Expl Agric. (2019), volume 55 (4), pp. 621–636 © *Cambridge University Press 2018*
doi:10.1017/S001447971800025X

DESIGNED EXPERIMENTS: DO YOU KNOW WHAT POPULATION YOU ARE SAMPLING FROM?

By MARCIN KOZAK†§‡ *and* HANS-PETER PIEPHO¶

†Department of Botany, Warsaw University of Life Sciences – SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland, §Department of Qualitative and Quantitative Studies, University of Information Technology and Management in Rzeszow, Sucharskiego 2, 35-225 Rzeszów, Poland and ¶Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70593 Stuttgart, Germany

(*Accepted 16 May 2018*; *First published online 21 June 2018*)

SUMMARY

Consider a field experiment laid out in a randomized complete block design in which you study three types of fertilizers for two winter wheat cultivars. One year, one location – the experiment is not repeated. You design it and then spend a lot of time and money to conduct it. You cultivate the soil and take care of the plants; you worry about them; you never know what can happen, so you cannot wait for the crop to be harvested. And that day finally comes. The crop is harvested, and everything is fine. And here you are, all went great, you have the data in hands, and now a simple thing to do – analyse them. Well, yes, the experiment was conducted in one year, and you are aware you cannot be sure the outcome would be the same next year or elsewhere, but whatever – suffice it to consider the conclusions as preliminary and get on with the interpretation. Why should you not? It was a properly designed experiment that took samples from the underlying infinite populations of the two winter wheat cultivars in the three water regimes studied. Statistics is here to help you out, is not it? Well, it is not. Statistics will not help you out whether the experiment was poorly designed. Agricultural science literature seldom explains what populations are studied and what types of samples are taken in designed experiments. To fill this gap, we discuss various aspects of the sampling process in designed experiments. In doing so, we look at the survey sampling methodology, a statistical framework for studying finite populations – we do this because survey sampling has developed into the advanced theory of sampling processes, and this background can help us understand the intrinsic aspects of sampling in designed experiments.

INTRODUCTION

Designed experiments are perhaps the most common means for studying biological phenomena. They come in handy when a researcher needs to control conditions in order to study how a particular set of treatments affects the biological phenomenon of interest. Such controlling indeed helps the researcher learn the isolated effects of the treatments, an approach that is both advantageous and disadvantageous. Advantageous, because we can learn their isolated effects. Disadvantageous, because we cannot learn how the treatments behave in reality – in which they do interact with other factors.

[‡]Corresponding author. Email: nyggus@gmail.com

Developed since the beginning of the twentieth century, the theory of designed experiments is quite advanced. It focuses mainly on developing appropriate designs and statistical methods to analyse results of these experiments. Thanks to this development, researchers can choose from a variety of designs and statistical methods for analysis.

Given how popular designed experiments are and how advanced their theory is, however, it is surprising that the topic of defining and understanding populations from which samples are taken by means of designed experiments is so underdeveloped. Should not this understanding be an integral part of the design and the interpretation of experimental data? The rich literature on designed experiments does not offer much help: textbooks on designed experiments either ignore the subject altogether or merely mention it somewhere at the beginning, without giving it serious attention (for an exception, see Mead *et al.*, 2012, Chapter 19). For this reason, the literature does not give much advice on how to define the population from which samples are to be taken by means of a designed experiment, nor on how to understand more advanced aspects of the sampling process, such as randomness of a sample, probabilities of selection of a particular population element to a sample, what actually constitutes a sampling element and what actually is and what is not random.

Maybe the authors of these books consider this knowledge as obvious? But is it obvious indeed? For a particular biological designed experiment, is it obvious how to define the underlying population? How to understand the sampling process from this population? How to understand the random character of this process?

Perhaps, then, the best we can do is accept that understanding the sampling concept underlying designed experiments is actually unnecessary to both conduct and analyse designed experiments.

We challenge such thinking: in our opinion, understanding this concept, which is a statistical backbone of designed experiments, is crucial. It is not obvious, however.

Although indeed designed experiments are difficult to explain in terms of the various faces of the sampling process (like a population from which sampling is done or selection probabilities of the population elements), this is not so with survey sampling. In fact, as its name suggests, survey sampling – sometimes termed the *representative method,* a name used by Jerzy Neyman back in 1934 (Neyman, 1934) – is mainly about sampling. To conduct a sample survey, one

1. defines the population;
2. constructs the sampling frame (a list of population elements to sample from);
3. chooses the sampling scheme (in most schemes, each population element has a known probability of being selected to a sample) ;
4. samples from the population; and
5. makes inferences about the population, using the knowledge of the population and sampling scheme.

From this simplified outline of a sample survey, it is clear that sampling is essential in this context. This is why survey statisticians have focused on sampling processes in survey sampling, while statisticians working on experimental design focused

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

on randomization, the backbone of designed experiments. Comparing these two frameworks for obtaining data and focusing on how they treat the sampling processes can help us better understand those processes for designed experiments. Hence, this paper aims to discuss various facets of sampling by means of designed experiments, which we will do against the background of survey sampling.

DESIGNED EXPERIMENTS VERSUS SURVEY SAMPLING: BASIC COMPARISON

In a broad context, there is just one type of *statistics*, meant here as a branch of science. This definition includes both Bayesian and frequentist statistics, but we will deal with the frequentist branch of statistics only. We will start off by comparing the basics of designed experiments and survey sampling. However, we will not present details of both methodologies, which the readers can find in numerous textbooks. For survey sampling, examples are Kish (1967), Cochran (1977), Thompson (2002), Särndal *et al.* (2003) and Singh (2003a; 2003b). For designed experiments, examples are Hinkelmann and Kempthorne (1994), John and Williams (1995), Caliński and Kageyama (2000), Bailey (2009), Mead *et al.* (2012) and Welham *et al.* (2015).

Sample surveys and designed experiments are two different statistical methodologies to study populations by means of quantitative methods. Most researchers work either with sample surveys or with designed experiments, but few work with both. Statistical textbooks and courses usually focus on one of the methodologies, seldom joining them together. This might seem strange, but we do not think it actually is. Survey sampling and designed experiments are so complex methodologies that comprehensible but insightful textbooks on just one of the two topics run for hundreds of pages. What is perhaps more important, the two methodologies differ so much not only in statistical methods used, but also in aims that a comprehensive book covering both survey sampling and designed experiments would likely get overly complicated, perhaps even confusing. For these reasons, the knowledge of these two frameworks is seldom joined even though such joining might help us develop them better.

By and large, statistics aims to help researchers infer about a studied population based on either a sample from this population or a complete enumeration. The latter scenario amounts to studying the whole population, a rare occasion in real life, censuses offering an example. The former scenario, thus, is the essence of statistics to make inferences about a population based on a finite sample of elements from this population.

Most of what we wrote till now equally applies to designed experiments and survey sampling, the only exception being complete enumerations, impossible to conduct in designed experiments. As we already mentioned, however, there are two different types of statistical methodology. While they share the aim of analysing a population based on a sample, they differ in the following crucial aspects:

i. How are populations treated: as finite or infinite?
ii. How are samples taken from populations?

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

iii. How is statistical estimation and inference conducted?
iv. How are estimates interpreted and generalized?

We will compare these aspects in designed experiments and survey sampling, hoping that the clear understanding of the sampling processes in survey sampling will help us understand these processes in designed experiments.

First, let us brainstorm what connotations *survey sampling* and *designed experiments* have for us.

Survey sampling:

> first-thought connotations: random sampling, population parameter, finite populations, official statistics, statistical offices, census, survey, studying nations and studying citizens.
>
> second-thought connotations: sampling designs, stratified sampling, finite-population factor, quota sampling, internet panel, questionnaire, CATI (computer-assisted telephone interview), CAPI (computer-assisted personal interview) and CAWI (computer-assisted web interview).

Designed experiments:

> first-thought connotations: randomization, effect sizes (treatment differences), comparative experiments, infinite populations, statistical analysis, experiments, blocks, random error, biological experiments, medical experiments, psychological experiments, analysis of variance and factor.
>
> second-thought connotations: nested/hierarchical designs, crossed designs, interaction and linear model.

These connotations themselves suggest that the two methodologies differ a lot. Still, they both deal with samples from populations and are analysed with statistical methods. We will thus compare them with the aim of understanding various sampling aspects of designed experiments.

Finite versus infinite populations

Perhaps the main intrinsic difference between survey samples and designed experiments is that the former deals with finite populations, while the latter deals with infinite populations, an important difference in biology (Kozak, 2008).

The concepts of finite and infinite populations require one to precisely define a population of interest. Such a definition can be tricky, particularly in the case of infinite populations. As we discuss further below, the concept of infinite populations has rather theoretical meaning. This is because common ways of obtaining samples (like the one in a single designed experiment) do not enable us to draw truly random samples from infinite populations. This seldom-noted fact has important consequences: in practice we are usually limited to (i) vaguely defined infinite populations or (ii) precisely defined infinite populations from which we cannot draw random and independent samples or (iii) precisely defined infinite populations but limited to so small a scale that they are uninteresting for researchers. Yet, the analysis of data from randomized experiments almost invariably makes the implicit assumption of some infinite population from which observations obtained in an experiment were sampled (Welham *et al.*, 2015, p.28ff.).

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

Let us turn to finite populations. As their name suggests, finite populations have a finite number of elements. But in practice, each population is unstable over time, so it should be considered for a particular period, its length depending on the stability of the population. For example, the (finite) population of agricultural farms in a particular country can change on a daily basis. Therefore, this population might be considered for a particular day. This would be impractical, however, for various reasons, so national statistical agencies usually consider such populations for a particular year. In our example, it would be the population of agricultural farms in the country that were registered in this particular year.

Unlike finite populations, infinite populations have an infinite number of elements (units). That is what theory says. In practice, however, this theory can become difficult to apply – defining a population can seem trivial, but eventually can become quite tricky. Consider a population of plants of a particular wheat cultivar. Agricultural researchers usually treat such a population as infinite, either implicitly (more often) or explicitly (less often). In so doing, they do not limit the population to the plants that grow in a particular period and location when and where the experiment was conducted. Instead, they include all plants of this cultivar *in general*, so the population contains all plants of the cultivar that were grown in the past, plants that are being grown in the experiment and all *possible* plants that can be grown in the future. For laboratory experiments with plants, populations are usually defined in a similar way.

Note how vaguely such populations are defined. Any field experiment with plants from a population defined this way would have a limited scope, dealing with plants grown in one particular experiment or in several experiments (for example, conducted in different locations). A laboratory experiment with plants from a population defined this way would be equally limited in scope because it would deal with plants grown in the conditions of this particular laboratory (including its staff, equipment, etc.).

Now the question is: Can we treat plants grown in a single experiment – whether field or laboratory – as a random sample from an infinite population whose definition is not limited to plants that can be grown in this particular experimental field and in these particular weather conditions or in this particular laboratory's conditions? Such an approach, common in the interpretation of agricultural and biological experiments, implies extrapolation: the knowledge gained from the experiment is extrapolated to the infinite number of possible experiments for the studied population. Now, does not this extrapolation go too far?

In the above paragraph, we moved from the infinite population of plants to the infinite population of experiments. Although directly related, these two populations differ. We discuss this issue in the next section.

Sometimes an assumed infinite population is limited, thanks to which it seems to be less vague. It *seems* less vague, but it *is not* less vague. For example, the population of wheat plants of a given cultivar is limited to plants that could be grown in a particular year or location; again, this definition of the population does not refer to the plants that are grown, but that *might be grown*. Such limitation of the population does not change three things. First, such populations will still be infinite, with an infinite number of possible elements (units). Second, although limited, such populations will

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

still be vaguely defined. And third, *no* experiment will be able to draw random and independent samples that would be representative of the whole population, regardless of how it is defined.

From the above discussion, it is clear that while defining underlying populations in survey sampling is relatively easy, it is not so in designed experiments. We need to add one more aspect to this difficulty, however. Consider a designed experiment comparing the efficacy of three herbicides in wheat, with an untreated control. Does this experiment study one population? Or perhaps more – four (three treatments and the control), to be precise? Generally put, the one population would relate to all plants being represented by the experiment. The four populations would relate to all plants being represented by the combination of the experiment and the four treatments (the three herbicides and the untreated control).

In this example, some people might see just one population, while others see four. But the two opposing views can be reconciled. When thinking about a designed experiment, we can treat it as a way of obtaining samples from one general population that is divided into subpopulations due to treatments. Hence, in our example, we would have one general population of wheat plants being represented by the experiment. This population, however, consists of three subpopulations of wheat plants treated with one of the three herbicides; again, each subpopulation covers plants that are represented by the experiment. This is perhaps the easiest way of looking at the populations being studied in designed experiments, affecting neither statistical inference nor the generalization of results. Of course, in blocked experiments, we have additional subdivision of the population due to blocks, so we should be talking about subpopulations due to blocks *and* treatments. Such thinking might make it easier to understand assumptions of statistical methods used to analyse designed experiments. For example, in a linear model, we assume equal variances and normality in subpopulations defined by blocks and treatments.

In summary, designed experiments aim to study infinite populations. Such populations, however, are often difficult to define, unlike in survey sampling. Even if we do manage to define a population of interest, any experiment – considered a sample from this population – will unlikely cover it all but will rather cover its specific subpopulation or subpopulations that result from where and when and how the experiment is conducted. Such a single experiment cannot even be regarded as a representative sample from the population.

Thus, when designing an experiment, one should pay particular attention to its representativeness. First, the population (of the possible experiments) must be precisely defined. Second, the experiment should be designed in a way that does not make the experiment a one-element sample from the underlying population. Third, whenever possible, the locations (or, in general, the repeats of the single experiment) should be drawn randomly. Fourth, their number should not be too small – for instance, an experiment repeated in two subpopulations could not cover the whole population of such possible experiments, even when they were selected at random; two is just too small a sample.

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

TAKING SAMPLES FROM A POPULATION

Over the centuries, various rigorous methods have been developed to take samples from finite and infinite populations. These methods differ a lot. Methods for sampling from finite populations go under the general name of survey sampling. A variety of methods have been developed for random sampling, from the simplest ones such as *simple random sampling with or without replacement* to more advanced ones such as *stratified sampling* or *unequal probability sampling* to complex ones such as *sampling according to rotation patterns with varying overlap of sampling units in the subsequent sampling periods*. There are also non-random sampling designs, the most important being quota sampling. Exploited by many commercial agencies in market research, non-random designs come in handy when random sampling is difficult to conduct, which can be for various reasons. In such instances, quota sampling can be used to take a sample that reflects specified proportions of particular segments of the population (e.g., proportion of women and men, employed and unemployed or age groups). Here, however, we should focus on methods of random sampling, since designed experiments do not – or rather should not—use non-random sampling.

Methods for taking samples from infinite populations through designed experiments are different, which does not mean simpler. Designed experiments offer various ways of taking samples, from the simplest ones such as *the completely randomized design* (useful, for example, in some laboratory experiments), to *randomized complete block designs*, to *randomized incomplete block designs*, to *complex nested designs repeated in different conditions*.

Since finite populations are relatively easy to define, survey samples are usually taken with random sampling methods (excluding censuses and non-random sampling designs). But infinite populations are difficult to define, and so designed experiments cannot use equally simple measures to sample from them; instead, in designed experiments, samples are taken mainly based on *randomization*, a half-measure that is to substitute random sampling. The only randomness in such experiments is usually during the allocation of treatments to experimental units; in addition, when the experiment is repeated in several locations, however, there might be randomness in the selection of locations. In a single designed experiment – whether in experimental fields or greenhouses or laboratories or clinical trials – *there is no actual random sampling of elements from the population*.

That is an important observation, given that the underlying frequentist theory seems to assume random samples. Does this affect the inference from designed experiments in any way? Let us again consider field experiments, in which we are modelling plots as independent, an approach that seems to imply random sampling. Note, however, that we are *not* sampling the plots at random from any larger population of plots. Instead, we select one single trial area on an experimental station or a farmer's field (or several such areas). Given this area, the plots are fixed (so, not random) and constitute a one-element sample of possible areas/fields (henceforth, we will call such areas 'fields', to follow the standard terminology used in agricultural experimentation). We see then that our sample is *not* a random sample of all possible

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

plots; it is a one-element sample – usually non-random – from the population of areas/fields. What is random is just the allocation of treatments to the plots within the same single field, but this randomness must not be confused with random sampling of plots from the underlying infinite population of plots.

Note we are using two levels of units and their associated populations. A first level is one concerned with plots in a particular field or, say, Petri dishes (or any other experimental unit) in laboratory experiments. We might say this is a basic experimental unit. A second level is one concerned with the repeats of experiments – for instance, a field experiment can be repeated in various locations and/or various years, while a laboratory experiment can be repeated in various laboratories or in the same laboratory but in different runs. Experimental fields may or may not be randomly sampled. Once we have selected an experimental field, we have also selected all plots within that field, so there is no random sampling of plots from a larger population of plots. Instead, we use randomization to allocate treatments to the given plots.

Thus, since a single experiment (which is repeated neither in locations nor in years) is a one-sample experiment from the infinite population of possible experimental fields or weather regimes, we cannot assess the between-experiment variance. Instead, we are stuck with assessing the within-experiment variance, which is of little interest: it says practically nothing about the variability in the population we want to study. Likewise, a laboratory experiment conducted in one laboratory is a one-sample experiment from the infinite population of laboratories.

Based on the above discussion, we can conclude that such a non-repeated experiment *interpreted alone* will be of little value for interpretation of the underlying population. If the selection of an experimental area can be considered random (in other words, if the experimental site has been selected randomly or if the laboratory has been selected randomly, an unrealistic situation in agricultural and biological research), then at least this one-sample element is random. But this randomness does not change anything – because it is still a one-element sample. Even if we grant that the area or laboratory is selected at random (even though it almost never is), the sample of plots is a cluster sample with just one cluster. Thus, the assumption that the experimental area is selected randomly is inconsequential because it does not provide any basis for statistical inference about the comparison of the treatments studied. More formally speaking, with a single experiment, all the treatment information is in the plot (or Petri dish or whatever constitutes an experimental unit in an experiment) stratum, not in the experimental area stratum.

It is worth recalling what Jerzy Neyman reasoned about single designed experiments: the population of plants we are sampling from is bounded by the plots included in the sample. Again, this shows that in a designed experiment, sampling is *not* done from the whole population defined that way, but its subpopulation – the one limited to the plots included in the experiment (Speed, 1991).

Thus, randomization in randomized trials plays a similar role as random sampling in surveys, but the differences are clear. What are the consequences? We will study this in the next sections.

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

STATISTICAL INFERENCE

Survey statisticians need to know the probability of drawing each population element into a sample. Without these probabilities, they cannot make point estimates of the studied population parameters as well as estimate the estimates' variance, a crucial piece of information for national offices and all those who conduct representative sample surveys. Thus, surveys are usually designed with careful plans so that the statisticians know these probabilities.

What about designed experiments? As we already mentioned, agricultural researchers conducting a designed experiment usually treat it as a way of sampling from the underlying infinite population of plots. But do they need to know the probability of drawing each population element into the sample? Of course not, for the simple reason that it is impossible: the number of elements of the population is infinite, so it is impossible to say anything about the probability of drawing a particular population element to the sample. This does not mean, however, that we should ignore this subject altogether. Even though we need not know these probabilities, we would like them to be equal – and in fact, we assume them to be equal. In other words, we would like to know that all population elements have the same chance (probability) to be selected into the sample.

Now the question is: given our discussion of how populations are defined when designed experiments are conducted, can we assume that all the elements of the corresponding population indeed have the same probabilities to be included in the sample? The answer is: not only can we not assume that, but also we *know* that it is not true. We know that while conducting a designed experiment, we sample from a subpopulation of the original population, a subpopulation bounded by the conditions of the experiment (e.g., plots in the field experiment in a given year or laboratory conditions in the laboratory experiment).

In survey sampling, inference mainly aims to estimate population parameters of interest and assess this estimation's precision. For estimation, we need to know the above-mentioned selection probabilities, which we usually do know based on the knowledge of the sampling design used. Estimators of most common parameters—such as total value, mean and proportion—are intuitive and simple in construction, but a lot of energy has been spent to derive asymptotically unbiased and more precise estimators than the classical ones. Basic examples of such estimators are product, ratio, and regression estimators. Although biased, they can offer much more precise estimation – and, as mentioned, precise estimation is one of the most important virtues of a good survey. This precision is represented by either an estimator's variance (in the case of unbiased estimators) or an estimator's mean-square error, MSE (in the case of biased estimators).

It is estimation and its precision that lie at the heart of survey sampling. All around – like dealing with missing data or estimating parameters for so-called small areas, for which we have no or insufficient information – serves the purpose of making estimation possible or simpler or more precise. Other methodological issues, such as interval estimation or hypothesis testing, are seldom used in sample surveys. This is

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

not to mean that they are totally ignored, but that they are not as interesting as in other branches of statistics. For instance, even though one can estimate confidence intervals, they are uncommon in sample surveys – instead, as mentioned above, variance or MSE is usually used to represent the estimation's precision. As for hypothesis testing, some textbooks offer variants of simple tests for particular designs – we have to remember that because of the complexity of survey designs, classical statistical tests do not apply and must be tailored to a design's needs. This might be one reason for the limited popularity of hypothesis testing in survey sampling. Another one might be that survey sampling seldom calls for hypothesis testing — this is estimating what is of crucial importance.

In survey sampling, thus, statistical inference aims to estimate parameters and the precision of these estimates. In designed experiments, statistical inference is much richer. It still aims to estimate parameters and the precision of the estimates, but this is just the first step of a bigger plan. Designed experiments are comparative, which means they aim to compare. They can compare treatments, times, experiments, conditions, etc. So, unlike in survey sampling, in designed experiments often it is hypothesis testing that plays a crucial role: in most situations, it is considered the main statistical tool for making comparisons, although many statisticians have emphasized the importance of other statistical tools for comparisons, such as simultaneous confidence intervals (e.g., Gardner and Altman, 1986; Läärä, 2009; Wasserstein and Lazar, 2016; Greenland *et al.*, 2016; Kitsche and Schaarschmidt, 2015).

For statistical tests to work, various assumptions should be fulfilled. For example, the *t*-test for comparing means of two variables assumes the normality of the variables, the independence of observations and, depending on which version of the test we choose, equal or unequal variances of the variables. The test for Pearson's correlation coefficient between two variables assumes not only that the variables are normally distributed, but also that they come from the same population. Each statistical test assumes something. Even the jackknife and the bootstrap (Efron and Tibshirani, 1993), supposedly free of distributional assumptions, do make two important assumptions: that the observations are identically and independently distributed and that the estimators being used for the comparison of parameters are random variables and have their own probability distributions (but we do not have to know these distributions: it suffices to be aware that they exist). These probability distributions of estimators must result from something. In survey sampling, for example, they result from the randomness of samples. Although estimators based on non-random samples vary from sample to sample, their distribution is in fact non-random: in the frequentist (i.e., non-Bayesian) world, there is no statistical sampling distribution for a statistic we compute from a non-random sample, so, in other words, such a sample statistic would have no stochastic properties. These only come with random sampling, which generates this stochastic distribution. Thus, non-random samples violate the main assumption of statistical inference – the one of randomness.

Taking the randomization distribution as the point of departure and deriving the appropriate null distribution for an *F*- or *t*-statistic, it turns out, however, that that distribution is pretty close to the one which holds for a linear model with fixed

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

treatment effects and independent error effects. This lucky coincidence makes it possible to apply standard statistical techniques, such as linear modelling, to designed experiments (Hinkelmann and Kempthorne, 1994, Sections 5.4 and 6 in particular).

Till now, we have stressed that assumptions underlying designed experiments are crucial. They are indeed more important than assumptions related to statistical analysis, such as the assumptions of normality and variance homogeneity in ANOVA. This is because, while you can often do something to deal with violations of these statistical assumptions (one way is to transform the data, another is to use a different method of data analysis), usually you can do nothing about violated assumptions at the design level.

To summarize

i. designed experiments do not offer random samples from underlying infinite populations;
ii. the only randomness in designed experiments is due to the randomized allocation of treatments to experimental units (plots, pots, Petri dishes, etc.); in multi-environment experiments, additional randomness can be due to random sampling of fields/areas to repeat the experiment;
iii. a single experiment is actually a one-element, non-random sample from the underlying population of such experiments;
iv. only a repeated experiment (in various conditions, such as fields) offers a sample of size over 1; within-experiment replications cannot be considered a source of randomness to use in analysis on a large (experiment-wise) scale;
v. only a repeated experiment with random selection of these conditions (e.g., locations) can be used for generalization of the phenomena studied over a larger population.

GENERALIZATION OF THE ESTIMATES AND INTERPRETATIONS

Above we have discussed various statistical aspects of taking samples by designed experiments and sample surveys as well as making inferences based on these samples. The next step in any study – whether a sample survey or a designed experiment – is interpretation. This step shows another big difference between survey sampling and designed experiments. Finite populations are studied through survey sampling in the historical context: we study a finite population in a particular period and later interpret what was happening back then. That is what all national statistical offices do; for instance, they study some economical phenomena in 2017 to report and interpret it after the survey. While prediction is also of interest for governments, it is seldom done in national agencies – usually, prediction is done by someone else based on data published by the agencies. The point is, however, that prediction is not considered an aim when designing most surveys – estimation for this particular period is.

Designed experiments, on the other hand, aim to study infinite populations in the future context: we study an infinite population in order to know it better and recommend something for it. Agricultural experiments provide us with good

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

examples: We really don't care what was happening with a population when a particular study was conducted – instead, when designing and conducting such experiments, we think about the future, or rather about what seems to be the best choice in similar situations in the future. For instance, an agrotechnical experiment aims to choose the best agrotechnical treatment (from among the studied ones) for a given situation (a combination of cultivar, environmental conditions, and given other – those not studied in the experiment – aspects of agronomic management). A breeding experiment aims to choose the best cultivars, also for a given situation.

In other words, a common practice in survey sampling is to interpret estimates for a finite population as characteristics of this population for here and now, so, for example, the estimations are not generalized for the same population in the future (because this population will actually *not* be the same in the future: it will change, so it will be a different population). For an infinite population, a frequent (though not the only) practice is to generalize estimations and analyses for the whole population, which is *not* bounded by time and, often, space.

Let us focus on designed experiments. As we just said, they aim to generalize estimations and analyses for the whole population, which is not bounded by time and, often, space. From our discussion in the previous section, it follows that far too many agricultural experiments *fail* to offer such a generalization opportunity. This is because most of agricultural experiments *are bounded in time and space*, and what is more, *by conditions*. Conducted in one year, in one location, and with a particular set of agrotechnical measures, such an experiment offers no way of generalization of its results. Its replications enable us to draw conclusions about within-field variability, a piece of information that cannot be generalized beyond this single particular field and year – and so, is seldom interesting. It can be treated as a source of useful preliminary information for further experiments.

It does not mean that single controlled experiments have no role in scientific discovery. Experimentation under very controlled conditions is often the only means to establish cause-and-effect relations between two factors of interest. But once such a relationship has been established under very controlled conditions, it becomes both necessary and interesting to check if the same relations hold under a broader range of conditions, and at this stage, replication of the experiment in space and time comes into play. The bottom line is single controlled experiments do have their role in scientific method, but they are usually the first step.

One exception when a single experiment would be valuable in itself is when indeed a researcher is interested in learning the phenomena studied in a narrowly defined population, one that this single experiment can cover. In such instances, to increase the experimental variability, one can add an additional environmental factor to the experiment, such as that representing additional agrotechnical measures, or a noise factor (in whose main effects we are not interested altogether) (Mead *et al.*, 2012, Chapter 19).

A single experiment is limited in scope, but almost invariably we wish to extrapolate beyond this scope to other locations, seasons, etc. Such extrapolation is always risky and almost never based on firm grounds because interaction of treatments with

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

locations, seasons, etc. is likely to occur and cannot be ruled out on a priori grounds. Agricultural trials use few replications. Contrast two or three or four replications, a typical situation in agricultural field experimentation, with thousands or even tens of thousands of elements in survey samples. And yet, estimations and inference based on these few replications from block designs often appear to be very sound from a statistical point of view even though they represent an infinite population. Why is that? For the simple reason that they do not cover the whole population, but its small subpopulation – that related to the experimental field. And if so, the variation observed in such single experiments represents the within-field variation only, not the variation in the whole population. It is like looking in a mirror and finding a little variation in human hair colour. Generalizing the conclusion beyond this particular reflection in the mirror (to human hair of any larger population) does not make much sense, does it? Yet this is what generalizing over single experiments actually does.

Conducting such an experiment, are we indeed interested in this particular field? Not at all – we want to say something more about the underlying population. For instance, we are never interested in learning which particular cultivar will be the best yielder in a particular field. We might be interested in finding out which (say) barley cultivar will be the best yielder in varying conditions of fields on a given soil in a particular region of Germany – or in the whole country, given a set of agrotechnical measures recommended for barley. Or, we might be interested in finding out which corn cultivar will be the least susceptible to drought – not in a particular field, but in varying conditions of fields and weather in Poland. Such experiments should thus be carried out in numerous fields offering such variable conditions *and* several years, giving a chance to study the phenomena of interest in varying weather conditions and capture any treatment-by-environment interaction.

Mead *et al.* (2012) argued that the only situation when one particular experimental field can be considered sufficient is when it represents the corresponding local environment *and* when local farmers do believe in the results. Unlike most scientific experiments, such an experiment would fail to offer a wider view, being focused on practical aspects of plant production in a narrowly defined environment. In most situations, a single field experiment should rather be considered one stanza of a long poem.

From what we have discussed in the above section, it follows that to design and conduct a series of experiments, it does not suffice to just take a number of fields and conduct an experiment there. First, we do have to define the target population of environments (so-called TPE; Comstock 1977; Cooper *et al.*, 1997; Fox and Rosielle, 1982) about which we want to make inferences. Second, having defined the population, we need to be able to conduct experiments in all the environments the population covers. Third, we should construct a sampling frame of many fields and *randomly* choose a number of them in which we would conduct the experiment. Only such an approach would give us a chance to generalize the results obtained from such a multi-environment experiment.

As we already mentioned, a truly random selection of trial locations is usually difficult (Mead *et al.*, 2012) – yet most analyses of series of experiments essentially

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

make that very assumption, starting from the seminal paper by Yates and Cochran (1938); this assumption continues to be an important part of such analyses (e.g., Caliński *et al.*, 2009; Damesa *et al.*, 2017; Hu, 2015). Without this assumption, series of experiments cannot be interpreted as we would like to, which is to mean they can be interpreted solely in the context of the studied locations, without generalizing the conclusions to a wider population of locations. If so, it is clear that we should always strive for the random selection of environments from a target population of environments, no matter how hard such a task may seem.

If indeed environments have not been randomly selected, then we should *not* treat the environment factor as random in a linear model. Instead, we should treat it as a fixed factor. Note that if no other factor in the experiment is random, then the use of linear mixed effects models is illegitimate (unless the experimental design involved random effects, as, for example, split-plot does). This shows that one should always ponder how to apply a linear mixed effects model for a particular experiment. While doing so, one should pay special attention to which (if any) factor should be treated as random. In many situations, the most common approach to analysing multi-environment series of experiments—that is, using a linear mixed effects model with environments considered a random factor—may turn out to be incorrect.

When conducted at a large scale, designed experiments offer wide interpretational opportunities. Dealing with adaptation data from various subregions, for instance, we can obtain precise estimates for the whole region. Piepho and Möhring (2005) proposed a method of such estimation, which uses a concept directly taken from survey sampling – stratified sampling. Such large-scale experiments fulfil the requirements we discussed above. The experiments for each subregion do too, although their scales are smaller, and so the populations they deal with are smaller. These experiments use single-location experiments – and single-location experiments do *not* fulfil these requirements: they are just one piece of the puzzle, in itself uninformative, like one page of a hundred-page book or one verse of a poem. While one page and one verse can be interesting, they will not say much about the book or about the poem. In the very same way, one single-location experiment in a single year will not say much about any wider population, while agricultural scientists are almost never interested in just one location and year, rather wishing to have a wider scope.

CONCLUSION

Only a few statisticians have made substantive methodological contributions to both survey sampling and designed experiments, eminent examples were Jerzy Neyman, Frank Yates and William G. Cochran. Most others focus on one of these two frameworks. We do not claim it is a bad thing, however. Survey sampling and designed experiments differ so much and are used in so different situations for so different problems that probing into both of them poses a difficult task. But this does not mean we should avoid joining the knowledge of these two frameworks, at least from time to time.

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

As we have shown, such an approach might shed light on unresolved (or hidden) problems, just as looking at randomness in survey sampling helped us better understand the topic of randomness in designed experiments. From the above discussion, the following conclusions follow.

First of all, when designing an experiment, we should always define (i) what biological population and (ii) what population of environments we want to study. Without this, we will not be able to design an experiment which would help us respond the question we ask about the population (i).

Second, we must be aware that the only randomness associated with a designed experiment is that in *random* allocation of treatments to plots and in *random* allocation of experiments to locations.

Third, a single experiment (repeated in neither locations nor years) cannot be generalized to any population that might be of interest for a scientist. Even when treated as preliminary results, single experiments should be interpreted with great caution.

Fourth, a single field experiment refers to a narrowly defined population, and thus can be interesting only to offer practical advice about plant production in a confined region, if local farmers believe in the results (Mead *et al.*, 2012).

Fifth, only series of experiments (in multiple environments) can provide a solid basis for broadly based – both biologically and methodologically sound – inference, interpretation and conclusions. This, however, requires that the environments be *randomly* selected from a well-defined population of environments that well represent the physical and biological target population we aim to study.

REFERENCES

Bailey, R. A. (2009). *Design of Comparative Experiments*. Cambridge: Cambridge University Press.

Caliński, T., Czajka, S., Kaczmarek, Z., Krajewski, P. and Pilarczyk, W. (2009) Analyzing the genotype-by-environment interactions under a randomization-derived mixed model. *Journal of Agricultural, Biological, and Environmental Statistics* 14:224–241.

Caliński, T. and Kageyama, S. (2000). *Block Designs: A Randomization Approach. Volume I: Analysis*. Berlin: Springer. Lecture Notes in Statistics 150.

Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.

Cooper, M., Stucker, R. E., DeLacy, I. H. and Harch, B.D. (1997). Wheat breeding nurseries, target environments, and indirect selection for gain yield. *Crop Science* 37:1168–1176.

Comstock, R. E. (1977). Quantitative genetics and the design of breeding programs. In *Proceedings of the International Conference on Quantitative Genetics*, 705–718 (Eds E. Pollack, O. Kempthorne and T. B. Bailey, Jr.), 16–21 Aug. 1976. Ames, Iowa: Iowa State University Press.

Damesa, T., Worku, M., Möhring, J. and Piepho, H.P. (2017). One step at a time: Stage-wise analysis of a series of experiments. *Agronomy Journal* 109:845–857.

Efron, B. and Tibshirani, R. (1993). *Introduction to the Bootstrap*. London: Chapman & Hall.

Fox, P. N. and Rosielle, A. A. (1982). Reference sets of genotypes and selection for yield in unpredictable environments. *Crop Science* 22:1171–1175.

Gardner, M. J. and Altman, D. G. (1986). Confidence intervals rather than *p* values: Estimation rather than hypothesis testing. *British Medical Journal* 292:746–750.

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

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press

Hinkelmann, K. and Kempthorne, O. (1994). *Design and Analysis of Experiments. Volume I: Introduction to Experimental Design*. New York: Wiley.

Hu, X. (2015). A comprehensive comparison between ANOVA and BLUP to valuate location-specific genotype effects for rape cultivar trials with random locations. *Field Crops Research* 179:144–149.

John, J. A. and Williams, E. R. (1995). *Cyclic and Computer-Generated Designs*. London: Chapman & Hall.

Kish, L. (1967). *Survey Sampling*. New York: Wiley.

Kitsche, A. and Schaarschmidt, F. (2015). Analysis of statistical interactions in factorial experiments. *Journal of Agronomy and Crop Science* 201:69–79.

Kozak, M. (2008). Finite and infinite populations in biological statistics: Should we distinguish them? *Journal of American Science* 4:59–62.

Läärä, E. (2009). Statistics: Reasoning on uncertainty, and the insignificance of testing null. *Annales Zoologici Fennici* 46:138–157.

Mead, R., Gilmour, S. G. and Mead, A. (2012). *Statistical Principles for the Design of Experiments*. Cambridge: Cambridge University Press.

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* 97:558–625.

Piepho, H. P. and Möhring, J. (2005). Best linear unbiased prediction of cultivar effects for subdivided target regions. *Crop Science* 45:1151–1159.

Särndal, C. E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*. Kluwer Academic Publishers, The Netherlands: Springer Science & Business Media.

Singh, S. (2003a). *Advanced Sampling Theory with Applications: How Michael "Selected" Amy (Vol. 1)*. Kluwer Academic Publishers, The Netherlands: Springer Science & Business Media.

Singh, S. (2003b). *Advanced Sampling Theory with Applications: How Michael "Selected" Amy (Vol. 2)*. Kluwer Academic Publishers, The Netherlands: Springer Science & Business Media.

Speed, T. P. (1991). Comment to: Samuels ML, Casella G, McCabe GP 1991 interpreting blocks as random factors. *Journal of the American Statistical Association* 86:798–821.

Thompson, S. K. (2002). *Sampling*, 2nd ed. New York: Wiley.

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

Welham, S. J., Gezan, S. A., Clark, S. J. and Mead, A. (2015). *Statistical Methods in Biology. Design and Analysis of Experiments and Regression*. Boca Raton: CRC Press.

Yates, F. and Cochran, W. G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science* 28:556–580.

https://doi.org/10.1017/S001447971800025X Published online by Cambridge University Press