Usefully combining a series of unreplicated cheesemaking experiments

By MURRAY C. HANNAH

Agriculture Victoria, Dairy Research Institute, RMB 2460, Ellinbank, VIC 3820, Australia

(*Received 19 March 1998 and accepted for publication 6 January 1999*)

Summary. Applied dairy research is characterized by experiments for which financial and physical constraints permit only a small number of experimental units. With few units it is difficult to replicate treatments, and without replication experimental error cannot be estimated. The statistical analysis and interpretation of such experiments is problematic. However, if there have been several such experiments it may be possible to perform a combined analysis. Nine unreplicated experiments comparing effects of diet on the composition of cows' milk and on cheese characteristics were jointly analysed as an incomplete block design. This analysis method was contrasted with analyses of individual experiments. For cheese moisture, the key outcome measurement, the assessment of statistical significance concurred for the two methods in 13 out of 21 comparisons of treatments with the control. Sources of error variation allowed for under the two methods were delineated. The combined analysis paradigm provided stronger inference and a wider interpretation of results than could be achieved using analyses for individual experiments. Unequal replication of treatments and unequal concurrence of treatments within experiments over the series gave rise to a wide range of SED. The challenge of presenting results with unequal SED was addressed graphically using error bars. Attention to series design, in particular the apportioning of replication and treatment concurrence across the series of experiments, was shown to ameliorate presentation difficulties and, more importantly, to yield higher precision at no extra cost.

Questions concerning the effects of diet on cows' milk functionality for cheese manufacture are of obvious interest to the dairy industry. However, experiments directed towards such questions that involve the manipulation of diet and empirical measurement of effects on subsequently manufactured cheese are expensive and rarely performed. For precise treatment comparisons, such experiments need to be carefully controlled. Suitable animals must be matched and allocated to treatments and time allowed for animals to adjust to diets. Sufficient quantities of milk are required from each treatment group for cheese manufacture. Facilities must be available for the parallel collection and storage, and subsequent processing into cheese, of milk from cows offered different diets.

Animal experiments are frequently expensive, and genuine replication of treatments within experiments may be limited or absent. However, if several experiments have been performed, and these experiments share sufficient similarities in the way in which they were structured, new possibilities emerge for the analysis

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

of the pooled results. The present paper discusses analysis, interpretation, presentation and design issues surrounding a series of nine experiments conducted at Ellinbank between 1991 and 1994, which investigated effects of diet quality and stage of lactation on milk and cheese characteristics.

MATERIALS AND METHODS

In each experiment, lactating cows were carefully selected from a large herd and matched to form treatment groups (from here on referred to as 'herds') that were as similar as possible in terms of milk production, age, live weight, days in lactation and milk protein genotype. There were ten cows per herd in all experiments except Expt 3, which had nine cows per herd, and Expt 7, which had eight cows per herd. Animals were housed in stalls and given their rations individually. The number of treatments per experiment ranged from two to five (Table 1). Details of dietary treatments are reported in Kefford *et al*. (1995) for Expt 1, Christian *et al*. (1999*a*) for Expts 3, 4 and 5 and Christian *et al*. (1999*b*) for Expt 7. Specifications for treatments in Expts 2, 6, 8 and 9 remain unpublished. These, with the exceptions of the sunflower and maize treatments unique to Expt 2, were designed to repeat treatments in Expts 1, 4 and 5, using the same procedures.

Milk from each herd was collected in a separate vat and used to manufacture cheese on four occasions for each experiment. More than 30 milk and cheese variables were measured (Table 2). Milk measurements corresponded to milk sampled from vats. Cheese measurements corresponded to the batch of cheese manufactured from milk in each vat. Thus, the sampling unit status for milk measurements (vats) and cheese measurements (batches) was equivalent. In the following, for convenience of terminology, we refer to this unit as the batch. Thus, there were four batches for each herd (one batch for each cheesemaking occasion) within each experiment.

Statistical analyses

We considered two approaches to the analysis of experiments in the series.

Analysis A. The first approach, which has been employed in a previous paper (Kefford *et al*. 1995), was to analyse each experiment separately. The estimate of error was based on variance between batches. Measurements were analysed using ANOVA in Genstat 5.3.2 for Windows (Genstat 5 Committee, 1995, 1996) with blocking structure of occasion split for herd and a single factor indexing the treatments in the experiment for the treatment structure.

Analysis B. The second approach to statistical analysis involved analysis of all experiments simultaneously. This was achieved by regarding individual experiments in the series as 'blocks' in an incomplete block design (Mead, 1988). Here, the word 'incomplete' refers to the fact that each block (experiment) contained an incomplete set of treatments (Table 1), as distinct from the familiar randomized complete block design in which each block would contain a full set of treatments.

Although the herds were selected to be similar within experiments, they differed between experiments. Different experiments were subject to different environmental conditions and different qualities of base diets. Such experiment-to-experiment differences may be characterized as block effects and are explicitly allowed for in this analysis. At its simplest, the statistical model for analysis B was applied to the herd means (averaged over occasions within experiments) and consisted of fixed effects for treatment and for experiment, plus a random error term.

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

Table 1. *Summary of treatments applied in nine experiments on diet and milk functionality for cheese manufacture conducted at Ellinbank between 1991 and 1994*

	Herd	Diet†	Lactation		Herd	Diet†	Lactation
Expt 1	1	Sunflower–maize	Mid	Expt 2	5	Sunflower–maize	Late
	2	Sunflower–maize	Late		6	Sunflower	Late
	3	Control	Mid		7	Maize	Late
	4	Control	Late		8	Control	Late
Expt 3	9	Oats–sunflower	Mid	Expt 4	13	Canola–wheat	Mid
	10	Wheat–sunflower	Mid		14	Cottonseed–wheat	Mid
	11	Barley–sunflower	Mid		15	Medium lupin–wheat	Mid
	12	Control	Mid		16	Control	Mid
Expt 5	17	High lupin–low wheat	Late	Expt 6	21	Medium lupin–wheat	Mid
	18	Medium lupin–wheat	Late		22	Medium lupin–wheat	Mid
	19	Low lupin–high wheat	Late		23	Control	Mid
	20	Control	Late		24	Control	Mid
Expt 7	25	High pasture	Mid	Expt 8	30	Medium lupin–wheat	Mid
	26	Medium pasture	Mid		31	Control	Mid
	27	Low pasture	Mid				
	28	Medium lupin–wheat	Mid	Expt 9	32	Medium lupin–wheat	Mid
	29	Control	Mid		33	Control	Mid

† Details of dietary treatments are reported in Kefford *et al.* (1995) for Expt 1, Christian *et al.* (1999*a*) for Expts 3–5 and Christian *et al.* (1999*b*) for Expt 7. Specifications for the dietary treatments in Expts 2, 6, 8 and 9 are unpublished. These, with the exceptions of the sunflower treatment and the maize treatment which were unique to Expt 2, were designed to repeat treatments of Expts 1, 4 and 5 using the same procedures.

For an incomplete block analysis to be successful the design needs to satisfy certain conditions. Firstly, error variance between units (herds) within blocks should be comparable over the series. This means that the physical configuration and environmental conditions under which the component experiments were conducted should not be so different as to materially affect the variation of herds. In the Ellinbank series of experiments a similar physical structure pertained to each experiment: herd size and selection methods were similar, and the same feeding, milking and cheesemaking facilities were used.

Secondly, there should be sufficient repetition of at least some treatments in the series as a whole to permit the error variance to be estimated adequately. Equal replication of treatments is not essential. In the Ellinbank series, there were seven degrees of freedom for error variance estimation. Each experiment contained a 'hay–silage' control treatment, and there were several occurrences of 'medium lupin–wheat' and 'sunflower–maize' diets (Table 1).

Thirdly, it is generally desirable that a design should be 'connected'. This means that comparisons can be made, directly or indirectly, between all treatment pairs. For example, a design consisting of two experiments having treatments A and B in Expt 1 and treatments B and C in Expt 2 would be connected, since treatments A and C may be compared via the comparison of each with treatment B. Conversely, a similar design having treatments A and B in Expt 1 and C and D in Expt 2 would not be connected, since comparison between treatments in different experiments (A and C, for example) would be confounded with experiment (block) effects. It is not difficult to check that the Ellinbank series was indeed connected (Table 1).

Although analysis B is most simply performed on the herd means, we chose to analyse all the measurements on batches. This was done in order to estimate components of variance and thus to facilitate comparisons with method A. To achieve this the model was extended to include random effects for herd within experiment, occasion within experiment, and batch within herd. Measurements were

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

analysed using REML in Genstat 5.3.2 for Windows (Genstat 5 Committee, 1995, 1996.)

RESULTS AND DISCUSSION

Results from Expt 1 using method A are contained in Kefford *et al.* (1995). Analyses using method A for other experiments were also performed (results not presented). A summary of cheese moisture results for the nine experiments using method B in Fig. 1 shows adjusted means and error bars for each treatment. More extensive results from analysis B are given in Christian *et al.* (1999*a*, *b*). Treatment means under analyses A and B frequently differed. However, the majority of the *differences* between treatment means were exactly the same under the two methods. Exceptions occurred where the treatment pairs appeared together in more than one experiment. For example, the medium lupin–wheat and control treatments (mid lactation) occurred together in several experiments (Table 1). Under analysis A, this treatment difference would be reported separately for each experiment. Under analysis B, it is summarized just once, averaging the available information and taking into account simultaneously the pattern of treatment differences between all treatments.

Although differences between estimated means were similar, the assessment of experimental error, and thus the assessment of statistical significance of treatment differences, differed appreciably for the two analyses (Table 2). For cheese moisture the conclusions of Kefford *et al.* (1995), using analysis A, were supported in analysis B. Results for the effects of the medium lupin–wheat treatment, relative to the control, achieved statistical significance at the 5% level with both analyses. However, apparent statistically significant differences under analysis A between pasture treatments, pasture treatments and the control, and pasture treatments and medium lupin–wheat, were not statistically significant under analysis B. This was also true for each of the cottonseed–wheat, canola–wheat, barley–sunflower and a high-protein lupin–low wheat diet compared with the control.

Over the nine experiments, 21 within-experiment comparisons could be made between diet treatments and their respective late- or mid-lactation control treatments (Table 1). For cheese moisture, the assessment of statistical significance at the 5% level by the two analysis methods concurred for 13 of these. On seven occasions analysis A ascribed statistical significance where analysis B did not. On one occasion analysis B ascribed statistical significance where analysis A did not.

Sources of variance

In order to understand how the two analyses differ, consider the sources of variation in the measurements. Systematic effects of treatment aside, unsystematic (or random) variation in the measurements could have been brought about by (1) differences between experimental conditions that apply to whole experiments (e.g. different times of the year, weather conditions, type and history of experimental cows, different qualities of base diet, different experiment management practices), (2) genetic and environmental differences between herds, within experiments, (3) differences associated with the time of sampling within experiments (occasion, e.g. caused by short-term weather fluctuations) and/or (4) effects peculiar to individual batches (unique to a herd and occasion, e.g. fluctuations in feed quality, time of milking, low-level contamination of milk in a particular vat, batch to batch variation in cheese manufacture). A fifth, and important, source of variation concerns the slight differences in definition and execution of the same treatments in different

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

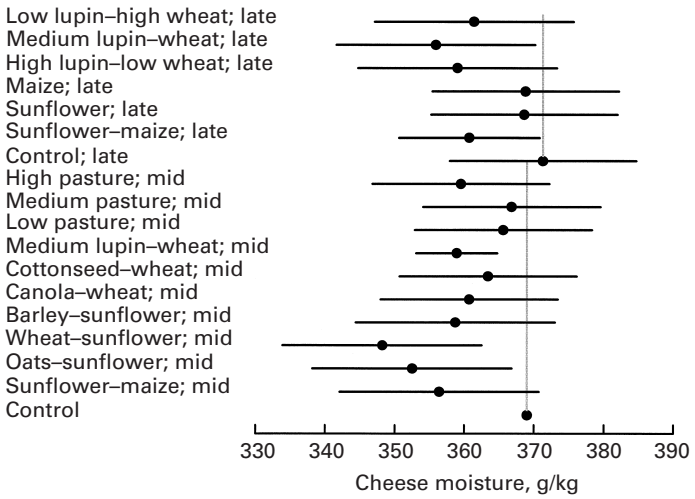


Fig. 1. Adjusted means and error bars for each treatment in nine experiments investigating cows' diet and the functionality of their milk for cheese manufacture. The error bars given are ± LSD (two sided least significant differences, at the 5% level of significance) specifically for testing the difference between each treatment and its corresponding (mid- or late-lactation) control. An error bar that does not intersect the vertical line from the control mean indicates a significant difference ($P < 0{\cdot}05$) between that treatment and the control. The error bar associated with the late-lactation control is specifically for testing the effect of mid *v.* late lactation between the controls. Error bars cannot be applied automatically to other treatment contrasts. Different sets of error bars would be required to compare different sets of non-control treatments. Late, late lactation; mid, mid lactation.

experiments. This, in conjunction with the fact that treatments may have produced different responses under different experimental conditions, gives rise to what may be described as a treatment × experiment interaction variance.

Both methods of analysis described under Materials and Methods allowed for differences between experiments, analysis method A by analysing experiments separately and analysis method B by allowing a separate mean (block effect) for each experiment. Likewise, differences between occasions within experiments were modelled in an equivalent manner in the two analyses, by blocking for occasion. The two analyses differed in the way three sources of variation were handled. These are variance between individual batches (σ_b^2), variance between herds within experiments (σ_h^2) and the treatment by experiment interaction variance ($\sigma_{t\times e}^2$). For simplicity, and as the other sources of variance do not affect the argument, the following discussion restricts attention to these three components of variance.

Milk from each herd assigned to a treatment was collected separately in vats. This implies that the herd, rather than the animal or the batch, was the experimental unit for comparing treatments. Hence the smallest error variance that would have been appropriate for assessing differences between treatments within experiments primarily involved variation between herds. It also involved variation between batches since herd means consisted of averages of four batch measurements. This minimal error variance is written as

$$\sigma_E^2 = \sigma_h^2 + \frac{\sigma_b^2}{4}.$$

This quantity could be estimated only in Expt 6, where replication of the two treatments allowed two degrees of freedom (df) for error (Table 1). It would have been inadvisable to base analysis of all experiments on an error estimate with such

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

Table 2. *Estimated error standard deviations used in two methods of statistical analysis in nine experiments on cows' diet and the functionality of their milk for cheese manufacture*

Variable	σ_A estimate†	σ_B estimate‡	Ratio of estimates σ_B/σ_A
True protein, g/kg	0·21	2·36	11·1
Casein, g/kg	0·16	0·86	5·4
Whey protein, g/kg	0·13	0·22	1·8
Casein number	0·28	0·43	1·5
Fat, g/kg	0·71	2·22	3·1
Total solids, g/kg	0·73	3·31	4·5
Lactose, g/kg	0·20	0·70	3·5
Milk pH	0·015	0·015	1·0
Log (somatic cell count, cells/ml)	0·047	0·092	2·0
Ca, g/l	0·0094	0·0308	3·3
Mg, g/l	0·0011	0·0029	2·7
Inorganic P, g/l	0·0078	0·0143	1·8
Citrate, g/l	0·022	0·049	2·3
Na, g/l	0·0068	0·0165	2·4
K, g/l	0·016	0·025	1·6
α_{s1}-Casein, g/kg total casein	0·028	0·021	0·7
α_{s2}-Casein, g/kg total casein	0·023	0·018	0·8
β-Casein, g/kg total casein	0·024	0·040	1·7
κ-Casein, g/kg total casein	0·0088	0·0097	1·1
γ-Casein, g/kg total casein	0·020	0·026	1·3
Cheese moisture, g/kg	1·7	4·3	2·5
Fat, g/kg	1·7	4·4	2·6
Moisture, g/kg fat-free matter	1·8	3·6	2·1
Protein, g/kg	1·2	1·6	1·4
Salt, g/kg	0·22	0·28	1·3
Salt, g/kg moisture	0·64	1·16	1·8
pH	0·013	0·014	1·0
Moisture:protein ratio	0·010	0·021	2·0
Fat recovery, %	0·29	1·01	3·5
Protein recovery, %	0·30	0·63	2·1
Total solids recovery, %	0·39	1·18	3·0
Solids-not-fat recovery, %	0·32	0·88	2·7
Yield, kg/100 kg milk	0·09	0·36	4·0

† Estimate σ_A is the error SD using method A, which analyses each experiment separately and takes into account only variation between subunits (batches within herds). The σ_A estimates were pooled over experiments for presentation.

‡ Estimate σ_B is the error SD using method B, which takes into account variation between subunits (batches), units (herds) and experiment × treatment interaction variance. All nine experiments were analysed simultaneously under method B. On average, the σ_B was about twice σ_A.

low df and which derived from just one experiment. The risk was high that it may not have been representative of the true variation between herds.

The error term used in analysis A involved variation between batches alone, and may be written

$$\sigma_A^2 = \frac{\sigma_b^2}{4}.$$

This was estimated independently for each experiment in the series.

The error term used in analysis B allowed for batch, herd and treatment by experiment sources of variance, and is written

$$\sigma_B^2 = \sigma_h^2 + \frac{\sigma_b^2}{4} + \sigma_{t\times e}^2.$$

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

This shows that the additional variance allowed for under analysis B, which was not accommodated under analysis A, was $\sigma_h^2 + \sigma_{t\times e}^2$.

Estimates of the error SD from the series, σ_A (pooled over experiments) and σ_B, are presented in Table 2. On average, σ_B was more than double σ_A. The ratio σ_B/σ_A was almost always $> 1{\cdot}0$ (Table 2) and had geometric mean 2·1. Consequently, SED (being proportional to the error SD) were usually much larger, and the assessment of statistical significance more conservative, under analysis method B than under analysis method A.

Consequences for analysis interpretation

Analysis A was based on subunit (batch) variance. This is sometimes referred to as pseudoreplication or repeated measures (Mead, 1988; Maindonald, 1992). The correct interpretation of a statistically significant (i.e. $P < 0{\cdot}05$) result from such analysis is that the apparent difference between treatment groups was very unlikely (i.e. $P < 0{\cdot}05$) to have been due solely to the variation between batches. This is a limited interpretation, as the possibility remains that differences were due to variation between the herds. Effects of treatments and effects of herds were indistinguishable under analysis A.

Differences due to herds may have been expected to be negligible since herds were initially chosen to be as similar as possible. However, there are at least three reasons why it would have been unwise to ignore herd effects. Firstly, the matching between herds can never be perfect. Thus some component of the apparent treatment effect (or the apparent lack of such effect) must have been due to herd. Secondly, several variables other than those used to match herds were measured as outcome variables. It was unlikely that herds were equally well matched for these variables as they were for the variables upon which the matching took place. Thirdly, many quantities change with time: variation present at the time of animal allocation need not be the same at the later time of measurement. Thus, although error variation between herds was expected to be small, without replication this variation remained unknown and confounded with treatment effects at the time of outcome measurement. We must therefore be circumspect about the interpretation of analyses of individual experiments using method A. At most, method A may serve to exclude variation between subunits as a plausible explanation for differences between treatment groups.

The choice of using σ_E^2, σ_A^2 or σ_B^2 as the error term in an analysis depends upon both availability of estimates of these and the interpretation desired. To avoid the limitation in interpretation possible using analysis method A, the error term in the statistical analysis must allow for both herd and batch effects. It can be seen from the above formulae that analyses based on σ_E^2 or σ_B^2 do this, whereas analysis based on σ_A^2 does not.

It is possible that treatments depend upon particular experimental conditions for their effect. A treatment effect evident in one experiment might be quite different in another experiment. Relative to σ_E^2, such an experiment-dependent treatment effect could be judged, quite rightly, as being 'statistically significant'. The treatment effect, although dependent for its expression on certain conditions prevailing during the particular experiment, may nonetheless be real. Whether we wish such ephemeral treatment effects to be detected as being statistically significant depends upon the purposes of the research. In basic research it may well be of interest.

In contrast, by the inclusion of $\sigma_{t\times e}^2$, the error term σ_B^2 took into account variation in treatment effects over different experiments. For a treatment effect to have been

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

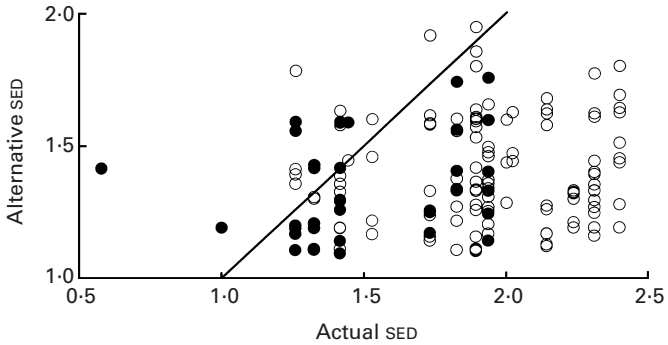


Fig. 2. Predicted SED values between pairs of treatment means under an alternative design plotted against the observed SED for the actual design in a series of nine experiments investigating cows' diet and the functionality of their milk for cheesemaking. ●, SED for contrasts between a treatment and a control treatment; ○, SED for two non-control treatments. The alternative design involved a reallocation of the same treatments to the existing design structure. The underlying error variance for the alternative design was assumed to be the same as was observed in the actual design. The graph shows that the majority of SED values from the alternative design would be smaller than the corresponding observed SED (since the majority of points fall below the line through the origin with a slope of 1), and that the alternative design SED would be less diverse (since points are more spread out horizontally than vertically). These observations hold for contrasts involving controls and for those not involving controls alike, even though repetition of controls was very much higher in the actual design.

statistically significant relative to σ_B^2, there would need to have been evidence that the effect was larger than can normally be explained in terms of variation in treatment effects over different experiments. Analysis B thus broadens interpretation from a narrow framework in which treatments were compared within the restricted context of each individual experiment, to one in which treatments were compared relative to the repeatability of treatment effects over a range of experimental conditions. Hence, effects judged to be statistically significant relative to this error term may reasonably be expected to persist into the future. A statistically significant result, while more difficult to achieve under analysis B, provided considerably stronger inference.

Presentation of results

In analysis A the risk of misinterpretation of 'statistical significance' was considerable. It is therefore arguable whether such analysis should be presented at all (Maindonald, 1992). Be that as it may, the mechanics of presentation of results from method A would be routine. As each experiment was analysed separately, comparisons would be limited to contrasts between treatments within each experiment. Publication, however, should either be accompanied by a disclaimer indicating the limited interpretation available under this analysis, or simply involve presentation of treatment means with no indication of experimental precision.

No such disclaimer would be required under analysis B. In addition all treatments may be compared with one another, not simply those that occurred together within an experiment. The presentation of treatment comparisons may, however, be more difficult. Table 1 shows that treatments were repeated unequally over the series and that the concurrence of treatments within experiments differed markedly. Under analysis B, treatments that were repeated and treatments that concurred (i.e. appeared together within the same experiment) shared smaller SED than other treatment pairs. Consequently, there were 17 distinct, widely differing, SED (Fig. 2). This made presentation of results (means and SED) cumbersome.

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

How the challenge of presenting results should be addressed depends upon the questions being asked. Rarely are all pair-wise treatment comparisons of interest. In Fig. 1 we presented all treatments individually contrasted with the control treatment only, using LSD. This made it possible to present all treatments once, relative to the control, on a single graph. A consequence of the inequality of SED was that the error bars in Fig. 1 could not be applied automatically to treatment contrasts other than those for which they were specifically designed.

Christian *et al.* (1999*a*, *b*) employed analysis B, but restricted presentation of contrasts to those between treatments within experiments, as these were the contrasts of interest. Results were presented graphically. Error bars associated with means were constructed so that statistical significance of treatment differences could be judged according to whether or not error bars overlapped. (Such error bars have been termed 'least significant intervals'; Snee, 1981.) When SED differ, exact error bars that accommodate several pair-wise comparisons do not always exist. Accordingly, approximate error bars for the estimated mean of treatment i were defined as $\pm\delta_i$, where the δ_i were determined as least squares solutions to $\text{LSD}_{ij} = \delta_i + \delta_j$, for all treatment pairs, i and j, within an experiment. The least squares solution for δ_i is given by

$$\delta_i = \frac{1}{k-2}\left(\sum_{j\neq 1} \text{LSD}_{ij} - \frac{1}{k-1}\sum_{1\leqslant l<m\leqslant k} \text{LSD}_{lm}\right),$$

where k is the number of treatments in the set being compared.

Restricting attention to the small number of treatments within experiments enabled a close (usually exact) approximation of $\delta_i + \delta_j$ to LSD_{ij}, as well as focusing specifically upon the comparisons of interest. Other approaches to the problem of defining least significant intervals are given in Goldstein & Healy (1995), and Hannah & Quigley (1996).

Options for more efficient series design

Considerable effort was employed in the allocation of animals to herds and in parallel milking of herds to ensure that within-experiment error (σ_E^2) would be minimized and thus afford a high level of precision for comparisons between treatments. In order to actually measure σ_E^2, replication of at least some treatments within experiments would have been necessary. For example, in the present series of nine experiments, if just one of the treatments per experiment had been replicated (preferably a different treatment in each experiment, so that a typical experiment consisted of, say, four herds and three treatments), a serviceable (9 df) pooled estimate of σ_E^2 would have been available.

However, a further advantage of using analysis B is that estimation of σ_B^2 does not require any replication of treatments to occur within experiments. Instead, the required replication is more profitably gained by repeating treatments over different experiments. By so doing, the number of treatment combinations that occur together within experiments is enlarged. This affects the precision of treatment comparisons and hence is worthy of careful consideration.

In the present design, a less uneven pattern both of replication and of treatment concurrence could have been achieved by trading off the high repetition of a few treatments, with modest replication of a larger set of treatments. For example, consider an alternative design given by reallocating the 18 treatments to the 9 experiments and 33 herds in Table 1. The alternative design could be constructed by first allocating the treatments to the first available 18 herds, and then selecting 15

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press

treatments to allocate to the remaining 15 herds in such a way as to spread treatment concurrence as evenly as possible. Each treatment would then appear with each other treatment in an experiment either 1 or 0 times. Fig. 2 shows that the design so constructed produces SED that are less diverse, and in most cases smaller, than in the present design. Increased experimental precision for essentially the same physical resource means more efficient use of research resources.

A little more resource (namely, three additional herds) would enable complete replication of treatments and even less diversity among SED values. It would then be possible to adequately summarize results using a single (say, average or maximum) SED. Presentation would then be quite uncomplicated, as well as precision being generally improved.

The author thanks Dr M. Christian for provision of measurements, experimental information and many helpful comments. The author also thanks Dr J. Reynolds and Mr G. Hepworth for their helpful critical review. We are grateful to Agriculture Victoria and the Dairy Research and Development Corporation, Australia for financial support.

REFERENCES

Christian, M. P., Grainger, C., Sutherland, B. J., Mayes, J. J., Hannah, M. C. & Kefford, B. 1999*a* Managing diet quality for Cheddar cheese manufacturing milk. 1. The influence of protein and energy supplements. *Journal of Dairy Research* **66** 341–355

Christian, M. P., Grainger, C., Sutherland, B. J., Mayes, J. J., Hannah, M. C. & Kefford, B. 1999*b* Managing diet quality for Cheddar cheese manufacturing milk. 2. Pasture *v.* grain supplements. *Journal of Dairy Research* **66** 357–363

Genstat 5 Committee 1995 *Genstat for Windows Reference Manual* (*Suppl.*) Oxford: Numerical Algorithms Group

Genstat 5 Committee 1996 *Genstat 5 Release 3.2 Command Language Manual*. Oxford: Numerical Algorithms Group

Goldstein, H. & Healy, M. J. R. 1995 The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A* **158** 175–177

Hannah, M. C. & Quigley, P. 1996 Presentation of ordinal regression analysis on the original scale. *Biometrics* **52** 771–775

Kefford, B., Christian, M. P., Sutherland, B. J., Mayes, J. J. & Grainger, C. 1995 Seasonal influences on Cheddar cheese manufacture: influence of diet quality and stage of lactation. *Journal of Dairy Research* **62** 529–537

Maindonald, J. H. 1992 Statistical design, analysis, and presentation issues. *New Zealand Journal of Agricultural Research* **35** 121–141

Mead, R. 1988 *The Design of Experiments: Statistical Principles for Practical Applications*. Cambridge: Cambridge University Press

Snee, R. D. 1981 Graphical display and assessment of means. *Biometrics* **37** 835–836

https://doi.org/10.1017/S0022029999003660 Published online by Cambridge University Press