A STATISTICAL PROGRAM FOR THE DATA EVALUATION OF A THERMAL IONISATION MASS SPECTROMETER

J.G. van Raaphorst
Energieonderzoek Centrum Nederland, Petten (N.H.), Nederland

In mass spectrometry statistics is often a loaded word. What I want to bring into discussion is a statistical program that we use in routine operation. First of all something about our machine: we have a Micromass 30B. The measurements are performed by a stepwise changing of the magnetic field. For instance, for non-spiked uranium the sequence is: 238-236-235-234, this is performed three times, followed by 238-233.5-233.5-233.5. The last three measurements are background measurements. This cycli is repeated 4 times to obtain a full measurement. All these data are stored in the memory of a PDP-8 computer.

At the end of each cycle the background is measured three times. To avoid the influence of memory effects, the first measurement, channel 14, is rejected. The other background data (10 in total) are used as a correction for the intensities in channels 1-13.

A remark may be made that also the intensities of the 236 and 234 are possibly not free from memory effects. I agree with that, when you want to perform very precise measurements on those minor isotopes, you have to use another methodology.

The statistical program starts with a student-t-test. This test is performed to check if there is statistical evidence for a difference between the series of intensity data and the background, the so-called null hypothesis. In our case the null hypothesis is that the difference between signal and background is zero.

The applicability of the student-t-test depends on three conditions.

a) The distribution of the data has to be normal. I know several colleagues put question marks around this. However, a number of background measurements showed that the distribution was normal. It is not possible to apply such a test to every measurement so in further experiments a normal distribution is assumed.
b) The samples have to be mutually independent. The way of measurement

makes this very likely.

c) As a third condition which is important in this case is the equality of the standard deviation of both populations. This was tested and it proved to be true in the case of comparability of signal and background, which is the only important case.

The evaluation of the t-factor is fairly simple. If one has two data series A and B with N_A and N_B determinations and M_A and M_B as their means, the variance of the sum of A and B is

$$Sp^2 = \frac{(N_A-1)S_A^2 + (N_B-1)S_B^2}{(N_A-1) + (N_B-1)}$$

The student-t-value is then achieved by

$$t = \frac{M_A-M_B}{Sp(1/N_A+1/N_B)^{\frac{1}{2}}}$$

When the calculated t-value exceeds a certain tabulated value, the null hypothesis has to be rejected. For this rejection we apply a 95% confidence interval. This means there is a 5% chance to reject the null hypothesis wrongly. The critical value is calculated with the help of a formula, which we developed from tables with critical values. The formula is:

$$t_{crit} = \frac{N_A+N_B-2}{0.51(N_A+N_B-2)-0.613}$$

After the student-t-test there are three possibilities:

a) There is no difference between the intensities of a certain mass and the background. The measured values are rejected from further mathematical treatment.

b) There is a negative difference. The operator is warned, there is something wrong with the background setting.

c) There is a positive difference. These values are used.

The intensities with a positive difference are corrected for background by subtraction.

As a next step in the statistical program the measured values are corrected for signal drift. We apply a linear interpolation between the different 238 measurements. A linear correlation is calculated between the 238 intensities in channel 1 and 5. The other intensities are corrected back on the time of the measurement of channel 1.

Then the same is done between the channels 5 and 9 and everything is calculated back on the time of channel 5. The same is performed between channel 9 and 13.

From the corrected intensities the ratios are calculated, all against the main isotope in this case usually 238. For uranium it delivers three sets (or less) of 15 ratio values for each full measurement.

The existence of the presence of outliers is checked by the so-called Dixon-test. To this purpose the calculated values are set out in a descending sequence. The value tested is the ratio r calculated from the distance of an extreme value to its next neighbour and the distance between all or nearly all values. The calculation depends on the number of ratios.

n	ratio r	formula for critical value
$3 \leq n \leq 7$	$r_{10} = x_2-x_1/x_n-x_1$	$r_{crit} = 0.181062 + 2.29645/n$
$8 \leq n \leq 10$	$r_{11} = x_2-x_1/x_{n-x}-x_1$	$r_{crit} = 0.169509 + 3.07777/n$
$11 \leq n \leq 13$	$r_{21} = x_3-x_1/x_{n-1}-x_1$	$r_{crit} = 0.218345 + 3.93352/n$
$14 \leq n \leq 25$	$r_{22} = x_3-x_1/x_{n-2}-x_1$	$r_{crit} = 0.226632 + 4.47516/n$

After calculation of the ratios a comparison is made with a critical value. When the experimental value of r exceeds the critical value the null hypothesis has to be rejected. The null hypothesis reads: an extreme value belong to the population. The confidence interval used is 95%.

It is clear that when a large number of data are rejected as outliers, the remaining data have no analytical significance and the whole measurements has to be repeated. In our case the maximum number of outliers

is 2-3. As a last part in the program the means of all ratios are calculated and the coefficient of variation.

The computer program is written in Basic and is available for anyone who is interested.

Discussion after the introduction of J.G. van Raaphorst

Barnes: I and a number of my colleagues have maintained for many years that if you cannot explain the data point you are going to throw out, you have no right to throw it out. These tests are really unfair.

Beyrich: The rejection of outliers is something of a philosophy you may discuss for hours and hours. But, in principle, I completely agree that it is bad to reject something if you are not absolutely sure about the reason, if you do not know that something special has happened. If you try to apply things like variance analysis, you are only allowed to apply them if the data material is homogeneous. That it is homogeneous you have to check by the application of outlier criteria and then, when you find some data which would be outliers, you do not know what to do. You can keep them; then you cannot proceed with your evaluation or you reject them and you may calculate something which, if too many outliers are rejected, is no longer realistic to describe the actual situation of measurement. I think it does not make too much sense to develop a philosophy; one should better try to be as reasonable as possible to reach the aims, to come to conclusions and results which still describe the actual situation.

Barnes: I would certainly agree with you. You can always calculate something. It is describing what it means after you have calculated it. That is the problem.

De Bièvre: I would indeed challenge fundamentally outlier criteria being calculated from the measurement itself. I do not object using outlier criteria like ten times the usual standard deviation of the combination analyst-machine-procedure and I do not mean the standard deviation of the method but of the above mentioned combination. Though I prefer ten times the standard deviation I may also go for 5 times and priori given criterion. That takes care of events which are really abnormal. If you throw out something on other grounds, calculated grounds, one is really fooling himself slightly -as well the colleagues in meeting and literature. It gives oneself the idea that the quality of the data is better than it really is. One should give himself a real objective opinion how good one's measurements are.

De Bièvre: The second point is that it constitutes an artificial improvement of the precision. A third remark I want to make is especially to the statisticians. Statistical models assume knowledge of the population. The question is whether the classical statistical model do apply to our measurements. I simply do not think so and when statisticians are trying to correct the experimental people and tell them how to measure, I want to throw the ball back and ask then to develop statistical models which fit the measurement world.

Agterdenbosch: There is one possibility with regard to the t-test, which was mentioned. You can use the Wilcoxon test which does not assume a normal distribution. The Wilcoxon test is free from any condition in this point and you lose very few information. With regard to the application of the Dixon test I agree with most of the things which were said before. One of the points of the selection of a good test is that it should give as good discrimination as possible and I think the Dixon test is not very favourable in this respect. I think there are other ones which give more probability for detecting what you want to detect.

Fudge: I think the use of statistics is to find out where we are really going wrong and put that right. If we could do that we should have achieved something with the use of statistics. If we have enough observartions that we are not as good as we think we are. When we are going wrong somewhere, we may be able to put a finger on where we have gone wrong and look at that point in more detail. We really have to go back at the beginning and look what each stage contributes to your final error. What in the introduction has been said is mainly what we are getting out of the machines and the variations which we get in the machines. These are perhaps responding to a bit more statistical treatment than many of the other ones. You have quite rightly systematic and random errors all the time. Systematic errors you should find out, understand and put right, at least know the magnitude of them. Random errors you only can put right by very careful observation of details. Let us use statistics in the first place to find out where we are going wrong and how good or how poorly we are performing and then work from there to put it right.

Bremmer: An intersting point for us is to know what different operators do. Dr. Barnes has said he never rejects results, that is one approach.

We certainly do the same procedure as here in Petten. How many people do reject outliers, do they roughly a procedure as presented or what?

Barnes: It is perfectly fair to reject outliers which are caused by influences from outside, as in our laboratory happened the start of an elevator. I think there are other things we have all seen. We are going along a set of peaks and we are drawing nice straight lines and all of a sudden something happens. This can perfectly be explained as your filament control jumped or something like that. You can use statistical methods to find these and reject them and nobody will quarrel with that. It is my opinion that this is the reason why a computer will never replace a mass spectroscopist. A mass spectroscopist sees this, knows what happens and can readily explain this. So I do not think rejecting these kinds of things we have any quarrel about at all. But I would like to make still one argument. When we check a method we run a standard n times where n is a very large number. If we have done this we can calculate a statistical value for how well we can run standard. And now our problem comes. We get one sample from our customer and he wants an analysis. We run his sample and we get one result. None of us is going to tell that guy that it is an infinitely poor analysis so what do we do. We apply thus statistics to that sample, that is only fair if you really give him that number. That is our position for not rejecting data, in fact we do not reject data. Who is applying statistical tests? It is one side of the room against the other half. I should like to hear some comments from people who apply some test, given this frame how do you justify this statistical test.

Unsworth: Have you any evidence to show that a statistical test will not reject your elevators or other things you easily get detected by an operator?

Barnes: Oh no I think they will. But the point is that we do not know by a statistical test that that is what in fact happened. We are averaging across the top of a peak. All of a sudden we get a peak that is bad and you have no way by statistics of telling that is a bad analysis or something went wrong with the machine.

Unsworth: You only should have a problem when things have been rejected that should not have been rejected.

Barnes: That is true, how do you see the difference? The human makes a decision about what is rejectable but he has some evidence why he rejected that and not just a statistical test. I maintain that is inadequate, I maintain statistics do no apply to our business.

Beyrich: In many cases the statistical test for an outlier is mainly performed to get aware that there has something happened and then to look what was the reason. The question whether or not you really reject the value is another problem (exposé on the blackboard about the Pafex experiment).
(Concerning the question whether it is meaningful to consider statistical mean values as best estimates for true values);
Just by interlaboratory comparisons you have first to find out whether there is an error source which causes all laboratories to have a bias in the same direction. This, for instance, we found in the 235-U determination in uranium hexafluoride by gas mass spectromety if no double standard technique was applied. Then you have always a small memory effect and this tends always in the same direction. In the interlaboratory comparison indeed we got the result that the mean of all laboratory results was significantly below the characterization value. But when you have found for a certain type of analysis that the mean value from a number of laboratories with its error limits lies within the error range of the characterization value, I do not see a reason why you cannot use it as an estimate for the true value, why you cannot reject an individual value as an outlier.

Barnes: The purpose of this experiment is to find out how the community is doing in making this particular analysis. In our safeguard experiment we depend on a single laboratory making an analysis. Therefore the value of this experiment comes from showing how well an individual laboratory might well be expected to do. I do not really see what is the point.

Beyrich: The point is that we pursue a different aim. There were, for instance, two different spiking techniques used and now we tried to find out by which techniques you will have a greater or a smaller spread among laboratories.

Barnes: The important thing is how well do I do with this different spiking techniques. Then I know that I ought to use. I do not deny that you do not get some information like that from here, but I maintain the

value comes from how well we did in this experiment. And say well the experiment indicates that the laboratory community is within control is false. It is making us look better than we are.

Beyrich: Yes, but they are not looking for the laboratory performance but rather for the method performance. If you want to see whether a certain method includes more or less error sources, then you have to look at a whole crew of laboratories.

Fudge: The big trouble with the mentioned experiment was that they did not send any standards out. Really everybodies result is as good as everybody elses. We had no feel wether the stability of the solution is all right, whether the treatment the people it had given was all right. Everybody assumes that the mean is the right answer. Without sending standards around you have to accept the situation that anyones results may be the right ones.

Van Raaphorst: We have to come back to the question of mr. Bremner. I should like to know how mr. de Bièvre deals with his raw data, what selection do the operators make.

Hebeda: I am against data handling at all - besides the elevator effect. You are handling statistics between n is 3 and 7, that is one problem. You have to apply a t-test where sigmas are quoted. Can anybody tell me which sigma I should take for this test? Just the sigma from the electronic equipment or from the standard?

De Bièvre: A pregiven criterion f.i. 5 sigma would be an acceptable situation. We apply that and for the rest we are using all the data comming out of our machines.
I would like to ask a question to dr. Beyrich. You were assuming that the results of the laboratories taking part in the PAFEX experiment had a normal distribution. In my opinion you cannot make that assumption, each laboratory is too different. So therefore some of the conclusions based on these wrong assumptions cannot be used. I take now the defence of your method of handling the data. It is a very nice and useful method and so far the best description of the actual state of affairs in what we may call comparison of declaration and verification data. Describing the situation is one thing but what we are really up to, especially in safeguards or in any nuclear material management how far are these values from absolute?

De Bièvre: This is a slide for a normal uranium determination by isotope dilution where the mean of the laboratories is 0.5% off the true value. The fact was perfectly explainable afterwards. Most laboratories probably used for the measurement of the 233/238 ratio an electron multiplier. The systematic error from this multiplier puts the 233/238 ratio wrong and hence the 238-U concentration calculated from that ratio.

Beyrich: Unfortunately, there are no error limits indicated in this figure. If we would add error limits on the characterization value and put sigma ranges or something like that, we would probably see that both overlap. It would be very difficult to see whether there is a really significant difference between the characterization value and the mean. The other thing I wanted to show was obviously leading to a misunderstanding. I am not happy at all about functioning of statistics, assuming you have a normal distribution and things like that which in practice do not exist. What I want to show with this curve is that we are just trying to make an approach in a very empirical way, more or less without statistics. We draw the curve on empirical data and we take only statistics to have some background in order to know where we can look to estimate our sigma. This method is only valuable to get an estimate of the variance without prior rejection of outliers. It does not say anything about calculation of the mean, that is something completely different. Unfortunately, we have absolutely no idea what else we could dot to find a true value. I mean that you have many measurements so that you can get a mean or you have a laboratory which takes some effort to furnish a characterization value and you can believe it. These are the two possibilities but in normal practice you cannot establish characterization values for all samples. The only way you can handle the problem is to make sure by intercomparisons with characterized materials that there does not exist such a systematic error source. After this has been proved, I think it is justified to consider the mean as the best estimate of the true value if no characterization value exists.

Barnes: We have to come back to our subject the data treatment in laboratories.

Tyrell: In your original presentation you said you stated a background

correction for uranium at mass 233.5 and plutonium at 240.5. Our experience is particularly with an integration system that this can vary considerably with the place you take the background correction and this causes a fundamental error. Have you looked at background correction at different mass positions?

Van Raaphorst: Yes we have noted too that there are differences in the different positions for background. We used the 233.5 position but nowadays we use also other ones.

De Regge: We have noticed this too and in many cases we only measure two isotopes, one isotope ratio at the time. The background values are measured on each side of the peaks. This is of special importance for the minor isotpes of uranium- and higher plutonium isotopes. Returning now to statistics we have nobody to watch elevators and we use statistics to find out if something occurred. Usually if a data reading is thrown out by the statistics it corresponds on the graph to a small pulse or a cosmic entering or a spike or something like that. There is a visual evidence that it should have been thrown out. On the other hand we use a two sigma criterion and I think there is some argument for that. Because in a series of 10-16 measurements it will throw out one value, it will never throw out two values, then the sigma will tend to be too large. If two values are outliers in a series of sixteen, there is really something to say about that measurement.