Fresenius Z. Anal. Chem. 292, 34–38 (1978)

Fresenius Zeitschrift für Analytische Chemie

Application of Statistical Methods to Activation Analytical Results Near the Limit of Detection

Arsenic in Human Skin Biopsies*

K. Heydorn** and B. Wanscher***

Isotope Division, Risø National Laboratory, DK-4000 Roskilde, Denmark

Anwendung statistischer Methoden auf aktivierungsanalytische Ergebnisse nahe der Nachweisgrenze: Arsen in menschlicher Haut

Zusammenfassung. Die Angabe tatsächlicher Zahlen anstelle von oberen Grenzen für analytische Ergebnisse bei oder unter der Nachweisgrenze kann zuverlässige Werte ergeben, wenn diese Zahlen in geeigneter Weise statistisch aufbereitet werden. Besonders bei radiometrischen Verfahren, wie z.B. die Aktivierungsanalyse, wo individuelle Standardabweichungen bestimmt werden können, kann eine verbesserte Unterscheidung aufgrund der Präzisionsanalyse erreicht werden. Dieses Prinzip wird am Beispiel der Arsenbestimmung in menschlicher Haut demonstriert.

Summary. Reporting actual numbers instead of upper limits for analytical results at or below the detection limit may produce reliable data when these numbers are subjected to appropriate statistical processing. Particularly in radiometric methods, such as activation analysis, where individual standard deviations of analytical results may be estimated, improved discrimination may be based on the Analysis of Precision. Actual experimental results from a study of the concentrations of arsenic in human skin demonstrate the power of this principle.

Key words: Best. von Arsen in Haut; Aktivierungsanalyse, Neutronen; statist. Methoden an der Nachweisgrenze

The presentation of analytical results close to the limit of detection was discussed in some detail by Kaiser [5] from a predominantly spectroscopical viewpoint, but his recommendations have been widely accepted in many other analytical disciplines. Thus, analytical signals that are less than 3 times the analytical noise are reported as *not detected*, or alternatively presented as results *less than* a limit of guarantee for purity.

This method of presentation eliminates the risk of giving credence to individual numbers, subject only to random variation; thereby it helps to prevent unjustified interpretation of analytical data. On the other hand, the meaningful interpretation of groups of data by systematic application of statistical methods is only possible when actual numbers are available.

Analytical variation is usually assumed to follow a normal, Gaussian distribution, but in general this is only true for results exceeding the detection limit by at least 2 orders of magnitude [8]; the actual distribution of analytical results in the vicinity of the detection limit is often very skew. The use of 3 standard deviations between the detection limit and the limit of guaranteed purity is justified, even if the distribution is not exactly normal.

The use of classical statistical methods based on the assumption of normality, however, is not justified in the general case of analytical data near the detection limit. Interpretation of such data should be based on distribution-free statistical tests, although in exceptional cases normality may be assumed.

Activation analysis represents such an exception, and both types of statistical test might therefore be applied. In addition, the underlying Poisson statistic permits the calculation of the precision of each individual result, and together with the *a priori* precision this forms the basis for the analysis of precision of duplicate results [4]. For results near the detection limit, the *a priori* precision is not significant, and when the

* Presented in part at the 6th Nordic Conference on Trace Elements, Södertälje, August 1977

** Address for correspondence

*** Department of Dermatology, Finsen Institute, Copenhagen

0016-1152/78/0292/0034/$ 01.00

analytical method is in statistical control [3] the variability of results is fully accounted for by counting statistics.

In the present investigation of arsenic in human skin almost 90% of the samples analysed earned the predicate: *not detected* or *less than* 17 μg/kg. The remaining samples averaged 10 μg/kg, but not even a grand mean value can be calculated from this information.

The processing of individual, analytical data not only produces a significant mean value, but also permits probing of the structure of the data material by appropriate statistical tests and by the analysis of precision of the associated standard deviations.

Experimental Results

In co-operation with the Finsen Memorial Institute a study was carried out to establish whether concentrations of arsenic in human skin correlated with the presence of psoriasis. Skin biopsies were performed on 16 psoriatic patients for comparison with 16 healthy persons, and samples with a thickness of 1 – 3 mm skin depth yielded an average mass of approx. 20 mg fresh weight.

Neutron activation analysis with radiochemical separation [2] was the only method available for the determination of As in these small samples. As this is an expensive and time-consuming analytical method, the number of samples analysed was kept at a minimum by careful experimental design.

Half of the group of psoriatic patients was sampled in duplicate, representing the same skin areas left and right; one was clearly psoriatic skin, the other apparently normal.

No particular interest was connected with the variability between the 16 normal persons, and samples were therefore combined into 5 pools so that the analytical effort was reduced and the precision possibly improved because of the larger amount analysed.

Individual values for arsenic in all samples are presented in Table 1 in such a way as to facilitate the extraction of maximum information by appropriate statistical and mathematical methods.

Calculation of the variation within any one group in the data yields relative standard deviations exceeding 40%; an assumption of normality is therefore only justified after appropriate verification.

Table 1. Measured values for arsenic in human skin biopsies

Psoriatic and/or un-psoriatic skin samples			Normal skin samples	
Duplicates		Single	Pooled	
μg/kg	μg/kg	μg/kg	number	μg/kg
1.5	8.1	9.9	2	5.7
5.1	4.8	6.0		
5.2	10.6	3.1	4	4.0
6.5	3.4	3.1		
2.8	0.6	9.5	5	2.8
5.6	5.2	2.7		
4.9	3.0	7.8	3	2.1
8.5	10.9	3.9	2	2.4

Table 2. Estimated standard deviations for arsenic in human skin biopsies

Psoriatic and/or un-psoriatic skin samples			Normal skin samples	
Duplicates		Single	Pooled	
μg/kg	μg/kg	μg/kg	number	μg/kg
2.7	3.3	2.4	2	1.0
2.2	2.3	3.0		
1.9	3.1	2.4	4	0.7
1.9	2.4	1.9		
1.8	2.2	2.1	5	0.5
2.3	1.8	3.6		
1.5	2.0	2.8	3	1.1
2.8	4.0	3.1	2	1.2

Statistical Considerations

The fundamental question to answer by means of statistical methods is whether all – or only some – groups of data are likely to be members of the same population, regardless of their particular distinction. If some a priori information is available on this population, such as its distribution (Gaussian, Poisson, Bernoulli, etc.) or location, more precise answers can be given for a particular set of data.

In the present study we tested the data under three different assumptions concerning their distribution:

(a) Unknown, but common distribution;

(b) Normal distribution with unknown, but common variance;

(c) Normal distribution with individual variance estimates.

For the first two cases, appropriate statistical techniques were selected from statistical textbooks [7] without regard to the problems associated with the detection limit.

In the last case the values presented in Table 1 are weighted with their estimated standard deviations presented in Table 2, and alternative statistical methods must be applied.

The hypotheses tested were the following:

(i) Psoriatic and non-psoriatic skin do not differ with respect to arsenic concentration;

(ii) All psoriatic patients have the same average concentration of arsenic;

(iii) Normal and psoriatic patients do not differ with respect to arsenic concentration.

These hypotheses are accepted or rejected according to the decision rules shown in Table 3.

Non-Parametric Tests of the Measured Values

For small samples, the efficiency of distribution-free tests is often comparable with – and may even exceed – tests based on the normal distribution. Their general

Table 3. Decision rules

Probability of no effect*	Symbol	Conclusion
$P > 0.05$	n.s.	Effect disregarded
$0.01 < P < 0.05$	*	More information required
$P < 0.01$	**	Effect established

* Two-sided levels of significance

applicability and ease of calculation make them an obvious starting-point for statistical inference [1].

(i) The simplest test for the identity of paired observations is the *sign test*, which requires pairs to be independent, but they need not belong to the same underlying distribution.

The outcome is binomially distributed with a probability of 0.5, and in the present case of $n = 8$, an outcome of 1 or 0 of either sign has a probability of $< 1\,\%$. It would therefore indicate a significant difference between psoriatic and non-psoriatic skin from the same patient. The actual outcome of 3 to 5 either way is the most probable outcome ($P = 0.43$) under the assumption of no difference, and the first hypothesis is therefore accepted.

(ii) With no perfect, non-parametric counterpart to the F-test, the individuality of the patients cannot be tested directly.

If individual differences are greater than differences between samples from the same patient, duplicate results from two different persons would not match as well as genuine duplicates. The resulting increased variability of differences between duplicates can be tested by *Mood's test* for dispersion differences.

This situation can be created by subtracting the results in the second column in Table 1 from the results in both the first and the third column. By ranking the 16 differences, a difference of dispersion is tested by the statistic W, calculated for the two columns

$$W = \sum_{1}^{8} \left(r_i - \frac{17}{2} \right)^2 . \qquad (1)$$

With an expectation value of 170 and an approximately normal distribution and a standard deviation of 39, the actual outcome of $W = 120$ or 220 for the two columns is not significant. The second hypothesis was accepted – that no individual arsenic concentrations could be detected among the psoriatic patients.

Undoubtedly, the same is true for normal persons, and the combination of several samples into one before the analysis does not reduce the generality of the investigation.

(iii) The most effective and simple non-parametric test for location difference is the *U-test*, introduced by Wilcoxon and tabulated by Mann and Whitney. Its asymptotic relative efficiency exceeds 95 %, and it is often used instead of the equivalent t-test, even for normal distributions.

In the present case of 24 indistinguishable values for psoriatic patients compared with 5 for normal persons, an outcome of $U = 32$ is not significant ($P > 0.05$).

The last hypothesis that no significant difference could be found between normal and psoriatic patients was therefore accepted. The distribution-free tests were thus unable to detect any significant structure in the experimental data.

With all 3 hypotheses accepted, all 29 results belong to the same population and should be pooled to represent the distribution of arsenic concentrations in human skin.

Tests of the Measured Values Based on the Normal Distribution

The assumption of a normal distribution of the 29 results in Table 1 was confirmed by a Range Test, as well as by simple distribution tests. The use of the classical, parametric tests might therefore be justified, although their proper sphere of application is orders of magnitude above the detection limit.

Under the assumption of homoscedasticity, the conventional statistical tests have the highest efficiency, and at the same levels of significance as in Table 3 the risk of committing errors of the second kind is reduced when testing the specified hypotheses.

(i) was tested by *Student's t-test* for paired measurements, and the hypothesis was accepted with a value of $t = 0.64$ at 7 degrees of freedom.

(ii) was tested by *Snedecor's F-test* for variance ratios, and the hypothesis was accepted with a value of $F = 1.05$ with 15 over 8 degrees of freedom.

(iii) was tested by the *analysis of variance* between the normal and the patient groups, and the hypothesis was accepted with a value of $F = 2.50$ with 1 over 27 degrees of freedom.

However, the variances of the two groups are probably different, which means that we have a Fisher-Behrens problem with no exact solution. In practice, *Welch's test* is useful [7], and for the present case a value of $\hat{t} = 2.40$ at 15 degrees of freedom is calculated. This has a probability of between 5 % and 1 %, which means that additional information is required before a decision can be made in accordance with Table 3.

Analysis of Precision of the Weighted Values

Additional information is available in the form of estimated standard deviations, $\hat{\sigma}_i$, in Table 2 for each individual measurement, y_i.

When measurements are based on the counting of radioactive samples, their distribution is determined by the Poisson statistic governing the process of radioactive decay.

The Poisson distribution is characterized by one parameter only (λ), in contrast to, e.g., the normal distribution requiring two (μ; σ^2). Both mean value and variance are equal to λ, which means that the standard deviation due to counting statistics can be calculated for a single measurement. In the vicinity of the detection limit, other sources of variability are usually negligible, and the precision of a result is determined by counting statistics only.

The three hypotheses may now be tested by the *analysis of precision*, which compares the observed and the estimated variability of analytical data by means of the statistic T, which is closely approximated by a chi-squared distribution [4].

(i) is tested by the same type of statistic as used to ascertain that the method is in statistical control [3]

$$T = \sum_{1}^{8} \frac{(y_{1i} - y_{2i})^2}{\hat{\sigma}_{1i}^2 + \hat{\sigma}_{2i}^2}. \quad (2)$$

In the present case, a value of $T = 7.1$ with 8 degrees of freedom shows excellent agreement between the calculated variance estimates for normally distributed results and the actual variability of results. The hypothesis is thus confirmed that the difference between duplicates is only caused by analytical uncertainty, and duplicates can be replaced by their *weighted mean* without loss of information content.

(ii) is tested by a slightly different expression [4]

$$T = \sum_{1}^{16} \frac{(y_i - \bar{y})^2}{\hat{\sigma}_i^2} \quad (3)$$

with $\bar{y}$ the weighted mean of the 16 results, and T closely approximated by a chi-squared distribution with 15 degrees of freedom.

With $T = 22.1$, individual differences between patients were not found to cause any significantly increased variability of results, and the same was found for the control group. The hypothesis was therefore accepted in agreement with the decision rules in Table 3.

(iii) is tested by the same statistic as (2) applied to the combined group of patients and controls, each representing 16 individuals. A systematic difference between the two groups gives rise to additional variability in the combined group of 21 results.

With $T = 41.3$ at 20 degrees of freedom, a significant difference was established ($P < 0.5\%$), and the hypothesis was rejected in accordance with the decision rules.

The final conclusion is that patients and controls belong to different groups, but no individual differences were detected. The weighted mean values of the two groups are

Patients 5.1 ± 0.5 µg/kg
Normal 3.3 ± 0.3 µg/kg

Table 4. Statistical testing of analytical results for As near the detection limit

Assumed type of distribution	Hypothesis tests		
	(i)	(ii)	(iii)
(a) Unknown distribution	Sign-test	Mood's test	U-test
	n.s.	n.s.	n.s.
(b) Normal distributions	t-test	F-test	ANOVA
homoscedasticity	n.s.	n.s.	n.s.
(c) Normal distributions	Analysis of precision [4]		
individual variance estimates	n.s.	n.s.	**(*)

Discussion

The preceding investigations are summarized in Table 4, from which it is seen that a significant structure could only be established by the *analysis of precision* of the weighted data. The systematic difference between patients and controls is no artifact, because all members of the patient group have at one time or another taken the arsenical drugs previously used in the treatment of psoriasis.

The precision of the mean values is quite satisfactory with relative standard deviation of 10%, and their accuracy is substantiated by the agreement with results for tissue samples weighing 50 times as much [6].

The distribution of the measured values is not markedly different from a normal distribution, and without the additional information in Table 2 the use of traditional statistical tests based on homoscedasticity would seem quite natural. As it is, however, the estimated standard deviations vary by almost an order of magnitude, and the conditions for carrying out these tests are not present.

The conditions for carrying out the analysis of precision are automatically verified when hypothesis (i) is accepted by the T-test; this confirms that the analytical method is in statistical control.

Conditions for performing distribution-free tests are also fulfilled, and in cases where individual variance estimates are not available, the non-parametric methods should be chosen in spite of their slightly lower asymptotic, relative efficiency [1].

The superior efficiency of the analysis of precision in the present case is associated with a better utilization of results from the controls, which were pooled before the analysis to yield a larger sample with better

precision. Nevertheless, differences in the weight of the single samples also influence their precision, and conventional statistical tests do not take advantage of this additional information.

Conclusion

Reporting measured values instead of upper limits for results near or below the detection limit makes possible a statistical testing of the data. In the present case of arsenic in human skin, mean values for two different groups were determined with standard deviations of 10%, even though the majority of measured values were below the detection limit.

The use of distribution-free tests instead of tests based on a normal distribution did not reduce the efficiency noticeably, and the estimation of individual standard deviations of the measured values showed that the assumption of homoscedasticity was not justified.

The increased information content from the individual variance estimates was utilized by the analysis of precision, whereby a significant structure in the data was established that could not be detected by conventional testing. This method of data analysis takes advantage of individual results with improved precision, achieved by the pooling of samples *before analysis.*

Analytical results near the detection limit are not likely to exhibit variance homogeneity, regardless of the analytical method involved. Distribution-free statistical methods should therefore be chosen for testing of measured values.

Analytical results for which individual standard deviations are available should be tested by the analysis of precision, which represents a much more efficient utilization of the information content of the data.

Acknowledgement. The support of the Daell Foundation is gratefully acknowledged. T. Campbell participated in part of the analytical work.

References

1. Gibbons, J., Dickinson: Nonparametric statistical inference. New York: McGraw-Hill 1971
2. Heydorn, K.: Clin. Chim. Acta **28**, 349 (1970)
3. Heydorn, K.: NBS Special Publication **422**, 127 (1976)
4. Heydorn, K., Nørgaard, K.: Talanta **20**, 835 (1973)
5. Kaiser, H.: Fresenius Z. Anal. Chem. **209**, 1 (1965)
6. Larsen, N. A., Nielsen, B., Pakkenberg, H., Christoffersen, P., Damsgaard, E., Heydorn, K.: Nucl. Act. Techniques Life Sciences, p. 561. Wien: IAEA 1972
7. Sachs, L.: Angewandte Statistik, 4. Aufl. Berlin-Heidelberg-New York: Springer 1974
8. Thompson, M., Howarth, R. J.: Analyst **101**, 690 (1976)

Received December 20, 1977