Unbalanced Analysis of Variance, Design, and Regression
Ronald Christensen, fletcher@stat.unm.edu
Copyright 2013

Chapter 17

Basic experimental designs

In this chapter we examine basic experimental designs: completely randomized designs (CRDs), randomized complete block (RCB) designs, Latin square (LS) designs, balanced incomplete block (BIB) designs, and more.

17.1 Experiments and Causation

The basic object of running an experiment is to determine causation. Determining causation is difficult. We regularly collect data and find relationships between “dependent” variables and predictor variables. But this does not imply causation. One can predict air pressure extremely well from the boiling point of water, but does the boiling point of water *cause* the air pressure? Isn’t it the other way around? We found that females scored lower in a particular statistics class than males, but does being female *cause* that result? Doesn’t it seem plausible that something that is correlated with sexes might cause the result? Interest in statistics? Time devoted to studying statistics? Understanding the instructor’s teaching style? Being Native American in Albuquerque in 1978 was highly associated with lower suicide ages. But to claim that being Native American *caused* lower suicide ages would be incredibly simplistic. Causation is fundamentally tied to the idea that if you change one thing (the cause), you will change something else (the result). If that is true, can sex or race ever cause anything, since we cannot really change them?

In constructing an experiment we *randomly assign treatments to experimental units*. For example, we can randomly assign (many kinds of) drugs to people. We can randomly assign which employee will operate a particular machine or use a particular process. Unfortunately, there are many things we cannot perform experiments on. We cannot randomly assign sexes or races to people. As a practical matter, we cannot assign the use of illegal drugs to people.

The key point in determining causation is randomization. If we have a collection of experimental units and randomly assign the treatments to them, then (on average) there can be no systematic differences between the treatment groups other than the treatments. Therefore, any differences we see among the treatment groups must be caused by the treatments.

Alas, there are still problems. The randomization argument works on average. Experimental units, whether they be people, rats, or plots of ground, are subject to variability. One can get unlucky with a particular assignment of treatments to experimental units. If by chance one treatment happens to get far more of the “bad” experimental units it will look like a bad treatment. For example, if you want to know whether providing milk to elementary school students improves their performances, you cannot let the teachers decide who gets the milk. The teachers may give it to the poorest students in which case providing milk could easily look like it harms student performances. Similar things can happen by chance when randomly assigning treatments. To infer causation, the experiment should be repeated often enough that chance becomes a completely implausible explanation for the results.

Moreover, if you measure a huge number of items on each experimental unit, there is a good chance that one of the treatment groups will randomly have an inordinate number of good units for some variable, and hence show an effect that is really due to chance. In other words, if we measure enough variables, just by chance, some of them will display a relationship to the treatment groups, regardless of how the treatment groups were chosen.

A particularly disturbing problem is that the experimental treatments are often not what we think they are. *An experimental treatment is everything we do differently to a group of experimental units.* If we give a drug to a bunch of rats and then stick them into an asbestos filled attic, the fact that those rats have unusually high cancer rates does not mean that the drug caused it. The treatment caused it, but just because we call the treatment by the name of the drug does not make the drug the treatment.

Alternatively, suppose you want to test whether artificial sweeteners made with a new chemical cause cancer. You get some rats, randomly divide them into a treatment group and a control. You inject the treatment rats with a solution of the sweetener combined with another (supposedly benign) chemical. You leave the control rats alone. For simplicity you keep the treatment rats in one cage and the control rats in another cage. Eventually, you find an increased risk of cancer among the treatment rats as compared to the control rats. You can reasonably conclude that the treatments caused the increased cancer rate. Unfortunately, you do not really know whether the sweetener or the supposedly benign chemical or the combination of the two caused the cancer. In fact, you do not really know that it was the chemicals that caused the cancer. Perhaps the process of injecting the rats caused the cancer or perhaps something about the environment in the treatment rats's cage caused the cancer. A treatment consists of *all the ways* in which a group is treated differently from other groups. It is crucially important to *treat all experimental units as similarly as possible so that (as nearly as possible) the only differences between the units are the agents that were meant to be investigated.*

Random assignment of treatments it fundamental to conducting an experiment but it does not mean haphazard assignment of treatments to experimental units. Haphazard assignment is subject to the (unconscious) biases of the person making the assignments. Random assignment uses a reliable table of random numbers or a reliable computer program to generate random numbers. It then uses these numbers to assign treatments. For example, suppose we have four experimental units labeled u_1, u_2, u_3, and u_4 and four treatments labeled A, B, C, and D. Given a program or table that provides random numbers between 0 and 1 (i.e., random samples from a uniform(0,1) distribution), we associate numbers between 0 and .25 with treatment A, numbers between .25 and .50 with treatment B, numbers between .50 and .75 with treatment C, and numbers between .75 and 1 with treatment D. The first random number selected determines the treatment for u_1. If the first number is .6321, treatment C is assigned to u_1 because $.50 < .6321 < .75$. If the second random number is .4279, u_2 gets treatment B because $.25 < .4279 < .50$. If the third random number is .2714, u_3 would get treatment B, but we have already assigned treatment B to u_2, so we throw out the third number. If the fourth number is .9153, u_3 is assigned treatment D. Only one unit and one treatment are left, so u_4 gets treatment A. Any reasonable rule (decided ahead of time) can be used to make the assignment if a random number hits a boundary, e.g., if a random number comes up, say, .2500.

By definition, treatments must be amenable to change. As discussed earlier, things like sex and race are not capable of change, but in addition many viable treatments cannot be randomly assigned for social reasons. If you want to know if smoking causes cancer in humans, running an experiment is difficult. In our society we cannot force some people to smoke a specific amount for a long period of time and force others not to smoke at all. Nonetheless, we are very interested in whether smoking causes cancer. What are we to do?

When experiments cannot be run, the other common method for inferring causation is the "What else could it be?" approach. For smoking, the idea is that you measure *everything else* that could possibly be causing cancer and *appropriately adjust* for those measurements. If, after fitting all of those variables, smoking still has a significant effect on predicting cancer, then smoking must be causing the cancer. The catch is that this is extremely difficult to do. How do you even identify, much less measure, everything else that could be causing cancer? And even if you do measure everything, how do you know that you have adjusted for those variables appropriately. The key to this argument is independent replication of the studies! If there are many such *observational studies* with many different ideas of what other variables could be causing the effect (cancer) and many ways of adjusting for those variables, and if the studies consistently show that smoking remains an important predictor, at some point it would seem foolish to ignore the possibility that smoking causes cancer.

I have long contended that one cannot infer causation from data analysis. Certainly data analysis speaks to the relative validity of competing causal models but that is a far cry from actually determining causation. I believe that causation must be determined by some external argument. I find randomization to be the most compelling external argument. In "What else can it be?" the external argument is that all other variables of importance have been measured and appropriately considered.

My contention that data analysis cannot lead to causation may be wrong. I have not devoted my life to studying causal models. And I know that people study causation by the consideration of counterfactuals. But for now, I stand by my contention.

Although predictive ability does not imply causation, for many (perhaps most) purposes, predictive ability is more important. Do you really care why the lights go on when you flip a switch? Or do you care that your prediction comes true? You probably only care about causation when the lights stop working. How many people really understand the workings of an automobile? How many can successfully predict how automobiles will behave?

17.2 Technical Design Considerations

As a technical matter, the first object in designing an experiment is to construct one that allows for a valid estimate of σ^2, the variance of the observations. Without a valid estimate of error, we cannot know whether the treatment groups are exhibiting any real differences. Obtaining a valid *estimate of error* requires appropriate replication of the experiment. Having one observation on each treatment is not sufficient. All of the basic designs considered in this chapter allow for a valid estimate of the variance.

The simplest experimental design is the *completely randomized design*. With four drug treatments and observations on eight animals, a valid estimate of the error can be obtained by randomly assigning each of the drugs to two animals. *If the treatments are assigned completely at random to the experimental units* (animals), *the design is a completely randomized design*. The fact that there are more animals than treatments provides our replication.

It is not crucial that the design be balanced, i.e., it is not crucial that we have the same number of replications on each treatment. But it is useful to have more than one observation on each unit to help check our assumption of equal variances.

A second important consideration is to construct a design that yields a small variance. A smaller variance leads to sharper statistical inferences, i.e., narrower confidence intervals and more powerful tests. The basic idea is to examine the treatments on homogeneous experimental material. The people of Bergen, Norway are probably more homogenous than the people of New York City. It will be easier to find treatment effects when looking at people from Bergen. Of course the downside is that you end up with results that apply to the people of Bergen. The results may or may not apply to the people of New York City.

A fundamental tool for reducing variability is *blocking*. The people of New York City may be more variable than the people of Bergen but we might be able to divide New Yorkers into subgroups that are just as homogeneous as the people of Bergen. With our drugs and animals illustration, a smaller variance for treatment comparisons is generally obtained when the eight animals consist of two litters of four siblings and each treatment is applied to one randomly selected animal from each litter. With each treatment applied in every litter, all comparisons among treatments can be performed *within* each litter. Having at least two litters is necessary to get a valid estimate of the variance of the comparisons. *Randomized complete block designs (RCBs) : 1) identify blocks of homogeneous experimental material (units) and 2) randomly assign each treatment to an experimental unit within each block*. The blocks are complete in the sense that each block contains all of the treatments.

The key point in blocking on litters is that, if we randomly assigned treatments to experimental units without consideration of the litters, our measurements on the treatments would be subject to all of the litter to litter variability. By blocking on litters, we can eliminate the litter to litter variability so that our comparisons of treatments are subject only to the variability within litters (which, presumably, is smaller). Blocking has completely changed the nature of the variability in our observations.

The focus of block designs is in isolating groups of experimental units that are homogeneous: litters, identical twins, plots of ground that are close to one another. If we have three treatments and four animals to a litter, we can simply not use one animal. If we have five treatments and four animals to a litter, a randomized complete block experiment becomes impossible.

A *Balanced Incomplete Block (BIB)* design is one in which every pair of treatments occur together in a block the same number of times. For example, if our experimental material consists of identical twins and we have the drugs A, B, and C, we might give the first set of twins drugs A and B, the second set B and C, and the third set C and A. Here every pair of treatments occurs together in one of the three blocks.

BIBs do not provide balanced data in our usual sense of the word "balanced" but they do have a relatively simple analysis. RCBs are balanced in the usual sense. Unfortunately, losing any observations from either design destroys the balance that they display. Our focus is in analyzing unbalanced data, so we use techniques for analyzing block designs that do not depend on any form of balance.

The important ideas here are replication and blocking. RCBs and BIBs make very efficient designs but keeping their balance is not crucial. In olden days, before good computing, the simplicity of their analyses was important. But simplicity of analysis was never more than a side effect of the good experimental designs.

Latin squares use two forms of blocking at once. For example, if we suspect that birth order within the litter might also have an important effect on our results, we continue to take observations on each treatment within every litter, but we also want to have each treatment observed in every birth order. This is accomplished by having four litters with treatments arranged in a Latin square design. Here we are simultaneously blocking on litter and birth order.

Another method for reducing variability is incorporating covariates into the analysis. This topic is discussed in Section 8.

Ideas of blocking can also be useful in observational studies. While one cannot really create blocks in observational studies, one can adjust for important groupings.

EXAMPLE 17.2.1. If we wish to run an experiment on whether cocaine users are more paranoid than other people, we may decide that it is important to block on socioeconomic status. This is appropriate if the underlying level of paranoia in the population differs by socioeconomic status. Conducting an experiment in this setting is difficult. Given groups of people of various socioeconomic statuses, it is a rare researcher who has the luxury of deciding which subjects will ingest cocaine and which will not. □

The seminal work on experimental design was written by Fisher (1935). It is still well worth reading. My favorite source on the ideas of experimentation is Cox (1958). The books by Cochran and Cox (1957) and Kempthorne (1952) are classics. Cochran and Cox is more applied. Kempthorne is more theoretical. Kempthorne has been supplanted by Hinkleman and Kempthorne (2008, 2005). There is a huge literature in both journal articles and books on the general subject of designing experiments. The article by Coleman and Montgomery (1993) is interesting in that it tries to formalize many aspects of planning experiments that are often poorly specified. Two other useful books are Cox and Reid (2000) and Casella (2008).

17.3 Completely randomized designs

In a completely randomized design, a group of experimental units are available and the experimenter randomly assigns treatments to the experimental units. The data consist of a group of observations on each treatment. Typically, these groups of observations are subjected to a one-way analysis of variance.

EXAMPLE 17.3.1. In Example 12.4.1, we considered data from Mandel (1972) on the elasticity measurements of natural rubber made by 7 laboratories. While Mandel did not discuss how the data were obtained, it could well have been the result of a completely randomized design. For a CRD, we would need 28 pieces of the type of rubber involved. These should be randomly divided into 7 groups (using a table of random numbers or random numbers generated by a reliable computer program). The first group of samples is then sent to the first lab, the second group to the second lab, etc. For a CRD, it is important that a sample is not sent to a lab because the sample somehow seems appropriate for that particular lab.

Personally, I would also be inclined to send the four samples to a given lab at different times. If the four samples are sent at the same time, they might be analyzed by the same person, on the same machines, at the same time. Samples sent at different times might be treated differently. If samples are treated differently at different times, this additional source of variation should be included in any predictive conclusions we wish to make about the labs.

When samples sent at different times are treated differently, sending a batch of four samples at the same time constitutes *subsampling*. There are two sources of variation to deal with: variation from time to time and variation within a given time. The values from four samples at a given time collectively help reduce the effect on treatment comparisons due to variability at a given time, but samples analyzed at different times are still *required* if we are to obtain a valid estimate of the error. In fact, with subsampling, a perfectly valid analysis can be based on the means of the four subsamples. In our example, such an analysis gives only one 'observation' at each time, so the need for sending samples at more than one time is obvious. If the four samples were sent at the same time, there would be no replication, hence no estimate of error. Subsection 19.4.1 and Christensen (2011, Section 9.4) discuss subsampling in more detail. □

EXAMPLE 17.3.2. In Chapter 12, we considered the suicide ages. A designed experiment would require that we take a group of people who we know will commit suicide and randomly assign one of the ethnic groups to the people. Obviously a difficult task. □

17.4 Randomized complete block designs

In a randomized complete block design the experimenter obtains (constructs) blocks of homogeneous material that contain as many experimental units as there are treatments. The experimenter then randomly assigns a different treatment to each of the units in the block. The random assignments are performed independently for each block. The advantage of this procedure is that treatment comparisons are subject only to the variability within the blocks. Block to block variation is eliminated in the analysis. In a completely randomized design applied to the same experimental material, the treatment comparisons would be subject to both the within block and the between block variability.

The key to a good blocking design is in obtaining blocks that have little within block variability. Often this requires that the blocks be relatively small. A difficulty with RCB designs is that the blocks must be large enough to allow all the treatments to be applied within each block. This can be a serious problem if there is a substantial number of treatments or if maintaining homogeneity within blocks requires the blocks to be very small. If the treatments cannot all be fitted into each block, we need some sort of *incomplete block* design.

Table 17.1: *Spectrometer data*

	Block		
Treatment	1	2	3
New-clean	0.9331	0.8664	0.8711
New-soiled	0.9214	0.8729	0.8627
Used-clean	0.8472	0.7948	0.7810
Used-soiled	0.8417	0.8035	

Table 17.2: *Analysis of variance for spectrometer data*

Source	df	SS	MS	F	P
Block	2	0.0063366	0.0031683	62.91	0.000
Treatments	3	0.0166713	0.0055571	110.34	0.000
Error	5	0.0002518	0.0000504		
Total	10	0.0232598			

The typical analysis of a randomized complete block design is a two-way ANOVA without replication or interaction. Except for the experimental design considerations, the analysis is like that of the Hopper Data from Example 15.3.1. A similar analysis is illustrated below. As with the Hopper data, block by treatment interaction is properly considered to be error. *If the treatment effects are not large enough to be detected above any interaction, then they are not large enough to be interesting.*

EXAMPLE 17.4.1. Inman, Ledolter, Lenth, and Niemi (1992) studied the performance of an optical emission spectrometer. Table 17.1 gives some of their data on the percentage of manganese (Mn) in a sample. The data were collected using a sharp counterelectrode tip with the sample to be analyzed partially covered by a boron nitride disk. Data were collected under three temperature conditions. Upon fixing a temperature, the sample percentage of Mn was measured using 1) a new boron nitride disk with light passing through a clean window (new-clean), 2) a new boron nitride disk with light passing through a soiled window (new-soiled), 3) a used boron nitride disk with light passing through a clean window (used-clean), and 4) a used boron nitride disk with light passing through a soiled window (used-soiled). The four conditions, new-clean, new-soiled, used-clean, used-soiled are the treatments. The temperature was then changed and data were again collected for each of the four treatments. A block is always made up of experimental units that are homogeneous. The temperature conditions were held constant while observations were taken on the four treatments so the temperature levels identify blocks. Presumably, the treatments were considered in random order. Christensen (1996) analyzed these data including the data point for Block 3 and Used-soiled.

The two-factor additive effects model for these data is

$$y_{ij} = \mu + \beta_i + \eta_j + \varepsilon_{ij},$$

$i = i, 2, 3,\ j = 1, 2, 3, 4$, however the $i = 3,\ j = 4$ observation is missing. Here β_i denotes a block effect and η_j a treatment effect. As usual, we assume the errors are independent and $N(0, \sigma^2)$.

Unlike the analysis in Chapter 14, *in blocking experiments we always examine the treatments after the blocks.* We constructed the blocks, so we know they should have effects. The only relevant ANOVA table is given as Table 17.2.

For now, we just perform all pairwise comparisons of the treatments.

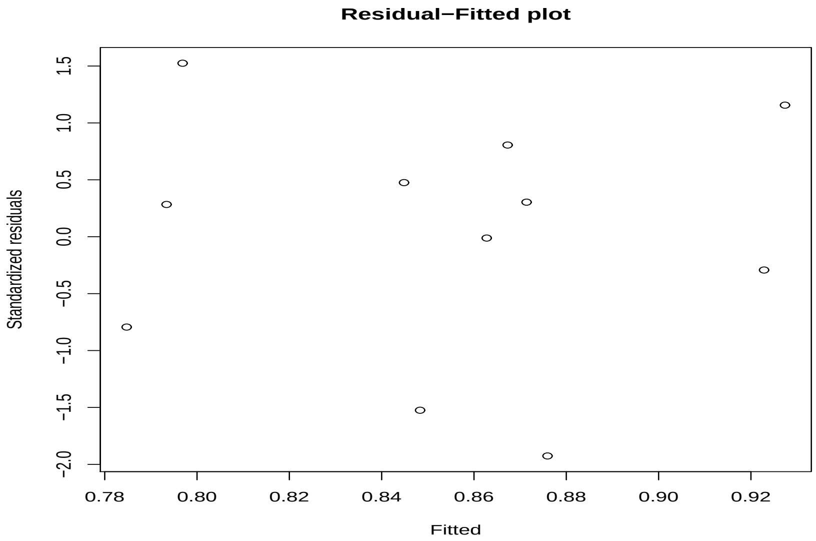

Figure 17.1: *Plot of residuals versus predicted values, spectrometer data.*

Parameter	Est	SE(Est)	t	Bonferroni P
$\eta_2 - \eta_1$	-0.00453	0.005794	-0.78	1.0000
$\eta_3 - \eta_1$	-0.08253	0.005794	-14.24	0.0002
$\eta_4 - \eta_1$	-0.07906	0.006691	-11.82	0.0005
$\eta_3 - \eta_2$	-0.07800	0.005794	-13.46	0.0002
$\eta_4 - \eta_2$	-0.07452	0.006691	-11.14	0.0006
$\eta_4 - \eta_3$	0.003478	0.006691	0.5198	1.000

The one missing observation is from treatment 4 so the standard errors that involve treatment 4 are larger. Although we have different standard errors, the results can be summarized as follows.

$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_4$	$\hat{\eta}_3$
0	-0.00453	-0.07906	-0.08253

The new disk treatments are significantly different from the used disk treatments but the new disk treatments are not significantly different from each other nor are the used disk treatments significantly different from each other. The structure of the treatments suggests an approach to analyzing the data that will be exploited in the next chapter. Here we used a side condition of $\eta_1 = 0$ because it made the estimates readily agree with the table of pairwise comparisons.

Table 17.2 contains an F test for blocks. In a true blocking experiment, there is not much interest in testing whether block means are different. After all, one *chooses* the blocks so that they have different means. Nonetheless, the F statistic $MSBlks/MSE$ is of some interest because it indicates how effective the blocking was, i.e., it indicates how much the variability was reduced by blocking. For this example, $MSBlks$ is 63 times larger than MSE, indicating that blocking was definitely worthwhile. In our model for block designs, there is no reason not to test for blocks, but some models used for block designs do not allow a test for blocks.

Residual plots for the data are given in Figures 17.1 through 17.4. Figure 17.1 is a plot of the residuals versus the predicted values. Figure 17.2 plots the residuals versus indicators of the treatments. While the plot looks something like a bow tie, I am not overly concerned. Figure 17.3 contains a plot of the residuals versus indicators of blocks. The residuals look pretty good. From Figure 17.4, the residuals look reasonably normal. In the normal plot there are 11 residuals but the

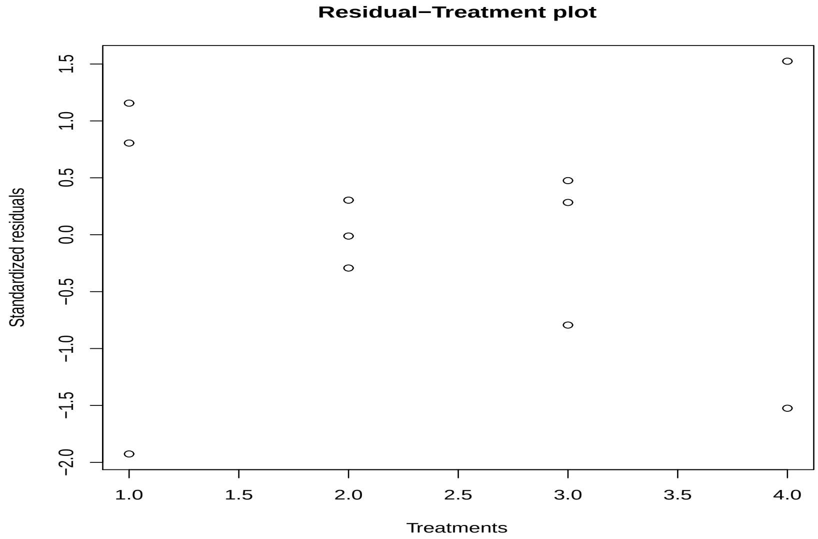

Figure 17.2: *Plot of residuals versus treatment groups, spectrometer data.*

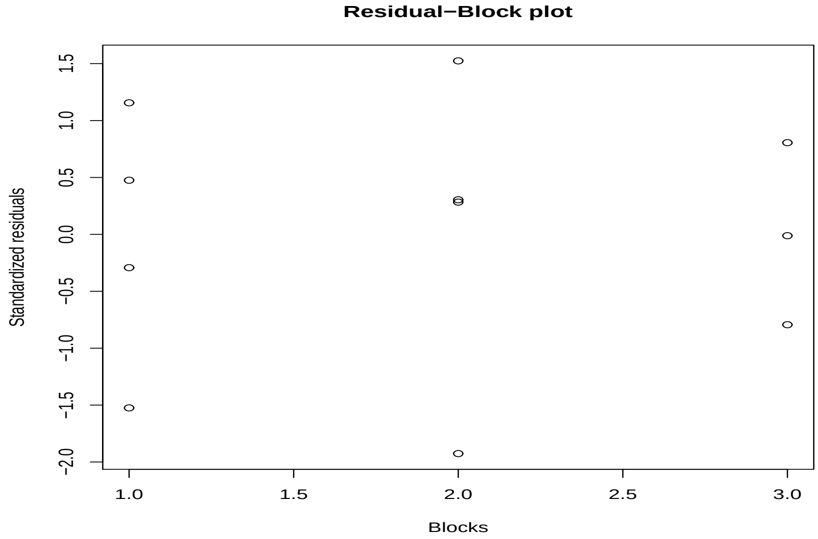

Figure 17.3: *Plot of residuals versus blocks, spectrometer data.*

analysis has only 5 degrees of freedom for error. If you want to do a W' test for normality, you might use a sample size of 11 and compare the value $W' = 0.966$ to $W'(\alpha, 11)$, but it may be appropriate to use the dfE as the sample size for the test and use $W'(\alpha, 5)$.

The leverages (not shown) are all reasonable. The largest t residual is -3.39 for Block 2, Treatment 1 which gives a Bonferonni adjusted P value of 0.088. □

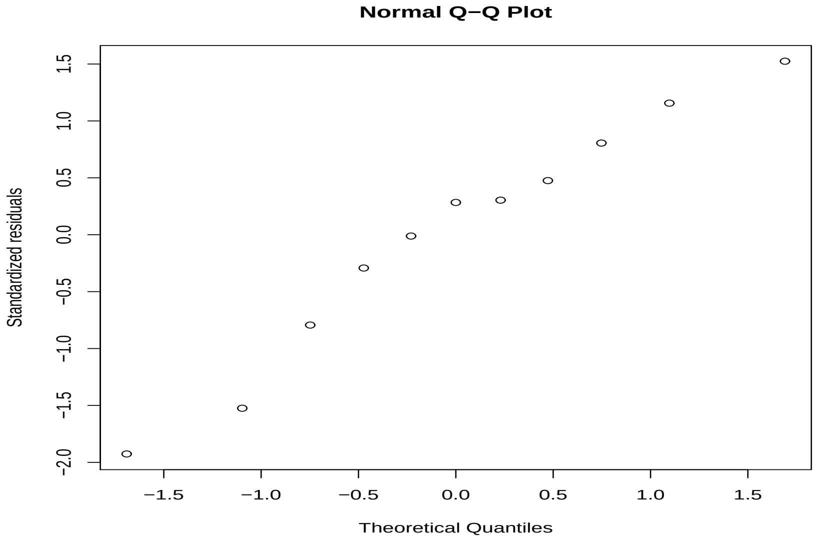

Figure 17.4: *Normal plot of residuals, spectrometer data,* $W' = 0.966$.

Minitab commands

The following Minitab commands generate the analysis of variance. Column c1 contains the spectrometer data, while column c2 contains integers 1 through 4 indicating the appropriate treatment, and c3 contains integers 1 through 3 that indicate the block. The predicted values are given by the 'fits' subcommand.

```
MTB > names c1 'y' c2 'Trts' c3 'Blks'
MTB > glm c1 = c3 c2;
SUBC>   Pairwise Treatment;
SUBC>     Bonferroni;
SUBC> sresid c10;
SUBC> fits  c11.
```

17.4.1 Paired comparisons

An interesting special case of complete block data is paired comparison data as discussed in Section 4.1. In paired comparison data, there are two treatments to contrast and each pair constitutes a complete block.

EXAMPLE 17.4.2. *Shewhart's hardness data.*
In Section 4.1, we examined Shewhart's data on hardness of two items that were welded together. In this case, it is impossible to group arbitrary formless pairs of parts and then randomly assign a part to be either part 1 or part 2, so the data do not actually come from an RCB experiment. Nonetheless, the two-way ANOVA model remains reasonable with pairs playing the role of blocks.

The data were given in Section 4.1 along with the means for each of the two parts. The two-way ANOVA analysis also requires the mean for each pair of parts. The analysis of variance table for the blocking analysis is given in Table 17.3. In comparing the blocking analysis to the paired comparison analysis given earlier, allowance for round-off errors must be made. The *MSE* is exactly half the value of $s_d^2 = 17.77165$ given in Section 4.1. The Table of Coefficients (from Minitab) gives

Table 17.3: *Analysis of variance for hardness data*

Source	df	SS	MS	F	P
Pairs(Blocks)	26	634.94	24.42	2.75	0.006
Parts(Trts)	1	2164.73	2164.73	243.62	0.000
Error	26	231.03	8.89		
Total	53	3030.71			

Table 17.4: *Mangold root data*

	Columns				
Rows	1	2	3	4	5
1	D(376)	E(371)	C(355)	B(356)	A(335)
2	B(316)	D(338)	E(336)	A(356)	C(332)
3	C(326)	A(326)	B(335)	D(343)	E(330)
4	E(317)	B(343)	A(330)	C(327)	D(336)
5	A(321)	C(332)	D(317)	E(318)	B(306)

a test for no Part effects of

$$t_{obs} = \frac{6.3315}{0.4057} = 15.61\,.$$

This is exactly the same t statistic as obtained in Section 4.1. The reference distribution is $t(26)$, again exactly the same. The analysis of variance F statistic is just the square of the t_{obs} and gives equivalent results for two-sided tests. Confidence intervals for the difference in means are also exactly the same in the blocking analysis and the paired comparison analysis. The one real difference between this analysis and the analysis of Section 4.1 is that this analysis provides an indication of whether pairing was worthwhile. □

17.5 Latin square designs

Latin square designs involve two simultaneous but distinct definitions of blocks. The treatments are arranged so that every treatment is observed in every block for both kinds of blocks.

EXAMPLE 17.5.1. Mercer and Hall (1911) and Fisher (1925, section 49) consider data on the weights of mangold roots. They used a Latin square design with 5 rows, columns, and treatments. The rectangular field on which the experiment was run was divided into five rows and five columns. This created 25 plots, arranged in a square, on which to apply the treatments A, B, C, D, and E. Each row of the square was viewed as a block, so every treatment was applied in every row. The unique feature of Latin square designs is that there is a second set of blocks. Every column was also considered a block, so every treatment was also applied in every column. The data are given in Table 17.4, arranged by rows and columns with the treatment given in the appropriate place and the observed root weight given in parentheses.

Table 17.5 contains the analysis of variance table including the analysis of variance F test for the null hypothesis that the effects are the same for every treatment. The F statistic $MSTrts/MSE$ is very small, 0.56, so there is no evidence that the treatments behave differently. Blocking on columns was not very effective as evidenced by the F statistic of 1.20, but blocking on rows was very effective, $F = 7.25$.

Many experimenters are less than thrilled when told that there is no evidence for their treatments having any differential effects. Inspection of the table of coefficients (not given) leads to an obvious conclusion that most of the treatment differences are due to the fact that treatment D has a much larger effect than the others, so we look at this a bit more.

We created a new factor variable called “Contrast” that has the same code for all of treatments

Table 17.5: *Analysis of variance for mangold root data*

Source	df	SS	MS	F	P
Columns	4	701.8	175.5	1.20	.360
Rows	4	4240.2	1060.1	7.25	.003
Trts	4	330.2	82.6	0.56	.696
Error	12	1754.3	146.2		
Total	24	7026.6			

Table 17.6: *Analysis of variance for mangold root data*

Source	df	SS	MS	F	P
Columns	4	4240.2	1060.1	8.89	0.001
Rows	4	701.8	175.5	1.47	0.260
Contrast	1	295.8	295.8	2.48	0.136
Error	15	1788.7	119.2		
Total	24	7026.6			

A, B, C, E but a different code for D. Fitting a model with Columns and Rows but Contrast in lieu of Treatments gives the ANOVA table in Table 17.6. The ANOVA table F statistic for Contrast is $295.8/119.2 = 2.48$ with a P value of 0.136. It provides a test of whether treatment D is different from the other treatments, when the other treatments are taken to have identical effects. Using our best practice, we would actually compute the F statistic with the MSE from Table 17.5 in the denominator giving $F_{obs} = 295.8/146.2 = 2.02$ which looks even less significant. This contrast was chosen by looking at the data so as to appear as significant as possible and yet it still has a large P value. Testing the two models against each other by using Tables 17.5 and 17.6 provides a test of whether there are any differences among treatments A, B, C, and E. The F statistic of 0.08 is so small that it would be suspiciously small if it had not been chosen, by looking at the data, to be small.

The standard residual plots were given in Christensen (1996). They look quite good.

If these data were unbalanced, i.e., if we lost some observations, it would be important to look at an ANOVA table that fits Treatments after both Columns and Rows. Fitted in the current order, the F test for Rows indicates that blocking on rows after blocking on Columns was worthwhile but the F test for Columns indicates that blocking on Columns alone would have been a waste of time. In an unbalanced experiment, if we cared enough, we might fit Columns after Rows to see whether blocking on Columns was a complete waste of time. Because the data are balanced, the two tests for Columns are the same and we can safely say from Table 17.5 that blocking on Columns was a waste of time.

□

Computing techniques

The following Minitab commands will give the sums of squares, means, and residuals necessary for the analysis. Here c1 is a column containing the mangold root yields, c2 has values from 1 to 5 indicating the row, c3 has values from 1 to 5 indicating the column, and c4 has values from 1 to 5 indicating the treatment.

```
MTB > names c1 'y' c2 'Rows' c3 'Cols' c4 'Trts'
MTB > glm c1 = c2 c3 c4;
```

Latin square models

The model for an $r \times r$ Latin square design is a three-way analysis of variance,

$$y_{ijk} = \mu + \kappa_i + \rho_j + \tau_k + \varepsilon_{ijk}, \quad \varepsilon_{ijk}\text{s independent } N(0, \sigma^2). \tag{17.5.1}$$

The parameter μ is viewed as a grand mean, κ_i is an effect for the ith column, ρ_j is an effect for the jth row, and τ_k is an effect for the kth treatment. The subscripting for this model is peculiar. All of the subscripts run from 1 to r but not freely. If you specify a row and a column, the design tells you the treatment. Thus, if you know j and i, the design tells you k. If you specify a row and a treatment, the design tells you the column, so j and k dictate i. In fact, if you know any two of the subscripts, the design tells you the third.

Discussion of Latin squares

The idea of simultaneously having two distinct sets of complete blocks is quite useful. For example, suppose you wish to compare the performance of four machines in producing something. Productivity is notorious for depending on the day of the week, with Mondays and Fridays often having low productivity; thus we may wish to block on days. The productivity of the machine is also likely to depend on who is operating the machine, so we may wish to block on operators. Thus we may decide to run the experiment on Monday through Thursday with four machine operators and using each operator on a different machine each day. One possible design is

	Operator			
Day	1	2	3	4
Mon	A	B	C	D
Tue	B	C	D	A
Wed	C	D	A	B
Thu	D	A	B	C

where the numbers 1 through 4 are randomly assigned to the four people who will operate the machines and the letters A through D are randomly assigned to the machines to be examined. Moreover, the days of the week should actually be randomly assigned to the rows of the Latin square. In general, the rows, columns, and treatments should all be randomized in a Latin square.

Another distinct Latin square design for this situation is

	Operator			
Day	1	2	3	4
Mon	A	B	C	D
Tue	B	A	D	C
Wed	C	D	B	A
Thu	D	C	A	B

This square cannot be obtained from the first one by any interchange of rows, columns, and treatments. Typically, one would randomly choose a possible Latin square design from a list of such squares (see, for example, Cochran and Cox, 1957) in addition to randomly assigning the numbers, letters, and rows to the operators, machines, and days.

The use of Latin square designs can be extended in numerous ways. One modification is the incorporation of a third kind of block; such designs are called *Graeco-Latin squares*. The use of Graeco-Latin squares is explored in the exercises for this chapter. A problem with Latin squares is that small squares give poor variance estimates because they provide few degrees of freedom for error. For example, a 3×3 Latin square gives only 2 degrees of freedom for error. In such cases, the

Latin square experiment is often performed several times, giving additional replications that provide improved variance estimation. Section 18.6 presents an example in which several Latin squares are used.

17.6 Balanced incomplete block designs

Balanced incomplete block (BIB) designs are not balanced in the same way that balanced ANOVAs are balanced. Balanced incomplete block designs are balanced in the sense that *every pair of treatments occurs together in the same block some fixed number of times*, say, λ. In a BIB the analysis of blocks is conducted ignoring treatments and the analysis of treatments is conducted after adjusting for blocks. This is the only order of fitting models that we need to consider. Blocks are designed to have effects and these effects are of no intrinsic interest, so there is no reason to worry about fitting treatments first and then examining blocks after adjusting for treatments. Blocks are nothing more than an adjustment factor.

The analysis being discussed here is known as the *intrablock* analysis of a BIB; it is appropriate when the block effects are viewed as fixed effects. If the block effects are viewed as random effects with mean 0, there is an alternative analysis that is known as the recovery of *interblock* information. Cochran and Cox (1957) and Christensen (2011, Section 12.11) discuss this analysis; we will not.

EXAMPLE 17.6.1. A simple balanced incomplete block design is given below for four treatments A, B, C, D in four blocks of three units each.

Block	Treatments		
1	A	B	C
2	B	C	D
3	C	D	A
4	D	A	B

Note that every pair of treatments occurs together in the same block exactly $\lambda = 2$ times. Thus, for example, the pair A, B occurs in blocks 1 and 4. There are $b = 4$ blocks each containing $k = 3$ experimental units. There are $t = 4$ treatments and each treatment is observed $r = 3$ times. □

There are two relationships that must be satisfied by the numbers of blocks, b, units per block, k, treatments, t, replications per treatment, r, and λ. Recall that λ is the number of times two treatments occur together in a block. First, the total number of observations is the number of blocks times the number of units per block, but the total number of observations is also the number of treatments times the number of replications per treatment, thus

$$bk = rt.$$

The other key relationship in balanced incomplete block designs involves the number of comparisons that can be made between a given treatment and the other treatments *within the same block*. Again, there are two ways to count this. The number of comparisons is the number of other treatments, $t-1$, multiplied by the number of times each other treatment is in the same block as the given treatment, λ. Alternatively, the number of comparisons within blocks is the number of other treatments within each block, $k-1$, times the number of blocks in which the given treatment occurs, r. Thus we have

$$(t-1)\lambda = r(k-1).$$

In Example 17.6.1, these relationships reduce to

$$(4)3 = 3(4)$$

and

$$(4-1)2 = 3(3-1).$$

Table 17.7: *Balanced incomplete block design investigating detergents; data are numbers of dishes washed*

Block	Treatment, Observation		
1	A, 19	B, 17	C, 11
2	D, 6	E, 26	F, 23
3	G, 21	H, 19	J, 28
4	A, 20	D, 7	G, 20
5	B, 17	E, 26	H, 19
6	C, 15	F, 23	J, 31
7	A, 20	E, 26	J, 31
8	B, 16	F, 23	G, 21
9	C, 13	D, 7	H, 20
10	A, 20	F, 24	H, 19
11	B, 17	D, 6	J, 29
12	C, 14	E, 24	G, 21

Table 17.8: *Analysis of variance for a BIB*

Source	df	Seq *SS*	*MS*	*F*	*P*
Blocks	11	412.750	37.523	45.54	0.000
Trts	8	1086.815	135.852	164.85	0.000
Error	16	13.185	0.824		
Total	35	1512.750			

The nice thing about balanced incomplete block designs is that the theory behind them works out so simply that the computations can all be done on a hand calculator. I know, I did it once, see Christensen (2011, Section 9.4). But once was enough for this lifetime! We will rely on a computer program to provide the computations. We illustrate the techniques with an example.

EXAMPLE 17.6.2. John (1961) reported data on the number of dishes washed prior to losing the suds in the wash basin. Dishes were soiled in a standard way and washed one at a time. Three operators and three basins were available for the experiment, so at any one time only three treatments could be applied. Operators worked at the same speed, so no effect for operators was necessary nor should there be any effect due to basins. Nine detergent treatments were evaluated in a balanced incomplete block design. The treatments and numbers of dishes washed are given in Table 17.7. There were $b = 12$ blocks with $k = 3$ units in each block. Each of the $t = 9$ treatments was replicated $r = 4$ times. Each pair of treatments occurred together $\lambda = 1$ time. The three treatments assigned to a block were randomly assigned to basins as were the operators. The blocks were run in random order.

The analysis of variance is given in Table 17.8. The F test for treatment effects is clearly significant. We now need to examine contrasts in the treatments.

The treatments were constructed with a structure that leads to interesting effects. Treatments A, B, C, and D all consisted of detergent I using, respectively, 3, 2, 1, and 0 doses of an additive. Similarly, treatments E, F, G, and H used detergent II with 3, 2, 1, and 0 doses of the additive. Treatment J was a control. We return to this example for a more detailed analysis of the treatments in the next chapter.

As always, we need to evaluate our assumptions. The normal plot looks less than thrilling but is not too bad. The fifth percentile of W' for 36 observations is .940, whereas the observed value is .953. Alternatively, the residuals have only 16 degrees of freedom and $W'(.95, 16) = .886$. The data are counts, so a square root or log transformation might be appropriate, but we continue with the current analysis. A plot of standardized residuals versus predicted values looks good.

Table 17.9 contains diagnostic statistics for the example. Note that the leverages are all identical for the BIB design. Some of the standardized deleted residuals (ts) are near 2 but none are so large

Table 17.9: *Diagnostics for the detergent data*

Block	Trt.	y	$\hat{y}$	Leverage	r	t	C
1	A	19	18.7	0.56	0.49	0.48	0.01
1	B	17	16.1	0.56	1.41	1.46	0.12
1	C	11	12.1	0.56	−1.90	−2.09	0.22
2	D	6	6.6	0.56	−0.98	−0.98	0.06
2	E	26	25.4	0.56	1.04	1.04	0.07
2	F	23	23.0	0.56	−0.06	−0.06	0.00
3	G	21	20.5	0.56	0.86	0.85	0.05
3	H	19	18.6	0.56	0.67	0.66	0.03
3	J	28	28.9	0.56	−1.53	−1.60	0.15
4	A	20	19.6	0.56	0.61	0.60	0.02
4	D	7	6.4	0.56	0.98	0.98	0.06
4	G	20	21.0	0.56	−1.59	−1.68	0.16
5	B	17	17.3	0.56	−0.49	−0.48	0.01
5	E	26	25.4	0.56	0.98	0.98	0.06
5	F	19	19.3	0.56	−0.49	−0.48	0.01
6	C	15	14.3	0.56	1.16	1.18	0.08
6	F	23	24.1	0.56	−1.77	−1.92	0.20
6	J	31	30.6	0.56	0.61	0.60	0.02
7	A	20	20.6	0.56	−0.92	−0.91	0.05
7	E	26	26.1	0.56	−0.18	−0.18	0.00
7	J	31	30.3	0.56	1.10	1.11	0.08
8	B	16	16.8	0.56	−1.29	−1.31	0.10
8	F	23	22.6	0.56	0.73	0.72	0.03
8	G	21	20.7	0.56	0.55	0.54	0.02
9	C	13	13.6	0.56	−0.92	−0.91	0.05
9	D	7	6.9	0.56	0.18	0.18	0.00
9	H	20	19.6	0.56	0.73	0.72	0.03
10	A	20	20.1	0.56	−0.18	−0.18	0.00
10	F	24	23.3	0.56	1.10	1.11	0.08
10	H	19	19.6	0.56	−0.92	−0.91	0.05
11	B	17	16.8	0.56	0.37	0.36	0.01
11	D	6	6.1	0.56	−0.18	−0.18	0.00
11	J	29	29.1	0.56	−0.18	−0.18	0.00
12	C	14	13.0	0.56	1.65	1.76	0.17
12	E	24	25.1	0.56	−1.84	−2.00	0.21
12	G	21	20.9	0.56	0.18	0.18	0.00

as to indicate an outlier. The Cook's distances bring to one's attention exactly the same points as the standardized residuals and the ts. □

17.6.1 *Special cases*

Balanced lattice designs are BIBs with $t = k^2$, $r = k + 1$, and $b = k(k+1)$. Table 17.10 gives an example for $k = 3$. These designs can be viewed as $k + 1$ squares in which each treatment occurs once. Each row of a square is a block, each block contains k units, there are k rows in a square, so all of the $t = k^2$ treatments can appear in each square. To achieve a BIB, $k + 1$ squares are required, so there are $r = k + 1$ replications of each treatment. With $k + 1$ squares and k blocks (rows) per square, there are $b = k(k+1)$ blocks. The analysis follows the standard form for a BIB. In fact, the design in Example 17.6.2 is a balanced lattice with $k = 3$.

Youden squares are a generalization of BIBs that allows a second form of blocking and a very similar analysis. These designs are discussed in the next section.

Table 17.10: *Balanced lattice design for 9 treatments*

Block				Block			
1	A	B	C	7	A	H	F
2	D	E	F	8	D	B	I
3	G	H	I	9	G	E	C
4	A	D	G	10	A	E	I
5	B	E	H	11	G	B	F
6	C	F	I	12	D	H	C

Table 17.11: *Mangold root data*

	Columns			
Row	1	2	3	4
1	D(376)	E(371)	C(355)	B(356)
2	B(316)	D(338)	E(336)	A(356)
3	C(326)	A(326)	B(335)	D(343)
4	E(317)	B(343)	A(330)	C(327)
5	A(321)	C(332)	D(317)	E(318)

17.7 Youden squares

Consider the data on mangold roots in Table 17.11. There are five rows, four columns, and five treatments. If we ignore the columns, the rows and the treatments form a balanced incomplete block design, every pair of treatments occurs together three times. The key feature of Youden squares is that additionally the treatments are also set up in such a way that every treatment occurs once in each column. Since every row also occurs once in each column, the analysis for columns can be conducted independently of the analysis for rows and treatments. Columns are balanced relative to both treatments and rows.

Table 17.12 contains the analysis of variance for these data. Rows need to be fitted before Treatments. As long as balance is maintained, it does not matter where Columns are fitted. If the data become unbalanced, Treatments need to be fitted last. From the ANOVA table, there is no evidence for a difference between treatments.

Evaluation of assumptions is carried out as in all unbalanced ANOVAs. Diagnostic statistics are given in Table 17.13. The diagnostic statistics look reasonably good.

A normal plot looks very reasonable. A predicted value plot may indicate increasing variability as predicted values increase. One could attempt to find a transformation that would improve the plot but there is so little evidence of any difference between treatments that it hardly seems worth the bother.

The reader may note that the data in this section consist of the first four columns of the Latin square examined in Example 17.5.1. Dropping one column (or row) from a Latin square is a simple way to produce a Youden square. As Youden square designs do not give a square array of numbers (our example had 4 columns and 5 rows), one presumes that the name Youden *square* derives from

Table 17.12: *Analysis of variance*

Source	*df*	Seq *SS*	*MS*	*F*	*P*
Rows	4	4247.2	1061.8	6.87	
Column	3	367.0	122.3	0.79	
Trts	4	224.1	56.0	0.36	0.829
Error	8	1236.7	154.6		
Total	19	6075.0			

Table 17.13: *Diagnostics*

Row	Col	Trt	y	$\hat{y}$	Leverage	r	t	C
1	1	D	376	364.5	0.6	1.46	1.59	0.27
2	1	B	316	326.8	0.6	−1.37	−1.47	0.24
3	1	C	326	323.9	0.6	0.27	0.25	0.01
4	1	E	317	322.0	0.6	−0.64	−0.61	0.05
5	1	A	321	318.8	0.6	0.28	0.26	0.01
1	2	E	371	367.7	0.6	0.42	0.40	0.02
2	2	D	338	345.9	0.6	−1.01	−1.01	0.13
3	2	A	326	340.3	0.6	−1.81	−2.21	0.41
4	2	B	343	332.1	0.6	1.38	1.48	0.24
5	2	C	332	324.0	0.6	1.02	1.02	0.13
1	3	C	355	360.8	0.6	−0.74	−0.71	0.07
2	3	E	336	330.9	0.6	0.65	0.63	0.05
3	3	B	335	326.1	0.6	1.14	1.16	0.16
4	3	A	330	331.5	0.6	−0.19	−0.18	0.00
5	3	D	317	323.7	0.6	−0.86	−0.84	0.09
1	4	B	356	365.0	0.6	−1.14	−1.17	0.16
2	4	A	356	342.4	0.6	1.73	2.04	0.37
3	4	D	343	339.8	0.6	0.41	0.38	0.02
4	4	C	327	331.3	0.6	−0.55	−0.53	0.04
5	4	E	318	321.5	0.6	−0.44	−0.42	0.02

Table 17.14: *Mangold root data: column(observation)*

	Treatments				
Row	A	B	C	D	E
1		4(356)	3(355)	1(376)	2(371)
2	4(356)	1(316)		2(338)	3(336)
3	2(326)	3(335)	1(326)	4(343)	
4	3(330)	2(343)	4(327)		1(317)
5	1(321)		2(332)	3(317)	4(318)

this relationship to Latin squares. Table 17.14 presents an alternative method of presenting the data in Table 17.11 that is often used. □

Minitab commands

The Minitab commands for the mangold root analysis are given below.

```
MTB > names c1 'y' c2 'Rows' c3 'Cols' c4 'Trts'
MTB > glm c1 = c2 c3 c4;
SUBC> means c4;
SUBC> fits c11;
SUBC> sresids c12;
SUBC> tresids c13;
SUBC> hi c14;
SUBC> cookd c15.
```

Balanced lattice squares

The key idea in *balanced lattice square designs* is that if you look at every row as a block, the treatments form a balanced incomplete block design and simultaneously if every column is viewed as a block, the treatments again form a balanced incomplete block design. In other words, each pair of treatments occurs together in the same row *or* column the same number of times. Of course

Table 17.15: *Balanced lattice square design for 9 treatments*

	Column				Column		
Row	1	2	3	Row	4	5	6
1	A	B	C	4	A	F	H
2	D	E	F	5	I	B	D
3	G	H	I	6	E	G	C

every row appears with every column and vice versa. Balanced lattice square designs are similar to balanced lattices in that the number of treatments is $t = k^2$ and that the treatments are arranged in $k \times k$ squares. Table 17.15 gives an example for $k = 3$. If k is odd, one can typically get by with $(k+1)/2$ squares. If k is even, $k+1$ squares are generally needed.

17.8 Analysis of covariance in designed experiments

In Section 2 we discussed blocking as a method of variance reduction. Blocks where then incorporated as a factor variable into an additive effects model with blocks and treatments, cf. Chapter 14. An alternative method of variance reduction is to incorporate a properly defined covariate into an additive ACOVA model with treatments and the covariate, cf. Chapter 15. This section focuses on choosing proper covariates.

In designing an experiment to investigate a group of treatments, concomitant observations can be used to reduce the error of treatment comparisons. One way to use the concomitant observations is to define blocks based on them. For example, income, IQ, and heights can all be used to collect people into similar groups for a block design. In fact, any construction of blocks must be based on information not otherwise incorporated into the ANOVA model, so any experiment with blocking uses concomitant information. In analysis of covariance we use the concomitant observations more directly, as regression variables in the statistical model.

Obviously, for a covariate to help our analysis it must be related to the dependent variable. Unfortunately, improper use of concomitant observations can invalidate, or at least alter, comparisons among the treatments. In the example of Section 15.1, the original ANOVA demonstrated an effect on heart weights due to sex but after adjusting for body weights, there was little evidence for a sex difference. The very nature of what we were comparing changed when we adjusted for body weights. Originally, we investigated whether heart weights were different for females and males. The analysis of covariance examined whether there were differences between female heart weights and male heart weights *beyond what could be accounted for by the regression on body weights*. These are very different interpretations. In a designed experiment, we want to investigate the effects of the treatments and not the treatments adjusted for some covariates. To this end, in a designed experiment we require that the covariates be logically independent of the treatments. In particular, we require that

the concomitant observations be made before assigning the treatments to the experimental units,

the concomitant observations be made after assigning treatments to experimental units but before the effect of the treatments has developed, or

the concomitant observations be such that they are unaffected by treatment differences.

For example, suppose the treatments are five diets for cows and we wish to investigate milk production. Milk production is related to the size of the cow, so we might pick height of the cow as a covariate. For immature cows over a long period of time, diet may well affect both height and milk production. Thus to use height as a covariate we should measure heights before treatments begin or we could measure heights, say, two days after treatments begin. Two days on any reasonable diet should not affect a cow's height. Alternatively, if we use only mature cows their heights should be unaffected by diet and thus the heights of mature cows could be measured at any time during the experiment. Typically, *one should be very careful when claiming that a covariate measured near the end of an experiment is unaffected by treatments*.

The requirements listed above on the nature of covariates in a designed experiment are imposed so that the treatment effects do not depend on the presence or absence of covariates in the analysis. The treatment effects are logically identical regardless of whether covariates are actually measured or incorporated into the analysis. Recall that in the observational study of Section 15.1, the nature of the group (sex) effects changed depending on whether covariates were incorporated in the model. (Intuitively, the covariate body weight depends on the sex "treatment".) The role of the covariates in the analysis of a designed experiment is solely to reduce the error. In particular, using good covariates should reduce both the variance of the observations σ^2 and its estimate, the *MSE*. On the other hand, one pays a price for using covariates. Variances of treatment comparisons are σ^2 times a constant. With covariates in the model, the constant is larger than when they are not present. However, with well chosen covariates the appropriate value of σ^2 should be sufficiently smaller that the reduction in *MSE* overwhelms the increase in the multiplier. Nonetheless, in designing an experiment we need to play off these aspects against one another. We need covariates whose reduction in *MSE* more than makes up for the increase in the constant.

The requirements imposed on the nature of the covariates in a designed experiment have little affect on the analysis illustrated in Section 15.1. The analysis focuses on a model such as (15.1.2). In Section 15.1, we also considered model (15.1.3) that has different slope parameters for the different treatments (sexes). The requirements on the covariates in a designed experiment imply that the relationship between the dependent variable y and the covariate z *cannot* depend on the treatments. Thus with covariates chosen for a designed experiment *it is inappropriate to have slope parameters that depend on the treatment*. There is one slope that is valid for the entire analysis and the treatment effects do not depend on the presence or absence of the covariates. If a model such as (15.1.3) fits better than (15.1.2) when the covariate has been chosen appropriately, it suggests that the effects of treatments may differ from experimental unit to experimental unit. In such cases a treatment cannot really be said to have *an* effect, it has a variety of effects depending on which units it is applied to. A suitable transformation of the dependent variable may alleviate the problem.

17.9 Discussion of experimental design

Data are frequently collected with the intention of evaluating a change in the current system of doing things. If you really want to know the effect of a change in the system, you have to execute the change. It is not enough to look at conditions in the past that were similar to the proposed change because, along with the past similarities, there were dissimilarities. For example, suppose you think that instituting a good sex education program in schools will decrease teenage pregnancies. To evaluate this, it is not enough to compare schools that currently have such programs with schools that do not, because along with the differences in sex education programs there are other differences in the schools that affect teen pregnancy rates. Such differences may include parents' average socio-economic status and education. While adjustments can be made for any such differences that can be identified, there is no assurance that all important differences can be found. Moreover, initiating the proposed program involves making a change and the very act of change can affect the results. For example, current programs may exist and be effective because of the enthusiasm of the school staff that initiated them. Such enthusiasm is not likely to be duplicated when the new program is mandated from above.

To establish the effect of instituting a sex education program in a population of schools, you really need to (randomly) choose schools and actually institute the program. The schools at which the program is instituted should be chosen randomly, so no (unconscious) bias creeps in due to the selection of schools. For example, the people conducting the investigation are likely to favor or oppose the project. They could (perhaps unconsciously) choose the schools in such a way that makes the evaluation likely to reflect their prior attitudes. Unconscious bias occurs frequently and should *always* be assumed. Other schools without the program should be monitored to establish a base of comparison. These other schools should be treated as similarly as possible to the schools with the new program. For example, if the district school administration or the news media pay a lot of attention to the schools with the new program but ignore the other schools, we will be unable to distinguish the effect of the program from the effect of the attention. In addition, blocking similar schools together can improve the precision of the experimental results.

One of the great difficulties in learning about human populations is that obtaining the best data often requires morally unacceptable behavior. We object to having our lives randomly changed for the benefit of experimental science and typically the more important the issue under study, the more we object to such changes. Thus we find that in studying humans, the best data available are often historical. In our example we might have to accept that the best data available will be an historical record of schools with and without sex education programs. We must then try to identify and adjust for *all* differences in the schools that could potentially affect our conclusions. It is the extreme difficulty of doing this that leads to the relative unreliability of many studies in the social sciences. On the other hand, it would be foolish to give up the study of interesting and important phenomena just because they are difficult to study.

17.10 Analytic and enumerative studies

In one-sample, two-sample, and one-way ANOVA problems, we assume that we have random samples from various populations. In more sophisticated models we continue to assume that at least the errors are a random sample from a $N(0,\sigma^2)$ population. The statistical inferences we draw are valid for the populations that were sampled. Often it is not clear what the sampled populations are. What are the populations from which the Albuquerque suicide ages were sampled? Presumably, our data were all of the suicides reported in 1978 for these ethnic groups.

When we analyze data, we assume that the measurements are subject to errors and that the errors are consistent with our models. However, the populations from which these samples are taken may be nothing more than mental constructs. In such cases, it requires extrastatistical reasoning to justify applying the statistical conclusions to whatever issues we really wish to address. Moreover, the desire to predict the future underlies virtually all studies and, unfortunately, one can never be sure that data collected now will apply to the conditions of the future. So what can you do? Only your best. You can try to make your data as relevant as possible to your anticipation of future conditions. You can try to collect data for which the assumptions will be reasonably true. You can try to validate your assumptions. Studies in which it is not clear that the data are random samples from the population of immediate interest are often called *analytic studies*.

About the only time one can be really sure that statistical conclusions apply directly to the population of interest is when one has control of the population of interest. If we have a list of all the elements in the population, we can choose a random sample from the population. Of course, choosing a random sample is still very different from obtaining a random sample of observations. Without control or total cooperation, we may not be able to take measurements on the sample. (Even when you can find people that you want for a sample, many will not submit to a measurement process.) Studies in which one can arrange to have the assumptions met are often called *enumerative studies*. See Hahn and Meeker (1993) and Deming (1986) for additional discussion of these issues.

17.11 Exercises

EXERCISE 17.11.1. Garner (1956) presented data on the tensile strength of fabrics. Here we consider a subset of the data. The complete data and a more extensive discussion of the experimental procedure are given in Exercise 11.5.2. The experiment involved testing fabric strengths on four different machines. Eight homogeneous strips of cloth were divided into four samples. Each sample was tested on one of four machines. The data are given in Table 17.16.

Table 17.16: *Tensile strength of uniform twill*

Fabric	Machines			
strips	m_1	m_2	m_3	m_4
s_1	18	7	5	9
s_2	9	11	12	3
s_3	7	11	11	1
s_4	6	4	10	8
s_5	10	8	6	10
s_6	7	12	3	15
s_7	13	5	15	16
s_8	1	11	8	12

Table 17.17: *Dead adult flies*

	Units of active ingredient			
Medium	0	4	8	16
A	423	445	414	247
B	326	113	127	147
C	246	122	206	138
D	141	227	78	148
E	208	132	172	356
F	303	31	45	29
G	256	177	103	63

(a) Identify the design for this experiment and give an appropriate model. List all of the assumptions made in the model.

(b) Analyze the data. Give an appropriate analysis of variance table. Examine appropriate contrasts using Bonferroni's method with $\alpha = .05$

(c) Check the assumptions of the model and adjust the analysis appropriately.

EXERCISE 17.11.2. Snedecor (1945b) presented data on a spray for killing adult flies as they emerged from a breeding medium. The data were numbers of adults found in cages that were set over the medium containers. The treatments were different levels of the spray's active ingredient, namely 0, 4, 8, and 16 units. (Actually, it is not clear whether a spray with 0 units was actually applied or whether no spray was applied. The former might be preferable.) Seven different sources for the breeding mediums were used and each spray was applied on each distinct breeding medium. The data are presented in Table 17.17.

(a) Identify the design for this experiment and give an appropriate model. List all the assumptions made in the model.

(b) Analyze the data. Give an appropriate analysis of variance table. Ccompare the treatment with no active ingredient to the average of the three treatments that contain the active ingredient. Ignoring the treatment with no active ingredient, the other three treatments are quantitative levels of the active ingredient. On the log scale, these levels are equally spaced.

(c) Check the assumptions of the model and adjust the analysis appropriately.

EXERCISE 17.11.3. Cornell (1988) considered data on scaled thickness values for five formulations of vinyl designed for use in automobile seat covers. Eight groups of material were prepared. The production process was then set up and the five formulations run with the first group. The production process was then reset and another group of five was run. In all, the production process was

Table 17.18: *Cornell's scaled vinyl thickness values*

	Production setting							
Formulation	1	2	3	4	5	6	7	8
1	8	7	12	10	7	8	12	11
2	6	5	9	8	7	6	10	9
3	10	11	13	12	9	10	14	12
4	4	5	6	3	5	4	6	5
5	11	10	15	11	9	7	13	9

Table 17.19: *Phosphorous fertilizer data*

	Laboratory				
Fertilizer	1	2	3	4	5
F	20.20	19.92	20.91	20.65	19.94
G	30.20	30.09	29.10	29.85	30.29
H	31.40	30.42	30.18	31.34	31.11
I	45.88	45.48	45.51	44.82	44.63
J	46.75	47.14	48.00	46.37	46.63

set eight times and a group of five formulations was run with each setting. The data are displayed in Table 17.18.

(a) From the information given, identify the design for this experiment and give an appropriate model. List all the assumptions made in the model.

(b) Analyze the data. Give an appropriate analysis of variance table. Examine appropriate contrasts using the Bonferroni method with an α of about .05.

(c) Check the assumptions of the model and adjust the analysis appropriately.

EXERCISE 17.11.4. In data related to that of the previous problem, Cornell (1988) has scaled thickness values for vinyl under four different process conditions. The process conditions were A, high rate of extrusion, low drying temperature; B, low rate of extrusion, high drying temperature; C, low rate of extrusion, low drying temperature; D, high rate of extrusion, high drying temperature. An initial set of data with these conditions was collected and later a second set was obtained. The data are given below.

	Treatments			
	A	B	C	D
Rep 1	7.8	11.0	7.4	11.0
Rep 2	7.6	8.8	7.0	9.2

Identify the design, give the model, check the assumptions, give the analysis of variance table and interpret the F test for treatments. The structure of the treatments suggests looking at average rates, average temperatures, and interaction between rats and temperatures.

EXERCISE 17.11.5. Johnson (1978) and Mandel and Lashof (1987) present data on measurements of P_2O_5 (phosphorous pentoxide) in fertilizers. Table 17.19 presents data for five fertilizers, each analyzed in five labs. Our interest is in differences among the labs. Analyze the data.

EXERCISE 17.11.6. Table 17.20 presents data on yields of cowpea hay. Four treatments are of interest, variety I of hay was planted 4 inches apart (I4), variety I of hay was planted 8 inches apart (I8), variety II of hay was planted 4 inches apart (II4), and variety II of hay was planted 8 inches apart (II8). Three blocks of land were each divided into four plots and one of the four treatments

Table 17.20: *Cowpea hay yields*

	Block			Trt.
Treatment	1	2	3	means
I4	45	43	46	$44.66\bar{6}$
I8	50	45	48	$47.66\bar{6}$
II4	61	60	63	$61.33\bar{3}$
II8	58	56	60	58.000
Block means	53.50	51.00	54.25	$52.91\bar{6}$

Table 17.21: *Hydrostatic pressure tests: operator, yield*

A	B	C	D
40.0	43.5	39.0	44.0
B	A	D	C
40.0	42.0	40.5	38.0
C	D	A	B
42.0	40.5	38.0	40.0
D	C	B	A
40.0	36.5	39.0	38.5

was randomly applied to each plot. These data are actually a subset of a larger data set given by Snedecor and Cochran (1980, p. 309) that involves three varieties and three spacings in four blocks. Analyze the data. Check your assumptions. Examine appropriate contrasts.

Exercise 17.11.7. In the study of the optical emission spectrometer discussed in Example 17.4.1 and Table 17.1, the target value for readings was .89. Subtract .89 from each observation and repeat the analysis. What new questions are of interest? Which aspects of the analysis have changed and which have not?

Exercise 17.11.8. An experiment was conducted to examine differences among operators of Suter hydrostatic testing machines. These machines are used to test the water repellency of squares of fabric. One large square of fabric was available but its water repellency was thought to vary along the length (warp) and width (fill) of the fabric. To adjust for this, the square was divided into four equal parts along the length of the fabric and four equal parts along the width of the fabric, yielding 16 smaller pieces. These pieces were used in a Latin square design to investigate differences among four operators: A, B, C, D. The data are given in Table 17.21. Construct an analysis of variance table. What, if any, differences can be established among the operators? Compare the results of using the Tukey, Newman–Keuls, and Bonferroni methods for comparing the operators.

Exercise 17.11.9. Table 17.22 contains data similar to that in the previous exercise except that in this Latin square differences among four machines: 1, 2, 3, 4, were investigated rather than differences among operators. Machines 1 and 2 were operated with a hand lever, while machines 3 and 4 were operated with a foot lever. Construct an analysis of variance table. What, if any, differences can be established among the machines?

Exercise 17.11.10. Table 17.22 is incomplete. The data were actually obtained from a Graeco-Latin square that incorporates four different operators as well as the four different machines. The correct design is given in Table 17.23. Note that this is a Latin square for machines when we ignore the operators and a Latin square for operators when we ignore the machines. Moreover, every operator works once with every machine. Using the four operator means, compute a sum of squares for operators and subtract this from the error computed in Exercise 17.11.9. Give the new analysis

Table 17.22: *Hydrostatic pressure tests: machine, yield*

2	4	3	1
39.0	39.0	41.0	41.0
1	3	4	2
36.5	42.5	40.5	38.5
4	2	1	3
40.0	39.0	41.5	41.5
3	1	2	4
41.5	39.5	39.0	44.0

Table 17.23: *Hydrostatic pressure tests: operator, machine*

B,2	A,4	D,3	C,1
A,1	B,3	C,4	D,2
D,4	C,2	B,1	A,3
C,3	D,1	A,2	B,4

Operators are A, B, C, D.
Machines are 1, 2, 3, 4.

of variance table. How do the results on machines change? What evidence is there for differences among operators. Was the analysis for machines given earlier incorrect or merely inefficient?

EXERCISE 17.11.11. Table 17.24 presents data given by Nelson (1993) on disk drives from a Graeco-Latin square design (see Exercise 17.11.10). The experiment was planned to investigate the effect of four different substrates on the drives. The dependent variable is the amplitude of a signal read from the disk where the signal written onto the disk had a fixed amplitude. Blocks were constructed from machines, operators, and day of production. (In Table 17.24, Days are indicated by lower case Latin letters.) The substrata consist of A, aluminum; B, nickel plated aluminum; and two types of glass, C and D. Analyze the data. In particular, check for differences between aluminum and glass, between the two types of glass, and between the two types of aluminum. Check your assumptions.

Table 17.24: *Amplitudes of disk drives*

	Machine			
Operator	1	2	3	4
I	Aa 8	Cd 7	Db 3	Bc 4
II	Cc 11	Ab 5	Bd 9	Da 5
III	Dd 2	Ba 2	Ac 7	Cb 9
IV	Bb 8	Dc 4	Ca 9	Ad 3