STATISTACY: VOCABULARY AND HYPOTHESIS TESTING

Alan McLean
Monash University
Australia

In this paper I consider the characteristics of a statistically literate (a "statisticate") person. I suggest that a statisticate person should be able to read and understand statistical arguments of moderate complexity, and to carry out statistical analyses to some degree. Significantly, the truly statisticate person should also have developed the habit of thinking quantitatively. Furthermore, he or she does not rely on rigid rules to make statistical decisions, but uses informed judgment. In particular, he or she should understand the concepts of modelling and selection between models, and recognise their importance. Consideration is given to one of the major barriers to developing statistacy: the vocabulary used, in particular, the common use of two words that should only be used with the greatest of care (if used at all). These words are "prove" and "true". An important illustration of the way that vocabulary hinders the development of understanding is the case of hypothesis testing, a vital statistical tool that is widely misunderstood. It represents a mode of thought that is fundamental to statistical analysis, and so belongs in the kit bag of any statisticate person.

INTRODUCTION

Since there will be a need to refer to the general theme of this conference, it would be useful to have a word to replace "statistical literacy", which is ambiguous. I propose "statistacy" with the adjective "statisticate"**. (**I realise that there is an excess syllable in the latter, but it sounds much better!

What are the characteristics of a statisticate person? What is statistacy?

It certainly amounts to more than "numeracy", the ability to work with numbers, to think and reason in quantitative terms. We can identify:

- The ability to read and understand statistical arguments
- The ability to carry out statistical analyses

Each of which represents a range of abilities:

- From *read and understand simple arguments* to *read and understand complex arguments*
- From *carry out simple analyses* to *carry out complex analyses*

They are interrelated. To some extent at least, one must be able to *do* in order to be able to *read and understand*, and equally to some extent, one must be able to *read and understand* in order to be able to *do*.

I would like to add a third *quality* that appears to be significantly different from these two abilities. It is possible that this quality is determined genetically. It is possible that if a person acquires the two abilities referred to above, this quality is likely to appear. It is possible that by trying to inculcate this quality, the other abilities can be developed. This quality is not an ability but a *habit*.

- The habit of using quantitative critical thinking wherever it is relevant

My conception of a person who is statisticate is one who has these abilities to a reasonable degree, and has this habit.

DEGREES OF STATISTICACY

It is clear that varying degrees of statistacy are required for different roles in life. In an ideal world, journalists, market researchers, lawyers and the like would be highly statisticate. In theory, our role as statistical educators is to produce graduates who are suitably statisticate for their professions. On the whole I would have to say that our success is at best patchy.

One question of implicit importance to statistical educators is: Can we identify a minimal level of statistacy that could reasonably be expected of (the well educated) Joe/Jane Citizen? If so, what is it? For very many of us, the amount of statistics that we really expect our students to learn in a first (frequently the only) course they take is no more than a minimal level! The graduates from this first course disappear into the general throng of Joe and Jane Citizens. Five years later, how statistically literate are they? Perhaps we should lower our expectations?

Let us suppose that we assume "statistical literacy" amounts to *ability to read and understand basic statistical arguments*. Hopefully some tendency to think quantitatively will develop as well. In this case, the three questions I would like to raise are:

- What topics/items would we require?
- What are the major difficulties/barriers to acquiring this statistacy?
- What are the implications, if any, for general statistical education?

ITEMS OF UNDERSTANDING

I imagine there is a considerable amount of variation - how could there not be, among statisticians? - in lists of what *items of understanding* (topics, concepts) should fill the statisticate citizen's kit bag. Mine reflects my view of statistics (McLean, 1998 - 2001). It includes:

- Probability as a formalisation of a common mode of thought
- The development and use of probability models
- The predictive use of probability models
- The roles of the different methods of quantification of probability
- The concept of a population, particularly as a model
- The concept of random sampling from a notional population
- Random (unpredictable) variation and how it is modeled
- Random sample design, and reasons for it
- Prediction intervals for a variable
- The probability model for variation of a sample statistic over a population of samples
- Confidence intervals for a parameter
- Selection between models on the basis of sample data
- Hypothesis testing as a form of model selection
- Models to account for more variation
- More selection between models

I suspect that my list would be different from that of others primarily in the language used to describe it, and matters of emphasis. One difference that I would expect to see in some others' is the omission of any reference to hypothesis testing (although some might include it for historical interest!). I will return to this later.

VOCABULARY

There are many obstacles to people learning statistics, most of them I think reasonably well understood, or at least recognised. Getting over these obstacles is another matter! I would like to focus on one such obstacle, because it is of particular interest to me. This is the matter of *vocabulary*.

The development of concepts is intimately interlinked with the vocabulary used for them. You cannot learn concepts without the vocabulary, and words represent concepts. They are names for them. (I occasionally run into a student who cannot understand the difference between 'frame' and 'population'. This is the same difficulty as confusing a man and his name. The person Alan is not the same as the name 'Alan'.) Concepts are built on simpler concepts; one needs a word for each of those simpler concepts, you then build up the new concept – and then you give it a name.

Students notoriously have difficulties with learning the concepts of statistics – they do not understand the basic concepts well. Equivalently, this can be expressed as: students do not learn the basic vocabulary well, so have great trouble developing the more complex concepts. Most (basic) statistical vocabulary uses everyday words, usually with more specialised meanings. Students have to do two things – learn the specialised meaning, and at the same time remember its meaning in everyday life. Further, most of these basic concepts are superficially simple, but are in reality much more complex and/or subtle. For example, "population" in everyday language can mean either the set of people who live in a country, or the number of these people. In statistics the meaning is restricted to the set. Indeed, its meaning is also tied up with the idea of a group *of interest*.

More importantly, in practice the population is usually either notional or ill defined or both. For example, in a trial of a new medical treatment, people undergoing the treatment are considered to constitute a sample from a notional population of *people who undertake this treatment*. Random allocation of sample members to the treatment or to a control treatment is assumed to make the sample random. In a typical introductory regression example, relating house prices to house size, the population of *all 20 square houses* is notional. In market research, the population of *customers for product X* is very ill defined.

The result is that except in very rare artificial applications the population is part of a model of the world. There is an element of the abstract in the concept that is rarely developed; but if it is not developed, there is something missing in the student's understanding.

The word population may be used even more abstractly, for the set of all possible values of the variable, rather than the set of all entities (fictional or otherwise) that are measured. The concept of population is essential to statistacy, because all statistical theory is based on the idea of random sampling from a population, even when in reality there is clearly no such population. If the population is an abstract model, so is random sampling. If the population of interest is ill defined, one cannot take a truly random sample. One may be able to use a surrogate population, but this is effectively a model for the real population of interest. If the population is notional, the process of "sampling" is likewise notional. In short, the concepts of population and random sampling are much more complex and abstract *in everyday applications* than most recognise.

The notion of "variable" and how it relates to a characteristic (or attribute) of interest is likewise very complicated. It is the core of the whole idea of measurement. To be statisticate a person clearly needs to understand that measurement of characteristics is much more complex (and more subjective!) than appears at first sight.

TRUTH AND PROOF

Two words that are great impediments to statistacy are "prove" and "true" (in the sense of *true value of the population parameter.*) Everyone is familiar with the saying, "You can prove anything with statistics!" The truth, of course, is the reverse – "You can prove nothing with statistics". It is true that there is a semantic problem here. To mathematicians (which includes many statisticians) to prove a result means to present a *totally* convincing argument; to others it means merely "convince me!" Here are some dictionary definitions of the word as a transitive verb, drawn from the Internet:

Macquarie Dictionary (http://www.macnet.mq.edu.au/):

1. to establish the truth or genuineness of, as by evidence or argument: to prove one's contention.
2. (*Law*) to establish the authenticity or validity of (a will or testament).
3. to give demonstration of by action.
4. to put to the test; try or test.
5. to show (oneself) to be as specified.
6. to determine the characteristics of by scientific analysis: to prove ore.

Infoplease Dictionary (http://ln.infoplease.com/dictionary.html):

1. to establish the truth or genuineness of, as by evidence or argument: to prove one's claim.
2. (*Law*) to establish the authenticity or validity of (a will); probate.
3. to give demonstration of by action.
4. to subject to a test, experiment, comparison, analysis, or the like, to determine quality, amount, acceptability, characteristics, etc.: to prove ore.
5. to show (oneself) to have the character or ability expected of one, esp. through one's actions.
6. (*Math*) to verify the correctness or validity of by mathematical demonstration or arithmetical proof.

Merriam-Webster (http://www.M-W.com/):

2a: to test the truth, validity, or genuineness of <the exception proves the rule> <prove a will at probate>
2b: to test the worth or quality of; specifically: to compare against a standard
2c: to check the correctness of (as an arithmetic result)
3a: to establish the existence, truth, or validity of (as by evidence or logic) <prove a theorem> <the charges were never proved in court>
3b: to demonstrate as having a particular quality or worth <the vaccine has been proven effective after years of tests> <proved herself a great actress>
4: to show (oneself) to be worthy or capable <eager to prove myself in the new job>

These definitions illustrate (apart from some cross over between sources) the variety of meanings, and the variety of nuances in meaning of the word. The meaning of interest here is probably the most usual – **"**to establish the truth or genuineness of, as by evidence or argument".

To "establish the truth" requires first that "truth" exists, second that it can be established, and third that it can be shown to be established. In absolute terms, these three conditions exist only in the artificial world of mathematics. As I have argued elsewhere (McLean 2000a) we work with models of the real world, particularly in statistics, where the models are probabilistic models. Within those models absolute statements can be made, but these statements are about *reality* only to the extent that the model reflects the real world – and that is a matter of judgement, sometimes guesswork.

Thus "proof" is absolute only in the mathematical sense. In general – and this includes the law – proof is not absolute. Nor for that matter is "truth". A statement may be considered "proved" if the evidence for it is strong enough that people accept it as true, in the sense of being a satisfactory basis for action. So if a person is "proved guilty of murder", it means that the evidence is strong enough that it is reasonable to commit the person to gaol or to the death penalty. "Proof" in this case means "proof beyond reasonable doubt". It does not mean "beyond doubt". In short, to "prove" a statement means "convince me to the extent that I believe it!"

Acceptance of the strength of evidence is subjective; so one person may consider a statement proved while her next-door neighbour thinks not. All that a statistical analysis of any sort does is to provide evidence. The evidence is to some extent objective – never totally, and sometimes not very much at all. If the statistical evidence is sufficiently strong, the reader may accept the result as "proved" in this weak sense. In particular, this argument applies to hypothesis testing.

VOCABULARY AND HYPOTHESIS TESTING.

There can be little doubt that a statisticate person must have some knowledge and understanding of hypothesis testing. Even if one does not agree with its use in research, one must take cognisance of the fact that hypothesis testing exists, is historically important and is very commonly used, as indeed it should be.

Understanding hypothesis testing as a mode of thought and argument is important. It should be understood that it is, first, the statistical equivalent of the scientific method; second, both hypothesis testing and the scientific approach are formalised developments of an everyday method of thought. I have made this point earlier (McLean 2001). Briefly, if a statement is made, for example in conversation, one automatically checks it with "sample data", in the form of past experience. The innovation of science is to introduce as great a degree of objectivity as possible into the sample data used.

A properly statisticate person will also fully understand what they are doing when they carry out a hypothesis test. This is not as common as one would hope. Too many people have learnt to carry out a test following a recipe, and to regard the test as a routine that will automatically – and conveniently - give a correct answer. (Students notoriously want rigid decision rules, so have difficulty with an approach which says, essentially: Use your judgment.)

The tendency to view a test in this mechanical way is emphasised in the minds of many people through their perception that statistics is mathematics. In hypothesis testing, nothing could be further from the truth. Certainly mathematical techniques are used as tools, and often very complex tools, but the situation is the same as in architecture, where mathematical calculations help in the design of the building.

The tendency towards this view is in my mind exacerbated by the vocabulary commonly used. To talk of true parameter values, of proof and decision rules and 1% or 5% significance levels can be very misleading. Briefly, a hypothesis test is a method of selecting between two probability models on the basis of sample data in certain circumstances. These circumstances are as follows: the two models are in some sense complements - there is a strand of research on tests in which the two models are not complements; my feeling is that the hypothesis testing approach is not appropriate to this model selection situation - the consequences of choosing one model differ from those of choosing the other; one of the models (the null model) is preselected on some external basis such as simplicity, cost or ethical considerations; the alternative model will be chosen only if the sample data support it sufficiently strongly.

The sample data always support the alternative model – if they did not, no test would be necessary, since the null model would automatically be chosen. The test provides a measure of the strength of the evidence in favour of the alternative model relative to the null model. The commonly used measure of the strength of the evidence in favour of the alternative model is the *p* value – the probability of obtaining the sample result (or more extreme) when using the null model. This approach to hypothesis testing is discussed at greater length in McLean (2000b).

There are two general areas in which hypothesis testing is used. One could be called production statistics, and includes quality control and the like; the other is general research. The distinction between these is quite critical. Broadly speaking, these two areas reflect the Neyman-Pearson and Fisher approaches to hypothesis testing. The first situation is exemplified by sampling to test if a process is in control or not. This is quite clearly a decision process. If it is decided that the process is (probably) in control, no action will be taken; if it is decided that the process is probably out of control, some action will be taken to correct the situation. It is preferable on grounds of cost that the action should not be taken, so the preferred null model will be accepted unless evidence for the alternative model is quite convincing. In this context the concepts of Type I and II errors make some sense, so it is appropriate to set this test up as a decision rule with a chosen significance level such as 5%: the sample evidence is identified as “strong enough” if the *p* value is less than this chosen level. In this type of situation, a decision has to be made – you have to proceed on the basis of one of the models. One of the models is “accepted” and acted upon. Thus there is a mechanical aspect to the test.

In the research situation a decision does have to be made, but it is more tentative. One does either accept or reject the null model – the phrase “fail to reject” is quite misleading – but it must always be born in mind that acceptance or rejection is tentative. Further studies may contradict the conclusion. Thus the emphasis has to be on the use of judgement in accepting the *p* value as a measure of the strength of the evidence. In this situation the more or less mechanical full-blown decision theory approach is inappropriate. Certainly the task is to decide which of the two models to choose, but it is certainly incorrect to use a rigid significance level rule to decide – the notorious “if $p < 0.05$, publish!”

The use of terminology such as “the true value of the mean” leads to such statements as “The null hypothesis is almost always false”. While this is true, it is irrelevant. Hypothesis testing is not directly concerned with the true value, but with models that are expected to work. To accept the null model means only that it is worthwhile using a model based on the specified value of the parameter; to reject it means that is preferable to use a model based on a different value. This does not necessarily mean even that the “true value”, if it could be known, is approximately equal to the specified value, but it is probably useful to express the idea in this way.

As can probably be guessed, I prefer to use the phrase “null model” rather than “null hypothesis”. I do this mainly to emphasise that we are comparing models, but also because it is strictly correct. The null hypothesis is that the null model can validly be used. The final point I wish to make about hypothesis testing is that in a research project a test is only used to help decide on matters internal to the statistical analysis. Furthermore, it is only one of a number of tools that can be used. That is, its usage is by definition limited.

JUDGEMENT

One of the characteristics of the statisticate person must be the ability to use judgement, rather than rigid rules, to make decisions. The role of statistics is to provide evidence for and against, with methods for assessing the strength of that evidence. Hypothesis testing is one form of this, which can be automated for production purposes, but generally should not. With this observation in mind, one of the characteristics of teaching statistics should be the careful avoidance of such rules as: "If $n > 30$, the sample mean is normal (and if $n < 30$ it is not)" - this particular rule is also confused with the choice between the normal and *t* distributions, so is doubly to be avoided - and "If $p < 5\%$ reject the null".

CONCLUSION

In this paper I have presented the following ideas:

1. A statisticate person should be able to read and understand statistical arguments of moderate complexity, and has developed some tendency to think quantitatively.
2. A list of items of understanding should include the concepts of modelling and selection between models.
3. A major barrier to developing statistacy is the vocabulary used. Two words that should only be used with the greatest of care are "prove" and "true".
4. Hypothesis testing is a useful statistical tool that is widely misunderstood. It represents a mode of thought that is fundamental to statistical analysis, and so belongs in the kit bag of any statisticate person.
5. The most important characteristic of the statisticate person is that he or she does not rely on rigid rules to make statistical decisions, but uses informed judgment.

REFERENCES

McLean, A.L. (1998), The forecasting voice: A unified approach to teaching statistics, In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, and W.K. Wong (Eds.), *Proceedings of the 5th International Conference on Teaching of Statistics* (pp.1193-1199). Singapore: International Association for Statistical Education and International Statistical Institute.

McLean, A.L. (1999). Hypothesis testing and the Westminster system. In *Proceedings of the 52nd Session of the International Statistical Institute*, *Contributed Papers* (Book 3, pp. 287-288). Helsinki, Finland: International Statistical Institute.

McLean, A.L. (2000a). The predictive approach to teaching statistics. In *Journal of Statistics Education*, *8*(3).

McLean, A.L. (2000b). *On the nature and role of hypothesis tests*. Department of Econometrics and Business Statistics Working Paper 4/2001. Available at http://www.buseco.monash.edu.au/Depts/EBS/

McLean, A.L. (2001). Statistics on the catwalk – the importance of models in training researchers in statistics. In C. Batanero (Ed.), *Training researchers in the use of statistics.* Granada, Spain. International Association for Statistical Education.