Objectivity, Subjectivity and Statistical Evidence

Michael EVANS

Statistical Society of Canada / Société statistique du Canada
Department of Statistics, University of Toronto, Canada

Statistics has applications in many fields. The point behind all of these applications is that there are questions for which there is no obvious way that we can obtain definitive answers. The reason for this lies in variation, which can arise for many reasons, and this leads to uncertainty. The health sciences provides an excellent example of this as the variation among patients, such as physical, genetic and lifestyle characteristics, lead to different responses to a treatment for a health problem. We are then left with the questions of whether or not a treatment works and, if so, how well.

To be a problem for which statistical methodology is applicable, this variation must exhibit some regularity which we can model. The archetypal example of this is a population Ω, perhaps consisting of all individuals suffering from a particular disease, and a measurement X which is measured for each member ω of Ω. For example, $X(\omega)$ could be the blood pressure, measured in appropriate units, of individual ω in Ω. The population Ω and the measurement X lead to a distribution of the characteristic over the population as given by f_X, where $f_X(x)$ is the proportion of individuals in Ω that have $X(\omega) = x$ and we record this for each possible value x of X. So in a statistical application we want to know the distribution f_X. If we can conduct a census, namely, obtain $X(\omega)$ for every ω in Ω, then we know f_X exactly and there is no need for statistics.

In general, however, we will not be able to conduct a census and so we cannot know f_X exactly. For example, suppose we consider distributions $f_{1,X}$ and $f_{2,X}$ of measurement X over where $f_{i,X}$ corresponds to giving each member of treatment i where this is supposed to result in a lowering of blood pressure. If treatment 1 corresponds to a standard, we might want to know if treatment 2 is different in the sense that $f_{1,X}$ and $f_{2,X}$ are materially different. Clearly, even if we could do a census, we cannot simultaneously know both distributions. The statistical solution to this problem is to select a subset from Ω say $\omega_1, \ldots, \omega_n$, apply treatment 1 to $\omega_1, \ldots, \omega_{n1}$ and apply treatment 2 to $\omega_{n1+1}, \ldots, \omega_n$. After measuring these individuals, we have a sample $x_1 = X(\omega_1), \ldots, x_{n2} = X(\omega_{n1})$ from $f_{1,X}$ and a sample $x_{n1+1} = X(\omega_{n1+1}), \ldots, x_n = X(\omega_n)$ from $f_{2,X}$. We will use these data to make inferences about differences between $f_{1,X}$ and $f_{2,X}$, which are correspondingly inferences about differences in the treatments.

An important point here concerns how we should select $\omega_1, \ldots, \omega_n$ from Ω and the answer from statistics is unambiguous: we should use a random mechanism. Various reasons can be put forward for this but the most compelling for me is that it guarantees the *objectivity* of the data, namely, the data were generated by a mechanism which the investigators had no way of controlling. This is an important contribution from statistics.

The process we have described for generating data represents a gold standard. We do our very best to achieve it in any application. It is well-known, however, that it is rarely achieved in its ideal form. For example, we may not be able to sample from the full population Ω but have to rely on local participants or, even worse, it may be that the data we use is the result of an observational study where no known random mechanism was applied to generate the data. In such circumstances qualifications have to be applied to any conclusions we reach based on a statistical analysis. For example, while we would like our conclusions to apply to the full population Ω the fact that the data was not generated by random sampling from Ω means that our conclusions really don't apply that broadly. This doesn't mean that the results should be summarily dismissed as useless, only that we must be wary of any conclusions drawn. We can still consider the results of an analysis as evidence concerning the questions of interest but just not at the highest level of evidence that we would have obtained through proper random sampling.

The important point here is that statistics works via establishing gold standards and in any applied statistical analysis we strive hard to reach this standard as closely as possible to present the highest possible form of evidence. Any consumer of a statistical analysis must, however, ask themselves, what could have gone wrong because of any deficiencies in the way the data were collected. While the gold standard for the data collection phase of a statistical analysis is fairly easy to establish, and it is one to which most statisticians adhere to, the statistical inference or analysis phase is more problematical. There are a wide variety of opinions about the appropriate approach to statistical inference and unfortunately different approaches sometimes conflict in the sense that they give contradictory answers. This is a fundamental ambiguity that statistics has yet to collectively resolve.

Perhaps the most well known point of contention about inference is the Bayesian versus frequentist argument. There are many variations of this, and this short essay can't delve into all of them. The essence of the debate, however, can be summarized by saying that, while the Bayesian approach acknowledges the *subjectivity* in a statistical analysis, and even tries to make a virtue of it, a frequentist make claims of objectivity for their inferences based upon the long-run behavior of these procedures. For example, a 95% confidence interval for an unknown characteristic of the distribution of X will cover the true value of the characteristic in 95% of future samples of the same size we imagine taking from the same population.

While the frequentist criterion seems reasonable, there are problems with this approach to such an extent that the writer of this essay is a Bayesian. There are several reasons for this.

Perhaps foremost concerns the reason statistics exists as a subject. At least for me, statistics exists to tell people how one is to reason in statistical contexts. The archetypal statistical context is just as described in the first few paragraphs. A theory that fails to tell us exactly how we are to reason in very simple situations like this cannot, in my view, be seen as an acceptable theory of statistical inference. And yet this is the case for frequentist approaches to statistical inference as there does not exist a compete theory of frequentist inference, free of ambiguities.

One could argue that further research could one day fill in the gaps in a way that most statisticians find acceptable. This is certainly possible, but a detailed study of the problem does not make me optimistic.

Another concern, with various approaches to statistical inference, lies with any claim of objectivity of the analysis. In reality, all statistical analyses depend on choices made by the analyst either explicitly or implicitly. For example, why do we often assume (choose) normality as a possible distribution of a quantitative variable X? Whenever a statistical analysis is dependent on a choice like this, it is inherently subjective as the conclusions are dependent on the option chosen. This is not necessarily bad as it is often the case that we can check such choices against the only truly objective part of a statistical analysis, at least when it is collected correctly, the data. Indeed in a frequentist analysis we can check the sampling model against the data to see if the model is reasonable. This is known as *model checking*. Even if the model passes its checks, however, this does not make the sampling model, or the inferences drawn from it, objective. We have only followed good scientific practice to see if our assumption is in a sense falsified by the data. There is a logical concomitant to this: we should not use ingredients in statistical analyses that can't be falsified by the data. This eliminates a number of ingredients commonly used in statistical analyses such as loss functions.

It is not well understood that the most controversial aspect of a proper Bayesian analysis, namely, the prior, which expresses beliefs about the true f_X, can be checked against the data. A 'falsified' prior is one where there is an indication that the truth lies in the tails of the prior. If such a prior has a big impact on the analysis, then surely we wouldn't want to use it just as we wouldn't use a sampling model where the data lay in the tails of every distribution in the model. A relevant reference for *checking for prior-data conflict* can be found at the following link.

http://ba.stat.cmu.edu/journal/2006/vol01/issue04/evans.pdf

Once we have a sampling model, a prior and the data, then it is possible to provide a measure of *statistical evidence* that leads to a complete theory of statistical inference. Details on this can be found at the following link.

http://ba.stat.cmu.edu/journal/2013/vol08/issue03/evans.pdf

All the ingredients used in this theory of statistical inference are falsifiable and no inference problems are left unanswered. Of course, this does not mean that all statistical problems are solved. In specific problems we still have to come up with relevant models, elicitation procedures for priors, and implement model checking, checking for prior-data conflict and inference based upon the measure of statistical evidence.

Beyond the data, statistical analyses are never objective as they are dependent on subjective choices made by the investigator. Perhaps this is true of all empirical scientific investigations. Part of the role of statistics is to tell us how to assess the effects and relevance of our subjective choices and, most importantly, give us a complete and logical approach to reasoning in statistical contexts.