A process engineer wishes to determine whether a change made to a chemical process has an important impact on the mean yield associated with a run of the process. A complicating issue in this regard is that each batch of raw material is sufficient to make only a few process runs, and different batches can be expected to have different characteristic yields.

Throughout this question, we will use a model for

$$y_{ijk} = \text{yield for run } k \text{ made using process } i \text{ and raw material batch } j$$

of the form

$$y_{ijk} = \mu_i + \beta_j + \varepsilon_{ijk} \qquad (*)$$

where the $\mu_i$ $(i = 1, 2)$ are unknown constants, the $\beta_j$ are iid $\mathrm{N}\left(0, \sigma_\beta^2\right)$ independent of the $\varepsilon_{ijk}$ that are themselves iid $\mathrm{N}\left(0, \sigma^2\right)$, and the variance components $\sigma_\beta^2$ and $\sigma^2$ are unknown constants.

Suppose initially that each raw material batch is sufficient to make 2 runs, and that 4 batches of raw material are available to the engineer for the study. Two possible plans for data collection are:

| **Plan I** | **Plan II** |
|---|---|
| 1 run from each raw material batch is made with each process (a total of 4 runs are made with each process) | 2 raw material batches are dedicated to each process (a total of 4 runs are made with each process) |

1. For

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

write out matrices $\mathbf{X}$ and $\mathbf{Z}$ so that the model (*) for 8 observations can be represented in usual matrix form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Do this first for **Plan I** and then for **Plan II**. In the first case, write the observations in the order

$$\mathbf{Y} = \begin{pmatrix} y_{111} \\ y_{211} \\ y_{121} \\ y_{221} \\ y_{131} \\ y_{231} \\ y_{141} \\ y_{241} \end{pmatrix}$$

In the second case, write the observations in the order

$$\mathbf{Y} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{231} \\ y_{232} \\ y_{241} \\ y_{242} \end{pmatrix}$$

2. What is the covariance matrix for $\mathbf{Y}$ (in the order indicated above) under **Plan 1**? Under **Plan 2**?

3. If all 4 unknown parameters were of some interest one might consider comparing **Plan 1** and **Plan 2** using appropriate $4 \times 4$ Fisher information matrices. Use the notation

$$\mathbf{D}\left(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2\right) = \text{the } 4\times 4 \text{ Fisher Information matrix for } \mathbf{U} \sim \text{MVN}_2\left(\begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \sigma_\beta^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma^2 + \sigma_\beta^2 \end{pmatrix}\right)$$

$$\mathbf{E}\left(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2\right) = \text{the } 4\times 4 \text{ Fisher Information matrix for } \mathbf{V} \sim \text{MVN}_2\left(\begin{pmatrix} \mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \sigma_\beta^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma^2 + \sigma_\beta^2 \end{pmatrix}\right)$$

$$\mathbf{F}\left(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2\right) = \text{the } 4\times 4 \text{ Fisher Information matrix for } \mathbf{W} \sim \text{MVN}_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \sigma_\beta^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma^2 + \sigma_\beta^2 \end{pmatrix}\right)$$

and write the Fisher Information matrices associated with **Plan I** and with **Plan II** in terms of these matrices.

4. Let $\overline{y}_{1..}$ be the arithmetic average of the process 1 observations and $\overline{y}_{2..}$ be the arithmetic average of the process 2 observations. What are the mean and standard deviation of

$$\overline{y}_{1..} - \overline{y}_{2..}$$

under **Plan 1**? Under **Plan 2**? (These can be found without appeal to matrix representations.)

5. For purposes of comparing $\mu_1$ and $\mu_2$, what does your answer to part 4 indicate about which of the two plans will typically be most effective?

Below are two hypothetical data sets, one corresponding to **Plan 1** and one corresponding to **Plan 2**.

| **Plan I** | | |
|---|---|---|
| Process | Batch | Yield |
| 1 | 1 | 82.0 |
| 2 | 1 | 78.6 |
| 1 | 2 | 71.8 |
| 2 | 2 | 75.6 |
| 1 | 3 | 80.0 |
| 2 | 3 | 78.8 |
| 1 | 4 | 77.6 |
| 2 | 4 | 77.8 |

| **Plan II** | | |
|---|---|---|
| Process | Batch | Yield |
| 1 | 1 | 82.0 |
| 1 | 1 | 79.2 |
| 1 | 2 | 71.9 |
| 1 | 2 | 76.3 |
| 2 | 3 | 79.3 |
| 2 | 3 | 78.9 |
| 2 | 4 | 77.0 |
| 2 | 4 | 77.8 |

6. For **both plans**, show the simple "by hand" calculations necessary to make valid/exact 95% $t$ confidence intervals for $\mu_1 - \mu_2$.

7. Simple valid/exact 95% $\chi^2$ confidence limits for $\sigma^2$ can be made from either set of hypothetical data above. Choose one of the plans and show the "by hand" calculations needed.

In the real application motivating this problem, practical constraints dictated that all runs from a given raw material batch had to be made consecutively, batches were of different sizes, and all runs from process 1 had to be made before runs from process 2. In fact, 4 small batches were dedicated to process 1, 1 larger batch was split between the two processes, and 1 batch of moderate size was dedicated to process 2. Attached to this question is an `R` printout useful in the analysis of the engineer's data. Use it in answering the following questions.

8. Is there a statistically significant difference between the processes? Explain, referring carefully to appropriate items on the printout.

9. How does run-to-run variability in yield appear to compare with batch-to-batch variability? Explain, again referring to appropriate items on the printout.

10. The engineer in charge of this study says to you “We need to redo this study. We’ll need to run process 1 before process 2. I can get raw material batches big enough to make as many as $r = 10$ runs per batch. We’ll run the same number of batches, $l$, with each process (splitting no batch between processes). I want to estimate $\mu_1 - \mu_2$ to within .5. I’d like to minimize the total number of runs made

$$\text{total runs made} = 2lr$$

in meeting this goal.” How many batches should we use for this study, and how many runs per batch should we make?

Find this person appropriate values of $r$ and $l$ on the basis of the estimates on the printout.

### R Printout

```
> data
   process batch     y
1        1     1 82.72
2        1     1 78.31
3        1     1 82.20
4        1     1 81.18
5        1     2 80.06
6        1     2 81.09
7        1     3 78.71
8        1     3 77.48
9        1     3 76.06
10       1     4 87.77
11       1     4 84.42
12       1     4 84.82
13       1     5 78.61
14       1     5 77.47
15       1     5 77.80
16       1     5 81.58
17       1     5 77.50
18       2     5 78.73
19       2     5 78.23
20       2     5 76.40
21       2     6 81.64
22       2     6 83.04
23       2     6 82.40
24       2     6 81.93
25       2     6 82.96

> Process<-as.factor(process)

> Batch<-as.factor(batch)

> output.1<-lme(y~1+Process,random=~1|Batch)

> summary(output.1)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  108.6438 113.1858 -50.32191

Random effects:
 Formula: ~1 | Batch
        (Intercept) Residual
StdDev:    2.927192 1.467032

Fixed effects: y ~ 1 + Process
               Value Std.Error DF  t-value p-value
(Intercept) 81.05442  1.260345 18 64.31128  0.0000
Process2    -0.67123  1.019483 18 -0.65841  0.5186
```

```
 Correlation: 
         (Intr)
Process2 -0.19 


Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max 
-1.901566862 -0.557726122 -0.005590905  0.505906835  2.018904889 

Number of Observations: 25
Number of Groups: 6 

> intervals(output.1)
Approximate 95% confidence intervals

 Fixed effects:
                lower       est.     upper
(Intercept) 78.406534 81.0544212 83.702308
Process2    -2.813087 -0.6712329  1.470621
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Batch 
                   lower     est.    upper
sd((Intercept)) 1.501453 2.927192 5.706776

 Within-group standard error:
   lower     est.    upper 
1.059645 1.467032 2.031042 


> predict(output.1,level=0:1)
   Batch predict.fixed predict.Batch
1      1      81.05442      81.09966
2      1      81.05442      81.09966
3      1      81.05442      81.09966
4      1      81.05442      81.09966
5      2      81.05442      80.62849
6      2      81.05442      80.62849
7      3      81.05442      77.69771
8      3      81.05442      77.69771
9      3      81.05442      77.69771
10     4      81.05442      85.31342
11     4      81.05442      85.31342
12     4      81.05442      85.31342
13     5      81.05442      78.61820
14     5      81.05442      78.61820
15     5      81.05442      78.61820
16     5      81.05442      78.61820
17     5      81.05442      78.61820
18     5      80.38319      77.94697
```

| | | | |
|---|---|---|---|
| 19 | 5 | 80.38319 | 77.94697 |
| 20 | 5 | 80.38319 | 77.94697 |
| 21 | 6 | 80.38319 | 82.29782 |
| 22 | 6 | 80.38319 | 82.29782 |
| 23 | 6 | 80.38319 | 82.29782 |
| 24 | 6 | 80.38319 | 82.29782 |
| 25 | 6 | 80.38319 | 82.29782 |