BASIC ANALYSIS FOR TRIAL DATA PART THREE

DAN W. JOYCE

Introduction

Unfortunately, we can't work with the original data from (Kane et al., 1988) because we can't access the patient-level data. However, using the figures and tables in Kane et al., we produced some simulated patient-level data for 265 patients. We saw this simulated data as `Kane-simulated.csv` in Part Two. Figure 1 shows Kane's results and our corresponding simulated data after treatment (note that we only simulated the total changes at the end of six weeks).

Nowadays, it's expected that published trials make participant-level data available so we can check the analyses used and conclusions

You should consult (Kane et al., 1988) to see why the actual number included in the analysis differs from the original number recruited and randomised

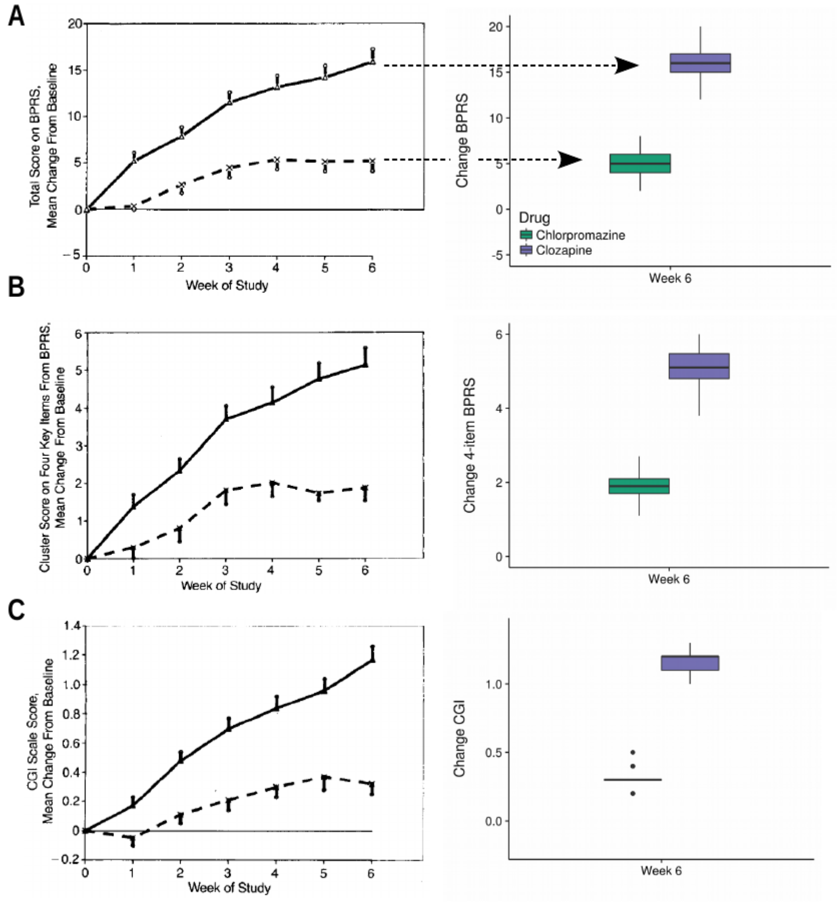

Figure 1: Results from (Kane et al., 1988) (left column) with corresponding simulated data (right column): A – Change in Mean BPRS, B – Change in Mean of the 4-item cluster BPRS score and C – Change in Mean CGI

In the simulated and real data there are changes in both groups evidenced by the week 6 outcomes being markedly different between clozapine and chlorpromazine groups. The Kane et al. graphs show relatively narrow error bars, and similarly, in our simulated data we show the BPRS change (using boxplots) and describe the spread of the data with 'whiskers' indicating the median plus/minus 1.5 times the interquartile range.

Now, we'll concentrate on the ways this data could be analysed emphasising different methods but without advocating for one or another. Any method of analysis needs to be justified depending on what hypotheses we need to test, the design of a study, suspected bias and the need to adjust for confounding factors in the study. At the very least, we will be able to show how (Kane et al., 1988) performed their analysis. Throughout, we will formulate analyses using the linear models described by equation 1 below.

Analysis by Repeated Measures ANOVA

Recall the equation for the linear model introduced in Part Two:

$$Y_{ijt} = \beta_0 + \beta_1 D_{ij} + \beta_2 T_{it} + \beta_3 D_{ij} T_{it} + \epsilon_{ijt} \tag{1}$$

Throughout, we have ignored the details of the error term ϵ_{ijt} which captures variation in Y_{ijt} that is *not* explained by the other terms in the linear model. To properly model these error terms, we would need a **hierarchical** model which is beyond the scope of this introduction – see (Gelman and Hill, 2007).

Where the variables are:

- Y_{ijt}: BPRS score (for patient i, in treatment group j at time t)
- $D_{ij} = 0$: chlorpromazine group
- $D_{ij} = 1$: clozapine group
- $T_{it} = 0$: before treatment
- $T_{it} = 1$: after treatment
- $D_{ij}T_{it}$: interaction between drug and the time

The formulation in equation 1 is a *repeated measures ANOVA*, it is a linear model, and we have framed it as a regression with discrete predictor variables for time, T, and drug treatment, D. This model lets us see if there is an *effect* of drug and time on the *means* of the BPRS values. To make this precise and concrete, examine Table 1 which describes each combination of time and drug and we have filled the cells with the corresponding means of BPRS values (the exercises in Part Two were to compute just these values).

		Drug	
		0	1
Time	0	$\mu_{00} = 62.23$	$\mu_{01} = 64.13$
	1	$\mu_{10} = 57.20$	$\mu_{11} = 48.20$

Table 1: Mean BPRS scores for each combination of time and drug

We denote the mean of Y_{ijt} (the BPRS value) as μ_{tj} for specific subsets of the data with j being the drug and t being the time. For example, if we want the mean BPRS for the drug chlorpromazine ($D_{i0} = 0$) at baseline ($T_{i0} = 0$) we collect together all rows of our data and compute their mean value of y and this would be labelled μ_{00} corresponding to the top-left cell in Table 1. Similarly, if we wanted the mean BPRS for the clozapine-treated patients at the end of the study, we would select all rows of our data where $D_{i1} = 1$ and $T_{i1} = 1$ and label it μ_{11}, located at the bottom-right of Table 1.

After computing the means for each cell in Table 1, after treatment, there seems to be a difference in the BPRS scores for patients treated with clozapine ($\mu_{11} = 48.20$) versus chlorpromazine ($\mu_{10} = 57.20$). There are now two questions that need answering:

1. Is the difference in BPRS scores between the two treatments *clinically* different ?
2. Assuming this is true, is the difference in BPRS scores attributable to chance (a 'fluke') or is the result likely to be reproducible – that is, are the results *statistically* significant ?

We know that the first question is addressed in Kane et al. using their inclusion criteria and their outcome is more sophisticated than simply change in the total BPRS score but as we are interested in understanding methods, we'll focus on statistical significance.

We'll use a statistics package to estimate the linear model in equation 1. The results we get are shown in Table 2. Let's revisit the terminology:

- the column Term in Table 2 refers to the variables in equation 1. So, for example, Drug is equivalent to D_{ij} and Time refers to T_{it}.
- when we set the variable $D_{ij} = 0$ or 1, we are examining chlorpromazine or clozapine respectively. Time behaves similarly with $T_{ij} = 0$ referring to pre-treatment, baseline or week 0 and $T_{ij} = 1$ refers to post-treatment, or week 6
- the term Drug:Time really means the interaction $D_{ij} \times T_{it}$ which describes how the BPRS scores differ when Time changes (from pre- to post-treatment) along with Drug (between chlorpromazine to clozapine treatments); we are looking at the effect of changes in time with respect to the two different drugs
- the column Beta in Table 2 are values estimated by the statistics package, and correspond to the coefficients β in equation 1. So, to find β_2 – the coefficient of the time variable T_{it} – we simply read off the term `Time`
- perhaps less obvious is the beta for (Intercept) which is equivalent to β_0 in equation 1

Different statistics packages will use different ways of formatting tabular output, as well as representing things like interaction terms

Fitted Model (Equation 1)		
	Estimated Beta	95% Conf. Int.
(Intercept)	62.23*	[60.56; 63.90]
Drug	1.90	[−0.53; 4.32]
Time	−5.03*	[−7.39; −2.66]
Drug:Time	−10.90*	[−14.33; −7.47]

* $p < 0.05$

Table 2: Typical output from a statistics package from estimating the repeated measures ANOVA model

To make interpretation of the output of the statistics package clearer, we re-tabulate Table 2 and insert the terms and coefficients from equation 1 alongside resulting in Table 3:

Output	Term in Eqn. 1	Coefficient	Estimated Beta	95% Conf. Int.
(Intercept)	–	β_0	62.23*	[60.56; 63.90]
Drug	D_{ij}	β_1	1.90	[−0.53; 4.32]
Time	T_{it}	β_2	−5.03*	[−7.39; −2.66]
Drug:Time	$D_{ij}T_{it}$	β_3	−10.90*	[−14.33; −7.47]

* $p < 0.05$

Table 3: Correspondence of outputs from a statistics package with linear model in equation 1

Interpreting Table 3, we note that the (Intercept) is 62.23 and the 95% confidence interval tells us this value could be as low as 60.56 or as high as 63.90. Also, note that this term's coefficient (β_0) is statistically significant at the $p < 0.05$ level

A simple definition of **confidence intervals**: if we repeat the Kane et al. study a large number of times (say, R times), and for each repetition we compute the linear model on the new data sample, we'll get R new values for the model βs. In 95% of these cases, the new βs will be between the lower and upper confidence interval reported, but in 5% of cases, they will not.

Next, notice how the top-left cell in our manually calculated table of means (Table 1), $\mu_{00} = 62.23$ is the same as the estimated value for (Intercept), corresponding to β_0 in equation 1 and Table 3. This is no coincidence, but a feature of the linear model. To summarise: the mean BPRS score for chlorpromazine at baseline was calculated to be μ_{00} and this is equivalent the estimated β_0 in the model described in equation 1.

We can now ask about the correspondence between the other coefficients β when we set the variables D_{ij} and T_{it} in equation 1 according to each of the cells in Table 1. Take the bottom-right cell, μ_{10} which corresponds to the mean BPRS for chlorpromazine $D_{ij} = 0$ after treatment $T_{it} = 1$ – substituting in equation 1:

$$\begin{aligned} Y_{i01} &= \beta_0 + \beta_1 \times D_{i0} + \beta_2 \times T_{i1} + \beta_3 \times D_{i0}T_{i1} \\ Y_{i01} &= \beta_0 + \cancelto{0}{\beta_1 \times 0} + \beta_2 \times 1 + \cancelto{0}{\beta_3 \times 0 \times 1} \\ Y_{i01} &= \beta_0 + \beta_2 \end{aligned} \tag{2}$$

Which tells us $\beta_0 + \beta_2$ should equal $\mu_{10} = 57.20$. Substituting the estimated values for β_0 and β_2 from Table 3:

$$\begin{aligned} \beta_0 + \beta_2 &= 62.23 + (-5.03) \\ &= 57.20 \end{aligned} \tag{3}$$

Repeating this procedure for all four combinations of D_{ij} and T_{it}, we see the correspondence of mean BPRS with estimated model coefficients shown in table 4. The additional calculations – similarly to equation 3 – are left as an exercise.

To summarise, we have seen how a linear model (equation 1) can be implement a repeated-measures ANOVA, and how the estimates obtained from a statistics package have a natural and intuitive interpretation that corresponds with the mean outcome (BPRS) computed 'manually' in Table 2. We should bear in mind that the relationships in Table 4 are *approximate* which is to say, the values obtained by adding the βs estimated by a statistics package will not be identical to the μs in Table 2. This is because our model (equation 1) assumes that each value of Y_{ijt} can be captured as a linear function – i.e. a sum of terms. This assumption will vary between patients and perhaps medications and time points, so we include an error term ϵ which captures additional variation. The algorithm implemented in a statistics package then tries to find the best estimate of each β given the model and it's assumptions. So, at the very least, you can check the assumptions of such models by looking at the cell means (Table 1) and seeing if they are reasonably close to the values given by the equivalences in Table 4.

		Drug	
		0	1
Time	0	$\mu_{00} : \beta_0$	$\mu_{01} : \beta_0 + \beta_1$
	1	$\mu_{10} : \beta_0 + \beta_2$	$\mu_{11} : \beta_0 + \beta_1 + \beta_2 + \beta_3$

Table 4: Correspondence of mean BPRS scores and βs

Questions and Exercises

1. Load `Kane-simulated.csv` into your preferred statistics package. Reproduce the 2×2 table of means for each combination of Time and Drug shown in Table 1 above (*Hint*: To compute the top-left cell, μ_{00}, you need to select all rows where the column Time = 0 and Drug = 0; then take the mean of the BPRS column for just those selected rows. Repeat this for each cell selecting instead rows which are selected by the different assignments of Time and Drug according to Table 1)

2. Using `Kane-simulated.csv`, plot the mean and 95% confidence interval on the mean – i.e. the graph showing the table of effects you computed above. It should look similar to Figure 2.

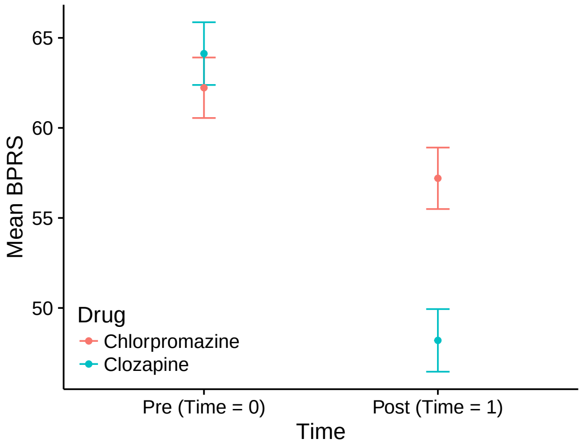

Figure 2: Plot of mean BPRS by Drug and Time

3. Attempt to fit the model given in equation 1 and compare the coefficients with the table derived above. Instructions are given below.

To implement equation 1 in SPSS, we have to cheat a little - we have to tell SPSS that Drug and Time are really numerical so that

SPSS refers to continuous numerical variables as *scale* variables they are entered directly as numbers into SPSS's algorithm for estimating the model. This ensures that the 'cancel out' mechanism described above (equations 2 and 3) works. SPSS normally tries to help users by working out which variables are nominal (categorical), and then the graphical user interface for different analyses helps further by using common terminology like 'factors' or 'covariates'. In fact, a factor or a covariate are really just terms (a variable and it's coefficient) on the right-hand side of equation 1. Given that we've been working with the 'raw' equations for each linear model, we don't need this help – rather we just want SPSS to take terms in our equations and estimate coefficients for them. For this reason, we use the Mixed Models tool in SPSS. The reasoning behind this was that by understanding the *actual* linear model, you can see the ways in which the models are similar and the differences in assumptions. Sometimes, statistics appears more complex because the mechanics of the underlying linear model are not spelled out.

To proceed in SPSS, you can load `Simulated-kane.sav` which is identical to `Kane-simulated.csv` only with the variables (columns) already configured correctly for analysis in SPSS (to save you some time)

Then, execute the following steps:

1. Select *Analyze*, *Mixed Models* then *Linear*

2. In the dialog box ('Linear Mixed Models: Specify Subjects and Repeated'), add variable 'Patient' to the *Subjects* list, and add 'Time' to the *Repeated* list, and click 'Continue'

3. In the next dialog box ('Linear Mixed Models'), add 'BPRS' as the *Dependent Variable*, and add both 'Drug' and 'Time' to the *Covariate(s)* list.

4. Click on the *Fixed* button, and then in the dialog, select both 'Drug' and 'Time' and hit *Add* so they appear in the *Model* list on the right – notice how SPSS adds entries for 'Drug', 'Time' and the interaction 'Drug*Time'. Click continue to return to the main dialog box.

5. Now, click *Statistics* and tick the *Parameter estimates* boxes – click continue.

6. Finally, back in 'Linear Mixed Models' dialog, hit *OK* and the model will be estimated

SPSS produces a lot of helpful diagnostic information, which we will ignore because in this tutorial, we are interested only in understanding the results. Scroll down to the 'Estimates of Fixed Effects' table, and compare with Table 2

This is not to say this is unimportant, but for this tutorial, we don't have scope to explore the details

References

Gelman, A. and Hill, J. (2007). *Data analysis using regression and hierarchical/multilevel models*. Cambridge University Press: Cambridge, UK.

Kane, J., Honigfeld, G., Singer, J., and Meltzer, H. (1988). Clozapine for the treatment-resistant schizophrenic: A double-blind comparison with chlorpromazine. *Archives of General Psychiatry*, 45(9):789–796.