Experimental Design and Data Analysis

UCL Statistical Design and Data Ethics (Spring 2025)

Grp D

Experimental Design

The experiment investigates the effect of rehabilitation method on days to recovery for patients who had a hip replacement. Treatment variable $Method$ has three levels: A, B, and C. The response variable $Days$ is the number of days to recovery.

The patients who participate in the experiment come from three hospitals. Each of the hospitals can undertake a maximum of 7 hip replacement surgeries per week. For each patient: the hospital where the hip replacement takes place also provides the physiotherapy rehabilitation. The three methods for physiotherapy rehabilitation differ in type of exercises, the intensity of the exercises, and dietary choices.

All the data for the experiment will be collected in one week.

Significance level is fixed at 5%. The experiment is set up to detect a difference of a week (7 days) in mean number of recovery days across the methods if this difference is present. A sample-size calculation was undertaken and a sample size of 18 patients was chosen. This calculation was based on a standard linear model (with interaction terms) for a two-way ANOVA and used $\hat{\sigma}^2 = (3.5)^2$ as a preliminary guess of the common variance.

Data were collected following a protocol that tried to control the experimental environment as much as possible. Hospital is a blocking variable, and the three hospitals are denoted by $H1$, $H2$, and $H3$.

Question (1)

Given the experimental setting, briefly explain why the hospitals are used as blocks in the design.

> George Box: "Block what you can; randomise what you cannot."

The experiment is a two-way design with one treatment variable. It is called a **randomised complete block design** (see §4.7 and §4.8 in the lecture notes): the hospitals form blocks and methods for physiotherapy rehabilitation are treatments. Within each block, all treatments are used. Hospital may cause variation in the number of days to recovery for patients (due to underlying differences in quality of care, facility resources, socio-economic factors relating to patients and etc.), hence we would like to include it in the model. We block on *Hosipital* because it is a known source of variation from the external environment (i.e. a nuisance variable) but we cannot possibly randomise the experimental units across its levels since each patient can only have their hip replacement surgery in one of the hospitals.

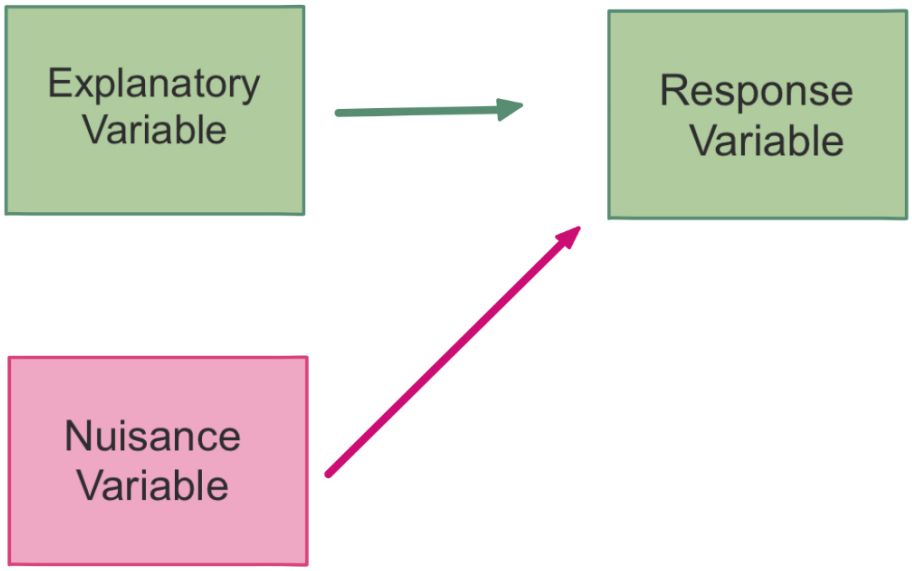

Figure 1: Nuisance variable effect on response variable[1]

Question (2)

Give an explicit example of how randomisation could have been applied in this setting. This example should provide concrete information on how to apply randomisation when the experiment is repeated.

Randomisation is applied within the blocks. At each of the hospitals, 6 patients who had a hip replacement surgery are chosen at random.

(Of course, here we assume that the three hospitals always have an abundance of patients who have just undergone a hip surgery and are waiting in line for physiotherapy)

[1]Wikipedia. https://upload.wikimedia.org/wikipedia/commons/f/fb/Nuisance_variable.jpg.

We then randomly assign 2 experimental units to each treatment group. The grouping can be done by a simple urn sampling without replacement.

Question (3)

Succinctly, give a protocol for controlling the experiment at the different stages; for example, about the timing of the physiotherapy, the measuring of the response, etc.

The duration of recovering from a hip replacement can vary depending on the patient's age and general health. At the recruitment stage, we may want to consider only accepting candidates within the same age group (e.g. young adults) or with similar general health status.

To measure number of days to recovery accurately, we should clearly define the date where recovery begins and the date of full recovery. The timer starts when a patient concludes their hip surgery. The timer stops when the said patient completes their last physiotherapy session and is ready to be discharged.

Question (4)

Show that given the information in the above section on sample size consideration, the choice of two repeated observations within each combination of treatment level and choice of hospital is indeed adequate.

The experiment concern a model for 3×3 with interaction terms and sum-to-zero constraints, which is expressed in

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{1}$$

where $i = 1, 2, 3$, $j = 1, 2, 3$, and $k = 1, \cdots, n$. And thus the sample size N is given by $9 \cdot n$.

Given the sum-to-zero constraints $\sum_i \alpha_i = \sum_j \beta_j = 0$, we have 4 main effects to be estimated. Given the additional sum-to-zero constraints $\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$, there are 4 interaction effects to be estimated. Including μ and σ^2 that are also to be estimated, there are 10 parameters in total. Hence, N must be greater than 10; n must be greater than 1. $n = 2$ is adequate to estimate all parameters in a standard linear model with interaction terms.

> Each of the hospitals can undertake a maximum of 7 hip replacement surgeries per week. [...] All the data for the experiment will be collected in one week.

Given this practical constraint, 2 repeated trials within each treatment level per hospital is the best we can have in the one-week's time span of the experiment.

Data Analysis

Data

The experiment investigates the effect of physiotherapy rehabilitation method on days to recovery for patients who had a hip replacement. Response *Days* is the number of days to recovery, denoted by `y` in code. Treatment variable $Method$ (denoted by `T` in code) has three levels: A, B, and C. Blocking variable $Hospital$ (denoted by `B` in code) has three levels: $H1$, $H2$, and $H3$. We are interested in knowing if the effect across treatment levels is different and if so, which method yields better outcome that is also contextually relevant. The data on the measured responses are given (in days) as follows:

		$H1$		$H2$		$H3$	
Method	A	21	22	23	20	29	31
	B	30	27	25	26	34	33
	C	29	32	31	29	40	38

The mean number of days to recovery across all treatment groups is 28.9 days. Figure 2 plots the average values of response for every combination of levels of $Method$ and $Hospital$ and shows the interaction effects between $Hospital$ and $Method$ on the number of days to recovery. We can observe that the average number of days to recovery increases from treatment level A to B to C. This holds true for all three hospitals. However, the difference between method A and B within $H1$ is more pronounced compared to $H2$ and $H3$; and the difference between method B and C within $H1$ is less pronounced. This suggests that there may be an interaction effect.

Model

In this analysis, we build an ANOVA model for a balanced two-way design with interaction terms. We define the model as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{2}$$

where $i = 1, 2, 3$, $j = 1, 2, 3$, and $k = 1, 2$. We have the sum-to-zero constraints $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$. We assume $\epsilon_{ijk} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

The average effects for each combination of levels of $Method$ and $Hospital$ on the "global" mean number of days to recovery (i.e. μ) can be seen in the table below.

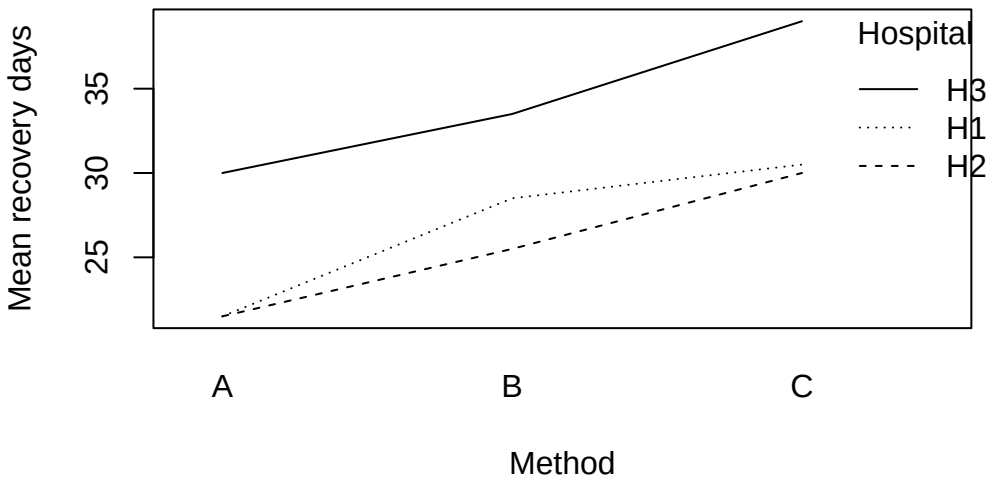

Figure 2: Mean recovery days by treatment and block variables

		$H1$	$H2$	$H3$
Method	A	$\alpha_1 + \beta_1 + (\alpha\beta)_{11}$	$\alpha_1 + \beta_2 + (\alpha\beta)_{12}$	$\alpha_1 + \beta_3 + (\alpha\beta)_{13}$
	B	$\alpha_2 + \beta_1 + (\alpha\beta)_{21}$	$\alpha_2 + \beta_2 + (\alpha\beta)_{22}$	$\alpha_2 + \beta_3 + (\alpha\beta)_{23}$
	C	$\alpha_3 + \beta_1 + (\alpha\beta)_{31}$	$\alpha_3 + \beta_2 + (\alpha\beta)_{32}$	$\alpha_3 + \beta_3 + (\alpha\beta)_{33}$

Results

The estimated coefficients of the two-way ANOVA model are as follows:

```
Full coefficients are

(Intercept):       28.88889
B:                      H1          H2          H3
                  -2.055556   -3.222222    5.277778
T:                       A           B           C
                  -4.555556   0.2777778   4.277778
B:T:                   H1:A        H2:A        H3:A        H1:B        H2:B
                  -0.77777778  0.38888889  0.38888889  1.38888889 -0.44444444
```

```
(Intercept):
B:

T:

B:T:                    H3:B        H1:C        H2:C        H3:C
                -0.94444444 -0.61111111  0.05555556  0.55555556
```

The intercept of the model is estimated by the overall mean recovery time across all method and hospital combinations, which is 28.9 days.

The model consists of the main effects α_is (i.e. estimated parameters under `A`, `B`, and `C` in `T`), the block effects β_js (i.e. estimated parameters under `H1`, `H2`, and `H3` in `B`), and the interaction effects $(\alpha\beta)_{ij}$s (e.g. estimated parameter under `H1:A` in `B:T`). The expected value of *Days* for each combination of levels of *Method* and *Hospital* can be calculated by $\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$. For example, the mean of cell $B \times H1$ is given by $\mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}$, which is 28.5.

The main effects and block effects can be interpreted as "average effects" on the expected value of the response when changing the treatment level (or block level)[2], while holding the other variable constant. The interaction effects address the effects that are "left over" due to a potential change in relationship between *Method* and *Days* across levels in *Hospital*.

We can compare the average effects of *Method* on *Days* according to the values of α_is. Method A generally reduces the number of days to recovery, while method B and C are associated with above-average recovery time. In addition, patients in $H2$ tend to have faster recovery compared to $H1$ and $H3$.

The ANOVA table for the fitted model is given below. The `R` output also includes the corresponding p-values of F-tests for the main effects and the interaction effect.

```
            Df Sum Sq Mean Sq F value   Pr(>F)
B            2 254.78  127.39  54.595 9.28e-06 ***
T            2 234.78  117.39  50.310 1.30e-05 ***
B:T          4   9.22    2.31   0.988    0.461
Residuals    9  21.00    2.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We are testing the null hypothesis that there is no main effect of *Method*, i.e. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$, against the alternative hypothesis that H_1 : at least one $\alpha_i \neq 0$. F-test has corresponding p-value: 9.28×10^{-6}. Hence, we reject the null hypothesis on a 5% significance

[2]Meier, L. (2022). ANOVA and Mixed Models: A Short Introduction Using R. Chapman and Hall/CRC. https://doi.org/10.1201/9781003146216.

level. There is at least one method that produces different mean number of days to recovery across all levels of *Hospital*.

F-test is also significant for the block effects *Hospital*. However, F-test provides no evidence of an interaction effect between *Hospital* and *Method* as the p-value is considerably large.

F-test provides strong evidence that *Method* affects *Days*. We can perform Tukey HSD to make comparisons between all possible pairs of levels in *Method*.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ B * T)

$T
    diff  lwr   upr p adj
B-A 4.83 2.37  7.30 0.001
C-A 8.83 6.37 11.30 0.000
C-B 4.00 1.54  6.46 0.004
```

There is evidence of a difference in each of the pairwise comparisons under *Method* using a significance level of 5%. We visualise the confidence intervals obtained in the `TukeyHSD` output in Figure 3. The biggest difference between group means is found in `C-A`. The 95% confidence interval for the difference between the mean number of recovery days under method A and C is given by $[6.37, 11.30]$.

We use diagnostic plots to check the model assumptions and ensure the validity of ANOVA inference. The normal Q-Q plot[3] for our residuals looks OK. Figure 4 suggests no departure from the normality assumption of the error term. Figure 5 suggests no evidence of a violation of the common variance assumption.

Discussion

In summary, there is evidence that the rehabilitation method affects recovery time. Tukey HSD suggests that method A is associated with faster recovery by approximately 5 and 9 days compared to B and C, respectively. The hospital where treatment was given also significantly affects recovery time. However, there's no evidence that treatment effects vary by hospital.

[3]The function `qqPlot` is a function of the package `car`. It draws normal Q-Q plots with a confidence envelope. Code is covered in Meier, L. (2022). ANOVA and Mixed Models: A Short Introduction Using R. Chapman and Hall/CRC. https://doi.org/10.1201/9781003146216. The package `car` is developed by Fox, J. and Weisberg, S. (2019). An R Companion to Applied Regression. Sage, Thousand Oaks CA. https://www.john-fox.ca/Companion/.

Figure 3: Tukey HSD's pairwise comparison

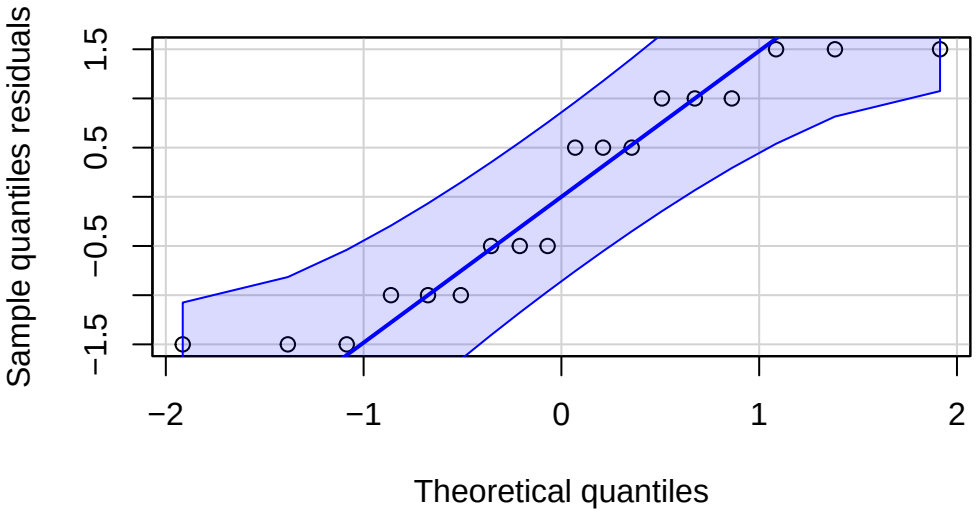

Figure 4: Normal Q-Q plot

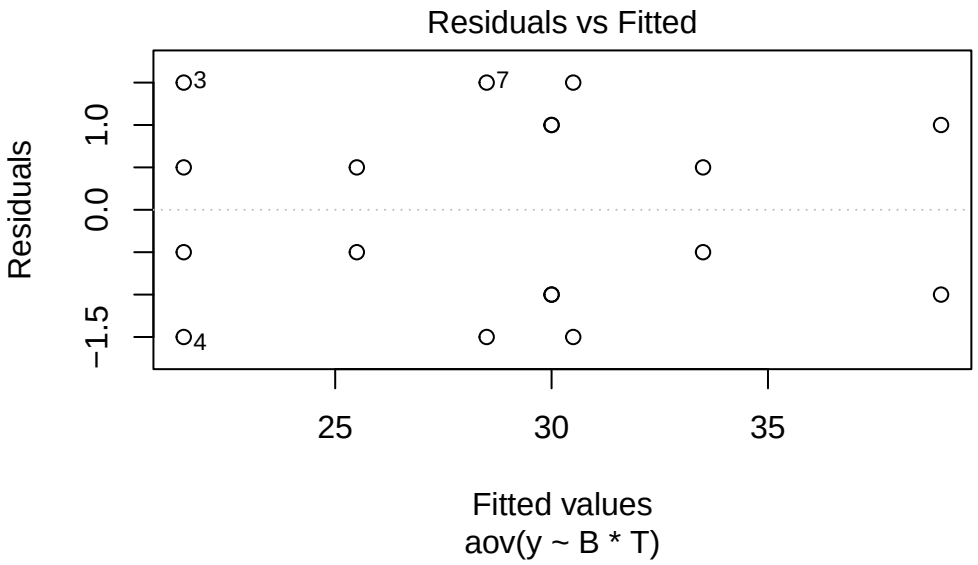

Figure 5: Tukey-Anscombe (residuals vs fitted values) plot

We argue that the differences between the mean recovery time across treatments are not only statistically significant, but relevant in the context. Method A effectively reduces the recovery time, on average, by 9 days compared to method C. This difference accounts for a third of mean recovery time under method C. Given that "each of the hospitals can undertake a maximum of 7 hip replacement surgeries per week", a reduction of 9 days in recovery time could improve patient turnover and efficiency of hospital operations in general. Hence, we conclude that the results are practically relevant.

Appendix

Use of AI Tools. ChatGPT is used for proofreading this report.