Australian Statistical Conference and Australian Conference on Teaching Statistics, 10-15 December 2023, Wollongong

Random sampling is a mathematical necessity beyond debate or opinion for valid statistical inferences

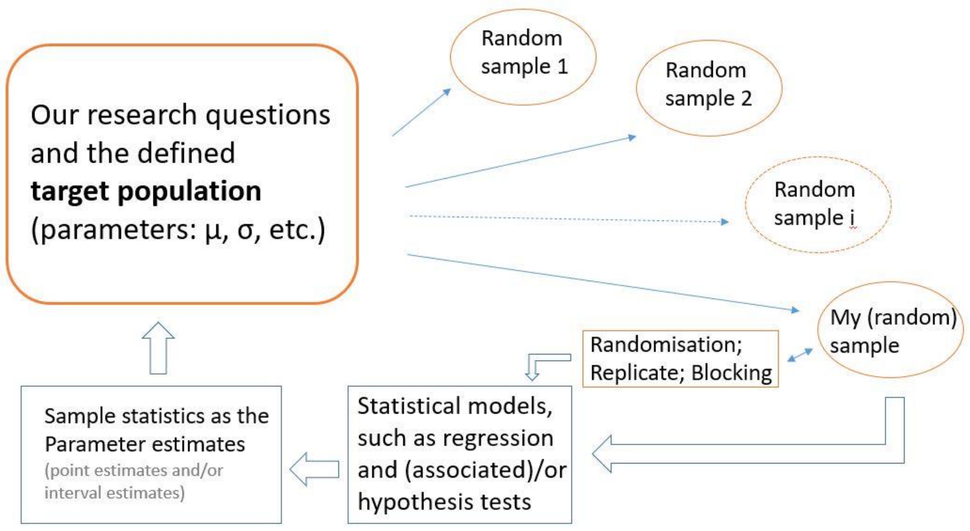
Author: Gang (John) Xie[1,*]
Affiliation: [1] Charles Sturt University, Wagga Wagga NSW 2678, Australia; * corresponding author

Introduction

Random sampling refers to the process of clearly defining a target population and then taking a probabilistic sample(s) from that population.

Most often a probabilistic sample is referred to as *a random sample* in standard statistics textbooks. More specifically, a target population is the collection of all sampling units relevant to some well-defined research questions and, by some objective chance mechanism, each sampling unit has an equal or known (non-zero) probability of being selected into the sample. This is referred to as 'a random sample'. Therefore, each sample unit can be weighted by the inverse of the selection probability to get unbiased estimates of the parameters that fully define the population. Since statistical inferences aim at making justifiable conclusions about a population based on sample data, the random sampling becomes a matter of mathematical necessity rather than a matter of debate or opinion for any valid statistical inferential analysis.

What we intend to do with statistical inferences[1]



The current practice with statistical inferences

- Many research questions were vaguely defined because the target population is often conceptual or hypothetical [3].
- As a matter of fact, it would be highly unlikely to have a real life study (experimental or observational) that employed a truly random sample. Pragmatic and/or ethical factors make it literally impossible in most instances to obtain random samples [3,4].
- Most statistics textbook authors unduly dismiss (or simply ignore) the concerns of random sampling issue with statistical fables (e.g., 'your sample of observations can be imagined to come (from a conceptual population or some superpopulations) and may be regarded as a random sample' [5]); hence most statistical analysis practices simply assume that the random sampling flaw in statistical inferences is negligible [4,6].

The necessary conditions for valid statistical inferences

- A literal population needs to be clearly defined so that research questions can be well-defined. This is the logical prerequisite for conducting random sampling.
- A random sample of size *n* is defined as '*n* random variables X1, X2, …, X*n* form a random sample from a given probability distribution if these random variables are independent and identically distributed (or at least with some known probability distributions)' [2]. The formal statistical inferences (e.g., formulas for the standard error of the estimated mean, t-test, F-test, etc.) depend on the assumption of random samples [2-4].

Doubts on the validity of statistical inferential analysis

- *Without a clearly defined population how can a research question be scientifically defined?* Population must be defined at the start of any study and its definition should include the spatial and temporal limits to the population and hence the spatial and temporal limits to the subsequent inference. For example, 'What is the gender ratio of university students?' is an ill-defined research question; instead, 'What is the gender ratio of Charles Sturt university students enrolled on the first day of 2023?' is a well-defined researchable question.
- *Without a clearly defined population how can we assess the representativeness of the sample data?* No sampling frame would be available without a defined population. A set of sample data obtained without referring to a sampling frame means there are an unknown number of legitimate sampling units with zero probability of being selected into our sample.
- *Without a clearly defined population is there any sense talking about the generalisation of the analysis results beyond the current sample?*
- *Can textbook authors who have appealed to a conceptual or an imaginary population or some superpopulations in place of a literal population tell us how our inferential analysis results can be applicable beyond the sample data in real life sense?*

Remarks and conclusions

- Authors of statistics textbooks were well aware of the fundamental concerns about the random sampling issue for statistical inferences [3-7]. The question is why most chose to cover up the issue with a logically unjustifiable argument of appealing to an imaginary population or pretending the issue did not exist by simply ignoring it. Two of the many reasonable answers could be (1) starting with the motivation to replace scientific inference with statistical inference; (2) then committing the psychological flaw 'It does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!' (quote from J. Cohen (1994), The Earth is Round ($p < 0.05$), doi:10.1037/0003-066X.49.12.997).
- Due to the inherent inability to practically achieve random sampling in real life research, the best formal statistical inference can do is nothing more than exploratory or 'what-if' analysis - NOT confirmatory analysis [6,7]. Thus, the focus of researchers' efforts should be on understanding and being able to describe their data, and providing plausible scientific interpretations about their data analysis results [7] as the grounds for generalisation beyond the sample data.
- **This poster aims to remind researchers of the unignorable fundamental concern of random sampling issue in statistical inferences.** The further concern about what we can do, or what we should do, about the random sampling issue, or more generally about the role of statistical inference related to scientific inference is beyond the scope of this poster. For the concern of what we can/should do, the author recommends reading Raymond Hubbard's book '*Corrupt Research, the case for reconceptualizing empirical management and social science*' (2016) and referring further to the rich references cited therein.

References:

[1] G. Xie (2021). A mini-literature-review: What have been said about Null Hypothesis Significance Test (NHST), presentation at Australian and New Zealand Statistical Virtual Conference, 05-09 July.
[2] M. H. DeGroot (1986). Probability and Statistics (2nd Edition), Addison-Wesley Publishing Company.
[3] R. A. Berk & D. Freedman (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg and S. Cohen (Eds.). Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger (2nd Edition) (pp. 235-254). New York, NY: Aldine.
[4] S. Gorard (2014). The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward?, Psychology of education review, 38(1). pp.3-10.
[5] G.E.P. Box, J.S. Hunter, and W.G. Hunter (2005), Statistics for Experimenters: Design, Innovation, and Discovery (2nd Edition). Wiley-Interscience
[6] Raymond Hubbard, Brian D. Haig, and Rahul A. Parsa (2019), The Limited Role of Formal Statistical Inference in Scientific Inference, The American Statistician, Vol. 73, No. S1, 91-98.
[7] Christopher Tong (2019), Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science, The American Statistician, Vol. 73, No. S1, 246-261.

Contact details: Dr John Xie
Statistics Support Officer,
Quantitative Consulting Unit

Phone: +61 2 69332229
Email: gxie@csu.edu.au
Website: https://www.csu.edu.au/qcu