PLANNING AND CONDUCTING SUCCESSFUL FIELD TRIALS TO TEST PRODUCT EFFECTS ON ANIMAL PERFORMANCE

H. H. Van Horn, C. J. Wilcox, M. B. Hall, C. R. Staples, and W. E. Kunkle
Dairy and Poultry Sciences and Animal Science Departments
University of Florida, Gainesville 32611-0920

INTRODUCTION

Proof of efficacy of products and management practices that affect animal performance is needed equally by industry, producers, and advisors of animal producers. Often, previous research is available but there is a desire to determine if similar performance gains hold up on-farm and if the gains are economical. An objective of any research study, field trials included, is to be able to partition the variation that might be attributed to the treatment (or management practice) in question so that the treatment effects can be measured independently from other effects. Our objective in this paper is to assist the user of the data to 1) recognize that variation other than that brought about by the treatment or practice being evaluated exists and 2) to question how well the design of the trial separated extraneous variation away from treatment effects. Additionally, we provide information that may be useful to those designing field trials and analyzing the data statistically.

STATISTICAL INTUITION

Obviously, animal performance varies tremendously. For example, milk production in a selected group of early-lactation cows averaging 80 lb milk/day may range from 65 to 125 lb/day and daily gain in 600-lb grazing cattle averaging 1.5 lb gain/day may range from 0.5 to 2.5 lb/day. A common example whereby variation due to treatment cannot be separated from other important components of variation is when the whole herd receives the test product through a period of time and the researcher attempts to estimate the treatment effect by comparing performance during the test period with performance prior to the start of the experiment. In addition to the treatment applied, any other variation that occurred during the experiment helped to create the observed results. It will not be possible to separate any time-associated source of variation from the treatment effect. These sources are confounded, i.e., intermingled without there being a possibility of separating the different effects.

From past experiences, we know that time effects with lactation studies can include variable effects from such things as stage of lactation, temperature changes, day-length changes, forage maturity, changes in management, changes in health status, etc. The best and most common way to account for variation other than that brought about by treatment is by comparing the treatment with a control. The control is as identical as possible with the treatment except for the administered treatment. With animals, the control is created by selecting animals randomly from groups the same

Presented at the 11th Florida Ruminant Nutrition Symposium, Gainesville, January 13-14, 2000

way the animals in the treated group were selected and managing them the same except for the treatment that is being evaluated. Thus, the control group can be used to represent the performance and variation that is expected in the treatment group during the time that the treatment is administered if there is no effect due to treatment. Statistical approaches permit applying a probability estimate as to whether or not the difference in variation between the groups can be accounted for by the difference in mean performance.

Many examples could be given where a control treatment was not or could not be used and the time effects that were confounded with treatment effects may have caused the observed changes rather than the treatment. One that comes up from time to time is the need to test manure additives that are promoted to increase degradation (fermentation) efficiency and reduce odor in lagoons. Farms usually have only one lagoon, certainly not two that are nearly identical in size, loading, and all other features. Thus, it is very difficult to provide a control to which the treated lagoon can be compared. If the field study is done when the farmer sees the lagoon covered with manure fiber and smelling worse than usual, there is a good chance this will be in the spring after buildup of solids during periods of low temperatures in winter. As summer comes on, fermentation rate increases and odors decrease. If the treatment was added concurrently, the treatment may get false credit for the improvement.

Some things to think about when designing an experiment

What is the question that we wish to answer with this experiment?
How long will it take to see difference?
How many animals will it take to detect a difference?
Can we properly deliver the treatment?
What records are needed to answer the question?
Can we collect the records needed to answer the question?
Are Control and Treated groups equivalent in all things but the treatment?
If not equivalent, can effects of differences be adjusted with covariance, e.g., differences in days-in-milk or beginning weights?
How will the data be analyzed statistically?
Has the consultant that will analyze the data reviewed the design beforehand?

What non-treatment factors may create differences and affect experimental results?	**Ambient temperature** **Weather (rainfall, humidity)** **Forage maturity** **Feed quality** **Feed mixing** **Equipment failure (function)**
What can you do to keep these types of factors from biasing your conclusions?	**Health status** **Days-in-milk** **Stage of lactation** **Beginning body weight** **?**

Inherent in experimentation is the decision as to what will be the experimental unit, i.e., the unit to use in our efforts to partition variation. In lactating cow experiments, often (but not always) it is the individual cow. In order to consider the cow as the experimental unit, we must be satisfied that she really received the treatment. For example, if the treatment were to be an injection of vitamin A, it would be easy to go into existing groups of cows in a herd and randomly select cows within those management groups to receive the injection and cows that do not (controls), and know that treated cows got the injection. If our objectives were to give an equal number of International Units (amount) of vitamin A by way of the ration, we would have to be able to feed cows individually and know that they consumed the desired amount. In both of these cases the cows probably would be the experimental unit. However, with group feeding, when we want to compare no-vitamin A supplementation with vitamin A supplementation, our treatment for the individual cow now becomes vitamin A concentration of the dry matter and not total intake because we will not be able to determine the vitamin A intake of individual cows. If we wanted the treatment to be an absolute amount (such as pounds or grams/day), we would be compelled to use the groups as the experimental units because we could only verify the amount of product that the group received.

In field trials, it is common to construct two groups of cows that have been selected randomly from a relatively uniform source (manipulative balancing is ill-advised), and choose a treatment that allows us to conclude that the individual cow is the experimental unit. In large Florida dairy herds, often it is possible to obtain use of two groups of cows with 100 to 200 cows per group, to assign cows to the groups at random, and then to randomly assign one of the groups to receive the treatment and the other to serve as the control. The numbers of cows are large and, thus, relatively small differences can be detected statistically that may appear to be caused by the treatment. These numbers help us to be able to find differences due to treatment when differences truly exist. Not finding true differences is one type of error that we wish to avoid statistically (Type II error). Unless we design the trial to use the individual animal or measures of treatment effects versus a control within the same animal, we may never be able to quantify treatment effects.

A key assumption when utilizing two groups (treatment and control) is that the difference in group effects is only due to treatment. However, it is worth considering that other group effects possibly could be confounded with treatment that would make groups appear to be affected by treatment when they were not. This illustrates the risk of a another type or error, i.e., statistically concluding that there is a difference due to treatment when, in fact, there is not a difference (Type I error). This is why it is important to replicate an experiment to confirm if the treatment effects can be repeated. Replication greatly reduces the possibility that unrecognized group differences will lead the researcher to a wrong conclusion.

Bad examples where groups were affected by differences other than treatment. We have seen cases where one group was fed in a feed barn and the other outside, where there were differences in shade and cooling, where there was one load of TMR mixed for the two groups and the control group fed first before the treatment was added and mixed in the last half of the load resulting in a TMR with reduced particle size, and many other differences. Were observed differences in groups in these experiments due to treatment or to these other effects that were confounded with the treatment being tested? Another example could be a test of a supplement treatment on growth of grazing cattle. Because the control and treatment supplement could not be fed separately to assigned animals if they all were maintained in the same group, control and treatment animals were separated and maintained on different pastures that were not equal. Were the pastures different enough to cause the groups to grow at different rates that were unrelated to the supplement? In a cooperative experiment with entomologists years ago, cows in two groups were not housed alike. Results indicated that being bitten frequently on the teats and udder by fire ants resulted in a slight and not significant increase in milk yield per cow. Did housing or fly biting create the difference in yield? In other words, do things that are confounded with treatment in the assigned groups affect our assessment of treatment effects? That is why it is important to have replications at several farms, if possible, and to consider reversing treatments within the groups that are available if the treatment lends itself to this protocol. Reversing treatments (changeover experiments) permits assessing treatment effects within cow, group, barn, etc. and, thus, many variables that might potentially be different between groups besides the treatment in question can be separated from treatment effects.

Changeover experiments. Responses to certain kinds of ration changes are relatively quick and effects often are reversible. For example, changes in dry matter intake (DMI) cause change in milk yield within a few days, maybe the next day. Changes in ration protein have been found to be observable within 2 wk and changes in milk fat percentage within a month. The reason there may be a longer time for change in milk fat percentage than a change in milk protein to be observed is that this change may be caused by changes in amounts and proportions of volatile fatty acids, and these changes may depend on adaptation of rumen microbes to the different diets. Experience has shown that it could take as much as a month for these conditions to stabilize but that reversing treatments after 4 wk or one month gives satisfactory comparisons of treatment effects. One month is an especially important length of time for lactation experiments because most DHI data are generated with monthly milk weights and, perhaps also, with monthly milk sampling. When the researcher can conclude that the treatment effects, if any, will be expressed within a given amount of time and, similarly, will dissipate as rapidly (not carry over after treatment is discontinued), reversing treatments in the treatment and control groups is extremely powerful.

The main advantage of changeover experiments is that the biggest source of variation, the cow herself, can be separated accurately from the treatment effect. Reversing treatments, e.g., one month receiving control and one month receiving the test treatment for half the cows and the other half receiving them in reverse order, permits measuring the treatment effect within the cows. Additionally, the average effect of time (month 1 versus month 2) can be determined and partitioned away from the treatment effect.

Changeover experiments are used because many unknown sources of variation are equalized and partitioned away from the measurement of the treatment effects but there is a key consideration. That is whether or not the treatment effect may carryover after the treatment is discontinued. This seldom is a worry but even if carry-over effects are not likely to be a problem, approaches can be utilized to test whether carry-over effects really occurred. Being able to say that a carry-over effect occurred is a very important finding in itself. A simple and efficient experimental design in which carryover (residual) effects can be estimated and tested for significance and in which treatment means can be adjusted for them is shown in Littell et. al. (1995).

Adjusting for pretreatment status (covariance analysis). If the decision is made that reversal trials are not appropriate because it is felt that the treatment effects may be slow in being expressed, the next best approach is to adjust the animals performance after the start of control and test treatments for performance prior to the start of the experiment. For example, a cow producing 100 lb milk/day just before the start of the experiment is expected to produce more over the next two or three months than one that was producing only 40 lb/day just before the experiment. A statistical method (covariance) adjusts the estimates of treatment effect relative to prior status based on the correlation of prior status with ending status in this experiment. For example, consider Figure 1 in which the average milk yield for treated cows is about 10 lb/day more than the controls. This appears dramatic until one sees that they averaged 10 lb/day more before the treatment started. The treated cows appear to have been higher producing cows and simply maintained their higher production after the treatment was administered. With covariance, the mathematical model permits estimation of the treatment and control means after adjusting for the preliminary milk production. In this example, there was no true effect due to treatment after adjusting production during the experiment for average production just before starting.

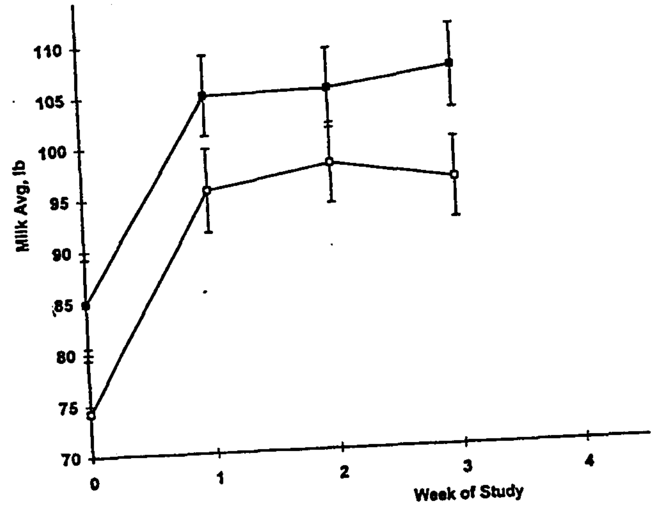

Figure 1. Comparison of Control group (lower line) and Treatment Group milk yields prior to treatment (week 0) and 1, 2, and 3 months after treatment was started.

Conclusion: Always have a control group that will account, as well as possible, for all variable effects other than the treatment being tested. Additionally, use other methods that are suitable for a particular experiment that will partition away as much extraneous variation as possible so that the treatment effects are measured more precisely and independent of factors that may lead to incorrect conclusions. And finally, repeat the experiment to see if the treatment effects are repeatable.

DESIGN AND STATISTICAL ANALYSIS

All animal research presents special problems. The experiment is difficult for the cooperating farmer to manage and expensive for the sponsor. Because of this, experiments must be designed carefully. It is quite easy to conduct an experiment from which no useful conclusions can be drawn. Worse yet, in many cases, this unpleasant result could have been predicted, before the experiment was started.

Prior to beginning any trial, the researcher should outline the data that will be collected and the statistical analysis that will be performed, including the sources of variation that will be accounted for in the mathematical model, the degrees of freedom, and the tests of significance. For example, consider a continuous trial in which the objective was to measure the treatment effect on heifer weight gains (two groups, treated and control) determined from weights obtained just before the treatment was administered and final weights taken when the experiment ended. The model would include only treatment as a main effect with the best dependent variable being final weight adjusted for beginning weight (covariate). Statistical analyses can be performed to determine if differences in beginning weights accounted for significant variation in the gains measured and the control and treatment mean daily gains can be adjusted to be independent of those effects.

Stat 101: Mathematically estimating variation. One needs to have a good estimate of the residual variation (often called the error variance) that will be used to test whether the treatment effects accounted for significant, added variation. In statistical methods, variation is estimated mathematically by summing the squares of deviations from different means or regression lines that are included in the analyses of the data. Variation also can be estimated by the range in the data recorded.

Squaring deviations from means results in extra weighting to deviations that are further from the mean. For example a deviation from the mean of 2.0 when squared weights that deviation four times as much as a deviation of 1.0, a deviation of 4.0 contributes 16 times as much weight as a deviation of 1.0, etc. A constructed example is shown in Figure 2. Note that in Figure 2A, the overall mean is 67.1, the sum of the squares of individual deviations from the overall mean is 3201.8 which can be divided by a number related to the total data points to get a mean of squares. Rather than dividing by 24 (the total number in this data set), statisticians use the total number less one (called degrees of freedom, d.f.) because this number avoids introduction of a bias into probability estimates derived from this and related calculations. For this overall

Figure 2. Understanding variation: Summing squares (S.S.) of deviations from means or regression lines.

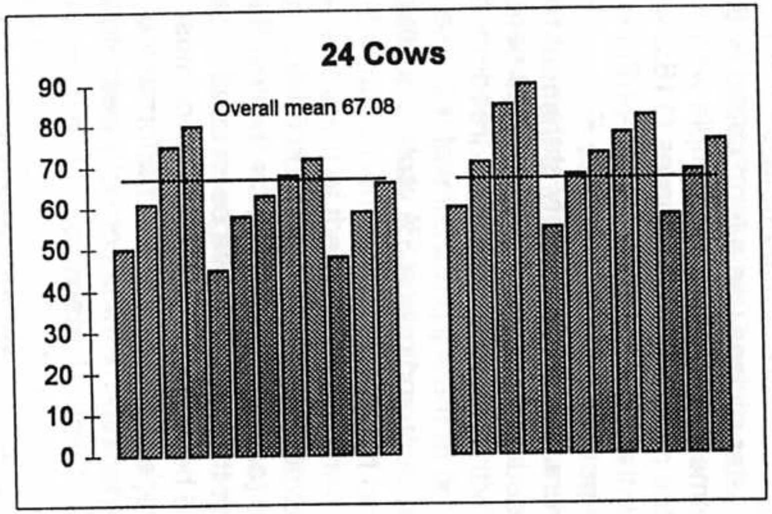

Figure 2A.
S.S. of 24 deviations from overall mean

SS	df	MS
3201.8	23	139.2

Standard deviation (square root of M.S.) = 11.8 lb milk/day
Standard deviation for cows within treatment (Figure 2B) = 10.9 lb milk/day

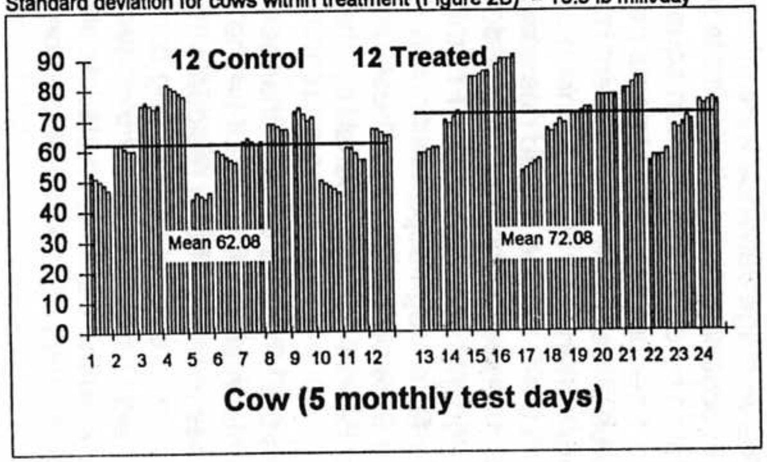

Figure 2C.

	SS	df	MS	*F*
Total (120 observations)	16191	119		
Treatment (Control and Treated)	3000	1	3000	5.1
Cows within treatment (24 cows)	13009	22	591.3	
Residual	182	96	1.89	

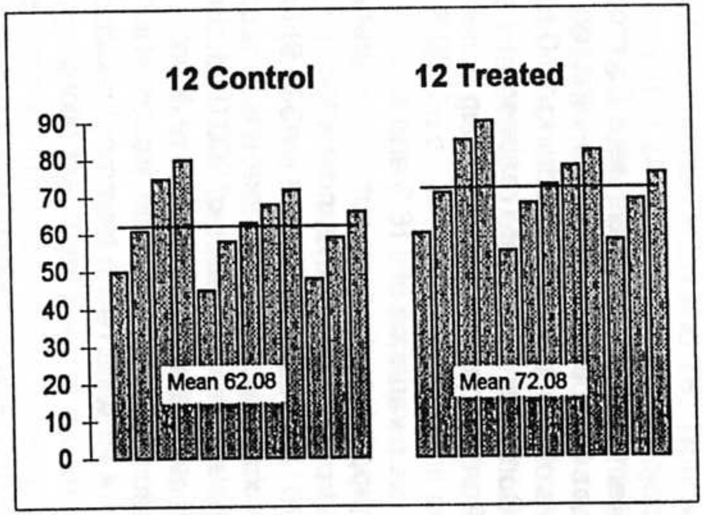

Figure 2B.

	SS	df	MS	*F*
S.S. of 12 deviations from Control mean =	1300.9	11	118.3	
S.S. of 12 deviations from Treatment mean =	1300.9	11	118.3	
TOTAL for cows within treatments	2601.8	22	118.3	
S.S. reduction from Fig. 2A (Treatment S.S.)	600	1	600	5.1

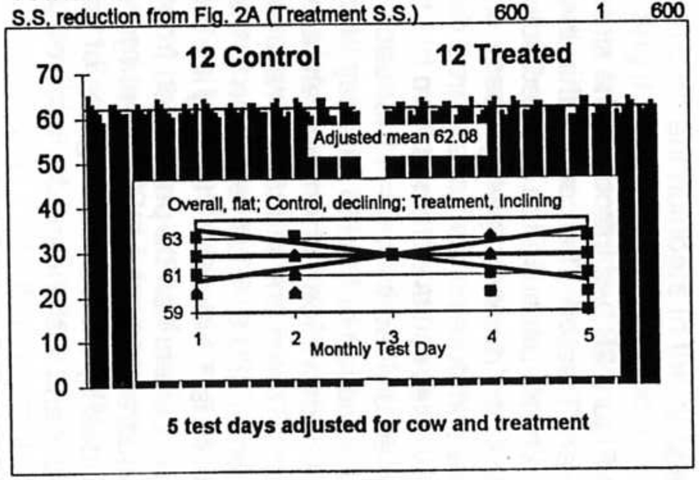

Figure 2D.

	SS	df	MS	*F*
SS of 120 deviations from adjusted mean	182.0	96		
SS reduction with common regression line	0.02	1	0	0
SS reduction with Control and Treatment lines	114.8	1	114.8	161
Residual	67.2	94	0.71	

example, the mean square (variance) is 3201.8/23 = 139.2 and the square root of that is the standard deviation, 11.8 lb milk/day. When data are normally distributed, the mean plus or minus the standard deviation will include, on average, two thirds of the observations.

In Figure 2B, we introduce the situation whereby 12 of the animals were Control and 12 were Treated. The cows within the two groups have exactly the same variation from each other because the Treated cow data were obtained simply by adding 10.0 to the Control cow numbers giving means of 62.1 for Control and 72.1 for Treatment. When one computes a sum of squared deviations from the mean for each group separately, the results obtained from the two groups are exactly the same, i.e., 1300.9 which, when divided by n-1 = 11 for each group, gives a variance of 118.3 and a standard deviation of 10.9 lb milk/day. Note what happens to the total sum of squares for the 24 observations when the data are deviated from the respective means for Control and Treatment, the total is reduced from 3201.8 to 2601.8 (1300.9 + 1300.9) for a reduction of 600.0. This reduction is the sum of squares due to treatment (i.e., Control versus Treatment). Stating it a bit differently, a sum of squares of 600.0 relating to treatment effect can be partitioned from 3201.8 leaving the 2601.8 as the residual sum of squares associated with average variation among cows within the groups which when divided by the correct degrees of freedom (11 + 11 within each treatment = 22) gives 118.3 and a standard deviation of 10.9, the same as in each treatment group.

The average variance within the treatment groups can be used to test if the variation accounted for by treatment is significant. Statisticians use a term called *F* that is calculated by dividing the mean squares for treatment (600.0 in this example with 1 degree of freedom, i.e., 2 treatments minus 1) by the residual mean squares (118.3 with 22 degrees of freedom). If treatment had no effect, the expectation is that *F* = 1.0, i.e., not different than the variation among cows within groups. For this example, *F* = 600.0/118.3 = 5.1. The *F* ratio has a distribution from which a probability statement can be derived based on the appropriate degrees of freedom for the numerator (treatment) and the denominator (residual). For this example with 2 and 22 degrees of freedom, the *F* = 5.1 is far enough departed from the expected 1.0 to have a probability of 0.03, i.e., there is a 97% probability that these means are different and only a 3% probability that a difference this large happened by chance and was not due to treatment.

As the structure of data changes from one experiment to another, the final residual mean squares may not be the appropriate residual (error term) for the test that one wishes to make. If that residual variance is less than should have been used to test if differences are significant, wrong conclusions will be drawn. This type of error most often is made when the researcher has collected repeated measures of performance and analyzed the data with a statistical computing package without correctly describing the mathematical model that should be used in the statistical analysis.

An example of a way to use repeated measures of performance correctly is to estimate trends in production apart from average production over the time in question. Consider that the mean daily productions for the 24 cows in Figure 2A and 2B were derived from five consecutive monthly test-day milk weights taken as part of the Dairy Herd Improvement data collection for that farm. Five separate values for each cow are plotted in Figure 2C; these were utilized to determine the mean production for each cow that was plotted in 2A and 2B. Now we have 120 different numbers in the data set (5 for each of 24 cows) but, for the objective of testing if the overall mean production for the Treatment group differs from the Control group, the appropriate divisor for the *F* test is still the mean squares associated with the cows within the treatments that is associated with the same 22 d.f. that we worked with in Figure 2B. Additionally, it will be interesting to test if there is a difference in the trends in production in the Control and Treated groups over the 5-month period. That is the reason that the monthly data are included in the statistical analysis. First, we need to compute the total variation (total sum of squares) in the data by accumulating the squares of deviations of all of the individual monthly production data from the overall mean of the data set, for this data set it is 16,191. We also need the sum of squares for the cows within the treatments like we generated before for Figure 2B. With statistical methods, the component sums of squares are generated from group totals (e.g., cow or treatment) rather than averages; each cow has five observations making the total five times as large as the average that was used in the Figure 2B example. The sum of squared deviations of each cow total from the overall mean is 13009 which when divided by the 22 d.f. for cows within treatment gives the mean squares for cows of 591.3 (note that this is exactly 5 times as large as the total mean squares for cows for Figure 2B, 118.3). Similarly, the sum of squares for treatment is five times as large, 3000, and, with 1 d.f., the treatment mean squares also is 3000. Thus, the *F* for this expanded model is exactly the same as before, 3000/591.3 = 5.07 with exactly the same probability that the means are different, *P* = .03.

The residual sum of squares (remainder) in this example, after subtracting the variance for treatment and cow within treatment, is 182 with 96 d.f. (119 total d.f. – 1 d.f. for treatment and 22 d.f. for cows) and a mean square of 182/96 = 1.89. Note that if 1.89 had been used to test the mean square for treatment, the *F* for treatment would be estimated incorrectly as 591.3/1.89 = 312 with 1 d.f for treatment (numerator) and 96 d.f. for error (denominator) and the probability that the treatment difference was different would be grossly mistaken to be *P* < .000001. The correct use of this component of the total variation in the data is to see if it can be further partitioned to determine the average trend of the data within cows over the 5-month data collection period and if the trends within Control and Treatment groups differed. The method that we have chosen is to determine the average linear regression over time for the combined data set and test if the linear regressions for the separate treatments differed from the average.

First, look at Figure 2C and see if there appears to be a difference in the slope of the average production trend within the data for individual cows. Does it appear that the average trend is downward within Control cows and upward within Treated cows? The variation associated with different test day measures (182 associated with 96 d.f.) is very small compared with other sources of variation identified previously, e.g. sums of squares of 3000 associated with treatment and 13,009 with cows within treatment. However, used appropriately, this variation may be partitioned in a way to help interpret treatment effects. Figure 2D illustrates that sums of squares of 182 are associated with the variation within cow after data are adjusted for cow and treatment effects. The test that we are going to make for trends is to see if there is an interaction of treatment with time. The average time trend was estimated by fitting a linear regression line through all of the data when they are arranged by test day month from 1 to 5 (see inset in Figure 2D. Squaring the deviations of individual observations from the overall linear regression line (the flat one in 2D with a slope that increases only .008 lb/day monthly) accounted for almost no additional variation than was obtained deviating data from the adjusted mean (residual sums of squares of 182 with 96 d.f.). However, when individual observations were deviated from the regression lines fitted within the respective Control and Treatment groups (crossing lines in 2D inset), the reduction was 114.8 with 1 d.f. The final residual is the sum of squares remaining after .02 and 114.8 are subtracted from 182, i.e., 67.2 with 96 d.f. which leaves residual mean squares of .71 (67.2/96). The reduction of 114.8 obtained by fitting separate regression lines for each treatment is tested by F = 114.8/ .71 = 161 with 1 and 96 d.f. This F = 161 lets us conclude that the trends (slopes) of the data over the 5 months are different in the two groups with a probability of $P < .0001$. The slopes associated with those lines are -.69 lb/day monthly for Control cows and +.69 lb/day monthly for the treated cows. Thus, not only can we conclude that the mean production was higher for treated than control cows ($P = .03$) but that daily production of treated cows was increasing month to month while daily production of control cows was decreasing ($P < .0001$).

Conclusion. The take home summary from this example is that it is important to identify variation wherever possible and partition as much as possible away from treatment; but be careful with repeated measures to make certain that differences in treatment means are tested correctly.

Designs. As mentioned earlier, animal experiments usually can be divided into two categories, continuous and changeovers (e.g., reversals, Latin Squares, switchbacks, etc.). Some experiments may involve both categories. A continuous trial is one in which the animal remains on the same treatment throughout the conduct of the experiment (as in the example in Figure 2). In a changeover, the animal is subjected to two or more treatments. Usually, differences between treatments in a continuous trial are tested statistically by comparing variation in treatment means with variation among animals within treatment. Covariance is a major mechanism for reducing error variance in this case; magnitude of the reduction depends on the correlation between the response variable, e.g., milk yield one month after treatment started, and the covariate, milk production just before the start of the treatment.

Even with covariance, error terms (residual variation) generally are large. Simply stated, for many things that we measure, animals which are treated the same vary considerably in their performance. A common measure of such error variability is the coefficient of variation, CV (error standard deviation divided by the mean for the experiment, multiplied by 100 to put it on a percentage basis). For milk yield of the dairy cow, the CV may be 20% for 305-day records, about 15% for 150-day trials, and perhaps 8-10% in shorter trials. The CV can be reduced somewhat further with changeover designs (Table 1). Many physiological measurements have somewhat larger CV, e.g., Table 2. In a particular trial, the CV may be somewhat higher or lower for various reasons. The effect of the magnitude of the error on the number of animals that need to be assigned to each treatment group is shown in Table 3.

Variation from day to day during a week is shown in Table 4 for milk yield, composition, and properties. Fat % is the most variable of the items listed. These estimates can be used to predict how many daily samples should be taken from each cow to characterize her performance during the week.

Analysis of binary data. At times a nutritionist is faced with a situation where the response variable is categorical, perhaps occurring in only two categories, 0 and 1 or 1 and 2. An example is reproduction data where cows may be recorded as pregnant or not-pregnant. In this case, the data are termed binary; their distribution is binomial. Regretfully, most common measures of reproductive performance are extremely variable (Table 5), and it is rare that real and important differences in reproductive performance can be detected in a single on-farm trial, particularly with less than 40 cows per group. Table 6 gives numbers of animals required in each of two groups with 5 or 10% differences in frequency percentages (e.g., percent pregnant) between groups. Note how large groups would have to be to measure pregnancy differences of 5%. Statistical analysis presents no real problem. Although ordinary least squares analysis of variance of binary data can result in biased tests of significance, the estimates of the effects are unbiased. If desired, statistical procedures exist and are available to perform unbiased tests of significance.

Blocking and balancing. These terms refer to different procedures, one which in general is desirable, and one which in general is undesirable. Blocking refers in animal experiments to the grouping of animals with some commonality. For example, breed, age group, location, etc., all are blocks. Block then is included in the statistical analysis along with its interactions with treatment and other sources of variation as appropriate.

Assignment of animals within block should be at random. This has been part of the foundation for statistical analysis from the beginning. Blocks should be something that you can name and know what it means, for example breed, lactation number, season of calving, etc. Balancing refers to the practice of re-assigning animals to different groups, based on their performance before the experiment starts. The objective is to have the groups start at the same, or nearly the same, level. It can be shown easily that this practice inflates the error variance of subsequent statistical analyses. The more nearly perfect the balancing, the greater the inflation. The actual magnitude of the inflation depends on the correlation between the preliminary measurement and the performance on trial. This undesirable practice is not needed anyway. If one knows individual animals' preliminary performance or status (e.g., body weight, days-in-milk) beforehand, the beforehand data should be used as a covariate to adjust performance measured during the experimental period, i.e., with covariance analysis. Such analyses will adjust the data for differences in group means at the beginning of the experiment without balancing.

Hence, blocking is desirable; the blocks should be called what they are. Balancing is an undesirable and unnecessary practice.

Length of the experiment and experimental efficiency. Researchers sometimes fall into the trap of believing that the longer the trial goes, the better. Once all of the planning, organization and other arrangements have been made, personnel assigned etc., and the experiment is going well, there may be a temptation to let the experiment continue longer than originally planned, particularly in the case of continuous trials. However, CV's for milk yield increase with increasing length of the trial so it is more difficult to detect small effects in long trials than shorter trials.

In growth of dairy calves from birth to 12 weeks, it is true that the longer the trial is, the fewer the number of calves needed to detect percentage differences between treatments because of reduced CV (Table 7). The numbers of calves needed in each of two treatment groups to detect mean differences of 20% at a probability level of $P < .05$ (Type I protection), with Type II protection of 80%, using a two-tailed t test, are in Table 8. The table also demonstrates that, on a percentage basis, it is more difficult to detect differences in some responses than others. In determining how many animals are needed for an experiment, the researcher must focus on the response of critical interest, realizing that this number may be too great for some responses and too small for others.

Certainly an experiment must be long enough to permit the treatment effect to manifest itself. At times longer trials are advantageous; at other times shorter trials are more efficient. Shorter trials, in general, are cheaper, conserve assets, and save time. Longer trials doubtless increase the disproportionality of the data in that the longer the trial, the greater the chances of losing animals (and data).

Thus, the researcher must weigh an appreciable number of factors in determining the length of a trial. Researchers should not be misled by what has been standard practice in the past.

Fixed and random effects. Determination by the researcher as to whether effects are fixed or random is important in performing tests of significance in Analysis of Variance (ANOVA). The material in Table 9 has been known for many years, e.g., for this simple example of an analysis of data from a 2-factor factorial design. It appears in essentially every textbook on Statistical Methods.

Once the decision has been made as to whether an effect is fixed or random the proper tests of significance (ratios of mean squares) can be determined. These procedures also have been developed over the years and appear in many statistical methods texts. Simply stated, a proper test is one in which the denominator contains the same components as the numerator, except that the numerator contains one additional item, the source of variation being tested. The denominator always is random.

Objective procedures for determining whether an effect is fixed or random also are in the literature but may be hard to find. See Henderson (1959) and Damron and Harvey (1987). With disproportionate subclass numbers, tests of significance are approximate unless special procedures are invoked.

An effect is considered to be fixed if all levels of the effect which exist are included in the experiment. An example is sex, when males and females are on the experiment. An effect is considered to fixed if it is selected arbitrarily from a very large (perhaps infinite) population of effects, but represents only the few of interest to the researcher. Examples are ration protein or energy levels. Random effects are those which were selected at random from a very large population. Animal generally is considered random, except possibly when a very small number of animals are on the experiment, and they were selected for specific attributes.

Interactions between random effects and between fixed and random effects are random. Those between fixed effects are fixed. Continuous independent variables are considered to be measured without error and are considered fixed. Even when this is not really true, e.g., body weight as a covariable, it is convenient to consider them so. Methods exist whereby the fact that they truly are not fixed can be handled statistically.

With fixed effects the researcher is limited as to the application of the experimental results to the real world, e.g., results of study of effects of two levels of ration protein apply only to those two levels. It often is acceptable to interpolate possible effects between those two levels (assuming linearity of response) but it is dangerous to extrapolate outside of these boundaries.

With random effects one hopes to apply results to a wide population, e.g., all Holsteins in the U.S., or all Jerseys in Florida. By considering animals to have been selected at random from a very large population, it is reasonable to do so. Animals on the experiment are considered to represent the population to which the results can be applied.

At times, it is not clear whether an effect should be declared fixed or random. Examples are year and farm. Clearly in an experiment involving only two or three years, year would be considered fixed. In some genetic studies, 30 to 50 years have been studied. They may represent all possible year effects and arguably could be considered random. Two or three farms, especially if they are at different locations, would be considered fixed. Thirty farms in Florida, however, probably do represent Florida farms, and could be considered random.

As a last resort, one can analyze data both ways, with an effect being called fixed and then being called random. If no real difference exists between the two models, in estimates of the effects or in their statistical significance, then no problem exists. However if differences occur that affect whether or not treatment other effect differences are considered to be significant, the researcher must make a decision which model to choose and, secondly, to defend it.

Statistical analysis software and mathematical model choices. The software most commonly used by researchers is SAS (SAS Inst., Inc., Cary, NC) with Harvey (Ohio State Univ., Columbus, OH) used by some. To use these software packages correctly, experience is needed. And to choose the right models to use in the analysis also requires considerable experience to avoid possible mistakes that may make interpretation of outcomes different than should be. Therefore, the researcher needs to either have gained that experience themselves or to use a statistical consultant who understands their experimental design and the correct options that can be used to analyze the data. It is possible to commit major Type I and Type II errors in the analysis of data from a single experiment.

Type I and Type II errors. Statisticians rightfully define these errors in terms of the testing hypotheses. Type I errors simply refer to the situation whereby the researcher claims that two treatments (or groups of treatments) differ, for example, when in fact they do not. If the researcher makes such an erroneous claim, a Type I error has been committed. The researcher tries to avoid this error by insisting on high levels of protection, e.g., 95 or 99%, corresponding to levels of significance of $P < .05$ and $P < .01$ (i.e., less than 5 or 1% chance of incorrect conclusion).

Type II errors represent the complementary situation whereby treatments truly differ but the researcher is unable to prove it. Unfortunately most researchers apparently ignore this source of error, since Type II probability levels rarely are presented in the scientific literature. Researchers perhaps should perform experiments in which their protection against committing Type II errors is 70 or 80%. After failing to detect significant effects, the researcher is advised to estimate how large the treatment differences would have had to have been to be detected. Note the probability statements for Type I and Type II errors in Tables 3, 6, and 8.

Prior to starting an experiment, the researcher is advised to consider the following (or a similar) question: How many animals must I assign to this experiment to be able to detect a difference of __ between treatments as being significant at ____ (Type I protection) with Type II protection of ____. In fact, very few researchers actually make this estimate; some discover, after the experiment is over, that their experiment had no chance of success and they could have predicted this in advance. Others discover that they have committed overkill, in that they have assigned too many animals and/or made too many measurements of each, thus wasting precious resources.

Type I errors can be categorized as experimentwise and comparisonwise. The former is estimated as the number of experiments in which one or more erroneous inferences have been made divided by the number of experiments which have been conducted. The latter represents the number of erroneous inferences which have been made divided by the number of inferences made.

Use of multiple range tests. Multiple range tests (MRT), in which every treatment mean is compared with every other treatment mean, can provide a very useful statistical tool for researchers. Yet the procedures are used improperly to a distressing degree. Researchers often confuse themselves and their readers as to the real statistical significance of the differences between treatments which they have detected.

Numerous studies have confirmed the misuse (abuse) of multiple range tests. It is rare in designed nutrition studies that use of MRT are needed. Since the design of the experiment is under control of the researcher, orthogonal contrasts can be devised *a priori* and conducted; then the appropriate comparisons can be determined. In general, MRT are contraindicated when the experimental design is factorial or the treatment levels are quantitative. For example, if ration protein levels of 12, 14, 16, and 18% are tested, orthogonal comparisons that test whether those responses are linear or curvilinear (quadratic or cubic) make much better sense than to make the six comparisons that could be made if each were compared with every other.

After the experiment has been completed and the appropriate orthogonal comparisons of treatment means have been performed, certainly the researcher can make a limited number of additional nonorthogonal comparisons, based perhaps on interesting or unexpected results of the experiment. Armed with the results of the orthogonal comparisons, the researcher is unlikely to reach any unwarranted conclusions. If such comparisons are made, such should be clearly stated.

The experimentwise error rates of several MRT are in Table 10. High error rates associated with the comparison of several means prompted Gill and Hafs (1971) to recommend that the Duncan MRT no longer be used. Since the LSD (least significant difference) is associated with even higher experimentwise Type I error rates, the inference is that the LSD also should not be used. In any event, if a MRT is used, the test used should be stated clearly by the researcher, and the researcher should be aware of the implications of its use. Oddly, one very rarely sees the DMRT used with probabilities of $P < .01$, only with $P< .05$.

Repeated measures. As discussed in the statistics examples in Figure 2A-D, oftentimes the researcher is interested in time trends during the trial. This is possible if measurements are made repeatedly at different times during the experiment. These repeated measure experiments, whereby animals stay on the same treatment during the experiment (i.e., a continuous trial) but are cross-classified with time, most often are split-plots. Treatments and time are fixed and animal is random; this is a mixed model. Advantages and disadvantages of such designs, and possible problems (and their solutions) in the statistical analyses are presented by Wilcox et al. (1990). By uses of the SAS program, PROC MIXED, it is possible to use covariance analysis in repeated measures experiments to adjust group means for differences in starting values (Littell et al., 1996; Littell et al., 1998).

When repeated measures over time are collected, one can compare curves and determine the probability that the curves differ (i.e., not parallel). In calf growth studies, for example, body weights are taken periodically so that rations can be adjusted. Gains can be evaluated more efficiently with regression analyses utilizing all weights measured than only utilizing starting and ending weight to compute overall gain. Linear regression, if changes over time are not curvilinear, has the advantage of giving average gain per day (or per selected time unit) adjusted for body weight of animals. Also, other examples are shown later where treatment effects were evaluated by determining the probability that curves were different rather than depending only on probability of means being different.

EXAMPLES FROM PREVIOUS TRIALS

Several examples follow of field experiments that utilized different design approaches. Statistical analyses of some form were utilized in all experiments to give a probability basis to interpretation. Several of these studies were presented and American Dairy Science Association meetings; published abstracts are referenced for those.

Example 1. Effects of feeding rumen protected methionine and modest reduction in dietary protein in a Florida dairy. H. H. Van Horn and M. B. Hall, University of Florida, Gainesville.

In trial 1, two groups of 180 early-lactation Holstein cows (110 DIM) in a large North Florida dairy were utilized to evaluate supplementation of rumen-protected methionine provided via 17 g/d per cow of Mepron® M85 (Degussa Corp.). Feeding was from Aug 5 to Sep 6, 1998. Milk weights were obtained from the Afikim electronic milk recording system utilized by the dairy and milk fat and protein percentages from DHI. The TMR were 18.6% CP, 36% NDF, 35% roughage mostly from corn silage and alfalfa, and contained many byproducts. Milk production for the final 20 d and milk fat and protein percentages from samples at end of trial were analyzed statistically with model including either 10-d average production preliminary to start or preliminary fat or protein, lactation number, and DIM as continuous independent variables. Feeding Mepron significantly increased milk yield (36.5 kg/d vs 35.3) and milk protein percentage (3.06% vs 2.97). Milk fat percentages were not different and averaged 3.0%. A follow-up trial compared performance from the control 18.6% CP TMR with lower CP (17.6%) after removal of 0.2 kg/d apparent excess RDP. No performance differences were detected which prompted a second study with rumen protected methionine in Dec 1997 in which Mepron was added to the lower CP (17.6%) diet and compared with the control (18.6%) CP diet. The design was similar except that the 180 multiparous cows selected to be equal to the control group were in a larger group of cows, all of which were fed Mepron at a rate of 18 g/d per cow. Milk yield and composition differences were not significant; milk yield averaged 41.0 kg/d, milk fat 3.1%, and milk protein 3.17%. Data show maintenance of production with modest reduction of dietary CP and potential for response to methionine supplementation when cows are fed in excess of 18% CP. ABSTRACT from *J. Dairy Sci.* 81(Suppl. 1):109.

Example 2. Protein supplement may improve gains of nursing calves. W. E. Kunkle, P. J. Hogue, E. W. Jennings, and S. Sumner.

Seven trials evaluating the effects of a cottonseed-salt supplement on the performance of nursing calves were conducted on cooperating ranches in Florida. Supplement consumption averaged .64 lb/head/day, calf gains were increased .27 lb/head/day and 2.9 lb of protein supplement was required for each pound of additional gain. Limit feeding a high-protein supplement to calves was profitable in five of the seven trials. Based on these studies and current costs of protein supplements and the value of the calf at weaning, the high-protein creep feed was projected to increase net returns $4/head.The seven trials were conducted on cooperating ranches during 1986-1995. The calves were offered 100% cottonseed meal until they started eating the supplement (usually 2 to 4 weeks) then were switched to a mixture containing 92% cottonseed meal and 8% salt. The salt was used to limit consumption to 1lb/head/day or less. The protein supplement was fed in creep feeders that prevented cows from consuming the supplement. The trials were initiated from June 19 to July 17 and ranged from 46 to 62 days in length. Animal Science Department, University of Florida, Cooperative Extension Service publication SS-ANS-11, July 1997.

Example 3: Utilization of heterogenity of regression to delineate effects of Zn-, Mn-, and Cu-proteinates on milk somatic cell counts, milk yields, and cow mobility in research conducted onfarm. H. H. Van Horn, J. K. Shearer, C. J. Wilcox, and W. de Groot, University of Florida, Gainesville.

Three hundred fifty-four cows from four groups (two primiparous and two multiparous) in a dairy near Jacksonville, FL were utilized in a field research study during summer months to determine effects of Zn-, Mn-, and Cu-proteinates (ZMC Optimin®, Ducoa Technical Products, Highland, IL) on milk yield, SCC, and foot health (as indicated by walking mobility or lameness on concrete). Proteinates supplying 400 mg Zn, 200 mg Mn, and 100 mg Cu/cow daily were provided in one of three daily feedings of a TMR to one primiparous and one multiparous group. Milk yields and SCC scores were obtained from all cows on days 0, 28, 65, 92, and 126 after initiation of feeding of proteinates. Foot health (mobility) was estimated by evaluation of lameness or tenderness while cows were walking on concrete on days -7, 5, 72, 104, and 137. Treatment effects were determined to be significant if curves for response variables differed significantly between treatments. Mathematical models included treatment, cow(treatment), stage of lactation (0 to 99 d, 100 to 199 d, 200 to 299 d, 300 to 399 d) (cubic), days on experiment (quadratic or cubic) and milk yield (cubic) for SCC or SCC (cubic) for milk yield or the same factors plus treatment X days on experiment. SCC were very low in this herd. However, SCC increased during the summer by >100,000/ml in control cows whereas cows fed mineral proteinates maintained relatively stable SCC (curves differed, P<.01). Milk yields probably were not affected by treatment. Curves from some models favored treatment by the end of the experiment whereas other models did not. There were no differences detected in mobility (as an indicator of lameness and foot health) scores. ABSTRACT from *J. Dairy Sci.* 77(Suppl. 1): 156.

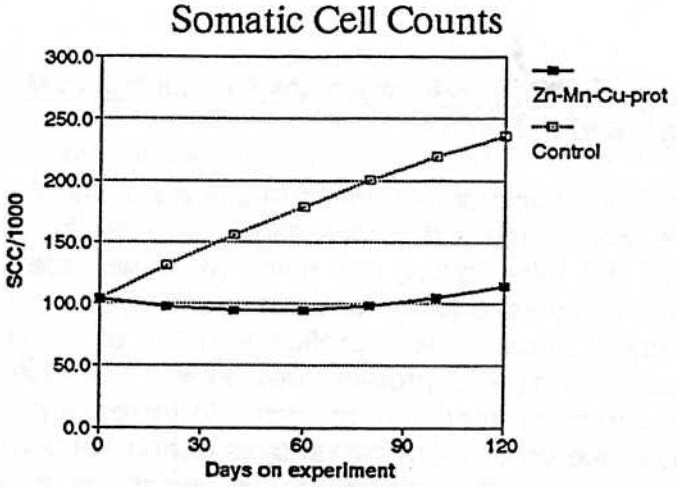

Figure 3. Effect of Zn-, Mn-, and Cu-proteinate on milk somatic cell counts; trend differences significant P<.001. The curves were generated from repeated measures of milk somatic cell counts over time on experiment (days 0, 28, 65, 92, and 126) to show time trend differences by use of heterogeneity of regression analyses. The mathematical model partitioned variation associated with cow within each treatment and stage of lactation from treatment (control versus proteinate) effects associated with time on experiment.

Example 4: Effect of supplementation with whole cottonseed on milk yield from grazing cows in Colombia, South America. C. R. Staples.

Five commercial dairy farms in Colombia, South America participated in an experiment from November 1, 1998 to April 14, 1999. Holstein cows (16 to 36 per farm; total = 129) were housed on pasture, grazing kikuyu grass as their only dietary forage. Two grain supplements, one containing whole cottonseeds, were compared on each farm. Milk production was measured weekly for each cow but averaged to give an overall average production per cow in the data set for statistical analysis using PROC GLM of SAS. The treatments were administered continuously throughout the trial. The model included treatment, farm, and parity and their interactions as independent variables. Milk production tended to be greater with whole cottonseed supplementation versus the control supplement (43.1 versus 40.3 lb/day) but those differences could not be attributed with statistical certainty (i.e., $P < .10$) to be due to treatment. However, solids-not-fat concentration in milk from cows supplemented with whole cottonseed was greater (8.10 versus 8.01%, $P = .05$).

Summary

An array of experimental designs exists that are available to meet the needs of on-farm trials. Knowledge of design and analysis of experiments, as well as the basic biology of the problem, is required because of the high cost of research and unique problems of large animals. Ready availability of computers has been particularly beneficial. Selection of a design, measurements and frequency of measurements to be made, number of animals to assign to the experiment, and length of the experiment generally involve a series of compromises. The researcher may sacrifice efficiency of comparisons in which there is minimal interest but must maximize the efficiency of the experiment for items of major importance. Knowledge of basic assumptions of analysis of variance, perhaps particularly in regard to error variance, is needed.

SUGGESTED REFERENCES

Cochran, W. G., and G. M. Cox. 1957. Experimental Designs. 2nd ed. John Wiley and Sons, New York, NY.

Damron, R. A., and W. R. Harvey. 1987. Experimental Design, ANOVA and Regression. Harper and Row, Publ. New York, NY.

Eaton, H. D., G. G. Gosslee, and H. L. Lucas. 1959. Effect of duration of experiment on experimental errors in calf nutrition growth studies. J. Dairy Sci. 42:1398.

Fleiss, J. L. 1973. Statistical Methods for Rates and Proportions. John Wiley & Sons. New York, NY.

Gill, J. L., and H. D. Hafs. 1971. Analysis of repeated measures in animals. J. Anim. Sci. 33:331.

Goodrich, R. D., B. P. Bradley, and A. D. Tallman. 1968. Importance of initial blood and plasma values. J. Anim. Sci. 27:247.

Henderson, C. R. 1969. Design and analysis of animal husbandry experiments. Monograph, Am. Soc. Anim. Prod. 55p.

Littell, R. C., P. R. Henry, and C. B. Ammerman. 1998. Statistical analysis of repeated measures data using SAS procedures. J. Anim. Sci. 76:1216.

Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfenger. 1996. SAS System for Mixed Models. SAS Inst. Inc., Cary, NC.

Littell, R. C., C. J. Wilcox, H. H. Van Horn, and A. P. Tomlinson. 1995. Estimation of and adjustment for residual effects in dairy feeding experiments using changeover designs. Proc. Kansas State Univ. Conf. Appl. Stat. Agric. Manhattan. Apr 23-25. 12p.

Roman-Ponce, H., H. H. Van Horn, S. P. Marshall, C. J. Wilcox, and P. F. Randel. 1975. Complete rations for dairy cattle. V. Interaction of sugarcane bagasse quantity and form with soybean meal, urea, and Starea. J. Dairy Sci. 58:1320.

Silva, H. M., C. J. Wilcox, W. W. Thatcher, R. B. Becker, and D. Morse. 1992. Factors affecting days open, gestation length, and calving interval in Florida dairy herds. J. Dairy Sci. 75:288.

Simerl, N. A., C. J. Wilcox, W. W. Thatcher, and F. G. Martin. 1991. Prepartum and peripartum reproductive performance of heifers freshening at young ages. J. Dairy Sci. 74:1724.

Snedecor, G. W., and W. G. Cochran. 1980. Statistical Methods. 7th ed. Iowa State Univ. Press. Ames.

St. Pierre, N. R., and L. R. Jones. 1999. Interpretation and design of nonregulatory on-farm feeding trials. J. Anim. Sci. 77(Suppl 2):177 and J. Dairy Sci 82(Suppl. 2):177.

Wilcox, C. J., and W. A. Krienke. 1964. Variability and interrelationships of composition and yield of daily milk samples. J. Dairy Sci. 47:638.

Wilcox, C. J., W. W. Thatcher, and F. G. Martin. 1990. Statistical analysis of repeated measurements in physiology experiments. Livestock Reproduction in Latin America. IAEC, Vienna. 15p.

Table 1. Coefficients of variation for production responses in an experiment with an incomplete Latin square (a changeover) design[1]

Milk yield	3	Fat %	12
4 % FCM	4	TS %	4
SCM yield	4	Feed intake	3
Protein yield	4	Body weight	3

[1] Roman-Ponce, et al., J. Dairy Sci. 58:1320. 1975.
Values are percentages

Table 2. Coefficients of variation (CV) for blood and plasma constituents of sheep[1]

Response variable	CV	Response variable	CV
Hemoglobin[2]	10	Plasma calcium	21
Hematocrit	11	Plasma phosphorus	14
Erthrocytes	11	Plasma molybdenum	60
Plasma copper	21	Plasma Vitamin A	19
Plasma protein	9		

[1] Goodrich, et al., J. Anim. Sci. 27:247. 1968.
See reference for units of measure.

Table 3. Animals required in each of two treatment groups to detect differences of various magnitudes[1]

Difference between means	Coefficient of variation					
	5	8	10	12	16	20
5	17	41	>50	---	---	---
10	5	11	17	24	41	>50
15	3	6	8	11	19	29
20	3	4	5	7	11	17
25	2	3	4	5	7	11
30	2	3	3	4	6	8

[1] Cochran and Cox (1962); Type I protection, 95%; Type II protection, 80%; two-tailed t test

Table 4. Daily variation in milk yield and composition during 6-day period

	Jersey		Holstein	
Response	C.V.	s.d.	C.V.	s.d.
SNF %	1	.138	2	.128
Fat %	10	.567	11	.448
pH	1	.042	1	.046
Acidity	6	.008	7	.009
Protein %	5	.164	4	.127
Chlorine %	5	.006	7	.031
Milk yield (lb)	8	1.913	8	3.040

[1] Wilcox and Krienke. J. Dairy Sci. 47:638. 1964

[2] C.V. = coefficient of variation; s.d. = error standard diviation.

Table 5. Means and coefficients of variation (CV) for dairy cattle reproductive traits[1]

Trait	Mean	CV
Age at first parturition (months)[1]	29	16
Days parturition to first service[1]	92	38
Days first service to conception[1]	35	154
Days open[1]	123	43
Gestation length[1]	280	2
Calving interval[1]	400	13
Age at first service[2]	441	10
Number of services per conception[2]	2.38	81

[1] Da Silva (1977); based on 7,000 to 15,000 observations for each trait of Jersey, Holstein and Guernsey cows in Florida.

[2] Simerl et. al. (1991); based on 1071 or 1144 observations of heifers at Univ. FL Dairy Research Unit.

Table 6. Animals required in each of two treatment groups to detect differences of various magnitudes, binary data[1]

Frequencies %		Animals	Frequencies %		Animals
Group 1	Group 2	per group	Group 1	Group 2	per group
10	15	764	40	45	1612
10	20	237	40	50	426
			40	60	116
20	25	1172			
20	30	332	50	55	1644
30	35	1455			
30	40	395	60	65	1549
30	45	153	60	70	395

[1] Based on Fleiss (1973); Type 1 protection 95%; Type 2 protection, 80%; two-tailed t test.

Table 7. Coefficients of variation (CV) for calf growth traits[1]

Variable	Length of experiment	
	2 wk	12 wk
Body weight	32	11
Wither height	33	11
Heart girth	70	11
Paunch girth	51	10

[1] Based on Eaton et al. (1959). Average of several estimates.

Table 8. Effect of length of experiment on number of calves needed to detect a 20% difference between two treatment means at P < .05, 80% assurance level.[1]

	Days on experiment			
Variable	42	56	70	84
Body weight	25	20	14	10
Wither height	62	33	23	17
Heart girth	38	31	20	14
Paunch girth	18	13	11	9

[1] Based on Eaton et al. (1959).

Table 9. Error terms for three models in ANOVA for a 2-factor factorial design.

Source of	Models		
variation	Fixed[1]	Random[2]	Mixed[3]
A	E	AB	AB
B	E	AB	E
AB	E	E	E
E	--	--	--

[1] A and B are fixed effects
[2] A and B are random effects
[3] A is fixed and B is random

Table 10. Experimentwise error rates for several multiple range tests

Number of means compared	LSD[1]	DMR[2]	NK/T[3]
2	.05	.05	.05
3	.12	.10	.05
4	.20	.14	.05
6	.40	.23	.05
10	.59	.37	.05
20	.86	.62	.05

[1] Least significant difference test
[2] Duncan multiple range test
[3] Newman-Keuls or Tukey test