LETTER FROM THE PRESIDENT

Time to get close

The president of ENBIS, Fabrizio Ruggeri, appeals to statisticians and companies to bond more closely

In my previous letter I discussed the need for statisticians to be open to new challenges arising from society. An example is the increasing awareness of the citizen's role and the opportunities offered by the internet to boost the new field of e-democracy, where statistical and decision analytic methods can be exploited to analyse problems and make decisions. Now, I want to address the opposite viewpoint: if it is true that statisticians should be open to novelties in society, society (mostly industry and business) has to be keen on what statisticians can offer.

As an anecdote, one of my students was enrolled in the top Italian Masters programme on applications of mathematics in industry and business. She was studying the effect of a campaign on the sales of a product of an important Italian company, where she was working as part of the Masters programme. Time series is the natural, textbook method that could be used to model the sales, but not in that company. Despite my efforts, the company officers wanted to 'invent' a summarising index whose properties were dubious, although my intervention made it sounder from a statistical viewpoint.

This is just one of the many examples of mistrust in statistical and mathematical methods I have met in industry (Alfa Romeo), then as a consultant and, eventually, as a researcher at the Italian National Research Council (CNR) and faculty in graduate programmes.

We, industrial statisticians, often discuss ways to spread statistics in companies and we are concerned about our limits. I believe our counterpart, the companies, should be similarly concerned.

Briefly, here is the attitude of many (not all, luckily) companies, at least in Italy, about research and higher studies. Companies rarely hire PhD graduates, especially in mathematical-related disciplines, and they invest almost nothing in research. Italy had top private research centres (such as Olivetti in computer science and Montecatini in chemistry) but almost all of them disappeared a long time ago. There is a tendency to ask the public research centres (such as CNR) for help at almost no cost to their own research centres. Furthermore, companies often prefer to invest in consultancy from engineers rather than from statisticians. We should explore why this happens.

These sketchy examples should show my point: companies have to act themselves to take advantage of the opportunities offered by statistics and statisticians. This is where ENBIS's role comes in …

ENBIS and companies have been reciprocally useful in the past and they could be even more so in the future. Our national representatives (see their list at the ENBIS website, www.enbis.org) have been invited to contact companies in their countries to promote their involvement in ENBIS. Although our annual meeting is well attended by people from companies, ENBIS must offer companies much more than the possibility of presenting and discussing their works, such as in the next ENBIS annual meeting, in Wroclaw, Poland, from 18 to 20 September.

ENBIS is providing various ways to promote continuing education, problem discussions and joint research, and we are keen to hear from companies about their needs and their suggestions. Here are just few of the ways companies could benefit from ENBIS's activities.

- Courses offered throughout Europe, like the three days course on '*Life length and reliability – needed for better design*' at Lyngby, Denmark, from 30 May to 1 June.
- Specialised workshops like the one on Data Mining in Gengenbach (Germany), from 5 to 7 April.
- Workshops before and after the annual meeting at a low price (last year they were on *Data Mining*, *Operational Risk Management*, *Simulation of Clinical Trials*, *Statistics for Innovation and the Design Process*, *Statistical Consulting Skills*).
- Workshops in the annual meeting where, for example, companies present their problems (such as *software quality and electricity customers' profiling* in the last ENBIS meeting) and statisticians propose models, with lively floor discussions.
- Publications. A review book on industrial statistics will be published shortly as part of the Pro-ENBIS project. We are investigating the possibility of starting a series of books for practitioners with a leading publisher.
- A newly-appointed ENBIS committee will look for funding opportunities, publicise them among ENBIS members and prepare outstanding groups for proposal submissions.
- A newly-appointed ENBIS committee will promote courses. Companies will be welcomed to contact the committee to arrange for courses to be delivered by outstanding ENBIS members.

Each year, ENBIS recognises the role of the managers who are contributing to the diffusion of statistical methods in industry and business by presenting the Best Manager Award, companion to the Young Statistician Award and the George Box Medal for outstanding contributions to industrial statistics.

Statisticians and companies can fill the gap between expertise and problems, provided each moves towards the other. ENBIS is here to help!

PS: I wrote this note just after returning home from Lima (Peru) where I attended ISBIS 5, the Fifth International Symposium on Business and Industrial Statistics. I expressed these opinions (and others) in a final round-table discussion on the 'Future of Quality Technology and Industrial Statistics'. This was an important promotion for the use of statistical methods in South America and I thank the organisers for it. Among the many people in the discussion, there were 30 from Peruvian companies. Perhaps the future is brighter than I depicted in my letter!

Big decisions felled by imprecision

John C Nash warns against assuming that your favourite package will correctly calculate means and variances. He offers defences

Any business that makes a critical decision on the difference between variances computed by dividing by n rather than $(n-1)$ clearly deserves to go bankrupt. But what about the business that gets those numbers wrong because it uses poorly-devised software?

Consider one of the commonest sets of calculations in statistics – the computation of mean and standard deviation.

Modern scenarios

- A large multinational organisation maintains data servers in a number of locations. These amass data from activities of the organisation (such as sales, production, flow of materials) within some region or division. The CEO likes to have daily reports that include statistics on such quantities.
- An internet communications provider needs to watch the level of traffic over its links and hubs. Local and aggregated statistics are wanted.

It is not easy to find out how the particular statistics in these cases are calculated. One way is to copy all information into a central data repository for calculation, and I suspect that this is the usual method.

Wherever decisions are made on the basis of data, means and standard deviations are central to calculating t and z statistics in hypothesis tests. If you get the mean and standard deviation wrong, these statistics are going to lead to inappropriate decisions. Even if your business uses good software, you could be affected by business partners who mis-compute quality measures, or zealous regulatory bureaucrats whose inaccurate statistics cause them to waste your time. People tend to cover up such embarrassments, and the only stories I can tell might upset a valued client.

The problem

We want to compute the mean and variance of a set of n numbers $y(i)$, $i=1:n$. We will ignore fancy problems involving grouped data, though I will be happy to exchange views with other workers on that subject. Using a pseudo-code notation, we define *(1)*:

y_bar = sum(y(i), i=1:n) / n

Having *y_bar*, we then define the variance as *(2)*:

var_y = sum ((y(i) – y_bar)^2 , i=1:n) / n

We define *var_y_sample* as:

*var_y * n / (n-1)*,

which aims to provide a better estimator of the population variance. Division by n simplifies the exposition, but in the calculations by standard packages that follow, division by $(n-1)$ is used. However, it is difficult to think of cases where $(n-1)$ division is more than a statistical nicety for most management issues.

When we have n in the millions, the time to pass through the data twice, once for the mean and once for the variance, becomes the main cost factor. Things get nasty when the data is on several, or perhaps hundreds, of different storage devices across a network.

A standard result is *(3)*:

var_y = average of the square - square of the average

This catchy formula is useful on many occasions, but it is not suitable for numerical computation, as the example below will show. However, it clearly reduces our work to one pass through the data.

Our goals:

1) Get the right answer for the mean and variance of a set of numbers; and

2) Pass only once through the data.

An example of potential numerical difficulties

Compute the mean and variance of the set of numbers 100001234, 100001243, 100001342, 100001432, 100001423.

Experienced data handlers will quickly re-code to 4-digit numbers: 1234, 1243, 1342, 1432 and 1423, by subtracting 100000000. The true variance and standard deviation are unaffected but the squares of the numbers are much smaller after subtracting. Indeed, the original numbers cause some calculators to fail. An unnamed professor gave such problems to students, saying the numbers were the weights of North American football players in milligrams. Table 1 shows results for some common spreadsheet programs and several added values. meansd3.xls is available online as http://macnash.admin.uottawa.ca/~nashjc/enbis/meansd3.xls if you want to try it. Or try the data in your favourite statistics package.

These results reveal differences between programs and program versions and provide guidance on where to trust or be wary of results. Changing low-order digits are of interest to number-crunchers like myself, but largely irrelevant to practical users. On the other hand, if there is a sudden failure as the addition goes from one level to another, we may want to check what is going on. If we are computing statistics from similar data, we are going to get things wrong and likely will make bad decisions.

Use these results to check the internal precision of the floating-point functions

used by different programs. They will reveal differences between programs and program versions and provide guidance on where to trust or be wary of results.

Software may format results to shorter lengths than available, but this seems unlikely to be a source of decision-altering errors. Clearly for maximum value, examples like this should bear some resemblance to the scale of the inputs in your own situation.

Note that large numbers also affect the mean. At some point adding a small number (your salary!) to the total income of a country that includes Queen Elizabeth or Bill Gates makes no change to the total. In fact, once the register that holds the sum reaches its maximum number of digits, we thereafter start to lose information. This is the large-plus-small problem.

Most people think this example is pretty artificial. However, many quantities we measure today feature limited variation; for example, the concentration of a component of a mixture or the width of a tape. The source of our numerical troubles is that the variation is small relative to the size of the inputs. Since this example uses integers, a better test may use numbers that force an input conversion.

A different criticism is that we used the built-in standard deviation functions. To focus the test on our data and not the program doing the work, I recommend computing equation (3) above, since we then work with just additions, multiplications and divisions. After all, the main reason for these tests is to find out where we may get into trouble when we use data to develop statistics to support decision-making.

The numerical errors shown in table 1 may be avoided with Åke Björck's revised form of sum of squares $S = n * var\_y$.

(4) $S\_adjusted = sum\ (\ (y(i) - y\_bar)\text{^}2\ ,\ i=1:n) - (sum(\ y(i) - y\_bar\ )\ )\text{^}2 / n$, where we can regard the second term as a correction. In exact arithmetic it is zero, but when we use real arithmetic it tells us something about our numbers.

From this discussion we have two tools – the textbook formula and the corrected two-pass method – that let us quickly get a rough idea of the size of the error we may make in computing the standard deviation.

The importance of staged algorithms

Besides the numerical issues, we have to face organisational and logistical ones. With data distributed over many sites, it is clearly not a great idea to have to pass through the data twice. And we would like to keep our mean and variance as close the definition as possible. One possibility is to use an algorithm that lets us combine the results of two sets of data. For these calculations, rather than use the mean and variance, we will use T_i, S_i, and n_i to represent the sum, the sum of squared deviations and the number of observations from a subset of the data. Thus

$S_i = sum(\ (data\_values\_in\_set\_i - Ti / ni)2\ )$

Take two collections of such data, call them i and j. Then clearly

$n_{i,j} = n_i + n_j$

$T_{i,j} = T_i + T_j$

while a little more algebra gives

$S_{i,j} = S_i + S_j + \{(\ n_i * T_j - n_j * T_i)^2\ \}// \{n_i * n_j * (n_i + n_j)\}$

As illustration, starting with an empty stack, we push a single element onto the stack. Clearly we have the triple:

n	T	S
1	x_1	0

since the variance of a single number is zero. Pushing another observation onto the stack gives

n	T	S
1	x_2	0
1	x_1	0

Noting now that the two top elements of the count or n column are equal, we use the combining rules and have

n	T	S
2	$T_{1,2}$	$S_{1,2}$

The process then continues until we have two more individual observations on the top of the stack. These are collapsed down to a single row with *count* = 2, which then is aggregated with the existing row which also has *count* = 2. Thus we mainly aggregate triples with equal numbers of observations. Large plus small cancellation could be further reduced by first sorting the data, but that is not an operation we want to do on distributed data, though it could be done locally. Moreover, local operations can be done without compromising security and privacy of individual observations. More details are given by Chan, Golub and LeVeque(ftp://reports.stanford.edu/pub/cstr/reports/cs/tr/79/773/CS-TR-79-773.pdf). The pairwise algorithm is awkward to code in a spreadsheet, but is easy in traditional programming languages. A BASIC program can be found at the site http://macnash.admin.uottawa.ca/~nashjc/

Program	Base	Add 1E+5	Add 1E+10	Add 1E+15
Excel 2000 9.0.2720	94.69266075	94.69266075	128	0
Excel 2003 11.5612.5606	94.69267725	94.69267725	94.69267725	94.69267725
Gnumeric 1.4.2 Linux	94.692660750452	94.692660750451	94.692660750451	94.692677251200
Gnumeric 1.6.1 Windows XP	94.692660750000	94.692660750451	94.692660750451	94.692677251200
OpenOffice 1.1.2 Linux	94.6926607505	94.6926607505	0.0000000000	0.0000000000
OpenOffice 2 1.9.125 Linux	94.6926607505	94.6926607545	0.0000000000	0.0000000000
Quattro Pro 9 9.0.0.883	94.6926607504510	94.692660750451	128.349522788361	ERR
Minitab 14.2	94.69266075	94.69266075	94.69266086	94.69993730

Table 1. Summary of standard deviation calculations for different programs using the (*n*-1) division sd function appropriate to each program. Formatting the cells in the spreadsheet and/or the word processor may affect the digits displayed.

enbis/pvar3.bas. If you need an interpreter to run this, see tips online at http://macnash.admin.uottawa.ca/nlpe/. In double precision it gets decent results on the test problem above for all of the example scalings.

The appropriate data files are in http://macnash.admin.uottawa.ca/~nashjc/enbis/pvardata.zip.

Issues for the near future

Beyond institutional and governmental statisticians, there are situations with point of sale data that is coming from large numbers of cash registers. A typical client had 1,000 to 2,000 stores with perhaps a dozen terminals per store across a wide geographic area. Multiple brand outlets in a single mall may share infrastructure and even staff. I believe that the POS data is either kept only in aggregate form, or else is archived centrally, possibly in partially-aggregated form. Details are important, particularly those concerning data access and storage, and such details are often kept guarded for commercial and consumer privacy reasons, but it is important that we still have methods that allow means and standard deviations to be computed. Some hints of things to come may be seen in http://www.niss.org/dgii/TR/technometics200511.pdf and also at the following site: http://www.dsprelated.com/showmessage/30504/1.php.

	Base	Add 1E+5	Add 1E+10	Add 1E+15
Textbook formula				
sample SD	94.6926608	94.6926608	0	0
Corrected two-pass method				
corrected sd	94.692660750451	94.692660750451	94.692660750451	94.692660750451

Table 2. Example (computed in Gnumeric 1.4.2 Linux) of use of the textbook and corrected two-pass method for standard deviation computation. See Table 1 for the built-in function results. Excel 2003 returns similar results.

The author:
John C. Nash, School of Management, University of Ottawa, P.O. Box 450, Stn A, Ottawa, Ontario, K1N 6N5 Canada email: nashjc@uottawa.ca

Truth beneath the figures

John Logsdon warns against button pressing analysis and urges you to understand the structure of your data.

Regressions. Love them or hate them, they control our lives. They are used in engineering to set the ignition timing in your car; they control your washing machine when it detects a heavy load; they are used in social sciences where the decisions taken affect the funding of your local hospital or school or the level of taxation the government thinks it requires.

By regression I mean a process that attempts to relate possibly influential variables to an outcome by purely statistical means. Sometimes the form of the relationship is imposed.

There are three main failures in regression:

- Failure to appreciate the structure of the data;
- Failure to check the diagnostics; and
- Failure to examine the predictive part of the model.

Note the order in which I have put these. They are all inter-related and the result of a change in one part of the model will almost certainly affect other parts, although the more robust your model is, the less this is true.

Modern technology makes it too easy to press a button without understanding the background. And as managers just want an answer, they often get what they deserve – the wrong answer. The unfortunate thing about statistics is that is often difficult to see that an answer is wrong.

Data structure is nearly always ignored. Hence, most regressions are wrong. For simplicity, I shall consider a case where there is no predictive model at all – I am just trying to estimate the mean. Some may recognise this as an analysis of variance, but these all sit within the same general framework.

Data structure is the natural grouping of the data. How do you know if there is a structure? Thought experiments are valuable. When you start to analyse any new dataset, ask yourself the following question: For each group, is it possible that the outcomes for all members of group are similar, even after taking into account some value associated with the group? If the answer is 'yes' then your data has a group structure. Even if some aspects may turn out not to be significant in the end they still need to be tested. Data are guilty until proved innocent.

What are the implications of this grouping? Let's start with an example that is really simple – and obviously wrong. Someone gives you some data and asks you to analyse it. The numbers are:

-1.118, -0.5, 0.5, 1.118, 10

Even a blind man on a galloping horse can see that the last of these is in a different group, assuming it is not a typographical error. But suppose the person who does this analysis doesn't think and just finds the 'mean' button or drop-down? He produces a mean and standard deviation, 2.00 and 4.56 respectively, and hands it back to the engineers, who fall about laughing. A little inspection shows that the first four readings are in a group, which we will label **one**, and the last is in group **two**, and there is a good reason for this. I did say it was a simple

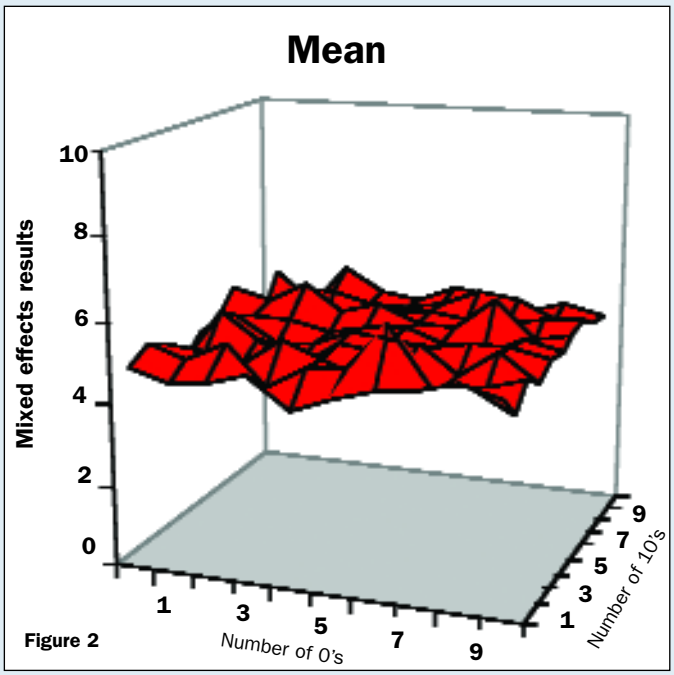

Figure 2

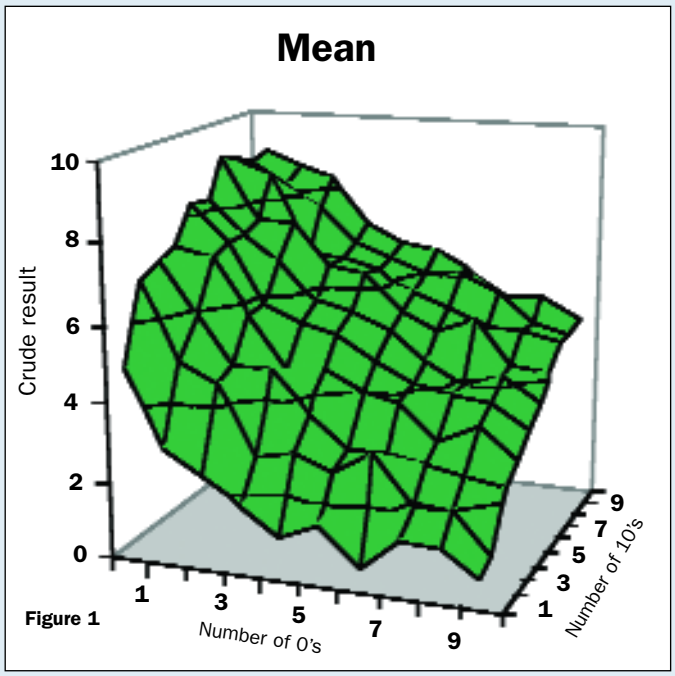

Figure 1

example, but it illustrates the effect of ignoring the structure.

How would an engineer solve this problem? He would split the data into groups and calculate the mean and standard deviation of each group. The engineer would ignore the fact that group two had only one value, so he couldn't calculate a standard deviation to compare or pool; he would just assume it was the same as for group one. So he would calculate a standard deviation of 1.00 from group one only, a mean of 5.00 and a between group standard deviation of 7.07. But he has only instinct to guide him. He cannot justify these results as being the most likely to arise from the data.

In the last issue, we saw that maximum likelihood is a powerful method for estimating the parameters to a model; we say calibrating the model. One of the fundamental assumptions in maximum likelihood is that each observation is statistically independent of all others. In other words, we do not expect the outcome of one item to depend in any way on the outcomes of any other items. This cannot be true if there is any group structure to the data. If two items are from different batches, different production lines, or are made by different operators, then their outcomes are not statistically independent. Oh dear – what do we do now?

Well, fortunately things are not so bad. Because maximum likelihood is so powerful, it is possible to include the correlations, or rather the covariances, as part of the model.

So, rather than having another problem to solve, we use the same philosophy, but add the covariances into the model. If the model correctly represents the data structure, we can then still use maximum likelihood.

Having brought himself into complete disrepute, our poor innocent rushes to the nearest shoulder to cry on, that of the always sympathetic statistician who immediately recognises a very unbalanced analysis of variance. To remove the bias, the statistician does a quick mixed-effects calculation by maximum likelihood, which gives the correct answer: overall mean of 4.96 with a standard deviation within the groups of 1.00 and a standard deviation between the groups of 7.03. A relieved innocent, in sackcloth and ashes, tables these numbers to the management and peace is restored.

The key problem here was that the data were not only in two clear groups, which had been ignored, but also that there were not equal numbers in these groups. One group had four members and the other only one. What happens as this balance changes? And how can a correct calculation of mean (for example) be made, whatever the balance? Surely we need a calculation that is insensitive, or robust, to the imbalance.

I have prepared a simple R program at www.enbis.org/SCW3.R. This assumes that there are two groups with means respectively 0.00 and 10.00 and standard deviations of 1.00. The data are simulated so the results will be slightly different every time, but:

runit(10,10,10)

will produce a 10 x 10 matrix of results for all combinations between one value from group one (mean of 0.00) and one value from group two (mean of 10.00), and 10 from group one and 10 from group two. Figure 1 shows the crude means in green and Figure 2 shows the mixed effects mean in red. See how much flatter is the red plot and close to the obvious answer – and how the green plot could lead you to completely wrong conclusions if you were unlucky. And luck should really play no part in your conclusions.

In this trivial example, the immediate answer given by the naïve analysis was obviously wrong, and we have shown how this can be accommodated within a maximum likelihood calculation. But what would happen if the model were more complex with a predictive model and a substantial structure? Even a slight increase in complexity will lead to a wrong answer that is difficult for an untrained analyst to spot. How do you know it is right then – or wrong? It is only by including the structure that we can estimate the correct values. You have been warned.