Sample representation in the social sciences*

Kino Zhao

March 5, 2020

Abstract

The social sciences face a problem of sample nonrepresentation, where the majority of samples consist of undergraduate students from Euro-American institutions. The problem has been identified for decades with little trend of improvement. In this paper, I trace the history of sampling theory. The dominant framework, called the design-based approach, takes random sampling as the gold standard. The idea is that a sampling procedure that is maximally uninformative prevents samplers from introducing arbitrary bias, thus preserving sample representation. I show how this framework, while good in theory, faces many challenges in application. Instead, I advocate for an alternative framework, called the model-based approach to sampling, where representative samples are those balanced in composition, however they were drawn. I argue that the model-based framework is more appropriate in the social sciences because it allows for systematic assessment of imperfect samples and methodical improvement in resource-limited scientific contexts. I end with practical proposals of improving sample quality in the social sciences.

1 Introduction

In 1936, the magazine *Literary Digest* set out to predict the US presidential election between Alfred Landon and Franklin D. Roosevelt. They surveyed more than 10 million people, of which 2.4 million responded, and concluded that Landon was going to win with 57% of the votes against Roosevelt's 43%. Instead, Roosevelt won with 62% against Landon's 38%.

This infamous incident is repeatedly cited to highlight the importance of selecting a sample that is representative. The *Literary Digest* employed a sampling procedure

*This is a post-peer-review, pre-copyedit version of an article published in *Synthese*. The final authenticated version is available online at: http://dx.doi.org/10.1007/s11229-020-02621-3

that favoured wealthy citizens over poor ones and did not correct for the vast majority of people who did not respond, resulting in a biased sample[1].

The problem of sample nonrepresentation remains prevalent in the social sciences. For example, Sears (1986) analyzed the sample composition of research papers published during the year 1980 in three mainstream social psychology journals, *Journal of Personality and Social Psychology* (JPSP), *Personality and Social Psychology Bulletin* (PSPB), and the *Journal of Experimental Social Psychology* (JESP), and found the percentage of studies using American undergraduate students as samples to be 70% for JPSP and 81% for both PSPB and JESP. Their subsequent analysis of these journals for the year 1985 revealed no significant change. Arnett (2008) analyzed six prestigious psychology journals in different areas for the years 2003-2007 and found that most of the samples are taken from the United States (68%), with the remaining largely composed of people from other English-speaking countries (14%) and Europe (13%). A closer look at JPSP in 2007 reveals that 67% of American studies had samples consisted of undergraduate psychology students. More recently, Pollet and Saxton (2018) report that 79% of samples in the journals *Evolution & Human Behavior* and *Evolutionary Psychology* in the years 2015-2016 are from North America or Europe; moreover, 70% of the samples were either online samples or student samples. An analysis of three 2017 issues of *Psychological Science* by Rad et al. (2018) shows similar patterns.

The prototypical psychology sample, consisting of Euro-American undergraduate students, has been coined as WEIRD (Western, Educated, Industrialized, Rich, and Democratic) by Henrich et al. (2010b). Echoing researchers before them (e.g. Peterson, 2001, Wintre et al., 2001), they argue that we have considerable evidence to believe that the WEIRD subjects are very different from other people whom these subjects are often taken to represent.

While most agree that a WEIRD sample is not representative, articulating the exact desiderata of a representative sample proves difficult. In fact, some advocate abandoning the concept "representation" altogether, preferring the more concrete concept of random selection. In this paper, I argue that one major obstacle faced by the improvement of sample quality in the social sciences is the unrealistic expectation of a randomness-based conception of sample representation, called the design-based approach. Instead, I argue that a model-based approach to sampling offers a more realistic framework for effective assessment and improvement of imperfect samples.

This paper is organized as follows. Section 2 presents a brief history of how, through the seminal work of Jerzy Neyman (1934), random sampling superseded purposive sampling and became the preferred method among survey samplers. Representative samples, according to this framework, are ones generated through random selection. Section 3 points to difficulties with random sampling in practice and how, at least in the context of social science, falling short of the ideal standards result in systematic

[1]The popular story told in statistics texbooks is that the *Literary Digest* used its own subscriber list, automobile registration and telephone books to choose its sample, and hence was biased towards wealthy Republicans (e.g., Likert, 1948; Scheaffer et al., 1971). This story is disputed by Bryson (1976), favouring instead the explanation from nonresponse bias.

selection bias. Section 4 revisits the method of purposive sampling, re-emerging under the names of "model-based" or "prediction-based" approach through the works of Royall and Herson (1973). Based on this alternative approach, I argue for the old purposive sampling idea where a representative sample is one *balanced* on all relevant features. Section 5 discusses consequences of adopting a model-based perspective of sample representation in the social sciences and make practical proposals for improvement. Section 6 concludes.

2 Design-based Representation

The Norwegian statistician Anders N. Kiær is often credited as the first to bring sample-based research – that is, investigative methods which utilize only part of the population – to the attention of the western statistics community (Rao and Fuller, 2017; Smith, 1976; Kruskal and Mosteller, 1980). Around the turn of the 20th century, Kiær delivered a series of speeches at the annual meetings of the International Statistical Institute, advocating for the use of samples as effective proxies for studying populations.

The major source of skepticism Kiær faced was a lack of justification for sample-based inference, called the "representative method" at the time, which is inferentially ampliative. Kiær believed that we could identify a set of "rational selection procedures", produce "miniature populations", and draw accurate conclusions without full enumeration. He justified this approach empirically: he demonstrated that the sample-based survey results could be accurate but did not provide a theory for why the process worked (Seng, 1951). Other statisticians followed suit. Although few had comprehensive theories regarding why the representative method worked, many were able to demonstrate, empirically, that it did. Sampling was widely used in European government survey efforts by the 1920s.

With the increased use of the representative method in survey work, a new point of contention emerged between random and purposive selections. For our immediate purpose, the difference between them concerns whether the selection of a sample needs to be sensitive to the sample's composition. In random selection, the inclusion of each member into the sample is governed by probability alone, which is supposed to be identical across all members of the population. In purposive selection, the sampler aims to pick a fixed number of subjects with different characteristics so that the sample has the same proportions of those characteristics as the population.

We can see that these two sampling approaches correspond to Kiær's two conceptions of what a good sample should be. On the one hand, a representative sample should be drawn through "rational procedures", such as a random procedure that leaves no space for personal bias. On the other hand, a representative sample should be a "miniature population" in the sense of matching certain aspects of the population. The most natural way to achieve this goal is to deliberately select samples to be like the population in desired ways through purposive sampling. Although Kiær had both senses of representation in mind, they do not always coincide. That is, a sample drawn

through a rational procedure may fail to be a miniature population. Consequently, samplers prefer one sense often had to let go of the other.

Although random and purposive sampling methods differ practically, they were not seen as direct competitors. Part of the reason may be that the dominant justification for the use of samples was still empirical – sampling with either method had been tried and true. In a 1926 Report for the International Statistical Institute, the English statistician Sir Authur Bowley distinguished the two sampling approaches and recommended them equally. All of these were changed by Jerzy Neyman's 1934 landmark paper.

Neyman's paper made two important contributions to the field of survey sampling. First, he provided a theoretical foundation for random sampling using his recently invented estimation method of confidence intervals. Second, he exposed an important flaw in purposive sampling. Although not everyone was convinced by Neyman's theory of confidence intervals[2], most were convinced enough to adopt random sampling as the superior method. Neyman's framework remained unchallenged within statistics until at least the 1960s. It is still very much the dominant paradigm among the social sciences today.

Neyman's definition of random sampling is elegantly summarized, in his own words, as follows (1934, p.585-586, emphasis original)

> Thus, if we are interested in a collective character X of a population π and use methods of sampling and of estimation, allowing us to ascribe to every possible sample, Σ, a confidence interval $X_1(\Sigma)$, $X_2(\Sigma)$ such that the frequency of errors in the statements
>
> $$X_1(\Sigma) \leq X \leq X_2(\Sigma)$$
>
> does not exceed the limit $1 - \varepsilon$ prescribed in advance, *whatever the unknown properties of the population*, I should call the method of sampling representative and the method of estimation consistent.

There are two important claims of generality here. One of them is emphasized by Neyman, namely that the inference should hold regardless of the population distribution on the characteristic in question. This means that the success of the inference does not depend on assumptions made about the population. This generality resides in the heart of the supposed superiority of random sampling over purposive sampling. As will be discussed further in section 4, Neyman's primary criticism of purposive sampling is the fact that one would need to make a number of assumptions, many of which are either rarely true or rarely known to be true.

The other claim of generality is not highlighted or discussed much, which is the claim that the method of sampling should allow us to "ascribe to every possible sample" this desired property. In other words, it is the *sampling design*, rather than the sample itself, that justifies the inference. In fact, the justification of the inference should not

[2]In particular, Bowley and Fisher remained skeptical, see Brewer et al. (2013).

refer to the specific characteristics of the sample at all. If an inference holds, it needs to hold for "every possible sample" drawn with the same method.

By putting the burden of justifying sample-based inference on the sampling method alone, to the explicit exclusion of referencing properties of the specific samples, Neyman's approach to sampling clearly follows the "rational procedures" line of Kiær's advocacy. Here, randomization is considered as the core of "rational" design, and it is in virtue of the power of design that the ampliative inference is justified.

Neyman's "design-based" approach to sampling remained unchallenged for decades. Later theorists developed more sophisticated schemes of sampling that allowed for uneven probability of inclusion across the population, but the basic idea remained. Randomization is the foundation of the representative method.

3 The Scientific Reality

According to the design-based framework, the inferential power of a sample comes from the sampling design, where the gold standard is random or probabilistic selection[3]. Theoretically, random sampling is often taken to contain two kinds of virtues. Smith (1983, p.394) explains,

> The arguments for randomization are twofold. The first, and most important for science, is that randomization eliminates personal choice and hence eliminates the possibility of subjective selection bias. The second is that the randomization distribution provides a basis for statistical inference.

I have explained that the statistical foundation of sampling is commonly considered to have begun with Neyman's 1934 paper[4], and yet sampling has been used widely before then. This is because random sampling, as a form of rational procedure, has a lot of intuitive appeal.

A central aspect of random sampling is the idea that the selection of elements is governed by probability, rather than scientists' intentions or other selection forces capable of causally influence the conclusions drawn from the sample. For example, suppose a group of surveyers is trying to estimate the average income of a country, then allowing the size of a person's house to influence the probability of that person

[3]I shall use the terms "random" and "probabilistic" interchangeably. Practically speaking, random selection implies that every element of the population has an equal chance of being included in the sample, whereas probabilistic selection allows that chance to differ from element to element. However, probabilistic sampling is almost always accompanied by a correction procedure where elements with greater chance of selection are weighed less in analysis. Theoretically, the two methods are the same.

[4]It seems that other statisticians, such as Bowley, have attempted to provide mathematical foundations for sampling before Neyman. However, Neyman does not discuss these alternative approaches in detail in his 1934 paper, and his paper is widely considered as the statistical landmark (see, e.g., Rao and Fuller, 2017 and Srivastava, 2016). It seems reasonable to conclude that whatever mathematical foundations of survey sampling existed before Neyman, whether or not they are adequate, have had limited historical influence.

being included in the sample is going to result in biased estimations. To guard against a tendency to preferentially sample people with big hourses or small ones, one needs to make sure that the size of someone's house cannot inform the probability of them being selected into the sample. The best way to achieve this goal regarding not only house size but all other forms of influence is to make the selection procedure maximally uninformative. Random selection is, at its core, a maximally uninformative selection procedure.

This intuitive appeal of random selection relies on the premise that maximal noninformation is sufficient in removing undesired interference to study results. As the phrase suggests itself, maximal noninformation precludes outside factors from systematically affecting ("informing") a sample's composition. However, this does not mean that the undesired biases would not occur.

To better appreciate this worry, consider again the problem of sample nonrepresentation discussed in section 1. In their influential attack on WEIRD samples, Henrich et al. (2010b) specifically argued that the worry with WEIRD samples is not simply that most of the world's population is not WEIRD, but that the behaviours of WEIRD samples may differ substantively from the rest of the population. They specifically sited results from Segall et al. (1966)[5] on how people from some cultures are not subject to the Müller-Lyer illusion and from Henrich et al. (2010a) on how people from different cultures respond to the Dictatorship and Ultimatum Games differently as reasons to be skeptical of the generalizability of results obtained from WEIRD samples. The point of contention from their critics is also that many behaviours are not subject to cultural influence. For example, Gächter (2010) argues that whether the use of student samples is problematic depends on the research question. In particular, since economic behaviours are taken to be universal, "any subject pool is in principle informative about whether theoretical predictions or assumptions contain behavioral validity" (p.2).

It is clear that the problem with biased samples is not so much that members of the sample "look" very different from members of the population. Instead, the worry is that these apparent differences translate to unappreciated behavioural differences and so the results obtained from the sample are not generalizable to the population. Similarly, the worry with "subjective selection bias" is not so much that a bias results in members of a sample more likely to have certain characteristics, but rather that *these characteristics interfere with drawing accurate conclusions from the sample.*

This form of systematic bias often occurs as a result of "personal choice" in the sense of preferential sampling, which may happen consciously or unconsciously. A researcher may consciously choose to sample wealthier citizens as a way to inflate the national average income estimation. Alternatively, the researcher may unconsciously

[5]A reviewer has pointed out that the validity and interpretation of Segall's results have been disputed. Indeed, it is a persistent difficulty to determine whether an observed difference is due to a difference in sample composition, methodological variation, or a number of other factors deemed irrelevant. One goal of the framework advocated in this paper is to help better systematize the variations in sample composition so as to facilitate better hypothesis testing regarding the source of a variation.

choose to sample only those who are dressed nicely, leading to the same effect. From the perspective of drawing conclusions from a sample, both forms of personal choice result in undesired systematic bias. Random selection eliminates both sources of influence.

However, the same bias may also occur as a matter of chance. Even if the sampling procedure is truly random, it is still possible that a particular sample happens to consist of members who are wealthier than the national average. To see why this is the case, consider how, even though a fair coin has a 50% chance of landing head, it is not the case that, for every 10 coins I flip, exactly 5 of them will land heads. If the coin is truly fair, the Law of Large Numbers guarantees that, as the number of flips goes to infinity, the proportion of heads converges to the true proportion – 50%. However, the Law of Large Numbers does not guarantee that the true proportion will be reached at any finite stage. In fact, it does not even guarantee that my estimation always improves with more flips[6]. Similarly, random selection only guarantees that, if the population is repeatedly sampled for infinitely many times, then the average of the sample means approximates the population mean. It does not guarantee that any single sample will have the same mean as the population.

The procedure standardly used to address the problem of chance bias is *post-stratification*. In stratified sampling, a population is divided into multiple mutually exclusive, collectively exhaustive "strata". Samples of different sizes are drawn, randomly or otherwise, from these strata. The resulting samples are weighed by the size of their strata relative to the population and combined to form the final sample. In post-stratification, the process is reversed. After a sample is drawn, it is partitioned into groups, often along some salient characteristics deemed important by the researchers. The associated strata are reversely constructed, their relative ratio computed from auxiliary full population data, and the groups weighed accordingly.

Consider the National Comorbidity Survey (NCS) as an example, which was launched in the US as "the first psychiatric epidemiologic survey to administer a broadbased research diagnostic interview to a nationally representative sample of the United States" (Kessler, 1994). The NCS uses a stratified, multistage area probability sample, which is common for survey efforts of its scale. The core sample contained 47.5% males. However, according to the National Health Interview Survey (NHIS) of 1989, a full enumeration of the population rather than a sampled survey, 49.1% of Americans were male. The NCS therefore post-stratified their sample by giving more weight to results obtained from the sampled males than that of the females.

It is worth noting that the NCS, despite adopting a method as close to random sampling as is feasible, still feels the need to adjust data to compensate for sample imbalance. This shows the limitations of random selection as a guard against chance bias.

[6]In certain special cases and with strong additional assumptions, a method may guarantee uniform convergence, where the estimation is always improved with increased sample size. When that happens, one can obtain an ε-δ bound on how far "off" we can be for a given confidence threshold. However, this option is only open for fields where it is easy to repeatedly, truly-randomly gather large samples, which is unrealistic for the social sciences.

More significantly, the method of post-stratification does not really fit in the design-based framework. Recall that the design-based conception of sample representation relies exclusively on the power of the selection process to justify sample-to-population inference. The idea is supposed to be that, as long as researchers adopt adequate sampling procedures, they should not feel the need to also analyze sample composition.

Besides, the design-based framework does not provide guidance for how sample composition should be analyzed. To see this, we can compare the NCS with similar sampling efforts from other countries. The NCS of America post-stratified against sex, age, marital status, race, education, region, and urbanicity (Mickelson et al., 1997, p.1095); the German National Health (GNH) survey post-stratified against sex, age (with a different range), marital status (in finer categories), and employment status (Jacobi et al., 2002); the Australia National Mental Health Survey (ANMHS), however, decided to not post-stratify at all (Henderson et al., 2000).

In addition to the inconsistencies across similar survey efforts, those that do post-stratify provide very little reasoning as to why they decide on the characteristics that they do. Post-stratification as a method depends on the existence of full enumeration demographics data like the NHIS, which is often called "auxiliary data" or "organic data" in this context. The existence of such data limits whether and how a sample survey can afford to post-stratify. That said, post-stratification also reflects conscious choices on the part of the research team. For example, the NCS chose to post-stratify against the NHIS rather than the US Census because "[the NHIS] includes a much wider array of sociodemographic variables for the purposes of poststratification" (Mickelson et al., 1997, p.1095). This is certainly not because the US Census did not gather a lot of data. In 1989, the Census Bureau gathered information as diverse as age differences between bride and groom, prevalence of AIDS, immigrative status, and average weekly expenditure (US Census Bureau, 1989). Instead, the US Census gathered *the wrong sorts of data*, at least from the perspective of the NCS.

It is clear that researchers make judgments about which characteristic imbalance is worth correcting in a randomly selected sample, and yet these judgments are rarely explicitly stated or argued for. Indeed, there is no theoretical space within the design-based framework for such corrections, so it only makes sense that corrections like these, when they do occur, are guided more by intuition than by arguments.

The discussion concerning post-stratification has highlighted two important observations. First, even the best random selection efforts result in sample imbalances deemed worthy of correction by researchers. The elimination of "subjective selection bias" guaranteed by random selection is clearly insufficient. Second, while post-stratification is frequently used to correct for chance bias, the practice is not principled. This is because post-sampling corrections of this form do not fit into the design-based understanding of how sampling is supposed to work.

Worse still, large-scale survey efforst like the NCS are relatively uncommon; most research teams within the social sciences do not have nearly as much resources to employ anything like an area probability sample over a nation. This is compounded by the fact that many research projects within psychology, anthropology, and economics

target the entire humanity as the intended population. If random sampling over a country is difficult, random sampling over the entire human race is practically impossible. This is especially true when the study procedures are very involved, such as in experiments, longitudinal studies, or when data are collected qualitatively.

Another major obstacle for samplers in the social sciences is the problem of nonresponse. When the sampled units are humans, there is always a chance that someone sampled will decline to participate. When that happens, the actual sample will differ from the theoretical sample envisionsed by design. Since nonresponse makes a probabilistically selected sample effectively nonprobabilistic, it is a serious problem. Indeed, many believe that failure to address nonresponse is the true culprit behind the epic failure of the *Literary Digest* poll.

The coping strategy developed in the 1950s, which continues to be the preferred strategy today, is two-phase sampling. In the first phase, the preferred measurement procedure is used for everyone theoretically selected in a sample. If some members of the sample do not respond, then a second phase is carried out where a different measurement procedure is used to reach nonrespondents. The idea is that the alternative measurement, while less ideal in other ways, may change the minds of nonrespondents. For example, the alternative method may be a more resource-intensive in-person interview as opposed to a paper-based questionnaire, or it may be a shortened version of the questionnaire which takes less time for subjects to complete. If the second phase elicits near-full response and the alternative measurement methods are considered empirically equivalent, then the problem of nonresponse is fully corrected. If significant nonresponse remains at the second phase, researchers would often assume homogeneity among nonrespondents and post-stratify as if second phase nonresponse is undersampling. Unsurprisingly, nonresponse remains a serious challenge today.

When the sampling procedure cannot plausibly be construed as random, the design-based framework ceases to provide guidance. Statistical foundations for the design-based approach, such as Central Limit Theorems, rely on the conceptualizaion of sampled elements as random variables. From this perspective, all non-random selection procedures are equally bad.

What this also means is that, if a study cannot obtain random selection, researchers lose any sense of how they might still improve their sample. The essence of probabilistic sampling is that every member of the population has a non-zero probability of being selected. Even if this probability varies from member to member, post-sampling correction methods such as post-stratification can adjust the weights such that the results are "as if" selection is truly random. However, if some members of the population have probability 0 of being selected, then there is nothing one can do to make the data look as if those members could have contributed. One cannot modify nonprobabilistic samples *post hoc* to make them probabilistic. I believe this is the main reason that, despite wide recognition of the problem of sample nonrepresentation, the proportion of studies employing undergraduate-only samples has not changed over the decades.

In the absence of principled ways of improvements, convenience becomes a major driving factor. Convenience sampling refers to the practice where members of the

sample are chosen because of ease of access and recruitment. The most common form of convenience sampling is using undergraduate students at the same institution where the researchers are based. Other forms include Amazon Mechanical Turk or community members recruited using posters or email advertisements.

Unsurprisingly, convenience sampling is the most common form of sampling within the social sciences. An analysis of sample composition in 5 journals in developmental science shows that 78-88% of all studies published in years 2007-2011 that use American samples use convenience sampling (Bornstein et al., 2013). Given the prevelance of undergraduate and online samples within the social sciences, the same is likely true of other fields as well.

In addition to being nonprobabilistic, convenience sampling often perpetuate a specific kind of systematic bias. Consider, again, the use of WEIRD samples, where E stands for educated and R for rich. It should not be surprising that people who are rich and educated are more likely to have the leisure to participate in odd psychological studies. This is especially true when the study offers very little compensation, as is the case in most resource-limited academic contexts.

The prevalence of convenience sampling highlights an important feature of sample design that is often overlooked in abstract discussions – sampling involves not only a decision about design, but also a series of actions associated with actually contacting and recruiting subjects. Without an explicit intention to guard against this tendency, subjects who are more "accessible" are likely going to dominate conveniently gathered samples. This is especially problematic because subjects who are less accessible are usually such because of other forms of marginalization. For example, one limitation identified by the ANMHS is the noninclusion of indigenous people who live in remote locations (Andrews et al., 2001). For another example, the persistent underrepresentation of African Americans in samples used in clinical psychology studies (Graham, 1992) is likely to be a major contributor to the persistent clinical malpractice disproportionatelly experienced by this population (Hall, 1997).

To summarize, the design-based framework of sample representation, where random sampling is considered the gold standard, faces two major problems in practice. First, while faithful execution of random sampling can eliminate intentional selection bias, it cannot eliminate chance bias. The scientific importance of chance bias can be witnessed by the wide use of post-stratification as a correction mechanism. However, the design-based framework provides no guidance for such corrections, which is why they are often carried out inconsistently and with little justification. Second, random selection is extremely difficult to achieve in resource-limited contexts. When random selection is out of the question, the design-based framework is again silent on how a sample may still be improved. Consequently, researchers rely on convenience as the dictating principle. Convenience sampling often introduces systematic bias of a particular kind that are likely to compound existing social gaps.

4 Balanced sampling

To briefly return to the history of sampling, recall that sample representation was used in two distinct senses by Kiær and his immediate followers: as samples obtained through "rational selection procedures" and as samples that are "miniature populations". Since these senses do not always coincide, differing priorities have led survey samplers to two different paths: random sampling and purposive selection. Neyman's 1934 seminal paper convinced the statistical community that random sampling rests on a solid statistical foundation, while purposive selection relies on contentious and unrealistic assumptions.

In purposive sampling, one has a variable of interest, X, and a number of control variables. For ease of illustration, assume there is only one control variable, Y. In the early form of purposive sampling targeted by Neyman, X and Y are assumed to be linearly correlated. Assume the characteristic Y is well known and easily measured, one can purposively select members such that the sample distribution on Y matches that of the population[7]. This sample is considered representative with respect to X.

Neyman's criticism of purposive sampling consists of two aspects. First, he pointed out that the then-recent Italian sample survey, which used purposive selection, was vastly inaccurate – this form of empirical argument has always carried a lot of weight with surveyers. Second, Neyman pointed out that the assumption of linear dependence between X and Y is often unrealistic. Random sampling, Neyman explains, is assumption-free.

As statisticians delved deeper into the foundation of sample-based estimation, they gradually reazlied that random sampling, or at least inferences based on it, are not as straightforward as Neyman had believed. For example, Godambe (1955) developed a unified account of a class of estimators commonly used around that time and showed that there does not exist a best linear unbiased estimator in this general class, contrary to what Neyman had claimed. Later, he showed how the likelihood function from the full sample data, theoretically understood to include a set of labels together with associated variables of interest, provides no information on the non-sampled values and hence on the population total or mean (Godambe, 1966; see also Rao and Fuller, 2017).

Against the backdrop of theoretical and practical challenges to random sampling, a new approach was developed by, primarily, the statistician Richard Royall (Royall 1968, 1970, 1992; Royall and Herson, 1973). Royall's basic observation is that sample-based inference can be conceptualized as a prediction problem, where results obtained from sampled individuals are used to predict what results we would obtain from the

[7]Neyman's original analysis was based on stratified versions of random and purposive sampling. In his rendition of purposive selection, each stratum was sampled such that the mean of Y in the stratum sample equaled the mean of Y in the overall stratum. Allowing the means of Y to differ among strata, Neyman's description of stratified purposive sampling is equivalent to sampling from the entire population in a way that the sample distribution of Y matches the population distribution of Y.

unsampled part of the population.

According to the design-based framework, the inferential power of the sample comes from the idea that, while *these* individuals were in fact sampled, the sample could very easily have contained *those* other individuals instead – those ones we are trying to estimate. In other words, the members in the sample are *interchangeable* with members outside of the sample in some sense[8]. From a prediction perspective, however, the relationship between sampled and unsampled individuals need not be nearly as strong. If I am using a person's wealth to predict their life expectancy, I do not need to assume that the tax return data, say, I have obtained from one person could have been from another person instead. What I do need to assume is that the person whose data I have is sufficiently similar in relevant ways to the person whom I'm trying to predict.

If the two tax returns are considered two instantiations of the same random variable, as in the case of design-based random sampling, then the assumption that they should be similar is in some sense warranted. However, if we know what it means for two people to be similar in this context, then we can check whether they indeed are similar in an *ad hoc* way. For example, if we think a person's country of residence affects their life expectancy, then we would want to make sure that the sampled person resides in the same, or a relevantly similar, country as the unsampled one.

This way of understanding sample-based prediction leads to the form of purposive sampling targetted by Neyman – if we believe that matching distribution of Y between sample and nonsample ensures that the individuals from these two groups are similar, then we should sample to match distribution of Y. If Y is indeed the only characteristic correlated to X, then the resulting sample would be a "miniature population" in Kiær's sense – it mirrors the population in a way relevant to the study target, X.

A major contribution by Royall is to develop a more general account of this style of inference where a sample that does not already match the population on the entire distribution of Y can be used in similar ways with extra assumptions. For example, in ratio estimation, the ratio between X and Y is considered constant along different values of Y. Suppose a person's wealth and their life expectancy are positively linearly correlated and that the sample we have consists mostly of people wealthier than the national average. In this case, we can compute the slope of the trendline relating X and Y from our sample of wealthy subjects and extrapolate this information for poorer ones. If we know the national average wealth, we can estimate the national average life expectancy accordingly. All of this is done without referencing the sample gathering process.

As is evident from the above example, this approach, called the model-based or prediction-based approach, relies on a number of auxiliary information. We need to first identify one or some control variable(s) Y, with the assumption that they relate

[8]Exchangeability is a Bayesian perspective on how random sampling works. The design-based framework is, by and large, developed and used under the frequentist paradigm, where random selection is defined as i.i.d. (independent and identically distributed) sampling, which grounds the application of the Law of Large Numbers. Exchangeability is presented here because it offers are more intuitive description of the inferential process.

to X strongly enough to serve their intended function. We also need certain kinds of population-level data concerning Y. If we cannot match Y across the entire distribution – which is almost always true in practice – we will need to make assumptions concerning the nature of the relationship between X and Y. In the case of ratio estimation, the model requires a linear dependence between X and Y that passes through the origin. With increased computational power, more complicated dependence relationships can be accommodated.

With as many assumptions as needed in even the simpliest cases, the problem identified by Neyman is a serious one. Just as we may be wrong about the relationship between X and Y being linear, we may also be wrong about any other assumed nature of this relationship, or that they are related at all. This problem is one of model misspecification, which is always a challenge in model-based inference. Indeed, model misspecification, especially the kind that is difficult to detect but can significantly bias the resulting estimation, has been the major challenge to the model-based approach (Hansen et al., 1983)[9].

Royall and Herson (1973) showed that sample balance can protect against model misspecification. Their definition of sample balance is as follows. Suppose $Y_1 \dots Y_n$ are all the variables upon which X is dependent. Then a balanced sample is one where the mean of each Y_i $(1 \leq i \leq n)$ of the sample equals that of the population. If a sample is balanced in this way, then many model-based estimators retain their optimality and unbiasedness under many instances of model misspecification[10].

Model misspecification remains a threat as long as sample balancing is practically difficult. In the social sciences, researchers often choose to study X precisely because they do not know how it relates to other variables. Questions of model misspecification and sample balance are intrinsically part of the unknown. In other words, if researchers knew that the model was adequate or that the sample was balanced, they would not have conducted the study in the first place.

The immediate consequence of this observation is that, like random sampling, the model-based approach does not offer an easy route to sample representation in the context of resource-limited social sciences. This should not come as surprise, however, as a change in perspective is not supposed to magically solve an intrinsically difficult problem. The benefit of the model-based approach is that it provides a framework capable of guiding sample improvement in systematic ways.

Recall that one important shortcoming of the design-based framework discussed in the previous section is that, once a research team cannot obtain probabilistic sampling,

[9] A design-model hybrid approach, called model-assisted sampling was developed not long after the development of the model-based approach. The hybrid approach aims to use properties of random selection to help guard against model misspecification (Cassel et al., 1976; see also Brewer, 1999). I will not discuss the hybrid approach for two reasons. First, the importance of purposive balancing, which is my main thesis, is equally emphasized in both the model-based and hybrid approaches. Second, the guarding power of the hybrid approach against model misspecification only appears in large samples with relatively good randomization, which is not part of my target.

[10] These estimators are approximately unbiased if the sample is approximately balanced.

it is difficult to see any other ways of improvement. In the absence of such principled guidance, convenience becomes the dominant consideration, leading to systematic bias. The benefit of the model-based framework is that one can explicitly state all the assumptions necessary to support the inference in question and discuss the evidence we have of them.

Return to the example of a person's wealth and life expectancy. In order to propose that our ratio estimator based on a biased sample of wealthy individuals is adequate in estimating the national average life expectancy, we need to make the following assumptions. First, variation in wealth accounts for most of the variation in life expectancy. Second, the relationship between wealth and life expectancy is positively linear, with the regression line passing through the origin. Third, our sample of wealthy individuals, albeit biased, contains enough data points to accurately estimate the slope of the regression line. Fourth, our information regarding the national average personal wealth is accurate.

Once explicitly laid out, skeptical researchers can challenge these assumptions methodically. For example, some may argue that wealth has only limited influence on life expectancy, or that the influence is moderated by the nature of the person's job. Others may argue that the contribution of wealth to life expectancy has a diminishing marginal return, where the increase in wealth produces less impact for wealthier individuals than for poorer ones. On the one hand, each of these challenges cast doubt on our claim that the ratio estimator is adequate. On the other hand, each of these doubts can be addressed with auxiliary evidence.

The same thought process can be used for preemptive improvements of samples, too. For example, I may believe that, in addition to wealth, the number of children a person has also predicts their life expectancy. If such information is not difficult to obtain, I may decide to have my sample of wealthy individuals balanced on the number of children they have even if I do not have perfect evidence concerning the nature and extent of the effect of children, with the knowledge that this additional act of balance is always beneficial. Furthermore, I can even perform a kind of cost-benefit analysis between convenience and theoretical improvements. Suppose, for example, that I have reasons to believe that the relative importance between job type and number of children to a person's life expectancy is comparable, and yet information regarding job type is much more difficult to obtain. I may choose to balance my sample against the number of children but not job type. This will make my inference less than perfect, but still better than using wealth alone.

This style of thinking is already present in the use of post-stratification, if only implicitly. Recall that post-stratification is a method aimed at balancing an already-drawn sample along some selected characteristics. The method is widely used in the design-based setting, but receives no theoretical guidance from the framework. Consequently, the choice of which characteristics to post-stratify against tends to be inconsistent across similar survey efforts. From the model-based perspective, however, post-stratification makes perfect sense. Indeed, ratio estimation from a biased sample can be seen as a form of post-stratification (Smith, 1991).

From the model-based perspective, researchers should post-stratify against characterstics they believe to be statistically relevant to the target variable. Limited by the availability of auxiliary data, researchers may choose to post-stratify against only variables they believe to contribute significantly or feel that they have strong enough evidence for believing so. Differences in such subjective thresholds can lead to inconsistencies in post-stratification decisions across similar survey efforts, as observed.

To summarize, model-based inference in sampling relies on assumptions concerning the relationship between control and target variables. To guard against possible inaccuracies in these assumptions, a sample should be balanced, either through purposive design or post-stratification. An ideally balanced sample – one that is representative in the "miniature population" sense – guards against many forms of model misspecifications, whereas an approximately balanced sample approximately guards against model misspecifications. The adequacy of the model-based framework is attested by its ability to account for both the intuitive justification and the practical inconsistencies observed with post-stratification.

In extremely resource-limited cases, even an approximate balance may not be feasible. Suppose I am using a sample to study how much people, in general, are willing to share their newly acquired wealth with a stranger. I may have some suspicions, evidentially justified or not, that certain characteristics could affect the extend of giving in a systematic way. For example, perhaps those who have gone through financial hardships themselves are more likely to empathize and share with strangers. Note that these suspicions appear a lot like independent variables in an experiment – indeed, one could systematically study altruistic behavioural differences between the rich and the poor. However, even when between-group differences are not of theoretical interest, cross-group balance is still important from the perspective of sample representation.

Nevertheless, I may not have a good understanding of how levels of wealth affect altruistic behaviour or a feasible way of balancing my sample across levels of wealth distribution. Moreover, it is highly likely that, even if wealth plays a role, its relationship with altruism accounts for only a small proportion of the total variation, and I may not have any idea at all what other variables are worth controlling for.

Situations like these are common in the social sciences. While balancing the sample against one more variable takes us closer to the ideal of full representation, controlling variables that only account for a minority of the total target variance is not sufficient to eliminate systematic bias. However, the benefit of the model-based framework is not about meeting the same standard with less content, but rather documenting and systematizing available and unavailabe information in a way that makes assessment possible.

Suppose I am able to secure participants at the top and bottom levels of the wealth hierarchy. This act of balancing accounts for wealth if wealth is linearly related to altruism, but not if they are quadratically related. If later research finds the relationship to be quadratic, then others can reasonably question the accuracy of my results. Similarly, if later research finds that age, a variable I have not controlled for, is also statistically related to altruism, then that would similarly constitute a weakness of my

initial estimation.

More importantly, reasoning like above allows for better synthesis of similarly aimed research. Suppose I conduct a study on altruism with a sample controlling for wealth only, and another research team conducts a similar study controlling for age only, and that our results are very similar. The model-based framework allows us to infer that, if someone had drawn a sample balanced on both wealth and age, they would have also gotten similar results. While multiple nonprobabilistic samples cannot be combined to form a probabilistic sample, multiple samples balanced in different ways can be combined to form a sample that is balanced on all of those ways. The model-based framework, therefore, allows for a systematic integration of resource-limited studies.

5 Sample representation in the social sciences

In an attempt to address the problem of questionable research practices in psychology, Simons et al. (2017) proposed that research papers should be required to include "Constraints on Generality" (COG) statements in their methods sections. They describe their vision as follows (p. 1124),

> A COG statement specifies your intended target population and the basis for believing that your sample is representative of it; it justifies why the subjects, materials, and procedures described in the method section are representative of broader populations.

Focusing on the sampling aspect alone, the proposal provides little guidance asside from *justifying why the sample is representative*. In this paper, I have discussed two senses of sample representation: the design-based approach where a representative sample is one drawn randomly and the model-based approach where a representative sample is balanced on all features relevant to the research target. I have further argued that the ideal versions of both senses of representation is infeasible for most research groups. Demanding a research team to explain why their sample, gathered with severe resource limitations, is representative is unlikely to lead to tangible improvements. We need proposals that are more feasible for small-scale research efforts.

Despite the dominence of the design-based framework in the social sciences, the idea that a representative sample is one that is balanced on relevant features is not foreign. Studies that use samples often report participants' demographics, which is how the literature was able to detect the lack of sample representation in the first place. However, few, if any, document reasons for why they choose to report the type of demographics that they do. As in the case of post-stratification, it seems reasonable to suppose that researchers are making these decisions based on the intuition that a more balanced sample is a better sample. If we take a model-based perspective, we can begin to unpack these implicit assumptions and question their adequacy.

Because of the dominance of the design-base framework, most metascientific studies on sample representation focus on sampling method rather than sample composition[11]. Nevertheless, a few studies have examined the practice of demographics reporting. In an analysis of all studies published in four pediatric psychology journals in 1997, Sifers et al. (2002) reported that "participants' ages, genders, and ethnicity were reported at moderate to high rates, whereas socioeconomic status was reported less often". Of the 260 papers they analyzed, gender was reported in 86.2% of the papers, whereas SES was reported in 46.5%. A similar review of studies published in *Psychological Science* in 2014 found that, while gender is reported in 75% of the studies, education levels is reported in only 52%, race/ethnicity is reported in 20%, and SES in only 8% (Rad et al., 2018).

Reporting sample demographics, even without explicit efforts at balancing or post-stratification, allows later researchers to better assess the overall coverage of the literature. That said, asking researchers to report "as many demographics as possible" is also infeasible. A lengthy demographics questionnaire attached to all studies is likely to cause cognitive fatigue in participants, harming study validity. There are also privacy concerns over potential reidentification through aggregated demographics data.

In other words, the control over sample demographics, be it actual balancing or mere reporting, requires deliberate planning. This is especially true in studies with small samples that are all gathered from the same location and through the same method, both of which instrinsically limit the diversity of the sample. Consequently, researchers need to be deliberate in choosing which control variables to report.

According to the model-based framework, an estimation is unbiased just in case all of the *statistically relevant* variables have been controlled for. This means what variables are worth controlling will change depending on what the research target is[12]. Instead of asking researchers to always report as many control variables as possible, it is more effective to report only a few that are considered statistically relevant to the target and explicitly justify them as such.

Furthermore, to assess the balance of a sample, researchers need auxiliary data concerning population-level composition. Balance is important for any study aimed at sample-to-population generalization, even when demographics is not part of the research interest. Consequently, researchers of human subjects in any discipline should pay attention to how individual characteristics systematically affect behaviour, as well

[11] Although the acronym "WEIRD" refers to a set of demographic features, the metascientific data Henrich et al. (2010b) relied on primarily concerned *where* samples were drawn, e.g., from undergraduate psychology classes at the researchers' universities, supplemented by secondary data on the demographics of students of such universities.

[12] Interpreted from this perspective, the preferential reporting of gender as a control variable brings up a series of questions concerning the presumed roles (and the presumed univocality of such roles) gender plays in shaping behaviour. Similar observations can also be made about the overreporting of some demographic variables and the underreporting of others. Indeed, since design-based principles cannot guide reporting or poststratification, culturally entrenched ideologies often substitute for this role. The philosophical implications of this dynamic are beyond the scope of the current paper but will be the subject of future work.

as how such characteristics are measured in full enumeration survey efforts. Sociologists are beginning to notice the mismatch between the changing societal understanding of sex and gender and the traditional ways of measuring them (Westbrook and Saperstein, 2015; Hart et al., 2019). Similar forms of close scrutiny of the methodology and assumptions underlying demographic surveys should become a bigger part of all areas of the social sciences, not just demography.

6 Conclusion

The social sciences face a persistent problem of sample nonrepresentation with no trend for improvement. I believe this is due to a lack of feasible proposals for resource-limited contexts. By tracing the history of sampling, I showed how the design-based framework for sampling where random selection is the gold standard, although good on paper, provides little practical guidance when the gold standard cannot be achieved. In contrast, the model-based framework provides a systematization of all assumptions, allowing them to be challenged and defended methodically. It also offers guidance on how small-scale studies with imperfect samples can be integrated for greater understanding.

Accordingly, I have made two practical proposals for the improvement of sample representation in the social sciences. First, instead of inconsistently reporting a more-or-less identical set of sample demographics, researchers should deliberately select a few that they believe to be statistically relevant to their research target and explicitly justify them as such. Second, there should be greater communication between scientists studying human behaviour and demographers designing full enumeration survey efforts.

Acknowledgements

I am grateful to Cailin O'Connor, Simon Huttegger, Michael Schneider, Greg Lauro, William Stafford, Jan-Willem Romeijn, the members of the philosophy of statistics reading group (in particular, Conor Mayo-Wilson, Samuel Fletcher, and Kathleen Creel), the audience at the Greater Cascadia HPS Workshop, and the audience at the University of Washington for invaluable discussion and encouragement, and two anonymous reviewers for the helpful comments.

Bibliography

Andrews, G., Henderson, S., and Hall, W. (2001). Prevalence, comorbidity, disability and service utilisation: overview of the australian national mental health survey. *The British Journal of Psychiatry*, 178(2):145–153.

Arnett, J. J. (2008). The neglected 95%: why american psychology needs to become less american. *American Psychologist*, 63(7):602.

Bornstein, M. H., Jager, J., and Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, 33(4):357–370.

Brewer, K. (1999). Design-based or prediction-based inference? stratified random vs stratified balanced sampling. *International Statistical Review*, 67(1):35–47.

Brewer, K. et al. (2013). Three controversies in the history of survey sampling. *Survey Methodology*, 39(2):249–262.

Bryson, M. C. (1976). The literary digest poll: Making of a statistical myth. *The American Statistician*, 30(4):184–185.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Gächter, S. (2010). (dis) advantages of student subjects: what is your research question? *Behavioral and brain sciences*, 33(2-3):92–93.

Godambe, V. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):269–278.

Godambe, V. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(2):310–328.

Graham, S. (1992). " most of the subjects were white and middle class": Trends in published research on african americans in selected apa journals, 1970–1989. *American Psychologist*, 47(5):629.

Hall, C. C. I. (1997). Cultural malpractice: The growing obsolescence of psychology with the changing us population. *American Psychologist*, 52(6):642.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384):776–793.

Hart, C. G., Saperstein, A., Magliozzi, D., and Westbrook, L. (2019). Gender and health: Beyond binary categorical measurement. *Journal of health and social behavior*, 60(1):101–118.

Henderson, S., Andrews, G., and Hall, W. (2000). Australia's mental health: an overview of the general population survey. *Australian and New Zealand Journal of Psychiatry*, 34(2):197–205.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., et al. (2010a). Markets, religion, community size, and the evolution of fairness and punishment. *science*, 327(5972):1480–1484.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):6183.

Jacobi, F., Wittchen, H.-U., Hölting, C., Sommer, S., Lieb, R., Höfler, M., and Pfister, H. (2002). Estimating the prevalence of mental and somatic disorders in the community: aims and methods of the german national health interview and examination survey. *International journal of methods in psychiatric research*, 11(1):1–18.

Kessler, R. C. (1994). The national comorbidity survey of the united states. *International Review of Psychiatry*, 6(4):365–376.

Kruskal, W. and Mosteller, F. (1980). Representative sampling, iv: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique*, pages 169–195.

Likert, R. (1948). Public opinion polls. *Scientific American*, 179(6):7–11.

Mickelson, K. D., Kessler, R. C., and Shaver, P. R. (1997). Adult attachment in a nationally representative sample. *Journal of personality and social psychology*, 73(5):1092.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of consumer research*, 28(3):450–461.

Pollet, T. V. and Saxton, T. K. (2018). How diverse are the samples used in the journals evolution & human behaviorand evolutionary psychology? *Evolutionary Psychological Science*, pages 1–12.

Rad, M. S., Martingano, A. J., and Ginges, J. (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, 115(45):11401–11405.

Rao, J. and Fuller, W. A. (2017). Sample survey theory and methods: Past, present, and future directions. *Survey Methodology*, 43(2):145–160.

Royall, R. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63(324):1269–1279.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387.

Royall, R. M. (1992). The model based (prediction) approach to finite population sampling theory. *Lecture Notes-Monograph Series*, 17:225–240.

Royall, R. M. and Herson, J. (1973). Robust estimation in finite populations. *Journal of the American Statistical Association*, 68(344):880–893.

Scheaffer, R. L., Mendenhall, W., and Ott, L. (1971). Elementary survey sampling. belmount.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, 51(3):515.

Segall, M. H., Campbell, D. T., and Herskovits, M. J. (1966). The influence of culture on visual perception.

Seng, Y. P. (1951). Historical survey of the development of sampling theories and practice. *Journal of the Royal Statistical Society. Series A (General)*, 114(2):214–231.

Sifers, S. K., Puddy, R. W., Warren, J. S., and Roberts, M. C. (2002). Reporting of demographics, methodology, and ethical procedures in journals in pediatric and child psychology. *Journal of Pediatric Psychology*, 27(1):19–25.

Simons, D. J., Shoda, Y., and Lindsay, D. S. (2017). Constraints on generality (cog): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6):1123–1128.

Smith, T. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society: Series A (General)*, 139(2):183–195.

Smith, T. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General)*, pages 394–403.

Smith, T. (1991). Post-stratification. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 40(3):315–323.

Srivastava, A. (2016). Historical perspective and some recent trends in sample survey applications. *Statistics and Applications*, 14(1-2):131–143.

US Census Bureau (1989). *Statistical abstract of the United States, 1989*. Bureau of the Census.

Westbrook, L. and Saperstein, A. (2015). New categories are not enough: Rethinking the measurement of sex and gender in social surveys. *Gender & Society*, 29(4):534–560.

Wintre, M. G., North, C., and Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology/Psychologie Canadienne*, 42(3):216.