The 3σ-rule for outlier detection from the viewpoint of geodetic adjustment

Author

Prof. Dr.-Ing. Rüdiger Lehmann
University of Applied Sciences Dresden
Faculty of Spatial Information
Friedrich-List-Platz 1
D-01069 Dresden
Tel +49 351 462 3146
Fax +49 351 462 2191
mailto:r.lehmann@htw-dresden.de

Abstract

The so-called 3σ-rule is a simple and widely used heuristic for outlier detection. This term is a generic term of some statistical hypothesis tests whose test statistics are known as normalized or studentized residuals. The conditions, under which this rule is statistically substantiated, were analyzed, and the extent it applies to geodetic least-squares adjustment was investigated. Then, the efficiency or non-efficiency of this method was analyzed and demonstrated on the example of repeated observations.

Keywords

Least-squares adjustment; Outlier detection; Internally/externally normalized/studentized residuals; Hypothesis tests.

Introduction

Geodetic observations are sometimes contaminated by outliers, particularly if they are plentiful. At the starting point, an outlier should be defined. Unfortunately, this is not easy because the definitions found in the literature are countless. Therefore, this paper is restricted to the most widespread definition, which goes back to Hawkins (1980): “An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.” This is not a precise definition because it does not say in which sense the observations are to fit together and when the suspicion that an observation does not fit with the rest is justified. Such a general definition does not exist. In geodesy, the term outlier is defined based on a statistical hypothesis test for the presence of gross measurement errors in the observations (Baarda 1968; Pope 1976; Heck 1981). Observations that are rejected by such an outlier test are called outliers. Therefore, an observation that is not grossly erroneous but is rejected by an outlier test can also be called outlier. It is sufficient that it arouses suspicion in the sense of Hawkins.

The question of what a different mechanism is within the meaning of Hawkins remains. In geodesy, the model of a gross error, i.e., blunder or a large measurement error that occurs rarely and is said to be avoidable, is referred to. The total avoidance of gross measurement errors, however, is connected to a now mostly economically unjustifiable expense. Therefore, it is better to allow for a small number of gross errors and to make them harmless by outlier detection.

The standard adjustment procedure by the least-squares method does not tolerate grossly erroneous observations (e.g., Baarda 1968). In recent years, two categories of advanced techniques for the treatment of observations contaminated by outliers have been developed:

1. Robust adjustment procedures are able to tolerate a small number of grossly erroneous observations without providing immediate grossly erroneous results. Many such methods have been developed in the last decades and their number is increasing at present [see Huber (1981); Hampel et al. (1986); Rousseeuw and Leroy (1987); and Wilcox (2012) for an overview]. The user is often confronted with the problem of choosing the proper procedure, if he has access to software in which such procedures are implemented at all.
2. The already mentioned outlier tests can detect and reject some grossly erroneous observations under the condition that the model design is reasonably good, i.e., the partial redundancy of the tested observation is not too low.

Besides the undoubted advantages of robust adjustment, the outlier tests are also used. The following advantages of outlier analysis are mentioned:

- Detected outliers provide the opportunity to investigate causes of gross measurement errors;
- Detected outliers can be re-measured; and
- If the outliers were discarded from the observations then the standard adjustment software, which operates according to the least squares principle, can be used.

Some robust estimation procedures like Huber's M-estimation by iterative re-weighting (Huber 1981) can also be viewed as a sort of outlier detection. One can recognize detected outliers by their low weights. Therefore, the first two advantages apply also here.

Robust estimation procedures can also be considered as preparatory tools for improved outlier testing. They should be applied in cases in which the standard outlier tests are expected to be insufficient [see, for instance, Koch (1999), Section 3.8.1]. The rationale for this is that robust estimates can be expected to be less distorted than least-squares estimates so that the robustly estimated residuals that correspond to outlying observations are larger than the least-squares residual would be; hence, the robustified outlier statistics can be expected to be more sensitive than the least-squares-based counterparts.

A simple and in geodetic practice (and not only there) widespread method for outlier detection is known as 3σ-rule. An observation is considered as an outlier if its least squares residual exceeds three times its standard deviation (SD). Some authors even refer to the 3σ-rule for the definition of outliers, e.g. Hekimoglu and Koch (2000).

In this paragraph, some selected scientific publications in the field of geodesy, which refer explicitly to the 3σ-rule, are listed. Kutterer et al. (2003) apply the 3σ-rule to the analysis of very long baseline interferometry (VLBI) observations. Both normalized residuals (see subsequently) as well as externally studentized residuals (see subsequently) are considered. If they exceed the critical value of 3, the observation is downweighted such that its impact is practically nulled. Neitzel (2004, p. 91) mentions the 3σ-rule and calls it well known but does not actually recommend or apply it. Instead, the choice of a probability of type I decision error α is preferred (see subsequently). Featherstone and Morgan (2007) use the 3σ-rule for validation of the AUSGeoid98 model in Western Australia. Together with their SDs, 435 differences between astronomic Helmert deviations and AUSGeoid98-derived deviations or deflections of the vertical are computed. Of these, 15 exceed the ratio of 3 and are rejected as outliers. van Loon (2008, p. 100) refers to the 3σ-rule as widely used and tries to apply it to the processing of CHAMP satellite gravity data. However, it is found that a method called cost-function estimation performs better.

Because initially it remains unclear how the SD is calculated, the 3σ-rule is actually a generic term for several simple methods for outlier detection, as is shown in the following. The different test statistics used when applying the 3σ-rule are reviewed. Then, the efficiency or nonefficiency of this method is analyzed and demonstrated on the example of repeated observations.

Residuals in Least Squares Adjustment

To evaluate the results of a least squares adjustment, the residuals v are of the greatest significance. If an observation generates an unexpectedly large (in magnitude) residual, then the suspicion that the assumptions of the adjustment model do not apply is justified. The only false assumption under investigation in this paper is that the observations are free of outliers.

The magnitude of a residual is unexpectedly large if it is outside of its confidence interval. This interval covers the true value of the residual with high probability 1-α, the confidence level. It is derived for each residual by its probability distribution implied by the adjustment model. Starting from a linear or linearized functional adjustment model (observation equations)

$$l = Ax + e \tag{1}$$

with the n-vector of observations l, the n-vector of observation errors e, the u-vector of adjustment parameters x and the $n \times u$-matrix A (matrix of observation equations) and from a stochastic adjustment model for normal distributed observation errors

$$e \sim N(0, \sigma^2 P^{-1}) \tag{2}$$

with positive definite symmetric weight matrix P and the a priori variance factor σ^2, the following is found for the least squares solution for the vector of residuals:

$$v = -Q_{vv} P l \tag{3}$$

the multivariate normal distribution

$$v \sim N(0, \sigma^2 Q_{vv}) \tag{4}$$

with cofactor matrix of the residuals

$$Q_{vv} = P^{-1} - A(A^T P A)^{-} A^T. \tag{5}$$

The superscript minus sign symbolizes a generalized matrix inverse. It will be requested for rank deficient adjustment models. If A and P have full column rank then the generalized matrix inverse is unique and coincides with the classical matrix inverse.

If one is interested only in a single residual v_i, then the confidence interval is derived from the marginal distribution of Eq. (4), which assumes the form

$$v_i \sim N\left(0, \sigma^2 q_{vv,ii}\right).$$

where $q_{vv,ii}$ denotes the ith diagonal element of Q_{vv}. The test of whether or not v_i is inside its confidence interval corresponds to the statistical test of the null hypothesis

$$H_0: E\{v_i\} = 0 \tag{6}$$

versus the alternative hypothesis

$$H_A: E\{v_i\} \neq 0. \tag{7}$$

It should be stressed that Eq. (7) refers to one outlier in the ith observation only. Therefore, the corresponding test cannot be expected to correctly detect multiple outliers. And in fact, it often fails to do so (see Hekimoglu and Koch 2000; Xu 2005; Baselga 2007, 2011). Detecting multiple outliers gets more complicated.

In contrast to the calculation of the adjusted parameters x and their SDs, for the calculation of confidence intervals and for hypothesis testing a distribution assumption, Eq. (2), is mandatory.

Normalized Residuals: Gauss Test

In geodetic adjustment textbooks (e.g. Koch 1999, Teunissen 2000), one finds for the test of H_0 versus H_A, the test statistic

$$T_{n,i} = \frac{v_i}{\sigma\sqrt{q_{vv,ii}}} \tag{8}$$

and refers to this as individual normalized (or standardized) residual. The residual is divided by its a priori SD, which only has the advantage that for $T_{n,i}$, the limits of the confidence interval are found in statistical lookup tables. Under the null hypothesis H_0 in Eq. (6), $T_{n,i}$ is found to be standard normally distributed

$$T_{n,i}|H_0 \sim N(0,1) \tag{9}$$

A rationale for the choice of test statistic, Eq. (8), is that for a systematically acting gross error (bias) in the observation l_i the corresponding test has maximum test power. Kargoll (2012) provides a comprehensive treatise on this topic. Besides the global or overall test, this so-called local or individual test is part of the famous data snooping according to Baarda (1968).

For a confidence interval $[-c, c]$ of $T_{n,i}$, with the cumulative distribution function Φ of $N(0,1)$, the probability of test decision error or significance level is

$$\alpha = 2\Phi(-\mathrm{c}) = 2 - 2\Phi(\mathrm{c}). \tag{10a}$$

Inversely, one derives c from α using the inverse function Φ^{-1}

$$c = \Phi^{-1}(1 - \alpha/2)$$

Implementations of both functions are widely accessible, e.g. in Microsoft Excel as NORMSDIST and NORMSINV.

The null hypothesis H_0 is to be rejected if $|T_{n,i}|$ exceeds the critical value c. This value is equal to the quantile of the corresponding distribution. Table 1 lists some common pairs (α, c). The general standard values for tests α = 0.05, where $T_{n,i}$ differs significantly from 0, and α = 0.01, where $T_{n,i}$ differs highly significantly from 0, which are occasionally recommended for normalized residuals, yield small critical values, c. The 3σ-rule suggests directly using a critical value of $c = 3$, which corresponds to a probability of type I decision error of

$$\alpha = 2\Phi(-3) = 0.0027. \tag{10b}$$

Such a very small probability of decision error α causes H_0 to be rejected rarely, if it is true, which is only with probability $\alpha = 0.0027$ (type I decision error). However, it is also more likely to be accepted even if it is false (type II decision error). This probability is denoted by β and depends on the size or the stochastic properties of the gross errors. Therefore, outliers remain unidentified more often and useful observations are discarded more rarely than, e.g., for α = 0.05.

In geodetic literature, β is sometimes used to denote the test power. This is the probability that a false H_0 is rejected. Here the line of the current statistical literature is followed, where β usually denotes the probability of type II decision error and 1-β denotes the test power.

Table 1. Critical values c for Individual Normalized Residual [Eq. (8)]

Significance level α	0.05	0.01	0.0027	0.001
Critical value c for Eq. (8)	1.96	2.58	3.00	3.29

Internally Studentized Residuals: τ Test according to Pope

If the variance factor σ² is not known a priori because there is insufficient experience with the measuring technology, then the calculation of normalized residual is impossible. But σ² can be replaced by an a posteriori estimated variance factor $\hat{\sigma}^2$. The new test statistic

$$T_{s,i} = \frac{v_i}{\hat{\sigma}\sqrt{q_{vv,ii}}} \tag{11}$$

has no normal distribution and is called individual studentized residual. In statistics studentization means the standardization with respect to an estimate of the variance. The name is derived from the pseudonym Student of the statistician William Sealy Gosset. The rationale for the choice of this test statistic is equivalent to that of normalized residuals (Kargoll 2012).

Depending on which estimate of σ² is computed, a different distribution of $T_{s,i}$ is obtained. The most common estimate is the best quadratic unbiased estimate

$$\hat{\sigma}^2 = \frac{v^T P v}{r} \tag{12}$$

with redundancy $r = n - \mathrm{rank}(A)$, which for regular adjustment problems equals $r = n - u$. Here, $r > 1$ is routinely assumed. Internally studentized residuals are then spoken of, and it is found that under H_0 in Eq. (6), the test statistic Eq. (11) has a so-called τ distribution with $r - 1$ degrees of freedom (Pope 1976; Heck 1981)

$$T_{s,i}|H_0 \sim \tau(r-1). \tag{13a}$$

This is the distribution of a random variable (Pope 1976, p.13)

$$\tau = \sqrt{\frac{r}{r-1+t_{r-1}^2}}\, t_{r-1} \tag{13b}$$

where t_{r-1} = random variable with Student's t distribution with $r - 1$ degrees of freedom (Fig. 1). It is found that the τ distribution has a bounded support, i.e. the random variable τ cannot assume arbitrarily small or large values (Baselga 2007). Namely

$$|\tau| = \left|\sqrt{\frac{r}{r-1+t_{r-1}^2}}\, t_{r-1}\right| \leq \left|\sqrt{\frac{r}{t_{r-1}^2}}\, t_{r-1}\right| = \sqrt{r}.$$

Moreover, this bound is also valid if H_0 is not true. This can be verified as follows: After the observation l_i has been discarded from vector l, the variance factor σ² can be estimated according to Heck (1981) from the rest of the observations by

$$\hat{\sigma}_i'^2 = \frac{1}{r-1}\left(v^T P v - \frac{v_i^2}{q_{vv,ii}}\right) \geq 0. \tag{14}$$

Taking Eq. (12) into account, this can be re-written as

$$r\hat{\sigma}^2 \geq \frac{v_i^2}{q_{vv,ii}}$$

and by Eq. (11) this is equivalent to $|T_{s,i}| \leq \sqrt{r}$. Magnitudes of internally studentized residuals of Eq. (11) can therefore not exceed $\sqrt{r}$, not even if arbitrarily large outliers are present in the observations. The application of the 3σ-rule to $T_{s,i}$, which must now more correctly be called the $3\hat{\sigma}$-rule, is therefore meaningless right from the beginning if $r \leq 9$. Not the most nonsensical observation is detected in this way.

Critical values for the τ distribution can be calculated from those of the t distribution by

$$c = F_\tau^{-1}\left(1-\frac{\alpha}{2}|r-1\right) = \sqrt{\frac{r}{r-1+c_t^2}}c_t \text{ with } c_t = F_t^{-1}\left(1-\frac{\alpha}{2}|r-1\right)$$

where F_τ^{-1} and F_t^{-1} denote the inverse cumulative distribution functions of the τ- und t-distribution. This relationship is a direct consequence of Eq. (13b). Critical values may be computed e.g. with Microsoft Excel using the function T.INV as an implementation of F_t^{-1}.

Table 2 gives some critical values. Conversely, one can calculate the significance level α, which belongs to $c = 3$ ($3\hat{\sigma}$-rule). It is increasing with redundancy r (see Table 2). The larger the redundancy, the more likely a true H_0 is rejected, i.e., the more likely useful observations are discarded. One could argue that with high redundancy, this loss is rather to be tolerated. The probability of type I error does fortunately not exceed 0.0027, but approaches this value for increasing r as $\hat{\sigma}^2$ in Eq. (12) approaches σ² (see Table 2).

Finally, it should be stressed that the studentized residuals, Eq. (11), should only be used if normalized residuals, Eq. (8), cannot be computed due to unknown σ.

Table 2. Critical values c for Individual Studentized Residual and Error Probabilities α When Using $3\hat{\sigma}$-rule

Redundancy r		2	3	4	5	10	15	20	25	30	40	50
τ-Test	c for α=0.05	1.41	1.65	1.76	1.81	1.9	1.93	1.94	1.94	1.94	1.95	1.95
	c for α=0.001	1.41	1.73	1.98	2.18	2.68	2.87	2.97	3.04	3.08	3.13	3.16
	α for c=3	0	0	0	0	0.0000	0.0004	0.0009	0.0012	0.0014	0.0017	0.0019
t-Test	c for α=0.05	12.71	4.3	3.18	2.78	2.26	2.14	2.09	2.06	2.05	2.02	2.01
	c for α=0.001	636.62	31.6	12.92	8.61	4.78	4.14	3.88	3.75	3.66	3.56	3.5
	α for c=3	0.2048	0.0955	0.0577	0.0399	0.015	0.0096	0.0074	0.0062	0.0055	0.0047	0.0042

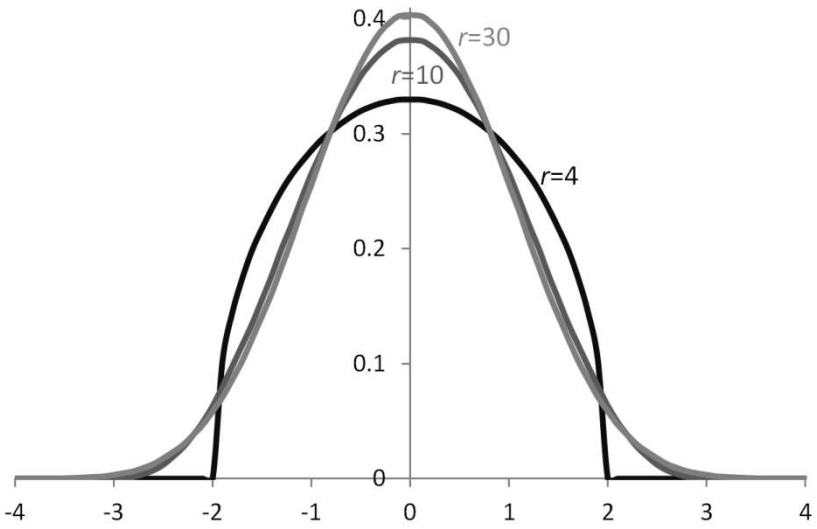

Fig. 1. Some probability density functions of the τ-distribution with r degrees of freedom

Externally Studentized Residuals: *t*-Test

The reason that the 3σ-rule applied to internal studentized residuals, Eq. (11), (3$\hat{\sigma}$-rule) at $r \leq 9$ cannot detect outliers, is rooted in the fact that the residual v_i is also used within the calculation of the estimate Eq. (12). Thus, the potentially large magnitude of the numerator is partly counterbalanced by a large magnitude of the denominator. The magnitude of the test statistic Eq. (11) cannot exceed the critical value of 3 when $r \leq 9$.

A remedy of this situation is to replace the estimate Eq. (12) with Eq. (14). The resulting test statistic

$$T'_{s,i} = \frac{v_i}{\hat{\sigma}'_i \sqrt{q_{vv,ii}}} \tag{15}$$

is called individual externally studentized residual. It turns out that under the condition Eq. (2), the numerator and denominator are stochastically independent and constitute a *t* statistic (Pope 1976, Heck 1981)

$$T'_{s,i}|H_0 \sim t(r-1). \tag{16}$$

The calculation of the critical value is now slightly easier

$$c = F_t^{-1}\left(1 - \frac{\alpha}{2}|r-1\right)$$

because F_t^{-1} is more widely accessible. However, this convenience comes at the cost of computing a new value $\hat{\sigma}_i'^2$ for each outlier-suspected observation l_i.

Table 2 gives some c values and the error probabilities α belonging to $c = 3$. It is noticeable that at low redundancy, the 3$\hat{\sigma}$-rule is connected with a very high significance level, α, even $\alpha > 0.2$. The reason for this is that for low redundancy $T'_{s,i}$ assumes significantly larger values than $T_{s,i}$, even without outliers. Here $|T'_{s,i}| > 3$ is then probably not a serious indication of one or more outliers.

The test statistics $T_{s,i}$ and $T'_{s,i}$ satisfy the identity

$$\frac{r-1}{T'^2_{s,i}} = \frac{r}{T^2_{s,i}} - 1.$$

If one is given, the other can be easily calculated. Both test statistics thus convey the same information about possible outliers. For the same significance level α and the corresponding critical value c both $T_{s,i}$ as well as $T'_{s,i}$ bring about the same test decision for H_0. But this is not valid if the 3$\hat{\sigma}$-rule is used, where in the first place c is fixed instead of α: H_0 is rejected more often using $T'_{s,i}$ than using $T_{s,i}$. Moreover, if $r \leq 9$ then H_0 can be rejected exclusively using $T'_{s,i}$. Here, the 3$\hat{\sigma}$-rule turns out to be too simple.

Ignorance of Outlier-Suspected Observations

Testing the hypotheses Eq. (6) versus Eq. (7) only makes sense if it is known a priori, which observation might be affected by a gross measurement error and is therefore candidate for an outlier. In case of Eq. (7), the outlier-suspected observation is l_i. This knowledge should not be derived from the observed values themselves, otherwise, a post-hoc hypothesis, a hypothesis suggested by the observations, is reached. If one already detects a feature in the observations and then tests the hypothesis that this feature is present, this hypothesis will almost never be rejected, even if the feature in the observations occurred only by chance. In using the residual with the extreme, i.e., largest magnitude, normalized or studentized value is not allowed in Eqs. (8), (11), or (15). Example: Assume that a vector of five repeated observation $l = (16{,}10{,}63{,}17{,}11)^T$ is obtained. By inspection of the observations it is suspected that the 3rd observation is an outlier. Simply setting $i = 3$ in Eqs. (8), (11), or (15) is not allowed because Eqs. (6) and (7) and therefore also i must be formulated independently of the observations.

An immediately admissible way to get knowledge of an outlier-suspected observation is only that irregularities occurred and were reported during the measurement, e.g., during a leveling campaign, and it could have been reported that the leveling rod was exceptionally not established on firm ground. At most, a reading that has an unusual number of zeros at the end may arouse suspicion that in the process of preevaluation, it was rounded inadmissibly and may therefore be grossly erroneous, or two observations surprisingly assume the same value, so a blunder in the processing of observations is suspected.

However, if no outlier-suspected observation is known a priori then only testing the null hypothesis remains

$$H_0: E\{v_1\} = 0 \wedge \ldots \wedge E\{v_n\} = 0 \tag{17}$$

No outliers exists in the observations. This is compared to the alternative hypothesis

$$H_A: E\{v_1\} \neq 0 \vee \ldots \vee E\{v_n\} \neq 0 \tag{18}$$

At least one outlier exists in the observations. The next procedure is to convert this hypothesis testing to a family of hypothesis tests as follows:

$$\begin{aligned} H_0^{(1)}: E\{v_1\} = 0 \text{ vs. } H_A^{(1)}: E\{v_1\} \neq 0 \\ \vdots \\ H_0^{(n)}: E\{v_n\} = 0 \text{ vs. } H_A^{(n)}: E\{v_n\} \neq 0. \end{aligned} \tag{19}$$

These are sequentially performed as described previously, respectively. If any $H_0^{(i)}$ is rejected then also H_0 must be rejected because if $H_0^{(i)}$ is false, then so is H_0.

Extreme Normalized and Studentized Residuals

Test statistics T_i in the n individual tests of the test family, Eq. (19), are preferably the normalized residuals, Eq. (8), or, if σ² is not known, either the internally or the externally studentized residuals, Eq. (11) or (15).

If l_i and l_j are both outliers, then it is expected that $H_0^{(i)}$and $H_0^{(j)}$ are both rejected, and therefore H_0 is rejected. This shows that Eqs. (17)–(19) are also valid for multiple outliers. But in this case, test statistics Eqs. (8), (11), and (15) are not optimal, such that the test often yields a decision error.

Because the redundancies are equal in all *n* tests, the same critical value *c* is obtained throughout. If in at least one of these tests

$$|T_i| > c$$

then $H_0^{(i)}$ must be rejected and consequently H_0 is also rejected.. This is equivalent to saying that H_0 must be rejected if

$$T := \max_{i=1\ldots n} |T_i| > c \tag{20}$$

holds. The new test statistic T for testing the hypothesis H_0 in Eq. (17) vs. H_A in Eq. (18) is thus either the extreme normalized or extreme studentized residual. At this point there seems to be a contradiction to the remarks of the last section, where it was said that these values should not be used to identify the outlier-suspected observation. But this was not done here, because the hypotheses H_0 and H_A, for which T is the test statistic, were established without reference to a special observation. Consequently they are no post-hoc hypotheses. Strictly speaking, with $T > c$ only the hypothesis H_A is accepted, but this does not yet say which observation is an outlier.

A remaining problem is the calculation of the critical value c of T or vice versa with the calculation of the significance level α associated with $c = 3$. This would require the determination of the probability distribution of $T|H_0$. There is no analytical solution and a numerical solution is

computationally demanding. With today's computer technology, computing such a solution is feasible (Lehmann 2012b), but from the literature a simple and popular approximation method is known (Pope 1976, Koch 1999). If $T_i|H_0, i = 1, \dots, n$ are stochastically nearly independent, then

$$\begin{aligned} 1 - \alpha &= P(T < c|H_0) \\ &= P(|T_1| < c \wedge \dots \wedge |T_n| < c|H_0) \\ &\approx \prod_{i=1}^{n} P(-c < T_i < c|H_0) = \prod_{i=1}^{n} (1 - \alpha') = (1 - \alpha')^n. \end{aligned} \tag{21}$$

if the correlations between T_i and $T_j, j \neq i$ are neglected. Here, α' denotes the significance level of the individual tests, Eq. (19), and α is the significance level of the initial test Eq. (17) versus Eq. (18). Baarda (1968) recommends α = 0.001 for geodesy (see also Mierlo 1983).

Lehmann (2012b) first tested the assumption of approximate stochastical independence of $T_i|H_0$. Based on a levelling network, it has been shown that there may be significant differences, at least if the redundancy is small, compared with the number of observations ($r \ll n$). The test statistics T_i are never completely independent because already the numerators are stochastically dependent.

Because $\alpha' \ll 1$ is typically chosen, the so-called Bonferroni equation (Abdi 2007) is derived from $(1 - \alpha')^n \approx 1 - n\alpha'$

$$\alpha \approx n\alpha'. \tag{22}$$

Because T_i has a known distribution, normal, τ, or t distribution, the relationship between c and α' is known and the relationship between c and α is also known with Eq. (22): The test of the extreme test statistic Eq. (20) corresponds approximately to the test of the individual test statistic T_i, where i is the index, for which the maximum in (20) is attained if the significance level α is further divided by n. If, however, c is fixed instead of α, as implied by the 3σ-rule or 3$\hat{\sigma}$-rule, then this difference disappears. This could lead to misunderstandings.

For extreme normalized residuals, $c = 3$ corresponds to a value $\alpha' = 0.0027$ [see Eq. (10a)], and from Eq. (22)

$$\alpha \approx n \cdot 0.0027$$

For extreme studentized residuals one would have to multiply the appropriate α value in Table 2 by n. This variant of the outlier test is the standard variant in most nongeodetic areas of application and is called the Grubbs outlier test (Grubbs 1969). If was originally designed for and is mostly applied to the detection of outliers in repeated observations (statistical samples).

The 3σ-rule applied to extreme test statistics, Eq. (20), means in each case that α is much larger again, especially for adjustment calculations with a large number of observations. Here, a true H_0 is probably rejected, which is equivalent to the loss of useful observations. One should realize that already $n = 200$ causes $\alpha \approx 0.5$.

The critical value

Whether $c = 3$ is generally suitable as a critical value is now considered. It can not be achieved in this way that both α and β become arbitrarily small. With a 2σ-rule more grossly erroneous observations were detected (β small), but also more useful observations will be lost (α large). At a 4σ-rule the relation would be reversed. This complementary behavior with respect to α and β is typical in parametric hypothesis tests. A smaller α implies a larger β and vice versa (Lehmann and Romano 2005, p.57).

A compromise must surely depend upon which loss weighs more heavily (Mierlo 1983). To find the best compromise, the loss function premium and the profit function protection, originally introduced by Anscombe (1960), are applied to geodesy for the first time by Lehmann (2013). Actually already

(Lehmann 2010) and (Lehmann and Scheffler 2011) worked with the protection-term, without naming it so, because at that time the term was not known to the authors. Fixed values for α and β, as can be found in the geodetic literature (e.g. Baarda 1968) should be scrutinized in the current state of computing technology.

Example: Repeated Observations

Adjustment Model

Consider for illustration the simplest conceivable example, namely the n-fold direct measurement of a scalar parameter x, i.e. $u = 1$. The system (1) assumes the form

$$l = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} x + e$$

with the stochastic model, Eq. (2), belonging to H_0 in the form

$$H_0: e \sim N(0, \sigma^2 I) \tag{23}$$

where σ^2 is based on long-standing experiences with this type of measurement and is therefore assumed known. From Eq. (5)

$$Q_{vv} = \frac{1}{n} \begin{pmatrix} n-1 & -1 & \dots & -1 \\ -1 & n-1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \dots & -1 & n-1 \end{pmatrix} \tag{24}$$

and consequently $q_{vv,ii} = (n-1)/n$.

Null Hypothesis Is True

Since an outlier-suspected observation is not known a priori, the extreme normalized residual is used as the test statistic. The relationship between α and c is established approximately by using Eqs. (10) and (22):

$$\alpha \approx 2n\Phi(-c) \tag{25}$$

By the method of Lehmann (2012b), an accurate calculation can also be performed. Fig. 2 shows this for $n = 3$. The null hypothesis H_0 is rejected, whenever $|T_{n,1}| > c$ or $|T_{n,2}| > c$ or $|T_{n,3}| > c$, where because of $v_1 + v_2 + v_3 = 0$

$$|T_{n,3}| = \frac{|v_3|}{\sigma\sqrt{q_{vv,ii}}} = \frac{|v_1 + v_2|}{\sigma\sqrt{2/3}} = |T_{n,1} + T_{n,2}|$$

and $1 - \alpha$ is the probability, that $(T_{n,1}, T_{n,2})$ falls into the region of acceptance (white area in Fig. 2)

$$\alpha = 1 - \iint_{\text{region of acceptance}} \varphi\,(T_{n,1}, T_{n,2}) dT_{n,2} dT_{n,1} \tag{26}$$

Here, φ is the probability density of $(T_{n,1}\ T_{n,2})$, which can easily be deduced from Eqs. (4) and (24)

$$(T_{n,1}\ T_{n,2})\ |H_0 \sim N\left(0, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right).$$

The integral in Eq. (26) can be easily calculated for this case with a numerical cubature formula. In higher dimensions, the calculation requires a Monte Carlo integration (Lehmann 2012b). Fig. 3 shows that there are no significant differences between Eq. (25) and the cubature, Eq. (26), so that the approximation, Eqs. (21) and (22), practically suffices here. This may be surprising because the correlation between the test statistics exhibits a correlation coefficient of -0.5, which does not seem to be negligible.

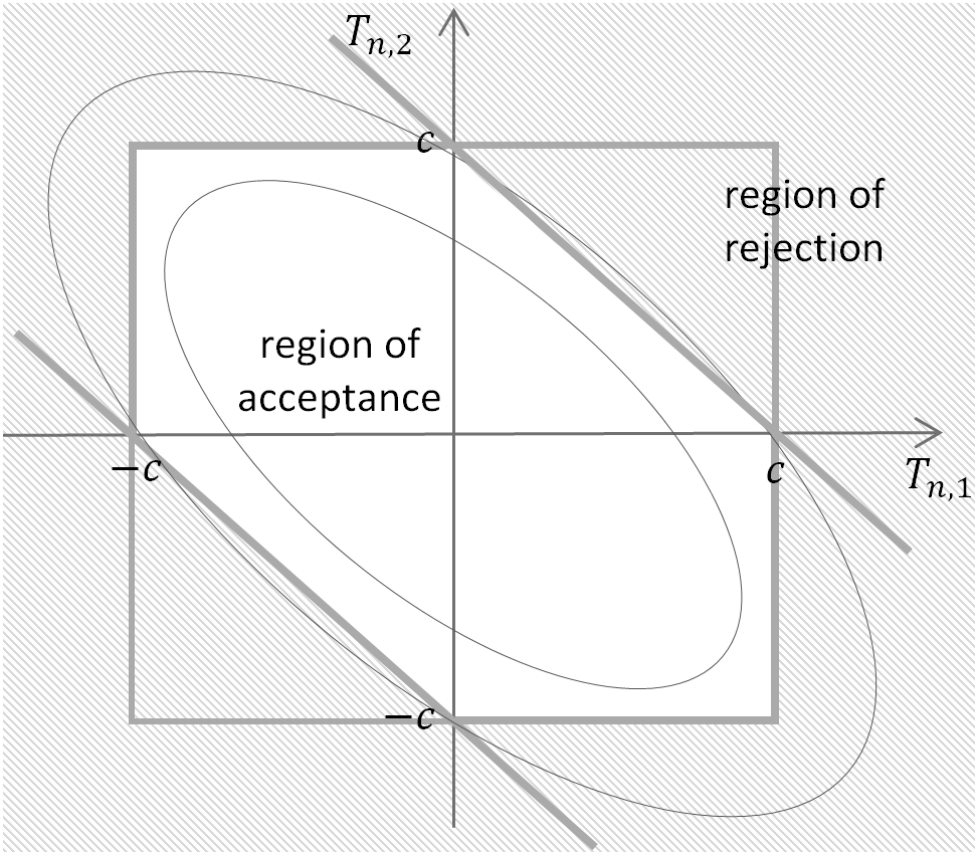

Fig. 2. Test for three repeated observations: H_0 is rejected whenever $|T_{n,1}|$ or $|T_{n,2}|$ or $|T_{n,1} + T_{n,2}|$ exceed the critical value c. The ellipses represent the contour lines of φ in Eq. (26).

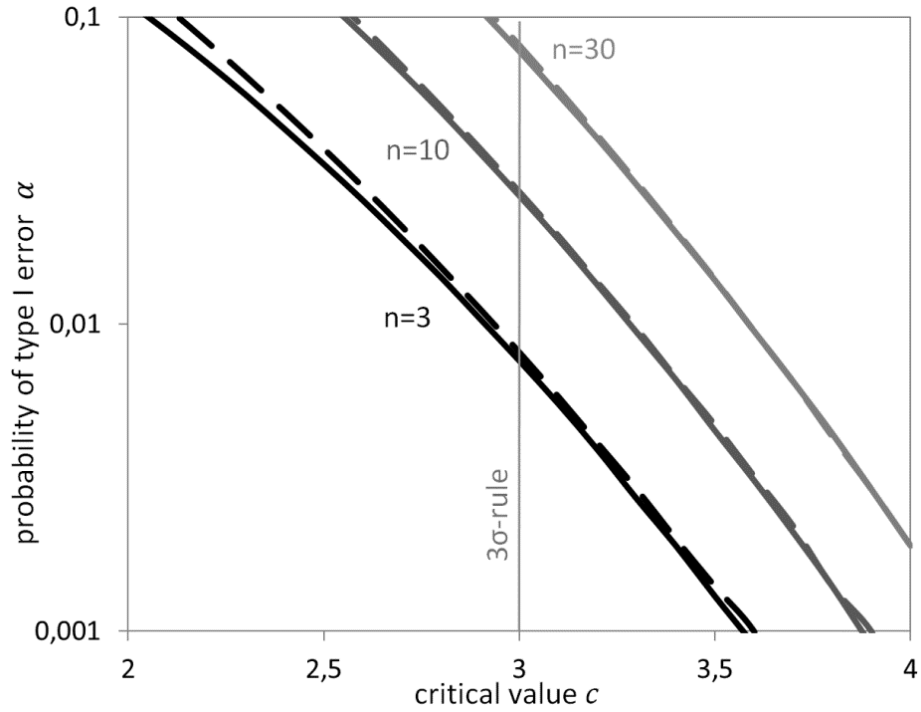

Fig. 3. Relationship between the probability of type I error α and the critical value c for n = 3, 10 und 30 repeated observations; dashed curves: approximation according to Eq. (25); solid curves: numerical calculation for n=3 by cubature of Eq. (26), otherwise by Monte Carlo integration

Alternative Hypothesis Is True: Gross Errors Act Systematically

The relationship between the probability of type II decision error β and the critical value c is not as easily established. In particular, a specific alternative hypothesis H_A is needed. This should generally be

H_A: At least one outlier is present in the observations.

This hypothesis is converted into a family of n hypotheses

$H_A^{(i)}$: The observation l_i is an outlier, $i = 1, \dots, n$,

of which at least one must be accepted to be able to reject H_0, i.e., to have one or more outliers detected. For each $H_A^{(i)}$ the appropriate test statistic is the corresponding $T_{n,i}$ in Eq. (8).

The variable β is now the probability that H_0 is accepted, although at least one $H_A^{(i)}$ is true. Under the already previously used assumption of the approximate statistical independence of the test statistics $T_{n,i}, i = 1, \dots, n$, which may also be valid under H_A, the following is obtained:

$$\beta = P(T < c|H_A) \approx \prod_{i=1}^{n} P\left(|T_{n,i}| < c|H_A\right) = \prod_{i=1}^{n} \left(1 - P\left(|T_{n,i}| > c|H_A\right)\right).$$

where T = extreme normalized residual. Now, $T_{n,i}$ is assumed to be large in magnitude mainly due to l_i being an outlier. This motivates the approximation

$$P\left(|T_{n,i}| > c|H_A\right) \approx P\left(|T_{n,i}| > c|H_A^{(i)}\right).$$

The latter probability is the same for each repeated observation. Therefore, it is sufficient to compute it for the first observation ($i = 1$) only

$$P\left(|T_{n,i}| > c|H_A^{(i)}\right) = P\left(|T_{n,1}| > c|H_A^{(1)}\right)$$

This is summarized as

$$\beta \approx \prod_{i=1}^{n} \left(1 - P\left(|T_{n,1}| > c|H_A^{(1)}\right)\right) = \left(P\left(|T_{n,1}| < c|H_A^{(1)}\right)\right)^n \tag{27a}$$

and using $q_{vv,ii} = (n-1)/n$, it is found that

$$\beta \approx P\left(|v_1| < \sigma c\sqrt{(n-1)/n}|H_A^{(1)}\right)^n. \tag{27b}$$

As an alternative to H_0 in Eq. (23), the first observation was falsified by an additional non-random, i.e. systematically acting, gross error e_g. This is the standard assumption in outlier detection, e.g. (Hekimoglu et al. 2012). Accordingly, the following is formulated:

$$H_A^{(1)}: e \sim N\left(\begin{pmatrix} e_g \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 I\right). \tag{28}$$

From that, the following is found with Eq. (3):

$$v|H_A^{(1)} \sim N\left(\frac{e_g}{n}\begin{pmatrix} 1-n \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \sigma^2 Q_{vv}\right).$$

To calculate the probability β in Eq. (27), the marginal distribution of v with respect to $v_1$is needed. This one reads directly from the joint distribution of the components of the vector v

$$v_1|H_A^{(1)} \sim N\left(\frac{1-n}{n}e_g, \frac{n-1}{n}\sigma^2\right)$$

and hence

$$T_{n,1}|H_A^{(1)} \sim N\left(-\sqrt{\frac{n-1}{n}}\frac{e_g}{\sigma}, 1\right).$$

These distributions are actually the same for all individual test statistics, thus independent of the index i of the outlier-suspected observation. Finally, the following is deduced from Eq. (27a)

$$\beta = \beta(c) \approx P\left(-c < T_{n,1} < c|H_A^{(1)}\right)^n = \left[\Phi\left(\sqrt{\frac{n-1}{n}}\frac{e_g}{\sigma} + c\right) - \Phi\left(\sqrt{\frac{n-1}{n}}\frac{e_g}{\sigma} - c\right)\right]^n. \quad (29)$$

Fig. 4 shows some functions $\beta(c)$. First it is understood that the size of the gross error e_g has an effect on β, (see Hekimoglu and Koch 2000). Surprising, perhaps, is how large this influence actually is. Example: For $n = 10$ observations, with the 3σ-rule , it is found that $\alpha = 0.027$ in Fig. 3 and $\beta = 0.82$ for $e_g = \sigma$, $\beta = 0.0031$ for $e_g = 3\sigma$ as well as $\beta = 1.2 \cdot 10^{-14}$ for $e_g = 5\sigma$ in Fig. 4. Fig. 4 does not display the last value. Small gross errors are predictive of an acceptance of H_0 and are thus rarely detected. On the other hand, gross errors of a size of 3σ and larger are almost certainly detected in a practically not too small set of observations.

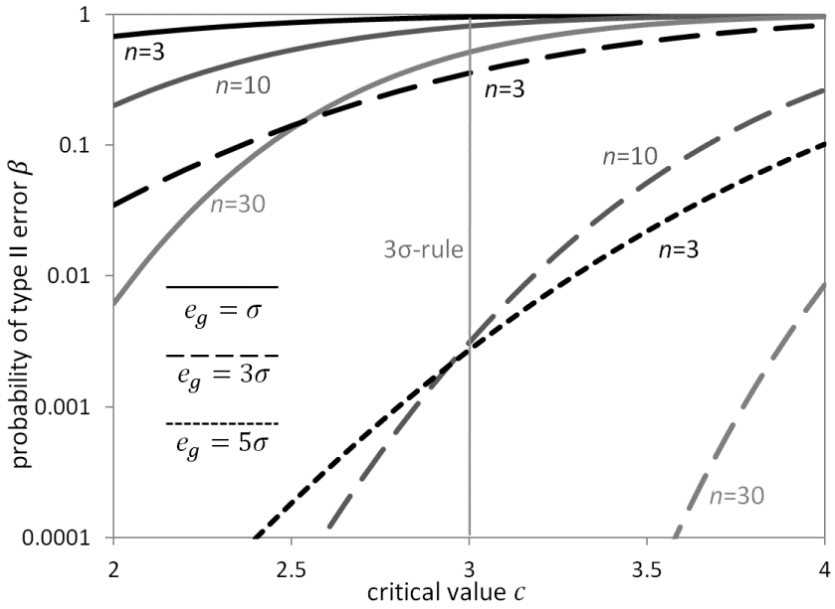

Fig. 4. Relationship between the probability of type II error β and the critical value c for n = 3, 10 und 30 repeated observations; solid curves: $e_g = \sigma$; dashed curves: $e_g = 3\sigma$; dotted curves: $e_g = 5\sigma$

In comparison with Fig. 3, Fig. 4 also shows the following:

1. The variable β depends on c in the opposite way as α does: $\beta(c)$ increases monotonically. An increase of c, i.e. in the direction of a 4σ-rule, will therefore not only detect good observations as outliers less frequently, but will also leave grossly erroneous observations more often undetected, with a decrease of c accordingly reversed.
2. Also the dependence of n is reversed for β. More observations cause a grossly erroneous observation to go less probably undetected (Hekimoglu and Koch 2000). This is due to the exponent n in Eqs. (27) and (29). The more individual tests are performed, the less probable it is that none reject H_0.

Alternative Hypothesis Is True: Gross Errors Act Randomly

Alternatively, the case that the gross errors act randomly, i.e., by repeating the observations, they assume a different value and possibly a different algebraic sign is also considered. The basic procedure remains the same as in the previous subsection. The alternative hypothesis $H_A^{(i)}$ does not mean that l_i is shifted in expectation by e_g anymore, but has an increased variance. According to the law of variance propagation, the previous variance σ^2 is added to the variance σ_g^2 of gross error e_g. Eq. (28) is replaced by

$$H_A^{(1)}: e \sim N\left(0, \begin{pmatrix} \sigma^2+\sigma_g^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2 \end{pmatrix}\right). \tag{30}$$

From that, the following is found with Eq. (3):

$$v|H_A^{(1)} \sim N\left(0, \sigma^2 Q_{vv} + \frac{\sigma_g^2}{n^2}\begin{pmatrix} (n-1)^2 & 1-n & \dots & 1-n \\ 1-n & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1-n & 1 & \dots & 1 \end{pmatrix}\right).$$

The marginal distribution of v with respect to v_1 assumes the form

$$v_1|H_A^{(1)} \sim N\left(0, \frac{n-1}{n}\sigma^2 + \left(\frac{n-1}{n}\right)^2 \sigma_g^2\right)$$

and hence

$$T_{n,1}|H_A^{(1)} \sim N\left(0, 1 + \frac{(n-1)\sigma_g^2}{n\sigma^2}\right).$$

These distributions are again the same for all individual test statistics, thus independent of the index i of the outlier-suspected observation. Finally, the following is deduced from Eq. (27a)

$$\beta = \beta(c) \approx P\left(-c < T_{n,1} < c|H_A^{(1)}\right)^n = \left[2\Phi\left(c\left(1 + \frac{(n-1)\sigma_g^2}{n\sigma^2}\right)^{-1/2}\right) - 1\right]^n.$$

Fig. 5 shows some functions $\beta(c)$. Of course, the magnitude of the gross error again influences β. However, the impact is now smaller than before. Even quite sizable gross errors cannot be detected very reliably.

Example

For $n = 10$ observations for the 3σ-rule now $\beta = 0.74$ for $\sigma_g = \sigma$, $\beta = 0.021$ for $\sigma_g = 3\sigma$ as well as $\beta = 0.00046$ for $\sigma_g = 5\sigma$ in Fig. 5. This, in comparison to Fig. 4, is a disappointing result. Otherwise, qualitatively, the same points apply that were made for Fig. 4.

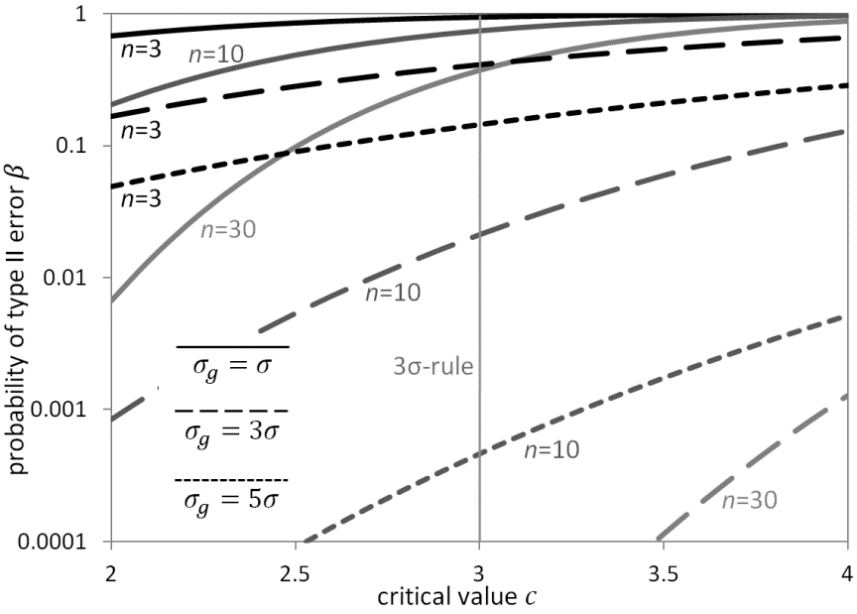

Fig. 5. Relationship between the probability of type II error β and the critical value c for n = 3, 10 und 30 repeated observations; solid curves: $\sigma_g = \sigma$; dashed curves: $\sigma_g = 3\sigma$; dotted curves: $\sigma_g = 5\sigma$

Conclusions

For gross measurement errors, which act randomly, thus at an imaginary repetition of the measurement process do not reproduce their magnitude and sign, the 3σ or similar rules seem to be working poorly. This has been shown in the example at the end of the last section. One explanation lies in the fact that the applied test statistic, Eq. (8), is not optimal for the alternative hypothesis, Eq. (30), but only for Eq. (28). This argument also applies to the other test statistics used, i.e., Eqs. (11), (15), and (20), and therefore, the same behavior is also expected there.

It is the author's opinion that the model of randomly acting gross errors is more relevant to geodetic practice [see Lehmann (2013) for a discussion]. Also, in this case, optimum test statistics in terms of the most powerful test would be able to be found. Unfortunately, they are much more complicated to calculate and numerically expensive. In the period from which the methodology for outlier detection originates, this was excluded because of the low processing power of available computers. If sticking to a simple rule for outlier detection such as the 3σ-rule is wanted, losses have to be accepted due to decision errors in outlier detection. Useful observations are lost, and, in fact, grossly erroneous observations remain undetected.

It can be definitely seen in Figs. 3–5 that an optimal critical value c cannot be determined irrespective of the number of observations n. In fact, c must be increasing with n. In current geodetic practice, adjustment problems with large amounts of observations are often found. The application of the 3σ-rule here means the loss of many useful observations.

Often, the situation is not as dramatic as it seems now, and here is why. Consider a properly formulated adjustment problem, Eqs. (1) and (2). The number of normalized residuals with a magnitude greater than 3 will, by Eqs. (9) and (10b), on average amount to $0.0027 \times n$, i.e., on average, approximately 0.27% of the useful observations is lost. This simple observation provides only a distorted picture because all observations are not immediately discarded at once, and the test is applied iteratively. After each test, only the one with the extreme normalized residual is discarded. A more accurate calculation is complicated.

It would be ideal if for each geodetic adjustment model, a realistic alternative hypothesis could be established and tested using an optimal test statistic, i.e., optimal in terms of the most powerful test or even better in terms of minimum premium and maximum protection. A very interesting possibility

is to assign a scale-contaminated normal distribution to the observation errors. This is basically a normal distribution but with a low probability contaminated by gross errors acting randomly (Lehmann 2012a). Critical values cannot be found in statistical tables. They cannot even be calculated using functions from standard statistical software libraries, and a computationally demanding Monte Carlo method is needed. The effort is nonetheless worthwhile because optimal test statistics can best distinguish between useful observations and outliers.

References

Abdi, H. (2007). "The Bonferroni and Sidák corrections for multiple comparisons." Encyclopedia of measurement and statistics, N. Salkind , ed., Sage, Thousand Oaks, CA.

Anscombe, F. J. (1960). "Rejection of outliers." Technometrics, 2(2), 123–147.

Baarda, W. (1968). "A testing procedure for use in geodetic networks." Publications on geodesy 9, 2(5), Netherlands Geodetic Commission, Delft, Netherlands.

Baselga, S. (2007). "Critical limitation in use of t test for gross error detection." J. Surv. Eng., 133(2), 52–55.

Baselga, S. (2011). "Nonexistence of rigorous tests for multiple outlier detection in least-squares adjustment." J. Surv. Eng., 137(3), 109–112.

Featherstone, W. E., and Morgan, L. (2007). "Validation of the AUSGeoid98 model in Western Australia using historic astrogeodetically observed deviations of the vertical." J. R. Soc. West. Aust., 90(3), 143–150.

Grubbs, F. E. (1969). "Procedures for detecting outlying observations in samples." Technometrics, 11(1), 1–21.

Hampel, F. R., Ronchetti, E. M., Roussseeuw, P. R., and Stahel, W. A. (1986). Robust statistics, Wiley, New York.

Hawkins, D. (1980). Identification of outliers, Chapman and Hall, London.

Heck, B. (1981). "Der Einfluß einzelner Beobachtungen auf das Ergebnis einer Ausgleichung und die Suche nach Ausreißern in den Beobachtungen." Allgemeine Vermessungsnachrichten, 01, 17–34 (in German).

Hekimoglu, S., Erdogan, B., and Erenoglu, R. C. (2012). "A new outlier detection method considering outliers as model errors." Exp. Tech., 10.1111/j.1747-1567.2012.00876.x.

Hekimoglu, S., and Koch, K. R. (2000). "How can reliability of the test for outliers be measured?" Allgemeine Vermessungsnachrichten, 7, 247–253.

Huber, P. J. (1981). Robust statistics, Wiley, New York.

Kargoll, B. (2012). On the theory and application of model misspecification tests in geodesy, Series C, Vol. 674, German Geodetic Commission, Munich, Germany.

Koch, K. R. (1999). Parameter estimation and hypothesis testing in linear models, Springer, Berlin.

Kutterer, H., Heinkelmann, R., and Tesmer, V. (2003). "Robust outlier detection in VLBI data analysis." Proc., 16th Working Meeting on European VLBI for Geodesy and Astrometry, W. Schwegmann and V. Thorandt, eds., Bundesamt für Kartographie und Geodäsie, Leipzig/ Frankfurt am Main, Germany, 247–255.

Lehmann, E. L., and Romano, J. P. (2005). Testing statistical hypotheses, 3rd Ed., Springer, New York.

Lehmann, R. (2010). "Normierte Verbesserungen - wie groß ist zu groß?" Allgemeine Vermessungs-nachrichten, 2, 53–61 (in German).

Lehmann, R. (2012a). “Geodätische Fehlerrechnung mit der skalenkontaminierten Normalverteilung.” Allgemeine Vermessungsnachrichten, 119(5), 143–149 (in German).

Lehmann, R. (2012b). “Improved critical values for extreme normalized and studentized residuals in Gauss-Markov models.” J. Geod., 86(12), 1137–1146.

Lehmann, R. (2013). “On the formulation of the alternative hypothesis for geodetic outlier detection.” J. Geod., 87(4), 373–386.

Lehmann, R., and Scheffler, T. (2011). “Monte Carlo based data snooping with application to a geodetic network.” J. Appl. Geod., 5(3–4), 123–134.

Microsoft Excel [Computer software]. Redmond, WA, Microsoft.

Neitzel, F. (2004). Identifizierung konsistenter Datengruppen am Beispiel der Kongruenzuntersuchung geodätischer Netze, Series C, Vol. 565, German Geodetic Commission, Munich, Germany.

Pope, A. J. (1976). “The statistics of residuals and the detection of outliers.” NOAA Technical Rep. NOS65 NGS1, U.S. Dept. of Commerce, National Geodetic Survey, Rockville, MD.

Rousseeuw, P. J., and Leroy, A. M. (1987). Robust regression and outlier detection, Wiley, New York.

Teunissen, P. J. G. (2000). “Testing theory; an introduction.” Series on mathematical geodesy and positioning, 2nd Ed., Delft Univ. of Technology, Delft, Netherlands.

van Loon, J. P. (2008). “Functional and stochastic modelling of satellite gravity data.” Publications on geodesy 67, Netherlands Geodetic Commission, Delft, Netherlands.

van Mierlo, J. (1983). “Problems of computing costs in decision problems.” Mathematical models of geodetic/photogrammetric point determination with regard to outliers and systematic errors, Series A, F. E. Ackermann, ed., Vol. 98, German Geodetic Commission, Munich, Germany.

Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing, Academic Press, Waltham, MA.

Xu, P. L. (2005). “Sign-constrained robust least squares, subjective breakdown point and the effect of weights of observations on robustness.” J. Geodesy, Berlin, 79(1–3), 146–159.