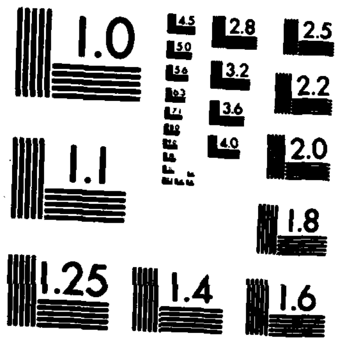
AD-A142 980 BI- AND POLY- OPTIMAL CONFIDENCE LIMITS FOR A LOCATION AND PARAMETER(U) PRINCETON UNIV NJ DEPT OF STATISTICS S MORGENTHALER NOV 83 TR-254-SER-2 ARO-19442.15-MA 1/1

UNCLASSIFIED DAAG29-82-K-0178 F/G 12/1 NL


1.0
4.5
5.0
2.8
2.5
3.2
2.2
3.6
4.0
1.1
2.0
1.8
1.25
1.4
1.6

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD-A142 980

DTIC FILE COPY

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ARO 19442.15-MA
2. GOVT ACCESSION NO.: N/A
3. RECIPIENT'S CATALOG NUMBER: N/A
4. TITLE (and Subtitle): "Bi- and poly- optimal confidence limits for a location parameter." Technical Report # 254, series 2.
5. TYPE OF REPORT & PERIOD COVERED:
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Stephan Morgenthaler
8. CONTRACT OR GRANT NUMBER(s): ~~DE-AC02-81ER10841~~ DAAG29-82-K-0178
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Statistics Department, Princeton University, Princeton, N.J. 08544
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709
12. REPORT DATE: November, 1983
13. NUMBER OF PAGES: 53
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

DTIC ELECTE JUL 13 1984 E

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this report we define poly-optimal confidence intervals for a locatio parameter. The formulas are given for the case of two shapes, but can easily be extended to the case of many shapes.

For the case of two situations, the Gaussian and the slash, the resulting family of confidence interval estimators is examined. These interval estimators are competitors of existing so-called robust procedures. A comparison to a few of these is included.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

84 07 12 042

Bi- and poly- optimal confidence limits
for a location parameter

by

Stephan Morgenthaler

Technical Report No. 254, Series 2
Department of Statistics
Princeton University
November 1983

*Prepared in connection with research at Princeton University, sponsored by the Army Research Office (Durham). The computing facilities were provided by the Department of Energy, Contract DE-AC02-81ER10841.

November 15, 1983

Bi- and poly- optimal confidence limits for a location parameter

Stephan Morgenthaler
Technical Report No. 254, Series 2
Department of Statistics
Princeton University
Princeton, New Jersey 08544

ABSTRACT

In this report we define poly-optimal confidence intervals for a location parameter. The formulas are given for the case of two shapes, but can easily be extended to the case of many shapes.

For the case of two situations, the Gaussian and the slash, the resulting family of confidence interval estimators is examined. These interval estimators are competitors of existing so-called robust procedures. A comparison to a few of these is, included.

1. Introduction.

This report deals with the issue of robustness in interval estimation for a location parameter. We will restrict attention to location-and-scale equivariant estimators. This puts us automatically into the theory connected with configurations (see Morgenthaler (1983)). Of special relevance to our problem are the conditional confidence distributions, which allow us to determine -- for the sampling situation(s) under consideration -- the conditional confidence coefficient given the configuration for any interval

Prepared in connection with research at Princeton University, sponsored by the Army Research Office (Durham). The computing facilities were provided by the Department of Energy, Contract DE-AC02-81ER10841.

November 30, 1983

estimator. We will see in the second section how these conditional confidence distributions can be employed to define "good" confidence limits. These poly-optimal -- or in our case bi-optimal -- interval procedures are then compared to existing robust methods. (Section 3).

Here we understand the essence of robustness in a sense similar to what it means in the point estimation case, i.e. high efficiency in a variety of underlying situations. And we will put special emphasis on small sample results instead of asymptotics. This allows us on one hand to be a lot more realistic but on the other hand we can not take an infinity of situations into simultaneous consideration. But -- as we will learn -- there is a lot of potential in this approach. It can teach us new things.

2. Bi-optimal confidence intervals for a location parameter.

We are interested in location-and-scale equivariant confidence limits. This means that our upper and lower bound statistics will satisfy

$$U(s(t\,\vec{1} + \vec{c})) = s(t + U(\vec{c})),$$

where $\vec{c} \in \mathbb{R}$ and $\vec{1}$ is the vector consisting of ones. Under location and scale changes of the configuration $\vec{c}$, the statistic behaves accordingly. From this equivariant behavior it follows immediately that for samples of the form

$$\vec{y}(s,t) = s(t\,\vec{1} + \vec{c})$$

the value of the statistic U is known if the value $U(\vec{c})$ alone is fixed. For each two-dimensional set of samples which only differ by location and scale changes we, therefore, select a representing element $\vec{c}$ -- <u>configuration</u> -- which serves as a base point in parametrizing the set of samples by $s \in \mathbb{R}$, and $t \in \mathbb{R}$ (see Morgenthaler (1983)). The conditional density given the configuration $\vec{c}$ can then be written as a function of s and t, which turns out to be

$$k_F(s,t|\vec{c}) = \frac{s^{n-1} \prod_{i=1}^{n} f(s(c_i+t))}{\int_0^{\infty} \int_{-\infty}^{\infty} s^{n-1} \prod_{i+1}^{n} f(s(c_i+t))\, dsdt},$$

where $(c_1, c_2, \ldots, c_n) = \vec{c}$ is the configuration and $F(\ )$ and is the probability distribution we sample from $(\frac{d}{dx}F(\ ) = f(\ ))$.

If we are interested in confidence limits, the <u>conditional confidence distribution</u> which gives us the conditional coverage probabilities is important. It has the form

$$\begin{aligned} Co_F(U) &= P_F[U(\vec{y}) > 0] = P_F[s(t+u) > 0] \\ &= \int_0^{\infty} \int_{-u}^{\infty} k_F(s,\ t|\vec{c})\ dsdt\ . \end{aligned}$$

Here we assumed that the distribution F is symmetric with center of symmetry at 0 -- then $Co_F(u)$ gives the conditional probability of the upper bound statistic $U(\ )$ actually being an upper bound for the true center of symmetry if $U(\vec{c}) = u$.

If $U(\ )$ and $L(\ )$ are upper and lower bound statistics with $U(\vec{c}) = u$ and $L(\vec{c}) = 1$, the conditional confidence distribution tells us the conditional coverage probability if we would sample from situation F . It would be

$$\alpha_F(\vec{c}) = 1-[(1-Co_F(u)) + Co_F(1)] = Co_F(u)-Co_F(1)$$

since $1-Co_F(u)$ is the conditional probability of missing the true parameter to the left and $Co_F(1)$ is the conditional probability of missing to the right.

We will now derive interval procedures which are single-situation optimal. Then we will go on to bi-optimal procedures and indicate how to proceed to poly-optimal methods. All of these methods are optimal in a small sample sense.

2.1. Single-situation-shortest confidence intervals

We might ask for the confidence interval -- in any given situation -- which has minimal expected length for this situation, and reaches a pre-fixed confidence coefficient. This leads in situation F to the following problem:

$$\text{minimize} \int E_F[s|\vec{c}]\ [U(\vec{c})-L(\vec{c})]\ d\mu_F(\vec{c})$$

with respect to $U(\vec{c})$ and $L(\vec{c})$ under the condition that

$$\int \{Co_F(U(\vec{c}))-Co_F(L(\vec{c}))\}\ d\mu_F(\vec{c}) = 1 - \alpha$$

We note that $d\mu_F(\ )$ is the (n-2)-dimensional measure across configurations induced by F and that
$E_F[s|\vec{c}]\ [U(\vec{c})-L(\vec{c})] = E_F[s(U(\vec{c})-L(\vec{c}))|\vec{c}]$ is the expected length

conditioned on the configuration of the confidence interval induced by $L(\vec{c})$ and $U(\vec{c})$. Introducing a Lagrange multiplier λ and assuming interchangeability of integration and differentiation, the solution to this problem is of the form:

$$E_F[s|\vec{c}] = \lambda\, co_F(U(\vec{c}))$$
$$E_F[s|\vec{c}] = \lambda\, co_F(L(\vec{c}))$$
$$\int\{Co_F(U(\vec{c})) - Co_F(L(\vec{c}))\}\, d\mu_F(\vec{c}) = 1 - \alpha$$

(where $co(u) = \frac{d}{du} Co(u)$).

In an experimental sampling setup the derivation of the corresponding solution is somewhat simpler and we include it here. The problem consists of the following:

$$\text{minimize } \frac{1}{N}\sum_{i=1}^{N} E_F[s|\vec{c}_i][U(\vec{c}_i)-L(\vec{c}_i)]$$

with respect to the numbers $U(\vec{c}_i)$, $L(\vec{c}_i)$, under the condition that

$$\frac{1}{N}\sum_{i=1}^{N}\{Co_F^i(U(\vec{c}_i)) - Co_F^i(L(\vec{c}_i))\} = 1-\alpha$$

In this notation $\{\vec{c}_1,\ldots,\vec{c}_N\}$ denotes the set of all configurations sampled from situation F -- hence we just replaced $d\mu(\ )$ by the empirical measure which puts a point mass $\frac{1}{N}$ on each of the $\vec{c}_i$'s. The solution is now straightforward and gives:

$$E_F(s|\vec{c}_k) = \lambda\, co_F^k(U(\vec{c}_k)) \quad ,k=1,\ldots,N$$
$$E_F(s|\vec{c}_k) = \lambda\, co_F^k(L(\vec{c}_k)) \quad ,k=1,\ldots,N$$
$$\frac{1}{N}\sum_{i=1}^{N}\{Co_F^i(U(\vec{c}_i)) - Co_F^i(L(\vec{c}_i))\} = 1-\alpha$$

In order to compute the bounds $L(\vec{c}_k)$, $U(\vec{c}_k)$, k=1,...N, one would fix

λ and find the inverses

$$L(\vec{c}_k) = co_{F,k}^{-1}\left(\frac{E_F[s|\vec{c}_k]}{\lambda}\right)$$

$$U(\vec{c}_k) = co_{F,k}^{-1}\left(\frac{E_F[s|\vec{c}_k]}{\lambda}\right) \tag{2.1}$$

and then check the overall coverage probability

$$\frac{1}{N}\sum_{i=1}^{N}\{Co_F^i(U(\vec{c}_i)) - Co_F^i(L(\vec{c}_i))\}$$

if this value is below $1-\alpha$ one has to try with a bigger value of λ and vice versa.

<u>remark</u>: $co_{F,k}^{-1}(\ )$ is not a well defined function. With $L(\vec{c}_k)$ we denote the smallest solution to the equation in x

$$co_F^k(x) = \frac{E_F[s|\vec{c}_k]}{\lambda}$$

and with $U(\vec{c}_k)$ the largest.

In order to get a short interval -- short measured by expected length -- we see from (2.1) that the coverage densities have to be <u>cut at equal height</u> adjusted by $E_F[s|\vec{c}_k]$, which takes care of the scale differences between the class-representing configurations $\vec{c}_k$.

In the Gaussian case the interval described above is identical with Student's t interval. in the next few sub-sections We will examine what sort of confidence intervals we get if we choose F = slash, i.e. a heavy-tailed symmetric distribution (see Rogers & Tukey (1972)).

2.1.1. Samples of size 20

The single-situation shortest interval with 95% coverage probability has an expected length of 2.237 in the slash situation. <u>This will be used throughout as a reference to compute efficiencies</u>. The conditional shortest (see Morgenthaler (1983)) has an expected length of 2.245 ($\pm$.02) and its slash "squared mean length efficiency", defined by

$$\frac{(\text{minimum expected length in slash})^2}{(\text{expected length of interval in slash})^2} \qquad (2.2)$$

is $\frac{(2.237)^2}{(2.245)^2}$ = 99.3% (see Horn(1981) on discussion of criterion for confidence intervals!).
We will often report excesses instead of efficiencies. These two are linked by

$$1 + \text{excess} = \frac{1}{\text{efficiency}} \qquad (2.3)$$

The conditional shortest has therefore an excess of $\frac{1}{.993} - 1$ = .7%. It is obvious that for a single situation we can without harm in terms of expected length, ask for a fixed conditional confidence coefficient.
The range of the distribution of conditional coverage for the 150 slash- drawn configurations is from 91.4% to 97.7% with an estimated standard deviation of about 1%.

2.1.2. Samples of size 10 and 5

Again we restrict attention to the procedure which will produce the shortest expected length in the slash case. In terms of swapping coverage probability between configurations, there is not much change if we go down the ladder of sample sizes. The 5-number summaries (see: Tukey(1977)) for the conditional coverage distributions across configurations are as follows:

		size=20	
#150			
M		95.2%	
H	94.7%		95.5%
	91.4%		97.7%

		size=10	
#150			
M		95.2%	
H	94.7%		95.5%
	88.3%		96.3%

		size=5	
#500			
M		95.2%	
H	94.9%		95.6%
	82.8%		97.7%

The three cases are close, the lower extreme is going down with decreasing sample size.

The minimal expected length for the slash in samples of size 10 is 3.604 ($\pm$.013) and in samples of size 5 it is 6.641 ($\pm$.025). These together with the Gaussian expected length of Student's t will be used in the next section as minimum expected length for the Gaussian and the slash situation. We should be aware that these confidence intervals are "single-situation" in their spirit. The slash optimum will be anticonservative if applied in the Gaussian situation, whereas Student's t will be conservative in the slash situation.

2.2. Bi-shortest confidence intervals

In this section we will derive confidence intervals which are robust in the sense that they will not be influenced unduly by outliers. In configurations which contain "outlying" points

Student's t-interval will be rather long and we plan to use the slash situation in order to provide guidance in shorteining Student's t-interval for such configurations. In doing this, we will of course have to pay a price. The conditional confidence coefficient for the Gaussian situation will be rather low and if we still want to reach 100(1-α)% overall confidence level, we will have to enlarge Student's t-interval in other configurations in order to have more than 100(1-α)% conditional coverage probability. So shortening confidence intervals naturally leads to exchange of coverage probability between configurations. Robustness of validity, i.e. of coverage probability and robustness of efficiency are two concepts which need balancing. If we understand the validity in an overall manner, we can ask for confidence intervals which are "short", but still reach an overall confidence coefficient of 100(1-α)% in both situations.

Let us define "shortest" in terms of expected length. This is by no means an obvious choice, since length distributions of confidence intervals are skewed and the expected value has no intuitive meaning. As we will see, this choice makes things simple for us. But as we will also learn it might be of interest to look at alternative definitions of "shortness". Any criterion which can be written as an expected value over the sample space can be handled in the same way as "expected length".

What can we expect from a confidence interval procedure, if we look at it from two sampling situations at the same time? Certainly there will be no procedure which is simultaneously optimal for both situations. If the optimality criterion is "convex", however, there is a one-dimensional family of procedures, any of which is such that it cannot be improved in both situations simultaneously.

In a decision theoretic framework (see Ferguson(1967)) our "parameter set" consists of two values {Gaussian, slash} and the risk of any interval procedure is defined through what we called a criterion.

remark: Since we look at equivariant "decisions" the risk within the Gaussian and the slash, i.e. under changes of the location and scale parameter, is constant or depends in a simple manner on the scale parameter.

If we use the expected length of the confidence interval, the value of the scale parameter σ will turn up as a multiplier.

Let us look at the "general picture" if we adopt expected length as our criterion. In order to avoid the trouble with the scale parameter σ, we will choose a canonical density in each of the two location-and-scale families and compute the expected length using these canonical forms. This results in no loss of generality. The risk set is:

$$R=\{(r_1,r_2):r_1=\text{Gaussian exp. length},r_2=\text{slash exp. length}\}\in\mathbb{R}^2$$

where the expected lengths are taken over the set of valid confidence intervals, i.e. intervals which reach overall at least $100(1-\alpha)\%$ coverage probability in both situations. For the usual reasons this risk set is convex: if we have two valid confidence interval procedures I_1 and I_2 the convex linear combinations $\lambda I_1 + (1-\lambda)I_2$ will be valid intervals too and the expected lengths will be convex linear combinations of the expected lengths of I_1 and I_2 for both the Gaussian and the slash.

If we want to get something which does not depend on the scale parameter within the Gaussian and the slash, we go to excesses, i.e. reciprocals of efficiencies minus 1. Since the Gaussian efficiency is the ratio

$$\mathrm{eff}_G(I) = \frac{(\text{min. length in Gaussian})^2}{(\text{Gaussian exp. length of } I)^2}$$

the scale parameter drops out. If we look at excess sets, they will be convex for the same reason.

remark: Furthermore the risks and efficiencies defined through expected length or through (expected length)^2 lead to exactly the same boundary procedures ("admissible solutions"), since we have merely performed a monotone transformation.

> This then leads to a one-dimensional family of bi-optimal procedures. Each of these confidence intervals has a right to be called optimal, since there is no other interval estimator which dominates it in the "two-situation world" according to the chosen criterion.

Any member of this one-dimensional family is characterized by a ratio of what economists call "shadow prices". Let p_g and p_s denote the shadow prices for the Gaussian and the slash, respectively. The bi-optimal confidence interval procedure corresponding to the shadow price ratio $\frac{p_s}{p_g}$ is then found as the solution to the following restrained minimization problem:

minimize p_g(Gaussian exp. length) + p_s(slash exp. length)

under the condition that the Gaussian and slash coverage is greater than or equal to $100(1-\alpha)\%$

or

$$\text{minimize } p_g \int E_g[s|\vec{c}\,][U(\vec{c}\,) - L(\vec{c}\,)]\,d\mu_g(\vec{c}\,) + p_s \int E_s[s|\vec{c}\,][U(\vec{c}\,) - L(\vec{c}\,)]\,d\mu_s(\vec{c}\,)$$

with respect to $U(\vec{c}\,)$ and $L(\vec{c}\,)$ under the condition that

$$\int \{Co_g(U(\vec{c}\,)) - Co_g(L(\vec{c}\,))\}\,d\mu_g(\vec{c}\,) \geq 1-\alpha$$
$$\int \{Co_s(U(\vec{c}\,)) - Co_s(L(\vec{c}\,))\}\,d\mu_s(\vec{c}\,) \geq 1-\alpha.$$

The subscript g refers to the Gaussian situation, the subscript s to the slash. $d\mu(\ )$ denotes the (n-2)-dimensional measure across configurations and the other functions and symbols are as described before.

In order to write down a sampling version of this minimization problem, it is essential to realize the fact that any configuration could arise from either the Gaussian or the slash (or actually any absolutely continuous sampling situation with infinite support). So if we have a stock of configurations drawn from the Gaussian, we can still learn something about the slash performance of any "statistical procedure" by applying a weight appropriate for the slash to the "slash answers" of these configurations. This kind of poly-sampling and the choice of reasonable relative weights is discussed in

Pregibon and Tukey (1981) and we will not elaborate on it here.

Now we are ready to rewrite the minimization problem in sampling terms, which will allow us to find approximate solutions.

$$\text{Minimize} \quad \frac{p_g}{N}\sum_{i=1}^{N} w_g^i E_g[s|\vec{c}_i][U(\vec{c}_i)-L(\vec{c}_i)] +$$

$$\frac{p_s}{N}\sum_{i=1}^{N} w_s^i E_s[s|\vec{c}_i][U(\vec{c}_i)-L(\vec{c}_i)] \tag{2.4}$$

with respect to $U(\vec{c}_i)$ and $L(\vec{c}_i)$ under the condition that

$$\frac{1}{N}\sum_{i=1}^{N} w_g^i\{Co_g^i(U(\vec{c}_i))-Co_g^i(L(\vec{c}_i))\}\geq 1-\alpha$$
$$\frac{1}{N}\sum_{i=1}^{N} w_s^i\{Co_s^i(U(\vec{c}_i))-Co_s^i(L(\vec{c}_i))\}\geq 1-\alpha.$$

The summation here runs over the whole set of sampled configuration, i.e. both Gaussian-drawn and slash-drawn. The relative weights w_g and w_s are used to correct for the fact that not all configurations are sampled from the correct situation -- they indicate the weight attributed to a certain configuration in answering questions about the Gaussian or the slash, respectively. All the other symbols are as above.

The step from the "continuous" formulation of the problem to the "sampling" formulation involves an approximation of

$$d\mu_g(\vec{c})$$

by putting point mass $\frac{w_g^i}{N}$ onto the "point" $\vec{c}_i$.

What constraints do the solutions of minimization problem (2.4) fulfill? The minimum will occur on the "boundaries" where the overall Gaussian coverage probability and the overall slash coverage probability are equal to $100(1-\alpha)\%$ except in special cases. If both restraints have to be met we need to introduce two Lagrange multipliers λ_g and λ_s and the solution takes the following form.

$$\frac{\lambda_g w_g^k co_g^k(U(\vec{c}_k))+\lambda_s w_s^k co_s^k(U(\vec{c}_k))}{\lambda_g w_g^k+\lambda_s w_s^k} = \frac{p_g w_g^k E_g[s|\vec{c}_k]+p_s w_s^k E_s[s|\vec{c}_k]}{\lambda_g w_g^k+\lambda_s w_s^k}$$

$$\frac{\lambda_g w_g^k co_g^k(L(\vec{c}_k))+\lambda_s w_s^k co_s^k(L(\vec{c}_k))}{\lambda_g w_g^k+\lambda_s w_s^k} = \frac{p_g w_g^k E_g[s|\vec{c}_k]+p_s w_s^k E_s[s|\vec{c}_k]}{\lambda_g w_g^k+\lambda_s w_s^k}$$

$$k=1,\ldots,N \qquad (2.5)$$

and

$$\frac{1}{N}\sum_{i=1}^{N} w_g^i\{Co_g^i(U(\vec{c}_i))-Co_g^i(L(\vec{c}_i))\} = 1-\alpha$$

$$\frac{1}{N}\sum_{i=1}^{N} w_s^i\{Co_s^i(U(\vec{c}_i))-Co_s^i(L(\vec{c}_i))\} = 1-\alpha$$

$co_g^i(x)$ denotes the derivative $\frac{d}{dx}Co_g^i(x)$.

This is a set of 2N + 2 equations which have to be satisfied simultaneously. The left hand side of the first 2N equations is the density of a mixture of the two coverage densities with weights $\lambda_g w_g^k$ and $\lambda_s w_s^k$. If we denote this mixture by $h_k(\ )$, the solution can be computed by inverting $h_k(\ )$ as in (2.1) which leads to

$$U(\vec{c}_k) = h_k^{-1}\left(\frac{p_g w_g^k E_g[s|\vec{c}_k]+p_s w_s^k E_s[s|\vec{c}_k]}{\lambda_g w_g^k+\lambda_s w_s^k}\right) \qquad k=1,\ldots,N$$

$$L(\vec{c}_k) = h_k^{-1}\left(\frac{p_g w_g^k E_g[s|\vec{c}_k]+p_s w_s^k E_s[s|\vec{c}_k]}{\lambda_g w_g^k+\lambda_s w_s^k}\right) \qquad k=1,\ldots N \qquad (2.6)$$

(see remark after (2.1))

and λ_g & λ_s such that

$$\frac{1}{N}\sum_{i=1}^{N} w_g^i\{Co_g^i(U(\vec{c}_i))-Co_g^i(L(\vec{c}_i))\} = 1-\alpha$$

$$\frac{1}{N}\sum_{i=1}^{N} w_s^i\{Co_s^i(U(\vec{c}_i))-Co_s^i(L(\vec{c}_i))\} = 1-\alpha.$$

This way we can find $U(\vec{c}_k)$ and $L(\vec{c}_k)$, $k=1,\ldots,N$ by an algorithm which adjusts the values of λ_g and λ_s after each iteration so that the overall coverage probability is equal to $100(1-\alpha)\%$ in both situations.
The solution is again of an "equal height" form similar to the single situation shortest confidence intervals. But now a mixture between the Gaussian and the slash "picture" - with the shadow prices p_g and p_s and the Lagrange multipliers λ_g and λ_s as weights -- is characterizing the conditional "knowledge".

The case where the minimum solution to (2.4) occurs on the single-boundary, where only the Gaussian coverage is at its lower bound, but the slash coverage is bigger than $100(1-\alpha)\%$, occurs, in particular with $p_g = 1$ and $p_s = 0$, i.e. shadow price ratio 0. This means that we are only concerned about the Gaussian expected length of the procedure and we know from that Student's t interval minimizes this length. However, the slash coverage is then bigger than $100(1-\alpha)\%$ at least for the common values of α. If we would force a solution with minimum Gaussian length and exact $100(1-\alpha)\%$ coverage in both situations, the Gaussian expected length would increase. This might at first sight seem paradoxical: if we want to bring down the slash confidence coefficient and thus "shorten" the intervals, the expected length in the Gaussian increases. But the idea that decreasing the confidence coefficient always results in shortening confidence intervals is wrong. In the above case we will force $100(1-\alpha)\%$ slash coverage by introducing asymmetry into Student's t intervals, which makes them on average longer in the Gaussian situation. The solution to the one-boundary case, i.e. only the Gaussian confidence coefficient is fixed, is as in (2.5) with $\lambda_s = 0$.

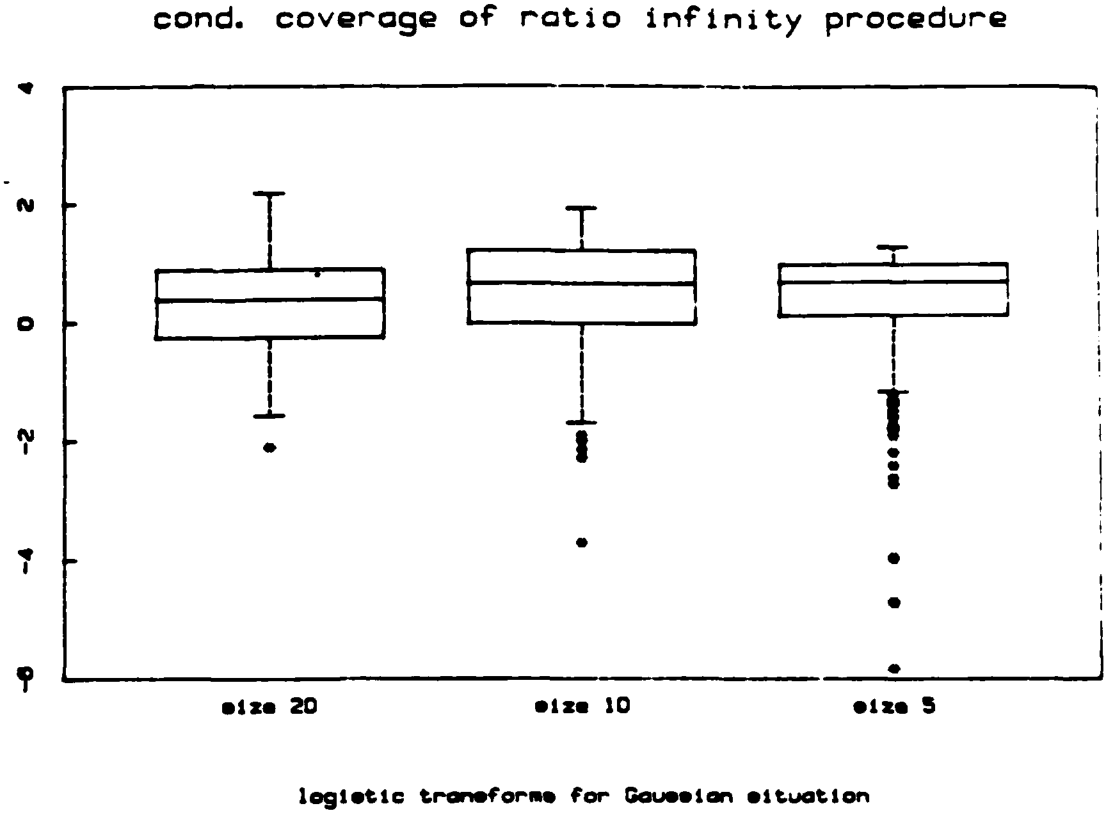
In solving the minimization problem (2.4) we will therefore always have to check whether putting λ_s equal to zero will improve on the "objective" function. This "paradoxical" behavior only happens if the shadow price ratio is sufficiently close to 0, i.e. if our "objective" function discounts the slash expected length sufficiently.

Let us first study the confidence interval procedure which results from putting $p_g = 0$, i.e. the shadow price ratio equal to infinity. These are the intervals which have shortest expected length in the heavy-tailed slash, but are also reaching $100(1-\alpha)\%$ coverage in the Gaussian. As always we will restrict attention to the 95% confidence level case.

2.2.1. $p_s = 1$ and $p_g = 0$: Shortest in slash (ratio infinity)

These intervals exhibit a considerable amount of coverage probability exchanging between configurations. Figures 2.1 and 2.2 show boxplots (see Tukey(1977)) of the conditional confidence coefficients for samples of configurations drawn from the Gaussian

Figure 2.1: Bi-shortest intervals in the Gaussian situation



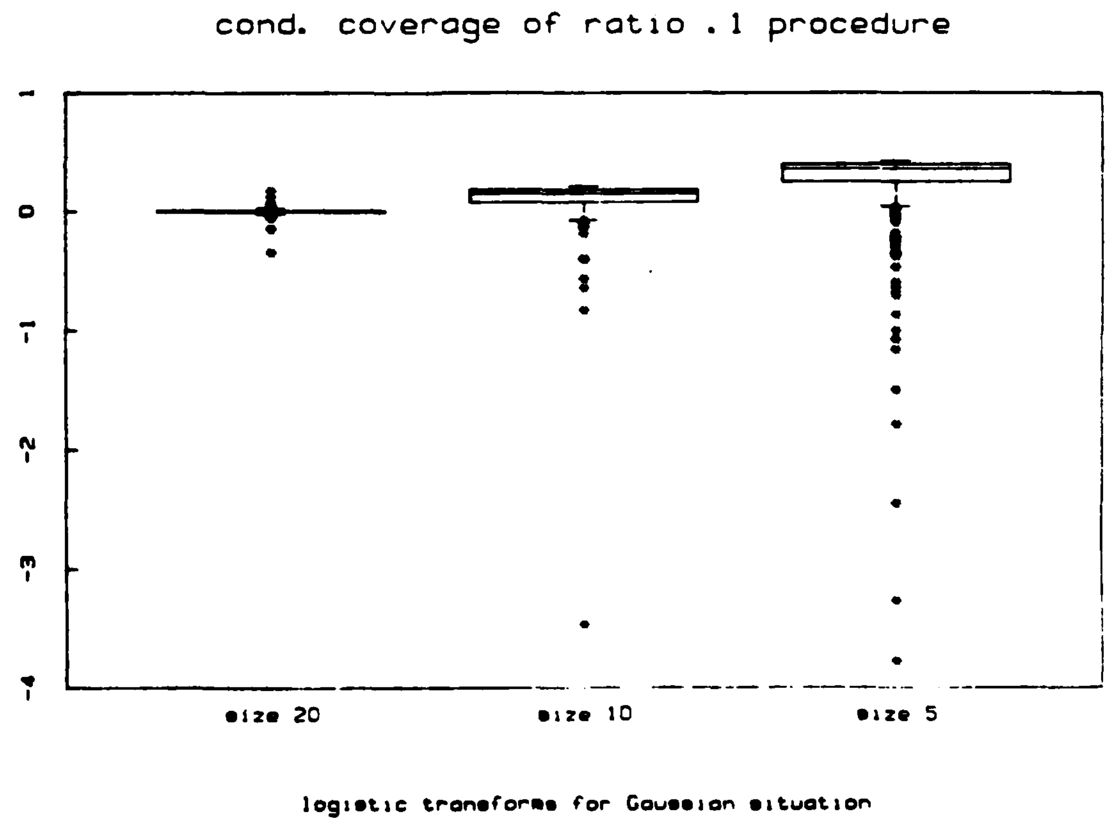
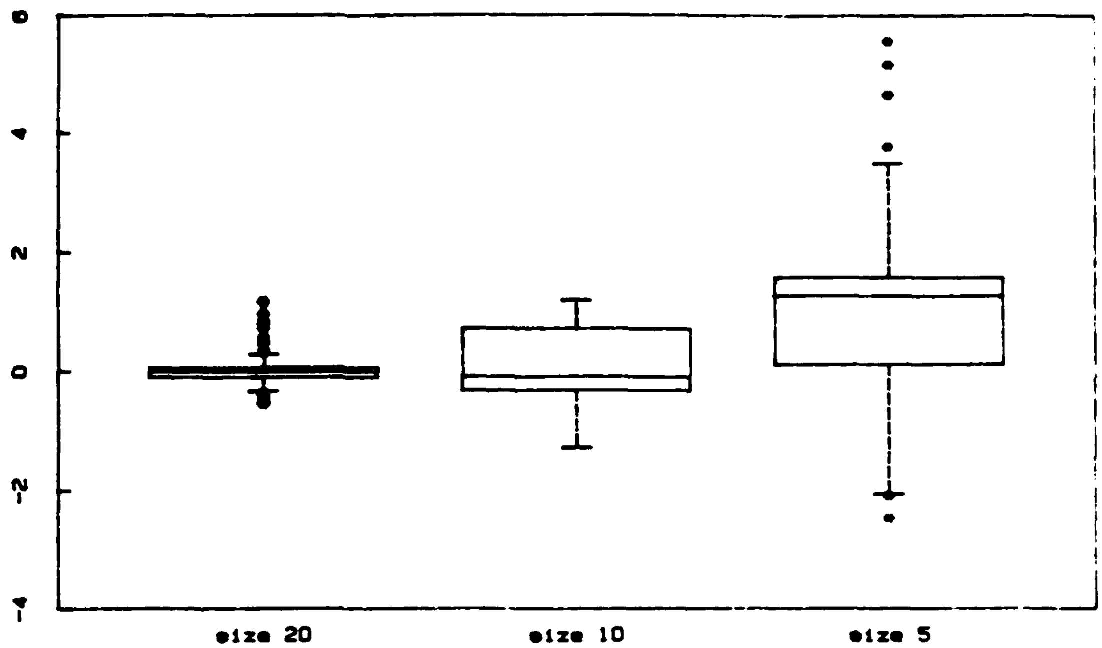


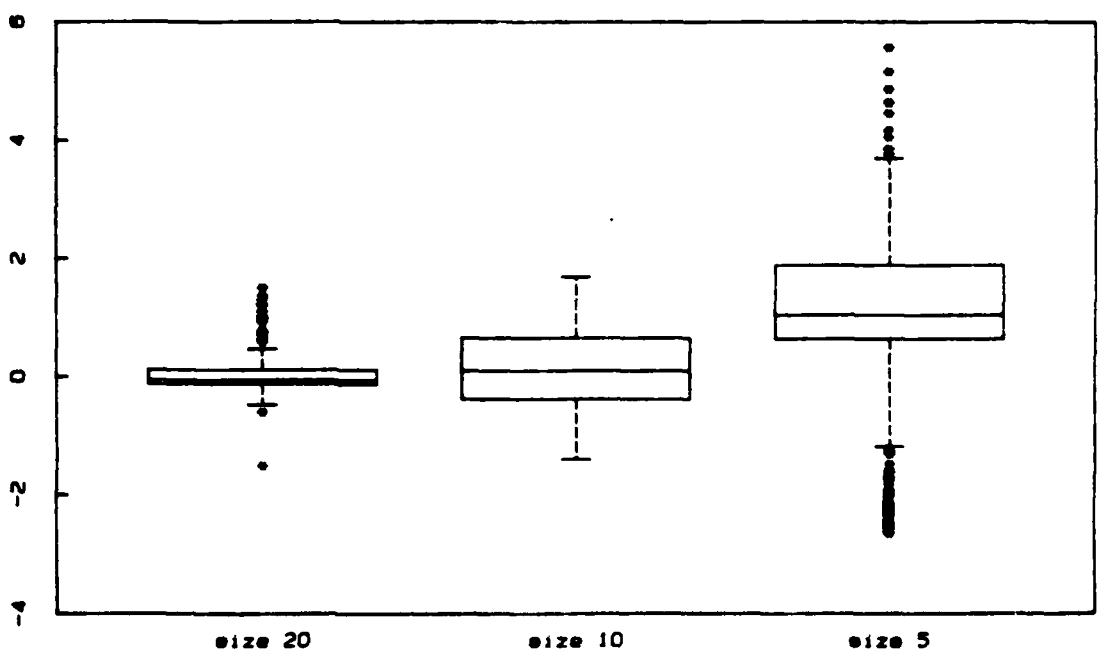
Figure 2.2: Bi-shortest intervals in the slash situation

cond. coverage of ratio infinity procedure



logistic transforms for slash situation

cond. coverage of ratio .1 procedure



logistic transforms for slash situation

and the slash, respectively. In order to make these exhibits more informative we do not display the raw confidence coefficients but rather a logistic transform

$$\log\left(\frac{\alpha(\vec{c})\ .05}{(1.-\alpha(\vec{c}))\ .95}\right)$$

of the conditional confidence level $\alpha(\vec{c})$.

In the Gaussian situation we see that as the sample size increases, the tail towards very low conditional confidence coefficients grows and at the same time the bulk of the distributions moves closer together. In the slash situation, Figure 2.2, the changes across sample sizes are more complex. In samples of size 20 we get a good slash-behavior if we modify for Gaussian coverage. If we decrease the sample size to 10, the tail of the distribution towards "over-coverage" has thickened considerably and as we get to 5, even the median coverage has moved to about 98.6%. The lower tail, i.e. towards "under-coverage" also grows with decreasing sample size.

We can think of these intervals as modified "shortest" slash -- modified to pick up additional Gaussian coverage in the most economical way. It is not surprising that modifying slash for additional Gaussian coverage is asking for more in smaller sample sizes (see Morgenthaler (1983)). The values of the Lagrange multipliers -- which together with the relative sampling weights are weighting the coverage densities in (2.5) -- are revealing in this respect. The ratio $\frac{\lambda_s}{\lambda_g}$ has the values 24.3, 1.8, 0.2 as the sample size goes from n=20, n=10 to n=5. In the case "n=20" the solution

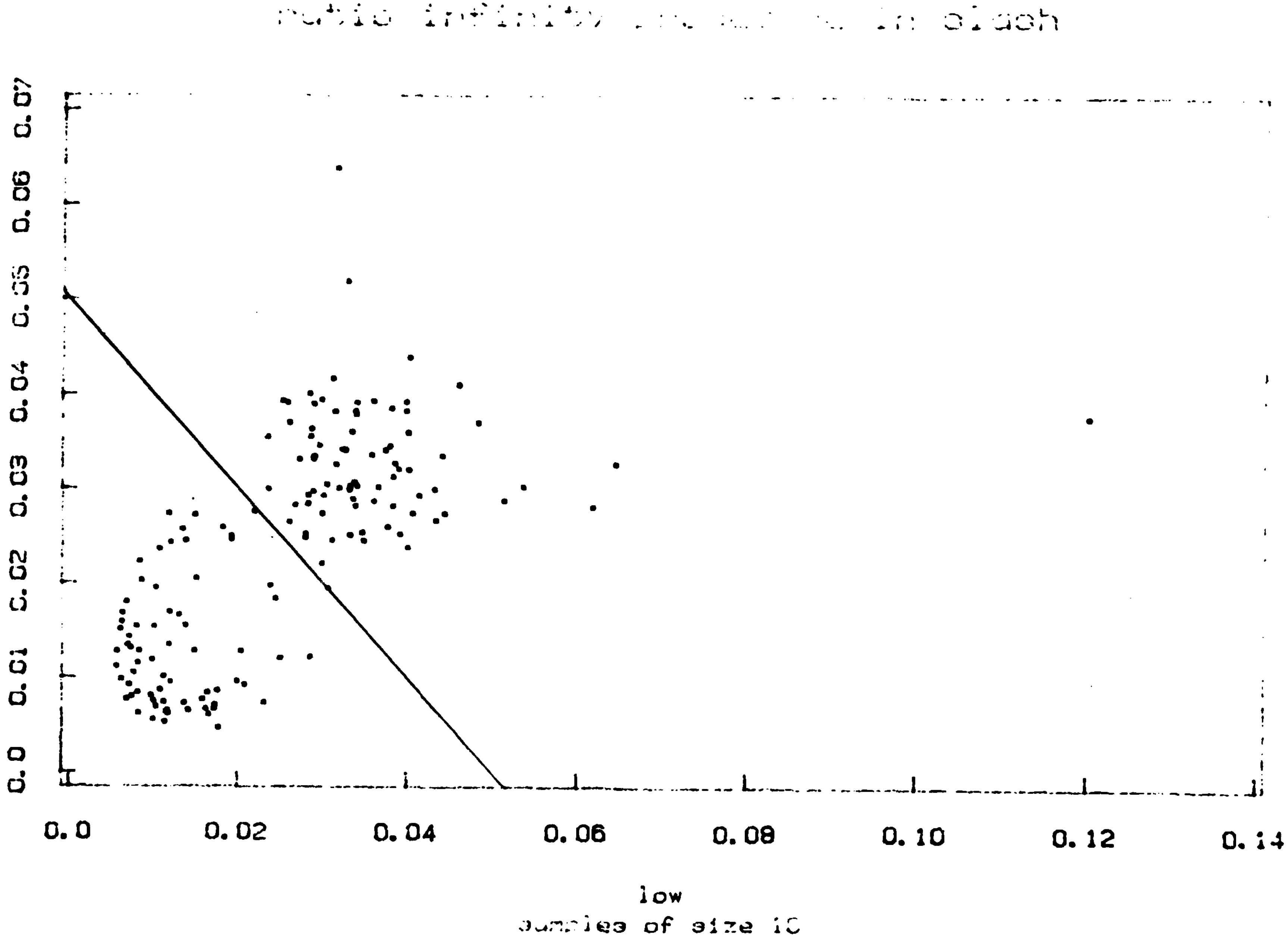
pays most attention to the slash whereas in "n=5" the Gaussian has to be taken into account with big weight. So even it we are solely interested in the slash expected length, the demand for 95% Gaussian coverage puts a lot of emphasis on the Gaussian. We expect to see this behavior also in the length distribution.
And indeed it turns out that the ratio of slash expected length to the single situation shortest has values 100.6%, 104.4% and 129.8% for the cases n=20, 10 and 5. So the penalty in terms of increased expected length we have to pay in order to get 95% Gaussian coverage increases with decreasing sample size.

Let us get back now to the slash coverage probability. Figure 2.3 shows the plot of the conditional probabilities of missing the true parameter value to the left vs. to the right. Clearly for samples of size 10 the slash configuration population is split into two parts -- roughly two halves. One half of the configurations gets long intervals, i.e. over-confidence, whereas the other half gets too short intervals. This feature can be understood by looking at equations (2.5). In a configuration with relative slash weight w_s big compared to w_g, the equation approximately reduces to

$$co_s^k(U(\vec{c}_k)) = \frac{1}{\lambda_s} E_s[s|\vec{c}_k] \quad k=1,\ldots,N$$

which is the same as (2.1) except that the Lagrange multipliers are possibly different. The Lagrange multiplier λ required in (2.1), i.e. for the single-situation-shortest slash intervals are 11.3, 20.7 and 41.3 for the cases n=20, 10 and 5. The values for λ_s in the above equation on the other hand are 10.7, 14.8 and 10.0, i.e. uniformly smaller. In a configuration with big relative slash weight the

Figure 2.3: Plot of the lower vs. the upper conditional missing probabilities of the bi-shortest interval with shadow price ratio infinity



interval will therefore be smaller -- due to the decreased Lagrange multiplier -- than for the single situation optimal solution. Figure 2.3 tells us that this happens in roughly half the slash-drawn configurations. Now we understand how the slash optimal intervals are modified to yield 95% Gaussian coverage. In configurations where we strongly "believe" in the slash sampling -- in terms of relative weight, i.e. compared to the Gaussian - the intervals are shortened. In other configurations the intervals have to be made longer, more nearly like Student's t. This effect of shortening seems very undesirable from the conditional coverage point of view, but is needed if we insist on 95% slash coverage. Maybe the more natural approach would leave these intervals at their single- situation optimum -- which would of course result in a slash overall confidence coefficient bigger than 95%.

As the comparison of the Lagrange multipliers suggests, this is not or in a limited way going on in the case "n=20". Indeed the single-situation-optimal intervals are there nearly the same as the modified ones which also guarantee 95% Gaussian coverage. In the case "n=5" on the other hand, the problem is getting really extreme but the configurations with relative slash weight dominating are getting rarer. But it is obvious that in order to have both the slash and the Gaussian confidence levels at exactly 95%, we need to make the intervals very short in configurations where we strongly "believe" (w_s big!) in slash sampling.

Why does the sample size have such a strong influence? In order to answer this question the (n-2)-dimensional measure

$d\mu_{F(\vec{c})}$ which goes across configurations and only depends on the shape has to be brought into the discussion. As the sample size increases the "overlap" between $d\mu_{\text{slash}}$ and $d\mu_{\text{Gaussian}}$ decreases, i.e. it is getting easier to discriminate between the two. This is the reason why the modification of the single-situation-optimal slash interval in order to gain Gaussian coverage does not influence the slash behavior very much, since the modifications take place in configurations quite far from the "core" of $d\mu_{\text{slash}}$.

The same plot as Figure 2.3 for the Gaussian case shows that the modified shortest slash intervals are in a few configurations very short and in the majority too long -- even from the Gaussian point of view. This can be explained by the "urge" of this interval estimator to be short in configurations which "look slash like" and long in others.

But over the whole we can certainly say that these modified shortest-slash intervals are not what we could call "good", mainly from the point of view of the conditional confidence behavior.

2.2.2. $p_s = .1$ and $p_g = 1$: "robust", but short in the Gaussian (ratio .1)

As we have already discussed, the solution for the case $p_s = 0$ and $p_g = 1$ just leads us to the familiar t-intervals. It should, however, be interesting to see how the interval procedures which solve problem (2.4) with small nonnegative ratio $\frac{p_s}{p_g}$ -- we choose the ratio .1 -- behave. Obviously these will be closer to Student's t

intervals without losing sight of the slash "requirements".
It turns out that in the case "n=20" the idea we have in the back of our mind - shortening Student's t intervals radically in "extreme" configurations while leaving them alone in others -- works. At ratio .1 there is very little variation left of the conditional confidence levels in the Gaussian situation. Figures 2.1 and 2.2 allow a comparison with the ratio = oo intervals. In the Gaussian case, both tails are moved in, whereas in the slash situation, the tail towards over-coverage grows.

In samples of size 10, one can still see the splitting of the slash drawn configurations into subpopulations, but the plot corresponding to Figure 2.3 has been loosened up.
If we care more for the Gaussian situation, these are clearly more sensible confidence intervals and in terms of expected length they have to be favored over Student's t.

2.2.3. The bi-optimal curves

Figures 2.4, 2.5 and 2.6 show the plots of the square mean length deficiencies, defined by

$$\mathrm{def}_F(I) = \left(\frac{\text{exp. length in situation F of interval I}}{\text{min. exp. length in situation F}}\right)^2 - 1$$

for the two situations Gaussian and slash.
The "*" denote the nonparametric procedures of discussed in Morgenthaler (1983), where the labels are "si" for the sign, "wi" for the Wilcoxon, "w#" for the winsorized Wilcoxon with a bound of # on the ranks. The bi-optimal procedures were computed for the ratio

Figure 2.4: Plot of deficiencies for nonparametric and bi-shortest interval procedures in samples of size 20

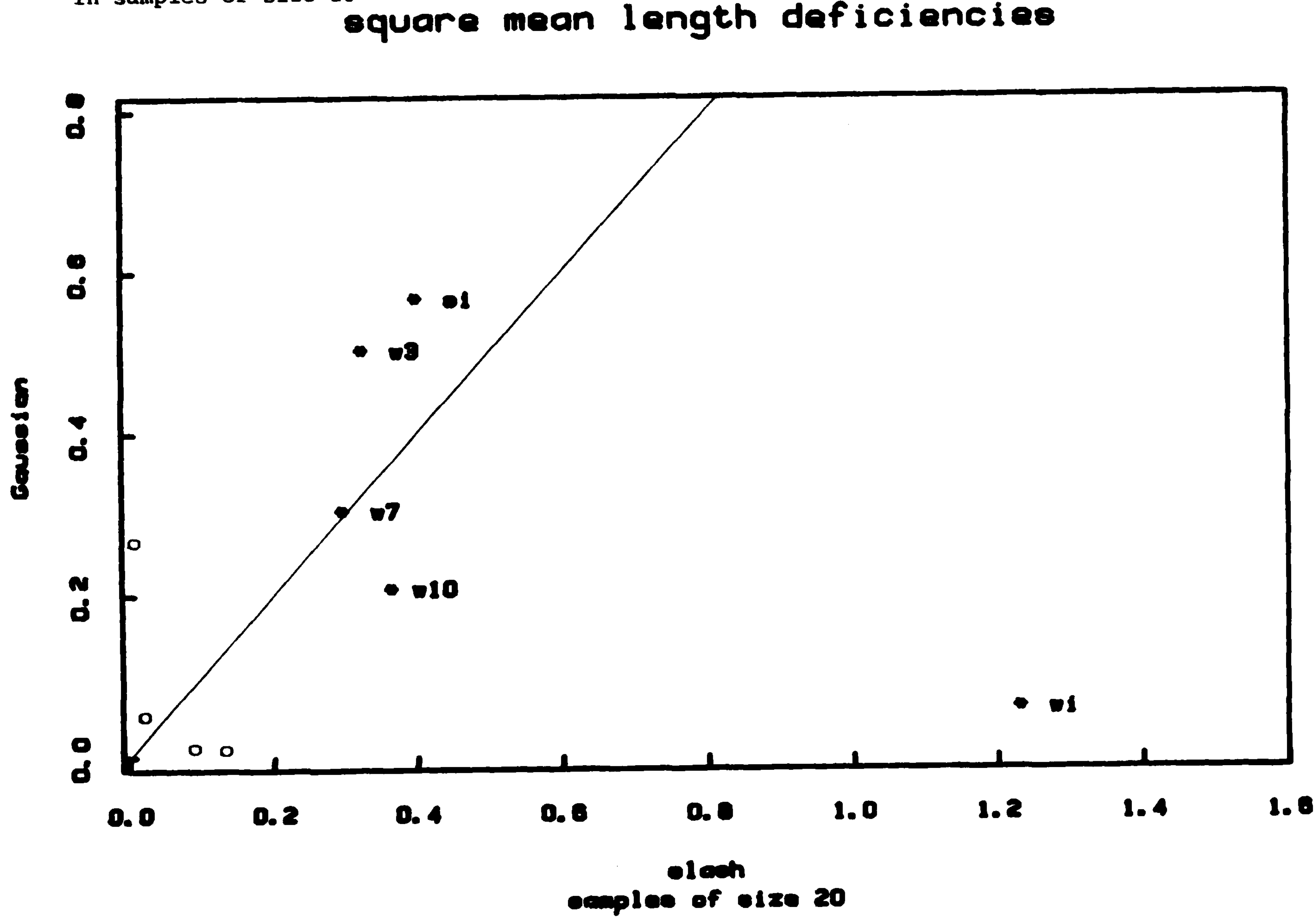


Figure 2.5: Plot of deficiencies for nonparametric and bi-shortest interval procedures in samples of size 10

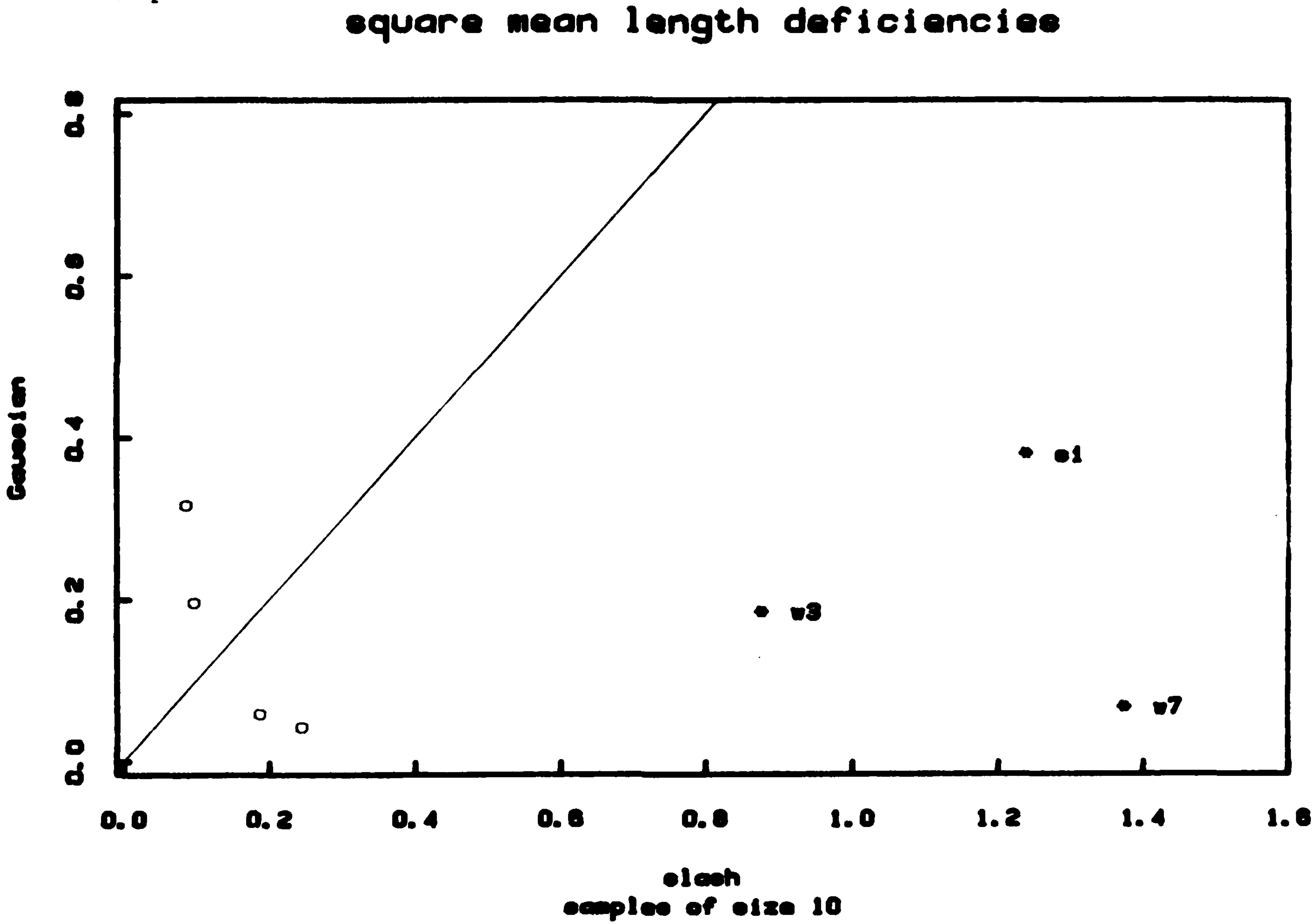
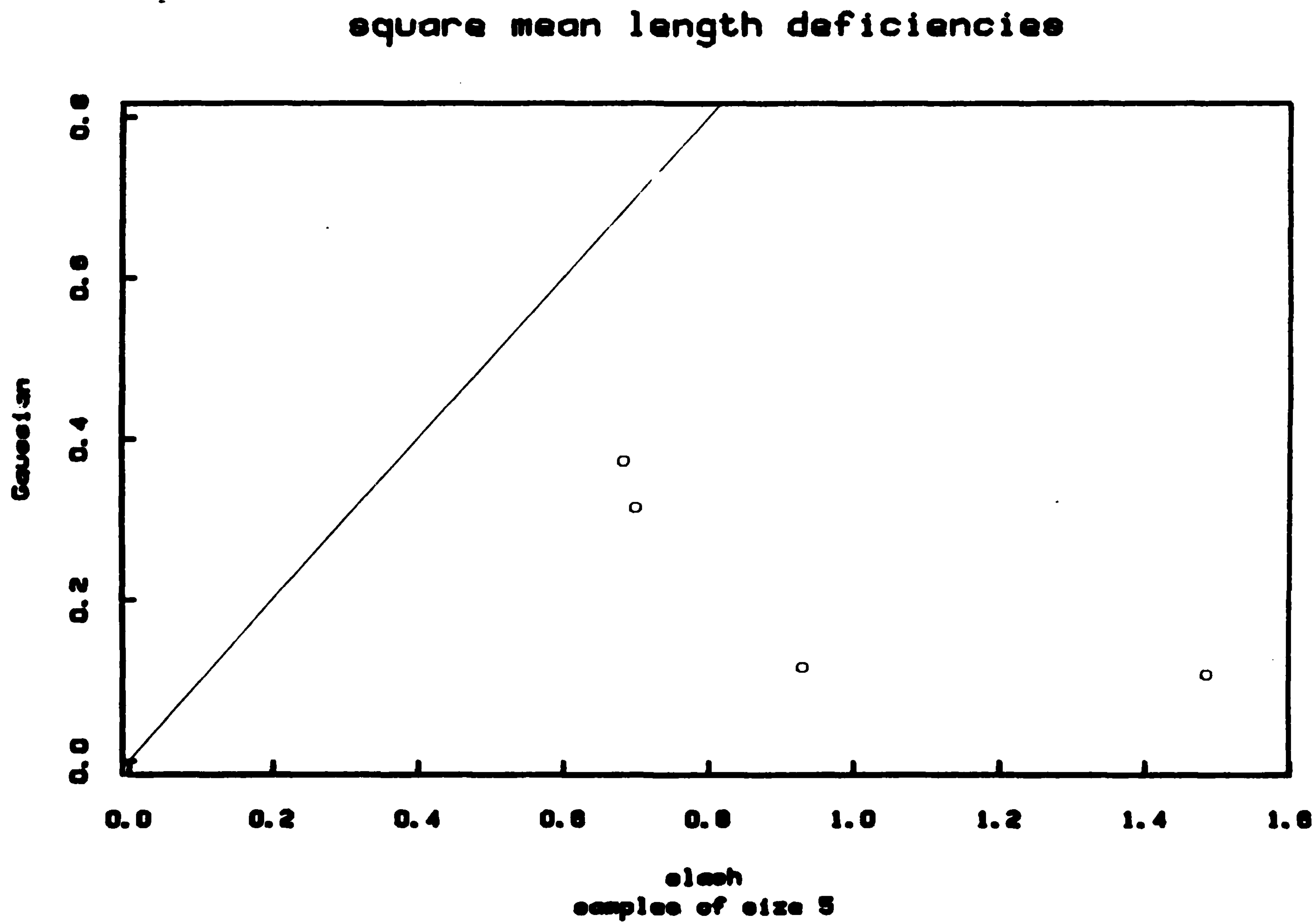


Figure 2.6: Plot of deficiencies for nonparametric and bi-shortest interval procedures in samples of size 5



values oo, 2, .2 and .1 and plotted with an "o". The diagonal -- corresponding to the minimax choice -- is included.

Clearly we are able to do an excellent job of compromising in samples of size 20. The minimax bi-optimal confidence interval reaches about 96.5% squared mean length efficiency, so that the expected length of this procedure is very close to the single-situation-optimal intervals. In samples of size 10 the curve moves towards the right, we have to pay a penalty in terms of increased slash expected length due to the fact that we require a confidence coefficient of 95% in the Gaussian situation. The minimax choice now has approximately 87.3% squared mean length efficiency. And a good compromise as far as expected length is concerned is still possible.

In samples of size 5 the slash-penalty we have to pay for gaining 95% Gaussian coverage probability is getting very large. The following Table gives the numbers:

Table 2.1: Estimated expected lengths of several confidence interval procedures

method	Gaussian			slash		
	size=5	size=10	size=20	size=5	size=10	size=20
Student's t	2.33	1.39	.93	-	-	-
opt. slash	-	-	-	6.64	3.60	2.24
ratio oo	2.74	1.60	1.05	8.62	3.76	2.25
ratio 2	2.68	1.52	0.95	8.66	3.76	2.26
ratio .2	2.47	1.43	0.95	9.23	3.93	2.34
ratio .1	2.45	1.42	0.93	10.46	4.02	2.38
si	-	1.64	1.16	-	5.39	2.65
w3	-	1.52	1.14	-	4.94	2.57
w7	-	1.44	1.06	-	5.55	2.55
w10	-	-	1.02	-	-	2.61
wi	-	1.46	0.955	-	24.5	3.34
hin	4.07	1.73	1.21	11.95	4.65	2.57
hub-1.5	3.74	1.63	1.03	11.88	4.25	2.57
hub-1.9	3.69	-	-	11.78	-	-
bi-9	3.82	1.52	0.97	12.48	4.29	2.68
bi-11	3.29	-	-	11.28	-	-
tp	2.17	1.35	0.92	11.95	4.14	2.41
wms	3.42	1.62	0.99	9.87	3.96	2.33

(the standard errors in this table are between $\frac{1}{2}$% and 3% of the estimates! The procedures labelled hin, hub-1.5, hub-1.9, bi-9, bi-11,tp and wms are discussed in section 3.)

The labels in this table stand for the following interval procedures: ratio # = bi-shortest with specified shadow price ratio, si = sign, w# = "winsorized" Wilcoxon score, wi = Wilcoxon, hin = pivot-t, hub = procedure based on Huber's ρ-function, bi = one-step biweight procedure, tp = three-point procedure and wms = procedure based on the weighted conditional mean-square-error curve.

Figures 2.4 and 2.5 also include the nonparametric confidence intervals discussed in (Morgenthaler (1983)). They too pay a slash penalty for reaching 95% Gaussian confidence level as the sample size decreases. The pictures only include the winsorized Wilcoxons where one puts a bound on the ranks, the trimmed Wilcoxon's (see Morgenthaler (1983)) give, however, nearly the same confidence intervals. The modified Wilcoxon's smoothly bridge the gap between the sign and the unmodified Wilcoxon and -- from the point of view of expected length -- are a preferable choice. As we have seen in (Morgenthaler (1983)), this is also true from the point of view of conditional confidence coefficients.

2.3. Discussion

We should be careful in interpreting the deficiency plots. Unfortunately the confidence interval estimation problem is more complex than the point estimation problem, where a deficiency plot derived from mean-square-errors, we believe, tells us nearly all. (However, we have not looked into matters in as much detail for the point estimate case!)
Here other aspects have to be taken into account. If we look at the variation of the conditional confidence levels across configurations, we get the following table.

Table 2.2: Hinge-spreads (see Tukey(1977)) for conditional coverage probabilities in %

		Gaussian	slash
ratio oo	size=20	4.23%	0.79%
	size=10	3.52%	4.28%
	size=5	2.51%	3.33%
ratio .1	size=20	0.06%	1.16%
	size=10	0.44%	4.50%
	size=5	0.52%	1.91%

It is obvious that the two choices of ratios have diametrically opposed consequences -- and that the small ratio makes more sense from the point of view of stability of conditional confidence levels. It is interesting to note that the bi-modal slash conditional confidence level distribution we get for samples of size 10 clearly stands out. If we compare this to to the numbers in Morgenthaler (1983), we note that the nonparametric confidence intervals seem to rely on somewhat more coverage-probability exchange between configurations. If we were to measure the variation in these tables by a statistic which uses more of the tail information - like the usual standard deviation -- we would, however, see that the bi- optimized intervals grow quite a heavy tail towards low conditional confidence levels. Trying to make the expected lengths of the confidence interval procedure small does of course have an impact on the distribution of the conditional coverage-probabilities across configurations.

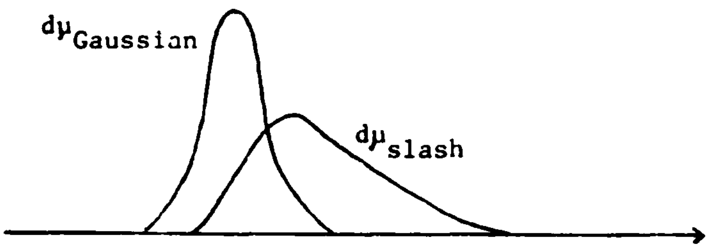
3. ROBUST CONFIDENCE INTERVALS

What should we mean by the term robustness if we use it in connection with confidence intervals? Most of the robustness literature (see Huber(1981)) is concerned with point estimation -- and the simplest case i.e. location parameter estimation with known scale parameter is certainly best understood.
If we deal with confidence intervals, or with the related tests, several complications arise. It is my belief that asymptotic theory loses some of its appeal when we apply it to confidence intervals. As the size of the sample goes to infinity, the problem of setting confidence limits gradually disappears. If we knew the population, our interval would be of zero length so that as the sample size gets big, most of the confidence interval estimation problem lies in finding a "good" center for it and we are really talking about point estimation. The usual way to get around this is of course to study powers of the tests at alternatives which tend towards the null hypothesis -- that way we can study the asymptotic length of the corresponding confidence interval procedure. _It seems to me that interval estimation is inherently a "sample problem"_.
P. Huber derives in his book an approach to get minimax intervals for the case of known scale (Huber(1981)), but this is too simple a situation to be helpful in practice.

If we use the center and width (or range) of the interval as co-ordinates, it certainly seems necessary for a "robust" interval estimator to have a robust center and a robust width -- but both alone do not satisfy us, since we also have to keep the validity of the procedure under control. This requires that the width gets large whenever the center is "weak" and in this sense the two co-ordinates center and width have to react in a matched way. Basically this means that their ratio has a distribution which does not change catastrophically much in the tails if the underlying situation changes. Student's t interval does satisfy this requirement -- and is robust in this sense.

In what way are the bi-optimal procedures superior? They basically try a small sample minimax approach for two underlying situations. In configurations where the two points of view that our situations supply are in disagreement, we use relative weights in compromising the two. This gives us a robust answer in the sense that for both "models" we do, globally, the best we can. The danger lies in configurations where some other situation -- not included in our two-situation analysis -- would have a high relative weight if it were included and would possibly give a very different answer. From the minimax point of view -- as advocated by P. Huber -- where in order to be realistic, one has a "real" neighborhood around the "model" (and not an infinitesimal "thing") and plays minimax inside, our approach of can be criticized. We took only two situations into account -- which were "far apart" -- and used essentially a minimax type of estimate. We are, however, not sure about the behavior between -- or to one side of -- the two chosen situations. Heuristically our proposed intervals from will safeguard us against many heavy-tailed underlying situations. Since the slash relative weight would dominate the Gaussian relative weight, we would be inclined to choose the slash answer in configurations drawn from any heavy-tailed situation. In this sense bi-optimal procedures are robust and safe to use.

What we have said above points towards the following features of the configural approach. The (n-2)-dimensional distributions $d\mu_F(\ )$ across configurations are giving "breadth" to our robustness claim.



The above picture shows in a schematic way what is going on. Both situations -- the Gaussian and the slash -- span a certain region of configurations and if we were to include other situations, we would probably end up spanning more and having a wider basis where our procedures work reliably. Note also that with increasing sample sizes the distributions over configurations get more and more concentrated -- we noticed for example how the overlap between the Gaussian and the slash pretty much disappears in samples of size 20.

The second important aspect of the configural approach is the conditional distribution given the configuration under varying situations. Here we could -- if only we knew how -- use a conditional minimax approach. By computing the conditional distributions for different situations -- which may lead to completely different answers -- we recognize the need for compromising and also get guidance in the direction and amount of the required adjustments. The relative weights, however, do seem to be important in order to find a working compromise and their practical usefulness depends on the situations we take into account.

There is a view of robustness as describing the stability of the inference process under changes of the underlying situation. In our setup we have the space of situations, i.e. location and scale families indexed by the shape and the space of configurations. We are interested in making inference about the location parameter based on the observed configuration. Asymptotical influence curves, which describe the changes introduced by infinitesimal perturbations near the assumed model, proved useful in the point estimation case (Hampel(1974)). Similar ideas might work in the configural setup. We could ask how stable the inference is conditioned on the configuration.

Figure 3.1 shows four plots of the coverage density for a specific configuration and shapes

$$(1 - \epsilon)\Phi + \frac{\epsilon}{2}(\delta_y + \delta_{-y})$$

which leads to a coverage density proportional to

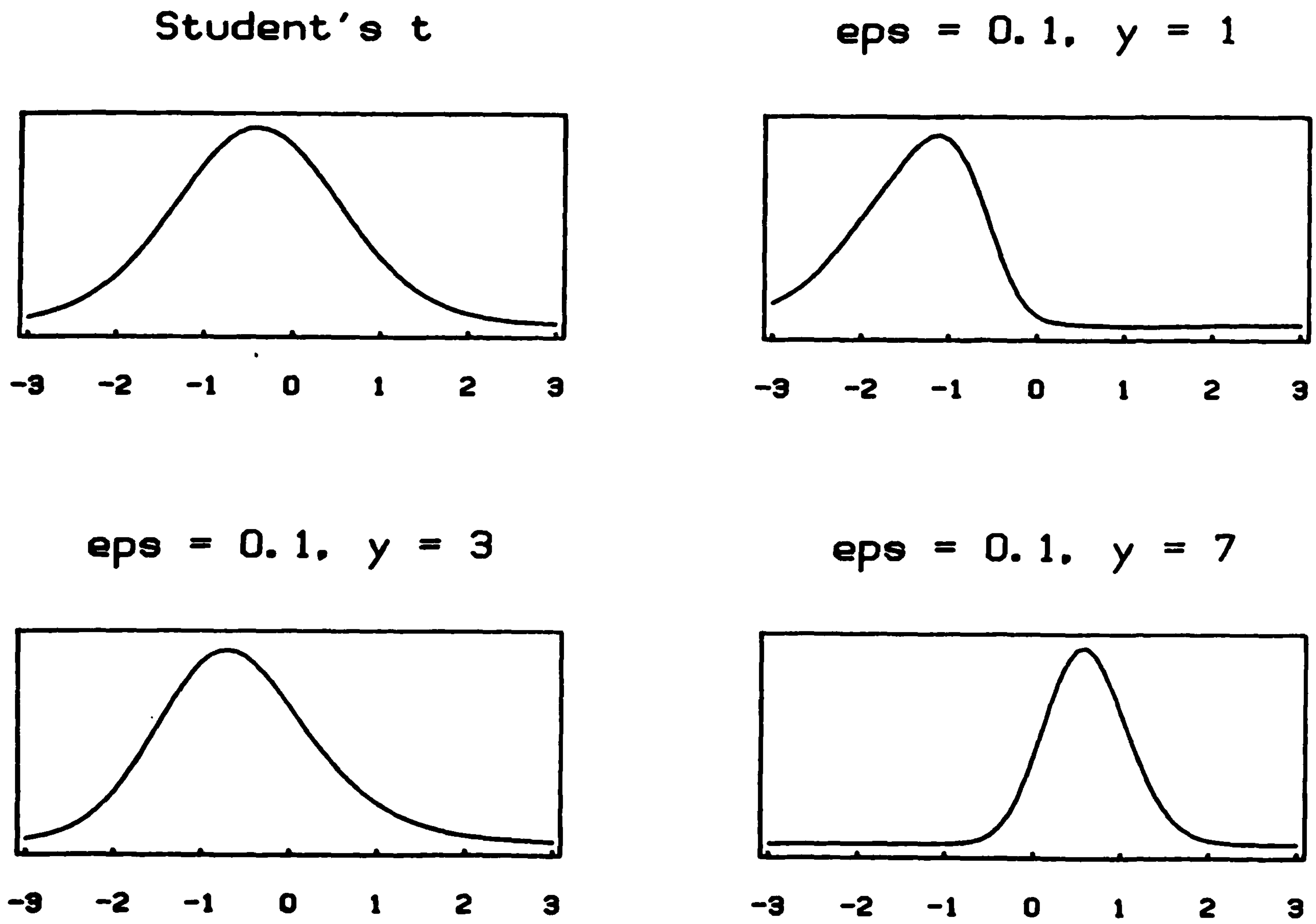
$$\left(\frac{1-\epsilon}{(2\pi)^{\frac{1}{2}}}\right)(n-2)(n-4)\ldots\left(\left(\frac{\pi}{2}\right)^{\frac{1}{2}} \text{ or } 2\right)\left(\frac{1}{\sum_{i=1}^{n}(c_i-x)^2}\right)^{\frac{n}{2}} +$$

$$\sum_{i=1}^{n}\frac{\epsilon}{2}\exp\left(-\frac{y^2}{2}\sum_{j\neq i}\left(\frac{c_j-x}{c_i-x}\right)^2\right)\frac{1}{\mathrm{abs}(c_i-x)}\left(\frac{y}{\mathrm{abs}(c_i-x)}\right)^{n-1}$$

where n is the sample size. Note that in this plot we have both heavy-tailed and light-tailed situations.

It is clear from these pictures that perturbations might very well teach us some things about the "robustness" of our inference

Figure 3.1: Conditional coverage densities under perturbed Gaussian situations (for configuration see Fig. 1.3 and 1.4)



conditioned on the given, i.e. observed configuration. We might for example study influences on the mean $ave_F(t|\vec{c})$ of the coverage density or on the expected length of confidence interval procedures which is determined by $ave_F(s|\vec{c})$. The last two can be viewed as mappings from the space of probability measures to the reals and therefore fit into the usual framework. Large influences would mean that the model is "dangerous" for the given configuration in the sense that nearby models would lead to different judgements. Since the configuration is a 2-dimensional class, the small sample approach seems feasible.

3.1. Robust confidence intervals derived from robust location estimators

The problem of confidence interval estimation in symmetric -- possibly heavy-tailed -- situations has been tackled in two papers by A. Gross (Gross(1976), Gross(1977)). He used ratios

$$\frac{n^{\frac{1}{2}}\ (T - \mu_o)}{S}$$

where T is a robust location estimate, S an estimate of its standard error and n the sample size to get intervals of the form

$$[T - \frac{\text{crit. value}}{n^{\frac{1}{2}}} S\ ,\ T + \frac{\text{crit.value}}{n^{\frac{1}{2}}} S]$$

The hope is of course that the critical values needed to get $100(1-\alpha)\%$ confidence is stable across situations. His conclusion was that a redescending estimate T with estimated asymptotic standard error S and the right tuning constant gives good intervals.

In his PhD thesis P. Horn (Horn(1981)) examined some simple confidence intervals based on 2 or 4 order statistics. His pivot and bi-pivot t-intervals are also designed to give a "robust" behavior.

Finally there are the exact finite sample minimax tests and corresponding intervals together with a somewhat arbitrary auxillary scale estimate, which were followed up by E. Ronchetti(Ronchetti(1982)) who showed us the right tests and corresponding intervals in an asymptotic infinitesimal sense for various robust location estimators. Ronchetti's approach carries over without problems to the more general regression case.

We will now see how well these intervals behave if we look at them more closely through our configural glasses, focusing on the two situations Gaussian and slash.

There are several ways in which one can specify any of the above confidence interval estimators. We will restrict attention to the following.

(A) One-step biweight interval

The center of this interval is a weighted mean of the observations y_i with weights

$$w_i = (1-u_i^2)^2 \quad -1 \le u_i \le 1 \text{ and } w_i = 0, \text{ otherwise}$$

where

$$u_i = \frac{y_i - \text{med}(y_j\text{'s})}{c \text{ MAD}} \quad \text{(c prefixed!)}.$$

The halfwidth is determined by an estimate of the asymptotic standard error s_{bi}, where

$$s_{bi}^2 = \frac{\Sigma(y_i - \mathrm{med}(y_j\text{'s}))^2 w_i^2}{[\Sigma w_i (1-5u_i^2)][\Sigma w_i (1-5u_i^2) - 1]}$$

(see Mosteller and Tukey (1977), p. 208).

This defines an interval estimator which on the configuration scale has the form

$$\text{center} \pm \text{critical value}(1-\alpha)\ s_{bi}$$

and which has still one constant (the multiplier of MAD) left at our disposal.

(B) pivot-t

We take the pivot-intervals as given in P. Horn's thesis (Horn(1981))

size=5 $[\frac{c_4+c_2}{2} \pm 2.02075(c_4-c_2)]$

size=10 $[\frac{c_8+c_3}{2} \pm 0.64875(c_8-c_3)]$

size=20 $[\frac{c_{16}+c_5}{2} \pm 0.39697(c_{16}-c_5)]$.

(C) Intervals based on the Huber function

To apply E. Ronchetti's intervals we choose the Huber -- ρ_c function (see Huber(1981)). The median absolute deviation MAD multiplied by the Gaussian bias correction of 1.484 will be our estimate $\hat{\sigma}$ of σ (see Ronchetti(1982), p. 74). The interval is then found by using

$$\{ \mu : \sum_{i=1}^{n} [\rho_c(\frac{y_i-\mu}{\hat{\sigma}}) - \rho_c(\frac{y_i\text{-Huberestimate}}{\hat{\sigma}})] \leq \text{cutoff} \}$$

Again we have one constant at our disposal. This constant has "traditionally" been chosen around the value 1.5 -- one argument being that this way the asymptotic loss in Gaussian efficiency is kept small, i.e. below 5%. We will concentrate on this choice. For the one-step biweight procedure it is known that the Gaussian efficiency is roughly at 95% for samples of size 20 (see Bell and Morgenthaler(1981)), if we choose a multiplier of c = 9.

3.1.1. The behavior of the conditional coverage probabilities

These three confidence interval procedures are automatically conservative in the slash case if tuned to reach 95% Gaussian confidence level. The one exception is the interval procedure based on the Huber's $\rho_{1.5}$, which has to be tuned for slash overall coverage in samples of size 20. This might indicate to some that this confidence interval will not perform well for that sample size.

For the smallest samples, i.e. size 5, none of these "robust" procedures does what we would want them to do. Especially the intervals based on a biweight-t seem to collapse. This fact has already been noticed in earlier work and is reported for example in P. Horn's thesis (Horn(1981)). The simple pivot-t intervals (B) seem to be as good as the more elaborate $\rho_{1.5}$ - intervals. Table 3.1 shows the hinge-spreads and medians of the conditional confidence coefficient distributions.

Table 3.1: Hinge-spreads (upper number) and medians of conditional coverage probabilities in %

method	Gaussian size=20	Gaussian size=10	Gaussian size=5	slash size=20	slash size=10	slash size=5
biweight-9	.68%	.90%	.23%	2.44%	2.33%	.51%
	95.75%	96.86%	99.22%	95.86%	97.25%	99.75%
pivot-t	2.73%	4.98%	3.48%	3.79%	2.98%	1.81%
	97.62%	97.50%	99.34%	97.14%	96.97%	99.53%
Huber-1.5	3.29%	3.51%	2.81%	3.43%	2.68%	1.10%
	96.45%	97.35%	99.10%	96.17%	96.98%	99.64%

(For samples of size 20 and 10 these values are based on 150 sampled configurations, for samples of size 5 on 500 configurations.)

Note that, in the size=5 columns, all procedures have a median coverage of over 99%, while reaching a mean confidence level of 95% in one of the two situations. We can conclude from this that in samples of size 5 the so called "robust" confidence intervals are most of the time overlong -- at least conditionally -- and sometimes too short. Also note the small values for the hinge-spreads of the biweight-t in this column.

We have of course already seen in previous sections that in the case of really small samples, i.e. size 5, a robust procedure will exhibit a somewhat unsatisfactory behavior of the conditional confidence coefficients. The large sample argument which leads us to the choice of tuning constants in both the biweight-t and the p_c - function is deceptive. For smaller samples the tuning constant has to be increased (Bell & Morgenthaler(1981)). If we compute the biweight intervals based on 11*MAD and the intervals based on $p_{1.9}$ for samples of size 5, we naturally get better behavior in the Gaussian situation. But somewhat surprisingly we also improve the slash behavior. This is again an indication that the Gaussian single situation optimal procedure, i.e. Student's-t intervals, are not very far from being a very good robust procedure in small samples.

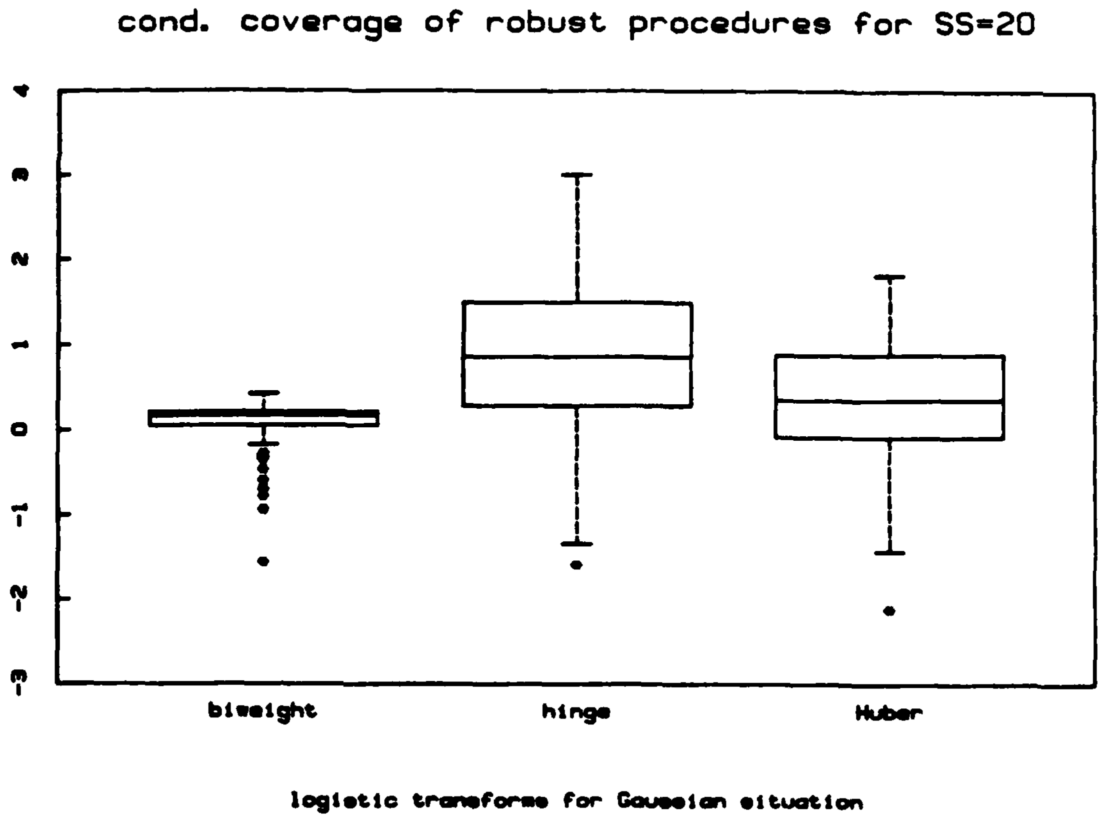
As the sample size increases, the three procedures under consideration improve. At the biggest sample size, 20, the biweight-t is doing best. For the intervals based on the Huber function and an "imitation of the classical residual sum of squares" it is probably quite crucial to use a "matched" scale estimate, but up to now this problem has not really been addressed.

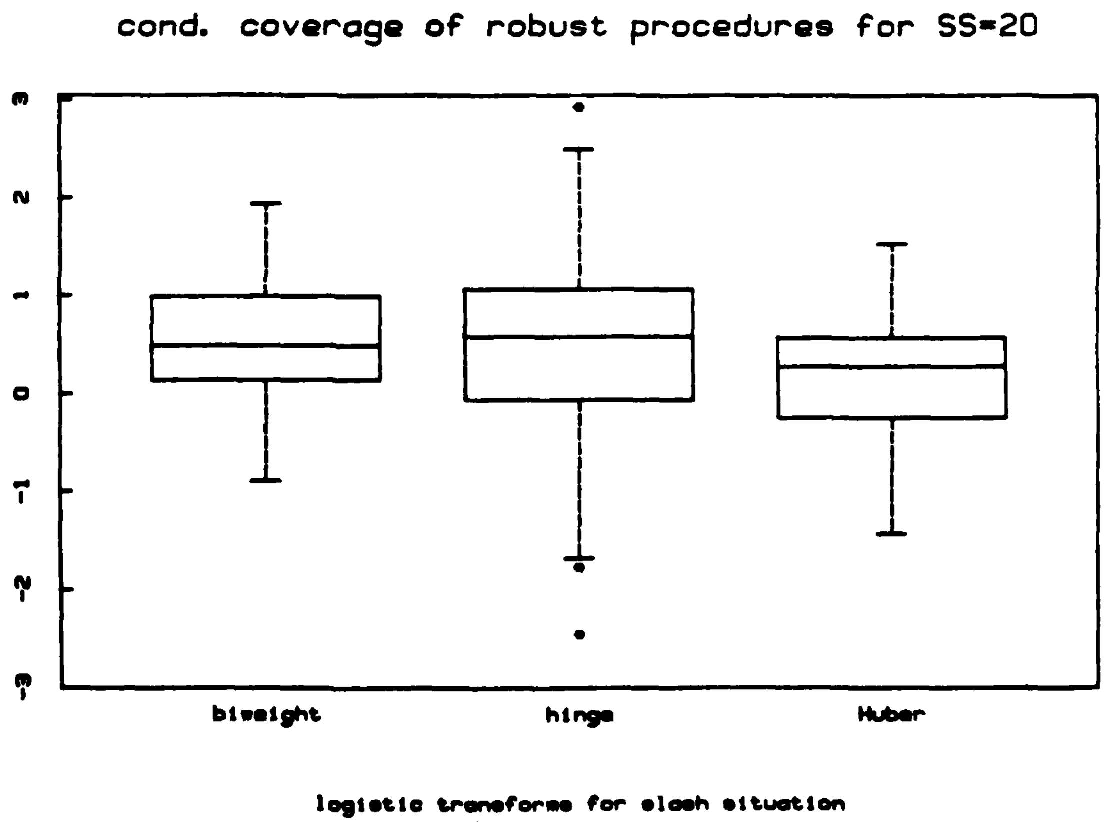
In the intermediate sample size, 10, the impressions based on Table 3.1 are somewhat mixed, but even here the biweight-9 seems to be the method of choice. It might surprise us that the intervals based on Huber-1.5 seem to behave better in the extreme slash situation than in the Gaussian. But remember that, if we want to be prepared against heavy-tailed situations, "robustness of validity" comes basically for free.

Figure 3.2 shows boxplots for the logistic transforms (see section 2) of the conditional coverage probabilities in samples of size 20. These plots can be compared with Fig. 2.1 and 2.2. The biweight-9 intervals are better behaved than the other two, but clearly worse than the bi-shortest intervals, which use information about the conditional coverage density.

In plots -- like Figure 2.3 -- where we plot the lower and upper conditional missing-probabilities, the robust procedures exhibit a very typical pattern -- especially in the Gaussian situation. The

Figure 3.2: Logistic transforms for 150 sampled configurations for three robust intervals





points lie inside a cone with vertex at (0,0) and symmetric around the diagonal. As the sample size increases, the cone opens up more and at sample size 20 the biweight-t interval does not exhibit this pattern any more. So a confidence procedure based on a robust center will at least result in balanced intervals of some sort. Of course we saw that the conditional coverage behavior is still not satisfactory -- except in samples of size 20 -- and this is due to the fact that the width of the intervals is considerably underestimated in a few configurations. In the majority of the configurations this forces us to make the intervals too long and in a plot of the conditional missing probabilities most of the points are near the vertex of the cone.

3.1.2. The square mean length efficiencies

Figures 3.3 through 3.5 show the square mean length deficiencies with the robust confidence intervals added to the procedures already discussed in section 2 (see: Figures 2.4 through 2.6). The biweight-t intervals are labeled "bi-#", where the # denotes the multiplier of MAD. The interval procedure based on Huber's $\rho_{1.5}$ function is labeled "hub" and the one based on $\rho_{1.9}$ by "hub-1.9". The pivot-t intervals finally are labeled by "hin". The method denoted by "tp" will be discussed in the next sub-section. Table 2.1 has the numbers.

In samples of size 5, Figure 3.5, it is obvious that the robust procedures pay a high price in terms of Gaussian efficiency in order to be "acceptable" in the slash. It is also clear from this picture that the less robust biweight-11 estimator leads to a better

Figure 3.3: Plot of deficiencies including nonparametric, robust and bi-shortest interval procedures in samples of size 20 (see p. 68)

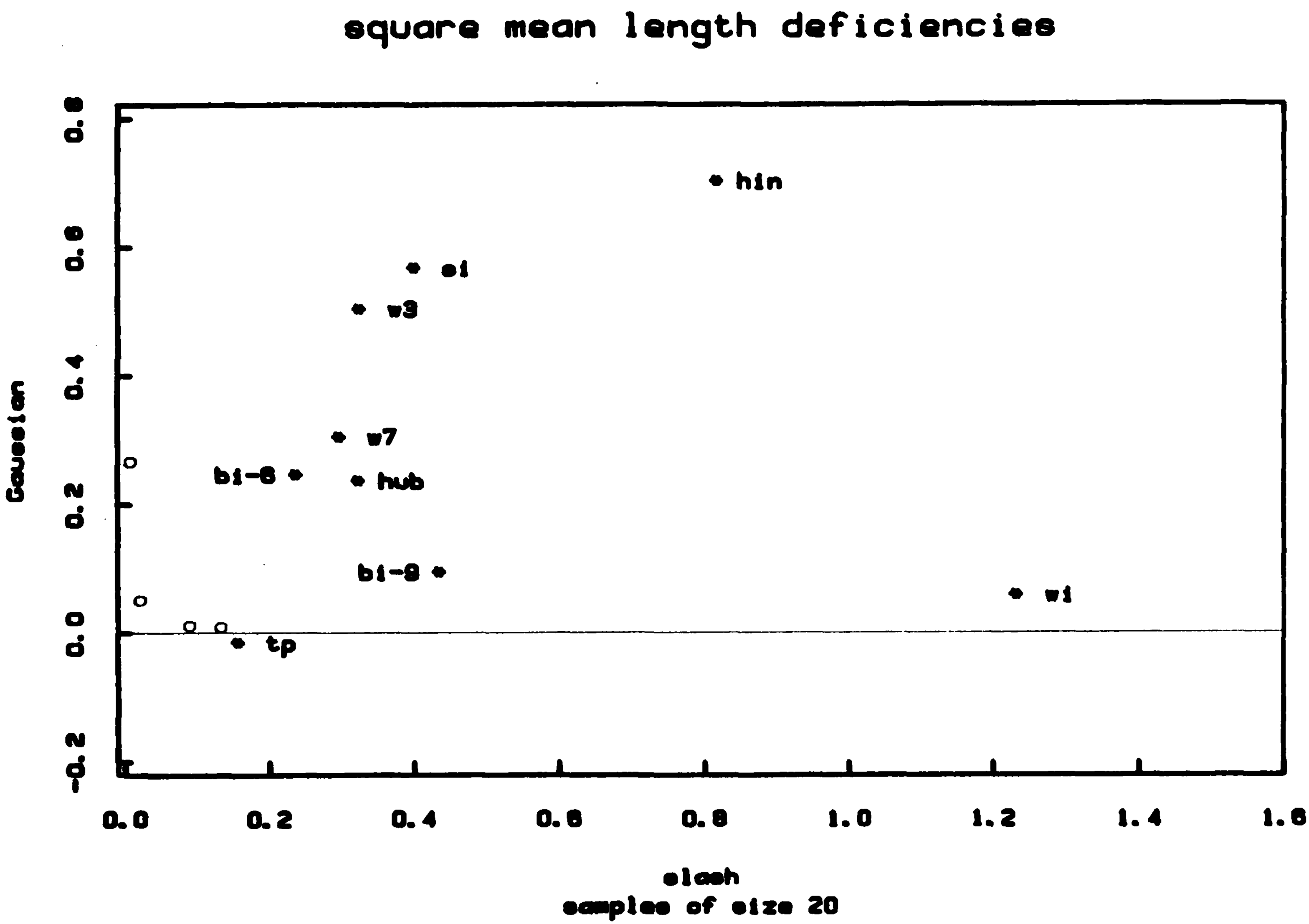


Figure 3.4: Plot of deficiencies including nonparametric, robust and bi-shortest interval procedures in samples of size 10

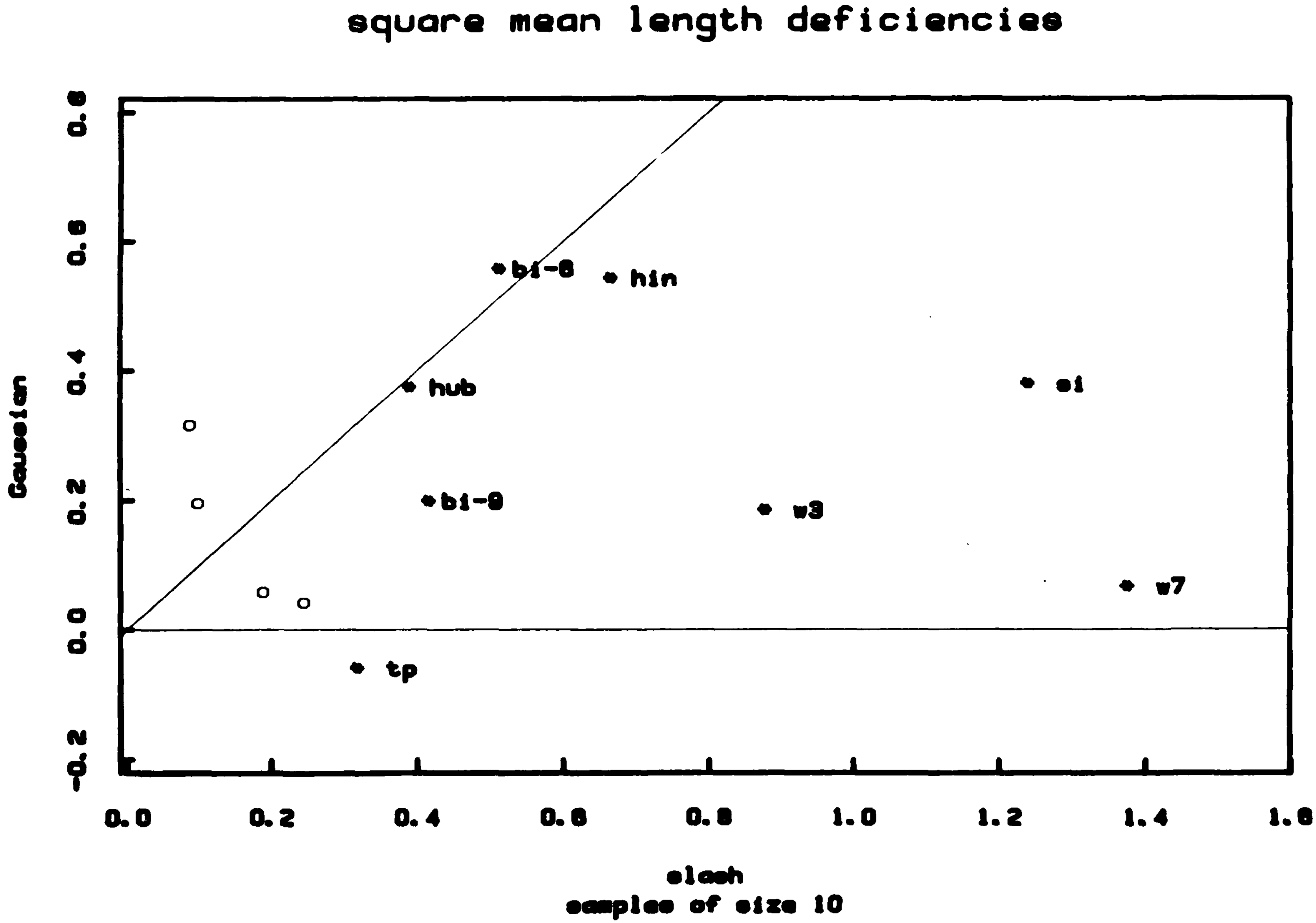
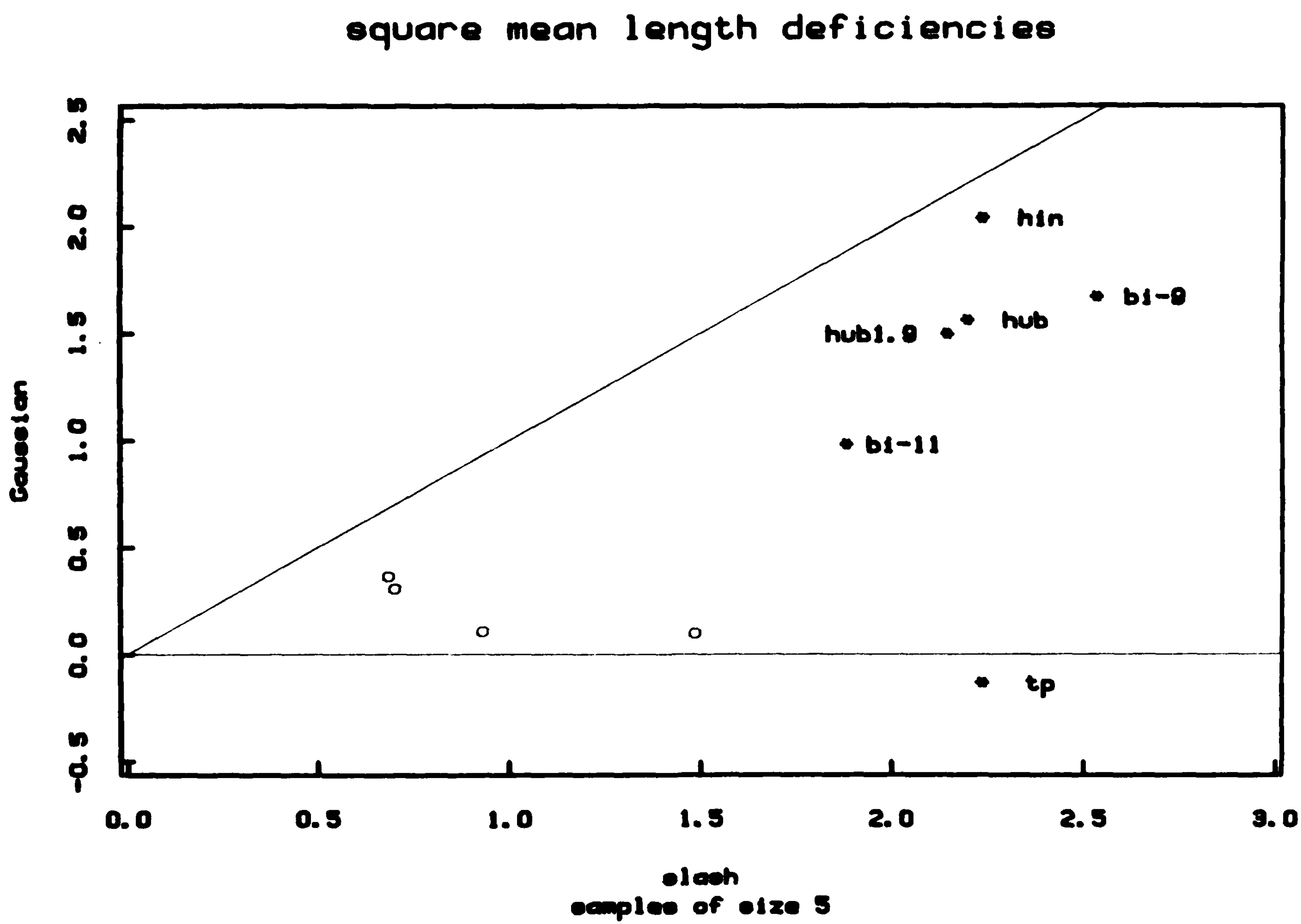


Figure 3.5: Plot of deficiencies including nonparametic, robust and bi-shortest interval procedures in samples of size 5



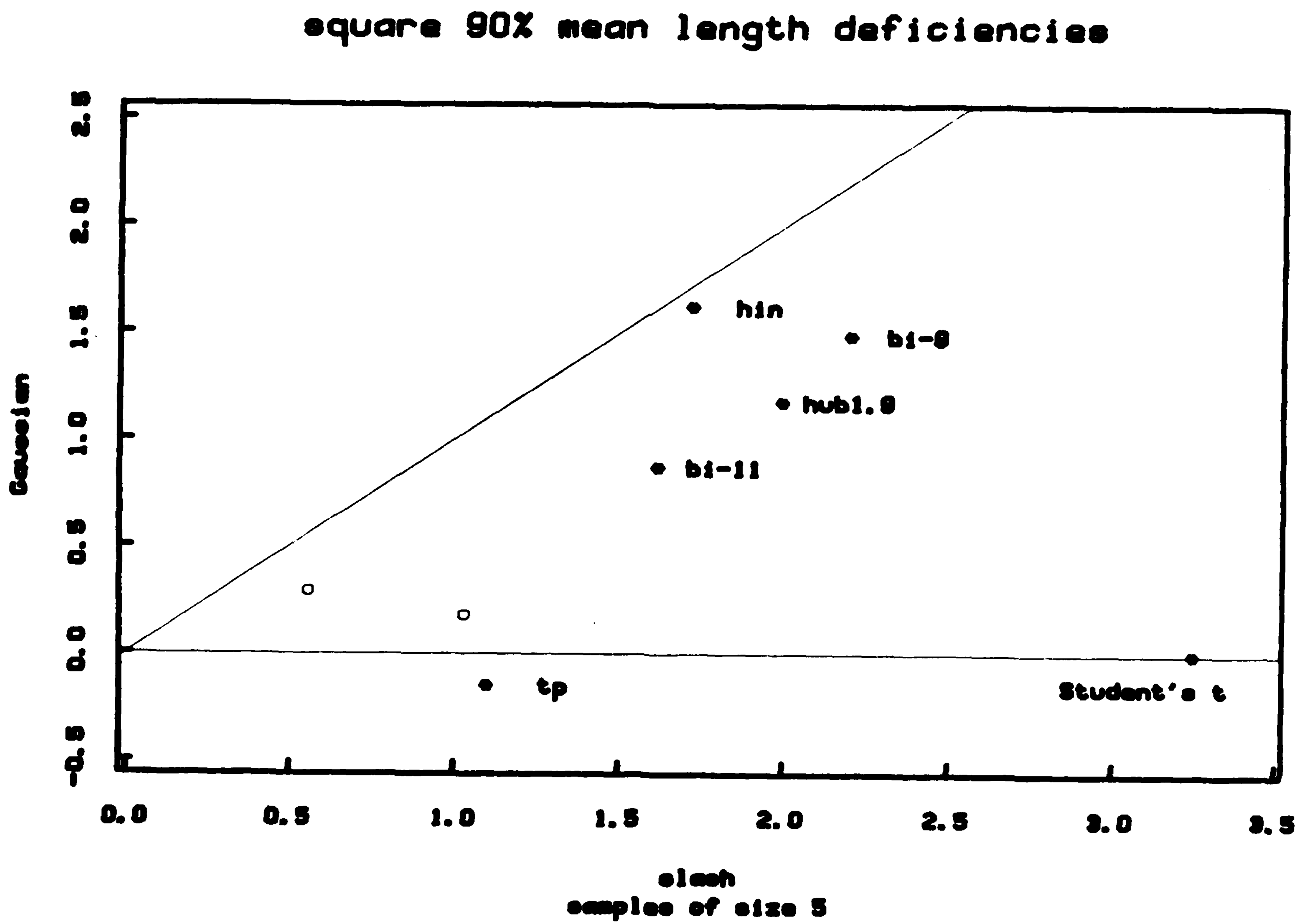
confidence interval. The same effect can be seen in the Huber interval, but it is much less obvious. The simple interval based on the second and fourth order statistic is dominated in terms of length efficiency by both hub-1.9 and bi-11, but -- as we discussed above -- none of them is satisfactory as far as the conditional coverage probabilities go. The increase in expected length over Student's t in the Gaussian situation is big, even with such "mildly robust" estimators as biweight-11. The bi-shortest intervals are a lot better than everything else, but again, for some people, the behavior of their conditional coverage probabilities will be nonacceptable. In that sense these are not practical confidence intervals.

When we go to samples of size 10 the picture gets more reasonable (note the change of the scale from Figure 3.5 to 3.3 & 3.4). Most striking is the improvement of the robust confidence intervals over the nonparametric ones. The "cloud" of robust procedures is moved along the slash axis without losing much in the Gaussian case. Again it is not advisable to base confidence intervals on very stringent robust estimators, bi-9 improves a lot over bi-6, which is not on the admissible part of the biweight curve. The simple pivot-t interval has to pay a price for its simplicity, it roughly balances its loss in the Gaussian and the slash.

In the largest sample size under consideration, 20, the pivot-t interval clearly is not competitive, the simple and distribution free sign-interval dominates the pivot-t. The nonparametric intervals are now reasonably "robust" themselves. If we look at the "winsorized" Wilcoxon scores (see Morgenthaler (1983)) it seems that we ought not go below the value of 7. The biweight-t intervals are superior to the ones based on Huber's ρ_c and the admissible part of the biweight curve stretches now to lower c-values.

In his thesis P. Horn (Horn(1981)) examines confidence interval procedures from the point of view of 90% - ECIL, i.e. the mean length after trimming off 10% of the upper tail of the length distribution. This is a natural way to go if we argue that a confidence interval procedure which most of the time produces reasonable intervals with just a few wild -- i.e. overlong -- ones should really be taken more seriously than its mean length, which of course is heavily influenced by the wild ones, suggests. It is not a trivial matter to compute 90% - ECIL values for any given confidence interval estimator. Conditioned on any configuratin there will be some (t,s) - points which are below the 90% point in terms of the *length distribution* and we ought to integrate out over these only. An approximation however is possible. If we take from our sampled configurations the 90% with the shortest conditional expected length and average over them, we will have an even more conservative effect than computing 90% - ECIL. If we adopt the above loss function and compute efficiencies, the plots corresponding to Figures 3.3 through 3.5 do not change drastically. In samples of size 20 and 10, the conclusions are similar to the expected length loss. It is the nonparametric and the hinge-t intervals which profit somewhat in the slash efficiency if we trim the upper 10% of their lengths. In the case of small samples, i.e. 5, the improvements in efficiency for Student's t are big. Figure 3.6 shows the new situation. Again we conclude that, while Student's t produces "long" intervals in 10% of the slash drawn

Figure 3.6: Plot of deficiencies based on the 90% configurations with shortest conditional expected lengths in samples of size 5



configurations, it is otherwise comparable to the robust procedures. Of course it has the ideal Gaussian behavior. On this plot only the procedures for shadowprice ratio infinity and 0.1 are included.

3.2. Conditional confidence intervals

We have seen that the bi-optimal intervals from section 2 are indeed superior to the existing robust confidence procedures. But from an applied point of view we would like to have confidence intervals which are easily interpreted and understood conditioned on the observed configuration. Furthermore we ought to keep the simplicity of our procedures in mind.

3.2.1. Three-point approximations

We saw in section 2 (equation 2.5) that the weighted combination of the confidence densities

$$h(\ ) = \frac{\lambda_g w_g co_g(\ ) + \lambda_s w_s co_s(\ )}{\lambda_g w_g + \lambda_s w_s} \tag{3.1}$$

where $co_g(\ )$ and $co_s(\ )$ are the coverage densities conditioned on the observed configuration, w_g and w_s the relative weights and λ_g and λ_s the Lagrange multipliers in the Gaussian and slash situation, plays an important role in combining the two situations. It seems rather natural to use the mixture (3.1) with $\lambda_g = \lambda_s = 1$ as a basis for approximate conditional confidence intervals. If we use the mixture (3.1) and "count in" $100\frac{\alpha}{2}$% from each tail, we find an interval which we might expect to have approximate level $100(1-\alpha)$% for both situations under consideration.

To simplify the procedure we will only compute three points on the conditional coverage distributions for the Gaussian and the slash and use linear logistic interpolation to find the actual confidence bounds. To get three points on the slash confidence distribution requires four numerical integrations which also yield the value of the relative slash weight. In the Gaussian situation we can rely on the tabulated t_{n-1} - critical values to get a convenient triplet. The Gaussian relative weight can be computed using the theoretical formula.

We can think of these three-point approximate intervals as using our two situations Gaussian and slash to correct an initial guess conditioned on the given configuration, if we choose the three points as lower bound, center and upper bound of a specified confidence interval.

It turns out that, if we again restrict attention to 95% confidence levels, the three-point-procedure is anti-conservative for the Gaussian and conservative for the slash. The 5-number summaries (see Tukey(1977)) for the conditional coverage probability distributions in the Gaussian situation are:

size=20

#150			
M		94.8%	
H	94.7%		94.8%
	86.1%		96.2%

size=10

#150			
M		94.9%	
H	94.0%		95.2%
	33.8%		95.5%

size=5

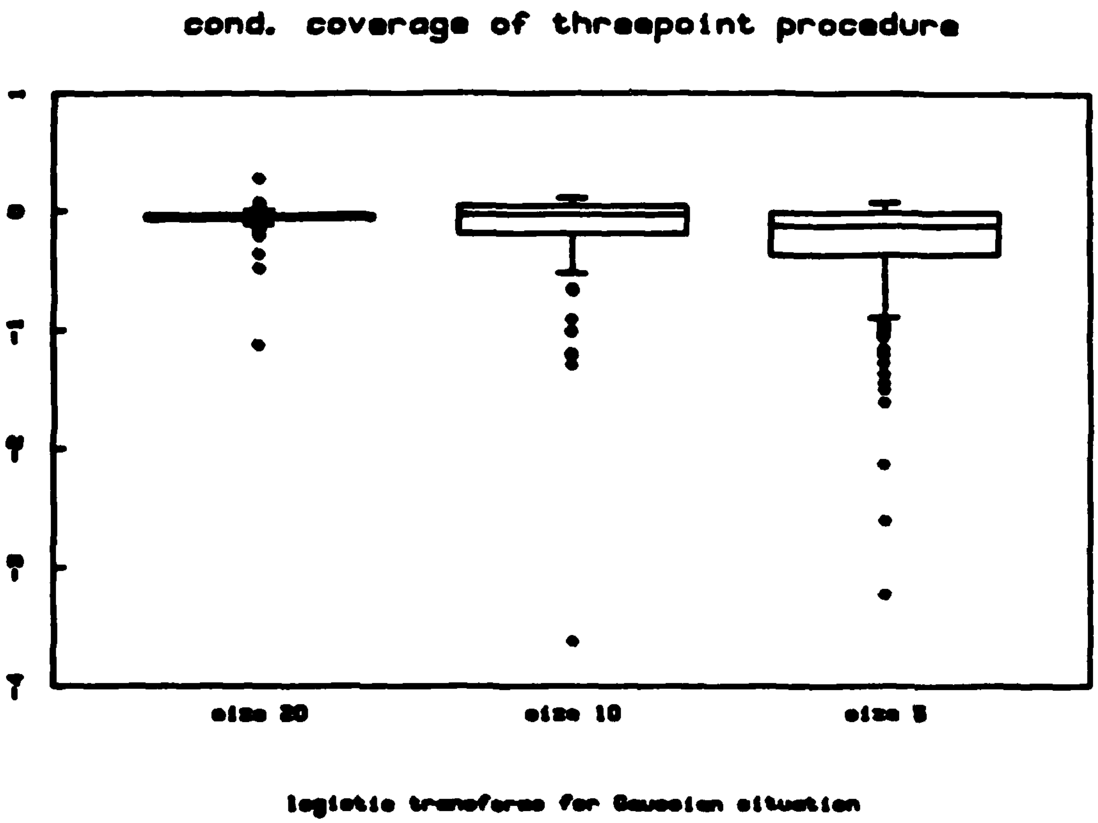
#500			
M		94.4%	
H	92.9%		94.9%
	42.9%		95.3%

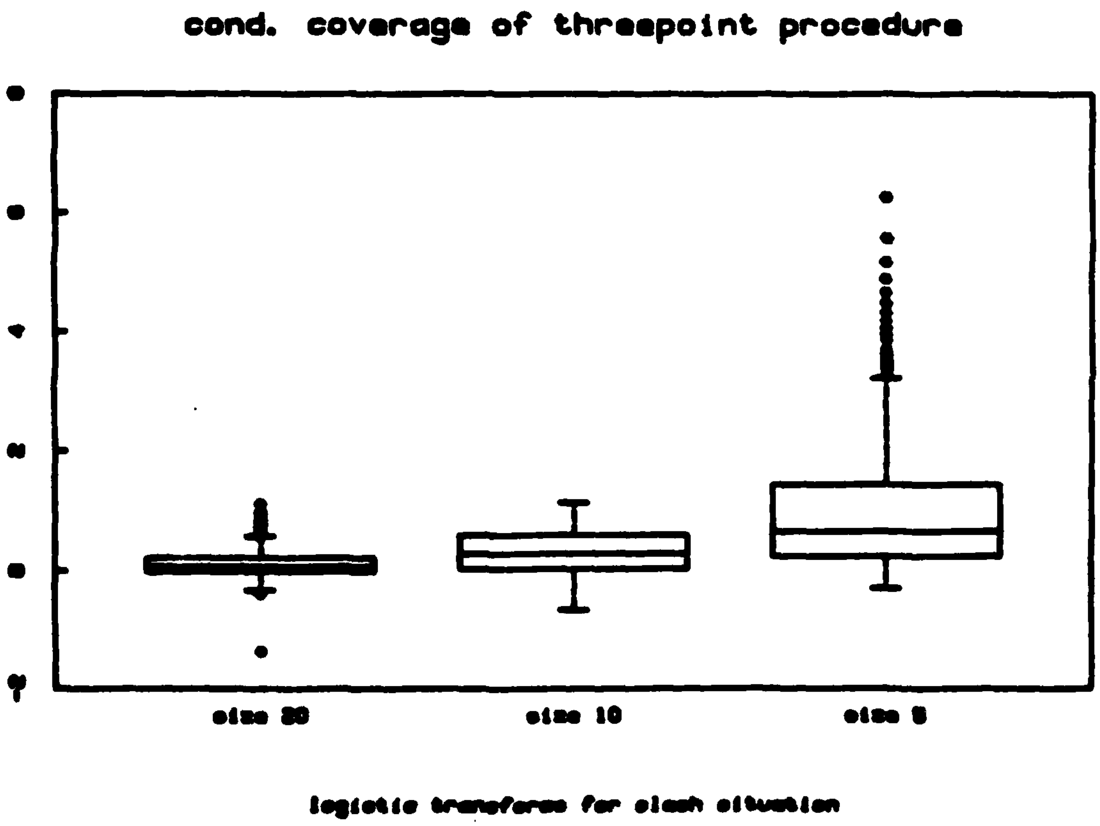
It is obvious that the conditional Gaussian confidence level hardly surpasses 95%, this once more demonstrates how -- if we are concerned about heavy-tailedness -- most of the configurations tend to shorten Student's t interval so that the weighted mixture used to get these three-point-confidence intervals tend to have shorter tails than Student's t distribution. Looking across sample sizes we notice how well behaved the three-point intervals are for samples of size 20. As the sample size decreases the conditional behavior gets worse. The estimated overall coverage probabilities are:

	size = 20	size = 10	size = 5
Gaussian	94.66%	93.78%	93.32%
slash	95.49%	96.12%	97.64%

Clearly for samples of size 20 we have a relatively cheap and very good confidence interval procedure. For the smaller sample sizes we might want to correct for the anticonservative Gaussian confidence level by introducing a "blow up" factor (for samples of size 5 we need a factor of 1.112 to reach 95% Gaussian coverage, for samples of size 10 a factor of 1.067 is sufficient). Figure 3.7 shows the boxplots for the logistic transforms of the conditional coverage probabilities in both situations. Using a weighted mixture of the

Figure 3.7: Logistic transforms for three-point procedure in three sample sizes





two conditional confidence distributions is having opposite effects in the Gaussian and the slash. In the Gaussian we use the "slash analysis" to shorten, in the slash we use the "Gaussian analysis" to lengthen the optimistically short slash confidence intervals. Note the thickening of the tail in these boxplots as the sample size goes down.

Having conservative slash confidence level eliminates the "split" behavior we observed for the bi-shortest confidence estimators in the slash situation (see Figure 2.3). The three-point interval does instead the natural thing by being somewhat conservative.

In all of the deficiency plots of the previous section the three-point approximate interval is included under the label "tp" (Figures 3.3 through 3.6). In all of the cases "tp" is on the average shorter than Student's t in the Gaussian and has, therefore, a negative deficiency. In samples of size 5 it is inside the cloud of "robust" interval estimators as far as the slash loss is concerned, but note how trimming off 10% of the longest intervals (Figure 3.6) moves the procedure away from the robust ones. This indicates that the "robust" confidence procedures -- in samples of size 5 -- have very many configurations where they are long, whereas both "tp" and Student's t have a tail towards "long configurations", but are most of the time a lot shorter.

On the whole we may say that the confidence procedure discussed in this section has an appealing behavior. The intervals from such a computation will be robust and it might be interesting to extend it to other pairs than Gaussian & slash or maybe even to triplets. This ought not create any new difficulties.

3.2.2. Confidence intervals based on the conditional mean-square-error curve

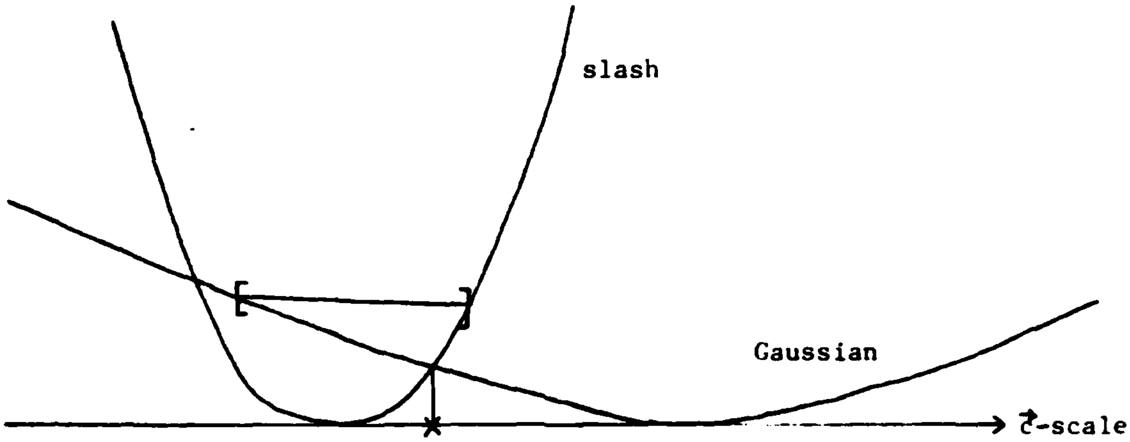
For the three-point confidence interval we need to calculate four integrals in each situation -- except in the Gaussian where the values are tabulated or formulas exist. Four integrals also come up naturally if we try to estimate the location parameter. In our parameter system for any given configuration $\vec{c}$ it follows that the conditional mean-square-error in situation F for a location estimate T is

$$\mathrm{mse}_F(T|\vec{c}) = \mathrm{ave}_F(t^2s^2|\vec{c}) - \frac{\mathrm{ave}_F^2(ts^2|\vec{c})}{\mathrm{ave}_F(s^2|\vec{c})} + (t_{opt,F} - T(\vec{c}))^2 \, \mathrm{ave}_F(s^2|\vec{c})$$

where $t_{opt} = \dfrac{-\mathrm{ave}_t(t^2s^2|\vec{c})}{\mathrm{ave}_F(s^2|\vec{c})}$ and $T(\ \vec{c}\ )$ denotes the value our location estimate T takes on the configuration level. All the expected values needed to get this quadratic curve in $T(\vec{c})$ are

$\mathrm{ave}_F(ts^2|\vec{c})$, $\mathrm{ave}_F(t^2s^2|\vec{c})$, $\mathrm{ave}_F(s^2|\vec{c})$ and the relative weight

These can be calculated again by four two-dimensional numerical integrations. They are somewhat simpler to get than the four integrals needed for the three points on the confidence distribution, since it is possible to economize somewhat. Based on the calculation of the four integrals we can compute an excellent robust location estimate by considering the weighted conditional relative excess curves in the picture below.



The weighted conditional relative excess curves for the Gaussian and the slash situation

The conditional relative excess is defined as

$$\text{cond. rel. exc.}_F(T) = (t_{opt,F} - T(\vec{c}))^2 \frac{\text{ave}_F(s^2|\vec{c})}{\text{cond. minimum in } F}$$

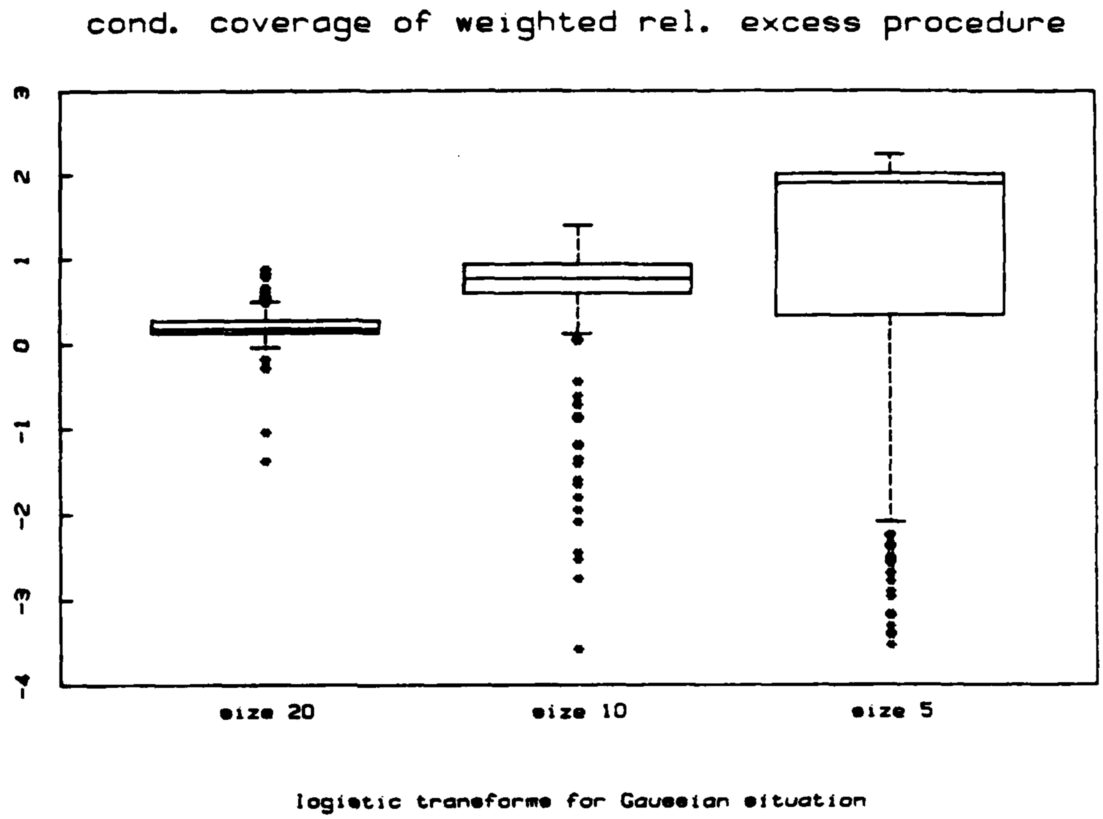
where the conditional minimum in F is

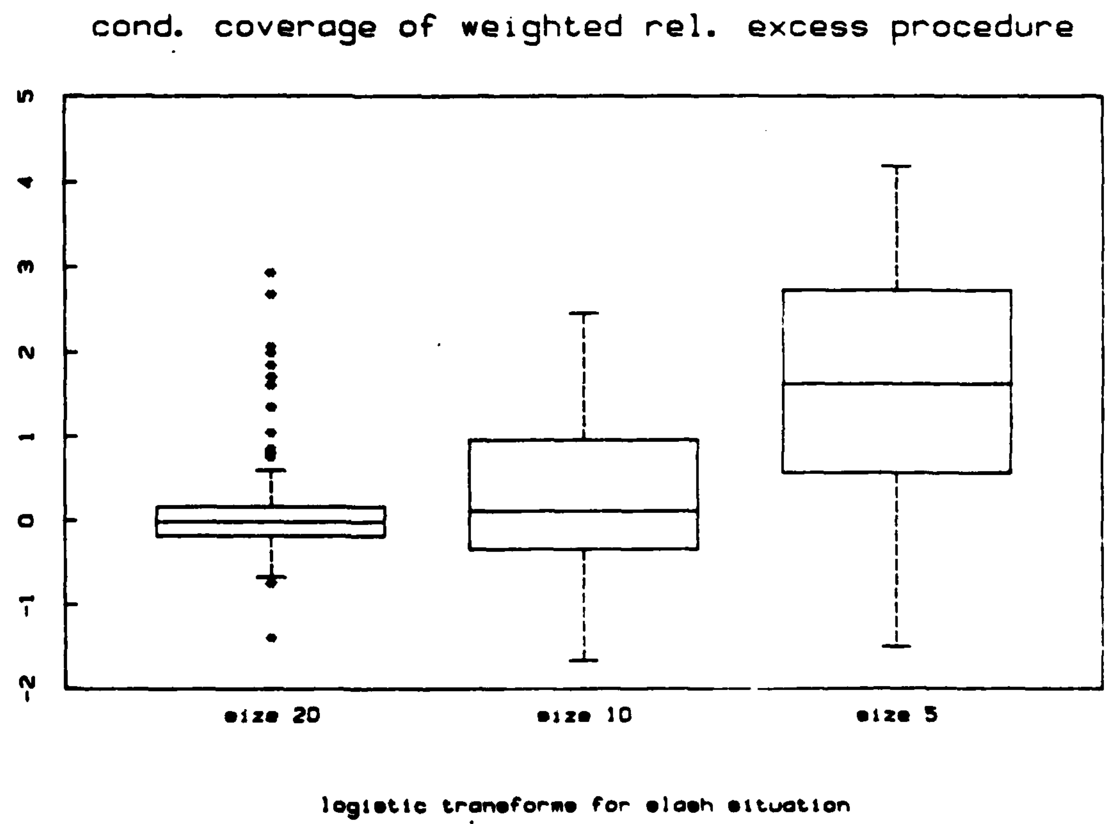
$$\text{ave}_F(t^2 s^2|\vec{c}) - \frac{\text{ave}_F^{\,2}(ts^2|\vec{c})}{\text{ave}_F(s^2|\vec{c})}.$$

The relative weight w_F for the given configuration under situation F is used to weight the conditional relative excess for the situation F. The point marked "x" in Figure 3.8 is a natural choice for the estimate $T(\vec{c})$ on the configuration level and the interval [---] drawn in seems to be a reasonable choice for a confidence interval on the configuration level based on these curves. The idea is to replace the two weighted conditional relative excesses by their maximum and define the interval bounds by a cutoff

$$\max_{\text{Gaussian, slash}} \{\text{weighted cond. rel. excess(upper bound)}\} =$$

Figure 3.8: Logistic transforms for procedure based on cond. mean-square-error curves in three sample sizes





$$\max_{\text{Gaussian, slash}} \{\text{weighted cond. rel. excess(lower bound)}\} = \text{cutoff}.$$

This is of course the same as using

$$L = \max \{L_g, L_s\}$$
$$U = \min \{U_g, U_s\}$$

where L_g & U_g and L_s & U_s are derived by the same cutoff from the single situation weighted conditional relative excess curve. The interval described above seems to rely on just how we represent the configuration, i.e. the choice of $\vec{c}$. However this is not true, because of the canonical changes in all the integrals involved under changes of the class-representing element $\vec{c}$.

What we propose here is a side product of an analysis whose primary purpose is the estimation of a location parameter. But even if we do have point estimation in mind, it is a small step to try and put a confidence interval around it.
In the above intervals we include the <u>the parameter values which if chosen as a parameter estimate on the configuration level would lead to small maximal mean-square-error relative to the minimum conditioned on the given configuration</u>.

Figure 3.8 shows the conditional confidence coefficients for the weighted mean-square-error interval. In samples of size 20 this procedure is slightly conservative in the Gaussian situation. A look at Figure 3.2 shows us that the Gaussian behavior is quite close to the biweight-t interval with tuning constant 9, but that in the slash situation it is somewhat better. The main difference for the

Gaussian seem to be in the direction of the skewness in the bulk of the distribution. As the sample size decreases the "coverage performance" of the weighted mean-square-error interval gets worse. It can certainly not compare itself to the three-point approximate intervals. However, its behavior is better than for the other robust confidence interval estimators. The overall confidence levels are

	size = 20	size = 10	size = 5
Gaussian	95.8%	95.0%	95.0%
slash	95.0%	95.1%	97.7%

Table 2.1 gives the expected lengths under the label "wms". In the slash situation the numbers are comparable to "ratio 0.2"; in the Gaussian they are more like the other robust procedures. This should give an idea where the "wms" point would fall in the deficiency plots.

The confidence interval based on a weighted mean-square-error seems to be doing about what other robust intervals do -- maybe slightly better. The very good behavior in the slash situation might be unduly influenced by the fact that the slash is one of the situations we took into consideration. It is interesting to notice that introducing the slash along with the Gaussian in this way -- i.e. by looking at weighted conditional mean-square-errors -- seems to put more emphasis than we would like on the slash.

If on the other hand we use the center of the three-point intervals as a "robust" location estimate, it has a high Gaussian efficiency, but is rather poor in the slash. Both approaches described in sub-section 3.2 have their merits.

4. What have we learned about confidence intervals for a location parameter

In the previous sections we discussed one possible way of approaching the problem of robust confidence interval estimators. It is based on the criterion of expected length. The ultimate interval estimator has the required coverage probability and at the same time is short. We learned that this approach has its drawbacks. The conditional confidence levels do not behave in a satisfactory way for samples of size 5 or 10, though they behave rather well in samples of size 20. A possible remedy might be in the choice of criterion. If we do not consider the expected length, but rather something which combines the behavior of conditional coverage probabilities and some aspect of the length distribution, we might very well improve over the bi-shortest procedures. However, the bi-shortest confidence interval procedures are superior to any of the methods proposed so far as solutions to the unconditioned confidence problem go. Relatively simple approximations, the three-point interval and the interval based on the conditional mean-square-error curve, are possible. The three-point interval has an excellent Gaussian behavior, is, however, rather bad in the slash situation. The opposite is true for the mean-square-error interval. The search for further simplifications, leading to nonlinear closed form formulas involving the configuration, might well be worthwile.
The viewpoint of this exposition is based on the behavior in small samples and we do not advocate the uncritical use or these ideas for larger sample sizes. As the sample size goes up, we learn more about

the underlying shape from our data and another "limited situation game" involving much more closely spaced situations might be more profitable. It remains to be seen, how well the methods proposed in the previous pages perform in situations other than the Gaussian and the slash. But we consider the small sample approach as a strength of our methods as opposed to procedures which are asymptotically justified.

It is interesting to note how much the problems we face change with changing sample sizes. We learned that in samples of size 5 a compromise between the Gaussian and the slash has more severe consequences on the conditional properties than when we deal with larger samples.

The use of numerical integration over configurations to get good statistical procedures is certainly worthwile doing and should be explored further. Such procedures are -- once we have a computer -- simple and cheap to calculate and they are potentially superior to existing techniques.

Some more ideas on how to implement all this in the case of confidence intervals can be found in Tukey (1981).

Bell Krystinik, K. and Morgenthaler, S. (1981). "comparison of the bi-optimal curve with curves for two robust estimators," Technical Report No. 195, Series 2, Department of Statistics, Princeton University, New Jersey 08544

Ferguson, T.S. (1967). Mathematical Statistics: A Decision Theoretic Approach, New York, Academic Press.

Gross, A.M. (1976). "Confidence interval robustness with long-tailed symmetric distributions," J. Am. Stat. Assoc., 72, 341-354.

Hampel, F.R. (1974). "The influence curve and its role in robust estimation," J. Amer. Stat. Assoc. 69, 383-393.

Horn, P. (1981). "On simple robust confidence procedures," Ph.D. Thesis, Department of Statistics, Princeton University, Princeton, New Jersey 08544.

Huber, P.J. (1981). Robust Statistics, Wiley Publishing Company, New York.

Morgenthaler, S. (1983). "Strong confidence intervals: A compromise between the Gaussian and the slash," Technical Report No. 253, Series 2, Department of Statistics, Princeton University, Princeton, New Jersey 08544.

Mosteller, F. and Tukey, J.W. (1977). Data Analysis and Regression: A Second Course in Statistics, Addison-Wesley Publishing Company, Reading, MA.

Pregibon, D. and Tukey, J.W. (1981). "Assessing the behavior of robust estimates of location in small samples: Introduction to configural polysampling," Technical Report No. 185, Series 2, Department of Statistics, Princeton University, Princeton, New Jersey 08544.

Rogers, W. H. and Tukey, J. W. (1972). "Understanding some longtailed symmetrical distributions," Statistica Neerlandica, 26, 211-226.

Ronchetti, E. (1982). "Robust testing in linear models: the infinitesimal approach," Ph.D. Thesis No. 7084, Swiss Federal Inst. of Technology, Zurich.

Tukey, J.T. (1977). Exploratory Data Analysis, Addison-Wesley, Publishing Company, Reading, MA.

Tukey, J.W. (1981). "Kinds of polyconfidence intervals for centers, and some thoughts on identification and selection of confidence procedures using configural poly-sampling," Technical Report No. 190, Series 2, Department of Statistics, Princeton University, Princeton, New Jersey 08544.

END

FILMED

8 84

DTIC