MATH3029/4082: Design of experiments in R

The main purpose of this session to use the `lm` and `aov` commands in `R` for fitting CRDs and CRBDs using the linear model formulation with identifiability constraints.

A company wishes to compare 3 package delivery carriers between two of its branches. To compare delivery times each shipment is sent using the three services and the delivery times are recorded in hours. *The objective is to assess if there is a difference in the mean delivery times of the carriers*. Data is given below. If shipments are considered to be homogeneous and hence

	Shipment				
Carriers	1	2	3	4	5
a	15.2	14.3	14.7	15.1	14.0
b	16.9	16.4	15.9	16.7	15.6
c	17.1	16.1	15.7	17.0	15.5

not contributing to variability of the observed times, a CRD can be used with 5 replicates for each Carrier. On the other hand, if we think that type of shipment would affect the mean delivery times then we would use shipment type as a blocking variable and carry out an RCBD with one replicate for treatment/block combination.

1. We need to first create a dataset in `R`. In the notation used in the notes $t = 3, n_i = n = 5, i = 1, 2, 3$.

   ```
   shipment=factor(rep(1:5,each=3))
   carrier=factor(rep(letters[1:3],times=5))
   times=c(15.2,16.9,17.1,
   14.3,16.4,16.1,
   14.7,15.9,15.7,
   15.1,16.7,17.0,
   14.0,15.6,15.5)
   Delivery=data.frame(times,shipment,carrier)

   Delivery
   ```

   Note the usage of the command `rep` when defining the factor variables. Match the structure of dataset created to the one in the table above.

2. We can get an idea about the appropriate design (CRD or RCBD) by first 'looking at the data': just some basic exploratory analysis. Boxplots of times for different shipments will inform us if the distribution of times varies depending on shipment; the same applies to Carrier types as well. Note that boxplots merely provide (important) graphical summaries of what we can *reasonably expect* to conclude in the formal statistical tests for significance of treatment or block effects.

   ```
   boxplot(times~shipment,data=Delivery)
   boxplot(times~carrier,data=Delivery)
   ```

Plots seems to indicate difference across shipment and carrier types. It *appears* that an RCBD might be appropriate. Let's, however, fit both models and then compare conclusions.

- Recall that we need a constraint (identifiability condition) since the number of parameters to estimate is one more than the number of linearly independent equations. In the lecture notes we used the constraint $\sum_i \alpha_i = 0$ to enable estimation; other constraints could have been used. In `R` the *default* is to consider the treatment 'Carrier' as a factor with 3 levels, choose the first level (Carrier 'a') as the base level, and compare the other two levels to it (you would have seen this when dealing with factor variables using `lm` command); this amounts to the constraint, or identifiability condition, $\alpha_1 = 0$. If we use the condition $\sum_i \alpha_i = 0$, then all treatments are compared to the overall mean, as in the notes. In order to use this condition in `R`, run the command

  ```
  options(contrasts = c("contr.sum", "contr.sum"))
  ```

- We will use the command `aov` in `R` to fit linear models for designed experiments.

Fitting a CRD

3. Under CRD there is no shipment effect, and we just treat the delivery times across shipments as replicates for a fixed treatment (Carrier).

```
mod.crd=aov(times~carrier,data=Delivery)
 ## summary of overall test of significance
summary(mod.crd)
## since crd is a special case of lm,
## we can use summary.lm to obtain coefficient info
summary.lm(mod.crd)
```

   - Note that Carrier 1 is Carrier 'a' and so on. The first thing to check is for a carrier effect on the delivery times. That is, perform a hypothesis test for each treatment mean to be equal to zero ($H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ in the lecture notes). The F test (last line of output) for carrier rejects the null hypothesis at $\alpha = 0.05$, indicating that the choice of the carrier does indeed have an effect on the *average* delivery time.
   - The `(Intercept)` is the estimate $\bar{y}_{++}$ of the overall mean μ and is thus 15.747. Recall that the estimates of α_i, that is $\hat{\alpha}_i$, are $\bar{y}_{i+} - \bar{y}_{++}$ and table thus provides estimates $\hat{\alpha}_i$; this is the estimate of *effect* α_i of treatment i and not its mean (which of course is $\mu + \alpha_i$).
   - What is the estimate of α_3, the mean treatment effect for carrier 'c' (Recall identifiability condition $\sum_i \alpha_i = 0$ and this evidently applies to the estimates $\hat{\alpha}_i$ as well)? Visually match these numbers to the information in the boxplot for carrier.
   - Are there significant individual treatment effects? (For each $i = 1, 2, 3$, test $H_0 : \alpha_i = 0$ v $H_1 : \alpha_i \neq 0$). Notice that the individuals tests are t-tests, as covered in the lectures.

The t-test (and estimate) for effect of carrier 'c' can be implemented using the fact, from lecture notes, that `Std.Error` estimates of standard deviations for each $\hat{\alpha}_i$ is the same. Thus value of the test statistic $\frac{\hat{\alpha_3}-0}{\sqrt{Var(\hat{\alpha_3})}} = \frac{-(-1.08667+0.553)}{0.221}$ under H_0 is compared to an appropriate percentile of a t_{12} distribution.

4. The QQ plot of residuals to check validity of assumption on errors suggests a reasonable fit in the middle but some deviation near the tails.

```
plot(mod.crd,which=2)
```

Fitting an RCBD

5. Let's now fit an RCBD with one replicate. Note that we have set the identifiability conditions to $\sum_i \alpha_i = \sum_j \beta_j = 0$. The boxplot for shipment against times seems to indicate an effect. We will use the command `aov` for this since with this we get an ANOVA table of the type used in the notes.

```
mod.rcbd=aov(times~carrier+shipment,data=Delivery)
summary(mod.rcbd)
```

   Notice the two F-tests: one each for the treatment effect and the block effect. Match this with the material in the notes, especially with the notation and degrees of freedom. Is the block effect significant? Does the carrier effect change in the presence of the blocks? Just use the p-values for the F-test, say, at 5% level.

6. Estimates of treatment and block effects? Significance of individual treatments? We can exploit the connection to linear models and use `summary.lm` command to obtain these.

```
summary.lm(mod.rcbd)
```

   The F-statistic at the end of the output is the statistic for testing $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$; that is, is there a treatment and/or block effect? We reject H_0 based on the small p-value computed by using an $F_{6,8}$ distribution: 6= (no. of treatments-1) + (no. blocks-1)=2+4, and 8=14-6 since DF of SS_{total} has df $3 \times 5 - 1 = 15 - 1$.

   The p-values for the carriers 1 and 2 are small and they are individually significant. We can test $H_0 : \alpha_3 = 0$ in an identical manner as for the CRD.

   Similar arguments apply when testing $H_0 : \beta_5 = 0$.