ESTIMATION OF BIAS USING EQA MATERIALS.

17-10-30 Anders Kallner. Dept Clin Chem Karolinska university hospital, Stockholm Sweden.
anders.kallner@ki.se.

Background

External Quality Assessment Schemes (EQAS) distribute samples which are as commutable as possible with patient material, collect results from participating laboratories and report the results in peer groups that use similar or identical measuring systems. EQA organisers may have an excess of previously distributed and measured materials. These have an assigned value for each peer group with information on the number of users and the distribution of values. These could be used for bias estimation and compensation in clinical laboratories. Since the concentrations of the test material will vary the differences will have to be normalized. A most convenient method is to calculate the z-score, i.e. expressing the difference in relation to the uncertainty (SD or SEM). This score can be used as a basis for a comparison between samples of different concentrations. The probability that the average of the *z*-score is not zero can be determined using Student's *t*-test and thus an assessment of whether there is a significant difference between the laboratory and the peer group can be established. A number of other pieces of information are calculated and displayed in graphs.

The idea of this technique is to take advantage of established statistical methods and materials measured by a large peer group to harmonize results of measurements. This would provide a potency of being superior to conventional EQA and offer a basis for recalibration. Since EQA schemes generally require only single measurements of the samples, the reported results include not only the bias of the measurement, but also the imprecision of the submitting laboratories' measurements. The magnitude of the imprecision or the size of the uncertainty contribution from the submitting laboratories is largely unknown and therefore, the results of the peer group may not correctly represent the true value.

Experimental design and software

Throughout the spreadsheet all cells are protected except those with blue or violet borders which are "input variable". All calculated results are in the red-bordered cells.

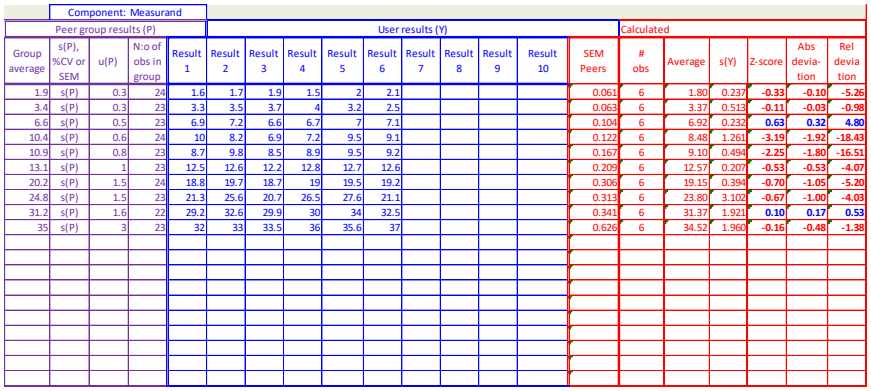
The laboratory should measure the samples at least in duplicates to improve the precision of the results. The software allows for up to ten replicates of laboratory measurements. To improve the power of estimating the within and between run imprecision it is suggested aiming at five replicates and five "peer" samples.

Results from the peer groups average and the associated uncertainty can be expressed either as a standard deviation, or coefficient of variation (*CV*) or standard error of the mean (SEM). Since the calculations will be based on the SEM, the number of participating peers should also be added if the SEM is not available. Up to 20 peer samples can be accommodated.

No calculations will take place unless the component is identified (Figure 1). Any entered data can be removed, exchanged or deleted.

The input cells should be sorted in ascending order of the group averages before analysis of the graphs. All calculations will be correct even if they have not been sorted but the graphs will then be more difficult to interpret. Sorting can be undertake at any time, make sure that the entire input data field is included in the sorting.

The uncertainty of the peer group's value is not expressed as *SEM* or confidence interval. The *z*-score represents the difference between the results and Student's *t*-test is used to assess whether the *z*-score is significantly different from zero.



	Component: Measurand																				
Peer group results (P)				User results (Y)										Calculated							
Group average	s(P), %CV or SEM	u(P)	N:o of obs in group	Result 1	Result 2	Result 3	Result 4	Result 5	Result 6	Result 7	Result 8	Result 9	Result 10	SEM Peers	# obs	Average	s(Y)	Z-score	Abs deviation	Rel deviation
1.9	s(P)	0.3	24	1.6	1.7	1.9	1.5	2	2.1					0.061	6	1.80	0.237	-0.33	-0.10	-5.26
3.4	s(P)	0.3	23	3.3	3.5	3.7	4	3.2	2.5					0.063	6	3.37	0.513	-0.11	-0.03	-0.98
6.6	s(P)	0.5	23	6.9	7.2	6.6	6.7	7	7.1					0.104	6	6.92	0.232	0.63	0.32	4.80
10.4	s(P)	0.6	24	10	8.2	6.9	7.2	9.5	9.1					0.122	6	8.48	1.261	-3.19	-1.92	-18.43
10.9	s(P)	0.8	23	8.7	9.8	8.5	8.9	9.5	9.2					0.167	6	9.10	0.494	-2.25	-1.80	-16.51
13.1	s(P)	1	23	12.5	12.6	12.2	12.8	12.7	12.6					0.209	6	12.57	0.207	-0.53	-0.53	-4.07
20.2	s(P)	1.5	24	18.8	19.7	18.7	19	19.5	19.2					0.306	6	19.15	0.394	-0.70	-1.05	-5.20
24.8	s(P)	1.5	23	21.3	25.6	20.7	26.5	27.6	21.1					0.313	6	23.80	3.102	-0.67	-1.00	-4.03
31.2	s(P)	1.6	22	29.2	32.6	29.9	30	34	32.5					0.341	6	31.37	1.921	0.10	0.17	0.53
35	s(P)	3	23	32	33	33.5	36	35.6	37					0.626	6	34.52	1.960	-0.16	-0.48	-1.38

Figure 1. Data input and initial calculations

If only one result from the laboratory is entered, nothing will be calculated. Similarly, if the average or the uncertainty of the peer group is not entered, calculation will not be performed. The uncertainty can be entered as a standard deviation, a relative standard deviation (%CV) or SEM. The uncertainty must be defined in column D unless the SEM is entered. The number of peers is also necessary to calculate the pooled uncertainty of the peer group.

Differences are calculated as the observed value minus the value of the peers. The results are colour-coded in columns U to W to identify negative differences (red) and positive (blue). If the difference is zero the numbers will be shown in green.

Calculations

The comparison procedure determines if the bias is different from zero and also if there is a significant difference between the observations made by the peer group and users. The significance is appraised using Student's dependent *t*, assuming a two-sided test. The false rejection rate can be changed but is set to 5 % by default. An interpretation of the *t*-tests is given in plain text below the table.

The regression between peer results and laboratory results is calculated. Ordinary linear regression and Deming regression are provided. Adjacent is a Bland-Altman inspired difference plot which also presents the residuals of the observations. An "allowable difference" can be set and resembles the A-zone of an error grid.

The within- and between series imprecision is calculated using the ANOVA approach to analysis of variance components.

The pooled standard deviation of the peer group and the measurements of the laboratory are calculated.

Report of results

Results are presented in tables and graphs. Most of the graphs are designed to present multiple results and the choice is made in the appropriate cells by choosing either "Y" or "N". All cells harbouring a choice have a comment attached which describes the choices and their implications.

For a rapid overview of the results differences and uncertainties are displayed in scattergrams. The normalized Gaussian curve is fixed in the graph and the average *z*-score moves and changes size according to its average and standard error of the average, respectively. The *z*-score is

displayed in the graph as a vertical solid red line overlaying a standard Gaussian distribution curve. The confidence interval is also shown and delineated by dotted red vertical lines (Figure 2). The bias or is thus expressed in *z*-values and should ideally be zero. It is zero if the confidence limit includes zero. A significant difference between the *z*-score and zero requires that the average *z*-score is at least *k x SEM* distant from zero. In the example of figure 2 the difference is less than 2 *x SEM* and thus there is no significant difference between the two. The significance is also tested by Student's *t*-test and reported in column R.

The false rejection rate can be changed but 1 % and 5 % are most frequently used, corresponding to a confidence interval of 99 % and 95 %, respectively.

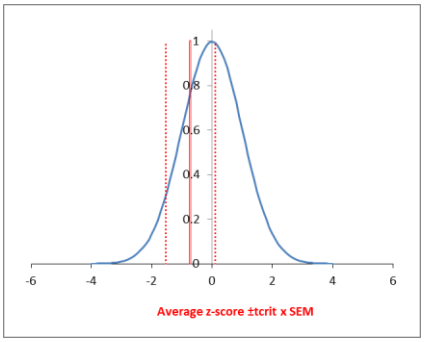


Figure 2. *z*-score and confidence limits projected on a Gaussian distribution curve. If the confidence interval includes zero (0) then the *z*-score of the difference is not significantly different from zero.

Results of comparison are reported in column R and V after Students t-test for paired samples. The critical and the actual *t*-values are shown in the table (Figure 3).

Average Z-score:	-0.72	Peer goup average:	15.75
s(Z):	1.15	SEM_{Peer}	3.65
SEM_z	0.36	Average user results:	15.11
t_{crit}:	2.262	SEM_{User}	3.65
t_{dep}:	-1.988	Diff:	0.6
df:	9	t_{crit}:	2.26
α:	5.0%	t_{dep}:	-2.594
p-value (two-sided):	0.078	α :	5.0%
Bias not different from zero		p-value (two-sided):	0.029
		User and peer group different	

Figure 3. Report of the Student's t-tests and interpretations.

In the first part of Table 3, the departure from zero of the bias is evaluated; in the second, the difference between the results of the peer group and the user is evaluated. In most cases the outcome is the same, but there may be apparently diverging results: one indicating that there is a bias judged by the *z*-score, and the other demonstrating that a corresponding significance between the groups cannot be shown, or *vice versa*. The reason for this is that the comparisons are different depending on the absolute value of the averages. For the *z*-score, the distribution is independent of the size of the average, whereas the absolute difference between the results may vary, particularly if the concentration interval is large. This might increase the calculated *t* so that it exceeds the critical value.

If there is no significant difference between the averages, then different measures need to be used to quantify the bias. The most obvious method is to report the difference between the

averages. This can be complemented by the pooled standard deviation which is an average of the variation between the results. The scattergram allows the detection of outliers.

The results of the peer group and the measured values from the laboratory are displayed in a scattergram with the equal-line indicated. Both the ordinary linear regression and the Deming linear regression lines and regression functions are calculated and can be displayed with the slope, intercept and attached uncertainties in a table (C31, Figure 4).

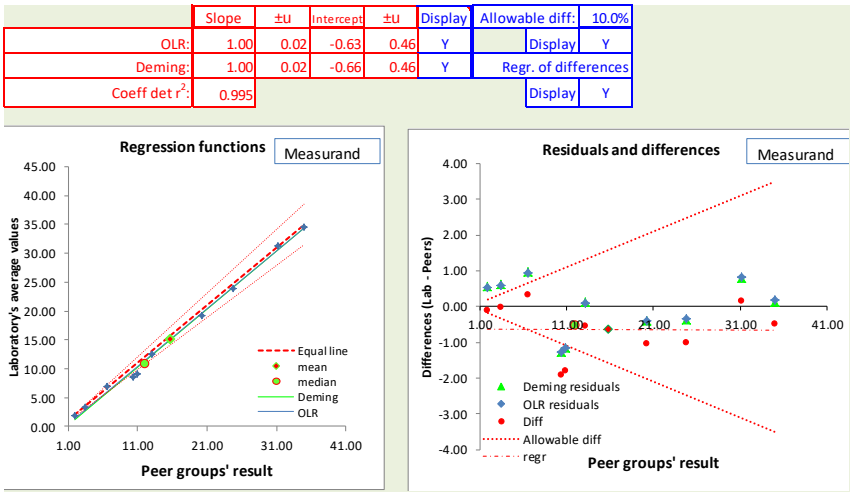


Figure 4. Regression analysis, scatter plot and difference graph. The sector representing an allowable difference of 10 is indicated.

The interpretations of the results are displayed as text, but all relevant results are also shown. A warning has been included that explains that there may be no indication on the Gaussian distribution graph if the average of the *z*-score falls outside the interval of ±4 (99,9 % of the cumulative area under the curve). One of the confidence limits, however, may still be shown.

Bland-Altman-type difference graphs are helpful to evaluate the difference between the group average and the user. Two such graphs are provided (Figure 5). Together, they constitute a crude uncertainty profile.

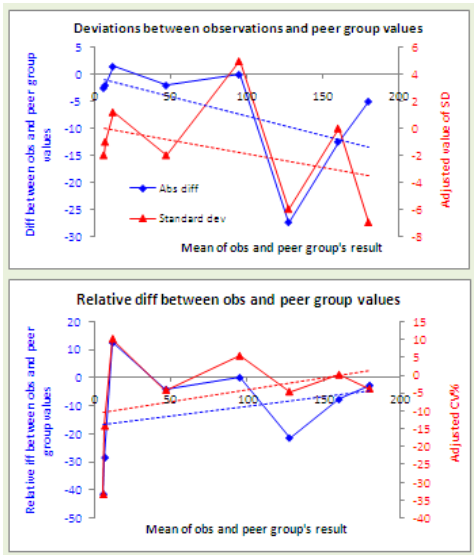


Figure 5. Difference graphs with regression lines indicating a trend of the differences in relation to the concentration of the peer group samples.

The absolute difference between the results and the relative differences are plotted in Figure 5a (top panel) and 5b (bottom panel). Provided the standard deviation of the peer group distribution has the same sign as the difference between the averages, then this can provide additional value to the graphs. If the input data has not been sorted the connecting lines will be difficult to follow. It is possible to sort the table at any time, if for instance additional results have become available.

All differences are displayed together with a 'trend line'. The standard deviation and coefficient of variation are plotted on the secondary Y-axis (on the right-hand side).

The uncertainty profile is supposed to rise and fall harmonically and continuously. The trend lines give a summary of the uncertainty profile. The different quantities normally develop together although differences occur at the low and high ends of the concentration interval.