40. Biometrisches Kolloquium der Deutschen Region. Münster, 15 -18 März 1994

Randomization and bootstrap tests in factorial experiments:

Does analysis follow from design?

Margriet Stapel & Cajo J. F. ter Braak

DLO-Agricultural Mathematics Group,
Box 100, NL-6700 AC Wageningen

It is common wisdom among biometricians that the correct analysis of designed experiments follows unequivocally from the experimental design, that is to say from the random allocation of units to treatments (Nelder, 1964). The classical F-test finds its justification in providing a good approximation to the randomization test result. The validity of the analysis thus depends on proper randomization - something the experimenter has control of - rather than on normality assumptions, which may be false in practice. Neyman, Pitman and Kempthorne did pioneering work for this view concerning randomized block designs with a single factor and Latin squares.

However, what happens in factorial experiments with or without blocks? The design determines that the randomization test should be based on the complete permutation of the observations or the permutation of observations within blocks (Manly, 1990). The resulting tests of main effects and interactions have the wrong size, even for normal errors, except if all effects are nil, as we will show. In the program for Randomization Testing (RT 1.01) by Manly (1990), mean squares are used as a computationally easier equivalent of the F-statistic. However, this equivalence does not hold in factorial experiments and mean squares yield tests that are conservative.

Apparently, proper randomization in itself does not always lead to valid randomization tests. We discuss other proposals which utilize the treatment structure more carefully, in particular, permutation of the levels of one factor within levels of other factors to test a main effect and permutation or bootstrapping of residuals (Fisher and Hall, 1990; ter Braak, 1992).

Fisher, N. I. and Hall, P. (1990). On bootstrap hypothesis testing. *Austral. J. Statist.* **32,** 177-190.

Manly, B. F. J. (1990). *Randomization and Monte Carlo methods in biology.* London: Chapman and Hall.

Nelder, J. A. (1964). The analysis of randomized experiments with orthogonal block structure. I & II. *Proc. Roy. Soc. A.* **283,** 147-178.

ter Braak, C. J. F. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and related techniques,* K. -H. Jöckel, G. Rothe, and W. Sendler (eds), 79-85. Berlin: Springer Verlag.

Welch, W. J. (1990). Construction of permutation tests. *Journal of the American Statistical Association* **85,** 693-698.

Addition to Margriet Stapel and Cajo J.F. ter Braak, 1994 on asymptotic power, 1996.

Example 4 x 2^2 factorial experiment with fixed factors A and B in 4 replications.

Classical ANOVA table

source	df	EMS
A	1	$\sigma^2 + 8\ K_A^2$
B	1	$\sigma^2 + 8\ K_B^2$
AB	1	$\sigma^2 + 4\ K_{AB}^2$
residual	12	σ^2
Total	15	$\sigma^2 + 8/15\ K_A^2 + 8/15\ K_B^2 + 4/15\ K_{AB}^2$

If the observations are randomized, the following 'null model' ANOVA-table holds true.

source	df	EMS
A	15	$\sigma^2 + 8/15\ K_A^2 + 8/15\ K_B^2 + 4/15\ K_{AB}^2$
B	15	$\sigma^2 + 8/15\ K_A^2 + 8/15\ K_B^2 + 4/15\ K_{AB}^2$
AB	15	$\sigma^2 + 8/15\ K_A^2 + 8/15\ K_B^2 + 4/15\ K_{AB}^2$
Residual	15	$\sigma^2 + 8/15\ K_A^2 + 8/15\ K_B^2 + 4/15\ K_{AB}^2$
Total	15	$\sigma^2 + 8/15\ K_A^2 + 8/15\ K_B^2 + 4/15\ K_{AB}^2$

For testing A ($H_0 : K_A^2 = 0$) in the presence of an effect of B ($K_B^2 > 0$) the EMS for A in the randomization is

$\sigma^2 + 8/15\ K_B^2 + 4/15\ K_{AB}^2$

whereas the EMS in the actual experiment is σ^2. Randomization thus leads to mean squares of the A-effect that are systematically greater than the mean squares in the classical ANOVA table. Consequently, the randomization test for A using the mean square of A is conservative.

If the F-statistic is used instead of the mean square, the numerator and denominator of the F for the effect of A are equal in expected value. If there is an effect of B, what is being randomized has a bimodal error distribution, that will not yield the desired critical points of the null distribution. Consequently, the randomization test for A using the F-statistic has the wrong size.

The bias in the size of the randomization F-test will be as for error distributions with a too high kurtosis, because the error distribution being randomized is bimodal. The bias is known to be small in balanced experiments. Because of the F-statistic is asymptotically pivotal (otherwise known as the robustness of the ANOVA for tests of means), the randomization F-test is asymptotically of the correct size. Under normality, the asymptotic power of the randomize-y-test using the F-statistic is asymptotically the same as the F-test in classical ANOVA, as can be seen as follows. Under normality the F-statistic has under the alternative hypothesis a noncentral F-distribution with noncentrality parameter relating to $8 K_A^2 / \sigma^2$. Both tests use this test statistic. Any difference in power is therefore solely determined by the critical values of the tests (i.e. by their size!). Asymptotically the critical values are identical because the F-statistic is asymptotically pivotal under the null; the randomization generates a randomization distribution of the F-statistic that converges in distribution to the normal F with the corresponding number of degrees of freedom, irrespective of the real effect of A.

Consequently, RT with randomizing observations gives the wrong results!
Conclusion 2: analysis does not unequivocal follow from design.