Statistical Inferences about Populations and Classes

Comment referring to the TFSI-article
'Statistical Methods in Psychology Journals: Guidelines and Explanations'

by
Albrecht Iseler
Freie Universität Berlin
E-Mail iseler@zedat.fu-berlin.de

Abstract: It is argued that statistics can be used legitimally not only for inferences about populations, but also about classes. However, the requirements upon the representation of a domain by a collection ('sample') of members are different under both kinds of inference. Therefore, the (partly explicit and partly implicit) restriction of inferential statistical to inferences about populations in the Sections 'Population' and 'Sample' of the guidelines and explanations should be replaced.

Before bringing forward a consideration concerning statistical inferences about populations *and* classes, I'ld like to express my admiration for the combination of wise moderateness and spirited implementation of recent methodological discussion, in particular concerning the role of null hypothesis significance testing in psychology. As a prototypical example of this approach, I consider the principal statement on hypothesis tests: 'It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p value or, better still, a confidence interval.' It implies a clear ranking, but even the least preferred option - the accept-reject decision - is given a chance to be suitable in special, presently unforeseen situations. I fully agree that this is the best summary of the results of the recent methodological discussion, which can be formulated for the addressees of the suggestion: Authors of psychological manuscripts, who are not methodologists themselves, but want a guideline how to keep their publications in line with the current methodological discussion.

Similarly, the distinction between inferences about populations and classes in the Section 'Population' raises due attention to a central (although less frequently discussed) methodological issue. However, the appropriateness of confining statistical inference to hypotheses about

populations and admitting only 'logical or other nonstatistical (!) methods' for inferences about a class may be questioned. Considering the respect for 'wise elders' testified by the report of the TFSI, I take allowance to recall and to generalise some results of a discussion in an almost venerable series of Psychological Bulletin articles by Sidman (1952), Bakan (1954) and Estes (1956).

Before discussing the issue in more detail, I'ld like to make sure that I'm not misunderstanding the intended distinction. I understand it in the context of two articles by Bakan (1955, 1967). The first one pointed out the difference between '*general-type propositions*', where something is asserted to be (presumably) 'true of each and every member of a designable class' and '*aggregate-type propositions*', whose assertion is presumed to be 'true of the class considered as an aggregate'. In Section 3 of the second article, Bakan criticised 'the confusion of induction to the aggregate with induction to the general'. I hope that I correctly understand the distinction between inferences about populations and about classes in Wilkinson et al. (1999), if I identify the first one with inferences about aggregate-type propositions and the second one with inferences about general-type propositions in the conceptual framework of Bakan.

I am aware, that many authors consider single case studies as the 'method of choice' for tests of hypotheses referring to each and every member of a class. (Bakan, 1955, p. 212: 'The "next" case presents a fundamental threat to the validity of a general-type proposition. ... The "next" case, for the aggregate-type proposition, simply increases the "power" of the test.') Indeed, I do not question that single case studies form the most appropriate test of a general-type proposition, if they can be performed properly, and I suppose that they are subsumed in Wilkinson et al. (1999) under 'logical or other nonstatistical methods'.[1] But there are at least two well known sources of problems with single case studies disabling a testable predictions: Measurement error and nonrepeatability of observations in a single case.

- In his article 'The Problem of Inference from Curves Based on Group Data', Estes (1956, p. 138) points out that individual curves can deviate from an ideal error free curve by 'experimental error', which may be assumed to average out in group curves. For instance, if a theory claims that each and every individual learning curve in a task under study is a strictly increasing, negatively

[1] Note, however, that even single case studies may need inferential statistics, if the claim referring to every member of the class (e.g., to every person) implies intraindividual aggregation (e.g., over situations).

accelerated, quadratic function of trial number up to an additive measurement error, then it is a testable prediction that average group curves will not deviate significantly from this shape.

- If a psychological hypothesis claims that a version *a* of some problem is easier to solve for each and every subsject (!) than version *b*, solution time serving as empirical indicator of difficulty, then the common practice of comparing average solution times of independent samples is forced not by the logical nature of the hypothesis (which is a general-type proposition), but by the simple fact that a problem solved under one of the two conditions isn't any longer a problem for the tested subject. So the only testable prediction is the difference in the average solution times of two groups of subjects.

In both examples, the 'statistical hypothesis' refers to an aggregate, but it is a prediction derived from a hypothesis referring to each and every member of a class, i.e., from a general-type proposition.

Now the following argument is brought forward by some proponents of single case studies as the only admissible method of testing hypotheses referring to every member of a class: If only aggregate hypotheses derived from a psychological hypothesis are available as testable predictions, then the reference of the hypothesis to each and every subject of a well defined class is a theoretical surplus without empirical content. So the hypothesis can be claimed with empirical content only as an aggregate hypothesis. But in my view, this argument neglects a fundamental difference in the requirements upon the sampling of cases, which has been pointed out implicitly by Bakan (1955)[2]: If a hypothesis is an aggregate hypothesis by its nature, then predictions can be made only for representative samples. On the other side, hypotheses about all members of a class may very well allow predictions for biased samples. Even more, daring predictions referring to deliberately biased samples may yield a stronger test of the hypothesis (in Popperian understanding) than those referring to the commonly used convenience samples. Since this idea has been published only fragmentarily in English (Iseler, 1996) and a more explicit discussion is presently available only in German (Iseler, 1997)[3], a demonstration by simple examples is enclosed as an appendix to this comment. The main conclusion is: For certain (but not all) hypotheses referring to each and every element of a class,

---

[2] Bakan (1955, p. 211): 'Whereas general-type propositions require a class of a plurality of members only for *validation*, aggregate-type propositions require a class of a plurality of members that they be sensible.

[3] See Iseler (1999) for an (admittedly much too long) draft of an English manuscript.

there is a universe of derivable aggregate hypotheses, and this 'universal aggregate hypothesis' is equivalent with the 'universal individual-related hypothesis'. But then a result of a comparison of two independent samples, which complies with some 'singular aggregate hypothesis' (i.e., with the hypothesis referring to particular selection distribution), is logically not more and not less (!) than a conforming result of one single case study, if the latter one can be performed properly at all.

Certainly, deliberations of this kind cannot be considered to form a part of the commonly agreed 'methodological canon', whose violation could be a sufficient reason to reject a manuscript submitted for publication, and I am aware that the formulation of such standards is one of the main objectives of the guidelines. However, there are two reasons to bring forward these considerations in the context of the guidelines. First (and most important in my view) is the fact that such guidelines shouldn't generally 'forbid' methods (like biased sampling), which are rational in the study of psychological hypotheses referring to all members of a class, even if their rationality in these situations isn't part of the 'undergraduate level' methodology. But in my view, the sections on 'Population' and 'Sample' in Wilkinson et al. (1999) seem to be almost tailor-made for an understanding, where only aggregate-type propositions are legitimally tested by statistical hypotheses in psychology, whereas 'we make inferences about a class through logical and other nonstatistical methods' (quoted from the explanation under 'Population').

A second reason to plead for the differentiation is a more or less didactical one. Of course, the widespread negligence and even thoughtlessness with sampling procedures is a deplorable fact. But couldn't this very fact reflect intuitive (and legitimate) doubts about the adequateness of the *general* and undifferentiated requirement of representative sampling, which is typically taught in undergraduate courses and textbooks on psychological methods? Formulated a bit colloquially: 'Somehow we all know that these rules are too strict, so why should just me follow them?' I see better chances of acceptance for a more differentiated guideline. In the above terminology, the message of a guideline with a greater chance of acceptance could be: Non-representative ('biased') sampling is almost deadly for the evidence resulting from data, if the claim of the hypothesis under study is an aggregate-type proposition. On the other side, if the psychological research hypothesis is a general-type proposition, then the statistical prediction tested by inferential statistics may follow deductively from the research hypothesis, if the sample is biased. But even then the widespread usage

of arbitrarily biased convenience samples is at least suboptimal and preferably replaced by deliberately biased sampling, i.e., by a sampling procedure with increased selection probabilities for subjects, who are most likely (by theoretical or empirical reasons) to be violators of the hypothesis, if such violators exist at all.

Now I am aware that a guideline requiring the above explicated background information would be misplaced, in particular, if a publication suitable for reference is still outstanding. Considering the remarkable ability of the TFSI-authors to communicate in a few simple sentences the sediment of complex methodological discussions, I hesitate to try it myself. In particular, I feel hindered by the obvious limitations of my English (for which I apologise).

Nevertheless, I dare to propose some formulations, which will need improvements by the TFSI-experts (certainly beyond the British orthography, which is more familiar to myself), even if the main idea seems worth mentioning in the guidelines. The proposal is to reformulate the Sections 'Population' and 'Sample' as follows:

Domains of Hypotheses

The interpretation of the results of any empirical study depends on the characteristics of the domain (of persons, persons in situations, stimuli, studies etc.) intended for analysis. Define this domain clearly, and specify whether your hypothesis refers to this domain as a 'population' or as a 'class'. If control or comparison groups are part of the design, present how they are defined.

Psychology students sometimes think that a statistical population is the human race or, at least, college sophomores. They also have some difficulty distinguishing a class of objects (where the claim of a hypothesis refers to each and every element of the class) versus a statistical population (i.e., an aggregate, where the hypotheses refer only to the aggregate as a whole, notwithstanding the existence of subpopulations with nonconforming distributions, e.g. with opposite effects of treatments). How a domain of persons, stimuli or even research articles is defined in an article and whether it is conceived as a population or as a class affects almost every conclusion in that article.

Sample

Describe the sampling procedures and discuss their adequateness for the intended interpretation of the referenced domain of subjects as a population or as a class. Emphasise any inclusion or exclusion criteria. If the sample is stratified (e.g., by site or gender) describe fully the rationale.

Note the proposed sample size for each subgroup.

It is frequently neglected that the requirements upon representativeness of sampling depend heavily on the interpretation of a domain as a population or a class. Non-representative ('biased') sampling is almost deadly for the evidence resulting from data, if the hypothesis under study refers to a domain of persons as a population. But the statistical prediction tested by inferential statistics may follow deductively from a research hypothesis referring to a class, even if the sample is biased. - Interval estimates ... (same text as in Wilkinson et al., 1999)... by implying that you used a random sample. Convenience samples are legitimate for the test of certain hypotheses referring to all members of a class. Furthermore, if representativeness is needed for the test of hypotheses referring to a population, the case for a convenience sample can sometimes be strengthened ...

References

Bakan, D. (1952). A generalisation of Sidman's results on group and individual functions, and a criterion. *Psychological Bulletin, 51,* 63-64.

Bakan, D. (1955). The general and the aggregate: A methodological distinction. *Perceptual and Motor Skills, 5,* 211-212.

Bakan, D. (1967). The test of significance in psychological research. Reprinted in: Morrison, D.E. & Henkel, R. E. (Eds.). (1970). *The Significance Test Controversy*. Chicago: Aldine.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134-140.

Iseler, A. (1996). A paradoxical property of hypotheses referring to the order of medians. *Methods of Psychological Reserach - Online, 1,* 4. URL: http://www.mpr-online.de.

Iseler, A. (1997). Populationsverteilungen von Merkmalen und Geltungsbereiche individuen-bezogener Aussagen als Gegenstand der Inferenzstatistik in psychologischen Untersuchungen. (Population distributions of variables and domains of individual-related propositions as objectives of inferential statistics in psychological studies.) In: In: H. Mandl (Ed.): *Bericht über den 40. Kongreß der Deutschen Gesellschaft für Psychologie in München 1996,* p. 699-708. Göttingen: Hogrefe. Available in the internet under http://userpage.fu-berlin.de/~iseler/papers/popdom.pdf

Iseler, A. (1999). Draft on Stability under Aggregation. Available in the internet under

http://userpage.fu-berlin.de/~iseler/papers/aggrstab.pdf

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.

Sidman, M. (1954). A note on functional relations obtained from group data. *Psychological Bulletin*, *49*, 263-269.

Wilkinson, L. & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Author's address :

Albrecht Iseler
Freie Universität Berlin
Habelschwerdter Allee 45
D-14195 Berlin
Germany
E-Mail: iseler@zedat.fu-berlin.de

Appendix:

Examples for the Test of Hypotheses about Classes by Deliberately Biased Samples.

Consider a psychological hypothesis claiming that a version *a* of some problem is easier to solve for each and every member of a class *D* of persons[4] than version *b*, solution time serving as indicator of difficulty. To formalise this claim, let τ_{ua} and τ_{ub} be the true scores of solution time for unit (individual, class member) *u* under condition *a* resp. *b*. Then the claim of the hypothesis may be formalised as

$$\forall\, u \in D\colon \tau_{ua} < \tau_{ub}.$$

More precisely, the true scores τ_{ua} and τ_{ub} are expectations of a dependent variable *Y* under probability distributions representing the behavioural dispositions of unit *u* under conditions *a* and *b*. Subsequently, these probability distributions will be represented by cumulative probabilities $P_{ua}(Y \le t)$ resp. $P_{ub}(Y \le t)$, i.e., probabilities that the value of the dependent variable *Y* (solution time) for unit *u* under condition *a* resp. *b* is not greater than the (positive) real number *t*.

Now consider the experiment of randomly selecting one element of the class and observing it under one of the conditions, understanding random selection such that it doesn't necessarily imply identical selection probabilities of all units. Rather, there is a selection probability $\pi_u \ge 0$ for every member *u* of the domain *D*, with $\sum_{u\in D} \pi_u = 1$. Taken together, these selection probabilities form the 'selection distribution' π. Now let $P_{\pi a}(Y \le t)$ resp. $P_{\pi b}(Y \le t)$ be the cumulative probabilities of obtaining a value up to (and including) *t* for the dependent variable *Y* in the experiment of selecting a unit with selection distribution π and testing it under condition *a* resp. *b*, and assume that this experiment is performed such that the equations

$$P_{\pi a}(Y \le t) = \sum_{u\in D} \pi_u \cdot P_{ua}(Y \le t)$$

---

[4] For the present demonstration, a domain of persons has an obvious advantage over other domains: Since the number of persons available for selection will always be finite, expectations can be conceived as sums. For generalisations to uncountable domains (e.g., of persons in situations), such expectations - i.e., sums of the form $\sum_{u\in D} \pi_u \cdot \ldots$ - have to be replaced by Lebesgue integrals of the form $\int \ldots d\pi$, where π is a probability measure on a suitable σ-algebra in the domain *D*. If (almost trivial) premissas about the existence of these expectations are added, the same results can be obtained.

and

$$P_{\pi b}(Y \leq t) = \sum_{u \in D} \pi_u \cdot P_{ub}(Y \leq t)$$

hold for every positive *t*. With the denotations $\mu_{\pi a}$ and $\mu_{\pi b}$ for the expectations of random variables with these probabilities, it is easily verified that this assumption implies

$$\mu_{\pi a} = \sum_{u \in D} \pi_u \cdot \tau_{ua}$$

and

$$\mu_{\pi b} = \sum_{u \in D} \pi_u \cdot \tau_{ub},$$

i.e.,

$$\mu_{\pi b} - \mu_{\pi a} = \sum_{u \in D} \pi_u \cdot (\tau_{ub} - \tau_{ua}) > 0,$$

the final inequality following from the assumption $\tau_{ua} < \tau_{ub}$ for every momber *u* of the domain *D*. Conversely, if there exist units with $\tau_{ua} \geq \tau_{ub}$, and π is any selection distribution assigning a non-zero selection probability only to such units, we obtain

$$\mu_{\pi b} - \mu_{\pi a} \leq 0.$$

With the denotation Π for the set of all conceivable selection distributions in the domain *D*,[5] the results can be summarised by the equivalence

$$\forall\, u \in D\colon \tau_{ua} < \tau_{ub} \;\Leftrightarrow\; \forall\, \pi \in \Pi\colon \mu_{\pi a} < \mu_{\pi b}.$$

With the denotations 'universal unit-related hypothesis' resp. 'universal aggregate hypothesis'[6] for the two sides of this equivalence, we can state that both universal hypotheses are equivalent.

Now a classical single case study is the test of a singular unit-related hypothesis referring to a

---

[5] In the language of probability spaces, Π can be defined to be the set of all probability measures on $\mathsf{P}D$ - the power set of the domain *D*.

[6] It should suffice to state once and for all that the term 'universal' has to be understood as 'universal within *D*' resp. 'universal within Π'.

specific unit. In the same way, the comparison of two randomised (!) independent samples is the test of a singular aggregate hypothesis (i.e., an aggregate hypothesis referring to the specific selection distribution π underlying both samples), this singular aggregate hypothesis following from the universal aggregate hypothesis in the same way as the singular unit-related hypothesis follows from the universal unit-related hypothesis. In other words, since the two universal hypotheses are equivalent, a result of a comparison of two independent samples, which complies with the singular aggregate hypothesis $\mu_{\pi a} < \mu_{\pi b}$, is logically not more and not less (!) than a conforming result of one single case study, if the latter one can be performed properly at all.

The parallell goes further. Many authors, particularly Meehl (1978), have deplored the lack of 'daring' predictions in psychological hypothesis testing. Although the prediction referring to the 'next' case resp. to a convenience sample is a legitimate test of a universal hypothesis (unit- resp. aggregate-related), more daring predictions are possible. Theoretical considerations or empirical informations (e.g. from ATI-studies with multivariate predictors) may suggest characteristics ('trait configurations') of persons, who are most likely to violate the hypothesis $\tau_{ua} < \tau_{ub}$, if such violators exist at all. Now it is a daring prediction that even the result of a single case study using a subject with these characteristics will comply with the hypothesis. Similarly, we can make the following prediction: If we recruit a large sample of subjects (maybe a convenience sample) and select (e.g., by measurement values in predictors derived from an ATI-study) for the main experiment those who are most likely to violate the hypothesis $\tau_{ua} < \tau_{ub}$, and assign them randomly to the two treatments, the comparison of the two independent samples will comply with the inequality $\mu_{\pi a} < \mu_{\pi b}$, where π is the selection distribution resulting from the general recruitment and from the confinement of the main experiment to subjects with certain values in the predictors. This kind of hypothesis testing is referenced by the concept of 'deliberately biased sampling'. In my view, it is an essentially more daring prediction derived from a general-type proposition than the commonly applied predictions for convenience samples.

The above demonstration refers only to the very simple case of hypotheses about the order of expectations. Similar predictions are possible for other hypotheses. Consider, e.g., the hypothesis of strictly increasing, negatively accelarated and quadratic learning curves as 'ideal error free individual curves'. This hypothesis would be violated systematically by persons whose error-corrected ('ideal')

learning curves start with a steep ascent and approach an asymptote well before the last trial. So 'likely violators' are 'fast learners', if the performance has a natural asymptote, or persons with a low individual asymptote, and predictors may exist for both types of violators. But if these subjects form a minority (in a domain conceived as a population), the departure of the individual curves of a few subjects in a sample may be undistinguishable from measurement error. However, the following daring prediction can be derived from the hypothesis, conceived as a general-type proposition: Even if a group is made up of subjects with the typical predictor values of fast learners or of persons with a low asymptote, the average group curve will not deviate significantly from the hypothesised shape of ideal individual curves.

However, it should be noted that this kind of prediction doesn't hold for all hypotheses referring to every member of a class *D*. Problems arise not only with hypotheses claiming a specific mathematical shape of a curve. (See the series of articles by Sidman, 1952, Bakan, 1954, and Estes, 1956, for examples of this kind.) For instance, if the above hypothesis referring to solution times is expressed in terms of the order of medians instead of expectations ('true scores'), there may occur a 'median paradox' (Iseler, 1996): Even if the median of the individual probability distributions (made up by the above introduced probabilities $P_{ua}(Y \leq t)$ and $P_{ub}(Y \leq t)$) is smaller under condition *a* than under condition *b* for *every* individual *u*, the reversed order may hold for the aggregate. Worse still, since there are typical individual pairs of probability distributions P_{ua} and P_{ub} producing this kind of paradox, the method of deliberately biased sampling may result in the selection of subjects with such distributions and produce a non-hypothesiscomforming order of medians in the distributions $P_{\pi a}$ and $P_{\pi b}$, although the hypothesised order holds for each individual.

I'm preparing a journal article and a monograph about such issues, a (much too long) draft version being partly available in the internet (Iseler, 1999). Basically, it is an extension of the results of Estes (1956), obtained by an imbedding of the problem into the mathematical theory of function spaces.