SOME REMARKS ON THE MISSING PLOT ANALYSIS

By SUJIT KUMAR MITRA

Indian Statistical Institute, Calcutta

SUMMARY. The analysis of variance of incomplete data from randomised block and latin square experiments is considered and the expected values, under the null hypothesis, of the treatment and error mean squares are obtained. For simplicity, only the case of a single missing observation is considered. The results could be similarly extended to the case of multiple missing observations in a more general situation where the null hypothesis need not be true.

1. Introduction

It is now recognised that the justification of the customary F-test in ANOVA for designed experiments has to be sought elsewhere and not in the normality and independence assumptions of the observed random variables (an assumption which is most certainly untrue). Several authors (Neyman (1935); Welch (1937); Pitman (1938); and more recently Kempthorne (1955); Wilk and Kempthorne (1955) among others) investigated the possibility of validating this test as an approximate randomisation test. According to them, the stochastic character of the observed variables is primarily due to the random assignment of the treatments to the experimental units and it is possible (theoretically at least) to write down their joint distribution as soon as the randomisation procedure R is specified. Consider an experiment involving N experimental units where it is desired to compare t treatments. Let $X_u(k)$ be the (hypothetical) yield of the u-th experimental unit when it receives treatment $k (u = 1, 2, \ldots, N,\ k = 1, 2, \ldots, t)$. The treatments are said to be equal in their effects if every plot gives the same yield irrespective of the treatment applied, i.e. if,

$$X_u(1) = X_u(2) = \ldots = X_u(t) \text{ for } u = 1, 2, \ldots, N. \qquad \ldots \quad (1.1)$$

Usually however we shall not be interested in establishing such a stringent hypothesis (1.1) and are satisfied in detecting deviations from (1.1) only in so far as they imply, differences in total yields (over all the N units), i.e., in deviations from

$$\sum_u X_u(1) = \sum_u X_u(2) = \cdots = \sum_u X_u(t). \qquad \ldots \quad (1.2)$$

To what extent this is achieved by the ANOVA F-test in some classical designs (like the Randomised Block Design, Latin Square etc.) has been rather thoroughly examined by all those authors and for a discussion on this subject the reader is referred to Kempthorne's book (1952). All their researches tend to show that the F-test is unbiased in a certain sense, namely, if (1.1) be true, both the numerator in F(the treatment m.s.) and its denominator (the error m.s.) have the same expected value. Conditions under which this is true under (1.2) also are known. The object of the present

paper is to demonstrate (in two simple situations) that this is no longer true with the ANOVA F-test when the yields on some of the units, in an otherwise well-designed experiment, are missing. For computing the expected values we consider independent repetitions of R, with the same set of experimental units reporting missing yields each time. They are derived making use of certain known results concerning the average values of mean squares in the analysis of such experiments with complete data.

2. RANDOMISED BLOCK DESIGN [ONE PLOT MISSING]

Here the experimental units (plots) are arranged in r blocks of t plots each and in each block the t treatments are assigned to the t plots completely at random. Let $X_{ij}(k)$ be the yield of the j-th plot in block i, when it receives treatment k and x_{ik} the observed yield of treatment k in block i. For simplicity of discussion we shall assume that the 1st plot in block 1 is missing which under R had received treatment m (itself a random variable), so that x_{1m} is reported to be missing. In such a case Yates' method of fitting constants (1933) for estimating the missing yield leads to the following estimate for x_{1m} :

$$\hat{x}_{1m} = \frac{rB_1' + tT_m' - G'}{(r-1)(t-1)} \qquad \ldots \quad (2.1)$$

where B_1' = total yield for the $(t-1)$ plots in block 1 for which yields were obtained

$$= \sum_{k \neq m} x_{1k}$$

T_m' = total yield for the $(r-1)$ plots of treatment m for which yields were obtained $= \sum_{i \neq 1} x_{im}$

G' = total of all the observed yields,

and the ANOVA table is obtained as follows:

TABLE 2.1. ANOVA FOR A RANDOMISED BLOCK DESIGN (ONE PLOT MISSING)

sources of variation	d.f.	sum of squares
treatment	$t-1$	$(T)_m$ (obtained by subtraction)
error	$(r-1)(t-1)-1$	$(E)_m$ (obtained as the error s.s. in the completed data inserting $\hat{x}_{1m}$ for the missing x_{1m})
treatment + error	$r(t-1)-1$	$(T+E)_m = \sum_{k \neq m} \left(x_{1k} - \frac{B_1'}{t-1}\right)^2 + \sum_{i=2}^{r} \sum_{k} \left(x_{ik} - \frac{B_i}{t}\right)^2$

$$B_i = \sum_{k} x_{ik}$$

Let $(T)_e$, $(E)_e$ and $(T+E)_e$ denote the sums of squares due to Treatment, Error and Treatment + Error respectively computed from the complete data, if x_{1m} were available. Then the following lemma can be easily established.

Lemma 2.1:

$$(E)_e-(E)_m = \frac{(r-1)(t-1)}{rt}(x_{1m}-\hat{x}_{1m})^2$$

$$(T+E)_e-(T+E)_m = \frac{t-1}{t}\left(x_{1m}-\frac{B_1'}{t-1}\right)^2$$

Hence
$$\mathcal{E}(E)_m = \mathcal{E}(E)_e-\mathcal{E}\left[\frac{(r-1)(t-1)}{rt}(x_{1m}-\hat{x}_{1m})^2\right] \quad \ldots (2.2)$$

and
$$\mathcal{E}(T+E)_m = \mathcal{E}(T+E)_e-\mathcal{E}\left[\frac{t-1}{t}\left(x_{1m}-\frac{B_1'}{t-1}\right)^2\right]. \quad \ldots (2.3)$$

Let us now assume that (1.1) is true, i.e.

$$X_{ij}(1) = X_{ij}(2) = \ldots = X_{ij}(t) = X_{ij} \quad \text{for all } (ij)$$

and write

$$e_{ij} = X_{ij}-X_{i\cdot} \text{ where } X_{i\cdot} = \frac{1}{t}\sum_j X_{ij}. \quad \ldots (2.4)$$

In this case

$$(T+E)_e = \sum\sum\left(x_{ik}-\frac{B_i}{t}\right)^2 = \sum_i\sum_j e_{ij}^2 = r(t-1)A \quad \text{(say)}$$

and
$$\left(x_{1m}-\frac{B_1'}{t-1}\right)^2 = \frac{t^2}{(t-1)^2}\left(x_{1m}-\frac{B_1}{t}\right)^2 = \frac{t^2}{(t-1)^2}e_{11}^2.$$

Hence
$$(T+E)_m = \sum_i\sum_j e_{ij}^2 - \frac{t}{(t-1)}e_{11}^2 = \mathcal{E}(T+E)_m. \quad \ldots (2.5)$$

Also since
$$\mathcal{E}(T_m') = \sum_{i=2}^{t} X_{i\cdot}, \text{ we have}$$

$$\mathcal{E}(\hat{x}_{1m}) = \frac{B_1'}{t-1} \quad \ldots (2.6)$$

and hence

$$\mathcal{E}(x_{1m}-\hat{x}_{1m})^2 = \left(x_{1m}-\frac{B'_1}{t-1}\right)^2 + V(x_{1m}-\hat{x}_{1m})$$

$$= \frac{t^2}{(t-1)^2}\, e^2_{11} + \frac{t^2}{(r-1)^2(t-1)^2}\, V(T'_m)$$

$$= \frac{t^2}{(t-1)^2}\, e^2_{11} + \frac{t}{(r-1)^2(t-1)^2} \sum_{i=2}^{r} \sum_{j} e^2_{ij} \qquad \ldots (2.7)$$

It is also known that

$$\mathcal{E}(E)_e = (r-1)(t-1)A. \qquad \ldots (2.8)$$

Hence
$$\mathcal{E}(E)_m = (r-1)(t-1)\,A - \frac{(r-1)t}{r(t-1)}\, e^2_{11} - \frac{A'}{r}$$

and
$$\mathcal{E}(T)_m = (t-1)A - \frac{t}{r(t-1)}\, e^2_{11} + \frac{A'}{r} \qquad \ldots (2.9)$$

where
$$A' = \frac{1}{(r-1)(t-1)} \sum_{i=2}^{r} \sum_{j} e^2_{ij}. \qquad \ldots (2.10)$$

The expected values of the treatment mean square and the error mean square are shown in Table 2.2.

TABLE 2.2. EXPECTED VALUES OF MEAN SQUARES IN TABLE 2.1

sources of variation	d.f.	expected value of mean square
treatment	$t-1$	$A^* + \frac{1}{r(t-1)}(A'-A^*)$
error	$rt-r-t$	$A^* + \frac{1}{r(rt-r-t)}(A^*-A')$

$$A^* = \frac{1}{r(t-1)-1}\left\{\sum_{i}\sum_{j} e^2_{ij} - \frac{t}{t-1}\, e^2_{11}\right\}$$

Hence the F-test would be unbiased if and only if $A^* = A'$.

Thus if the average error variance in the (r-1) complete blocks is larger than the error variance of the incomplete block 1, we have $A' > A^*$, and then the treatment mean square would have a larger expected value than the error mean square. Consequently we would expect a larger proportion of significant F values even under the null hypothesis (1.1), than what we normally anticipate at the nominal level of testing.

3. Latin square design [one plot missing]

Here $N = t^2$ and the t^2 plots are arranged in t rows and t columns. The t treatments are assigned at random to these plots in such a way that each treatment occurs once in every row and once in every column. The randomisation procedure R in a Latin square is discussed in Fisher and Yates Tables (1948). Let $X_{ij}(k)$ be the yield of plot (i, j) (i-th row and j-th column) when it receives treatment k and x_{ik} the observed yield of treatment k in row i. We shall assume that plot (1, 1) is missing which under R had received treatment m so that x_{1m} is reported to be missing. Here the estimate for the missing yield (Yates, 1936) is computed as

$$\hat{x}_{1m} = \frac{t\,R_1' + t\,C_1' + t\,T_m' - 2G'}{(t-1)(t-2)} \qquad \ldots \quad (3.1)$$

where

R_1' = total yield for the $(t-1)$ plots in row 1 for which yields were obtained,
C_1' = total yield for the $(t-1)$ plots in column 1 for which yields were obtained,
T_m' = total yield for the $(t-1)$ plots of treatment m for which yields were obtained, and
G' = total of all the observed yields.

The ANOVA table is obtained as follows :

TABLE 3.1. ANOVA FOR A LATIN SQUARE (ONE PLOT MISSING)

sources of variation	d.f.	sum of squares
treatment	$(t-1)$	$(T)_m$ (obtained by subtraction)
error	$(t-1)(t-2)-1$	$(E)_m$ (obtained as the error s.s. in the completed Latin Square inserting $\hat{x}_{1m}$ for the missing x_{1m})
treatment+error	$(t-1)^2-1$	$(T+E)_m$ (obtained as the (treatment+error) s.s. in the completed Latin Square inserting $\tilde{x}_{1m}$ for the missing x_{1m})

$$\tilde{x}_{1m} = \frac{t\,R_1' + t\,C_1' - G'}{(t-1)^2}$$

Let $(T)_e$, $(E)_e$ and $(T+E)_e$ be the sums of squares due to Treatment, Error and Treatment + Error respectively computed from the complete latin square if x_{1m} were available. Then the following result holds :

Lemma 3.1:

$$(E)_e - (E)_m = \frac{(t-1)(t-2)}{t^2}\,(x_{1m} - \hat{x}_{1m})^2$$

$$(T+E)_e - (T+E)_m = \frac{(t-1)^2}{t^2}\,(x_{1m} - \tilde{x}_{1m})^2.$$

Hence $$\mathscr{E}(E)_m = \mathscr{E}(E)_c - \mathscr{E}\left[\frac{(t-1)(t-2)}{t^2}(x_{1m}-\hat{x}_{1m})^2\right]$$

$$\mathscr{E}(T+E)_m = \mathscr{E}(T+E)_c - \mathscr{E}\left[\frac{(t-1)^2}{t^2}(x_{1m}-\tilde{x}_{1m})^2\right]. \quad \ldots (3.2)$$

Let us now assume that (1.1) is true, i.e.

$$X_{ij}(1) = X_{ij}(2) = \ldots = X_{ij}(t) = X_{ij} \text{ for all } (i,j)$$

and write $$e_{ij} = X_{ij} - X_{i\cdot} - X_{\cdot j} + X_{\cdot\cdot}, \quad \ldots (3.3)$$

where $$X_{i\cdot} = \frac{1}{t}\sum_j X_{ij},\ X_{\cdot j} = \frac{1}{t}\sum_i X_{ij},\ X_{\cdot\cdot} = \frac{1}{t}\sum_i X_{i\cdot}.$$

As before it can be easily seen that

$$x_{1m} - \tilde{x}_{1m} = \frac{t^2}{(t-1)^2}\, e_{11}$$

and that $$(T+E)_c = \sum_i \sum_j e_{ij}^2 = (t-1)^2 A \quad \text{(say)}.$$

Hence $$(T+E)_m = \sum_i \sum_j e_{ij}^2 - \frac{t^2}{(t-1)^2}\, e_{11}^2 = \mathscr{E}(T+E)_m. \quad \ldots (3.4)$$

Also since $$\mathscr{E}(T'_m) = \frac{1}{t-1}\sum_{i=2}^{t}\sum_{j=2}^{t} X_{ij},\ R'_i = \sum_{j=2}^{t} X_{1j},\ C'_1 = \sum_{i=2}^{t} X_{i1}$$

and $G' = \sum\sum_{(ij)\neq(11)} X_{ij}$, we have

$$\mathscr{E}(\hat{x}_{1m}) = \tilde{x}_{1m} \quad \ldots (3.5)$$

and hence $$\mathscr{E}(x_{1m}-\hat{x}_{1m})^2 = (x_{1m}-\tilde{x}_{1m})^2 + V(x_{1m}-\hat{x}_{1m})$$

$$= \frac{t^4}{(t-1)^4}\, e_{11}^2 + \frac{t^2}{(t-1)^2(t-2)^2}\, V(T'_m). \quad \ldots (3.6)$$

It is known that

$$V(T'_m) = \frac{1}{t-2}\sum_{i=2}^{t}\sum_{j=2}^{t} e'^2_{ij} \quad \ldots (3.7)$$

where $$e'_{ij} = X_{ij} - X'_{i\cdot} - X'_{\cdot j} + X'_{\cdot\cdot}$$

$$X'_{i\cdot} = \frac{1}{t-1}\sum_{j=2}^{t} X_{ij},\ X'_{\cdot j} = \frac{1}{t-1}\sum_{i=2}^{t} X_{ij},\ X'_{\cdot\cdot} = \frac{1}{t-1}\sum_{i=2}^{t} X'_{i\cdot}.$$

It is also known that

$$\mathcal{E}(E)_r = (t-1)(t-2)A. \qquad \ldots (3.8)$$

Hence
$$\mathcal{E}(E)_m = (t-1)(t-2)A - \frac{t^2(t-2)}{(t-1)^2}\epsilon_{11}^2 - \frac{A'}{t-1} \qquad \ldots (3.9)$$

and
$$\mathcal{E}(T)_m = (t-1)A - \frac{t^2}{(t-1)^2}\epsilon_{11}^2 + \frac{A'}{t-1} \qquad \ldots (3.10)$$

where
$$A' = \frac{1}{(t-2)^2}\sum_{i}^{t}\sum_{j}^{t}\epsilon_{ij}^2.$$

The expected values of the corresponding mean squares are shown in Table 3.2.

TABLE 3.2. EXPECTED VALUES OF MEAN SQUARES IN TABLE 3.1

source of variation	d.f.	expected value of mean square
treatment	$t-1$	$A^* + \frac{1}{(t-1)^2}(A'-A^*)$
error	t^2-3t+1	$A^* + \frac{1}{(t^2-3t+1)(t-1)}(A^*-A')$

$$A^* = \frac{1}{(t-1)^2-1}\left\{\sum\sum\epsilon_{ij}^2 - \frac{t^2}{(t-1)^2}\epsilon_{11}^2\right\}$$

Hence the F-test would be unbiased if and only if $A^* = A'$.

4. Conclusion

Unless the exact nature of the process, by which an observation is missed, is known, it is difficult to make any further comments on the missing plot analysis. It may be worthwhile to note in this connection that if one plot is missed at random from all the available plots, the bias disappears in both the cases considered in this paper.

The bias which we noticed in this paper possibly affects the test of equality of treatment effects only in so far as the customary use of the percentage points of the variance ratio distribution for judging the significance of the computed value of F will either overestimate or underestimate the level of the randomisation test. When the number of missing observations is relatively small, this distortion may be only of minor importance.

REFERENCES

FISHER, R. A. AND YATES, F. (1948): *Statistical Tables*, Oliver and Boyd, Edinburg, 3rd edition.

KEMPTHORNE, O. (1952): *Design and Analysis of Experiments*, John Wiley and Sons, New York.

——— (1955): The randomisation theory of experimental inference. *J. Amer. Stat. Ass.*, 50, 946-967.

NEYMAN, J. (with the co-operation of Iwaskiewicz, K. and Kolodzieczyk, St.) (1935): Statistical problems in agricultural experimentation. *J. Roy. Stat. Soc.*, (Supp.) 2, 107-180.

PITMAN, E. J. G. (1938): Significance tests which may be applied to samples from any population, III. The analysis of variance test. *Biometrika*. 29, 332-335.

WELCH, B. L. (1937): On the z-test in randomised blocks and latin squares. *Biometrika*, 29, 21-52.

WILK, M. B. AND KEMPTHORNE, O. (1955): Derived linear models and their use in the analysis of randomised experiments. *Final Report, Analysis of Variance Project*, Statistical Laboratory, Iowa State College, April 1955.

YATES, F. (1933): The analysis of replicated experiments when field results are incomplete. *Emp. Jour. Exp. Agri.*, 1, 129-142.

——— (1936): Incomplete latin squares. *Jour. Agri. Sc.*, 26, 301-315.

Paper received : March, 1959.