Jens Braband[1], Hendrik Schebe[2]

ASSESSMENT OF NATIONAL REFERENCE VALUES FOR RAILWAY SAFETY A STATISTICAL TREATMENT

We discuss the decision procedure used in the Commission Decision [1] for national reference values (NRV). According to the safety directive, every year seven safety indicators have to be computed for every member state. In the decision a fixed procedure has been presented for computing the safety indicators and to assess whether there is a possible deterioration in safety. In the safety assessment, the decision depends on a weighted sum in place of an arithmetic mean.

It is then of interest how such a decision procedure would behave and what would be the advantages and disadvantages of the particular method. In this paper, we study a slightly simplified version of the procedure by two means. First, we analyse the weighted sum and derives its characteristic as efficiency. Moreover, we compare it via a spread with an ordinary sample mean. We support the theoretical results with the help of a simple simulation study in order to estimate failure probabilities of the first and second kinds. In particular, we construct bad alternative distributions which the decision procedure cannot distinguish.

Keywords: *national reference values, railway safety indicator, robust estimators.*

1. Introduction

In the Commission Decision [1], a set of so-called NRV (national reference values) for each member state of the EU is estimated from accident data for several years. The accident data is measured in FWSIs (fatalities and weighted serious injuries) for different categories of persons: passengers, employees, level-crossing users, others, unauthorized persons, society as a whole. The choice of measures as the passenger FWSI is motivated by the fact that individual accidents have a relatively high frequency and low consequences compared with rare but potentially very high consequence accidents, such as high speed train collisions. However, for the purpose of this paper, we do not consider a particular NRV but concentrate on the generic approach and its properties. These NRV are then set as targets for the following years. In each year for each state and each NRV there the following hypothesis has to be assessed: "H_0: Is there evidence that the safety performance of the member state with respect to NRV has deteriorated?"

This hypothesis as such is well-known for problems of quality control. There, parameters as e.g. lengths, widths etc. are monitored. A sampling technique is applied to judge whether the process produces items that are out of specification. For this sake, a normal distribution with mean value μ and a spread σ is supposed

[1] Siemens AG, Брауншвейг, Германия
[2] TÜV Rheinland InterTraffic GmbH, Кёльн, Германия

to exist for the parameter to be monitored. Then, limits are computed and if a certain number of items are outside the limits, the process is assumed to be out of control. Details can be found in many books on statistical quality control, as e.g. [5]. In order to monitor the process, quality control charts are used, where the measurement values are registered together with a running number or the time, the measurement was carried out. Then, on the control charts the warning limits and the action limits are drawn according to the rules applied. Later on, these control charts have been refined and computerized.

So, the problem treated by the ERA (European Railway agency) in its directive [1] is similar to that of the well-known control charts.

However, the algorithm proposed by the ERA [1] is new and has never been analysed from a statistical point of view. This is subject of the present paper.

In section two we will describe the approach used by the ERA to monitor their indicators. In the third section we will provide a theoretical analysis of the main step of the decision procedure of the ERA. This will then be supported by a Monte-Carlo simulation of the rule in section four. The last section is dedicated to conclusions.

2. The Approach Used by the ERA

The ERA wishes to test the H_0:hypothesis:

‘The safety performance of the member state has not deteriorated with respect to the NRV”

by use of a procedure, which is in fact a complex statistical test consisting of several steps. We first give a generic definition of the problem we want to discuss in this paper:

Let the yearly values be denoted by X_i. Then, a particular moving weighted average (MWA) can be defined as a weighted sum:

$$\mathrm{MWA}_{q,p} = \frac{\sum_{i=p}^{q} w_{p,q}(i) X_i}{\sum_{i=p}^{q} w_{p,q}(i)}. \tag{1}$$

The weights are defined in the following manner:

$$W_{p,q}\,(i) = 1/\max(\varepsilon;|X_i - \overline{X}\ _{p,q}|) \tag{2}$$

$$\text{with } \overline{X}_{p,q} = \frac{1}{q-p+1} \sum_{i=p}^{q} X_i \tag{3}$$

as the arithmetic mean. The value ε has been set to a value of 0.01 in [1], probably based on a heuristic argument. Note that ERA calculates the MWA only based on the five most recent years, which we designate $\mathrm{MWA}_{n,'n-4}$ here. Then, we can express the multi-step decision rule which has in fact become law as a EU regulation in the following algorithm:

```
Algorithm
IF NRV≥$X_n$ OR NRV≥ $MWA_{n,n-4}$, THEN ACCEPT $H_0$ ELSE.
IF NRVx1.2 ≥ $MWA_{n,n-4}$, THEN ACCEPT $H_0$ ELSE.
IF “First deviation (in steps 1 or 2) in the last three years”, THEN ACCEPT $H_0$ELSE.
IF “Number of significant accidents not increased”, then ACCEPT $H_0$ELSE.
REJECT the hypothesis that safety performance has not deteriorated.
```

This algorithm has been generated directly from the description given in [1]. A justification for this algorithm is not provided in [1]. Therefore, we will give explanations on how the algorithm works from a statistical point of view. In this paper, we are particularly interested in the failure probabilities of the first and second kinds, namely that H_0 is rejected in case it holds true and that H_0 is accepted while being false. We denote the coinciding failures by α and β as the probability of error of first and second kind, respectively.

We start with some simple observations that we have made when analysing the algorithm consisting of steps 1-5:

Step 1 yields a more or less random result with a probability of acceptance in the order of 0.25 to 0.5 under reasonable assumptions; so, it is not an important factor and is discarded here.

In step 2, the constant tolerance band seems completely heuristic.

Step 3 could also be uninteresting as there will be dependencies in the data of subsequent years and it cannot be expected to contribute much.

In step 4, ERA explicitly calls for a statistical test to be applied and assumes a Poisson distribution for the number of significant accidents as the null hypothesis

On the basis of the last observation, we may conclude that ERA assumes a compound Poisson process as the null hypothesis with N_t depicting the number of accidents until time t and S_1, S_2,… an independent, identically distributed (i.i.d.) sequence of random variables describing the severity of the accidents as we have studied in [6]. The accidents occur with a rate of occurrence of λ(t) at time T_1, T_2, etc., where i is the index of the accident. For each accident with index i there exists a certain „jump height“ S_i, which is the severity. The severity is measured as the number of fatalities (or equivalent fatalities). So, for a fixed interval [0,t], the process X_t describes the cumulated number of fatalities and N_t describes the accumulated number of accidents.

This compound Poisson process can be used to derive characteristics for collective risk.

So, under the null hypothesis, the observables X_t is a random variables that can be assumed to be distributed as

$$\sum_{i=1}^{N_t} S_i \, . \tag{4}$$

With a constant accident rate λ(t) = λ.

Thus

$$E(X_t) = \lambda t E(S_1) \text{ and } V(X_t) = \lambda t E(S_1^2) \tag{5}$$

holds, where we assume that the moments of S_1 exist. Given a sufficient number of accidents per year, limit theorems for the distribution of X_t may apply, e.g. central limit theorem.

In the following section, we will analyze the estimator (1) from a theoretical point of view and afterwards explore the complete decision process by statistical simulation.

3. Theoretical Analysis of Step 2 of the algorithm

When looking at (1) from a theoretical point of view we see that

a) the number of observations in the sample is small, since only 5 values are used,

b) the mean (1) is a weighted mean, where the weight (2) itself is also computed from the sample,

c) the weight (2) is truncated from below, ensuring that it is bounded away from zero, where the choice of the truncation constant ε is somewhat arbitrary,

d) the weight (2) reduces the influence of large observations since it is proportionally to the inverse of the difference of the observation and the sample mean.

The form of the estimator (1) is well-known in robust statistics. It can be found in [2], Section 2.3d, as the so-called Tukey's W-estimator, as introduced by Tukey [3] with weight function (2). So, the results from [2] can be directly used for the specific form (1) of this W-estimator. It has been shown in [2] that the statistical behavior of (1) is the same as that of Huber's M-estimator with function ψ(u)=u w(u). It is to be recalled that Huber's M-estimator is defined as an estimator T_n satisfying the equation

$$\sum_{i=1}^{n} \psi\,(X_i; T_n) = 0\,, \tag{6}$$

See, for example, [2]. Now, the theory of robust estimators can simply be applied to (1) to see what its properties are. The two main properties are robustness and efficiency.

a) Robustness

An estimator is called robust if it is not susceptible to the presence of outliers in the sample. Outliers are elements of the sample that do not belong to the normal sample space which consists of identically distributed random variables.

Two main characteristics for the robustness of an estimator can be applied. The first is the so-called breakdown point. This is the fraction of outliers in the sample up to which the estimator still yields a consistent result. That means, although a certain fraction of data in the sample does not belong to the population, parameters of the population are estimated correctly. After some algebra, the breakdown point can be found to be at 50%. This holds for any sample size. Any observation, if it is large enough, is divided by its absolute difference from the sample mean. The sample mean is only influenced if more than half of the observations X_i are outliers.

The second important characteristic is the influence function, introduced by Hampel. It is defined for a statistic value T constructed on independent, identically distributed random variables with a common distribution function F and for an outlier y as

$$IF\,(y;T;F) = \lim_{t\to 0} \frac{T(\,1-t)F + t\Delta_y) - T(F)}{t}, \tag{7}$$

See [2, Section 2.1b]. By Δ_y, we denote a probability distribution putting mass 1 to point y. The influence function shows how much a single observation can influence the estimator itself. It can easily be computed by letting one observation tend to infinity and seeing how the estimator changes. If y is the value of the outlier and m denotes the sample mean of the values X_i, it can be seen that the influence function of the arithmetic mean (3) is

$$IF(y) = \lim_{t\to 0} \frac{(1-t)m + ty - m}{t} = y - m. \tag{8}$$

Usually, a robust estimator is characterized by a limited influence function. It may be noted that the influence function of the normal sample mean is not limited and increases together with the value of the outlier y. That is why the sample mean is not a robust estimator of the population mean.

The estimator (1) can be expressed in the form

$$T(F) = \frac{\int w(x) x dF(x)}{\int w(x) dF(x)},$$

with

$$w(x) = \frac{1}{\left| x - \int x dF(x) \right|}.$$

Here, and furthermore, the integrals are over the entire real axis.

For the estimator (1), we derive the influence function as

$$IF(y) = y \frac{\int w dF(x) \int \beta^2 x \,|\, x - y \,|\, dF(x) - \int w x dF(x) \int \beta^2 \,|\, x - y \,|\, dF}{\beta^2} + O(1), \tag{9}$$

with m being the mean of F, w being the weight function and β = |x-m|. Furthermore, O(1) denotes a term that is of order 1, compared with y when y tends to infinity.

The influence function (9) is not limited if y increases. Therefore, one single observation X_i can influence the estimator of the mean such that if a single observation would tend to infinity, the estimator of the mean would do the same.

The conclusion from this short study is that the estimator (1) is not robust.

b) Efficiency

In order to obtain the variance of estimator (1), simply the results of [2] are used for location-type estimators in Section 2.2b.The variance is given by

$$Var(\psi, F) = \frac{\int \psi^2 dF}{\left(\int \psi' dF\right)^2}. \tag{10}$$

Now, remembering ψ(u)= u w(u) with w(u) = max(εm;|u-m|), we see that

$$\psi'(u) = w(u) + u\, w'(u) = \begin{cases} \max(\varepsilon m, |\, u - m \,|) + u \text{ for } \varepsilon m < |\, u - m \,| \\ \max(\varepsilon m, |\, u - m \,|) \text{ otherwise} \end{cases} \tag{11}$$

If we neglect the influence of ε which is small anyway, and simply set it to zero, we obtain

$$Var(\psi,F)=\frac{\int x^2(x-m)^2\,dF}{\left(\int(u+|u-m|)dF\right)^2}. \tag{12}$$

We can neglect ε since it seems to be introduced to avoid division by zero and make the computation feasible.

Note that this value has to be divided by n to obtain the variance of an estimator constructed from a sample of size n. This quantity has to be compared with the variance of the simple arithmetic mean, viz.

$$Var(F)=\int(x-m)^2\,dF=\sigma^2. \tag{13}$$

Equation (12) yields

$$Var(\psi,F)=\frac{m_4-2m_3m+m^4}{\left(m+\int|x-m|dF\right)^2}, \tag{14}$$

where m_4 and m_3 are the ordinary fourth and third moments of the distribution. Since the values X are sums of other random variables, a normal distribution can be assumed as a good approximation. Remember, that we have assumed the existence of the moments of the S values so that weak convergence to the normal distribution is guaranteed. The approximation, however, would only be a good one, if the number of events is large enough. So for small numbers of events, it will be just a rough approximation. Now, for a normal distribution, (14) yields

$$Var(\psi;F)=(2m^4+6m^2\sigma^2+3\sigma^4)/(m+2\sigma)^2. \tag{15}$$

Dividing (15) by (13), we arrive at

$$\mathrm{Eff}=(2+6v^2+3v^4)/(v^2+4v^3+4v^4), \tag{16}$$

with v = σ/m as the coefficient of variation. Eff is the relative efficiency of the robust estimator, related to the normal sample mean. It may be noted that it tends to infinity as v tends to zero, i.e. the estimator (1) has a much larger spread than the ordinary sample mean. Even for v=1, it may be observed that Eff = 11/7, i.e. the variance of the robust estimator is about 57% larger than that of the ordinary estimator.

So, the larger the number of observations in a single year, the smaller the value of v and thus the larger the loss of efficiency of the estimator (1).

Notes:

1. We have neglected the fact that estimator (1) is not an iterated estimator, i.e. the average is computed once for estimator (1) and computation of the estimator is not repeated with the new mean, after (1) has been computed for improving the estimate. However, the failure introduced is of order 1/n, where n is the sample size.

2. When computing the efficiency, we have replaced the sample mean by the population mean in the weight function, thus introducing an error of order 1/n.

3. Integrals without indication of the limits are over the entire domain of definition of the distribution function F.

As a conclusion of our theoretical study, we see that estimator (1) has lost efficiency, applies smaller weights to larger observations but is not robust. What does that mean practically? A sudden change in one or several years of the statistics would not influence the statistics in the same manner as for an unweighted estimator, e.g. the sample mean. In addition, the spread of the estimator is much larger than that of the ordinary sample mean. That means we would observe much more "random noise" in our MWA, i.e. there might be unnecessary random fluctuations in the MWA compared with the use of the sample mean.

4. Simulation Results

A number of simulations with different models of distribution functions have been carried out to support the theoretical results of Section 2.

We simulate i.i.d. annual observations from a given distribution with the mean NRV to test the null hypothesis. We start with a normal distribution and vary the standard deviation σ. From some experiments, we have gained the impression that, for such a well behaved and symmetric distribution, not a great difference between the decisions based on the normal average (3) and the weighted moving average (1) was visible. So, as an upper limit for α, we can postulate the normal approximation

$$P(N > \frac{NRV}{\sigma\sqrt{5}}), \tag{17}$$

where N is a standard normal random variable. This approximation is derived from the fact that we average 5 values of NRV's according to (1) and assume a normal approximation. Our simulations confirm that in fact the distribution of the MWA is centered more tightly around the NRV, but the error is surprisingly small. This is only logic, since the MWA is built from several values of the NRV.

For a second experiment, we take the gamma distribution as a non-symmetric solution and compare the results of the skewness of the distributions for equal σ.

The results in table 1 show the great dependency of the failure probabilities for step 2 on σ. For small σ, α is virtually zero, but, for changes in the mean of less than 20%, β is close to 1, while, for larger σ, α is increasing and β is decreasing. Note that, the mean has been chosen to be one, so that σ is also the coefficient of variation.

Table 1 Results of simulation for first kind error probability α

Experiment	σ=20%	σ=30%	σ=50%
Normal approximation	0.013	0.068	0.19
Normal MWA	0.01	0.074	0.18
Normal steps 1 to 3	0.0043	0.032	0.08
Gamma MWA	0.029	0.084	0.15
Gamma steps 1 to 3	0.011	0.032	0.06

The range of σ values shown in table 1 can be regarded as quite representative, as a detailed evaluation of Swiss accident statistics has demonstrated. Unfortunately the detailed results are confidential. The results also show the adequacy of the normal approximation, in particular for larger values of σ. They also show that steps 1 and 3 when combined only contribute by a factor of between 2 and 3, indicating that there is a high degree of dependency. Note that we have not simulated step 4 as, for this case, α has been fixed by ERA to 0.05, but also with a high degree of dependency. Also, we would have had to make many more assumptions on the compound Poisson process.

Also other qualitative simulations have been performed to find some “bad” alternative hypothesis that the decision procedure can hardly distinguish. In these experiments the i.i.d. observations have been distorted by fixed or randomised linear trends or by periodically recurring outliers, representing large scale accidents. The general observation is that the decision procedure is quite slow in detecting a linear trend, it generally takes 4 to 5 years until the decision procedure rejects the null hypothesis, after the trend has become visible in the data. This effect can be explained by a combined effect from the suppression of outliers by the moving average in step 2 and the “one off” exception in step 3. As expected also outliers are suppressed quite well by the decision procedure, e. g. for low σ, say 10% of the NRV, large outliers are tolerated by the decision procedure occurring as often as every third year, which in practice could not be tolerated.

5. Conclusions

In this paper, we have discussed, both by theoretical argument and by Monte Carlo simulation, the decision procedure as introduced in [1]. It has turned out that the weighted sum (MWA) used in the procedure has more spread than the ordinary sample mean. . Note that, the maximum likelihood estimator is always asymptotically efficient, provided the regularity conditions are fulfilled. In our case, the maximum likelihood estimator is the sample mean and this is the efficient estimator. The estimator applies smaller weights to large observations, but is not robust. The MWA gives a good average value but loses efficiency compared with the sample mean. It has also turned out that, while the error probability of the first kind is quite small for typical parameters, the error probability of the second kind can be very high, in particular as large values would be discounted by small weights and therefore, trends in alternative hypotheses would not be detected early.

It is also uncommon for statistical decision procedures that the sample variation is not taken into account in the procedure and that no target value for the error probability of the first kind has been fixed in the design of the procedure.

If an estimator with limited robustness were needed, the so-called α-trimmed mean could be used. This estimator neglects the γ/*100% smallest and γ/*100% largest observations in a sample and computes the mean only with regard to the remaining elements of the sample. Such an estimator is much simpler and its breakdown point (2γ) could be changed, depending on the trimming. Then, no weighting would need to be applied and the loss of efficiency would be much smaller than with the present estimator.

Acknowledegement

Thre authors are grateful to two anonymous referees that helped to improve the paper.

References

1. Commission Decision of 5 June 2009 on the adaption of a common safety method for assessment of achievement of safety targets, as referred to in Article 6 of Directive 2004/49/EC of the European Parliament and the Council; 2009/460/EC.

2. **F.R. Hampel, E.M. Ronchetti, P.J. Rosseeuw, W.A.Stahel, Robust Statistics** – The Approach based on Influence Functions,. Wiley 1986.

3. **J.W. Tukey**, Exploratory Data Analysis, Addison Wesley 1977.

4. Directive 2004/49/EC of the European Parliament and of the Council of 29 April 2004 on safety on the Community’s railways.

5. A**.J. Duncan, Quality Control and Industrial Statistics,** Homewood / Illinois 1965.

6. **J. Braband, H. Schäbe,** The collective risk, the individual risk and their dependence on exposition time, ESREL 2011, Proceedings , Advances in Safety, Reliability and Risk Management, pp. 1783-1787).