Data castle building, *or an official statistician looks askance at Data Science*

The mission of the Australian Bureau of Statistics is to improve decision-making in the public sphere - or something similar. And it was ever thus, from the first time that people in power began counting their subjects, and their subjects' wealth, for the purposes of pursuing foreign wars, or taxing, building roads or monuments to glorify their splendour, or latterly to address needs of the population, or otherwise pursue policies of government.

From this, statistics as an activity was born. However, the founding of the profession of statistics as a discipline rests on observations concerning variability and chance, on having incomplete knowledge of outcomes, and on the notion of quantity of information conveyed by statements concerning measures or numbers. The convergence of these otherwise non-communicating spheres of thought and meaning, created the bounded field of endeavour that modern statisticians call home.

The organised collection of data - that is readings of a population process - has been a preoccupation of official statistics leading into the machine tabulation and computation age. The first computers were employees of statistical offices who transferred systematically collected data onto worksheets, from thence to published tables. With mechanical, then electrical and finally electronic computation, human computers gave way to statistical programmers, database administrators and managers, who turned discrete collection effort into systems for design, collection, processing and dissemination.

Data points supplied by informants become information concerning a population. The defining characteristic of official statistics however returns to the grounding of such authority in the stochastic distribution of recorded individual characteristics, and concurrently the measurability of the variation attached to this unitary response.

Other branches of statistics can show analogous constructions: combining design, data handling, conceptualisation and underlying phenomenal processes about which typically inference is required. In each case there is a mix of prior knowledge of the process - for instance, embedded in some scientific corpus or specification, body of data yielded, and post facto conclusions regarding any hypothesis built into the design.

What then is the data science world? One that came into contention when unit autonomy is modelled by a string of digits taking values characteristic of qualities abstracted from the unit. The datum is a realisation of a multivariate random variable, specific to the unit.

This specificity is the analogy of the autonomy of identity of a population member. A plurality of datums becomes a database, sitting somewhere in a memory machine within a computing engine. Any one electronic datum, as for unit records in a statistical database, is an observation from a process, insignificant in itself but powerful when bundled, contributing to patterns of correlation linking process values.

A data scientist explores these patterns, looking for meaningful connections and associations. Statisticians had been there before, optimising fit of data to distributions mathematically, or, in practical settings, using approximation techniques from applied mathematics.

Having put my cards on the table, I list the grey areas in the discussion, where misunderstandings seem to be compounding:

Volume

Big data is effectively cheap data, often data scraped from administrative or commercial procedures, given potency through originating from a real-world real-time transaction. However the data is accumulated and commodified, it rests resolutely on something that happened, tagged with whatever is gathered in the course of the transaction. It is neither inevitable nor necessarily useful. It is not a substitute for controlled observation. Yet it cannot be ignored

Model

“Something that works like something else, usually more malleable or visible or accessible.” In the present context models are used to describe patterns made by ensembles of otherwise autonomous units whose characteristics are captured in a database. These models can be used to estimate population quantities (or ‘statistics’) or to describe the hypothetical behaviour of units of the population on the basis of their profile. In non-statistical contexts models built out of scientific observation fashion the language of decision-makers - as evidence-based policy for instance. In a statistical context models have been used to refine or narrow hypotheses - in Bayesian analysis for instance. And deeper into mathematics, a model is a tool in defining the truth of a proposition. In the popular mind a model makes sense of something more complicated or esoteric, says something about something else, without all its often inconvenient qualities (think models of the cosmos).

In applying computer science to policy questions in direct competition with statistics, data science is modelling population variability using a static representation, and arguing that patterns in this representation correspond to latent patterns or laws in the population

Variability

What is true in one case, may not be in the next, other things being equal. If this were not so we would be wasting our time, as either data scientists or statisticians. We deal with indeterminacy. As of an engineering disposition (in either role) this is unnerving and has driven advances in our respective disciplines. We demand pattern, and can find pattern in the midst of randomness because of the laws of plurality, from the Central Limit Theorem, to the fundamental theorem of stochastic processes.

For statistics, we use the term error, not in the sense of a mistake, but rather as the deviance from predictability, the premium we allow in handling imperfect information. Our datums are not fully known, even the values recorded may be misleading or wrong. For data scientists variability at the level of a single transaction is washed out in the whole.

Sample

Sampling theory arose in official statistics to reduce the cost and increase the efficiency of collection. Evolving understanding of sample design could extend the coverage of official collections at no sacrifice in authority. With the advent of large statistical datasets - such as generated in the course of population censuses, or administrative byproduct panel datasets - new theories have been developed to strengthen sample survey inference. These run counter to the scenario in data science, whereby masses of raw data are assimilated without intervention or selection, any prior editing by a third party is anathema.

Chance

Statistics exploits chance; data science shuns it. Or so it seems to me. By allowing chance to play a role in the variability we are observing, a statistician can work through competing models, and reduce the spread of data to a signal family and finally a signal. By excluding chance a data scientist leans on the integrity of the data generator, and volume, to reveal all signal, and treat the data mass itself as simultaneous.

Algorithms take over the functions of chance, in a rapid learning sequence, passing as artificial intelligence in some contexts.

Algorithms

Originating in computer science as navigational programs, algorithms have been reinterpreted in statistics: in the first instance in developments in statistical computing - how to manage numerical approximations where analytical representation is not available - and secondly as a flag bearer for new methods for integrating diverse sources of data, under new demands and constraints, but

balancing what can be controlled through collection design with what is available outside of design that can lend weight to inferences.

Analysis

Unfortunately, analysis has become confused with modelling. Fitting a regression is not the same as an analysis of variance. A modelled hypothesis can be tested, and the result forms an analysis of the data. The two need to be conceptually separated in this debate.

Prediction and forecasting

As for causation, both are more or less outside the remit of statistics, but well covered in data science. This leaves ample room for statistics to be useful in exploring the dynamics of process, and providing envelopes of confidence.

Causation

A statistical analysis can identify factors accounting for variability in response. It can explore various hypotheses bearing on the direction of causation, and it can examine correlations. Little in statistical literature addresses causation as such; this is left to ‘subject experts’. That is only to underline the close working relationship between data experts (that is owners or contributors of primary information), and (statistical) data analysis.

Population

This is a divisive topic within statistics - are they fixed and large and unknowable; or are they random instantiations from a collection of all possible populations. Outside statistics, I imagine no one really cares. But it remains a distinction that data scientists ignore at their peril. Where are we drawing our data from? Is incomplete knowledge about everyone balanced by more complete knowledge of only some?

S.Horn 15 July 2022

(the beginnings of a) Reading List

Richard v Mises, *Probability, Statistics and Truth* (1957 translation of 1939 original German edition, 3rd German edition from 1950) Dover
Alain Desrosieres, *La politique des grands nombres, histoire de la raison statistiqu*e, la Decoverte, 1993
Bruce Hand, *3 lenses of public policy* 2010?
Olivier Kempf and Bruno Teboul, *La Donnee nést pas donnee-strategie et big data, seminaire sous la direction de Phillippe Davadie,*, edtns kawa, CESG 2015

Note: the last gives a tour of the horizon for using new data sources in public statistics: “*this publication inquires into the sense of data according to the various*

disciplines - economics, information science, philosophy, then its role in the numerical space. It goes on to describe different strategies for using data be it in the private or public sectors. It is a French perspective, with little reference to statistics.

On the other hand, Desrosiere's book is *a classic bringing together different domains never before connected in the history of science and politics: it retraces at the same time the history of the State, statistics, offices of administration and modelisation of the economy… To the extent that the laws of large numbers inspire games of chance, risks of vaccination, life assurance, the fixing of tariffs, the decisions of juries, the catastrophic effects of economic cycles and opinion polls*… or so says the blurb. The author was a senior executive at INSEE, and historian of science. There may well be equivalent books in English, but I have yet to come across them. The literature tends to be strictly statistics or strictly policy/ economics, bridged in the 1970s-80s by literature coming out of the social indicators movement