Elffers' method and Elffers' mistake

Richard D. Gill*

January 7, 2007

Abstract

I outline a justification of Elffer's approach to the Lucia de Berk data. This is only partly succesful: data from three wards should be combined using a standard method such as the Cochran-Mantel-Haenszel test for independence in stratified contingency tables, before making Elffers' posthoc correction; Elffers' combination—simply by multiplication of p-values—is seen to be incorrect. Reanalysing the original data correctly, yields a p-value of about one in a hundred thousand—small, but not so dramatically small as Elffers' one in 342 million.

Bonferoni

Consider $J = 27$ nurses and suppose for each nurse there's a statistic T_j. We believe that there have been murders committed or attempted at the Juliana Children's hospital, JKZ, and possibly elsewhere too, and wish to investigate if they are connected to one of the 27 nurses at JKZ. The null hypothesis H_0 is that all J nurses are innocent. Under H_0, the marginal distribution of all the T_j's are known. If nurse j is guilty, then T_j would tend to be larger. We fix a significance level α and announce "nurse j is guilty" if and only if T_j strictly exceeds the upper α/J quantile q_j of T_j's null distribution; i.e., the largest number q such that $\Pr(T_j \geq q \mid H_0) \geq \alpha/J$. So q_j satisfies $\Pr(T_j \geq q_j \mid H_0) \geq \alpha/J$, $\Pr(T_j > q_j \mid H_0) \leq \alpha/J$, and moreover, the probability on the left hand side of the first of these two inequalities is as small as possible.

Under H_0, the probability we (unjustly) accuse any nurse is not greater than α. The technique used here is the well-known Bonferoni method for dealing with multiple comparisons or for testing multiple hypotheses, when

*Mathematical Institute, Leiden University; http://www.math.leidenuniv.nl/∼gill

we wish to control the overall probability of error. In this way we control for the well-known problem that a fraction α of null hypotheses is going to be rejected by chance even if they are all true. In this case we can pretend that we did not in advance suspect the individual nurse Lucia de Berk, and did not set up the hypothesis testing problem *after* seeing the data.

The J nurses all work at JKZ. Each nurse may have worked at some other hostpitals also. We may define each T_j just as we like (before inspecting any data, of course), with a view to obtaining a large power against the alternative that nurse j is guilty. The only condition is that the marginal null distributions of all T_j's are known (distribution free tests). Consequently we may take T_j to be based on any of the wards at which nurse j has worked, if we so desire.

Discussion

I now discuss a number of issues concerning the implementation of this methodology in the case of nurse Lucia de Berk.

- *Before* observing any data, the statistician must make a principled choice for each T_j. This involves delineating both the data, and the statistic. Concerning the data, one might like to fix some period of time and then take all workplaces of all nurses during this period. Concerning the statistic, one possiblity is to use the Fisher method for combination of independent p-values, to build a single statistic for the data of a number of wards, based for instance on Fisher's exact test for each ward separately. A more sophisticated but more reasonable test statistic would be the Mantel-Haenszel test, which is designed to have high power against the alternative of a constant odds ratio across the different wards. The advantage of the Fisher method is that it is quick and easy; its disadvantage is that the alternatives against which it has high power depend on the margins of the different 2×2 tables as well as the odds ratios in all the tables. It puts the separate p-values on an equal footing though possibly one of the tables is much more informative than another. Two rhetorical advantages are firstly, that it shows that Elffers' multiplication of p-values is incorrect, since it always produces a larger p-value at the end, and secondly, that it comes from another chapter of the same book in which Fisher's exact test is described.

- Reanalysing the original data correctly with the Mantel-Haenszel (correctly that is, if we trust that the data gathering is unbiased), yields

a p-value of about one in a hundred thousand—small, but not so dramatically small as Elffers' one in 342 million.

- In the present case the null hypothesis distribution of whichever T_j we use, is based on the conditional distributions, given the margins, of all the 2×2 tables of each (nurse, ward) combination. The rationale for this is that we know in advance that murders have been committed and have it makes good sense to *condition* on the total number of deaths or emergencies ("incidents"). These total numbers are irrelevant to the question of who committed the murders and murder attempts, if we know that such were committed. However, if we do not know this in advance, the total numbers should be highly relevant.

- The validity of the method depends on the validity of the assumption, that when all nurses are innocent, incidents (deaths and emergencies) are randomly distributed over shifts, in particular, over the shifts of any particular nurse. There are reasons to doubt this assumption, and a rejection of H_0 could equally well be argued as reason to reject this assumption. Especially, if we are not even sure that many murders were committed or attempted, we should take this alternative possibility seriously. Here are some possibilities to consider.

  - A more experienced, confident, or ambitious nurse, might deliberately choose harder shifts than a more cautious nurse or a beginner. A better nurse might be earlier alert to a developing crisis situation, than a less good nurse. In both of these ways there is a completely innocent correlation between nurses and incidents.
  - A cluster of incidents in a particular ward might be caused by a structural change in a hospital, e.g., the closure of another ward, while at the same time, a particular nurse might coincidentally be working relatively more shifts at the same time (relocations, vacations, recruitments). Thus, trends and oscillations in the case-mix severity on a particular ward could coincide with trends or oscillations in the numbers of shifts taken by individiual nurses, leading to spurious correlations between nurses and incidents, when one looks at relatively short periods of time.

  Thus there can be innocent reasons for medium-term spurious correlations between incidents and nurses, and also for longer term correlations. These correlations in no way contradict the opinion of specialists that any two nurses are interchangeable, when we look at one particular shift. Replacing nurse j by j' should not influence the occurrence or

not of an incident, though perhaps it influences the moment at which the incident is registered. (In the present case there is actually concern that the *definition* of an incident depended on who was on duty – the statistician has to check that definitions are objective and fair).

- Elffers' method of combining wards by multiplying p-values is blatantly incorrect, since data from a large enough number of wards would make *any* nurse eventually guilty. Worse still, by disaggregating a fixed amount of data, the p-value can also be made almost arbitrarily small. The suggestion has arisen that two adjacent and essentially identical wards at the Red Cross Hospital, RKZ-41 and RKZ-42, have been separated merely to exploit this phenomenon. One could have split the data over months, and again decreased the p-value. A suspicion arises that the level of disaggregation, and the wards which are taken in the analysis, have simply been chosen to get a p-value in the region of JKZ director de Smit's initial and premature "one in 7 billion" (which he proudly referred to as "statistics of the cold ground", i.e., a simple farmer's statistics). The approach I propose, using for instance the Mantel-Haenszel test, will be far more robust to further stratification of the data.

- *Why is it important to evaluate Elffers' analysis*? Statisticians believed that there was medical evidence for murders, medical experts believed there was statistical evidence. These mutually-supporting beliefs biased the data gathering and data definition against the defendent. Neither class of expert considered themselves qualified to question the judge's conclusions drawn from the evidence of the other experts. It is crucial to understand how this could have happened and to take steps to prevent it happening again. Statisticians need to realise that their responsibility is not merely to place some stamp of approval on some calculation based on data, the interpretation of which is the responsibility of other scientists, or even of the judge. Not just the statistical analysis is in question, but also the statistical data. A statistician's evidence is only useful as scientific expert evidence if it is independent scientific expert evidence. This means that the statistician has the same responsibility when working on behalf, e.g., of the prosecution, as when collaborating with a medical scientist or a social scientist. The statistician has to question all his clients' assumptions and certainly not to jump to the conclusions which the client is aiming for. In this case, the statistical analysis of Elffers contained controversial features, which needed to be discussed openly and scientifically. Otherwise mis-

carriages of justice are going to occur again, and statisticians are going to be partly responsible again. In this case a key part of Elffers' methodology would not have stood up to peer review. However only the experts directly involved in the case were aware of his controversial (unfounded, and unfoundable) methodolodgy.

- Statisticians are also responsible when their calculations are deemed irrelevant or unimportant, and the lawyers are left to draw their own conclusions from direct inspection of data. A statistician who says that a quantative statistical analysis is impossible, has implicitly got criticisms on the data itself; he or she has meta-statistical grounds for arguing that the data cannot be taken at face value. These arguments need to be brought out into the open. Applied biostatisticians, medical statisticians, epidemiologists, psychometricians, ... know about the many pitfalls, and many ways bias can be introduced. Mathematical statisticians have hardly been trained in this direction and are not even aware of the extensive nomenclature.

- Elffers has repeatedly refused to elaborate on the controversial aspects of his analysis. Within his own paradigm he makes an objective serious error. In public he presents a sanitized and simplified version, which enables non-experts to get the idea that they understand what he does, and moreover that he apparently knows what he is doing since he is able to explain it so transparently. This is deceit, deliberate or not. The judge in the UK cot-death scandal turned down the offer of professional statisticians to analyse the data, since "this is not rocket science". The editor-in-chief of "Natuur en Techniek" refuses to publish a note criticizing Elffers' methodology because "everyone knows that you are allowed to multiply independent probabilities". Elffers is now saying "the case should be reconsidered, without statistics". The suspicion arises that he is anxious for his good name. I am anxious for future defendants. Whether or not Mrs de Berk is a murderer, it is clear that there is a lot wrong with the use of scientific evidence in her case, and in particular, with statistical analysis and with statistical data.

Conclusion

Scientific evidence is too complex for present-day Dutch judges. The scientific community has to take fuller responsibility for the use of scientific evidence in legal cases, an in particular, must be not afraid to cross specialistic boundaries. Scientific evidence, whenever controversial, should be subject

to public peer review.

References

Elffers' analysis is described in a preprint by Meester, Collins, Gill, and van Lambalgen, available from my home page. It has been posted on arXiv.org, http://arxiv.org/abs/math.ST/0607340.

The original data analysed by Elffers, and the data as revised by Derksen and de Noo, are also available, together with R scripts for analysing them in various ways, and further discussion:
http://www.math.leidenuniv.nl/∼gill/R.txt