Lantmäteriet
Lantmäteriverket - National Land Survey
S - 801 12 GÄVLE · SWEDEN

Tekniska skrifter - Professional Papers 1982:10

ISSN 0280:5731

Homogeneity Tests of the Swedish and Norwegian RETrig Data Using Modulated Normal Distribution Theory

by Lars Sjöberg

Gävle
september 1982

Abstract. Several subsets of the Scandinavian RETrig data do not pass the standard normal distribution test. Two possible explanations might be that the data are biased or inhomogeneous or combinations of these effects. Merely two subsets were found biased, and in the case of the Swedish EDM observations by MRA2 the normal distribution was accepted after removal of the bias.

The theory of modulation on the normal distribution is reviewed and used as a possible mean to explain the distribution of the remaining subsets. The origin of such an inhomogeneous population is probably, that it is a combination of two or more subpopulations of various variances. Summarizing, five data sets were found significantly modulated normal.

1 Introduction

In Sjöberg and Eliasson (1981) statistical tests were reported on the Swedish block of the Recomputation of the European Triangulation (RETrig), phase II (ED79). Normal distribution tests were carried out on the observation residuals of subblocks and for various sets of instruments. In some tests the hypothesis H_0 on normal distribution of observations failed. A specific test showed that all EDM observations were found biased, which could possibly explain the rejection of H_0 for two sets of instruments, namely geodimeters and MRA2.

Similar tests were reported on block Norway by Sjöberg (1982). All these observations were found unbiased. In spite of this fact the directions and some EDM observations (MRA1 and MRA3) did not pass the normal distribution test.

The tested residuals were normed according to the formula

$$w_k = v_k / s_{v_k} \tag{1a}$$

where

$$s_{v_k} = m_k s \sqrt{1-m/n} \tag{1b}$$

m = number of unknowns

n = number of observations

A constant error factor of the a priori standard error (m_k) of observation k will be assimilated as an a posteriori variance of unit weight (s^2) deviating from unity. Thus neither the standard error estimate of v_k {formula (1b)} nor the outcome of the normal distribution test would be affected by this constant error factor. The opposite will

be the result if we assume that m_k is false in a more complicated way than just by a constant.

Summarizing, the rejection of H_0: {the sample w_k is N(0,1)} can possibly be explained by:

a. the observations are biased

or

b. incorrect error model of observations (m_k).

or a combination of a and b.

These possibilities will be tested for the Swedish and Norwegian RETrig data. The test for bias is straight forward. The test of the standard error model (m_k) will be carried out by means of the theory of modulation on normal distribution developed by Romanowski (1979).

2 Characteristics of the data

The Swedish ED 79 data that failed in the normal distribution test is summarized in Table 1. (cf: Sjöberg and Eliasson, 1981). In a separate test all EDM observations were found biased. The geodimeter and MRA2 observations of Table 1 yielded the means

$\bar{w} = 0.151 \pm 0.045$ and $\bar{w} = 0.287 \pm 0.079$, respectively.

Table 1. Summary of rejected H_0 {the data is normally distributed with expectation 0} of block Sweden (See Sjöberg and Eliasson, 1981.) T = computed test parameter. χ^2 = theoretical test parameter. Risk level: 5%.

Block No.	Subset	No. of observations	T	χ^2
3	directions	251	24.7	18.3
4	"	239	22.5	18.3
5	"	80	16.3	15.5
1-5	geodimeter	343	34.0	18.3
1-5	MRA2	145	22.5	18.3

In Table 2 we summarize the Norwegian ED79 data that did not pass the normal distribution test. No bias was found in these sets of data.

Table 2. Summary of rejected H_0 { the data are normally distributed with expectation 0 } of block Norway. (From Sjöberg, 1982). T = computed test parameter. χ^2 = 18.3 (theoretical test parameter). Risk level: 5%.

Block No.	Subset	No. of observations	T
3	all data	498	23.0
4	-"-	120	28.9
5	-"-	383	29.4
1	directions	569	20.3
1-5	-"-	1284	30.2
1-5	MRA1	111	29.5
1-5	MRA3	561	31.8
3	"	79	29.0
5	"	373	29.2

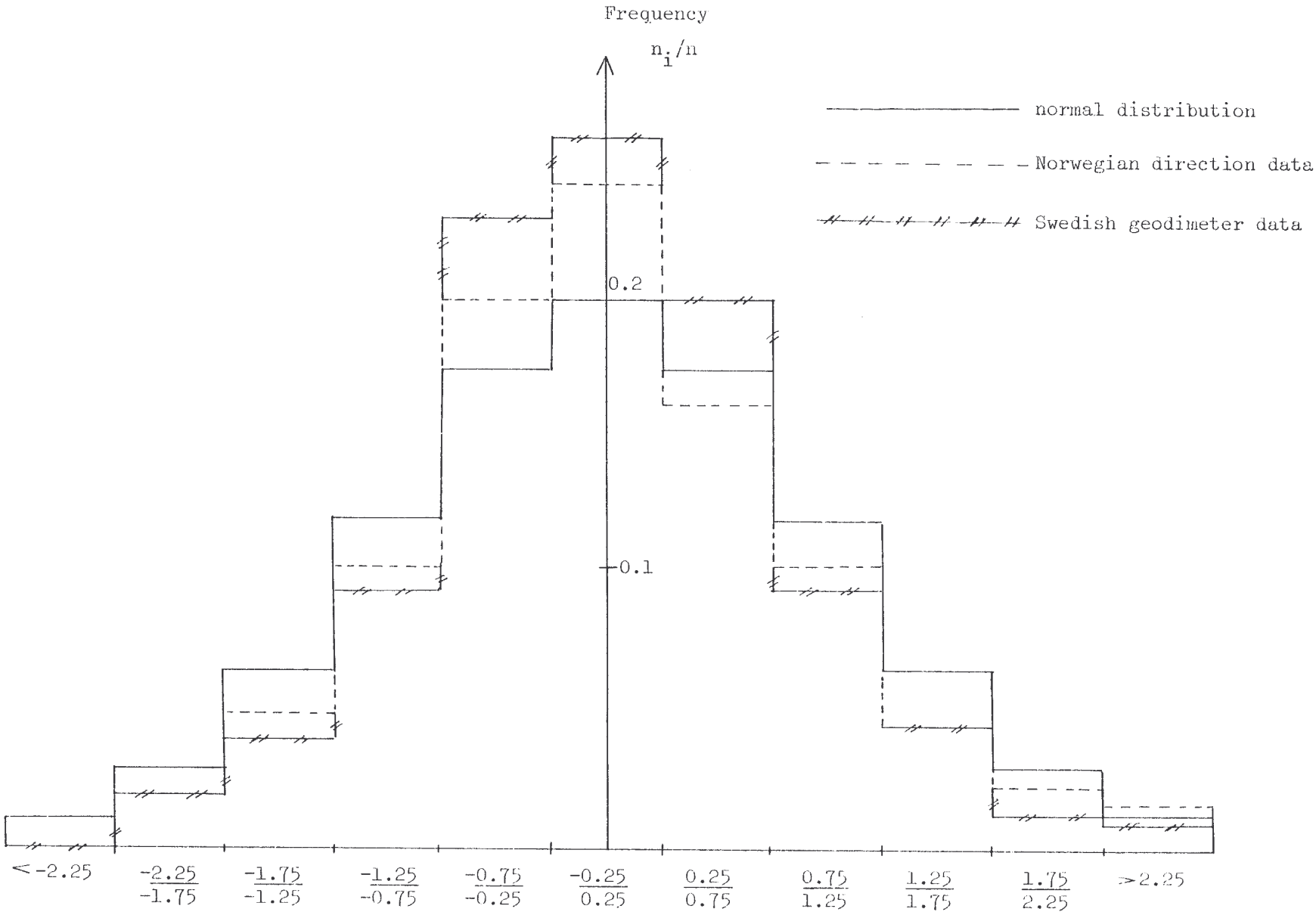

Figure 1. Histogram of the normal distribution function and some obtained distribution functions from Scandinavian RETrig residuals (w_i).

In Fig 1 some typical data sets are compared with the normal distribution. Note that the determined curves are above the normal curve in the central area and below in the periferi. This characteristic is typical for the model developed in sections 4 and 5.

3 Test of bias

If no unknown and unmodelled systematic effects are present in a linear least squares adjustment the expectation of the residuals (w_k) should be zero. A comparison of the mean ($\bar{w}$) and its standard deviation ($s_{\bar{w}}$) of a set of residuals will give you a rough idea of whether the data are biased or not. If a mean $\bar{w} \neq 0$ is obtained, we form the new residuals

$$w_k^1 = w_k - \bar{w}$$

If H_0 :{w_k^1 is N(0,1)} is accepted we may conclude that the distribution is normal with a non-vanishing expectation.

This test is summarized in Table 3. We conclude that the geodimeter observations are not normally distributed, while the opposite holds for the MRA2 observations.

Table 3. Normal distribution tests of Swedish EDM observations before and after subtraction of bias. H_0 {the residuals are normally distributed with expectation 0} is accepted if $T \leq \chi^2_{0.95}$, where $T = \sum_{i=1}^{11} \eta_i$, $\eta_i = (n_i - np_i)^2/np_i$

n_i = number of w within group i. Risk level: 5%

Data set	$\bar{w}$	$\bar{w}$ subtracted?	T	$\chi^2_{0.95}$	H_0 accepted?
Geodimeter	0.151	no	34.0	18.3	no
"	0.151	yes	31.8	16.9	no
MRA2	0.287	no	22.5	18.3	no
"	0.287	yes	7.2	16.9	yes

The rejection of H_0 of MRA2 observations reported in Sjöberg and Eliasson (1981) is completely explained by the bias of the data. However, for all other sets of data given in Tables 1 and 2 the reasons for the negative outcomes of the normal distribution tests are still hidden to us. In the next sections we try another way to explain this obscurity.

4 Mixture of normal variables

The theory of summation of normally distributed variables is well treated in the statistical literature. As a contrast the theory of combining variables is less commonly delt with. According to Hald (1952, p 152) a heterogeneous population is formed by combining two given populations in one given proportion. Thus, if a population is composed of two gaussian populations X_1 and X_2 $[N(\mu_1,\sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively] combined in the ratio a_1/a_2 (where $a_1 + a_2 = 1$), we get a heterogeneous population with the distribution function

$$f(x) = \frac{a_1}{\sigma_1}\phi(\frac{x-\mu_1}{\sigma_1}) + \frac{a_2}{\sigma_2}\phi(\frac{x-\mu_2}{\sigma_2})$$

where ϕ is the standardized normal distribution function. A comprehensive analysis on combination of observations was made by Smart (1958). As a particular case of combination we consider $\mu_1 = \mu_2 = 0$. This condition will be valid for the distribution of the residuals of unbiasedly estimated observations. The distribution function is then modified to

$$f(x) = \frac{a_1}{\sigma_1}\phi(\frac{x}{\sigma_1}) + \frac{a_2}{\sigma_2}\phi(\frac{x}{\sigma_2})$$

As $\phi(0) = 1/\sqrt{2\pi}$ it follows that the central ordinate of this distribution is

$$f(0) = (\frac{a_1}{\sigma_1} + \frac{a_2}{\sigma_2})\frac{1}{\sqrt{2\pi}} \qquad (2)$$

and the second moment m_2 is

$$m_2 = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{a_1}{\sigma_1}\int_{-\infty}^{\infty} x^2\phi(\frac{x}{\sigma_1})\,dx + \frac{a_2}{\sigma_2}\int_{-\infty}^{\infty} x^2\phi(\frac{x}{\sigma_2})\,dx$$

or $$m_2 = a_1\,\sigma_1^2 + a_2\,\sigma_2^2$$

The central ordinate g(0) of a normal variable with expectation 0 and variance m_2 is

$$g(0) = 1/\sqrt{2\pi m_2} \tag{4}$$

Let us compare the ordinates f(0) and g(0). We form the difference.

$$F = f(0) - g(0) \tag{5}$$

Inserting formulas (2) - (4), and noting that $a_1 + a_2 = 1$, we easily prove

$$F = \frac{1}{\sqrt{2\pi}} \frac{a_1 a_2 (\sigma_1 - \sigma_2)^2 (a_1\sigma_2^2 + 2\sigma_1\sigma_2 + a_2\sigma_1^2)}{\sigma_1\sigma_2 \sqrt{a_1\sigma_1^2 + a_2\sigma_2^2}} \tag{6}$$

As $a_1 \geq 0$ and $a_2 \geq 0$ it follows from (5) and (6) that

$$f(0) \geq g(0)$$

where equallity holds for $\sigma_1 = \sigma_2$ Thus we have proved a fundamental feature of a mixed distribution, namely that the central ordinate of this distribution function is greater than (or at least equal to) the central ordinate of the normal curve. As the areas under both curves has the measure 1 and as in the centre the mixture curve is above the normal curve the opposite must be the case away from the centre. This feature, typically displayed in Fig 1, has been named "leptokurtosis" by its discoveror K Pearson (see Romanowski, 1979, p 56).

5 Modulated Normal Distribution

The theory of modulation on the normal distribution was developed by Romanowski (1979). Here we give a short review of his presentation.

According to the original theory of errors by Hagen (1837) each observation is contaminated by an "infinite" number of causes

of errors. In the development of the theory this number (k) is assumed very large but finite. Each cause of error produces a positive or negative elementary error of constant magnitude $\varepsilon/2$ ($\varepsilon>0$). Thus the total random error ranges from $-k\varepsilon/2$ to $k\varepsilon/2$, where these extremes being very rare to occur, but the error equal to zero is of highest probability.

We now assume that in a set of k elementary errors X are positive and Y are negative. Hence the total resulting error becomes

$$H = X\frac{\varepsilon}{2} + Y\left(-\frac{\varepsilon}{2}\right) = \frac{\varepsilon}{2}(X-Y)$$

Assuming that the probabilities for positive and negative elementary errors are equal (1/2) it can be shown that the probability density function (distribution function) of the discrete variable H is

$$f_H = \frac{1}{S\sqrt{2\pi}} \exp(-H^2/2S^2) \qquad (7)$$

where

$$S = \varepsilon\sqrt{k}/2$$

Now, let ε approach zero in such a way that

$$\lim_{\substack{\varepsilon \to 0 \\ k \to \infty}} \frac{\varepsilon^2 k}{4} = \sigma^2 \qquad (8)$$

where σ^2 is a constant. Then the discrete variable H becomes the continuous variable x, obeying

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-x^2/2\sigma^2) \qquad (9)$$

which is the wellknown distribution function of a normally distributed variable with variance σ^2 and expectation 0.

We now extend the possible outcomes of elementary errors to include also zero. This means the elementary error can be $-\varepsilon/2$, 0 or $+\varepsilon/2$. Suppose that among the total set of k elementary errors n differs from $0 (0 \leq n \leq k)$. We postulate also that n is a variable with distribution function

$$\Phi(n) = An^a \tag{10}$$

where A and a are constants. The constant A is a normalizing factor, satisfying

$$\sum_{n=o}^{k} \Phi(n) = 1$$

Assuming that k is very large we can write this equation

$$\int_o^\infty \Phi(n)\, dn = 1$$

or, from (10)

$$A \int_o^\infty n^a dn = 1$$

with the solution

$$A = (a+1)k^{-(a+1)} \tag{11}$$

Finally we assume also that the probability of a positive elementary error is always equal to that of a negative error. Thus the distribution of all total errors corresponding to the same value n (the number of "non-zero" errors) is normal with the variance

$$\sigma_n^2 = \varepsilon^2 n/4 \tag{12}$$

This means that the total population of random errors is a mixture of a large number (k) of gaussian subpopulations of variance σ_n^2, $0 \leq n \leq k$. Thus the total distribution function of x whatever the value of n becomes

$$f(a,x) = \int_o^k \Phi(n) \frac{1}{\sigma_n \sqrt{2\pi}} \exp(-x^2/2\sigma_n^2)\, dn$$

Inserting (10) - (12) we arrive at

$$f(a,x) = \frac{2(a+1)}{\varepsilon\, k^{a+1}\sqrt{2\pi}} \int_0^k n^{a-\frac{1}{2}} \exp(-x^2/2\varepsilon^2 n)\, dn \quad (13)$$

or, after inserting the variance τ^2 of x

$$\tau^2 = \frac{a+1}{a+2}\,\frac{\varepsilon^2 k}{4}$$

into (13)

$$f(a,x) = \frac{(a+1)\sqrt{a+1}}{\tau\sqrt{2\pi}\sqrt{a+2}} \int_0^1 t^{a-\frac{1}{2}} \exp\{(-x^2(a+1)/(2\tau^2 t(a+2))\}\, dt \quad (14)$$

where $t = n/k$

The central ordinate of (14) is

$$f(a,0) = \frac{1}{\sqrt{2\pi}\tau}\,\frac{2a+2}{2a+1}\sqrt{\frac{a+1}{a+2}}$$

The ratio between this ordinate and that of a normal variable of equal variance (τ^2) is

$$\omega(a) = \frac{2a+2}{2a+1}\sqrt{\frac{a+1}{a+2}} \quad (12)$$

In Table 4 we give the ratio $\omega(a)$ for some particular values of the "modulator" a.

Table 4 The index of leptokurtosis $\omega(a)$ for some "modulators" a.

a	$\omega(a)$
0.00	1.09
0.25	1.16
0.50	1.24
1.00	1.41

These ratios clearly shows the leptokurtic feature of the modulation on the normal distribution (cf. section 4).
For the numbers of a given above the integrated distribution function

$$\int_0^y f(a,\lambda)\, d\lambda \qquad (13)$$

where $\lambda = x/\tau$, is tabulated in the Appendix of Romanowski (ibid.) for $0 \leq y \leq 4.5$ with the step 0.01. These tables will be useful for testing the modulated normal distribution on the RETrig data.

6 Tests of Modulation on the Normal Distribution

We now use the χ^2 -test to investigate whether the distributions of the Swedish data of Table 1 and the Norwegian data of Table 2 can be characterized as modulations on normal distributions. The index of modulation $\{\omega(a)\}$ is determined directly from the ratio between the central ordinates of the data function and the normal curve. Knowing $\omega(a)$ the most probable "modulator" a is given by Table 4. Then the test parameter to be compared by the theoretical $\chi^2_{0.95}(f)$ is determined by

$$T = \sum_{i=1}^{11} (n_i - nP_i)^2/nP_i \qquad (14)$$

where

n_i = number of outcomes within group i (cf. Fig 1)

$n = \sum_{i=1}^{11} n_i$

P_i = probability of an outcome within group i (from Tables in Romanowski, 1979).

The number of degrees of freedom f of χ^2 is one less than in the normal distribution test due to the estimation of the parameter a.

Table 5. Summary of χ^2-test of modulation on normal distribution. H_0 = {the data are modulated normal}. H_0 is rejected if $T \geq \chi^2_{0.95}(f)$, where T is given in formula (14) and f is the number of degrees of freedom. Risk level: 5%.

Country	Block No	Subset	ω(a)	a used	T	f	$\chi^2(f)$	H_0 accepted?
Sweden	3	directions	1.09	1.00	20.9	9	16.9	no
"	4	"	0.64 [2]	1.00	24.8	9	16.9	no [2]
"	5	"	0.76 [2]	1.00	19.8	9	16.9	no [2]
			1.34	0.00	22.6	9	16.9	no
"	1-5	geodimeter	1.31 [1]	0.25	19.3 [1]	8	15.5	no
Norway	3	all data	1.21	0.25	13.5	9	16.9	yes
"	4	-"-	1.82	0.00	13.4	9	16.9	yes
"	5	-"-	0.97 [2]	1.00	25.5	9	16.9	no [2]
"	1	directions	1.13	0.50	12.7	9	16.9	yes
"	1-5	"	1.23	0.25	9.3	9	16.9	yes
"	1-5	MRA1	0.78	1.00	24.5	9	16.9	no
"	1-5	MRA3	1.10	1.00	17.8	9	16.9	no
"	3	"	1.60	0.00	22.4	9	16.9	no
"	5	"	0.98 [2]	1.00	25.8	9	16.9	no [2]

1) The bias $\bar{w}$ =0.151 is subtracted

2) The distribution is skew.

From Table 5 we conclude that the non-normal distribution of any Swedish subsets of data cannot be explained by the theory of modulation at the 5% risk level. The distribution of the directions of blocks 4 and 5 are too skew to be accepted. The geodimeter data are accepted after reduction of bias as modulated normal at the risk level 1%($\chi^2_{0.99}$=20.1).

Among the 9 Norwegian subsets tested 4 are accepted as modulated normal distributations. This tendency is particularly clear in the total set of directions. Also in the Norwegian data two sets were found skew.

7 Discussion and concluding remarks

It has not been distinguished whether the biases of the Swedish RETrig data are due to constant "zero-errors" or to scale errors (Sjöberg and Eliasson, 1981). The MRA2 and geodimeter observations failed to pass the standard normal distribution test. When the bias was removed from the MRA2 data the null-hypothesis was accepted (at the 95% level). The geodimeter data shows an additional leptokurtosis (see Fig 1). And indeed, after removal of the bias this data was significantly modulated normal at the one per cent risk level. The distribution of the directions of the Swedish blocks 3-5 cannot be characterized in any of these ways, but the last two data sets are typically skew. Also among the Norwegian data the distribution of 4 subsets can be explained as modulated normal.

As outlined in sections 4 and 5 the origin of the leptokurtic behaviour of the data might be a mixture of various populations of variable variances. It is quite natural that the combination of measurements from various observers and instruments observed under different conditions should have some negative impact on the homogeneity of the data. However, only for five subsets of data did this feature of inhomogeneity prove to be significant.

References

Hagen, G. H. L., Grundzüge der Wahrscheinlichkeitsrechnung. Berlin, 1837.

Hald, A., Statistical Theory with Engineering Applications. John Wiley and Sons, Inc., New York and London, 1952.

Romanowski, M., Random Errors in Observations and the Influence of Modulation on their Distribution. Verlag Konrad Wittwer, Stuttgart, 1979.

Sjöberg, L. and Eliasson L., Swedish National Report on the Computations RETrig phase II, 1980-1981. In the Report on the Symposium of the IAG Subcommission for the New Adjustment of the European Triangulation, London, 11-14 May, 1981.

Sjöberg, L., Statistical Tests on ED79 Residuals for Norway. National Land Survey, Gävle, 1982.

Smart, W. M., Combination of Observations. Cambridge University Press, Cambridge, 1958.