ELSEVIER

Mathematical Biosciences 180 (2002) 23–28

Mathematical Biosciences

www.elsevier.com/locate/mbs

Fisher's randomization test and Darwin's data – A footnote to the history of statistics

John A. Jacquez [a], Geoffrey M. Jacquez [b,*]

[a] *Departments of Biostatistics and Physiology, The University of Michigan, Ann Arbor, MI 48109, USA*
[b] *BioMedware, 516 North State Street, Ann Arbor, MI 48104, USA*

Received 24 August 2001; received in revised form 9 April 2002; accepted 25 June 2002

Abstract

In the presentation of his randomization test for paired data, Fisher used Darwin's data on the relative growth rates of cross- and self-fertilized corn to motivate the development. On reading Darwin's description of his experiment, it appears clear that the experiment did not use true paired comparisons. Although the statistical foundation of Fisher's randomization test is sound, it is of historical interest that it does not suit the design of the motivating experiment.

Keywords: Darwin; Fisher; Randomization tests

1. Introduction

In chapter III of his book, Fisher [1] [1] introduced what is now called Fisher's randomization test in his analysis of Darwin's data [2] on the difference between cross- and self-fertilized plants of *Zea mays*. That was the first non-parametric test developed and is widely used; indeed in their analysis of paired experimental designs, Kempthorne and Doerfler [3] showed that in general it is better than the Wilcoxon rank test and the sign test.

* Corresponding author. Tel.: +1-734 913 1098; fax: +1-734 913 2201.
E-mail address: jacquez@biomedware.com (G.M. Jacquez).

[1] There was a minor error in Fisher's calculations. Fisher corrected that in the second edition and made a few small changes in wording. After that, chapter III remained unchanged until the seventh edition when Fisher added a short final section on non-parametric tests. We have examined the first edition (1935), the third (1942), the fifth (1949), the seventh (1960) and the eighth (1966, posthumous) editions.

0025-5564/02/$ - see front matter
PII: S0025-5564(02)00123-2

On reading Darwin's description of his experiment, it appeared to us that Darwin's experiment did not satisfy the requirements for a paired design. If so, we have an example of a very useful and widely used test which was developed for the analysis of data for which the test was inappropriate. That of course does not in any way invalidate the test but it adds an interesting footnote to the history of the development of statistical methods.

In what follows, we present Darwin's experiment and the data on *Zea mays*, Fisher's analysis and development of his randomization test and then present alternative randomization tests that Fisher might have used for the analysis of the data.

2. Darwin's data

Darwin raised some cross-fertilized plants and some self-fertilized plants; he planted equal numbers of each in four different pots but not the same numbers of each in every one of the pots. Darwin's description of the experiment is not very detailed. In his own words [p. 234],

Some plants were raised in the greenhouse, and were crossed with pollen taken from a distinct plant; and a single plant, growing quite separately in a different part of the house, was allowed to fertilise itself spontaneously. The seeds thus obtained were placed on damp sand, and as they germinated in pairs of equal age were planted on the opposite sides of four very large pots; nevertheless they were considerably crowded. The pots were kept in the hothouse. The plants were first measured to the tips of their leaves when only between 1 and 2 feet in height, as shown in the following table:-

There follows his Table XCVII. Table 1 gives Darwin's results converted to eighths of an inch plus an extra column giving the differences by line of entry.

Table 1
Darwin's results for the heights of cross- and self-fertilized plants of *Zea mays*, reported in 1/8th's of an inch

Pot	Crossed	Selfed	Difference
I	188	139	49
	96	163	−67
	168	160	8
II	176	160	16
	153	147	6
	172	149	23
III	177	149	28
	163	122	41
	146	132	14
	173	144	29
	186	130	56
IV	168	144	24
	177	102	75
	184	124	60
	96	144	−48

Darwin's experiment can be criticized on a number of grounds. First he does not tell us anything about the parentage of the plants that were cross- and self-fertilized. Furthermore, more than one plant was used to obtain cross-fertilized seed but only one was used for the self-fertilized seed. In addition, the cross-fertilized plants were fertilized by hand but the self-fertilized plants were allowed to self-fertilize naturally. Furthermore, there is no evidence for randomization in assignment of members of a pair to position in a pot. None of that accords with what we would nowadays require in a balanced design with randomization. That aside, what of the pairing? There is no true pairing of like seeds that received different treatments. The cross- and self-fertilized seeds were paired as they germinated so as to give pairs that were at about the same stage of germination. The fact that they were put in different pots that have different numbers of pairs suggests that each pot contained seeds that germinated at about the same time. Under this experimental design the two seeds in each 'pair' necessarily came from different plants (Fig. 1).

Paired comparisons are appropriate for two types of experimental designs (chapter 4 of Snedecor and Cochran [4]). First, when observations are made on the same individuals under two different treatments. Second, when individuals are paired (litter mates for example) to control for extraneous factors that might influence the outcome of the treatments. The latter requires careful consideration of possible effects of extraneous factors. The second appears to be the basis for Fisher's analysis of Darwin's data. However, for a paired design in this context, one would take a number of seeds from a number of parent plants. For each such parent plant, some of the progeny would then be cross-fertilized and others would be self-fertilized. The pairing would then be to

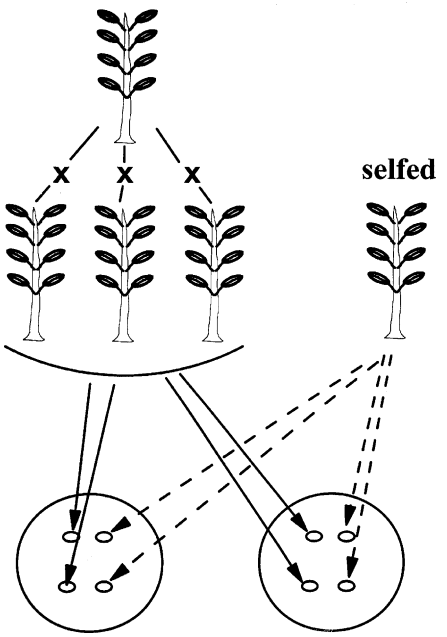

Fig. 1. Experimental design as described by Darwin. Four seeds (ellipses) are shown in each of two pots (circles). Solid arrows indicate crossed seeds, dashed arrows indicate selfed seeds. In each pot the crossed and selfed seeds are from different parent plants.

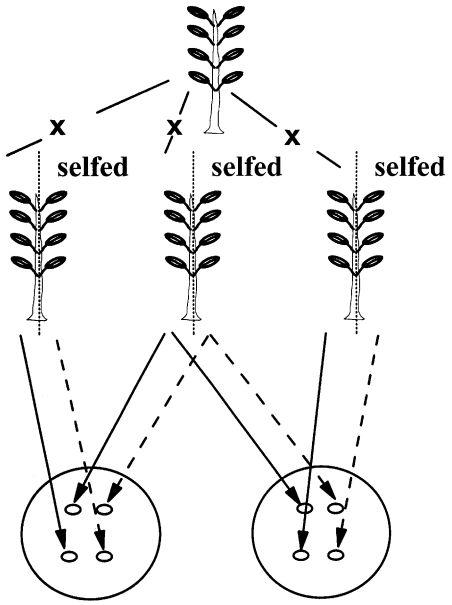

Fig. 2. Experimental design appropriate for true paired comparisons. Four seeds (ellipses) are shown in each of two pots (circles). Solid arrows indicate crossed seeds, dashed arrows indicate selfed seeds. In each pot the crossed and selfed seeds are now from the same parent plant.

compare pairs of plants obtained from the crossed and selfed plants that were derived from the same original parent, and that under the same environmental conditions, as shown in Fig. 2.

3. Fisher's analysis

In his description of Darwin's data, Fisher refers only to Darwin's introduction in which Darwin described Galton's analysis; he does not mention Darwin's description of the experiments which appears on pages 234 and 235 of Darwin's book. Darwin had asked Galton to examine his data and Galton had combined the data from all four pots, rank ordered the cross-fertilized plants and the self-fertilized plants, and then looked at the paired differences. That was, of course, improper and Fisher criticized it in some detail. As for Darwin's experimental design, Fisher points to the need for randomization in assignment of the crossed and selfed plants of each pair to locations in each pot and ends his discussion on validity and randomization with the statement on page 49, *''The one flaw in Darwin's procedure was the absence of randomisation''*. He does not criticize other aspects of Darwin's experimental design, and interestingly, Fisher, the man who played such a large role in the development of randomized plot designs, does not raise the possibility that Darwin's pots play the role of plots in a split plot design.

The problem of proper randomization aside, Fisher accepts Darwin's pairing of cross- and self-fertilized plants and forms the differences between the pairs. He then compares the t-test and randomization test. The null hypothesis is that there is no difference between crossed and selfed plants. The mean difference is $314/15 = 20.933$ and the standard error of the mean is 9.746 giving

$t = 2.148$; adjusted for continuity, $t = 2.139$. For 14 degrees of freedom, a one-sided test gives a probability of 2.485% (2.529% adjusted) of finding a difference as large or larger than that found between the crossed and selfed plants. Under the null, the t-test assumes the differences are normally distributed so Fisher calls that the analysis under the normal hypothesis.

Fisher then develops his randomization test under what he calls the general hypothesis about distribution of the differences. There are 15 differences with a total difference of 314 between the crossed and selfed plants. On the null hypothesis that the crossed and selfed plants are random samples from the same distribution, each difference could have appeared with positive or negative sign with equal probability. There are a total of $2^{15} = 32\,768$ possible arrangements of signs with the 15 differences obtained. Assuming all are equally likely, only 863 of those arrangements give differences of 314 or more, giving a probability of 2.634% for that one-sided test. That is the now widely used and famous Fisher randomization test.

Incidentally, knowing the robustness of the t-test it is perhaps not surprising that the two tests give results that are close.

4. Other randomization tests

Since there is considerable doubt that Darwin's pairings provide a rational basis for paired comparisons, what other randomization tests might Fisher conceivably have developed for the analysis of the data? There are two.

4.1. Two random samples

Let us dispose of the less likely one first. If the pairing of the crossed and selfed plants in each pot is invalid, as we strongly suspect it is, *and* there are no differences between pots, then we can take as the null that we have two random samples of 15 from the same population. In that case, the randomization test would be to pick a random sample of 15 out of the actual 30 heights and assign them to one group and the remaining 15 to the other group. Now there are

$$^{30}C_{15} = \frac{30!}{15!15!} = 155\,117\,520$$

possible randomizations. For each, calculate the difference of means and, assuming all randomizations are equally likely, what fraction give a difference of means $\geqslant 314$? That would have to be done by Monte Carlo sampling. Clearly the difference in design gives a randomization test different from Fisher's. However, one cannot ignore the possibility that there are differences between the pots.

4.2. Randomization within pots

There might actually be some differences between pots so they play the role of plots in an agricultural experiment; but there is no particular pairing within pots. On the assumption that the pots are independent repetitions, we can do the randomizations of the last section within pots. Table 2 summarizes the means, the differences between means, the number of possible

Table 2
Pot identifier ('Pot'), Number of pairs of seeds ('No.'), mean height of crossed ('Av. X') and selfed plants ('Av. Self'), Difference between those mean ('Diff.'), number of possible randomizations in that pot ('Rand.'), and probability of the observed difference in means ('Prob.') under the assumption that the pots are independent repetitions

Pot	No.	Av. X	Av. self	Diff.	Rands.	Prob.
I	3	150.667	154	−3.333	20	0.55
II	3	167	152	15	20	0.1
III	5	169	135.4	33.6	252	0.00794
IV	4	156.25	128.5	25.75	70	0.0857

randomizations and the probability attached to the possible samples that have a difference between the means greater than or equal to the actual difference found, all within pots.

Looking at the data in Table 1 and the means in Table 2, pot I seems to differ considerably from the others. Going back to Darwin's description, we find on page 235:

Shortly afterwards (the measurements in Table 1) one of the crossed plants in Pot I. died; another became diseased and stunted; and the third never grew to its full height. They seemed to have been all injured, probably by some larva gnawing their roots. Therefore all the plants on both sides of this pot were rejected in the subsequent measurements. When the plants were fully grown they were again measured to the tips of the highest leaves, and the eleven crossed plants now averaged 68.1, and the eleven self-fertilised plants 62.34 inches in height; or as 100 to 91.

Accepting Darwin's conclusion and rejecting pot I, it looks as though there might be a significant difference between the crossed and self-fertilized plants. The appropriate parametric test is now an analysis of variance for a comparison of two treatments (crossed and selfed) in three independent experiments with different numbers of replicates in each experiment.

5. Conclusion

It is of historical interest that, although Fisher devised his randomization test for paired data to test Darwin's data on growth of cross- and self-fertilized corn, the data do not meet the rigorous criteria for paired data. Nonetheless, Fisher's randomization test, historically the first non-parametric test, remains one of the important tests for truly paired data. But he could also have developed randomization tests for non-paired data to analyze the same data set.

References

[1] R.A. Fisher, The Design of Experiments, 1st Ed., Oliver & Boyd, London, 1935.
[2] C. Darwin, The Effects of Cross and Self Fertilisation in the Vegetable Kingdom, John Murray, London, 1876.
[3] O. Kempthorne, T.E. Doerfler, The behaviour of some significance tests under experimental randomization, Biometrika 56 (1969) 231.
[4] G.W. Snedecor, W.G. Cochran, Statistical Methods, 6th Ed., The Iowa State University, Ames, IA, 1967.