Experimental Design - Homework Solutions

1 To Block or Not to Block?

A statistics instructor at Iowa State University wants to test whether an online tutorial (treatment 1) is more effective than a textbook (treatment 2) at helping students learn statistics. During class, she will have each student take a pre-test, complete the online tutorial or textbook exercises, and then complete a post-test. The response of interest is the increase in test score from pre-test to post-test. Because she teaches three separate sections of the class she is considering using a randomized complete block design (RCBD), and treating each class section as a block. There are 30 students in each section of her class. Based on previous studies, she expects that $\sigma^2_{CRD} = 35$, and $\sigma^2_{RCBD} = 32$. The RCBD has the form

$$y_{ijk} = \theta_i + \beta_j + \epsilon_{ijk}, i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, r$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$, and all $\epsilon' s$ are independent. Here, r is the number of replications per treatment within each block, so that the total number of observataions is $N = abr$. Note that this situation is slightly different than the one we discussed in class because we now have replicates in the RCBD. The error degrees of freedom for this design is $abr - a - b + 1$, which is equal to the degrees of freedom from SS_{AB} and SS_E in the two-factor factorial design. Note also that we are assuming that there is no interaction effect between treatment and block.

If the instructor wants a $(1 - \alpha)\%$ confidence interval around $\theta_1 - \theta_2$ that is as small as possible, should she use a CRD or a RCBD? Please justify your answer.

Solution

In class, we learned that an RCBD will result in a smaller (1-α)% confidence interval if

$\sigma_{RCBD}/\sigma_{CRD} < t_{n-a,1-\alpha/2}/t_{n-a-b+1,1-\alpha/2}$,

where n is the total number of observations, a is the number of treatments, and b is the number of blocks. (Actually, this is true only if the number of observations per treatment is the same in both designs. This was true in the notes, where we had $r_i = r = b$ observations per treatment in both designs, for all values of i.) In this problem, $n = 90$, $a = 2$, and $b = 3$.

The choice of α is up to you. I will consider the common α values of 0.1 and 0.05, which correspond to 90 and 95% confidence intervals.

$\alpha =\ 0.1 : \sigma_{RCBD}/\sigma_{CRD} = \sqrt{32}/\sqrt{35} = 0.956 < t_{90-2,1-.1/2}/t_{90-2-3+1,1-.1/2} = 1.662/1.663 = 0.999$,
$\alpha = 0.05 : \sigma_{RCBD}/\sigma_{CRD} = \sqrt{32}/\sqrt{35} = 0.956 < t_{90-2,1-.05/2}/t_{90-2-3+1,1-.05/2} = 1.987/1.988 = 0.999$,

and the inequality holds in both cases. Therefore, a RCBD should be used to achieve a narrower confidence interval.

2 Laundry Detergent Experiment (2^2 factorial design)

A researcher for a laundry detergent company wants to look at the effect of detergent concentration and stain type on the time it takes for the stain to be removed. In a preliminary study, he sets up a 2^2 factorial experiment to look at 2 levels of detergent concentration (factor A), either 3 teaspoons or 5 teaspoons, and 2 levels of stain type (factor B), either blue ink or tomato sauce.

Results from the experiment are in the text file: http://www.public.iastate.edu/∼gdancik/stat430/detergent.txt

Each row of the table corresponds to one run of the experiment, where column A contains the coded value of factor A (with -1 being 3 teaspoons), column B contains the coded value of factor B (with -1 being blue ink), and column y contains the time until the stain is removed (in seconds).

The standard 2^2 factorial effects model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, i = 1, 2, j = 1, 2, k = 1, \ldots, r$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and $\epsilon' s$ are independent.

You will probably want to use R to help you complete the problems below.

Solution

```
### read in the data ######
> data = read.table("http://www.public.iastate.edu/~gdancik/stat430/detergent.txt", header=TRUE)
> y = data[,3]

### create the design matrix ##
> mu = rep(1, dim(data)[1])
> A = data[,1]
> B = data[,2]
> AB = A*B
```

1. Calculate the main effects of factors A and B as well as the AB interaction effect

   ```
   ### calculate main effects ###
   > mainA = sum(A*y) / (length(y) / 2)
   > mainB = sum(B*y) / (length(y) / 2)
   > intAB = sum(AB*y) / (length(y) / 2)
   ```

   The main and interaction effects are as follows:
   A: 444.8
   B: -620.9
   AB: 294.2

2. Find estimates for the parameters $\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij}$ for $i = 1, 2$, $j = 1, 2$.

```
### calculate parameter estimates ###
> mu.hat = sum(mu*y) / length(y)
> alpha1.hat = -1*sum(A*y) / length(y)
> alpha2.hat = +1*sum(A*y) / length(y)
> beta1.hat = -1*sum(B*y) / length(y)
> beta2.hat = +1*sum(B*y) / length(y)
> ab11.hat = +1*sum(AB*y) / length(y)
> ab12.hat = -1*sum(AB*y) / length(y)
> ab21.hat = -1*sum(AB*y) / length(y)
> ab22.hat = +1*sum(AB*y) / length(y)
```

parameter estimates:
$\hat{\mu} = 2499.85$

$\hat{\alpha_1} = -222.4$	$(\hat{\alpha\beta})_{11} = 147.1$
$\hat{\alpha}_2 = 222.4$	$(\hat{\alpha\beta})_{12} = -147.1$
$\hat{\beta}_1 = 310.45$	$(\hat{\alpha\beta})_{21} = -147.1$
$\hat{\beta}_2 = -310.45$	$(\hat{\alpha\beta})_{22} = 147.1$

Using the fitted model, how long do you expect it would take to remove a blue ink stain using 3 teaspoons of the detergent?

```
### blue ink stain and 3 teaspoons correspond to the "low" values of each factor ###
> y.pred = mu.hat + alpha1.hat + beta1.hat + ab11.hat
```

$\hat{y}_{11} = \hat{\alpha}_1 + \hat{\beta}_1 + (\hat{\alpha\beta})_{11} = 2735$ seconds

3. Perform an F-test that tests the null hypothesis
H$_0$: $(\alpha\beta)_{ij} = 0$ for all i, j (there is no interaction effect) versus
H$_1$: $(\alpha\beta)_{ij} \neq 0$ for some i, j

To calculate SS_{AB}, note that

$$\overline{y}_{ij} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...} = \overline{y}_{ij} - (\overline{y}_{i..} - \hat{\mu}) - (\overline{y}_{.j.} - \hat{\mu}) + \overline{y}_{...} - \hat{\mu} - \hat{\mu} = \overline{y}_{ij.} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) = \hat{\alpha\beta}_{ij}$$

Therefore, $SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} r\left(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} r\left((\hat{\alpha\beta})_{ij}\right)^2$, with $(a-1)(b-1) = 1$ degree of freedom.

```
> r = length(y) / (2*2)  ## number of replicates, from N = abr, r = 10 here

## All interaction effects are the same magnitude (but different sign).
## Since we only need the square, this simplifies calculations.
SS.AB = 4*r*ab11.hat**2   ## SS.AB = 865536.4
```

This is obviously not the only way to calculate SS_{AB}. If you use the formula from the ANOVA table, you would probably use code similar to

```
y1bar =  mean(y[A==-1])
```

to calculate $\overline{y}_{1..}$, for example.

Using the definitions of each effect, one can similarly show that

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk}-\overline{y}_{ij.}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk}-\hat{y}_{ij}\right)^2,$$

where $\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + (\hat{\alpha\beta})_{ij}$.

Again, this is not the only way to do this, but I find that it is easier for calculations. I first get a vector of predicted responses, by using the fact that the values in the columns of the design matrices (the elements in the vectors A, B, and AB in my R code) are either $\pm$ 1 corresponding to the high or low level of each treatment, and these values are the same as the -1^t from the formula that gives us parameter estimates. The vector of predicted responses is then used to calculate SS_E. The R code is given below.

```
y.preds = sum(mu*y) / length(y) +
A * sum(A*y) / length(y) +
B * sum(B*y) / length(y) +
            AB * sum(AB*y) / length(y)
SSE = sum((y - y.preds)**2) ## SSE = 4968.2
```

Now under H_0,

$$F = \frac{SS_{AB}/\left((a-1)(b-1)\right)}{SS_E/\left(ab(r-1)\right)} \sim F_{(a-1)(b-1),ab(r-1)}$$

Using R,

```
F = (SSAB / 1)  / (SSE / (4*(r-1)))
```

and $F = 6271$.
Let's calculate a p-value for this F statistic

```
1 - pf(F, 1, 4*(r-1))
```

and this is (approximately) 0.

What is your conclusion, and practically, what does this result mean?

Therefore, we reject the null hypothesis that there is no interaction effect, and conclude that there is at least one interaction effect present. Practically, this means that it is misleading to look at the main effect of factor A, for example, because the effect of factor A depends on the level of factor B. In this problem, the main effect of factor A (444.8) indicates that on average the time until stain removal will be 444.8 seconds faster when going from the high level (5 teaspoons) to the lower level (3 teaspoons). (I accidentally reversed the high and low levels of this factor, so this result doesn't really make sense). Given this information, one might argue that using 3 teaspoons of detergent instead of 5 will allow a user to get rid of the stain 444.8 seconds faster. However, this improvement is very dependent on the value of B. When B is at it's low level, improvement in time is (main effect of A) - (AB interaction effect) = 150.6 seconds, and when B is at it's high level, improvement in time is (main effect of A) + (AB interaction effect) = 739 seconds. Arguing that a 444.8 second improvement in speed is possible is therefore misleading....it is possible that the improvement is much less. In certain cases, it is even possible that the main effect of A is positive at one value of B but negative at another.

I should probably mention that it is possible to do this entire analysis quickly in R by fitting a linear model using the coded values of the treatments (the $\pm 1's$ that we use), and then calling the anova function. However, it is useful to be able to answer this question "by hand" because R does not know how to analyze more complicated designs, such as BIBDs. Also note that the parameters in this model are slightly different (though directly related) to the parameters of the factor effects model. The code is given below:

```
> fit = lm(y~A+B+AB)
> fit
Call:
lm(formula = y ~ A + B + AB)

Coefficients:
(Intercept)            A            B           AB
     2499.8        222.4       -310.4        147.1

> anova(fit)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
A          1 1978470 1978470 14336.2 < 2.2e-16 ***
B          1 3855168 3855168 27934.9 < 2.2e-16 ***
AB         1  865536  865536  6271.8 < 2.2e-16 ***
Residuals 36    4968     138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3 Balanced Incomplete Block Designs

The randomized complete block design (RCBD) requires that all treatments are observed within each block. However, it is not always possible that block sizes are large enough to accomodate all treatments. An example of this situation is when an experiment requires an oven or machine that should be treated as a block but can only hold a limited number of experimental units. In these cases, a balanced incomplete block design (BIBD) is typically used. In a BIBD, there are a treatments, each treatment is observed a total of r times, there are b blocks each of size k, each treatment appears once in a block or not at all, and each pair of treatments appear within a block λ times. In a BIBD, the total number of observations is $N = ar = bk$, and b must be $\geq a$.

The model for a BIBD has the form

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} = \theta_i + \beta_j + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$. The index i can take on values $1, \ldots, a$, and j can take on values $1, \ldots, b$, but not all combinations of (i, j) appear in the design.

Consider the following BIBD, with $a = 3, b = 3, k = 2$, and $\lambda = 1$.

Block 1	1 2
Block 2	1 3
Block 3	2 3

In this design, block 1, for example, contains treatments 1 and 2 (which would be assigned randomly to the experimental units within the block).

In a RCBD, we saw that for a contrast $c = (c_1, \ldots, c_a)$, $\sum_{i=1}^{a} c_i \overline{y}_{i.}$ is an unbiased estimator of $\sum_{i=1}^{a} c_i \theta_i$, where by the definition of a contrast, $\sum_{i=1}^{a} c_i = 0$. However, this result does not hold for BIBD's.

Solution

1. In the BIBD above, show that $\sum_{i=1}^{a} c_i \overline{y}_{i.}$ is a *biased* estimator of $\sum_{i=1}^{a} c_i \theta_i$. This is true for all BIBD's, but for this exercise you only need to show it for the BIBD above.

   The BIBD above gives us the observations $y_{11}, y_{21}, y_{12}, y_{32}, y_{23}$, and y_{33}. These observations, along with their expected values, are given in the table below:

y_{ij}	$\mathrm{E}[y_{ij}]$
y_{11}	$\theta_1 + \beta_1$
y_{21}	$\theta_2 + \beta_1$
y_{12}	$\theta_1 + \beta_2$
y_{32}	$\theta_3 + \beta_2$
y_{23}	$\theta_2 + \beta_3$
y_{33}	$\theta_3 + \beta_3$

   Average responses for each treatment are

$$\overline{y}_{1.} = \tfrac{1}{2}(y_{11} + y_{12})$$
$$\overline{y}_{2.} = \tfrac{1}{2}(y_{21} + y_{23})$$
$$\overline{y}_{3.} = \tfrac{1}{2}(y_{32} + y_{33})$$

   and these have expected values

$$\mathrm{E}[\overline{y}_{1.}] = \theta_1 + \tfrac{1}{2}(\beta_1 + \beta_2)$$
$$\mathrm{E}[\overline{y}_{2.}] = \theta_2 + \tfrac{1}{2}(\beta_1 + \beta_3)$$
$$\mathrm{E}[\overline{y}_{3.}] = \theta_3 + \tfrac{1}{2}(\beta_2 + \beta_3)$$

   From these expectations, it follows that

$$\mathrm{E}\left[\sum_{i=1}^{a} c_i \overline{y}_{i.}\right] = \sum_{i=1}^{a} c_i \theta_i + \frac{1}{2}\left[c_1(\beta_1 + \beta_2) + c_2(\beta_1 + \beta_3) + c_3(\beta_2 + \beta_3)\right], \text{ or equivalently}$$
$$\mathrm{E}\left[\sum_{i=1}^{a} c_i \overline{y}_{i.}\right] = \sum_{i=1}^{a} c_i \theta_i + \frac{1}{2}\left[(c_1 + c_2)\beta_1 + (c_1 + c_3)\beta_2 + (c_2 + c_3)\beta_3\right]$$

   This estimator is clearly biased if block effects are present. For example, for $c = (1, -1, 0)$,

   $\mathrm{E}[\overline{y}_{1.} - \overline{y}_{2.}] = \theta_1 - \theta_2 + \frac{1}{2}(\beta_2 - \beta_3) \neq \theta_1 - \theta_2$ in general.

2. In a BIBD, an unbiased estimator for $\sum_{i=1}^{a} c_i\theta_i$ is

$$\frac{k}{\lambda a}\sum_{i=1}^{a} c_iQ_i,$$

where $Q_i = T_i - \frac{1}{k}\sum_{j=1}^{b} n_{ij}B_j$, $T_i = \sum_{j=1}^{b} n_{ij}y_{ij}$, and $B_j = \sum_{i=1}^{a} n_{ij}y_{ij}$, and n_{ij} is 1 if treatment i is observed in block j, and 0 otherwise. The extra notation is necessary because not all combinations of (i,j) are present. But note that T_i is the sum of all observations on treatment i, and B_j is the sum of all observations in block j. In the BIBD above, show that $\frac{k}{\lambda a}\sum_{i=1}^{a} c_iQ_i$ is an unbiased estimator of $\sum_{i=1}^{a} c_i\theta_i$. Again, this is true for all BIBD's, but for this exercise you only need to show it for the BIBD above.

Let's start with finding the expected values for the $T_i's$ and the $B_j's$, which are found by adding the expected values of the individual responses (given in the table for part (1) of this problem). In order to provide some intuition about why $\sum_{i=1}^{a} c_iQ_i$ is an unbiased estimator for all BIBDs, I will use r to refer to the number of times each treatment is observed ($r=2$ in this example), and make use of the variables k (the size of each block, $k=2$ in this example), and λ (the number of times each pair of treatments appear together, $\lambda = 1$ in this example).

	T_i	E$[T_i]$
T_1	$y_{11}+y_{12}$	$r\theta_1+\beta_1+\beta_2$
T_2	$y_{21}+y_{23}$	$r\theta_2+\beta_1+\beta_3$
T_3	$y_{32}+y_{33}$	$r\theta_3+\beta_2+\beta_3$

	B_j	E$[B_j]$
B_1	$y_{11}+y_{21}$	$k\beta_1+\theta_1+\theta_2$
B_2	$y_{12}+y_{32}$	$k\beta_2+\theta_1+\theta_3$
B_3	$y_{23}+y_{33}$	$k\beta_3+\theta_2+\theta_3$

Now, since $Q_i = T_i - \frac{1}{k}\sum_{j=1}^{b} n_{ij}B_j$,

$$\begin{aligned} E\left[Q_i\right] &= E\left[T_i - \tfrac{1}{k}\sum_{j=1}^{b} n_{ij}B_j\right] \\ &= E\left[T_i\right] - E\left[\tfrac{1}{k}\sum_{j=1}^{b} n_{ij}B_j\right] \\ &= E\left[T_i\right] - \tfrac{1}{k}\sum_{j=1}^{b} n_{ij}E\left[B_j\right], \end{aligned}$$

where n_{ij} is 1 if treatment i appears in block j, and is 0 otherwise.

This gives us the following expected values for the $Q_i's$.
$E[Q_1] = r\theta_1 + \beta_1 + \beta_2 - \frac{1}{k}(k\beta_1 + \theta_1 + \theta_2 + k\beta_2 + \theta_1 + \theta_3)$
$E[Q_2] = r\theta_2 + \beta_1 + \beta_3 - \frac{1}{k}(k\beta_1 + \theta_1 + \theta_2 + k\beta_3 + \theta_2 + \theta_3)$
$E[Q_3] = r\theta_3 + \beta_2 + \beta_3 - \frac{1}{k}(k\beta_2 + \theta_1 + \theta_3 + k\beta_3 + \theta_2 + \theta_3)$

Note that all of the block effects (the $\beta_j's$) cancel out in all $Q_i's$. Also note that for each Q_i, the B_j's (the term multiplied by $\frac{1}{k}$) that get subtracted from the $T_i's$ include all the responses in all blocks that contain treatment i. Since each treatment appears r times, for a given Q_i, the summation in the term multiplied by $\frac{1}{k}$ will contain r $\theta_i's$. Each treatment i is also paired with every other treatment λ times, so the summation in the term multiplied by $\frac{1}{k}$ will contain λ $\theta_p's$, for all $p \neq i$.

We can then rewrite the $E[Q_i]' s$ as follows:
$E[Q_1] = r\theta_1 - \frac{1}{k}(r\theta_1 + \lambda(\theta_2 + \theta_3))$
$E[Q_2] = r\theta_2 - \frac{1}{k}(r\theta_2 + \lambda(\theta_1 + \theta_3)))$
$E[Q_3] = r\theta_3 - \frac{1}{k}(r\theta_3 + \lambda(\theta_1 + \theta_2)))$

Now rewrite the expectations with a common denominator of k to obtain:
$E[Q_1] = \frac{r(k-1)\theta_1 - \lambda(\theta_2+\theta_3)}{k}$
$E[Q_2] = \frac{r(k-1)\theta_2 - \lambda(\theta_1+\theta_3)}{k}$
$E[Q_3] = \frac{r(k-1)\theta_3 - \lambda(\theta_1+\theta_2)}{k}$

Now to each $E[Q_i]$, add and subtact $\lambda\theta_i$ to the numerator to obtain:
$E[Q_1] = \frac{r(k-1)\theta_1 + \lambda\theta_1 - \lambda(\theta_1+\theta_2+\theta_3)}{k}$
$E[Q_2] = \frac{r(k-1)\theta_2 + \lambda\theta_2 - \lambda(\theta_1+\theta_2+\theta_3)}{k}$
$E[Q_3] = \frac{r(k-1)\theta_3 + \lambda\theta_3 - \lambda(\theta_1+\theta_2+\theta_3)}{k}$

From the BIBD aside posted on the wiki, we know that $r(k-1) + \lambda = \lambda a$, and we have
$E[Q_1] = \frac{\lambda a\theta_1 - \lambda(\theta_1+\theta_2+\theta_3)}{k}$
$E[Q_2] = \frac{\lambda a\theta_2 - \lambda(\theta_1+\theta_2+\theta_3)}{k}$
$E[Q_3] = \frac{\lambda a\theta_3 - \lambda(\theta_1+\theta_2+\theta_3)}{k}$

and in general, $E[Q_i] = \frac{\lambda a\theta_i - \lambda(\theta_1+\theta_2+\theta_3)}{k}$

Taking it to the next step (by adding the contrasts),

$$E\left[\sum_{i=1}^{a} c_i Q_i\right] = \frac{\lambda a \sum_{i=1}^{a} c_i\theta_i - \lambda(\theta_1+\theta_2+\theta_3)\sum_{i=1}^{a} c_i}{k}$$
$$= \frac{\lambda a \sum_{i=1}^{a} c_i\theta_i}{k}$$

since by definition of a contrast, $\sum_{i=1}^{a} c_i = 0$.

Finally, it follows that
$E\left[\frac{k}{\lambda a}\sum_{i=1}^{a} c_i Q_i\right] = \sum_{i=1}^{a} \theta_i$, which shows that the estimator is unbiased.