STA305/1004 - Design of Scientific Studies

(Class 21 note)

Nathan Taback

Mar 20, 2017

UNIVERSITY OF TORONTO

Today's Class

- Linear model for factorial design
- Advantages of factorial designs over one-factor-at-a-time designs
- Randomized block designs

Linear model for factorial design

Let y_i be the yield from the i^{th} run,

deviation coding

3 factors

each has 2 levels

needs 1 dummy

In total, we need 3 dummy variables

$$x_{i1} = \begin{cases} +1 & \text{if } T = 180 \\ -1 & \text{if } T = 160 \end{cases}$$

$$x_{i2} = \begin{cases} +1 & \text{if } C = 40 \\ -1 & \text{if } C = 20 \end{cases}$$

$$x_{i3} = \begin{cases} +1 & \text{if } K = B \\ -1 & \text{if } K = A \end{cases}$$

A linear model for a 2^3 factorial design is:

$$y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{\text{main effect}} + \underbrace{\beta_4 x_{i1}x_{i2} + \beta_5 x_{i1}x_{i3} + \beta_6 x_{i2}x_{i3} + \beta_7 x_{i1}x_{i2}x_{i3}}_{\text{2-way + 3-way interaction effect.}} + \epsilon_i.$$

The variables $x_{i1}x_{i2}$ is the interaction between temperature and concentration, $x_{i1}x_{i3}$ is the interaction between temperature and catalyst, etc.

Linear model for factorial design

$Y = X\beta + \varepsilon \implies \hat{\beta} = (X^T X)^{-1} X^T Y$

The table of contents for a 2^3 design is the design matrix X from the linear model above.

$TK = 10$ $\quad$ $TC = \frac{3}{2}$ $\quad$ 0 $\quad$ $TKC = \frac{1}{2}$

design matrix X

Mean	T	K	C	T:K	T:C	K:C	T:K:C	yield average
1	-1	-1	-1	1	1	1	-1	60
1	1	-1	-1	-1	-1	1	1	72
1	-1	-1	1	1	-1	-1	1	54
1	1	-1	1	-1	1	-1	-1	68
1	-1	1	-1	-1	1	-1	1	52
1	1	1	-1	1	-1	-1	-1	83
1	-1	1	1	-1	-1	1	-1	45
1	1	1	1	1	1	1	1	80
β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	$= Y$

- All factorial effects can be calculated from this table.
- Signs for interaction contrasts obtained by multiplying signs of their respective factors. $T{:}K = \text{col } T * \text{col } C$
- Each column perfectly balanced with respect to other columns.
- Balanced (orthogonal) design ensures each estimated effect is unaffected by magnitude and signs of other effects. eg $\langle T, K \rangle = \sum_{i=1}^{8} T_i K_i = 0$
- Table of signs obtained similarly for any 2^k factorial design. $k=3$

Linear model for factorial design - calculating factorial effects from parameter estimates

$\hat{\beta} = (X^T X)^{-1} X^T Y$

The parameter estimates are obtained via the `lm()` function in R.

- Estimated least squares coefficients are one-half the factorial estimates.
- Therefore, the factorial estimates are twice the least squares coefficients.

$\boxed{2\hat{\beta}_i = \text{effect} \quad i = 1, 2, \ldots 7.}$

$$\hat{\beta}_1 = 11.50 \Rightarrow T = 2 \times 11.50 = 23$$

$$\hat{\beta}_2 = 0.75 \Rightarrow K = 2 \times 0.75 = 1.5$$

$$\hat{\beta}_4 = 5.00 \Rightarrow TK = 2 \times 5.00 = 10.00$$

```r
fact.mod <-lm(y~T*K*C,data=tab0502)
round(summary(fact.mod)$coefficients,2)
```

		Estimate	Std. Error	t value	Pr(>|t|)
β_0	(Intercept)	64.25	NaN	NaN	NaN
β_1	T	11.50	NaN	NaN	NaN
β_2	K	0.75	NaN	NaN	NaN
β_3	C	-2.50	NaN	NaN	NaN
β_4	T:K	5.00	NaN	NaN	NaN
β_5	T:C	0.75	NaN	NaN	NaN
β_6	K:C	0.00	NaN	NaN	NaN
β_7	T:K:C	0.25	NaN	NaN	NaN

- For each factor-level combination, only one observation. And we have 8 unknown β's to estimate.
- No replicates, no extra infor. for s.e./t-value/p-value.

Linear model for factorial design - significance testing

- When there are replicated runs we also obtain p-values and confidence intervals for the factorial effects from the regression model.
- For example, the p-value for β_1 corresponds to the factorial effect for temperature

$H_0: T=0$ vs $H_1: T \neq 0$

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

why equivenlent?
since $2\beta_1 = T_{\text{effect}}$

If the null hypothesis is true then
$\beta_1 = 0 \Rightarrow T = 0 \Rightarrow \mu_{T+} - \mu_{T-} = 0 \Rightarrow \mu_{T+} = \mu_{T-}$.

- μ_{T+} is the mean yield when the temperature is set at 180° and μ_{T-} is the mean yield when the temperature is set to 160°.

To obtain 95% confidence intervals for the factorial effects we multiply the 95% confidence intervals for the regression parameters by 2. This is easily done in R using the function `confint.lm()`.

```
fact.mod <-lm(y~T*K*C,data=tab0503)
round(2*confint.lm(fact.mod),2)
```

```
             2.5 % 97.5 %
(Intercept) 125.54 131.96
T            20.04  26.46
K            -1.46   4.96
C            -7.96  -1.54
T:K           7.04  13.46
T:C          -1.46   4.96
K:C          -2.96   3.46
T:K:C        -2.46   3.96
```

$\boxed{\beta_i \ \text{CI} : (L_i, U_i)}$

$P(L_i < \beta_i < U_i) = 0.95$

$\boxed{\text{effect} = 2\beta_i}$

$P(2L_i < \text{effect}_i < 2U_i) = 0.95$

$\boxed{95\% \ \text{CI for effect}_i \quad (2L_i, 2U_i)}$

(*): CIs that do not contain 0.

→ significantly different from 0.

Advantages of factorial designs over one-factor-at-a-time designs

- Suppose that one factor at a time was investigated. For example, temperature is investigated while holding concentration at 20% (-1) and catalyst at B (+1).
- In order for the effect to have more general relevance it would be necessary for the effect to be the same at all the other levels of concentration and catalyst.
- In other words there is no interaction between factors (e.g., temperature and catalyst).

One-factor-at-a-time approach

① Suppose want to compare Temperature: + VS −

fix concentration $C =$ "−" and catalyst $K =$ "+"

Expt #1: compare T: "+" VS "−" and fix $C =$ "−", $K =$ "+"

⇒ Find $T =$ "+" is better.

② Take result from #1 and compare C

Expt #2: compare C: "+" VS "−" and fix $T =$ "+", $K =$ "−"

⇒ Find $C =$ "+" is better.

③ Take result from #2 and compare K

Expt #3: compare K: "+" VS "−" and fix $T =$ "+", $C =$ "+"

⇒ Find $K =$ "+" is better

In class note

- Suppose two factors $A\,(a_1, a_2)$ and $B\,(b_1, b_2)$
- one-at-a-time approach
  - Fix level of B and then compare levels of A

obs	Trmt #1 at $B=b_1$ $a_1 b_1$	Trmt #2 at $B=b_1$ $a_2 b_1$
1	x_1	y_1
2	x_2	y_2
3	x_3	Y_3
4	x_4	Y_4
	$\bar{X} = \frac{1}{4}(X_1 + \cdots + X_4)$	$\bar{Y} = \frac{1}{4}(Y_1 + \cdots + Y_4)$

assume $V(X_i) = V(Y_j) = \sigma^2 \Rightarrow V(\bar{X}) = V(\bar{Y}) = \frac{1}{16}(\sigma^2 + \sigma^2 + \sigma^2 + \sigma^2) = \frac{1}{4}\sigma^2$

In class note

- suppose $V(x_i) = V(y_i) = \sigma^2$

$$\Rightarrow \quad Var(\bar{x}-\bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma^2}{4} + \frac{\sigma^2}{4} = \frac{\sigma^2}{2}$$

Now if a_1 is better than a_2, then do the experiment again and fix $A=a_1$, then compare B: b_1 vs b_2

- For $A=a_1$, use 4 obs for each level of B.

$\Rightarrow$ Two factors, use one-at-a-time approach: need 16 obs

A	B	y
-1	-1	y_1
-1	-1	y_2
-1	+1	
-1	+1	⋮
+1	-1	
+1	-1	
+1	+1	
+1	+1	y_8

$\rightarrow$ But a factrial design, $Var(\bar{y}_+ - \bar{y}_-) = \frac{\sigma^2}{2}$

This would only take 8 runs vs 16 runs in one-at-a-time approach.

Advantages of factorial designs over one-factor-at-a-time designs

- Suppose that one factor at a time was investigated. For example, temperature is investigated while holding concentration at 20% (-1) and catalyst at B (+1).
- In order for the effect to have more general relevance it would be necessary for the effect to be the same at all the other levels of concentration and catalyst.
- In other words there is no interaction between factors (e.g., temperature and catalyst).
- If the effect is the same then a factorial design is more efficient since the estimates of the effects require fewer observations to achieve the same precision. → variance : the smaller, the better.
- If the effect is different at other levels of concentration and catalyst then the factorial can detect and estimate interactions.

① no interaction effect: Factorial design is more efficient since we need fewer observations

② interaction effect exists: Factorial design ⇒ main/interaction effect.

Randomized Block Designs

Randomized block designs

- Blocked designs extends the principle of paired comparisons to more than two treatments.
- This uses randomized designs with larger block sizes.

with 3 treatments
we need larger block size to
accommodate treatments.

Where do block design fit into what we have learned so far?

	unblocked	blocked
2 treatments	randomized unpaired	randomized paired
3 or more treatments	randomized one-way	randomized block

one-way ANOVA

key: how many treatments?

Randomized block designs

- In blocked designs two kinds of effects are contemplated:
  - treatments (this is what the experimenter is interested in).
  - blocks (this is what the experimenter wants to eliminate the contribution to the treatment effect).
- Blocks might be: different litters of animals (extension of twin idea); blends of chemical material; strips of land; or contiguous periods of time.

Example: penicillin yield

- In this example a process for the manufacture of penicillin was investigated and yield was primary response of interest.
- There were 4 variants of the process (treatments) to be compared.
- An important raw material corn steep liquor varied considerably.
- It was thought that corn steep liquor might causes significant differences in yield.

Y : Yield

Treatments : 4 variants of the process

Block : corn steep liquor

Example: penicillin yield

\# of Blocks

- Experimenters decided to study 5 blends of corn steep liquor.
- Within each blend the order in which the four treatments were run was random.
- Randomization done separately within each block. Within each blend the order in which the treatments were run were randomized.

Blend	A	B	C	D
1				
2				
3				
4				
5				

← • random order of A,B,C,D within each block.

• do the randomization separately.

Example: penicillin yield

- In a fully randomized one-way design blend differences might not be balanced between the treatments A, B, C, D. This might increase the experimental noise.
- But, by randomly assigning the order in which the four treatments were run within each blend (block), blend differences between the groups were largely eliminated.

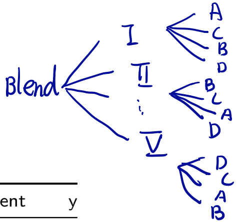
reduce: noise

Separate randomization

randomization

eg Block
- Female: A, C, B
- male: B, C, A

Example: penicillin yield



do randomization separately

The results of the experiment for blend 1

run	blend	treatment	y
1	1	A	89
3	1	B	88
2	1	C	97
4	1	D	94

order: A C B D

The results of the experiment for blend 2

run	blend	treatment	y
4	2	A	84
2	2	B	77
3	2	C	92
1	2	D	79

order: D C A B

Randomization of treatments was done separately within each block.

The ANOVA identity for randomized block designs

$Y_{ij} - \bar{Y}_{..} = (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}) + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..})$

The total sum of squares can be re-expressed by adding and subtracting the treatment and block averages as:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}}_{df=ab-1=n-1}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\left[(\bar{y}_{i.}-\bar{y}_{..})+(\bar{y}_{.j}-\bar{y}_{..})+(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})\right]^2.$$

After some algebra ... $SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2$ is equal to

$$\underbrace{b\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}_{SS_{trmt}}^{df=a-1} + \underbrace{a\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y}_{..})^2}_{SS_{Block}}^{df=b-1} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2}_{SSE}^{df=(N-1)-(a-1)-(b-1)}$$

$= (a-1)(b-1)$ since $N = ab$

So,

$$SS_T = SS_{Treat} + SS_{Blocks} + SS_E$$

understand d.f.

① $X_1, X_2, X_3 \Rightarrow df = 3$

② X_1, X_2, X_3, $\bar{X} = 7 \Rightarrow df = 2 = 3-1$

$\Rightarrow$

trmt: $\bar{Y}_1, \bar{Y}_2, \cdots \bar{Y}_a$, $\sum_{i=1}^{a}(\bar{Y}_{i.} - \bar{Y}_{..})^2 \leftarrow a-1$ (df), where $\bar{Y}_{..} = \frac{1}{a}\sum_{i=1}^{a}\bar{Y}_i$

block: $\bar{Y}_{.1}, \bar{Y}_{.2}, \cdots \bar{Y}_{.b}$, $\sum_{j=1}^{b}(\bar{Y}_{.j} - \bar{Y}_{..})^2 \leftarrow b-1$

Degrees of freedom

- There are N observations so SS_T has $\underset{ab}{N} - 1$ degrees of freedom.
- There are a treatments and b blocks so SS_{Treat} and SS_{Blocks} have $a - 1$ and $b - 1$ degrees of freedom, respectively.
- The sum of squares on the left hand side the equation should add to the sum of squares on the right hand side of the equation. Therefore, the error sum of squares has

$$df(SSE) = (N - 1) - (a - 1) - (b - 1) = (ab - 1) - (a - 1) - (b - 1) = (a - 1)(b - 1)$$

degrees of freedom.

source	df	SS
Treatment	a-1	SS_{trmt}
Block	b-1	SS_{block}
Error	(a-1)(b-1)	SSE
Total	N-1	SST

Poll Question

The goal of a certain field experiment is to test the effect of the amount of potash on the strength of cotton. There are 5 levels of potash (the treatments). A large section of a field will receive the treatments. Which of the following is closest to a randomized block design.

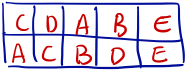
Respond at **PollEv.com/nathantaback**

Text **NATHANTABACK** to **37607** once to join, then **A or B**

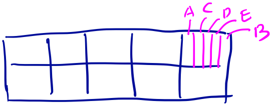
The field is divided into 10 plots and the 5 treatments are randomly assigned to the plots with each treatment in exactly 2 plots. **A**

The field is divided into 10 plots and 5 smaller sections of each plot is randomly assigned to receive the 5 treatments. **B**

A:



B:



Penicillin Manufacturing Example

The block averages are:

```
block.ave <- sapply(split(tab0404$y,tab0404$blend),mean); block.ave
```

```
 1  2  3  4  5
92 83 85 88 82
```

The treatment averages are:

```
trt.ave <- sapply(split(tab0404$y,tab0404$treatment),mean);trt.ave
```

```
 A  B  C  D
84 85 89 86
```

The grand average is:

```
grand.ave <- mean(tab0404$y);grand.ave
```

```
[1] 86
```

Penicillin Manufacturing Example

The block deviations from the grand average and the sum of squares of block deviations are:

```
block.devs <- block.ave-grand.ave; block.devs; sum(block.devs^2)*4
```

```
 1  2  3  4  5
 6 -3 -1  2 -4
```

$\leftarrow \bar{Y}_{.j} - \bar{Y}_{..}$

```
[1] 264
```

$\leftarrow a\sum_{j=1}^{b}(\bar{Y}_{.j} - \bar{Y}_{..})^2$

The treatment deviations from the grand average and the sum of squares of treatment deviations are:

```
treatment.devs <- trt.ave-grand.ave; treatment.devs; sum(treatment.devs^2)*5
```

```
 A  B  C  D
-2 -1  3  0
```

$\leftarrow \bar{Y}_{i.} - \bar{Y}_{..}$

```
[1] 70
```

$\leftarrow b\sum_{i=1}^{a}(\bar{Y}_{i.} - \bar{Y}_{..})^2$

Penicillin Manufacturing Example

The sum of squares of deviations from the grand average are:

```
all.devs <- tab0404$y-grand.ave; sum(all.devs^2)
```

```
[1] 560
```

$\leftarrow \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$

So, the error sum of squares is:

```
sum(all.devs^2)-sum(treatment.devs^2)*5-sum(block.devs^2)*4
```

```
[1] 226
```

$\leftarrow SSE = SS_{tot} - SS_{trmt} - SS_{Block}$

Poll question

If blocking was not incorporated into the design then what would happened to the value of SSE?

Respond at **PollEv.com/nathantaback**

Text **NATHANTABACK** to **37607** once to join, then **A, B, or C**

Increase	A
Decrease	B
Not change	C

$$\boxed{SS_{tot}} = \boxed{SS_{trmt}} + \underbrace{SS_{Block} + SSE_1}_{}$$

$SS_{Block} > 0$, $SSE_1 > 0$

$$\boxed{SS_{tot}} = \boxed{SS_{trmt}} + SSE_2$$

$SSE_2 > 0$

same once Y_{ij} given

$SSE_2 > SSE_1$

since $SSE_2 - SSE_1 = SS_{Block} > 0$

Linear Model for Randomized Block Design

- The linear model for the randomized block design is

$$y_{ij} = \mu + \underbrace{\tau_i}_{\text{trmt effect}} + \underbrace{\beta_j}_{\text{block effect}} + \epsilon_{ij},$$

$\begin{cases} i = 1, \ldots a \\ j = 1, \ldots b \\ Y_{ij} : \text{obs at } T = i, B = j \end{cases}$

  where $E(\epsilon_{ij}) = 0$.

- The model is completely additive.
- It assumes that there is no interaction between blocks and treatments.
- An interaction could occur if an impurity in blend 3 poisoned treatment B and made it ineffective, even though it did not affect the other treatments.

Linear Model for Randomized Block Design

- Another way in which an interaction can occur is when the response relationship is multiplicative

$$E(y_{ij}) = \mu\tau_i\beta_j.$$

Take log

- Taking logs and denoting transformed terms by primes, the model then becomes

$$y'_{ij} = \mu' + \tau'_i + \beta'_j + \epsilon'_{ij}$$

- Assuming that ϵ'_{ij} were approximately independent and identically distributed the response $y'_{ij} = log(y_{ij})$ could be analyzed using a linear model in which the interaction would disappear.

Linear Model for Randomized Block Design

- Interactions often belong to two categories:
  1. **transformable interactions**, which are eliminated by transformation of the original data, and
  2. **nontransfromable interactions** such as a treatment -blend interaction that cannot be eliminated via a transformation.

The ANOVA table for a randomized block design can be obtained by fitting a linear model and extracting the ANOVA table. Using R the penicillin example has ANOVA table

```
pen.model <- lm(y~as.factor(treatment)+as.factor(blend),data=tab0404)
anova(pen.model)
```

```
Analysis of Variance Table

Response: y
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(treatment)  3     70  23.333  1.2389 0.33866
as.factor(blend)      4    264  66.000  3.5044 0.04075 *
Residuals            12    226  18.833
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No evidence that trmt means are diff.

$\begin{cases} H_0: \mu_{T1} = \mu_{T2} = \mu_{T3} = \mu_{T3} \\ H_a: \text{at least one is diff from others.} \end{cases}$

$H_0: \mu_{B1} = \cdots = \mu_{B5}$

$H_a: \exists\, i,j \text{ s.t. } \mu_{Bi} \neq \mu_{Bj}$

If it's assumed that $\epsilon_{ij} \sim N(0, \sigma^2)$ then

$MS_{Treat}/MS_E \sim F_{a-1,(a-1)(b-1)}$, $MS_{Blocks}/MS_E \sim F_{b-1,(a-1)(b-1)}$.

Testing Treatment group means

Penicillin example - interpretation

- There is no evidence that the four treatments produce different yields.
- Q : How could this information be used in optimizing yield in the manufacturing process?
- ① Trmt cost: Is one of the treatments less expensive to run?
- If one of the treatments is less expensive to run then an analysis on cost rather than yield might reveal important information.
- ② block: The differences between the blocks might be informative.
- In particular the investigators might speculate about why blend 1 has such a different influence on yield.
- Perhaps now the experimenters should study the characteristics of the different blends of corn steep liquor. (Box, Hunter, Hunter, 2005)

Linear Model for Randomized Block Design

- The linear model for the randomized block design is

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij},$$

  where $E(\epsilon_{ij}) = 0$.
- The model is completely additive.
- It assumes that there is no interaction between blocks and treatments.
- An interaction could occur if an impurity in blend 3 poisoned treatment B and made it ineffective, even though it did not affect the other treatments.

- Another way in which an interaction can occur is when the response relationship is multiplicative

$$E(y_{ij}) = \mu\tau_i\beta_j.$$

Take log

- Taking logs and denoting transformed terms by primes, the model then becomes

$$y'_{ij} = \mu' + \tau'_i + \beta'_j + \epsilon'_{ij}$$

- Assuming that ϵ'_{ij} were approximately independent and identically distributed the response $y'_{ij} = log(y_{ij})$ could be analyzed using a linear model in which the interaction would disappear.

- Interactions often belong to two categories:
  1. **transformable interactions**, which are eliminated by transformation of the original data, and
  2. **nontransfromable interactions** such as a treatment -blend interaction that cannot be eliminated via a transformation.

The ANOVA table for a randomized block design can be obtained by fitting a linear model and extracting the ANOVA table. Using R the penicillin example has ANOVA table

```
pen.model <- lm(y~as.factor(treatment)+as.factor(blend),data=tab0404)
anova(pen.model)
```

```
Analysis of Variance Table

Response: y
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(treatment)  3     70  23.333  1.2389 0.33866
as.factor(blend)      4    264  66.000  3.5044 0.04075 *
Residuals            12    226  18.833
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$
H_a: at least one μ_i is different.
No evidence that trmts produce different yield.

$H_0: \mu_{B1} = \mu_{B2} = \cdots \mu_{Bb}$
H_a: at least one is different
We have evidence the different Block leads different yield

If it's assumed that $\epsilon_{ij} \sim N(0, \sigma^2)$ then $MS_{Treat}/MS_E \sim F_{a-1,(a-1)(b-1)}$, $\frac{MS_{Blocks}}{MSE} \sim F_{b-1,(a-1)(b-1)}$.

Penicillin example - interpretation

- There is no evidence that the four treatments produce different yields.
- Q: How could this information be used in optimizing yield in the manufacturing process?
- Is one of the treatments less expensive to run?
- If one of the treatments is less expensive to run then an analysis on cost rather than yield might reveal important information.

Trmt is non-sig consider cost

- The differences between the blocks might be informative.
- In particular the investigators might speculate about why blend 1 has such a different influence on yield.
- Perhaps now the experimenters should study the characteristics of the different blends of corn steep liquor. (Box, Hunter, Hunter, 2005)

Block is sig. investigate them