 4: do the procedures lead to different conclusions when employed to real-world data? Subsequently, the use of smooth, ideal simulated data is discussed within the framework of Micceri (1989). Finally, the whole study will be discussed very briefly in comparison with earlier robustness studies and suggestions for further research will be given.

### §4.1 The three simulation studies

For the first three studies a Monte-Carlo study is performed to investigate the performance of the six procedures for comparing two or more groups, under several circumstances (Questions 1 and 2). In Study 1 six procedures for comparing two samples were compared on coverage percentages in different conditions, in Study 2 six procedures for comparing four samples in a contrast analysis were compared on coverage percentages in different conditions, and in Study 3 six procedures for comparing two samples were compared on coverage percentages with cardiovascular simulated data.

The Student's  $t$  and the ANOVA procedure perform well in case of equal population standard deviations with equal sample sizes over all different distributions. The Student's  $t$  and the ANOVA procedure fail when unequal population standard deviations are accompanied with unequal sample sizes. With negative pairing (e.g.  $\sigma_1 > \sigma_2$  with  $n_1 < n_2$ ) the coverage percentages are far below 95% (liberal). With positive pairing (e.g.  $\sigma_1 > \sigma_2$  with  $n_1 > n_2$ ) the coverage percentages are above 95% (conservative). The Student's  $t$  and the ANOVA procedure also fail when all population distributions are extremely skewed with unequal population standard deviations, with the exception that

the ANOVA procedure does perform well when  $\sigma_1=4$ ,  $\sigma_2=3$ ,  $\sigma_3=2$ ,  $\sigma_4=1$  with equal sample sizes in case of extremely skewed distributions. This might be caused by the fact that the differences between population standard deviations are gradual and not large enough, which causes the Student's  $t$  and the ANOVA procedure to perform as if the population standard deviations are equal.

The Welch and the Satterthwaite procedures perform well in almost all conditions. The Welch and the Satterthwaite procedures only fail when the distributions are extremely skewed with different population standard deviations, with the exception that the Satterthwaite procedure, like the ANOVA, also performs well when  $\sigma_1=4$ ,  $\sigma_2=3$ ,  $\sigma_3=2$ ,  $\sigma_4=1$  in case of extremely skewed distributions.

The bootstrap-Welch and bootstrap-Satterthwaite procedures are not robust to non-normal distributions while they were expected to be robust. The coverage percentages are a little liberal for leptokurtic and extremely skewed distributions and a little conservative for platykurtic distributions. However, the bootstrap-Welch performs well with the simulated cardiovascular data.

The bootstrap percentile mean procedure has in all studies for almost all conditions coverage percentages that are below 94 % (liberal).

The bootstrap percentile median procedure has in all studies coverage percentages that are a little conservative when the population standard deviations are equal. The coverage percentages get closer to 95% with increasing differences in population standard deviations. It seems that a canceling effect occurs: the general conservative effect of the bootstrap percentile median procedure is canceled by the coverage percentages reducing effect of unequal population standard deviations.

The B&P procedure is mostly influenced by the kurtosis of the distributions: for leptokurtic distributions the coverage percentages are a little conservative and for platykurtic distributions a little liberal. The coverage percentages get closer to 95% with increasing total sample size. The B&P procedure also has lower coverage percentages when population standard deviations differ.

When the population distributions are known to have equal standard deviations, the Student's  $t$  and the ANOVA procedure could be used, because both have good coverage percentages. However, the Welch and the Satterthwaite procedure perform as

well in those cases and much better when the populations are known to have unequal standard deviations. In experimental research the distributions of the populations are unknown. Therefore a procedure is needed that is not much influenced by unequal population standard deviations or non-normality. On the basis of Studies 1, 2 and 3 the conclusion would be that the Welch and the Satterthwaite procedure are the best procedures in those cases because these procedures are robust to unequal population standard deviations, and robust (to a great extent) to non-normality. Further, balancing the design (equal sample sizes) is not necessary for the Welch and the Satterthwaite procedures to have good coverage percentages. The Welch and the Satterthwaite procedure are, regarding the content, equal to the familiar Student's  $t$  and the ANOVA procedure, with the distinction that the Welch and the Satterthwaite procedure take the differences in variances into account. In terms of coverage percentages the Welch and the Satterthwaite procedure are the best procedures to use in experimental research.

The power of a procedure increases when for a condition the confidence interval becomes narrower with the coverage percentage remaining the same. In general, all procedures get narrower confidence intervals with increasing total sample size for all distributions in case the population standard deviations are equal. When the population standard deviations are unequal the confidence intervals get narrower or remain the same with increasing total sample size for unequal samples sizes. For Study 3 the confidence intervals are for all procedures always narrower with larger total sample size. However, the Student's  $t$  and the ANOVA procedure show bad coverage percentages when the population standard deviations and the sample sizes are unequal. The three bootstrap procedures and the B&P procedures have for various conditions coverage percentages that are either too conservative or too liberal. Only the Welch and the Satterthwaite procedure have good coverage percentages for all conditions (except for extremely skewed distributions). Therefore, it can be concluded from the decreasing width of the confidence interval that the power of the Welch and the Satterthwaite procedure increases with increasing sample size, even in case of unequal sample sizes. Thus a researcher should always try to use as many subjects without leaving subjects out, even if this leads to unequal sample sizes (unbalanced design). The problems encountered with the

Student's  $t$  and the ANOVA procedures with using an unbalanced design are overcome by the Welch and the Satterthwaite procedures.

#### **§4.2 Logarithmic transformation of cardiovascular data**

Cardiovascular variability data are theoretically Chi-squared distributed (Van Roon, 1998) and are therefore usually logarithmically transformed to approximate a normal distribution. For the Student's  $t$  and the ANOVA procedure normality is assumed. The question is whether logarithmic transformation of the data is necessary and whether transformation leads to better coverage percentages (Question 3). Is the theoretical basis sufficient to use transformation, are there procedures that are unaffected by transformation, and does transformation always lead to better results?

Logarithmic transformation of the data is not always necessary, as was seen in Study 3. The untransformed variables 'HRV mid' and 'BPV high' were almost normally distributed. Transformation caused these variables to be more non-normally distributed. For the other cardiovascular variables transformation did lead to a more normal distribution. When transformation did lead to a more normal distributions, the coverage percentages for the Student's  $t$  procedure were about the same for equal sample sizes and better (but still bad) for unequal sample sizes compared to no transformation. However, the improvement in coverage percentages is probably caused by the fact that the differences between the population standard deviations were smaller and not by the fact that the populations were more normally distributed. That the variables 'HRV mid' and 'BPV high' lead to more non-normal distributions can be attributed to chance. In a next experimental research other variables can lead to more non-normal distributions.

From Studies 1 and 2 can be concluded that non-normality is not the main problem for the Student's  $t$  and the ANOVA procedure, but that unequal population standard deviations with unequal sample sizes is the main cause for bad coverage percentages. The Welch and the Satterthwaite procedure do not have such problems, and are therefore the best alternative procedures to use (see Studies 1 and 2). Furthermore, for the Welch procedure and the other alternative procedures transformation does not lead to better coverage percentages (see Study 3). The coverage percentages for the alternative procedures are practically insensitive to transformation.

However, based on Study 4, transformation did have some effect on the confidence intervals of the procedures. In particular, the mean based procedures showed a considerable shift in confidence interval due to transformation. The shift in confidence interval does not influence the coverage percentage as was seen in Study 3. However, the location of the confidence interval, in this case compared to zero, is different. Therefore the conclusions based on transformed variables can be different. Furthermore, in cardiovascular research the difference between rest and task scores is mostly used. Therefore, in Study 4 the used variables were difference variables between (logarithmically transformed) rest-scores and (logarithmically transformed) task-scores. However, for the robustness question, the distribution of the difference variable is actually only interesting. Unfortunately, no insight in the distribution of the difference variables is obtained. This should be investigated in further research.

In short, (a) transformation does not always lead to more normal distributions, (b) transformation does not always lead to better coverage percentages, (c) the coverage percentages of the Welch and the Satterthwaite procedure for cardiovascular data with transformation are as good as without transformation, and (d) due to transformation the confidence intervals of the mean based procedures are differently located compared to no transformation. Furthermore, untransformed variables are a little easier to interpret than transformed variables. All together, one could use the Welch (or the Satterthwaite) procedure on cardiovascular data without logarithmic transformation to get reliable results. The choice to use logarithmically transformed cardiovascular data could then only be founded on theoretical issues or may be related to other practical reasons.

#### **§4.3 Smooth ideal simulated data versus experimental simulated data**

Micceri (1989) mentioned that conclusions from robustness studies that use smoothed ideal data or mathematical functions cannot be applied in educational and psychological settings. Micceri pleaded that research to real-world data would be more useful to get a better understanding of the use of procedures in psychological experimental research. Therefore, the procedures were investigated on both ideal simulated data and on realistic experimental simulated data that are not ideally distributed (Study 3).

This research showed that smoothed, ideal simulated data can be used in educational and psychological settings. The conclusions drawn from Study 3 are very much in line with those from Study 1 that is based on smooth, ideal simulated data. Although the distributions of the two populations for the simulated experimental data were different, the performance could largely be deduced from what is known about the influence of non-normality, unequal population standard deviations, and unequal sample sizes from Study 1. Two typical examples: the Student's  $t$  procedure showed bad coverage percentages when the two populations were different and the sample sizes were unequal. The Welch procedure showed good coverage percentages in all conditions for all cardiovascular variables, as was expected from Study 1.

#### **§4.4 To conclude**

The goal of this study was to get more insight in the use of six procedures for comparing two or more samples. The influence of non-normality, unequal population distributions, and unequal sample sizes was investigated extensively. Therefore, better insight is obtained in the relation between the performance of a procedure in comparison to other procedures. Compared to most other robustness studies reported in the literature, this study used either more procedures or more conditions. Further, in this study confidence intervals and the associated coverage percentages are used instead of the probability of Type I error ( $\alpha$ ) as was used in most robustness studies. Confidence intervals are used rather than actual probability Type I errors, because they are currently considered the best reporting strategy (APA, 2001).

The results are very much in line with earlier robustness studies (e.g. Gans (1981); Gibbons & Chakraborti, 1991; Glass et al, 1972; Penfield, 1994; Ramsey, 1980; Scheffé, 1959; Wilcox, 2001; Zimmerman, 1987). The Student's  $t$  and the ANOVA procedure show a clear trend in their coverage percentages, as was expected from earlier studies. The Student's  $t$  and the ANOVA procedure perform well when the population standard deviations or the sample sizes are equal and bad when both are unequal. The Welch and the Satterthwaite procedures are in most cases the best procedures to use. The Student's  $t$ , the ANOVA, the Welch, and the Satterthwaite procedures appear more robust to deviations from normality than expected. The bootstrap procedures are no good

alternatives because several shortcomings are found. In earlier studies (e.g. Noreen, 1989; Wasserman & Bockenholt, 1989; Wilcox, 2001) this was already concluded about the bootstrap procedures, but now the shortcomings are clearly demonstrated in the comparison with other procedures in many conditions. Maybe the use of bootstrap procedures with robust estimators will perform better (e.g. Keselman, Wilcox & Lix, 2003; Lix & Keselman, 1998; Wilcox, 1997, 2001).

The design of this study was still limited. Many other alternative procedures can be used and also the use of rank scores and robust estimators are not taken into account. Furthermore, in Studies 1 and 2 the samples are always drawn from populations with the same type of non-normality. However, in Study 3 the samples were drawn from two differently distributed populations but not systematically varied. Study 3 did show results that pointed at an effect of using different non-normal distributions. The precise influence of comparing populations with different non-normal distributions on the procedures should be investigated in further research.

To conclude, for the conditions investigated in this study, the Welch and the Satterthwaite procedure perform as good as or much better than the Student's  $t$  and the ANOVA procedure. Therefore, it is concluded that the Welch procedure for comparing two samples and the Satterthwaite procedure for comparing more than two samples are the best procedures to use in experimental research.

## Appendices

### Appendix A: The procedures in formulas

Student's  $t$  procedure (Moore & McCabe, 2001):

The Student's  $t$  procedure uses pooled variances and a  $t$ -distribution with  $n_1+n_2-2$  degrees of freedom (df).

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad , s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\text{Confidence interval: } (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$t_{\alpha/2}$  look up in a  $t$ -table with  $df = n_1 + n_2 - 2$

ANOVA with linear contrasts (Snedecor & Cochran, 1980, p.228):

The ANOVA contrast analysis also uses pooled variances and a  $t$ -distribution with  $N-J$  degrees of freedom (df).

$$\Psi_A = \sum_{j=1}^J c_j \bar{x}_j \quad , SE(\Psi_A) = \sqrt{s_p^2 \sum_{j=1}^J \frac{c_j^2}{n_j}} \quad , s_p^2 = \frac{\sum_{j=1}^J (n_j - 1)s_j^2}{\sum_{j=1}^J (n_j - 1)}$$

$$\text{Confidence interval: } \Psi_A \pm t_{\alpha/2} SE(\Psi_A)$$

$t_{\alpha/2}$  look up in  $t$ -table with  $df = N - J$ , with  $N =$  subjects and  $J =$  groups.

Welch procedure (Moore & McCabe, 2001; Algina & Olejnik, 1984):

The Welch procedure uses weighted variances instead of pooled variances and a  $t$ -distribution with adjusted degrees of freedom ( $df_w$ ).

$$W = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}}, \quad df_w = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)}}$$

$$\text{Confidence interval: } (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}$$

$t_{\alpha/2}$  look up in  $t$ -table with  $df_w$ .

Satterthwaite procedure (Snedecor & Cochran, 1980, p.228):

The Satterthwaite procedure is a generalization of the Welch procedure, and uses weighted variances instead of pooled variances and a  $t$ -distribution with adjusted degrees of freedom ( $df_s$ ).

$$\Psi_s = \sum_{j=1}^J c_j \bar{x}_j, \quad SE(\Psi_s) = \sqrt{\sum_{j=1}^J \frac{c_j^2 s_j^2}{n_j}}, \quad df_s = \frac{\left(\frac{\sum_{j=1}^J c_j^2 s_j^2}{n_j}\right)^2}{\sum_{j=1}^J \frac{\left(\frac{c_j^2 s_j^2}{n_j}\right)^2}{(n_j - 1)}}$$

$$\text{Confidence interval: } \Psi_s \pm t_{\alpha/2} SE(\Psi_s)$$

$t_{\alpha/2}$  look up in  $t$ -table with  $df_s$ .

Bonett & Price (2002) procedure:

The samples' scores are ordered. For calculating the confidence interval the median and a distribution-free estimate of the variance of the median are determined. For calculating the median's distribution-free estimate of the variance, value  $z_j$  from Table 1 in Bonett & Price (2002) and a lower ( $Y_{(a_j)j}$ ) and an upper score ( $Y_{(n_j-a_j+1)j}$ ) are determined from the ordered scores.

$$\text{Median when odd: } \eta_j = Y_{\frac{(n+1)}{2}} \quad , \quad \text{median when even: } \eta_j = \frac{(Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1})}{2}$$

$$\Psi_{B\&P} = \sum_{j=1}^J c_j \eta_j \quad , \quad \text{with } \sum_{j=1}^J c_j = 0 \quad , \quad \text{var}(\eta_j) = \left( \frac{Y_{(n_j-a_j+1)j} - Y_{(a_j)j}}{2z_j} \right)^2$$

$Y_{(a)j} = a^{\text{th}}$  largest score of group  $j$ ,

$$a_j = \left( \frac{n_j + 1}{2} - \sqrt{n_j} \right),$$

$z_j$  (see Table 1 in Bonett & Price (2002)).

$$\text{Confidence interval: } \Psi_{B\&P} \pm z_{\alpha/2} \sqrt{\sum_{j=1}^J c_j^2 \text{var}(\eta_j)}$$

Bootstrap percentile procedure (Wilcox, 2001, p.94-99):

From the original sample  $B$  new samples are drawn with replacement, bootstrap samples. For each of the  $B$  bootstrap samples the location statistic (mean  $\bar{x}_j$  or median  $\eta_j$ ) is calculated. In case of comparing two groups the statistics are subtracted from each other resulting in differences of mean or medians. The  $B$  differences, give an approximation for the distribution of sample differences. The middle 95% of these differences are taken to get a 95% percentile interval which can be considered an approximation to a 95% confidence interval.

Bootstrap-Welch procedure (Wilcox, 2001, p.99-104):

For the bootstrap Welch procedure, B bootstrap samples are drawn with replacement from the original sample. For each of the B bootstrap samples the mean ( $\bar{x}_j^*$ ) and standard deviation ( $s_j^*$ ) are used to calculate the Welch value ( $W^*$ ). Next the B  $W^*$  values are ordered. The Welch procedure uses the samples' standard error and the new critical values to get the 95% confidence interval. The lower ( $W_L^*$  at 2.5%) and upper ( $W_U^*$  at 97.5%) values are set to be the new critical values for the 95% confidence interval based on the Welch procedure.

$$W_k^* = \frac{(\bar{x}_{1k}^* - \bar{x}_{2k}^*) - (\bar{x}_{1k} - \bar{x}_{2k})}{\sqrt{\frac{s_{1k}^{*2}}{n_1} + \frac{s_{2k}^{*2}}{n_2}}},$$

$\bar{x}_{jk}^*$  = mean of group  $j$  of the  $k^{\text{th}}$  bootstrap

$s_{jk}^*$  = standard deviation of group  $j$  of the  $k^{\text{th}}$  bootstrap

$$\text{Confidence interval} = \left[ (\bar{x}_1 - \bar{x}_2) - W_U^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) - W_L^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Bootstrap-Satterthwaite procedure for contrasts:

For the bootstrap Satterthwaite procedure with contrasts, the same steps are carried out as for the bootstrap Welch procedure.

$$\Psi_S = \sum_{j=1}^J c_j \bar{x}_j, \quad , SE(\Psi_S) = \sqrt{\sum_{j=1}^J \frac{c_j^2 s_j^2}{n_j}}$$

$$\Psi_k^* = \sum_{j=1}^J c_{jk} \bar{x}_{jk}^* \quad , SE(\Psi_k^*) = \sqrt{\sum_{j=1}^J \frac{c_{jk}^2 s_{jk}^{*2}}{n_j}} \quad , \psi^* = \frac{\Psi_k^* - \Psi}{SE(\Psi_k^*)}$$

$$\text{Confidence interval} = \left[ \Psi_S - \psi_U^* SE(\Psi_S), \Psi_S - \psi_L^* SE(\Psi_S) \right]$$

## Appendix B: Population distribution properties for Study 1

Study 1: Population distribution properties (npop=1.000.000) based on Ramberg' GLD (Ramberg et al.1979).

Distrib.*	Cond.**	Pop.	Mean	Median	$\sigma$	Skew.	Kurt.
1	1	1	0.0012	0.0021	1.0002	-0.0028	2.9992
1	1	2	0.8005	0.8011	0.9996	0.0050	3.0000
1	2	1	0.0032	0.0029	2.0007	0.0022	3.0076
1	2	2	0.8003	0.7995	1.0002	-0.0011	2.9960
1	3	1	0.0044	0.0000	3.9966	0.0028	2.9944
1	3	2	0.8003	0.8008	1.0018	-0.0024	2.9986
1	4	1	0.0011	0.0048	2.0015	-0.0053	2.9957
1	4	2	0.0013	0.0012	1.0003	0.0025	3.0011
2	1	1	0.0004	0.0003	1.0005	-0.0171	8.9182
2	1	2	0.8003	0.8000	1.0005	0.0224	9.1164
2	2	1	0.0028	0.0013	2.0021	-0.0142	8.8454
2	2	2	0.7997	0.7999	1.0013	0.0072	9.1150
2	3	1	-0.0005	-0.0003	3.9924	-0.0106	8.8295
2	3	2	0.8001	0.8009	0.9993	-0.0301	9.4746
2	4	1	0.0003	0.0018	2.0001	-0.0138	8.5928
2	4	2	-0.0021	-0.0021	1.0015	0.0010	9.4580
3	1	1	0.0001	0.0013	0.9997	-0.0016	1.8008
3	1	2	0.8006	0.8017	0.9999	-0.0010	1.7989
3	2	1	0.0004	0.0023	1.9991	-0.0013	1.8008
3	2	2	0.7993	0.7984	0.9998	0.0008	1.8012
3	3	1	0.0027	0.0030	4.0004	-0.0014	1.7989
3	3	2	0.7987	0.7983	0.9998	0.0006	1.7992
3	4	1	0.0014	0.0016	2.0003	-0.0008	1.7989
3	4	2	-0.0003	-0.0018	0.9992	0.0013	1.8022
4	1	1	0.0004	-0.1158	1.0004	0.6487	3.8054
4	1	2	0.7989	0.6822	0.9990	0.6534	3.8037
4	2	1	0.0023	-0.2309	2.0003	0.6501	3.7961
4	2	2	0.7991	0.6823	0.9987	0.6467	3.7878
4	3	1	-0.0033	-0.4697	3.9966	0.6504	3.7926
4	3	2	0.8002	0.6834	0.9996	0.6500	3.8072
4	4	1	-0.0031	-0.2350	1.9982	0.6496	3.8031
4	4	2	-0.0006	-0.1181	0.9997	0.6516	3.8007
5	1	1	-0.0020	-0.2784	0.9962	1.9904	9.5964
5	1	2	0.7966	0.5188	0.9971	1.9911	9.4930
5	2	1	-0.0043	-0.5593	1.9973	2.0082	9.6984
5	2	2	0.7999	0.5204	1.0032	2.0157	9.7542
5	3	1	-0.0014	-1.1137	3.9990	2.0050	9.7185
5	3	2	0.7983	0.5206	0.9986	1.9986	9.5747
5	4	1	0.0005	-0.5559	2.0020	2.0000	9.5893
5	4	2	-0.0022	-0.2811	0.9989	1.9934	9.4753

\* 1=normal, 2=leptokurtic, 3=platykurtic, 4=moderately skewed, 5= extremely skewed

\*\* see Table 2.2a in section Methods

## Appendix C: Population distribution properties for Study 2

Study 2: Population distribution properties (npop=1.000.000) based on Ramberg' GLD (Ramberg et al.1979).

Distrib.*	Cond.**	Pop.	Mean	Median	$\sigma$	Skew.	Kurt.
1	1	1	0.0012	0.0021	1.0002	-0.0028	2.9992
1	1	2	0.3005	0.3011	0.9996	0.0050	3.0000
1	1	3	0.6000	0.6007	0.9999	-0.0015	3.0065
1	1	4	0.9000	0.8985	1.0010	0.0010	2.9984
1	2	1	0.0063	0.0059	4.0014	0.0022	3.0076
1	2	2	0.3003	0.2995	1.0002	-0.0011	2.9960
1	2	3	0.6006	0.6013	1.0013	-0.0043	3.0019
1	2	4	0.9000	0.9001	1.0007	0.0039	2.9976
1	3	1	0.0044	0.0000	3.9966	0.0028	2.9944
1	3	2	0.3009	0.3025	3.0055	-0.0024	2.9986
1	3	3	0.6015	0.6002	1.9985	0.0018	3.0019
1	3	4	0.9007	0.9009	0.9994	-0.0017	2.9990
2	1	1	0.0002	0.0018	1.0002	0.0155	8.7781
2	1	2	0.3014	0.3009	1.0000	-0.0120	8.5739
2	1	3	0.5998	0.5997	1.0017	-0.0642	10.3640
2	1	4	0.9006	0.9008	0.9999	-0.0195	9.0891
2	2	1	0.0016	0.0012	4.0021	-0.0171	8.9182
2	2	2	0.3003	0.3000	1.0005	0.0224	9.1164
2	2	3	0.6002	0.5996	1.0002	0.0244	8.6620
2	2	4	0.9006	0.9009	0.9984	-0.0275	8.6359
2	3	1	0.0056	0.0026	4.0042	-0.0142	8.8454
2	3	2	0.2991	0.2997	3.0039	0.0072	9.1150
2	3	3	0.6007	0.6002	2.0006	-0.0098	9.1493
2	3	4	0.8988	0.8990	1.0001	0.0271	9.0122
3	1	1	-0.0003	-0.0001	0.9991	0.0015	1.8014
3	1	2	0.3004	0.3017	1.0000	-0.0010	1.7998
3	1	3	0.6001	0.5999	1.0000	0.0005	1.7995
3	1	4	0.9016	0.9024	0.9998	-0.0017	1.7992
3	2	1	0.0014	0.0066	4.0028	-0.0003	1.7983
3	2	2	0.2979	0.2961	1.0003	0.0022	1.7987
3	2	3	0.6002	0.5997	0.9999	-0.0002	1.8010
3	2	4	0.8998	0.8993	0.9999	0.0011	1.7998
3	3	1	0.0005	0.0051	3.9986	-0.0016	1.8008
3	3	2	0.3017	0.3050	2.9996	-0.0010	1.7989
3	3	3	0.5979	0.5959	1.9997	0.0026	1.8002
3	3	4	0.8986	0.8995	1.0008	0.0009	1.7979
4	1	1	-0.0002	-0.1161	0.9992	0.6483	3.8048
4	1	2	0.2992	0.1820	0.9993	0.6478	3.7894
4	1	3	0.6004	0.4834	1.0002	0.6502	3.8022
4	1	4	0.9018	0.7855	1.0000	0.6529	3.8187
4	2	1	0.0019	-0.4656	3.9985	0.6490	3.8040
4	2	2	0.2982	0.1819	0.9980	0.6440	3.7839
4	2	3	0.6011	0.4833	0.9999	0.6522	3.8054
4	2	4	0.8992	0.7826	0.9990	0.6509	3.8063
4	3	1	0.0033	-0.4655	4.0002	0.6524	3.8098
4	3	2	0.2994	-0.0544	2.9975	0.6527	3.8041
4	3	3	0.5987	0.3668	1.9990	0.6486	3.7910
4	3	4	0.9003	0.7815	1.0017	0.6513	3.7930
5	1	1	-0.0008	-0.2781	0.9999	2.0043	9.6243
5	1	2	0.2975	0.0203	0.9984	2.0045	9.6296
5	1	3	0.5992	0.3213	1.0004	2.0145	9.7547
5	1	4	0.8991	0.6209	1.0003	2.0033	9.6085
5	2	1	-0.0007	-1.1110	3.9997	2.0001	9.6212
5	2	2	0.2976	0.0203	0.9969	1.9918	9.5434
5	2	3	0.5998	0.3210	1.0014	2.0103	9.6580
5	2	4	0.8993	0.6206	0.9998	1.9892	9.4105
5	3	1	-0.0088	-1.1177	3.9917	1.9899	9.4267
5	3	2	0.2966	-0.5362	2.9979	2.0108	9.7605
5	3	3	0.5973	0.0401	2.0021	2.0173	9.8346
5	3	4	0.8973	0.6199	0.9972	2.0009	9.5784

\* 1=normal, 2=leptokurtic, 3=platykurtic, 4=moderately skewed, 5= extremely skewed

\*\* see Table 2.3a in section Methods

## References

- Algina, J., Olejnik, S.F. (1984). Implementing the Welch-James procedure with factorial designs. *Educational and Psychological Measurement*, 44, p.39-48.
- Althaus M., Roon A.M. van, Mulder, L.J.M., Mulder, G., Aarnoudse, C.C. & Minderaa, R.B. (2004). Autonomic response patterns observed during the performance of an attention-demanding task in two groups of children with autistic-type difficulties in social adjustment. *Psychophysiology*, 41, p.893-904.
- APA (2001) *Publication Manual of the American Psychological Association* (5<sup>th</sup> ed.). APA, Washington.
- Boneau, C.A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 57, p.49-64.
- Bonett, D.G. & Price, R.M. (2002). Statistical inference for a linear function of medians: confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, 7, p.370-383.
- Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, p.144-152.
- Cochran W.G. (1954). Some methods for strengthening the common  $\chi^2$  tests. *Biometrics*, 10, p.417-451.
- Efron B., Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.
- Gans, D.J. (1981). Use of a preliminary test in comparing two sample means. *Communications in Statistics, Simulation and Computation*, B10 (2), p. 163-167.
- Gibbons, J.D. & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student's *t*, and alternate *t* tests for means of normal distributions. *The Journal of Experimental Education*, 59, p.258-267.
- Glass, G.V., Peckham, P.D. & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, p.237-288.

- Harwell, M.R., Rubinstein, E.N., Hayes, W.S., Olds, C.C. (1992). Summerizing Monte Carlo results in methodological research: The one-and two-factor effects ANOVA cases. *Journal of Educational Statistics*, 17, p. 315-339.
- Keselman, H.J., Huberty C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C. & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, p.350-386.
- Keselman, H.J., Wilcox, R.R., Lix, L.M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology*, 40, p.586-596.
- Lix, M.L., Keselman, H.J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, 58, p. 409-429.
- Lix, M.L., Keselman, J.C., Keselman, H.J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, 66, p. 579-619.
- Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Reviews*, 23, p.151-169.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, p.156-166.
- Miller, R.G. (1986). *Beyond ANOVA, basics of applied statistics*. New York: John Wiley & Sons.
- Moore, D.S. & McCabe, G.P. (1999). *Introduction to the practice of statistics* (3<sup>o</sup> ed.). New York: W.H. Freeman.
- Moser, B.K. & Stevens, G.R. (1992). Homogeneity of Variance in the Two-Sample Means Test. *The American Statistician*, 46, p.19-21.
- Noreen, E.W. (1989). Computer intensive methods for testing hypotheses. An introduction. John Wiley & Sons:New York.

- Pearson, E.S. & Please, N.W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, 62, p.223-241.
- Penfield, D.A. (1994). Choosing a Two-Sample Location Test. *Journal of experimental education*, 62, p343-360.
- Ramberg, J.S., Dudewicz, E.J., Tadikamalla, P.R. & Mykytka, E.F. (1979). A probability distribution and its uses in fitting data. *Technometrics*, 21, p.201-214.
- Ramberg, J.S. & Schmeiser, B.W. (1974). An appropriate method for generating asymmetric random variables. *Communications of the Association for Computing Machinery*, 17, p.78-82.
- Ramsey, P.H. (1980). Exact type 1 error rates for robustness of student's t test with unequal variances. *Journal of Educational Statistics*, 5, p.337-349.
- Roon, A.M. van (1998). *Short-tem cardiovascular effects of mental tasks. Physiology, experiments and computer simulations*. Proefschrift, Rijksuniversiteit Groningen.
- Roon, A.M. van, Mulder, L.J.M., Althaus, M. & Mulder, G. (2004). Introducing a baroreflex model for studying cardiovascular effects of mental workload. *Psychophysiology*, 41, p.961-981.
- Sawilowsky, S.S., Clifford Blair, R. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin*, 111, p.352-360.
- Scheffé, N. (1959). *The Analysis of Variance*. John Wiley & Sons, New York.
- Snedecor, G.W. & Cochran, W.G. (1980). *Statistical methods* (7<sup>th</sup> ed.). Ames: Iowa State University Press.
- SPSS (2001). *SPSS for windows: Statistical package for the social sciences* (version 11.0.1).
- Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3<sup>rd</sup> ed.). Lawrence Erlbaum Associates, Mahwah.

- Sutton, C.D. (1993). Computer-intensive methods for tests about the mean of an asymmetrical distribution. *Journal of the American Statistical Association*, 88, p.802-810.
- Wal, W.M. van der (2004). Een Monte Carlo-vergelijking van statistische toetsen voor twee onafhankelijke steekproeven met behulp van gesimuleerde reactietijden. Masterthesis, Rijksuniversiteit Groningen.
- Wasserman, S. & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology*, 26, p.208-221.
- Wilcox, R.R. & Charlin, V.L. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics*, 11, p.263-274.
- Wilcox, R.R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.
- Wilcox, R.R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer.
- Zimmerman, D.W. (1987). Comparative power of Student's t test and Mann-Whitney U test for unequal sample sizes and variances. *The Journal of Experimental Education*, 55, p.171-174.
- Zimmerman, D.W. (1987). Failure of the Mann-Whitney test: A note on the simulation study of Gibbons and Chakraborti (1991). *The Journal of Experimental Education*, 60, p.359-364.