STATISTICAL MODELLING

PRACTICAL V
SOLUTIONS

V.1 It is desired to run a wine-tasting experiment in which the differences between six wines are to be evaluated by scoring them on a 20-point scale. It is decided to have 6 expert judges evaluate the wines by evaluating a glass of wine on each of six consecutive occasions. It is desired to be able to isolate judge differences in scoring and differences between occasions so that a Latin square is to be employed.

A standard Latin square layout is given below.

A	B	C	D	E	F
B	A	F	E	C	D
C	F	B	A	D	E
D	C	E	B	F	A
E	D	A	F	B	C
F	E	D	C	A	B

Use R to obtain a randomized layout for the experiment, using a `seed` of 559 to set the random number generator seed.

The following expressions, produces the randomized layout:

```
> t <- 6
> n <- t*t
> LSWine.unit <- list(Judges=t, Occasions=t)
> Wines <- factor(c(1,2,3,4,5,6, 2,1,6,5,3,4, 3,6,2,1,4,5,
+                   4,3,5,2,6,1, 5,4,1,6,2,3, 6,5,4,3,1,2),
+                 labels=c("A","B","C","D","E","F"))
> LSWine.lay <- fac.layout(unrandomized=LSWine.unit, randomized=Wines, seed=559)
> remove("Wines")
> LSWine.lay
   Units Permutation Judges Occasions Wines
1      1          29      1         1     B
2      2          27      1         2     C
3      3          30      1         3     E
4      4          26      1         4     A
5      5          28      1         5     F
6      6          25      1         6     D
7      7          35      2         1     E
8      8          33      2         2     A
9      9          36      2         3     F
10    10          32      2         4     D
11    11          34      2         5     C
12    12          31      2         6     B
13    13          11      3         1     C
14    14           9      3         2     F
```

15	15	12	3	3	D
16	16	8	3	4	B
17	17	10	3	5	E
18	18	7	3	6	A
19	19	23	4	1	A
20	20	21	4	2	B
21	21	24	4	3	C
22	22	20	4	4	F
23	23	22	4	5	D
24	24	19	4	6	E
25	25	17	5	1	F
26	26	15	5	2	D
27	27	18	5	3	B
28	28	14	5	4	E
29	29	16	5	5	A
30	30	13	5	6	C
31	31	5	6	1	D
32	32	3	6	2	E
33	33	6	6	3	A
34	34	2	6	4	C
35	35	4	6	5	B
36	36	1	6	6	F

For example the first judge will taste wines in the order B, C, E, A, F, D.

V.2 In the lecture, a Latin square example was discussed that involved 4 drivers and 4 cars in testing the effects of 4 additives on the pollution produced by the cars. Discuss the circumstances in which the factors Drivers and Cars are likely to be regarded as fixed and those in which they are likely to be regarded as random.

Cars would be fixed if one expects a systematic difference between them such as if each car is a sample from a different brand to the other cars and there is no thought of these being representative of a larger group of cars. One expects the pattern in the deviations of the car means from the grand mean to be quite irregular and their distribution is likely to uninformative. The best model would appear to be that each brand has a different mean value.

Cars might be random if all were of the same brand and model and are supposed to be representative of all cars of that brand and model. In this case one expects the deviations to be randomly above and below the grand mean and could be described using a probability distribution with some variance.

Drivers might be random if they are just four different drivers selected to be representative of a wide range of drivers. In this case one expects a driver's mean to deviate from the grand mean somewhat randomly and could be modelled using a probability distribution with some variance.

Drivers might be fixed if they represent four different categories of driver with different training and experience. In this case the best model may well be that each driver has a different mean.

In this example, we do not have enough information to decide which factors should be fixed and which random.

V.3 The following data are from a Latin square experiment designed to investigate the moisture content of turnip greens. The experiment involved the measurement of the percent moisture content of five leaves of different sizes from each of five plants. The treatments were time of measurement in days since the beginning of the experiment.

		Plant									
		1		2		3		4		5	
Leaf Size (A = smallest, E = largest)	A	5	6.67	2	5.40	3	7.32	1	4.92	4	4.88
	B	4	7.15	5	4.77	2	8.53	3	5.00	1	6.16
	C	1	8.29	4	5.40	5	8.50	2	7.29	3	7.83
	D	3	8.95	1	7.54	4	9.99	5	7.85	2	5.83
	E	2	9.62	3	6.93	1	9.68	4	7.08	5	8.51

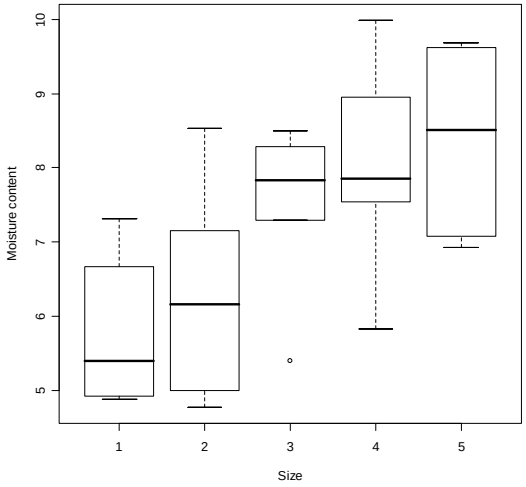
Classify the factors Leaf Size and Plant as either fixed or random.

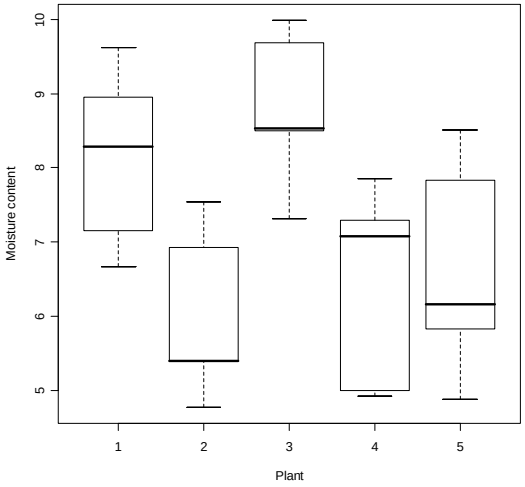
It is most likely that Leaf Size will be fixed whereas Plants will be random. There may well be systematic differences between leaves of different sizes and so this is best modelled using different means for each size. Plants on the other hand are most likely just 5 plants selected from many plants of this type and a probability distribution with some variance is likely to be an appropriate model.

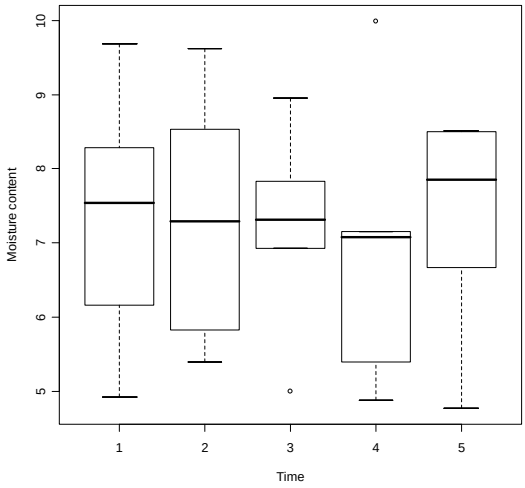
The factors Size, Plant and Time and the Moisture contents have been saved in the data.frame file *LSTurn.dat.rda* available from the available from the Statistical Modelling resources web site.

Analyze the data using R, including diagnostic checking and the examination of mean differences. Note that If you designated any one of the factors Leaf Size and Plant as random, the corresponding function $q(\boldsymbol{\psi})$ should be replaced by $5\sigma_S^2$ or $5\sigma_P^2$ in the E[MSq] for the line for the factor that is random (5 is the number of replicates of each Leaf Size and of each Plant).

```
> load("LSTurn.dat.rda")
> attach(LSTurn.dat)
> boxplot(split(Moisture, Size), xlab="Size", ylab="Moisture content")
> boxplot(split(Moisture, Plant), xlab="Plant", ylab="Moisture content")
> boxplot(split(Moisture, Time), xlab="Time", ylab="Moisture content")
```







There appears to be differences between the plants and the leaf sizes, but not between the times.

In the following output you will notice that Size and not Plant has terms included outside the Error function because Size is fixed but Plant is random.

```
> LSTurn.aov <- aov(Moisture ~ Size + Time + Error(Size*Plant), LSTurn.dat)
> summary(LSTurn.aov)

Error: Size
     Df Sum Sq Mean Sq
Size  4 23.708   5.927

Error: Plant
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  4 28.8853  7.2213

Error: Size:Plant
          Df Sum Sq Mean Sq F value Pr(>F)
Time       4 0.6273  0.1568  0.2327 0.9147
Residuals 12 8.0879  0.6740
> # Compute F and p for Size and Plant
> Size.F <- 5.927/0.6740
> Size.p <- 1-pf(Size.F, 4, 12)
> Plant.F <- 7.2213/0.6740
> Plant.p <- 1-pf(Plant.F, 4, 12)
> data.frame(Size.F,Size.p,Plant.F,Plant.p)
    Size.F      Size.p  Plant.F      Plant.p
1 8.793769 0.001482854 10.71409 0.0006232247
> #
> # Diagnostic checking
> #
> res <- resid.errors(LSTurn.aov)
> fit <- fitted.errors(LSTurn.aov)
> plot(fit, res, pch=16)
>
> qqnorm(res, pch=16)
> qqline(res)
> tukey.1df(LSTurn.aov, LSTurn.dat, error.term="Size:Plant")
$Tukey.SS
[1] 0.1318738

$Tukey.F
[1] 0.182329

$Tukey.p
[1] 0.6776171

$Devn.SS
[1] 7.956014
```

Step 1: Set up hypotheses

a) H_0: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5$ *(or* $\mathbf{X}_T\boldsymbol{\tau}$ *not required in model)*
H_1: *not all population Time means are equal*

b) H_0: $\sigma_P^2 = 0$
H_1: $\sigma_P^2 > 0$

c) H_0: $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5$ *(or* $\mathbf{X}_S\boldsymbol{\gamma}$ *not required in model)*
H_1: *not all population Size means are equal*

Set $\alpha = 0.05$.

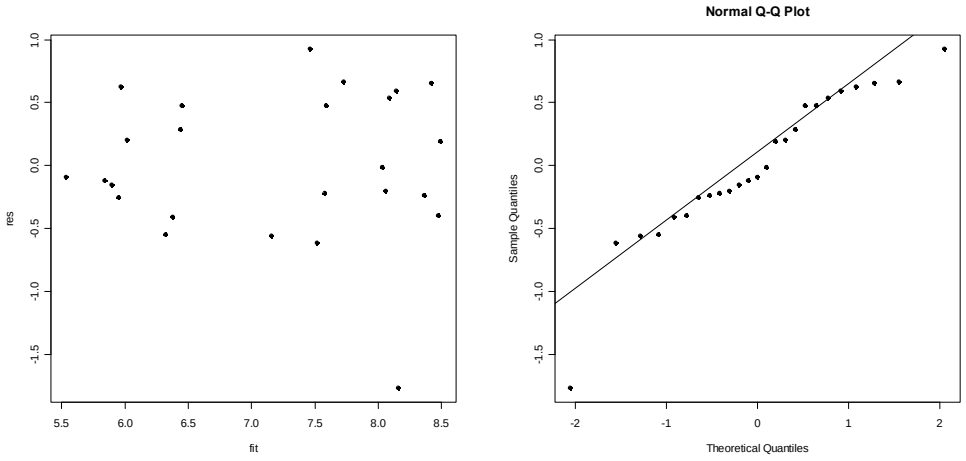
Step 2: Calculate test statistics

The analysis of variance table for a Latin square is:

Source	df	SSq	MSq	E[MSq]	F	Prob
Plant	4	28.8853	7.2213	$\sigma^2_{PS} + 5\sigma^2_{P}$	10.71	<0.001
Size	4	23.7081	5.9270	$\sigma^2_{PS} + q_{S}(\psi)$	8.79	0.001
Plant#Size	16					
Times	4	0.6273	0.1568	$\sigma^2_{PS} + q_{T}(\psi)$	0.23	0.915
Residual	12	8.0879	0.6740	σ^2_{PS}		
Non-additivity	1	0.1319	0.1319		0.18	0.678
Deviations	11	7.9560	0.7233			
Total	24	61.3806				

Step 3: Decide between hypotheses

There analysis indicates that there are no significant differences between the treatments, in spite of a very effective Latin square (both Plant and Size are significant). The expectation model that appears to best describe the data is the model $\psi_S = E[\mathbf{Y}] = \mathbf{X}_S\gamma$ *and* $\sigma^2_P > 0$. *Tukey's one-degree-freedom-for-nonadditivity is not significant indicating that there is not evidence of nonadditivity. However, the plot of residuals-versus-fitted-values indicates that there is an observation for which has an extremely low residual. Similarly, the normal probability plot is drawing attention to an outlier. From the list of residuals it is seen that the observation with the lowest residual is Plant 5 and Size 4 or the 20th observation. This observation needs to be investigated.*



V.4 The following layout is that appropriate to an experiment in which the same four drivers and the same four cars are used to repeat testing of four petrol additives. The assignment of additives to cars and drivers was accomplished using a Latin square on each of the two occasions the testing is conducted.

	Occasions							
	1				2			
Car	1	2	3	4	1	2	3	4
Driver								
1	A	B	C	D	A	B	C	D
2	C	D	A	B	C	D	A	B
3	D	C	B	A	D	C	B	A
4	B	A	D	C	B	A	D	C

Use R to verify that the analysis given in class, assuming all unrandomized factors are random, is the appropriate one for this experiment. You will need to set up the factor information, generate some random numbers from say a normal distribution using the `rnorm` function, and use the `aov` and `summary` functions to analyse the random data. For the `aov` function, the model formula should include an `Error` function whose argument is the unrandomized structure. Because all unrandomized factors are random, it is not necessary to include, outside the `Error` function, any terms from inside the `Error` function.

```
> r <- 2
> t <- 4
> n <- r*t*t
> LSRepeat1.dat <- fac.gen(generate = list(Occasion=r, Drivers=t, Cars=t))
> LSRepeat1.dat$Additives <- factor(c(1,2,3,4,3,4,1,2,4,3,2,1,2,1,4,3,
+                                     1,2,3,4,3,4,1,2,4,3,2,1,2,1,4,3),
+                                     labels = c("A", "B", "C", "D"))
> LSRepeat1.dat <- data.frame(LSRepeat1.dat, Data=rnorm(n))
> LSRepeat1.aov <- aov(Data ~ Additives + Error(Occasion*Drivers*Cars),
+                                                          LSRepeat1.dat)
```

```
> summary(LSRepeat1.aov)

Error: Occasion
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  1 1.5480  1.5480

Error: Drivers
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  3 1.96806 0.65602

Error: Cars
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  3 7.1813  2.3938

Error: Occasion:Drivers
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  3 1.15603 0.38534

Error: Occasion:Cars
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  3 3.9526  1.3175

Error: Drivers:Cars
          Df Sum Sq Mean Sq F value Pr(>F)
Additives  3 0.3176  0.1059  0.1145 0.9484
Residuals  6 5.5491  0.9249

Error: Occasion:Drivers:Cars
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  9 11.5510  1.2834
```

Note that the above ANOVA table is of the same form as that for case I given in section V.E, Sets of Latin squares. In particular, Additive is indented under just Drivers#Cars. Note how, that once you have the unrandomized and randomized structure formulae, they can be incorporated into an `aov` *function and R will produce the analysis.*