Hirschauer, Norbert; Grüner, Sven; Mußhoff, Oliver; Becker, Claudia; Jantsch, Antje

Can p-values be meaningfully interpreted without random sampling?

Statistics Surveys

Provided in Cooperation with:
Leibniz Institute of Agricultural Development in Transition Economies (IAMO), Halle (Saale)

Suggested Citation: Hirschauer, Norbert; Grüner, Sven; Mußhoff, Oliver; Becker, Claudia; Jantsch, Antje (2020) : Can p-values be meaningfully interpreted without random sampling?, Statistics Surveys, ISSN 1935-7516, Cornell University Library, Ithaca, NY, Vol. 14, pp. 71-91, https://doi.org/10.1214/20-SS129 , https://projecteuclid.org/euclid.ssu/1585274548

Statistics Surveys
Vol. 14 (2020) 71–91
ISSN: 1935-7516
https://doi.org/10.1214/20-SS129

Can p-values be meaningfully interpreted without random sampling?

Norbert Hirschauer

Martin Luther University Halle-Wittenberg, Faculty of Natural Sciences III, Institute of Agricultural and Nutritional Sciences, Chair of Agribusiness Management, Karl-Freiherr-von-Fritsch-Str. 4, D-06120 Halle (Saale), e-mail: norbert.hirschauer@landw.uni-halle.de

Sven Grüner

Martin Luther University Halle-Wittenberg, Faculty of Natural Sciences III, Institute of Agricultural and Nutritional Sciences, Chair of Agribusiness Management, Karl-Freiherr-von-Fritsch-Str. 4, D-06120 Halle (Saale), e-mail: sven.gruener@landw.uni-halle.de

Oliver Mußhoff

Georg August University Goettingen, Department for Agricultural Economics and Rural Development, Farm Management, Platz der Göttinger Sieben 5, D-37073 Göttingen, e-mail: Oliver.Musshoff@agr.uni-goettingen.de

Claudia Becker

Martin Luther University Halle-Wittenberg, Faculty of Law and Economics, Institute of Business Studies, Chair of Statistics, Große Steinstraße 73, D-06099 Halle (Saale), e-mail: claudia.becker@wiwi.uni-halle.de

and

Antje Jantsch

Leibniz Institute of Agricultural Development in Transition Economies (IAMO), External Environment for Agriculture and Policy Analysis (Agricultural Policy), Theodor-Lieser-Str. 2, D-06120 Halle (Saale), e-mail: jantsch@iamo.de

Abstract: Besides the inferential errors that abound in the interpretation of p-values, the probabilistic pre-conditions (i.e. random sampling or equivalent) for using them *at all* are not often met by observational studies in the social sciences. This paper systematizes different sampling designs and discusses the restrictive requirements of data collection that are the indispensable prerequisite for using p-values.

JEL codes: C10, C83.

Keywords and phrases: Sampling design, propensity scores, p-values, random sampling, selection bias, inference.

Received August 2019.

1. Introduction

Statistical inference in the study of observational data is concerned with random (sampling) error, i.e. the fact that even a random sample does not exactly reflect the properties of the population and that there is often considerable sample-to-sample variation when we repeatedly draw random samples from a population. Inferential tools based on standard errors, such as *p*-values and confidence intervals, are to help make inductive inferences beyond the confines of a particular sample to a population. Due to their probabilistic foundation, they *require* a random process of data generation that can hypothetically be replicated. When *p*-values or confidence intervals are displayed, a plausible argument should be given that the studied sample meets the underlying probabilistic assumptions, i.e. that it is or can be treated as a random sample. Otherwise, there are no grounds for using these inferential tools and they become essentially uninterpretable (Copas and Li, 1997; Vogt et al., 2014). When using them, researchers should therefore transparently describe how and from which parent population the sample was drawn and, consequently, to which inferential (target) population they want to generalize (Hirschauer et al., 2019).

This paper focuses on *p*-values and inductive statistical inference in terms of estimating population quantities and generalizing from a sample to its parent population.[1] For example, we might have measured the average per capita income in a sample of 10,000 randomly selected residents, but be ultimately interested in the average per capita income in a city of one million residents. To understand the purpose of statistical inference, one must clearly distinguish between sample quantities (estimates) and population quantities. Otherwise one runs the risk of lurching all-too readily from the description of some sample data that are conveniently at hand to overconfident generalizations regarding population quantities (Matthews et al., 2017). Failing to be explicit regarding the data generation process and the inferential target population fuels the risk of rash statements regarding real-world regularities. Berk and Freedman (2003) emphasize that statistical assumptions are empirical *commitments* and that acting as if one obtained data through random sampling does *not* create a random sample. Complying with the empirical procedure "random sampling" permits using *p*-values as inferential aids for inductive generalizations from the probability sample towards its parent population. Non-compliance, in contrast, precludes a meaningful use of inferential statistics, except when deviations from random-sampling are demonstrably ignorable or adequately corrected for.

Despite delusive significance declarations that are often attached to *p*-values, their inferential content is limited even when they are applicable in principle. In the words of McCloskey and Ziliak (1996: 98) we must realize that "[t]he uncertainty of the estimate that arises from sampling error [is] the only kind of uncertainty the test of significance deals with." In other words, the *p*-value

[1] While being an issue beyond this paper's scope, it should be noted that a probabilistic data generation mechanism (randomization) is also the prerequisite for a meaningful interpretation of *p*-values in experiments aimed at identifying causal effects *within* a given sample of study (cf. Hirschauer et al., 2020).

compares a point estimate from a particular random sample with an estimated noise-only distribution of outcomes that would result if random error were at work alone (Ludwig, 2005). This distribution is inherently based on the thought experiment of statistical replication, i.e. we assume noise-only and presumably apply the same econometric model to many other equal-sized random samples from the same population. In brief, the p-value is the conditional probability of re-finding an observed effect (or a larger one) in random replications if we assumed the statistical model including the null hypothesis of zero effect to be true. Per definition, a p-value cannot work inversely and inform us on the underlying reality; i.e. "it is logically impossible to infer from the p-value whether the null hypothesis or an alternative proposition is true. We cannot even derive probabilities for hypotheses based on what has delusively become known as 'hypothesis *testing*.' In the usual sense of the word, a p-value cannot 'test' or 'confirm' a hypothesis, but only describe data frequencies under a certain statistical model including the null" (Hirschauer et al., 2019: 712).

Unfortunately, in the economic and social sciences, we rarely have random samples. We often use pre-existing observational data or survey data obtained from convenience samples that are non-random but easy to come by at low costs (students in a classroom, volunteers on Web-platforms, people on mailing lists assembled for other purposes, etc.). Even when we can avoid non-coverage problems and succeed in randomly assembling subject pools (sampling frames) from defined populations, we usually do not escape sample selection bias because we cannot force subjects to participate. Due to non-coverage and self-selection, participants and non-participants in a survey might be systematically different.[2] In longitudinal studies, we face the additional problem of attrition when study participants "get lost" over time. The problem of non-probabilistically sampled data can be framed as a data problem where data are missing, but missing not at random (cf. Little and Rubin, 2002; Mercer et al., 2017). As a consequence, estimates of population quantities including regression coefficients as well as the standard errors and p-values of those estimates might be biased in unknown ways. Non-random sample selection is often obvious. Nonetheless, many studies tacitly proceed as if they had a random sample and follow a misguided routine of always displaying p-values for group differences and regression coefficients.[3] Resulting conclusions run the risk of being grossly misleading.

Simple random sampling clearly represents a chance-based sampling design that allows for a meaningful interpretation of p-values.[4] When data collection

[2]Throughout this paper, we distinguish two main reasons why samples may not match parent populations: "non-coverage," when already the sampling frame omits certain portions of the parent population, and "self-selection" or "non-response," when subjects selected from the frame can freely choose to participate or not.

[3]While providing a comprehensive list of studies that use p-values but do not meet the empirical commitment of random sampling is not feasible, Gomez et al. (2019); Chen and Crown (2019), and Massenot and Pettinicchi (2018) represent some haphazardly identified examples from a range of academic disciplines.

[4]Even so, one should be cautious. Over the last decades, a vast literature has evolved warning against inferential errors associated with p-values. For an introduction, see, for example, Ziliak and McCloskey (2008); Krämer (2011); Gigerenzer and Marewski (2015); Greenland et

deviates from simple random sampling, things become less obvious and it is often unclear under which conditions p-values can be used as an inferential aid at all. There is a vast variety of complex and mixed sampling designs whose respective features need to be considered to obtain unbiased estimates of population quantities and standard errors (for an introduction see, for example, Lohr, 2009). For systematic reasons, however, it seems useful to distinguish five ideal-type sampling contexts regarding the use of frequentist tools such as p-values: (i) simple random sampling, (ii) stratified sampling, (iii) cluster sampling, (iv) sampling contaminated with selection bias, and (v) full population surveys. All of them are discussed in this paper to facilitate the understanding of the probabilistic pre-conditions for estimating standard errors and the corresponding fact that using p-values is only meaningful when these conditions are met.

It should be noted that formal representations throughout this paper focus on the mean because the notation for the estimation of standard errors of other quantities such as regression coefficients tends to become messy. The universal argument, however, that standard errors and associated inferential tools strictly presuppose a random process of data generation and that deviations from simple random sampling need to be considered when estimating standard errors applies to all estimates.

2. Sampling designs that facilitate the use of inferential statistics

2.1. Simple random sampling – the generic benchmark for statistical inference

The most basic case of a probability sample is a simple random sample (SRS), i.e. a subset of a population drawn with the same probability as any other conceivable subset of identical size, such that each unit had equal probability of being selected into the subset. Being an unbiased representation of the population, a SRS permits not only the use of conventional point estimates such as the sample mean as unbiased estimate of the population mean but also the use of inferential statistics based on standard errors such as p-values. The standard error of the mean for a SRS is given by:

$$\widehat{SE}_{SRS} = \left(\left(1 - \frac{n}{N} \right) \cdot \frac{s^2}{n} \right)^{0.5}, \tag{1}$$

where n is the sample size, N the population size, and s^2 the variance of the sample.

al. (2016); Berry (2017); Gelman and Carlin (2017); Trafimow et al. (2018); Hirschauer et al. (2018; 2019). The topicality of this issue is reflected in the unprecedented p-value warning of the *American Statistical Association* in spring 2016 (Wasserstein and Lazar, 2016), its p-value symposium in fall 2017, the p-value Special Issue on "Statistical inference in the 21st century: A world beyond $P < 0.05$" of *The American Statistician* (Wasserstein et al., 2019), and the p-value petition to retire statistical significance in *Nature* (Amrhein et al., 2019).

Contrary to standard textbook representations, equation (1) includes the finite population correction factor $1 - n/N$. The finite population correction (*fpc*) accounts for the fact that sampling error not only decreases with growing sample size but also when sample size becomes large *relative* to the population (Hirschauer et al., 2020). The *fpc* reduces the standard error and researchers are advised to use it when a sample comprises more than 5% of the population (Knaub, 2008). When 5% of the population are included in the sample, the *fpc* reduces the standard error by 2.5% compared to the default standard error that considers the absolute size of the sample but ignores its relative share. For a share of 50% (75%), the reduction of the standard error increases to 29.3% (50.0%).

2.2. Stratified sampling

In many surveys, researchers deviate from simple random draws of the population because other sampling designs are statistically or financially more efficient. In stratified sampling, we first divide a population of size N into H mutually heterogeneous but internally as homogeneous as possible subpopulations ("strata"). When deemed appropriate, we might for example divide a given population into several income classes. We then randomly sample from each stratum h ($h \in \{1,\ 2, \ldots, H\}$) independently. The simplest case is proportionate stratified sampling where we sample an identical fraction of each stratum ($n_1/N_1 = \ n_2/N_2 = \cdots = n_H/N_H$), ensuring that the stratum sample size n_h is proportional to stratum size N_h. Disproportionate stratified sampling, in contrast, intentionally oversamples certain strata – for example those that exhibit more variability than others – to reduce sampling error.

Contrary to proportionate stratified sampling, where each unit has equal probability of being included into the sample, we need to consider differential weights when a sample is systematically unrepresentative of the population such as in disproportionate stratified sampling (Solon et al., 2013: 5). In other words, we need to "reconfigure the sample as if it was a simple random draw of the total population" (Friedman, 2013). The weight $w_{hi} = w_h = N_h/n_h$ that is assigned to a sampled unit i ($i \in \{1,\ 2, \ldots, n_h\}$) in stratum h is the reciprocal of the probability that this unit is included in the sample. It indicates how many units of the full population are represented by a sampled unit. If we sample a 10%-fraction in stratum 1 and a 20%-fraction in stratum 2, then each sampled unit in stratum 1 has weight $w_1 = 10$ (represents 10 units), whereas each sampled unit in stratum 2 has weight $w_2 = 5$ (represents 5 units). When estimating a population regression from such a sample, we need to apply, in the simplest case, weighted least squares instead of ordinary least squares to the sample regression to obtain unbiased point estimates (Solon et al., 2013).

Stratified sampling needs also to be considered when estimating sampling variability. Compared to a SRS of size n, stratified sampling, where $n = \sum n_h$, reduces sampling error. The estimated standard error of the mean, for example,

is given by (cf. Lohr, 2009: 79):

$$\widehat{SE}_{Strat} = \left(\sum \frac{N_h^2}{N^2} \cdot \frac{N_h - n_h}{N_h} \cdot \frac{s_h^2}{n_h} \right)^{0.5}, \tag{2}$$

where s_h^2 is the sample variance within each stratum h.

When estimating the standard error for stratified samples, we consider the within-strata variance but *not* the variance between strata because we *independently* draw a random sample from *each* stratum. This is why, compared to simple random sampling, the reduction of the standard error is the more pronounced the smaller the variance within the strata and the greater the variance between them. Equation (2) uses a finite population correction $(N_h - n_h)/N_h = 1 - n_h/N_h$. This correction considers that sampling error not only decreases with growing stratum sample size but also when stratum sample size becomes large *relative* to stratum size. Since the correction applies *within* each stratum, it can be used to make stratified sampling still more efficient: it reduces the standard error when strata with high variability are oversampled. Such oversampling can also be seen as an intentional use of low weights for strata with high variability. Using $w_h = N_h/n_h$, this can be shown by rewriting equation (2) as:

$$\widehat{SE}_{Strat} = \frac{1}{N} \cdot \left(\sum N_h \cdot (N_h - n_h) \cdot \frac{s_h^2}{n_h} \right)^{0.5} = \frac{1}{N} \cdot \left(\sum w_h \cdot (N_h - n_h) \cdot s_h^2 \right)^{0.5}$$

We may summarize that stratified sampling is a special case of random sampling. Not only are the probabilistic pre-conditions for estimating standard errors and using p-values met, but standard errors and p-values can be adjusted downwards. If we neglected this downward adjustment and used the default standard error for a SRS, we would base our inferential reasoning on inflated p-values and therefore be too conservative in our judgement (for example, in conventional "hypothesis testing") – insofar as we find it useful to resort to p-values and p-value thresholds as inferential aids at all.

2.3. Cluster sampling

Cluster sampling (e.g. area sampling) is used in practice because randomly sampling from preselected segments of the population ("clusters") is usually much cheaper and more convenient than randomly sampling from the full population (Lohr, 2009: 167). There is a superficial similarity of cluster sampling to stratified sampling because both subdivide a population of size N into exhaustive subpopulations (segments). However, earmarking the fundamental difference in the sampling design, we use G to denote the number of segments that are now called "clusters." The difference to stratified sampling is that, instead of randomly drawing observational units from *each* segment of the population, we now adopt a hierarchical approach to data collection: in a primary step, we draw a random sample of g clusters from the population of G clusters, which

are therefore called "primary sampling units" (psus). In a second step, we randomly select observational units, which are called "secondary sampling units" (ssus), from the previously selected clusters. Both stages of data collection need to be accounted for when estimating population quantities and standard errors. Lohr (2009: 168) warns that "[o]ne of the biggest mistakes made by researchers using survey data is to analyze a cluster sample as if it were an SRS. Such confusion usually results in the researchers reporting standard errors that are much smaller than they should be."

A simple variant of cluster sampling is one-stage cluster sampling where *all* observational units (ssus) within the selected clusters (psus) are surveyed. When this is too costly, two-stage cluster sampling is applied; i.e. instead of fully surveying selected clusters, we draw a SRS of units from each selected cluster. A frequently used form of cluster sampling is two-stage area sampling. Let's assume we are interested in measuring per capita income in a country with 50 provinces. Area sampling would imply, for example, that in a first step we randomly select ten provinces. In the second step, we then draw random samples of residents merely in the ten previously selected provinces.

Analogous to stratified sampling, we need to consider differential weights when estimating population quantities from cluster samples that are systematically unrepresentative of the population. The appropriate weight is still the reciprocal of a unit's probability of being included into the sample. However, this probability now derives from the probability that a cluster is selected in the first stage *and* the (conditional) probability that a unit within a selected cluster is sampled in the second stage. For dealing with cluster samples in the regression context, adequate methods account for between-cluster and within-cluster variability. Keywords in this context are hierarchical, multilevel, random effects or mixed effects models (McNeish and Harring, 2017: 856; see McCulloch et al., 2008 for methodological details).

Whereas stratification decreases sampling error compared to a SRS, the opposite is generally true for cluster sampling. Lohr (2009: 166) illustrates why:

"Members of the same cluster tend to be more similar than elements selected at random from the whole population – [. . .]; residents of the same nursing home tend to have similar opinions of the quality of care. These similarities usually arise because of some underlying factors that may or may not be measurable – [. . .]. Thus, we do not obtain as much information about all nursing home residents in the United States by sampling two residents in the same home as by sampling two residents in different homes, [. . .]."

The increase of the sampling error in cluster sampling occurs even when we use equal selection probabilities for all ssus that facilitate a conventional ("self-weighting") calculation of unbiased point estimates analogous to a SRS.[5] There

[5]The literature cautiously states that cluster sampling "generally" increases standard errors. This is because units in natural clusters such as geographical areas often share environmental influences that make them more homogeneous than units that are randomly selected from the full population. Measuring this homogeneity, we would find that cluster members exhibit positive intra-cluster correlations. The opposite can be imagined in principle; i.e. the

are two approaches of obtaining equal probabilities for all observational units (ssus): (i) we use probabilities proportional to cluster size when selecting clusters (psus) and then sample the same *absolute number* of ssus in each psu, or (ii) we use a fixed probability for selecting psus and then sample an identical *fraction* of ssus in each psu (one-stage clustering is a special case of this approach). Imagine a population of size 2,000 that is divided into 5 large psus à 200 and 10 small psus à 100. Approach (i) implies that sampling a large psu is twice as likely as sampling a small psu (e.g. 0.2 vs. 0.1). An equal number of ssus (e.g. 50), is then sampled in each psu – corresponding to a fraction of 0.25 in large psus and a fraction of 0.5 in small psus. The resulting probability of ssus being selected would be 0.05, both for ssus in large psus ($0.05 = 0.2 \cdot 0.25$) and for ssus in small psus ($0.05 = 0.1 \cdot 0.5$). In approach (ii), an equal weight could be obtained, for example, by using an equal probability of 0.1 for selecting psus and sampling an identical 50%-fraction of ssus within each psu. While both approaches are self-weighting, Lohr (2009: 220) notes that approach (ii) is expected to produce an even larger sampling error than approach (i).

In one-stage clustering with clusters of equal size N/G that are selected with equal probability, estimating the standard error of the mean is straightforward (cf. Lohr, 2009: 171):

$$\widehat{SE}_{Clust\_1} = 1/\frac{N}{G} \cdot \left(\left(1 - \frac{g}{G}\right) \cdot \frac{s_t^2}{g} \right)^{0.5}, \tag{3}$$

where s_t^2 is the between-psu variance of cluster totals.

Equation (3) shows that in one-stage clustering, the standard error depends on the between-psu variance but not the within-psu variance. This is because there is no within-psu sampling error when we measure *all* ssus in the preselected psus. Even when we maintain the assumption of equal-sized clusters that are selected with equal-probability, the standard error formula becomes complex in two-stage clustering: we now need to consider the between-psu *and* the within-psu variance because sampling error results from two sources: from sampling g clusters in the first stage *and* from sampling n/g units within each selected cluster in the second stage (cf. Lohr, 2009: 185):

$$\widehat{SE}_{Clust\_2} = 1/\frac{N}{G} \cdot \left(\left(1 - \frac{g}{G}\right) \cdot \frac{s_t^2}{g} + \frac{1}{g^2} \sum_{i \epsilon \text{ sampled clusters}} \frac{N^2}{G^2} \cdot \left(1 - \frac{n/g}{N/G}\right) \cdot \frac{s_i^2}{n/g} \right)^{0.5}, \tag{4}$$

where s_i^2 is the within-psu variance of sampled observations from the ith sampled cluster, and n is the total sample size.[6]

If we analyzed the data as if they were obtained from a simple random draw of the full population, we would mistakenly estimate the standard error according

units within a cluster could be less homogeneous than randomly selected units and exhibit negative intra-cluster correlations. While this is unlikely in natural clusters, it might occur in artificial clusters (Lohr, 2009: 173ff).

[6] When *all* clusters are selected in the first stage, we, again, have stratified sampling. Aligning the notation ($g = G = H$, $N/G = N_h$, $n/g = n_h$, $s_i^2 = s_h^2$), equation (4) can correspondingly be reduced to equation (2).

to equation (1). Basing the estimation of standard errors in a cluster sample – be it for a mean or a regression coefficient – on the SRS assumption[7] generally leads to an underestimation. This underestimation is sometimes quantified by the so-called "design effect," i.e. the ratio of the sampling variance that results from the cluster sampling design to the sampling variance of a SRS of the same size (Kish, 1965: 161). Often, correct standard errors are several times larger than the default standard errors that presume SRS. While the design effect tells us by how much we would underestimate the variance if we erroneously used the formula for a SRS of the same n, "it is not a way to avoid calculating variances: You need an estimate of the variance from the complex design to find the design effect" (Lohr, 2009: 309).

It should be noted that the confusingly similar term "cluster-robust estimation of standard errors" is used in a different research context. For example, we might know that data were collected in a SRS design, but an *ex-post* analysis of the data might nonetheless reveal intra-cluster correlations. The literature on "cluster-robust estimation" focuses on this ex-post identification of clusters. For an introduction see Cameron and Trivedi (2005: 829–845, 2009: 323–329 and 306–311); Cameron and Miller (2015) or MacKinnon (2019) and the literature referenced therein. Contrary to the consideration of the data collection design "cluster sampling" where the number of clusters must be finite and known, the cluster-robust estimation of standard errors is based on the assumption that the total number of clusters is very large (goes to infinity).[8] The two approaches do generally not coincide. It is therefore essential to distinguish the two research contexts. A small example illustrates why: when data were collected by others, researchers might ignore the particular design. While it is common in such circumstances to try to circumvent the problem by using cluster-robust standard errors, such approaches might be flawed. Imagine a researcher who ignores that the data resulted from stratified sampling and finds ex post that observations are similar in certain segments (e.g. provinces). These provinces might then be considered as clusters and a cluster-robust estimation based on the usual assumption that the total number of clusters goes to infinity would find standard errors that are larger than conventional standard errors. They should be smaller, however, because in the stratified sampling design *each* province was selected.

We may summarize that a cluster sampling design is another special case of random sampling where the probabilistic pre-conditions for estimating standard errors and using p-values are met. However, cluster samples usually produce

[7] It should be noted that even in a SRS design, we need to use "heteroscedasticity-robust" standard errors when the dispersion of observations is different in different segments of the population.

[8] Cameron and Miller (2015: 23) note that, when this is not the case, even cluster-robust standard errors can be *downwards-biased*. At the same time, the number of observed clusters should be small relative to the population of clusters. Noting that many economic datasets do not meet this condition, Abadie et al. (2017) find that cluster-robust standard errors are often *upwards-biased*. This raises the question of how cluster-robust standard errors are to be estimated in the SRS case when there is a finite population of clusters (e.g. 20 geographical areas) and a number of observed clusters (e.g. 10) that is *not* small relative to the population (cf. Cameron and Miller, 2015).

larger sampling errors than SRSs. Hence, standard errors and p-values need to be adequately adjusted upwards. If we neglected this upward adjustment and used the default standard error for a SRS, we would base our inferential reasoning on wrongly deflated p-values and therefore be too confident and, even worse, be too confident to an *unknown degree* in our judgements (for example, in conventional "hypothesis testing") – insofar as they are based on p-values and p-value thresholds in the first place.

3. Sampling designs that preclude the use of inferential statistics

3.1. Convenience samples contaminated by bias from non-coverage and self-selection

Observational data are often tacitly analyzed as if they were obtained through random sampling even when a non-random selection mechanism was at work. Such approaches can be seriously flawed. Trafimow (2019: 344) unmistakably warns that "it is the rare study that uses sampling that is completely random and where each participant sampled is independent of the other participants sampled. Without making assertions about exactly how often these assumptions are importantly violated, there can be little doubt that the violation prevalence is considerable. In that case, no inferential statistical procedures, [. . .], are completely valid. [. . .] The notion that inferential statistical procedures may sometimes, and even often, be inappropriate, may be tough medicine for reviewers and editors to swallow. But the medicine nevertheless is therapeutic. Another option is to use methods for addressing violations of independence or random sampling [. . .], while keeping in mind that these have their own assumptions and drawbacks."

Using data from non-probability convenience samples is common in econometric studies (Miller, 2017). For example, researchers often ask people who happen to be present in certain venues to participate in a survey. In behavioral economics and psychology, the most notorious instance are students from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) countries who happen to be in a particular researcher's classroom (Henrich et al., 2010). Other examples are passers-by in shopping-malls, users of social media, people who happen to be on some mailing list or who explicitly agree to be included in some data base that assembles volunteers for surveys. On Web- platforms, for example, respondents can be recruited at low costs from non-probability panels of individuals who volunteer to participate in many different surveys for money. Amazon's Mechanical Turk is a prominent example. To put it more systematically: non-coverage, i.e. when certain segments of the population are omitted from the sampling frame, is one instance that leads to a non-probabilistic sample composition. But even when we succeed in avoiding non-coverage and obtaining a sampling frame that contains a random pool of subjects, we usually do not escape self-selection because people can freely decide whether to respond to a survey or not. Selection problems can arise in all sampling designs and need to be

accounted for *in addition* to the design-specific estimation procedures described in Section 2. Because inclusion into a sampling frame and response are rarely random (i.e. data are missing but missing not at random), the characteristics of participants and non-participants may be systematically different (selection bias). Therefore, sample quantities may tell us little about the population quantities of interest. In regression analysis, we run the risk of *mis*estimating coefficients and standard errors in unknown ways unless we adequately correct for the selection bias that results from the violation of independent random sampling (cf. Rosenthal and Rosnow, 2009: book 3).

Given the trend of decreasing survey participation (often not more than 20% or even much less) and the increasing recruitment of respondents from volunteer Web-platforms, *Public Opinion Quarterly* published a Special Issue on "Survey Research, Today and Tomorrow" in 2017 (Vol. 81). The Special Issue puts a particular focus on the question of how to deal with the non-probability samples resulting from these developments. Its editor notes:

"To survive, surveys need to maintain their 'value proposition.' One important part of this proposition is that surveys gather information that is generalizable from the measured sample to a well-defined population" (Miller, 2017: 205–207).

With a view to the increasing use of non-probability samples obtained from volunteer platforms, he states that "[t]here have been prominent clashes between advocates for the probability and nonprobability approaches. Those who support the probability sampling method observe that they have probability theory to support their population inferences, while the advocates for volunteer panels rely on inductive modeling that has no comparable theoretical foundation." In response, the supporters of non-probability panels stress that their approaches work well enough in many practical research contexts and counter that actually achieved samples do very rarely come up to the presumed probabilistic requirements anyway.

What we do in the attempt to correct for selection bias is best understood when looking back to stratified and cluster sampling. In both sampling designs, we knew ex ante the selection probabilities and the ways in which the data generating process was systematically different from a simple random draw of the full population. We therefore also knew how to "reconfigure" the sample to make it comparable to a simple random draw and ensure that the variable distribution in the sample matches the distribution in the parent population on average. When the composition of the sample is influenced by non-coverage and/or self-selection, we do *not* know these probabilities ex ante. We only know *that* (certain types of) people might be included into the sample with differing probabilities, depending on individual characteristics (age, sex, education, income etc.) that may be observable or not. Sample selection models are used to ex post control for these selection probabilities. They are based on the assumption that we know all confounding variables which might play a role in subjects' participation in a survey and that we have observations for these variables, both from participants *and* non-participants (no unmeasured confounding).

From the late 1970s, an extensive literature developed that concerned itself with selection problems.[9] Two important research contexts were addressed: (i) experimental research where the identification of causal relationships is compromised when randomization is not fully successful (e.g. due to dropouts after randomization or quasi-experimental designs), and (ii) survey research where the unbiased estimation of population quantities is impaired when random sampling is not warranted due to non-coverage and/or non-response. In the last two decades, a particular focus was put on selection problems in Web-surveys and, methodologically, on propensity score methods, which were imported to the survey context from (quasi-)experimental research. Propensity score models are probably the most intuitive way to address selection bias. In experimental research, propensity score models focus on causal inferences within the group of experimental subjects (internal validity). In this context, they are used to deal with the problem of unbalanced confounders across treatment groups. Mercer et al. (2017: 253) explicitly liken selection processes in survey research and experimental research:

"A probability-based survey is essentially a randomized experiment where the pool of subjects is the set of units on the sampling frame and the treatment is selection into the survey. Unlike experiments where we observe outcomes on both treated and untreated subjects, in surveys we observe outcomes only on the selected units, with the expectation that there should be no difference between selected and non-selected units."

In the survey context, propensity score models ideally use all selection-relevant variables to estimate subjects' probability (the "propensity score") of participation as opposed to non-participation (Heckman et al., 1997). Similar to the ex-ante known selection probabilities in stratified and cluster sampling, these ex-post estimated probabilities are then used to reconfigure the sample by correcting for any imbalances between those who are in the final sample and those who are not. Propensity scores are therefore also called "balancing scores," *conditional* on which the distribution of confounders in the participation-group is assumed to match the distribution in the non-participation group (Austin, 2011). This can be associated with the notion of "missing data" (cf. Mercer et al., 2017): in random sampling, where "data are missing completely at random" (MCAR), unit selection is *unconditionally* independent of the variable of interest (no confounding). No corrections are therefore needed. Using propensity scores to "remedy" selection problems, in contrast, is an attempt to make unit selection independent of the variable of interest *conditional* on observed confounders. This corresponds to the notion of "data missing at random" (MAR). When not all confounders are observed, it is not possible to adequately correct for selection bias. This corresponds to the notion of "data

[9] A review of this vast literature is beyond this paper's scope (see e.g. Heckman, 1976; 1979; Greene, 1981; Berk, 1983; Winship and Mare, 1992; Heckman et al., 1997; Stolzenberg and Relles, 1997; Vella and Verbeek, 1999; Lee and Marsh, 2000; Shadish et al., 2002; Kalton and Flores-Cervantes, 2003; Luellen et al., 2005; Rosenbaum, 2010; Austin, 2011; Brick, 2013; Valliant et al., 2013; Guo and Fraser, 2015; Imbens and Rubin, 2015; Mercer et al., 2017).

missing not at random" (MNAR), which rules out the use of inferential statistics.

Let us look at a stylized example to illustrate how propensity score methods work in principal: we are interested in the mean of a target variable y (e.g. per capita income) in a population with $N = 10,000$ subjects. We presumable constructed an appropriate sampling frame that comprises 1,000 subjects (500 males and 500 females) and addressed a survey to all subjects on this sampling frame. The response rate was only 15%, resulting in a sample size of $n = 150$. In this sample, we find $n_m = 100$ males (a 20%-share of the male population on the sampling frame), but only $n_f = 50$ females (a 10%-share of the female population on the sampling frame). Assuming that gender is the only variable that affects selection (this is a critical assumption that we make for the sake of simplicity), we would equate the two gender shares with selection probabilities (propensity scores). That is, we would act on the assumption that males selected themselves into the sample with propensity 0.2 (= 100/500), whereas females did so with propensity 0.1 (= 50/500). To put it the other way round, each male in the sample has a weight $w_m = 5$ (represents five males on the sampling frame) whereas each female has a weight $w_f = 10$ (represents ten females on the sampling frame). Let's further assume that the mean among the 100 sampled males is $\overline{y}_m = 7$, whereas the mean among the 50 sampled females is $\overline{y}_f = 4$. This corresponds to a conventional sample mean $\overline{y} = 6 = (100 \cdot 7 + 50 \cdot 4)/150$. To correct for the fact that gender is not equally distributed across participants and non-participants (oversampling of males, undersampling of females), we resort to the weights derived from the propensity scores. Doing so, we obtain an estimate for the population mean of $\hat{\overline{y}} = 5.5 = (5 \cdot 100 \cdot 7 + 10 \cdot 50 \cdot 4)/1,000$ (weights-corrected sample mean).

This highly stylized example corresponds to post-stratification; i.e. we use the selection-relevant characteristic "gender" to define population strata (i.e. a "male stratum" and a "female stratum") and then put the differing self-selection probabilities of males and females, which are beyond our control, on a level with intentional selection probabilities that we might have used in a disproportionate stratified sampling design. Hence, the same formulas as in the stratification case apply and equation (2) can be used to estimate the standard error of the mean in our stylized example (cf. de Leeuw et al., 2008: 317–341) for the estimation of standard errors under various versions of weight corrections). Our example was simple because gender was the only variable affecting selection and because there was *one* selection probability for all males and *one* selection probability for all females on the sampling frame. Compared to that, participation in real surveys may be much more complex and depend on multiple, and potentially interacting, confounding variables (including the target variable itself) that might furthermore affect participation in a non-linear fashion. Individual participation propensities will therefore often differ between many subjects. As a consequence, simple weighting schemes as the one described above do not suffice any more to account for the biased sample composition introduced by self-selection. More elaborate propensity score models try to accommodate these complexities by considering all potentially relevant characteristics to calculate

individual participation probabilities, which, in turn, are considered when estimating population parameters and standard errors. Contrary to the outcome in our stylized example, standard errors that are adjusted for self-selection are often larger than those resulting from the assumption that there is no problem with a biased composition of the sample (Copas and Li, 1997).

While the formal presentation of propensity score models or other sample selection models that deal with the problem of data that are missing but missing not at random is beyond this paper's scope (cf. e.g. Little and Rubin, 2002), it should be recognized that we need to be generally wary of miss-specifying the selection model. When groups with distinct characteristics are completely missing in the sample or when we do not know, or are not able to measure, all selection-relevant variables, we cannot properly correct for selection bias. While emphasizing that correction methods should be used whenever sufficient data are available, Cuddeback et al. (2004: 22) note that "failing to use these methods when appropriate is different than failing to use them when data for non-participants are unavailable, which is common. In this latter case, sample selection models are obviously of no use." Of course, researchers can never be completely sure that they know all selection-relevant variables. And often the necessary data from non-participants will simply be lacking. Even when considerable amounts of data are available from non-participants, one should remain skeptical whether all confounding variables that affect selection were adequately considered. Going beyond conventional routines, this needs to be explicitly reflected in the interpretation of results to avoid overconfident inferential reasoning.

We may summarize that we often use convenience samples because we are unable to comply with the "empirical commitment" of randomly sampling units from a defined parent population. This is due to two obstacles: first, lacking accessibility to the population or insufficient budgets may prevent researchers from constructing appropriate sampling frames that cover the entire target population. Second, even in appropriate sampling frames, uncontrolled self-selection is likely to produce biased samples. Ignoring whether and how a sample was probabilistically composed from a defined population rules out the use of standard errors and p-values because no sampling distribution can be envisaged. In some rare cases, we may have enough information from non-participants to correct for selection bias, which, in turn, rehabilitates the probabilistic foundations for using inferential statistics (Levay et al., 2016). However, when we are not able to perform the necessary corrections, which is likely to be more often the case in practical research than the literature implies, we should refrain from delusively insinuating that the composition of the sample is unbiased. Instead, we should openly communicate that we have to do without inferential statistics, and limit ourselves to descriptive analysis (see the above quote by Trafimow, 2019). While inductive inferences based on comprehensible scientific arguments are of course feasible in such circumstances, p-values are meaningless and cannot be used as inferential aids in the inductive exercise of generalizing from the sample to the parent population.

3.2. Full population surveys (100% samples)

Since statistical inference is concerned with sampling error and generalizations from random samples to parent populations, there is neither need nor room for statistical inference when we already have data for an entire population. Vogt et al. (2014: 243) note that this is quite common, for example, when analyzing aggregated economic data. They also provide an illustrative example of a study that looks at the association between mean education levels and birth rates in *all* UN member states. Since there is nothing to infer, displaying p-values does not make sense. Instead, one should simply describe the population quantities including regression coefficients. This intuitive fact is formally reflected in the finite population correction (*fpc*) $1 - n/N$, as used in equation (1). 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 very large population of size N – the *fpc* considers that sampling error decreases when the fraction of the population that is contained in the sample becomes large (Hirschauer et al., 2020). Having data for a full population ($n = N$) corresponds to a *fpc* of zero, which leads a corrected standard error of zero. This is consistent because there is no sampling error when the "sample" covers 100% of the population.

Nonetheless, the *fpc* is frequently ignored and p-values are displayed in regressions that analyze data from entire populations. To justify this procedure, the frequentist statistician must somehow introduce the notion of a sampling distribution and a sampling error. This implies imagining an infinite "unseen parent population" (or "superpopulation") and a generating process from which one has presumably observed one noisy random realization (Hirschauer et al., 2018). In so doing, the observed full population becomes in imagination a sample of a (parent) superpopulation. Even back in the 80s, Denton (1988: 167) noted that this is a rhetorical device (also known as "great urn of nature") that does not evoke wild enthusiasm from everybody. However, some random process of data generation – and not just a record of empirical data – has to be presumed for frequentist inferential tools such as p-values to make sense.

When we have observations for an entire population of interest, adopting the notion of a superpopulation may sometimes be comprehensible. Researchers could imagine, for example, a quasi-stable real-world system whose operative regularities, while being possibly limited to a finite time interval, exceed the time interval over which the population data were collected. In the above UN example, this would imply envisaging regularities that not only produced the association between mean education levels and birth rates in *all* UN member in the period for which the data were collected but that will also be at work in the next period(s). Obviously, the plausibility of such an assumption and thus the validity of generalizing claims supported by p-values can only be assessed when the time interval over which the data generating system is presumably at work is clearly defined.

Sometimes a random process of data generation is even assumed in the case of non-probability *samples*. Imagine, for example, a researcher who addresses a survey to the 100 students who happen to be in the classroom on a Monday

morning. We know that such convenience samples might be seriously biased and tell us little about the population quantities of interest. Assuming in such research contexts that there is a random process of data generation does not facilitate statistical inference in any conventional sense of the word. It requires considering the sample as the finite inferential target population. That is, instead of generalizing towards a broader (numerically larger) parent population of students, inferences would be limited to the unseen superpopulation in terms of a random process that is valid for exactly these 100 students and from which one has presumably observed one realization. No statistical inferences beyond the sample of the 100 students can be made (Hirschauer et al., 2019).

Unfortunately, the presumption of a superpopulation often remains implicit. Sometimes its necessity may even not be realized by researchers themselves who, due to engrained disciplinary habits, engage more or less automatically in "statistical significance testing" whenever comparing groups or running a regression model. Researchers who display p-values in the analysis of non-random samples or full-populations data should explicitly discuss why and how they base their inferential reasoning on the notion of a superpopulation. This is but a specification of the general desideratum that researchers explicitly describe the sampling process as well as the population of interest from which the random sample comes and to which they want to generalize (cf. Abadie et al., 2014).

4. Conclusion

The p-value is sometimes seen as conclusive piece of evidence. However, in recent years a plethora of publications have warned against inferential errors associated with such views. This includes the p-value warning of the *American Statistical Association* (Wasserstein and Lazar, 2016), a Special Issue in *The American Statistician* (Wasserstein et al., 2019), the widely supported p-value petition to retire statistical significance in *Nature* (Amrhein et al., 2019), and the Consensus Report on Reproducibility and Replicability in Science of the National Academies of Sciences (2019). Following up on these warnings, this paper focused on the often overlooked fact that the very pre-conditions for using p-values are not met when specific sampling designs are ignored or when empirical studies are based on non-probability convenience samples. Uncorrected non-probability samples rule out using p-values because inferential statistics concerned with generalizing towards a broader population are conceptually based on the notion of repeated random sampling (statistical replication) and a resulting sampling distribution. When there is no random sampling, there is no sampling distribution and no random sampling error. Random data generation is therefore a *necessary* condition for a meaningful application of standard errors and p-values. When data do not satisfy this probabilistic requirement, p-values are essentially uninterpretable.

Critically reflecting on the conceptual pre-conditions for using p-values is important for two reasons: first, it calls to mind the nature of statistical inference, which – even if applicable – solely deals with the uncertainty of estimates

caused by random error, and which is therefore only a part of scientific inference. Second, it is important because much of econometric research uses data from convenience samples that are not probabilistically obtained from well-defined populations. Convenience samples are affected by one of the two following drawbacks (or both): first, the units on the sampling frame are conveniently chosen but not probabilistically selected from the target population (non-coverage). Due to non-coverage, there are unknown differences between the units who are on the sampling frame and those who are not. Second, the units that are on the sampling frame select themselves with unknown probability into the final convenience sample which, if not corrected for, leads to self-selection bias. While these two features often occur jointly in practical research, each of them suffices to rule out a meaningful use of p-values. To ensure successful scientific communication, non-probability samples must therefore be clearly distinguished from probability samples: (i) When researchers succeed in achieving probability samples, they should transparently state from which population the sample was drawn, and consequently, to which population they want to make inductive inferences (generalizations). Without providing such basic information, inference statements are nebulous, at best – and without value, at worst. (ii) When researchers are limited to using non-probability samples that preclude the use of inferential statistics they should be clear about it and refrain from displaying p-values. Displaying inferential statistics in circumstances where population inferences cannot be supported by probability theory is likely to cause overconfident inductive conclusions. Alas, it still seems a widespread spontaneous reflex among researchers who often do not explicitly question whether there is a chance model upon which to base statistical inference.

Acknowledgment

The authors would like to thank the editor Wendy Lynn Martinez and the anonymous reviewers for their helpful comments that led to a significant improvement in the paper. We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding.

References

Abadie, A., Aghey, S., Imbens, G., Wooldridge, J.M. (2014): Finite population causal standard errors. NBER Working Paper No. 20325.

Abadie, A., Athey, S., Imbens, G.W., Wooldridge, J.M. (2017): When should you adjust standard errors for clustering? NBER Working Paper 24003.

Amrhein, V., Greenland, S., McShane, B. (2019): Retire statistical significance. Nature 567: 305–307.

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

Berk, R.A. (1983): An introduction to sample selection bias in sociological data. American Sociological Review 48(3): 386–398.
Berk, R.A., Freedman, D.A. (2003): Statistical assumptions as empirical commitments. In: Blomberg, T.G., Cohen, S. (eds.): Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger (2nd ed.). New York, de Gruyter: 235–254.
Berry, D. (2017): A p-value to die for. Journal of the American Statistical Association 112(519): 895–897. MR3735344
Brick, J.M. (2013): Unit nonresponse and weighting adjustments: a critical review. Journal of Official Statistics 29(3): 329–353.
Cameron, A.C., Miller, D.L. (2010): Robust inference with clustered data, Working Paper, No. 10-7, University of California, Department of Economics, Davis, CA.
Cameron, A.C., Miller, D.L. (2015): A practitioner's guide to cluster-robust inference. Journal of Human Resources 50(2): 279–292.
Cameron, A.C., Trivedi, P.K. (2005): Microeconometrics: Methods and Applications. Cambridge, Cambridge University Press. MR0579788
Cameron, A.C., Trivedi, P.K. (2009): Microeconometrics using Stata. College Station, Stata Press.
Chen, J.J., Crown, D. (2019): The gender pay gap in academia: Evidence from the Ohio State University. American Journal of Agricultural Economics 101(5): 1337–1352.
Copas, J.B., Li, H.G. (1997): Inference for non-random samples. Journal of the Royal Statistical Society Series B 59(1): 55–95. MR1436555
Cuddeback, G., Wilson, E., Orm, J.G., Combs-Orme, T. (2004): Detecting and statistically correcting sample selection bias. Journal of Social Service Research 30(3): 19–33.
Denton, F.T. (1988): The significance of significance: Rhetorical aspects of statistical hypothesis testing in economics. In: Klamer, A., McCloskey, D.N., Solow, R.M. (eds.): The Consequences of Economic Rhetoric. Cambridge, Cambridge University Press: 163–193.
de Leeuw, E.D., Hox, J.J., Dillman, D.A. (2008): International Handbook of Survey Methodology. New York, Taylor & Francis Group.
Friedman, J. (2013): https://blogs.worldbank.org/impactevaluations/tools-of-the-trade-when-to-use-those-sample-weights.
Gelman, A., Carlin, J. (2017): Some natural solutions to the p-value communication problem – and why they won't work. Journal of the American Statistical Association 112(519): 899–901. MR3735346
Gigerenzer, G., Marewski J.N. (2015): Surrogate science: The idol of a universal method for statistical inference. Journal of Management 41(2): 421–440.
Gomez, E.M., Young, D.M., Preston, A.G., Wilton, L.S., Gaither, S.E., Kaiser, C.R. (2019): Loss and loyalty: Change in political and gender identity among Clinton supporters after the 2016 U.S. presidential election. Self and Identity 18(2): 103–125. MR3834282
Greene, W.H. (1981): Sample selection bias as a specification error: Comment. Econometrica 49(3): 795–798. MR0619483

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G. (2016): Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31(4): 337–350.

Guo, S., Fraser, M.W. (2015): Propensity Score Analysis. Statistical Methods and Applications (2nd ed.). Thousand Oaks, Sage Publications.

Heckman, J.J. (1976): The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimation for such models. Annals of Economic and Social Measurement 5(4): 475–492.

Heckman, J.J. (1979): Sample selection bias as a specification error. Econometrica 47(1): 153–161. MR0518832

Heckman, J.J., Ichimura, H., Todd, P.E. (1997): Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies 64: 605–654. MR1623713

Henrich, J., Heine, S.J., Norenzayan, A. (2010): Behavioral and Brain Sciences 33(2/3): 1–75.

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. (2018): Pitfalls of significance testing and p-value variability: An econometrics perspective. Statistics Surveys 12(2018): 136–172. MR3860867

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. (2019): Twenty steps towards an adequate inferential interpretation of p-values in econometrics. Journal of Economics and Statistics 239(4): 703–721.

Hirschauer, N., Gruener, S., Mußhoff, O., Becker, C. (2020): Inference in economic experiments. Economics, The Open-Access, Open-Assessment E-Journal 14(2020-7): 1–14.

Imbens, G.W., Rubin, D.B. (2015): Causal Inference for Statistics, Social, and Biomedical Sciences. New York, Cambridge University Press. MR3309951

Kalton, G., Flores-Cervantes, I. (2003): Weighting Methods. Journal of Official Statistics 19(2): 81–97.

Kish, L. (1965): Survey Sampling. New York, Wiley.

Knaub, J. (2008): Finite population correction (fcp) factor. In: Lavrakas, P. (ed.): Encyclopedia of Survey Research Methods. Thousand Oaks, Sage Publications: 284–286.

Krämer, W. (2011): The cult of statistical significance – what economists should and should not do to make their data talk. Schmollers Jahrbuch 131(3): 455–468.

Lee, B., Marsh, L.C. (2000): Sample selection bias correction for missing response observations. Oxford Bulletin of Economics and Statistics 62(2): 305–323.

Levay, K.E., Freese, J., Druckman, J.N. (2016): The demographic and political composition of mechanical turk samples. SAGE Open, January-March 2016: 1–17 (DOI: https://doi.org/10.1177/2158244016636433).

Little, R.J.A., Rubin, D.B. (2002): Statistical Analysis with Missing Data (2nd ed.). Hoboken, New York, Wiley. MR1925014

Lohr, S.L. (2009): Sampling: Design and Analysis (2nd ed.). Boston, Brooks/Cole. MR3057878

Luellen, J.K., Shadish, W.R., Clark, M.H. (2005): Propensity scores: An introduction and experimental test. Evaluation Review 29: 530–558.

Ludwig, D.A. (2005): Use and misuse of p-values in designed and observational studies: Guide for researchers and reviewers. Aviation, Space, and Environmental Medicine 76(7): 675–680.

MacKinnon, J.G. (2019): How Cluster-Robust Inference is Changing Applied Econometrics. Queen's Economic Department Working Paper no. 1413.

Massenot, B., Pettinicchi, Y. (2018): Can firms see into the future? Survey evidence from Germany. Journal of Economic Behavior & Organization 145(2018): 66–79.

Matthews, R., Wasserstein, R.L., Spiegelhalter, D. (2017): The ASA's p-value statement, one year on. Significance 14: 38–41.

McCloskey, D.N., Ziliak, S.T. (1996): The standard error of regressions. Journal of Economic Literature 34(1): 97–114.

McCulloch, C.E., Searle, S.R., Neuhaus, J.M. (2008): Generalized, Linear, and Mixed Models (2nd ed.). New York, Wiley. MR2431553

McNeish, D.M., Harring, J.R. (2017): Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. Communications in Statistics – Simulation and Computation 46: 855–869. MR3606856

Mercer, A.W., Kreuter, F., Keeter, S., Stuart, E. (2017): Theory and practice in nonprobability surveys. Parallels between causal inference and survey inference. Special Issue 2017: Survey Research, Today and Tomorrow. Public Opinion Quarterly 81: 250–279.

Miller, P.V. (2017): Is there a future for surveys? Editorial to the Special Issue 2017: Survey Research, Today and Tomorrow. Public Opinion Quarterly 81: 205–212.

National Academies of Sciences, Engineering, and Medicine (2019): Reproducibility and Replicability in Science. Consensus Study Report. Washington, DC, The National Academies Press (https://doi.org/10.17226/25303).

Rosenbaum, P.R. (2010): Design of Observational Studies. New York, Springer. MR2561612

Rosenthal, R., Rosnow, R.L. (2009): Artifacts in Behavioral Research. Oxford, Oxford University Press.

Shadish, W.R., Cook, T.D., Campbell, D.T. (2002): Experimental and Quasi-experimental Designs for Generalized Causal Inference. Boston, Houghton Mifflin.

Solon, G., Haider, S.J., Wooldridge, J. (2013): What are we weighting for? NBER Working Paper 18859 (http://www.nber.org/papers/w18859).

Stolzenberg, R.M., Relles, D.A. (1997): Tools for intuition about sample selection bias and its correction. American Sociological Review 62(3): 494–507.

Trafimow, D. et al. (2018): Manipulating the alpha level cannot cure significance testing. Frontiers in Psychology 9 (https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00699/full).

Trafimow, D. (2019): Five nonobvious changes in editorial practice for editors and reviewers to consider when evaluating submissions in a post $p < 0.05$

universe. The American Statistician 73(sup1): 340–345. MR3925740

Valliant, R. Dever, J.A., Kreuter, F. (2013): Practical Tools for Designing and Weighting Survey Samples. New York, Springer. MR3088726

Vella, F., Verbeek, M. (1999): Two-step estimation of panel data models with censored endogenous variables and selection bias. Journal of Econometrics 90(2): 239–263.

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

Winship, C., Mare, R.D. (1992): Models for sample selection bias. Annual Review of Sociology 18: 327–350.

Wasserstein, R.L., Lazar N.A. (2016): The ASA's statement on p-values: Context, process, and purpose. The American Statistician 70(2): 129–133. MR3511040

Wasserstein, R.L., Schirm, A.L., Lazar, N.A. (2019): Moving to a world beyond "$p < 0.05$". The American Statistician 73(sup1): 1–19. MR3511040

Ziliak, S.T., McCloskey D.N. (2008): The Cult of Statistical Significance. How the Standard Error Costs Us Jobs, Justice, and Lives. Michigan, The University of Michigan Press. MR2730043