J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

STARTER

Statistics is What Statistics Does

Anil Gore*

Abstract | This paper provides background to articles in the special issue. It is aimed at readers who are not experts in statistics and may not be familiar with sequence of developments in topics covered. Reader will find very non-technical material about the role of statistics in genetics, health, evolution, survival analysis, etc.

Statistics can be regarded as a particular kind of inferential logic. It helps us jump from known to unknown, from sample to universe, from effect to cause.

Typical scientific propositions help us identify the effects if we know the cause that is operational. But what about going backwards? A hole in the ground could be due to a bomb or a meteor or a half-done job by a farmer. Can we guess which is the cause in the particular case? We know that meteorites are an extremely rare event. A bomb explosion is also rather unlikely, in times of peace. So, our prior suspicion may be that a farmer wants to plant a tree. If the newspaper reports finding of some unexploded bomb in a nearby village, will our judgment change? Of course. Among highbrow statisticians this is called Bayesian logic. So, to put it simply, we have a prior judgment about likelihood of some event, as we go on getting more evidence, we keep revising this judgment, the resultant being called a posterior judgment. If I am thinking about height of an adult Indian male, then it is more appropriate to talk about prior distribution which would represent my judgment. The prior for me may be a normal distribution with mean five and a half feet (pardon my laziness, I am yet to migrate to centimeters) and a coefficient of variation of say 15%. If I know that the person is a basketball player, I will revise my prior to a normal distribution with mean 6 feet. If I have to think about a population that I am not familiar with (say Australian aboriginals) I may be hard pressed to come up with a prior distribution. Then if I find data on heights of 50 such individuals reported in an anthropology journal, I can use that as a prior. In that case, I become a practitioner of 'Empirical Bayes' approach. Selection of priors can be controversial. In pre-independence India, police routinely assumed that a robbery is very likely attributable to a 'criminal' tribe camping outside the town. In modern USA, police are repeatedly accused of assuming that culprit is a black American. An empirical Bayes approach seems to fail to correct these priors (prejudices). Court related software like COMPAS[6] which bases all decisions strictly on empirical grounds and data fed to the software does not include information on ethnicity, still ends up in being discriminatory. Such biases can be corrected using non-informative/uniform priors. While majority of statisticians continue to use the frequentist approach (not using prior beliefs), Bayesian approach is making increasing inroads. It is therefore in the fitness of things that there are multiple review articles in the issue highlighting different features of Bayesian methodology. Bayesian statisticians come in a variety. Savage Bayesians are those who point out that frequentist interpretation of probability is very unsatisfactory and there is no alternative to Bayesian approach. More pragmatic variety accepts a blending of the two approaches. Rubin seems to have suggested (see the article by Kazuo Shigemasu) as a via media, use of frequentist approach to evaluate long-term behavior of Bayesian solutions. But everyone has to accept that in some situations Bayesian approach yields a better solution. One such case is the problem of estimating the mean vector of a multivariate (at least three variates) normal distribution. In this case the intuitive frequentist estimator namely sample mean vector is inadmissible, in the sense that the so-called James–Stein estimator is always better than the mean vector. The article by Rasines and Young gives many details of this fascinating story.

[1] Cytel Statistical Software and Services Pvt. Ltd., Pune, India.
*Anil.Gore@cytel.com

The subject of statistics has (especially in social science) two faces. Statistics as information and statistics as inference. Kautilya's Arthashastra,[1] a 2000-year-old monograph about conduct of state, gives detailed advice about collection and use of statistics related to the economy. But it also raises the issue of how to judge if a tax remittance from a transaction is too small. In common parlance the aspect of statistics as information gets greater emphasis whereas in scientific writing it is the inferential use that receives more attention. A third aspect that deserves to receive a lot more attention than it does is that of causality. Traditional statisticians drop the subject like a hot potato and say that correlation is not causation. A small group of statisticians argue that we should emphasize, not causes of effects but measurement of effects of causes and suggest ways of doing that.

Perhaps this deserves further explication. We have a population of interest and there are two variables x and y measured on each unit of the population. If y is the variable of interest, say performance in an examination and x is another variable, say daily hours of study, traditional statisticians readily offer sample correlation coefficient as a measure of association, discuss its sampling distribution and ways of inferring about population value of that coefficient from the sample value. The question of interest is to what extent y (score in the exam) is 'caused' by x (effort put in). It is of course implicit that 'cause' precedes 'effect', in the sense that definition of variable x has a time aspect (effort before the exam). To simplify matters, suppose effort is at two levels (low and high). Further suppose we can measure y for the same subject at low and high levels of x. Then we can say that the difference y(low effort)-y(high effort) is 'caused' by x, effort. Unfortunately, of these two values only one is observable. This seemingly insurmountable problem has a nice statistical solution. While individual value of the difference y(low effort)-y(high effort) is unobservable, what about the expected value (population average) of this difference? Note that expected value of difference is equal to difference of expected values. Further the latter is measurable. Expected value of y(low effort) is simply the average of y values for all subjects who invested low effort. This is not a sleight of hand. I invite you to delve into the article by Shigemasu for a deeper discussion of this exciting topic.

Bhagavat Geeta in chapter 2 verse 27 states an eternal truth: jatasya hi dhruvo mrutyu.

Everyone born is destined to die. But when? That is uncertain. Perhaps for the first time in history, in England, John Graunt developed in seventeenth century a life table giving probabilities of survival at each age.[4]. Now, the life insurance business is based such probabilities. Graunt's book 'Natural and Political Observations Made upon the Bills of Mortality' (published 1662) analyzed data available from the Bills of Mortality. Graunt, calculating with the Rule of Three and using ratios obtained by comparing years in the Bills of Mortality, was able to estimate the size of the population of London and England, birth rates and mortality rates of males and females, and the rise and spread of certain diseases. That is why he is regarded as a pioneer mathematical demographer and epidemiologist.

Link between fertility rates and mortality rates of humans is interesting. In medieval societies, both rates were high. Then developments in microbiology and medicine caused a major fall in death rates (due to control over acute diseases) while birth rates remained the same. This led to a huge population growth. Then came many technologies of family planning and avoiding pregnancy. A gradual fall in birth rates was recorded. This so-called demographic transition begs a clear explanation. Partly it is due to developments in science and technology but partly due to social attitudes. Economic development is the best (contraceptive) pill, they said. For planners this was a critical matter. Naturally, there has been considerable effort to develop stochastic models of fertility. Statistics group in the Banaras Hindu University has committed a long-term effort to develop such models. In his article, R C Yadav affords us a peep into this journey through the eyes of one member of the group.

Study of the distribution of length of life faces a major hurdle. It is that there can be systemic difficulties in recording very low or very high values. In astronomy, an optical telescope may have a minimum level of light intensity below which it fails to notice the source. So, presence of sources with very dim light is censored (not that there is anyone stopping us from observing them, but our tools have limits). In case of a cricket match that is stopped by rain, consider the runs scored by a batsman who is 'not out'. He could have scored more if rain had not interrupted. How much more? We don't know. In case of people the time to death is very long and it is very difficult to keep the observation system on alert for all that time. As a special case of course, we can arrange it. As an example, The Centennial

Light[5] is the light bulb, continuously burning since 1901, located at 4550 East Avenue, Livermore, California, and overseen by the Livermore-Pleasanton Fire Department. Hopefully, it will be possible to know the precise duration for which the bulb functions before it breaks down. But this is an exception. Most of us mortals with limited resources have to live with the so-called censored data in which the act of observation stops before the person dies and all we know is that the length of life was higher than the recorded time. Another reason for censoring is that the subject under observation drops out of the study or dies due to a road accident or some such cause unrelated to the study. If I did not have this problem of censoring, I would have perhaps regressed the length of life upon values of potential causes to figure out the impact of each cause upon length of life. With censoring this regression approach gets tricky. One way out is using hazard rate. This is a one-to-one-onto mapping of the distribution of length of life. Simply put, it is the instantaneous chance of death at a point of time given that a person has survived up to that age. One can use censored data quite effectively while estimating the hazard rate. Shape of a hazard function can reveal some interesting properties of the variable of interest namely length of life. A constant hazard function suggests death due to random accident. A rising hazard rate of course reflects aging. In natural populations, one expects a bathtub shaped hazard which falls initially in infancy, remains flat during adult life, and rises during old age. Clearly, relating hazard rate to potential causal factors can be useful. Fifty years ago, D R Cox proposed that we express hazard rate as a product—reference hazard rate multiplied by a correction factor involving causal variables. If we do that, ratio of hazard rate of interest and a reference hazard rate can be written as a linear function (on log scale) of causal variables. Bingo! No wonder this approach has found very extensive use. I draw your attention to the article by P K Andersen for a review of the exciting history of this tool.

Statistics is the discipline that tries to understand and manage variation. Genetic variation is good because it allows evolution. Variation in manufactured goods is undesirable and must be reduced. Sometimes just anticipation of variation can be of use for planning or action. Another important facet of statistics is as a tool to judge if variation of ground reality from a theoretical expectation is so large as to discredit the theory. If no theory exists and the phenomenon is very complex involving many potentially causal variables, one can try to bring some order into the madness by fitting regression models. One can begin by characterizing variation using tools such as a probability distribution.

If we regard Gregor Mendel's experiments in the nineteenth century, on height of sweet pea plants as the beginning of modern genetics, publication in 1900 of Karl Pearson's goodness of fit test that examines discrepancy between observed counts and counts expected under a theory can be regarded as birth of modern inferential statistics. For a single locus with two alleles and simple dominance, in a segregating F2 generation, Mendelian logic predicted a ratio of 3:1. For two loci the prediction becomes 9:3:3:1. These predictions can very well be checked using the goodness of fit chi-square test. In case of Mendel's data, the overall discrepancy seemed very small (the value of chi square was quite low) and hence seemed to support the theory. The logic here is that if the theory is valid, very large discrepancy has a very low probability. Combined with the unstated assumption that very unlikely things do not happen, the test rejects a theory if the value of chi square is too large. Otherwise, it 'accepts' the theory (or suggests that the data do not contradict it). R A Fisher later turned this argument around. His point was that if a very large discrepancy is unlikely for a theory that is correct, so also is a very small discrepancy. In other words, very close matching between observed and expected counts should be regarded as 'too good to be true'. Perhaps Mendel's assistants doctored the data to give their mentor what he hoped to see.[5].

Another related aspect that linked genetics with statistics is the issue of discrete and continuous traits. Karl Pearson, Francis Galton, and others were interested in heritability of continuous traits such as human stature and crop yields and not in discrete traits such as eye color. Hence a theory meant to explain variation in discrete traits seemed less exciting. They developed statistical tools such as correlation and regression which could handle observations on continuous traits. Again, R A Fisher demonstrated[3]that there is no essential difference between the two approaches. All you need to do is to postulate that a continuous trait is a manifestation of many genes working together.

Parallel to Mendelian genetics, the other upheaval taking place in nineteenth century biology was Charles Darwin's theory of evolution through natural selection. How do we link these two? Think of evolution as change in gene frequencies (proportions) across generations and impact of environmental factors as natural selection. Add to this the notion of mutations as random DNA copying errors and the two fit together like hand and glove. If an individual does not reproduce, any mutations inherited by that individual are lost. How do we estimate chance of such gene extinctions? The article on salmon evolution by Stanford physicist turned evolutionary modeler Shripad Tuljapurkar shows us the way.

Welcome to this special issue on review of recent developments in statistics.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 13 September 2022

References

1. Rangarajan LN (ed) (1987) Kautilya-The Arthashastra. Penguin Books, New Delhi
2. Galton DJ (2012) Did mendel falsify his data? QJM Int J Med 105(2):215–216
3. Visscher P, Goddard M (2019) From R. A. Fisher's 1918 paper to GWAS a century later genetics. 211(4): 1125–1130.
4. Connor H (2022) John Graunt F.R.S. (1620–74): the founding father of human demography, epidemiology and vital statistics. J Med Biogr. https://journals.sagepub.com/doi/full/10.1177/09677720221079826
5. Dunstan M (2011) https://www.centennialbulb.org/ctb-press1.htm
6. Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the COMPAS recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Anil Gore received PhD in statistics from the University of Kentucky, USA, in 1072. He retired as Professor from the Department of Statistics, Pune University, India, in 2007, but continued to work in the field of pharmaceutical trials and was Vice President Statistical Services in CYTEL Statistical Software and Services Pvt Ltd, Pune, from where he retired in 2018. At present, he works as a freelance statistical consultant and trainer. He is Fellow of the Indian Academy of Science, Elected Member, International Statistical Institute, former Member, National Statistical Commission, Government of India.