HW3

1(a)

power	σ^2
p_1	σ_1^2
p_2	σ_2^2
p_3	σ_3^2
p_4	σ_4^2

1(b). Assume $\sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \neq \sigma_4^2$
calculate power

1(c). Compare p's

4(a)

Treatment

a_1b_1	a_2b_1
y_1	x_1
y_2	x_2
y_3	x_3
y_4	x_4

$Var(\bar{y}_i - \bar{x}_i)$

STA305/1004-Class21

March 21, 2016

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,

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

$$x_{i2} = \begin{cases} +1 & \text{if } C = 40 \\ -1 & \text{if } C = 20 \end{cases} \qquad \text{factor 2}$$

$$x_{i3} = \begin{cases} +1 & \text{if } K = B \\ -1 & \text{if } K = A \end{cases} \qquad \text{factor 3}$$

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

$$y_i = \beta_0 + \overbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}^{\text{main}} + \overbrace{\beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{i1} x_{i2} x_{i3}}^{\text{interaction}} + \epsilon_i .$$

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

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

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

For T $:= \frac{-60+72-54+68-\cdots+\cdots}{4}$

- All factorial effects can be calculated from this table.
- Signs for interaction contrasts obtained by multiplying signs of their respective factors.
- 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.
- Table of signs obtained similarly for any 2^k factorial design. $K=3$

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

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.

$$\hat{\beta}_1 = 11.50 \Rightarrow T = 2 \times 11.50 = 23.26$$
$$\hat{\beta}_2 = 0.75 \Rightarrow K = 2 \times 0.75 = 1.5$$
$$\hat{\beta}_4 = 5.00 \Rightarrow TK = 2 \times 5.00 = 10.00$$

This can be shown by deriving the least square est. from 1st principles

See question in class notes

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

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    64.25        NaN     NaN      NaN
T              11.50        NaN     NaN      NaN
K               0.75        NaN     NaN      NaN
C              -2.50        NaN     NaN      NaN
T:K             5.00        NaN     NaN      NaN
T:C             0.75        NaN     NaN      NaN
K:C             0.00        NaN     NaN      NaN
T:K:C           0.25        NaN     NaN      NaN
```

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

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

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°.

Linear model for factorial design - significance testing

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)
```

multiply ×2

```
              2.5 % 97.5 %
(Intercept) 125.24 131.76
T             19.74  26.26
K             -1.76   4.76
C             -8.26  -1.74
T:K            6.74  13.26
T:C           -1.76   4.76
K:C           -3.26   3.26
T:K:C         -2.76   3.76
```

look for CI that do not contain 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).
- 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, smaller var more precise
- If the effect is different at other levels of concentration and catalyst then the factorial can detect and estimate interactions.

① Sps ~~want to~~ compare Temp + vs −

fix conc = −, catalyst = +

② Then take res of . . .

Expt #1 Compare Temp + vs −
and fix conc = −,
catalyst = −

4 obs/trt

Find temp = + is better

Expt #2. Fix Temp = +
Catalyst = −
and compare conc = − & +
find conc = + is better

4 obs/trt

Expt #3 Compare conc = + vs. conc = −
Fix Temp = +, conc = +

4 obs/trt

↑ basically one-at-a-time

↓ without interactions

↓ factorial design
more effecient

Sps 2 factors A (levels a_1, a_2)
B (levels b_1, b_2)

one at a time approach
Fix level of B at b_1 then
Compare levels of A

A		
obs	Trt #1 a_1b_1	~~Trad~~ Trt #2 a_2b_1
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4

∴ Now if a_1 is better than a_2 then do the experiment again & fix the level of A at a_1 then compare b_1 to b_2 for $A = a_1$. use 4 obs for each trt.

Sps $Var(x_i) = Var(y_i) = \sigma^2$

$Var(\bar{x} - \bar{y}) = \frac{\sigma^2}{2}$ ↑ need to show

$\therefore$ Two single factor expt use 16 obs.

But a replicated factorial approach

$Var(\bar{x} - \bar{y}) = \frac{\sigma^2}{2}$ (same precision)

(see last class note)

this would only take 8 runs.

vs. 16 runs in one-at-a-time approach

(but more efficient runs)

Randomized block designs

pair 1   Right A   Left B } Block size=2

randomiz a treatment

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

with 3 trts for example, we need larger blocks

Randomized block designs

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

key is how many trts?

1-way ANOVA

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. A B C D
- 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.

Example: penicillin yield

Blend
1
2
3
4
5

A B C D ← randomize the order

~~END~~

male
female

Randomize A B C

Randomize A B C

- Experimenters decided to study 5 blends of corn steep liquor. → 5 blocks
- 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.
- In a fully randomized one-way treatment classification blend differences might not be balanced between the treatments A, B, C, D. This might increase the experimental noise. → since we have separate randomizations
- 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.

one-way

Randomize A B C D

Blend I A B C D
II A B C D
III . . .
IV . .
V . .

Do separate randomization for each blend.

Example: penicillin yield

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

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

Randomization of treatments was done separately within each block.

The ANOVA identity for randomized block designs

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

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

After some algebra ...

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2=b\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2+a\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y}_{..})^2+\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2$$

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

Degrees of freedom

- There are N observations so SS_T has $N-1$ 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

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

degrees of freedom.

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**

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
```

```
[1] 264
```

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
```

```
[1] 70
```

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
```

So, the error sum of squares is:

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

```
[1] 226
```

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**

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.

Linear Model for Randomized Block Design

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

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

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

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

and 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 such as a treatment -blend interaction that cannot be eliminated via a transformation.

Linear Model for Randomized Block Design

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
```

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

Penicillin example - interpretation

- There is no evidence that the four treatments produce different yields.
- 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.
- 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)