Statistical Science
2017, Vol. 32, No. 3, 352–355
DOI: 10.1214/16-STS600
Main article DOI: 10.1214/16-STS571

Inference from Randomized (Factorial) Experiments

R. A. Bailey

Abstract. This is a contribution to the discussion of the interesting paper by Ding [*Statist. Sci.* **32** (2017) 331–345], which contrasts approaches attributed to Neyman and Fisher. I believe that Fisher's usual assumption was unit-treatment additivity, rather than the "sharp null hypothesis" attributed to him. Fisher also developed the notion of interaction in factorial experiments. His explanation leads directly to the concept of marginality, which is essential for the interpretation of data from any factorial experiment.

Key words and phrases: Factorial design, marginality, randomisation, unit-treatment additivity.

1. WHAT SHOULD FISHER'S NAME BE ATTACHED TO?

This interesting paper compares so-called "Neymanian inference" (Section 2.2) with the so-called "Fisherian randomisation test" (Section 2.3). These are linked to two different null hypotheses, which have Neyman's and Fisher's names attached to them. I believe that it is inappropriate to use Fisher's name for the second null hypothesis, and that it is over-stating the case to attach his name to the randomisation test.

Fisher's ground-breaking work on the design of experiments, and the ensuing analysis of data, was laid out in Fisher (1925, 1926, 1935a). Throughout these, it is notable that he did not use any notation for the response on unit i under treatment t, which is written as $Y_i(t)$ by Ding (2017); nor did he write down any equation expressing this quantity in terms of any others. It is therefore problematic to identify his name with the so-called "sharp null hypothesis" described in Section 2.3.

It is true that the discussions by Fisher (1935a) of Darwin's data and of the tea-tasting experiment appear to be consistent with the methodology presented in Section 2.3. However, the seventh edition of this book (Fisher, 1960) contains a new subsection called "Section 21.1 'Non-parametric' Tests", in which Fisher refuted the idea of randomisation tests. In this new subsection, he said, "The reader will realise that [the randomisation test] was in no sense put forward to supersede the common and expeditious tests based on the Gaussian theory of errors. ... [It] has been an attraction to some mathematicians who often discuss experimentation without personal knowledge of the material. ... Experimenters should remember that they and their colleagues usually know more about the kind of material they are dealing with than do the authors of text-books written without such personal experience".

Yates was Fisher's colleague at Rothamsted Experimental Station from 1931 to 1933, and they collaborated for a further 30 years. In Yates (1964, 1965), he explained Fisher's thinking about the randomisation test. He attributed part of the above quotation to Fisher (1960) in Yates (1964) but to the second edition Fisher (1937) in Yates (1965). He also said that "Fisher, I think, tended to lay undue emphasis on the importance of formal tests of significance in experimental work" (Yates, 1965), and "... in many types of experimental work estimates of the treatment effects, together with estimates of the errors to which they are subject, are the quantities of primary interest" (Yates, 1964). Fisher (1935b) complained that Neyman was confusing estimation with tests of significance.

R. A. Bailey is Professor of Mathematics and Statistics, School of Mathematics and Statistics, University of St. Andrews, St. Andrews, Fife KY16 9SS, United Kingdom, and Professor emerita of Statistics, School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom (e-mail: rab24@st-andrews.ac.uk).

The paper by Ding (2017) pays no attention to estimation.

2. A FIXED FINITE POPULATION OR AN ASSUMPTION OF ADDITIVITY?

Ding (2017) begins by stating that the potential-outcomes framework has been widely used in randomised experiments ever since 1923. I have practised as a statistician for 40 years, helping to design experiments and analyse the data therefrom, in subjects such as agriculture, ecology, pre-clinical trials and human-computer interaction. I have never used the potential-outcomes framework.

In those subjects, the experimental unit might be a plot in a field or a pot in a greenhouse; a microcosm in a temperature-controlled laboratory; a dog or a monkey, or a limb or tissue of such a animal; a university student for a single afternoon. In no case were the experimental units a random sample from a fixed finite population. They were convenient, and were deemed to be representative enough that results on them could be extrapolated to other units, such as real farmers' fields. Such extrapolation is impossible unless we can assume that

$$Y_i(t) = \tau_t + Z_i, \tag{1}$$

where τ_t depends only on the treatment t (and is usually assumed to be constant) and Z_i depends only on the experimental unit i (and is usually assumed to be a random variable). Of course, $Y_i(t)$ needs to measured on a scale which makes such additivity plausible.

Equation (1) is given in Bailey (2008), Section 1.5. I explained it in more detail in Bailey (1981). However, I certainly did not invent it. It is given in Nelder (1965), Section 3. Cox [(1958b), Chapter 2], sets out the argument with exemplary clarity.

I do not think that Yates (1964, 1965) was rewriting history. Statisticians who learnt their subject at Rothamsted Experimental Station, or from Fisher and his academic descendants at the Universities of London and Cambridge, or at C.S.I.R.O. in Adelaide, where Fisher spent his final years (1959–1962), all use the additive model (1). This is a weaker assumption than the linear model given in Section 7.2 of Ding (2017).

How is this linked to randomisation? It is argued by Bailey (1981, 1991) and Bailey and Brien (2016) that assumption (1): combined with a suitable method of randomisation, allows us to assume a fairly straightforward joint distribution for $(Z_1, \ldots, Z_N)$. This is implicit, in different words, in various early papers of Yates, starting with Yates (1933). In Yates (1935a), he stated explicitly that "the randomisation process effectively generates the distribution of z".

The famous disagreement between Fisher and Neyman began with the paper Neyman, Iwaskiewicz and Kołodziecyk (1935) read to the Royal Statistical Society, which claims to include a proof that any experiment designed as a Latin square gives biased results. Wilk and Kempthorne (1957) developed this argument further. In 1957, Kempthorne chaired a six-week I.M.S. Summer Institute on the topic at Boulder, Colorado. Cox attended this; as a result, he published Cox (1958a) arguing that the wrong question had been addressed and that Fisher had been correct to state that there is no bias in a Latin-square experiment. The acknowledgement at the end of Cox (1958a) makes it clear that Cox and Kempthorne had had positive, stimulating discussions on the topic. In later years, Kempthorne also used assumption (1), for example, see Kempthorne (1975b).

That workshop was nearly 60 years ago. Why are some authors still claiming that Fisher did not assume unit-treatment additivity?

3. A LITTLE MORE HISTORY

Further insight into the differences between Fisher's and Neyman's approaches can be found in the books by Fisher Box (1978) and Reid (1982), as well as the more recent paper by Senn (2004). All of these are well worth reading.

As for the paradox described in Section 3.2 of Ding (2017), George Barnard discovered something similar and explained it in a 1955 letter to Fisher, given in Bennett (1990), pages 29–30. He said, "We therefore seem to be in a situation where we can believe A, but not B, although B is a logical consequence of A. ... such a paradox as this one inevitably arises whenever we have a test of a wider hypothesis (B), and of a narrower hypothesis (A)".

4. FACTORIAL TREATMENT STRUCTURE

Fisher [(1935a), Section 38], did indeed introduce factorial treatment structure. There he explained that if F_1 and F_2 are two two-level treatment factors then there is an interaction between F_1 and F_2 if the effect of F_1 depends on the level of F_2. Putting this in different words, there is a nonzero interaction between F_1 and F_2 if the effects of F_1 and F_2 are not additive. In the notation used in Ding [(2017), Section 4.2], this means that $\bar{\boldsymbol{Y}}$ is not in the three-dimensional vector

space $V_1 + V_2$ spanned by $\boldsymbol{g}_1$, $\boldsymbol{g}_2$ and the all-1 vector $\mathbf{1}$. But it does not make sense to consider the vector space spanned by $\boldsymbol{g}_{12}$ and $\mathbf{1}$: as Yates (1935b) said, "If interaction exists, then usually information will be required on the responses to each factor in the presence and absence of the other".

More generally, let $\mathcal{S}$ be any subset of $\{1, 2, \ldots, K\}$, where K is the number of treatment factors. If the jth factor has n_j levels, then there are $d_\mathcal{S}$ combinations of levels of factors F_j for j in $\mathcal{S}$, where $d_\mathcal{S} = \prod_{j\in\mathcal{S}} n_j$. Let $V_\mathcal{S}$ be the subspace consisting of vectors which are constant on each of these combinations, so that $\dim(V_\mathcal{S}) = d_\mathcal{S}$. If $\mathcal{R} \subset \mathcal{S}$, then $V_\mathcal{R} < V_\mathcal{S}$.

Further, define the subspace $V_\mathcal{S}^-$ by

$$V_\mathcal{S}^- = \bigoplus_{\mathcal{R}\subsetneq\mathcal{S}} V_\mathcal{R}.$$

For example, if $n_1 = n_2 = n_3 = 2$ then V_{123} is spanned by the vectors $\mathbf{1}$, $\boldsymbol{g}_1$, $\boldsymbol{g}_2$, $\boldsymbol{g}_3$, $\boldsymbol{g}_{12}$, $\boldsymbol{g}_{13}$, $\boldsymbol{g}_{23}$ and $\boldsymbol{g}_{123}$, while $V_{123}^- = V_{12} + V_{13} + V_{23}$, which does not contain $\boldsymbol{g}_{123}$.

Suppose that $\bar{\boldsymbol{Y}}$ is in $V_\mathcal{S}$. Then to say that $\bar{\boldsymbol{Y}}$ exhibits $\mathcal{S}$-interaction means that $\bar{\boldsymbol{Y}}$ is not contained in the subspace $V_\mathcal{S}^-$, or, rather, $\bar{\boldsymbol{Y}}$ is not sufficiently close to this subspace.

Under the assumption (1), this definition of interaction should be applied to the vector $\boldsymbol{\tau}$. Thus the existence of a nonzero $\mathcal{S}$-interaction means that the assumption that $\boldsymbol{\tau} \in V_\mathcal{S}$ cannot be simplified to the assumption that $\boldsymbol{\tau} \in V_\mathcal{S}^-$, and so no testing is done on any proper subset $\mathcal{R}$ of $\mathcal{S}$. Thus it is not true that "analogous discussion also holds for general factorial effects due to symmetry". The vectors $\boldsymbol{g}_1$ and $\boldsymbol{g}_{12}$ do indeed have the same mathematical properties, but the concepts of "main effect" and "two-factor interaction" are not so easily interchangeable.

In Bailey (2015), I define the F_1F_2 interaction to be the difference between the orthogonal projection of $\boldsymbol{\tau}$ onto V_{12} and its projection onto $V_1 + V_2$. Figure 5.11 of Bailey (2008) shows all the subspaces that need to be considered when $K = 3$. These should make it clear that there is no symmetry between main effects and interactions.

In fact, this observation is not new. Nelder (1977) called the $\mathcal{R}$-effect *marginal* to the $\mathcal{S}$-effect if $\mathcal{R} \subsetneq \mathcal{S}$, and lamented that some statisticians neglect marginality. He made the point again in Nelder (1994, 1998). As Kempthorne (1975a) said bluntly: "The testing of main effects in the presence of interaction, without additional input, is an exercise in fatuity." It is unfortunate that some recent papers about factorial designs, including Dasgupta, Pillai and Rubin (2015), Ding (2017), Pauly, Brunner and Konietschke (2015), ignore marginality completely.

5. ONE PARTICULAR FACTORIAL EXPERIMENT

The factorial experiment described in Ding [(2017), Section 6.3], could serve as a textbook example of what not to do in the design and conduct of an experiment and the processing of its data. In the first place, we are told that the purpose of the experiment is "to identify the optimal combination of (levels of) the four factors". This is a perfectly valid purpose, but it has nothing to do with hypothesis testing; it is closer to estimation. Moreover, the factorial structure of the treatments is irrelevant to this purpose; the aim is simply to select the best one of 16 treatments.

Although the paper by Ding (2017) is about randomisation, nothing is said about how the 16 treatments were randomised to the 32 experimental units in this experiment. The experimental units can be identified by their sequence in time, and the data should be presented in this order, to aid identification of time trends or sudden changes. Moreover, although three different students were involved, we are not told their roles. If some helicopters were made or thrown by one student, and other helicopters by another, then the students give a blocking factor: this information should be taken into account during the randomisation and presented with the data.

Any practising statistician will cast her or his eyes over the data before embarking on formal analysis. This enables obvious mistakes or anomalies to be spotted. With these data, the first glaring omission is the lack of measurement units. Were the flight times measured in minutes and seconds? Have they been converted to the decimal format shown? If so, there is a danger of spurious accuracy.

The second obvious feature of the data is that the two measurements for each treatment are, in most cases, surprisingly close. Did the students really make 32 helicopters in a completely randomised design? Did they make the second helicopter for each treatment immediately after the first? Did they actually make only 16 helicopters and fly each one twice? If they used the second or third procedure then this experiment has *pseudo-replication* and so it is impossible to test for any treatment effects. Pseudo-replication is one of the oldest, and most common, faults in designed experiments; see Hurlbert (1984, 2009), Sparks, Bailey and Elston (1997).

Even without the foregoing problems, if we want to use these data to investigate which factors, and their

combinations, affect flight times, then we have to respect marginality. We are told that the interactions F_1F_3 and $F_1F_2F_4$ are both nonzero. This means that the vector $\boldsymbol{\tau}$ cannot be assumed to belong to a proper subspace of $V_{13} + V_{124}$. There is therefore no point in testing for the interactions F_1F_2, F_1F_4 and F_2F_4, or for any of the main effects.

REFERENCES

BAILEY, R. A. (1981). A unified approach to design of experiments. *J. Roy. Statist. Soc. Ser. A* **144** 214–223. MR0625801

BAILEY, R. A. (1991). Strata for randomised experiments. *J. Roy. Statist. Soc. Ser. B* **53** 27–78. With discussion and a reply by the author. MR1094275

BAILEY, R. A. (2008). *Design of Comparative Experiments*. *Cambridge Series in Statistical and Probabilistic Mathematics* **25**. Cambridge Univ. Press, Cambridge. MR2422352

BAILEY, R. A. (2015). Structures defined by factors. In *Handbook of Design and Analysis of Experiments* (A. Dean, M. Morris, J. Stufken and D. Bingham, eds.) 371–414. Chapman & Hall/CRC Press, Boca Raton.

BAILEY, R. A. and BRIEN, C. J. (2016). Randomization-based models for multitiered experiments: I. A chain of randomizations. *Ann. Statist.* **44** 1131–1164. MR3485956

BENNETT, J. H. (ed.) (1990). *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. The Clarendon Press, Oxford. With a preface by J. H. Bennett. MR1076366

COX, D. R. (1958a). The interpretation of the effects of non-additivity in the Latin square. *Biometrika* **45** 69–73.

COX, D. R. (1958b). *Planning of Experiments*. Wiley, New York; Chapman & Hall, London. MR0095561

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753. MR3382595

DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345.

FISHER, R. A. (1925). *Statistical Methods for Research Workers*, 1st ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **33** 503–513.

FISHER, R. A. (1935a). *The Design of Experiments*, 1st ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1935b). Contribution to discussion of "Statistical problems in agricultural experimentation" by J. Neyman, with the help of K. Iwaskiewicz and St. Kołodziecyk. *J. Roy. Statist. Soc. Suppl.* **2** 154–157.

FISHER, R. A. (1937). *The Design of Experiments*, 2nd ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1960). *The Design of Experiments*, 7th ed. Oliver and Boyd, Edinburgh.

FISHER BOX, J. (1978). *R. A. Fisher: The Life of a Scientist*. Wiley, New York. MR0500579

HURLBERT, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54** 187–211.

HURLBERT, S. H. (2009). The ancient black art and transdisciplinary extent of pseudoreplication. *Journal of Comparative Psychology* **123** 436–443.

KEMPTHORNE, O. (1975a). Fixed and mixed models in the analysis of variance. *Biometrics* **31** 473–486. MR0373183

KEMPTHORNE, O. (1975b). Inference from experiments and randomisation. In *A Survey of Statistical Design and Linear Models* (*Proc. Internat. Sympos., Colorado State Univ., Ft. Collins, Colo.*, 1973) (J. N. Srivastava, ed.) 303–331. North-Holland, Amsterdam. MR0375664

NELDER, J. A. (1965). The analysis of randomised experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proc. Roy. Soc. Ser. A* **283** 163–178. MR0174156

NELDER, J. A. (1977). A reformulation of linear models. *J. Roy. Statist. Soc. Ser. A* **140** 48–76. With discussion. MR0458743

NELDER, J. A. (1994). The statistics of linear models: Back to basics. *Stat. Comput.* **4** 221–234.

NELDER, J. A. (1998). The great mixed-model muddle is alive and flourishing, alas! *Food Qual. Prefer.* **9** 157–159.

NEYMAN, J. (1935). Statistical problems in agricultural experimentation. *J. Roy. Statist. Soc. Suppl.* **2** 107–154. With the help of K. Iwaskiewicz and St. Kołodziecyk.

PAULY, M., BRUNNER, E. and KONIETSCHKE, F. (2015). Asymptotic permutation tests in general factorial designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 461–473. MR3310535

REID, C. (1982). *Neyman—From Life*. Springer, New York. MR0680939

SENN, S. (2004). Added values. Controversies concerning randomisation and additivity in clinical trials. *Stat. Med.* **23** 3729–3753.

SPARKS, T. H., BAILEY, R. A. and ELSTON, D. A. (1997). Pseudoreplication: Common (mal)practice. *SETAC* (*Society of Environmental Toxicology and Chemistry*) *News* **17**(3) 12–13.

WILK, M. B. and KEMPTHORNE, O. (1957). Non-additivities in a Latin square design. *J. Amer. Statist. Assoc.* **52** 218–236. MR0088137

YATES, F. (1933). The formation of Latin squares for use in field experiments. *Empire Journal of Experimental Agriculture* **1** 235–244.

YATES, F. (1935a). Contribution to discussion of "Statistical problems in agricultural experimentation" by J. Neyman, with the help of K. Iwaskiewicz and St. Kołodziecyk. *J. Roy. Statist. Soc. Suppl.* **2** 161–166.

YATES, F. (1935b). Complex experiments. *J. Roy. Statist. Soc. Suppl.* **2** 181–223.

YATES, F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics* **20** 307–321.

YATES, F. (1965). A fresh look at the basic principles of the design and analysis of experiments. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.* 4 777–790. Univ. California Press, Berkeley, CA.