Statistical significance - meaningful or not

Statistical significance - meaningful or not

C.G.G. Aitken and A. Wilson,
School of Mathematics and Maxwell Institute, The University of Edinburgh, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom

R. Sleeman,
Mass Spec Analytical Ltd., Building 20F, Golf Course Lane, Bristol BS34 7RP, United Kingdom

Abstract

Statistical tests with large sample sizes can have large power. Power is the ability to detect an effect. Detection is indicated by a result which is statistically significant. A test with large power will detect a very small effect. This very small effect may not be meaningful in the context of the analysis being conducted. The courts have the perception that for an effect to be meaningful it is necessary for the effect to be statistically significant. However, statistical significance is not a sufficient condition for an effect to be meaningful. This can lead to a difficulty where testimony of no meaningful effect is interpreted by counsel as one of no statistically significant effect. Should the difference between a meaningful effect and a statistically significant effect be explained in reports and if so, how? Some possible answers are proposed.

Keywords

Banknotes, Cocaine, Statistical significance, Meaningful support.

1 Article

1.1 Introduction

Statistical tests with large sample sizes can have large power. Power is the ability to detect an effect. Detection is indicated by a result which is statistically significant. A test with large power will detect a very small effect. This very small effect may not be meaningful in the context of the analysis being conducted. The analysis may be fortunate enough to have a large dataset available, with a correspondingly large number of degrees of freedom. Any statistical test conducted using such a dataset has correspondingly high power. Very small differences in mean quantities in a comparison of responses to different treatment groups can be statistically significant but would not be judged meaningful by scientists.

Often with problems of inference the question of statistical significance is of no interest. A statistical analysis is conducted to understand the random process that generates the data and the possible sources of variation. This will then help the decision-maker.

However, in court statistical significance can be of interest. The court is interested in differences in responses of some variable to different treatments. There is a general understanding in the courts that for a difference to be meaningful it has to be significant in a statistical sense. Some courts will understand enough to be interested in the level of significance, recognising that the lower the level, the greater the difference in the responses to the different treatments.

If the data set available for analysis is very large then there are a large number of degrees of freedom and associated tests will have a large power. As a consequence, a very small significance probability can be associated with a very small difference in response. The very small difference may not be meaningful in that the difference in response, whilst being very small, is

sufficiently small that there is no discernible difference in response. The court notices, or is told, that the difference is statistically significant. The court then thinks the difference is meaningful and makes a judgement accordingly.

The expert is then in a quandary. They have to explain to the court that the difference is not meaningful. Such an explanation may lead more to confusion than to enlightenment. An example is given in Section 1.2. Section 1.3 suggests solutions to the quandary.

1.2 A case study design and analysis

The problem of interpretation associated with a test with large power is discussed in the context of the variation by location in England of quantities of cocaine on banknotes. This variation is studied with use of a large dataset, with a large number of degrees of freedom, of banknotes in general circulation in England and Wales. Any statistical test conducted has correspondingly high power. Very small differences in mean quantities amongst locations can be statistically significant but would not be judged meaningful by scientists.

Evidence evaluation with the likelihood ratio requires a background dataset from a relevant population in order that an appropriate statistical model may be chosen, with choices of parameters if necessary. Often the choice of the relevant population is uncontroversial. However, in cases that involve the discovery of cocaine on banknotes this is not the case. Background data for banknotes are the quantities of cocaine on banknotes in general circulation. The controversy concerns the choices that are made of the locations from which the banknotes are sampled. The current database in use by the company that analyses banknotes for drugs contains many banknotes taken from banks around Bristol. An alternative suggestion is that the background database be case-specific. For each case, a sample of banknotes is taken close in location to the place where the crime was committed, and close to the environment in which the crime was committed. Thus, for a crime committed

in a night club in a particular town notes from general circulation should be taken from night clubs in that town. The use of the current database assumes that there is no variation in the quantity of cocaine across the country. A test of this assumption (Aitken *et al.*, 2017) had very high power. The difficulty this caused for interpretation is discussed below.

The choice of which banknotes to sample in order to have a representative sample of banknotes in general circulation is by no means straightforward. The notes used in the MSA database were accepted by Appeal Court rulings in 2002 and 2004: *Compton and Compton*[2002] EWCA Crim 2835 and *Benn and Benn* [2004] EWCA Crim 2100. However, this choice of notes for a sample of a population relevant as background to crimes elsewhere in the UK was questioned in a judgement in England in 2015 (*R. v. Rashid and others*, [T20147216], 19 January 2015). The court was very critical of the composition of the database currently used. Remarks included: '[i]t is a database of pure convenience' and '[t]he assertion that notes from banks are typical is not supported by any evidence and is illogical'.

Some consideration of the meaning of the word *relevance* is pertinent. A letter in the *Journal of Forensic Sciences* in 2017 gave the following definition of relevance which is particularly appropriate in the context of sampling banknotes from general circulation:

> The relevant population ... is arguably those who had access to the crime scene or who inhabited the geographic area where the crime was committed. Moretti and Budowle (2017)

In the context of sampling banknotes in general circulation as part of an investigation into a drugs-related crime, this definition could be edited to read:

> The relevant population (of banknotes in general circulation) is

> arguably those that were in the geographic area where the crime was committed when the crime was committed.

The phrase 'access to the crime scene' has been melded into the phrase 'geographic area' and a temporal aspect has been included. The definition that the background database be case-specific is unrealistic. The crime may not be investigated until some considerable time after it was committed or the crime may relate to a long period of time. In both cases, notes would have moved in and out of the area through natural circulation. Similarly notes in the environment of the crime, such as a night club, will also have left the environment through deposit at a bank or with other customers.

The current database has many notes taken from banks around the Bristol area. In order to counter criticism that this procedure was unrepresentative of notes in general circulation, research was undertaken to consider the variation in quantities of cocaine on banknotes in general circulation across the country.

It is important to know if the variation was sufficiently great that the location and environment of the crime under investigation should be of relevance when sampling from the population of banknotes in general circulation. A study (Aitken *et al.*, 2017) was conducted and the results showed that the variation was not sufficiently great and the comment in *Rashid* could be rebutted. However, the analysis also provided an interesting problem concerning a distinction between meaningful and statistical significance.

Banknotes were obtained from eight redistribution centres chosen to represent a source of banknotes with a good geographical spread across England and Wales in terms of area in which they last circulated, following a suggestion by experts at the Bank of England. There are no redistribution centres in Wales and notes from Wales are sent to centres in England. In total, 1950 notes were analysed, each twice. Thus there were 3900 observations and a correspondingly large number of degrees of freedom for the analysis.

A mixed effects model gave statistically significant differences in responses

amongst the eight locations at a very low level. There were two reasons for this. First, there is auto-correlation in the measurements which can cause estimates of standard deviations of regression coefficients to be too small and hence lead to a reduction in the p-values of significance tests. Second, there were a large number of degrees of freedom so the power of the test was very high. However, inspection by expert analytical chemists determined that the differences were not meaningful in the sense that their conclusions would not differ for notes analysed from different locations. Some scientists use a verbal scale to interpret a numerical likelihood ratio (e.g. ENFSI, 2015). In such a paradigm, a meaningful difference may then be thought to be one that leads to a change in the verbal description of the support for the likelihood ratio.

1.3 Discussion

The conclusion of the analysis was that the differences in mean responses for each of the eight locations were not meaningful. Thus the following situation can arise. The statistician as expert witness testifies that there is no meaningful support for a difference in responses for the locations. Counsel then asks for confirmation that the results are not statistically significant. However, they are statistically significant. The statistician cannot deny this. However, the statistician has testified that there is no meaningful support. It could be argued that the statistician has strayed outside their area of expertise in their testimony.

There is a difference between meaningful support and statistical significance and it is one of interpretation. The difference is difficult for a statistical layman to understand. The difficulty cannot be resolved as an evaluation of evidence through the use of a likelihood ratio. It would not be helpful, for example, to summarise a result of no meaningful support with the statement that the evidence provides ... support for the proposition of no geographical variation *versus* the proposition of geographical variation. Fact-finders and

counsel will almost certainly have difficulty in understanding that statistical significance is necessary but not sufficient for meaningful significance. Note that the described study was conducted to investigate the variation of quantities of cocaine on banknotes across England and Wales. The results would then assist with the definition of the relevant population in order to assist with the calculation of the likelihood ratio to evaluate evidence in support of the prosecution or defence propositions in a particular case. The analysis of the results of the study does not involve likelihood ratios. The analysis is an investigation to determine the best-fitting model for the description of the underlying process that generates the results.

A resolution to the problem of how the difference between meaningful and statistical significance be described is provided by Weinbach (1989). First, it should be explained that statistical significance should not be confused with importance or meaningful support. Second, only practitioners and decision-makers, not statisticians, can make the final decision as to whether an association and its strength are meaningful. Interpretation of results needs input from a practitioner as well as a statistician. Thus in a criminal trial it is not the statistician that should testify to the lack of meaningful support but the practitioner. Alternatively, the difference can be presented in the report and the judge or jury can make the decision.

In many cases, the magnitude of the relationship can be reported rather then the significance probability. Perhaps the magnitude that would be meaningful can be decided before any data analysis is conducted. The statistician would then have a well-defined cut-off point. However, this approach is similar to the effect associated with significance testing known as the 'fall-off-the cliff' effect (Robertson *et al.*, 2016). If the difference is just one side of the point, meaningful support is declared; if the difference is just the other side of the cut-off point, no meaningful support is declared. The possible existence of a cut-off point is not the only problem that may arise with an attempt to determine a numerical solution to the determination of the level of

support that might be meaningful. A meaningful response is one that would be a difference in quantity, with perhaps an unusual pattern of contamination, sufficient to change the opinion of the analysts about how they would report the findings from a sample of notes. One objective of an analysis of quantitative measurements might be to provide a measure of support for one proposition over another using the likelihood ratio. In order for the difference to be meaningful the difference has to have a substantial impact on the value of the likelihood ratio (where substantial means a large enough impact to change the view of the judge or jury). A test as to whether a difference has a substantial impact would be to alter the measurements in the background database by that difference and observe the change in the likelihood ratio (or use alternative databases if these are available and note the changes in the likelihood ratio). Further discussion of the role of unusual patterns of responses in the decision as to whether a response is meaningful or not is beyond the scope of this paper. The purpose of this paper is consideration of the interpretation for the courts of a result which is statistically significant but not meaningful.

There are more general comments that are relevant to the role of statistics and probabilistic reasoning in the administration of criminal justice, exemplified by the difficulty described here of explaining the distinction between meaningful support and statistical significance, again based on ideas in Weinbach (1989). Professional judgement is essential to decision-making. A decision should not be delegated to the value of a significance probability or a meaningful response defined in advance of a study. However, care has to be taken that the absence of delegation is balanced against the requirement to be fair and balanced when assessing the results of a study. For example, if it had been defined in advance of a study what it meant for a result to be meaningful, it would not be right for an analyst to change their mind about what is meaningful having seen the results. The contribution of a statistician to the administration of criminal justice will have most impact when there is

a practical mix of statistical expertise, the expertise of the forensic scientist with whom the statistician is working, legal expertise and sound judgement. In the legal environment, when there is a need for an expert witness who applies their research to the case in hand, the expert witness, forensic scientist or statistician, should be sensitive to the decision-making needs of the lawyers, conduct their work for the case and communicate their findings in a way that can be used by practitioners. Practitioners in many criminal cases will need to interpret their results for jurors. By the nature of the selection process, these jurors will have a wide range of backgrounds. Communication needs to be sensitive to this range.

1.4 Conclusion

There is a problem with presentation of evidence in the courts over the difference between meaningful significance and statistical significance. For tests with large power it is possible for a test result to be statistically significant at a very low level and yet have no meaningful significance. The solution to the problem is, first, to explain that statistical significance should not be confused with importance or meaningful support. Second, only practitioners, not statisticians, can make the final decision as to whether an association and its strength are meaningful. Interpretation of results needs input from a practitioner as well as a statistician. Thus in a criminal trial it is not the statistician that should testify to the lack of meaningful support but the practitioner. Expert witness reports which discuss such differences should be submitted with dual authorship, that of the statistician and that of the practitioner. As always, a longer term solution is to include appropriate statistical training in the curricula of law schools and in courses in continuing professional development.

Education is crucial. Statisticians need to be educated in the ways of criminal law. Lawyers need to be educated in the ways of statistics and also

probabilistic reasoning, especially in an understanding of natural variation. Both lawyers and statisticians are concerned with decision-making under uncertainty. There is a good case to be put that all law schools should include a course in statistics in their curriculum. Conversely, it is not clear that all statistics degree programmes should have a law course in their curriculum. Statisticians who recognise sufficiently early in their career that they have an interest in the law should be encouraged to attend law courses.

The International Conference on Forensic Inference and Statistics has an important role to sustain communication amongst statisticians, forensic scientists and lawyers so as to ensure the highest quality of administration of justice as possible. It is a pleasure to note that on the occasion of the tenth conference that the conference is in a very healthy state and well-placed to fulfil this role.

Funding

This work was supported by the Leverhulme Trust, grant number EM2016-027, and the Swiss National Science Foundation, grant number BSSGI0_155809.

Acknowledgements

The authors are grateful to the very helpful comments from the reviewers. The support of the Leverhulme Trust, grant number EM2016-027, and the Swiss National Science Foundation, grant number BSSGI0_155809 is gratefully acknowledged. One of us (CGGA) is grateful to the National Institute of Science and Technology of the USA and another (AW) is grateful to the School of Mathematics, The University of Edinburgh, who funded their attendance at the 10th International Conference of Forensic Inference and Statistics hosted by South Dakota State University in Minneapolis, 6th to 8th September, 2017

References

Aitken, C.G.G., Wilson, A., Sleeman, R., Morgan, B., and Huish, J. (2017) Distribution of cocaine on banknotes in general circulation in england and wales. *Forensic Science International*, **270**, 261–266.

ENFSI (2015) European Network of Forensic Science Institutes guideline for evaluative reporting in forensic science, Strengthening the evaluation of forensic results across Europe (STEOFRAE), Dublin.

Moretti, T. and Budowle, B. (2017) Letter to the editor: Reiteration of the statistical basis of dna source attribution determinations in view of the attorney general's directive on 'reasonable scientific certainty' statements. *Journal of Forensic Sciences*, **62**, 1114–1115.

Robertson, B., Vignaux, G.A., and Berger, C.E.H. (2016) *Interpreting evidence*, John Wiley and Sons Ltd.

Weinbach, R.W. (1989) When is statistical significance meaningful? a practice perspective. *The Journal of Sociology*, **16**. Http://scholarworks.wmich.edu/jssw/vol16/iss1/4/.