Inference using non-random samples?

Stop right there!

Statistical inference allows researchers to learn things about a population using only a sample of data from that population. But if it isn't a random sample, inference becomes tricky or outright impossible, as **Norbert Hirschauer**, **Sven Grüner**, **Oliver Mußhoff**, **Claudia Becker** and **Antje Jantsch** explain

Statistical inference is a powerful concept. Among other things, it allows us to infer information about a population based on a sample of data from that population. To make appropriate inferences from sample to population, certain pre-conditions need to be met. One of these pre-conditions is that data come from a random sample.

You do not need to be an expert in inference to suspect that an estimate of national voting intention based on a survey of 1,000 people emerging from a political party conference may not be as trustworthy as a survey of 1,000 people polled at random across the country. And yet, without making precise assertions, there is little doubt that the pre-conditions for using inferential statistical procedures are violated by a considerable proportion of studies in the social sciences – think of surveys based on *convenience samples* such as students in classrooms or voluntary participants recruited from web platforms such as Amazon Mechanical Turk.

In studies such as these, inferential statistics should not be used. However, in many cases they are. A lack of awareness of the need for random sampling among researchers leads them to go through the motions of conducting statistical significance tests and reporting *p*-values and confidence intervals without realising they have breached a fundamental assumption. Even when researchers themselves do know this, they can find themselves compelled to perform inference by ignorant referees.

Statistics are sample quantities

For any set of data – whether obtained through random sampling or not – we can compute summary statistics that describe certain properties of the data. Widely used quantities include the mean ($\bar{x}$) and the standard deviation (s) of observations for a variable, the difference between the means of two groups, and regression slope coefficients that describe relationships between variables in the sample.

Generally speaking, these sample-derived statistics tell us only about the sample they came from. However, some of them gain inferential meaning if the sample is randomly drawn from a broader *parent population*, which must also be the *inferential target population*. For instance, in *a simple random sample*, the sample mean ($\bar{x}$) becomes the unbiased estimate for the unknown population mean (μ); and the sample statistic $s/\sqrt{n}$ becomes the estimated standard error of the mean. This is but another label for the standard deviation of the (*sampling*) *distribution* of all sample means that we would find if we very frequently drew equal-sized simple random samples from the same parent population.

Statistical assumptions are empirical commitments,[1] and complying with the empirical procedure of "random sampling" permits using the standard error for assessing the uncertainty of an estimation caused by random sampling error. Non-compliance, in contrast, implies that all sample quantities remain pure data descriptions. They are devoid of inferential meaning (except when deviations from random sampling are adequately corrected within a sample selection model).[2] More generally speaking, all statistical inferential procedures based on the standard error – including statistical significance tests – are inappropriate for making sample-to-population inferences when studies are based on non-random samples.

Due to deeply ingrained habits, however, statistical significance testing is often "ritualistically" performed without consideration of the specific data and research context. Inappropriate teaching in conjunction with flawed journal policies has effectively propagated significance testing as a routine that must be performed in *all* circumstances, even though it causes egregious inferential errors and obstructs critical inferential reasoning.[3] As a consequence, we face an uninterrupted stream of overconfident proclamations of scientific discoveries (yes/no conclusions) based on misleading significance declarations, although in the majority of cases there is not even a chance model upon which to base statistical inference.[4]

How can statistics help make inferences?

Statistical inference presupposes random sampling and then concerns itself with *random sampling error* – that is, the fact that even a random sample does not exactly reflect the properties of its parent population (see box, "Random sampling error"). Because of random sampling error, sample quantities such as a mean or a regression slope deviate to a greater or lesser extent from the corresponding population quantities. However, unbiased estimators

Random sampling error

Random sampling error occurs even when studies are flawless. Due to the *law of large numbers*, the uncertainty of an estimation resulting from random sampling error decreases when sample size increases. Random sampling error completely disappears when we can study full populations. That is, there is neither need nor room for statistical inference when we already have data for an entire population. Vogt *et al*. note that reporting *p*-values in the analysis of full population data is nonetheless quite common.[9]

When no sample-to-population inference is necessary, interpreting summary data statistics as inferential statistics does not make sense. This is formally reflected in the *finite population correction* (fpc) factor $1 - n/N$, which is used for adjusting the *squared* standard error. Instead of implicitly assuming that a sample was drawn from an infinite population – or at least that a small sample of size n was drawn from a large population of size N – the fpc considers that random sampling error decreases not only with growing sample size but also when the fraction of the population that is sampled becomes large. Researchers are advised to use the fpc when a sample share is greater than 5%.[10] When 5% of the population is sampled, the fpc reduces the standard error by 2.5%. For a share of 50%, the reduction increases to 29.3%. Having data for an entire population ($n = N$) results in an fpc of zero, which, in turn, leads to a corrected standard error of zero. This is consistent because there is no random sampling error when the "sample" covers 100% of the target population.

In contrast, *non-random errors* can cause serious validity problems even when we are in the comfortable position of having access to full population data.

Varavin88/Unsplash.com

Norbert Hirschauer is professor of agribusiness management at the Institute of Agricultural and Nutritional Sciences of the Martin Luther University Halle-Wittenberg, Germany.

Sven Grüner is a postdoctoral researcher at the chair of agribusiness management, Institute of Agricultural and Nutritional Sciences of the Martin Luther University Halle-Wittenberg, Germany.

estimate correctly on average over frequently repeated draws of random samples from the same population. Therefore, it does not make sense to ask whether an individual study's estimate is "true" or not. Instead, we need to consider the body of evidence and include the knowledge contribution from each properly conducted study in our inferential reasoning even if, on its own, the single study only produces an estimate with high uncertainty.[5]

Estimation uncertainty caused by random sampling error is the only type of error that statistical inferential quantities and procedures – including *p*-values and statistical significance tests – deal with.[3] At best, we can extract two meaningful pieces of information from a random sample: an unbiased point estimate (*signal*) and an unbiased estimation of the uncertainty of this point estimate, expressed through the standard error (*noise*); no more, no less.

Downgrading the signal and noise information first into a quotient (such as a *t*-ratio) and then into a *p*-value (based on the usually meaningless null hypothesis of zero effect), and finally into a dichotomous significance declaration (based on an arbitrary threshold) is not in itself wrong. We can, of course, perform the mathematical manipulations. But even when there is a random sample, the procedure is not useful. It not only causes a substantial loss of information but virtually instigates overconfident conclusions, both when *p* is below and when it is above the "significance" threshold (conventionally set at 0.05).

The capacity of sample-derived inferential statistics to help answer the question of what we should most reasonably believe after seeing the results of a study must not be overvalued. Statistical inference is only part of the much larger enterprise of scientific inference. Scientific inference means drawing the most reasonable conclusions regarding a real-world state of interest given all available evidence (e.g., from previous studies) *and* the incremental information that was extracted from a particular sample. In this larger, ongoing enterprise of accumulating knowledge, statistical inference is limited to helping evaluate a single study's knowledge contribution under consideration of the noise resulting from random sampling error.

The sampling design needs to be considered

To obtain unbiased estimates of population quantities and standard errors, we need to consider the probabilistic mechanism through which population members are selected into the sample.[6] While there is a variety of random sampling designs, it is useful to distinguish three generic types: simple random sampling, stratified sampling, and cluster sampling.

Imagine we are interested in the mean per capita income of people in Phantasia City. When we use *simple random sampling* (SRS), each member of the parent population (here, *N* residents of Phantasia City) has equal probability of being selected into the sample (proportionate sampling). Therefore, we can use the (unweighted) mean income ($\bar{x} = \sum x_i/n$) of the *n* sample members as an unbiased estimate of the mean income of all residents in Phantasia City. SRS also facilitates a relatively easy estimation of the standard error of the mean: $\widehat{SE}_{\bar{x}} = s/\sqrt{n}$.

In *stratified sampling*, we first divide the population into segments ("strata") – income classes, for example. Next, we randomly sample residents from each stratum. In proportionate stratified sampling, we sample an identical fraction of each stratum. Analogous to SRS, this gives each resident of Phantasia City an equal probability of being included in the sample. In contrast, disproportionate stratified sampling oversamples certain strata, leading to a sample that is systematically unrepresentative of the population. Weights must be used to correct for this systematic unrepresentativeness, and the weight that is assigned to a sampled resident is the reciprocal of the probability that this resident is included in the sample. If we sample a 10% fraction in Stratum 1 and a 20% fraction in Stratum 2, each resident sampled from Stratum 1 has weight 10 (represents 10 residents), whereas each resident sampled from Stratum 2 has weight 5 (represents 5 residents). Rather than using the simple arithmetic mean, the average of the weighted observations in the sample must be used to obtain an unbiased estimate of the mean income in Phantasia City. The stratified sampling design and the weights need also to be considered when estimating the standard error.

Statistical inference is only part of the much larger enterprise of scientific inference

Cluster sampling bears superficial similarity to stratified sampling because it also subdivides the population into segments. However, it uses a hierarchical approach to data collection. Area sampling is an example. Imagine Phantasia City

Oliver Mußhoff is professor of farm management at the Department of Agricultural Economics and Rural Development of the Georg August University, Göttingen.

Claudia Becker is professor of statistics at the Department of Economics of the Martin Luther University Halle-Wittenberg, Germany.

is divided into 20 urban districts. In area sampling, we might first randomly select, say, eight districts ("primary sampling units"). In a second step, we would then randomly select residents ("secondary sampling units") from only these eight districts. Analogous to stratified sampling, an unbiased estimation of the mean income in Phantasia City must be based on the average of the weighted observations in the sample. Again, the appropriate weight is the reciprocal of a resident's probability of being sampled. But now this probability results from the probability that a district is selected in the first stage *and* the probability that a resident within a district is sampled in the second stage. The multi-stage sampling approach and the weights must also be considered when estimating the standard error.

Convenience samples preclude statistical inference

In the Phantasia example, we made two very unrealistic assumptions. The first was that we had access to all residents of Phantasia City or, at least, to a sampling frame in the form of a random pool of residents. While we are rarely in such a position, we made a second assumption that is even more doubtful: we assumed that each randomly selected resident participated in the survey. Because some people have better things to do than filling out surveys, a full response rate will be achieved only if people can be forced to participate. This is rarely the case in democratic societies, but it would be good news from a statistical point of view because it would prevent *self-selection bias*. In democratic Phantasia City, where people can freely decide whether or not to participate in the study, the inclusion of residents in the sample would not be random (i.e., data would be missing not at random). Consequently, participants and non-participants may be systematically different, and the mean income in the sample may tell us little about the mean income in Phantasia City.

Variables such as age, education, gender, or even income might influence the probability of survey participation. For example, residents with high incomes might be less inclined to respond. Among the vast array of *sample selection models*, propensity score models are the most intuitive way to tackle this problem. Ideally, they use all selection-relevant variables to estimate individual participation probabilities. Similar to the ex-ante known selection probabilities in stratified and cluster sampling, these ex-post estimated probabilities ("propensity scores") are used to reconfigure the sample

Correcting for selection bias

The need and the possibility to correct for selection bias are related to the notion of "missing data".[11] In SRS, no corrections are needed because *data are missing completely at random*, meaning selection into the sample is unconditionally independent of the variable of interest (no confounding).

Somewhat difficult to grasp at first, the label *data missing at random* is attached to a situation in which selection bias can be "remedied" by making unit selection independent of the variable of interest conditional on observed confounders (no unmeasured confounding).

For example, assume that we have access to a sampling frame in the form of a truly random pool of 1,000 residents of Phantasia City (500 men and 500 women). Assume also that we addressed the survey to the entire pool, but the response rate was only 15%. In the sample, we find 100 men (20% of the men on the sampling frame) but only 50 women (10% of the women on the sampling frame). We observe a mean income in the sample of $\bar{x} = 6$. Taking a closer look, we find a mean of $\bar{x}_m = 7$ among the 100 men and a mean of $\bar{x}_f = 4$ among the 50 women. It now seems natural to equate the gender-specific participation shares with selection probabilities ("propensity scores") and act on the assumption that men selected themselves into the sample with a propensity of 0.2, and women did so with a propensity of 0.1. Consequently, each man in the sample would have a weight of $w_m = 5$ and represent five men, whereas each woman would have a weight of $w_f = 10$ and represent ten women. We could now use the weights-corrected sample mean $\bar{x}_w = 5.5 = (5 \cdot 100 \cdot 7 + 10 \cdot 50 \cdot 4)/1{,}000$ as the estimate for the mean income in Phantasia City.

When not all confounders are known or can be observed, it is impossible to adequately correct for selection bias (unmeasured confounding). This corresponds to the notion of *data missing not at random*, which rules out the use of inferential statistics.

Antje Jantsch is a postdoctoral researcher at the Department of Agricultural Policy of the IAMO (Leibniz Institute of Agricultural Development in Transition Economies), Halle (Saale), Germany.

We should accept the hard truth that inference is not always possible

and correct for imbalances between those who are in the sample and those who are not. Therefore, propensity scores are also called "balancing scores".[7] Similar to cluster sampling, the consideration of self-selection usually causes standard errors to be much larger than those that would be obtained from analysing the sample as if it were a simple random sample.[8] But still worse, the high data requirements of sample selection models often cannot be met. For example, if gender, age, education, and income affect participation, we need to know the distribution of these variables not only among participants but also in the parent population.

We may summarise that we often use convenience samples because we are unable to comply with the "empirical commitment" of random sampling. Even when we follow the *procedure* of random sampling, the *result* will often not be a random sample because of self-selection. Ignoring whether and how a sample was probabilistically composed from a defined population rules out estimating standard errors because no sampling distribution can be envisaged. In some rare cases, we may have enough information about non-participants to correct for selection bias, which then rehabilitates the probabilistic foundations for using inferential statistics (see box, "Correcting for selection bias", page 23). However, we must be generally wary of misspecifying the model that is used to correct for self-selection. When groups with distinct characteristics are completely missing in the sample, or when we do not know or are unable to measure all selection-relevant variables because data for non-participants are unavailable, we cannot properly correct for selection bias.

In short, if we do not start with a random sample, turning what we have into one is challenging or even impossible. In such cases, we should accept the hard truth that statistical inference is not possible. We must simply report what the data show – and refuse to push them statistically further. ■

Note

This short article is based on the paper "Can *p*-values be meaningfully interpreted without random sampling?", published in *Statistics Surveys*.[2]

Acknowledgements

We gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG). We further acknowledge the valuable comments and suggestions of the editors of *Significance*.

Disclosure statement

The authors declare no competing interests.

References

1. Berk, R. A. and Freedman, D. A. (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 edn). New York: de Gruyter.

2. Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. and Jantsch, A. (2020) Can *p*-values be meaningfully interpreted without random sampling? *Statistics Surveys*, **14**, 71–91.

3. Ziliak, S. T. and McCloskey, D. N. (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

4. Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (2019) Moving to a world beyond "$p < 0.05$". *The American Statistician*, **73**(sup1), 1–19.

5. Hirschauer, N., Grüner, S., Mußhoff, O. and Becker, C. (2018) Pitfalls of significance testing and *p*-value variability: An econometrics perspective. *Statistics Surveys*, **12**, 136–172.

6. Lohr, S. L. (2019) *Sampling: Design and Analysis* (2nd edn). Boca Raton, FL: CRC Press.

7. Austin, P. C. (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, **46**, 399–424.

8. Copas, J. B. and Li, H. G. (1997) Inference for non-random samples. *Journal of the Royal Statistical Society, Series B*, **59**(1), 55–95.

9. Vogt, W. P., Vogt, E. R., Gardner, D. C. and Haeffele, L. M. (2014) *Selecting the Right Analyses for Your Data: Quantitative, Qualitative, and Mixed Methods*. New York: Guilford Press.

10. Knaub, J. (2008) Finite population correction (fpc) factor. In P. Lavrakas (ed.), *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: Sage.

11. Mercer, A. W., Kreuter, F., Keeter, S. and Stuart, E. (2017) Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, **81**(S1), 250–279.

Statistical inference: a step-by-step approach

1. **Start by unambiguously defining the parent population from which the random sample will be taken. This population must coincide with the target population about which inferences are to be made.**
2. **Collect a sample from the parent population following a clearly defined probabilistic sampling design (e.g., SRS, stratified sampling, cluster sampling) that describes how members of the population are selected into the sample.**
3. **Extract a sample quantity (effect size) – for example, a mean, a mean difference, or an association between two variables such as a regression slope coefficient.**
4. **In order to make generalising inferences, check if selection bias is apparent and can be remedied by a sample selection model. If not, simply report descriptive statistics; do *not* state quantities (e.g., *p*-values) implying inferences are possible.**
5. **If the probabilistic preconditions for statistical inference are met, use the observed sample quantity as a point estimate for the population quantity of interest.**
6. **The validity of findings beyond the confines of an idiosyncratic study is impaired by more than random error (e.g., measurement, model specification, and non-random selection error), so stress that statistical inference deals only with uncertainty caused by random error.**
7. **Accounting for the specific quantity of interest (e.g., mean, mean difference, regression slope) and the specific sampling design (e.g., SRS, stratified sampling, cluster sampling), estimate the standard error to quantify the uncertainty of the point estimate.**
8. **State the "signal" and "noise": the point estimate represents the signal, and the standard error represents the noise from random sampling.**
9. **Use these two intuitive pieces of information – signal and noise – to assess the single study's knowledge contribution within a comprehensive scientific reasoning that makes reasonable inferences in light of all of the available information, including the single study's findings.**