

2. Parallel groups design and its data analysis

§2.1. Parallel groups design and randomization schemes

The parallel groups design for comparing g treatments is as follows: n (a multiple of g) patients are randomly divided into groups of equal size, and the g treatments are randomly assigned to the g groups.

- **Why is randomization necessary?**

Randomization is to make the groups assigned to different treatments as homogeneous as possible so as to avoid bias.

Bias might be caused by assigning a particular treatment to a particular group of patients because other considerations, such as race, severity of the disease, etc.

When bias presents the treatment effects will be confounded with the effects of other (unknown) factors.

- **Simple randomization scheme and its drawbacks**

A simple randomization scheme is obtained by randomly permute the n patients and then divide the permute patients in order into g groups.

An example: 45 patients are randomly permuted and divided into 3 groups.

44	42	10	21	28	38	6	31	27	30	17	35	29	41	15
9	25	14	16	37	45	40	36	8	24	32	5	20	34	2
33	43	13	23	12	4	1	3	18	22	11	26	19	7	39

If the trials stops earlier than planned, it might result in an imbalance of patients among the groups, some bias might occur.

- **Randomly permuted blocks schemes**

Instead of permuting all patients together, patients are permuted in batches with size as a smaller multiple of g .

The simplest one is to permute the patients in batches of size g .

— This scheme is easier to implement, but might still cause bias if the trial cannot be completely double-blinded.

A modified one is to permute the patients in batches of size gr with $r \leq 3$ or 4 chosen randomly for each batch.

— This scheme is a good compromise between simplicity and unbiasedness.

§2.2. Analysis of variance and multiple comparisons

- Summarization of data from a parallel groups study

n_i : sample size of group i ,

X_{ij} : the measurement of patient j in group i .

Let $n. = \sum_{i=1}^g n_i$. Define

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad \bar{X}. = \frac{1}{n.} \sum_{i=1}^g n_i \bar{X}_i,$$
$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2, \quad S^2 = \frac{1}{n. - 1} \sum_{i=1}^g \sum_{j=1}^{n_i} (X_{ij} - \bar{X}.)^2.$$

Example: In a surgical trial involves 29 patients. Each patient received one of four anesthetics: Ether, Cyclopropane, Thiopental and Spinal. The measurement is the level of inorganic serum phosphorus after surgery.

The data is summarized as follows:

General				Example			
Sample				Sample			
Group	size	Mean	Var	Group	size	Mean	Var
1	n_1	\bar{X}_1	s_1^2	Ether	5	4.64	1.2080
2	n_2	\bar{X}_2	s_2^2	Cyclo	7	4.63	0.7390
:				Thio	9	3.53	0.2025
g	n_g	\bar{X}_g	s_g^2	Spinal	8	3.08	0.5479
Total	$n.$	$\bar{X}.$	S^2	Total	29	3.86	0.9915

• **Analysis of variance — Are there differences among groups?**

To test whether or not there is any difference at all among different treatments, an F -test from analysis of variance is applicable.

General analysis of variance table for data from a parallel groups study

Source	Between	Within	Total
df	$g - 1$	$n. - g$	$n. - 1$
SS	$\sum n_i(\bar{X}_i - \bar{X}.)^2$	$\sum (n_i - 1)s_i^2$	$(n. - 1)S^2$
MS	BMS	WMS	
F-ratio	BMS/WMS		

$$F = \text{BMS}/\text{WMS} \sim F_{g-1, n.-g}.$$

Analysis of variance table for the example

Source	Between	Within	Total
df	3	25	28
SS	13.04	14.72	27.76
MS	4.35	0.59	
F-ratio	7.37		

P-value = $P(F_{3,25} > 7.37) \approx 0.001$.

Conclusion: There are significant difference among the four anesthetics.

- **Multiple comparison — What are the differences?**

After it has been confirmed that there is difference among the treatments, it is more important to find what the difference is. There are two situations: a) there are no pre-hypothesized specific differences to investigate and a general exploring needs to be made to discover

the differences; b) there are certain scientifically important specific differences which are pre-hypothesized. Situation a) is tackled by Scheffe's method. One case of situation b) is tackled by Tukey's method.

Scheffe's method — general exploring

A general structure of difference is expressed by a *Contrast*. Let μ_i denote the effect of treatment i . A contrast is defined as

$$C = \sum_{i=1}^g c_i \mu_i \quad \text{with} \quad \sum_{i=1}^g c_i = 0.$$

All contrast together describe the totality of the difference among the treatments, because of the following equivalence:

$$\begin{aligned}
& \mu_1 = \mu_2 = \cdots = \mu_g \\
(1) \quad & \iff \\
& \sum_{i=1}^g c_i \mu_i \text{ for all } C .
\end{aligned}$$

In data-snooping, it is not possible to investigate all contrasts, the contrasts of interest are often suggested by the data. For instance, in the example, some contrasts suggested by the data are:

$$\begin{aligned}
C_1 & : (1, 0, -1, 0) \\
C_2 & : (0, 1, 0, -1) \\
C_3 & : (0.5, 0.5, 0, -1) \\
C_4 & : (0.5, 0.5, -0.5, -0.5)
\end{aligned}$$

To test whether a contrast (suggested by data) is significant, the Scheffe's method claims sig-

nificance at level α , if

$$L = \frac{\hat{C}}{\text{se}(\hat{C})} \geq \sqrt{(g-1)F_{g-1, n.-g, \alpha}},$$

where

$$\hat{C} = \sum_{i=1}^g c_i \bar{X}_i,$$

$$\text{se}(\hat{C}) = \sqrt{\text{WMS} \cdot \sum \frac{c_i^2}{n_i}},$$

and $F_{g-1, n.-g, \alpha}$ is the upper α -quantile of the F -distribution with df $g-1$ and $n.-g$.

Rationale of Scheffe's method:

Because of the equivalence (1), to control the Type I error of wrongly declaring significant difference among the treatments, it requires

to find a c_α such that

$$P\left(\max_{\text{all } C} \frac{\hat{C}}{\text{se}(\hat{C})} \geq c_\alpha\right) \leq \alpha.$$

It can be shown that

$$\max_{\text{all } C} \frac{\hat{C}}{\text{se}(\hat{C})} = (g - 1) \frac{\text{BMS}}{\text{WMS}},$$

where

$$\frac{\text{BMS}}{\text{WMS}} \sim F_{g-1, n.-g}.$$

Testing data suggested contrasts by Scheffe's method in the example:

	\hat{C}	$\text{se}(\hat{C})$	L	p-value
1	1.11	0.428	2.59	0.1089
2	1.55	0.398	3.89 ^a	0.0072
3	1.56	0.353	4.42 ^a	0.0021
4	1.33	0.292	4.55 ^a	0.0015

Remark:

(i). It is possible that the overall F -test is significant but none of the contrasts considered is significant.

(ii). For any single contrast, $\frac{\hat{C}}{\text{se}(\hat{C})}$ follows a t -distribution with df $n. - g$.

(iii) If the test for each contrast is based on the t -distribution, it is possible that even if the overall F -test is not significant at level α , some of the t -tests are significant at level α .

Why?

Tukey's method — pairwise comparison

Tukey's method is used for the situation that one is only concerned with pairwise differences.

Define the *studentized range* as:

$$q_{g,n.-g} = \frac{\max \bar{X}_i - \min \bar{X}_i}{\sqrt{\text{WMS}}}.$$

According to *Tukey's criterion*, the difference between group i and group j is significant at level α if

$$Q_{ij} = \frac{\sqrt{n} |\bar{X}_i - \bar{X}_j|}{\sqrt{\text{WMS}}} > q_{g,n.-g,\alpha},$$

when $n_1 = \dots = n_g = n$;

$$Q_{ij} = \frac{\sqrt{\tilde{n}_{ij}^{(H)}} |\bar{X}_i - \bar{X}_j|}{\sqrt{\text{WMS}}} > q_{g,n.-g,\alpha},$$

otherwise,

where $\tilde{n}_{ij}^{(H)} = \frac{2n_i n_j}{n_i + n_j}$.

For values of $q_{g,n.-g,\alpha}$, use Table A.5 of Fleiss.

Pairwise comparison by Tukey's method in the example:

Comparison	C_{ij}	$\tilde{n}_{ij}^{(H)}$	Q_{ij}
1 vs. 2	0.01	5.83	0.03
1 vs. 3	1.11	6.43	3.66
1 vs. 4	1.56	6.15	5.04 ^a
2 vs. 3	1.10	7.88	4.02 ^a
2 vs. 4	1.55	7.47	5.52 ^a
3 vs. 4	0.45	8.47	1.71

$${}^a Q_{ij} > q_{4,25,0.05} = 3.89$$

§2.3. Equality of variance, normality and transformations

- **Assumptions of analysis-of-variance model and consequence of their violation.**

The ANOVA model assumes

$$y_{ij} = \mu_i + \epsilon_{ij},$$

where ϵ_{ij} are iid $N(0, \sigma^2)$.

The most important features of the normal distribution are: 1) it is symmetric about its mean and 2) its variance does not depend on its mean.

The tests involved in the ANOVA and multiple comparison such as F -test are robust with respect to minor departure from the normality assumption.

But if the underlying distribution is seriously skewed, especially, when its variance depends on its mean, the accuracy of the tests will be adversely affected.

- **Transformation as a remedy.**

For a random variable with variance depending on mean, a transformation can be found

such that the variance of the transformed variable is independent of its mean. Suppose that, for the original variable X , $\sigma^2 = V(\mu)$, the proper transformation is derived as

$$h(x) = \int \frac{1}{\sqrt{V(x)}} dx.$$

The transformation is called *Variance Stabilization* transformation. It makes the transformed variable more like a normal variable.

- **Transformations for some particular responses.**

Proportion as response

Example: Two methods are applied to train patients with senile dementia to care for themselves. After the completion of the training, patients are asked to take 20 tests involving activities of daily living. The response from

each patient is the proportion of his or her tests that are successful.

A common feature for proportions: if the mean is close to either zero or 1, the variance is smaller; if the mean is close to 0.5, the variance is larger. The variance of such response can be approximated by

$$V(p) = cp(1 - p),$$

where c is a constant, p is the mean proportion.

The variance stabilization transformation is given by

$$h(p) = \arcsin \sqrt{p}.$$

Count as response

Examples: Counted number of micro-organisms in a sample of a patient's blood or saliva; counted

number of attacks of angina pectoris that a patient experience in a specified period of time, etc.

Most count response can be approximated by a Poisson random variable. The variance is proportional to the mean. The variance stabilization transformation is given by

$$h(\mu) = \sqrt{\mu}.$$

Time to event as response

Examples: disease free time after a treatment; survival times, etc.

For time to event, the variance is usually dependent of mean. A reciprocal transformation is appropriate when the variance is proportional to the fourth power of the mean.

The reciprocal transformation usually has a

physical meaning, e.g., the reciprocal of time to death is the death rate, the reciprocal of time to the occurrence of a reaction is the speed of the reaction, etc.

Response with log transformation

A log transformation is appropriate in the following situations:

1. Mean values are more sensibly compared in terms of their ratios than in terms of their differences.
2. The variance of the responses is proportional to the square of their mean.
3. The responses have a log-normal distribution.

Box-Cox transformation

A general family of transformations, Box-Cox

transformation, is given by

$$h(\mu) = \frac{\mu^\lambda - 1}{\lambda}.$$

The parameter λ can be determined by the data.

An example: Lysozyme levels in the gastric juice of 29 patients with peptic ulcer and of 30 normal controls.

Patient group:

0.2 0.3 0.4 1.1 2.0 2.1 3.3 3.8 4.5 4.8 4.9
5.0 5.3 7.5 9.8 10.4 10.9 11.3 12.4 16.2 17.6 18.9
20.7 24.0 25.4 40.0 42.2 50.0 60.0

Control group:

0.2 0.3 0.4 0.7 1.2 1.5 1.5 1.9 2.0 2.4 2.5 2.8
3.6 4.8 4.8 5.4 5.7 5.8 7.5 8.7 8.8 9.1 10.3 15.6
16.1 16.5 16.7 20.0 20.7 33.0

The t -test for the comparison of the two groups is significant at level 0.05. But the test is doubtful.

Mean and standard deviation of the two groups:

$$\bar{X}_1 = 14.31, \bar{X}_2 = 7.68,$$

$$s_1 = 15.74, s_2 = 7.85.$$

$$s_1/s_2 = 2.0051.$$

The unequal standard deviations suggest that a transformation is in order.

Standard deviations with various transformations

	s_1	s_2	s_1/s_2
$\sqrt{\mu}$	1.9648	1.3649	1.4395
$1/\mu$	1.1348	1.1070	1.0251
$\log(\mu)$	1.4814	1.3162	1.1255

An appropriate transformation can be either the reciprocal or the log transformation. However, the t -test based on the transformed data, say, log transform, is no longer significant at level 0.05.

§2.4. Non-parametric methods

Non-parametric methods should be used if responses are non-normally distributed, especially, when the parametric tests based on the original data and on the transformed data do not coincide.

Non-parametric tests are distribution-free; that is, the distribution of the test statistics does not depend on the underlying distribution of the responses.

• Kolmogorov-Smirnov test

This test can be used for the comparison of two groups.

Test statistic:

$$M = \max_x |\hat{F}_1(x) - \hat{F}_2(x)|,$$

where

$$\hat{F}_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} I\{X_{ij} \leq x\}, \quad i = 1, 2.$$

The difference between the two groups is judged significant at level α if

$$M \sqrt{\frac{n_1 n_2}{n_1 + n_2}} > \kappa_\alpha.$$

Large-sample critical values for
Kolmogorov-Smirnov test:

α	κ_α
0.20	1.073
0.1	1.224
0.05	1.358
0.025	1.480
0.01	1.628
0.005	1.731
0.0025	1.828
0.001	1.949

- **Mann-Whitney-Wilcoxon and Kruskal-Wallis tests**

Mann-Whitney-Wilcoxon (MWW) and Kruskal-Wallis (KW) tests are based on ranks of the responses. MWW test is for the comparison of two groups. KW test is for the comparison of more than two groups.

MWW test statistic is given by

$$\chi_{\text{mww}} = \frac{12n_1n_2(\bar{R}_1 - \bar{R}_2)^2}{n^2(n. + 1)}.$$

KW test statistic is given by

$$\chi_{\text{kw}} = \frac{12 \sum n_i(\bar{R}_i - \frac{n.+1}{2})^2}{n.(n. + 1)}.$$

In the above, \bar{R}_i is the average ranks of the responses in group i when all the observations are ranked together.

If several observations are tied, each is given the average of their ranks they would have received if they had not been tied. When there are ties, the denominator of both MWW and KW statistic should be adjusted by a factor f defined as

$$f = 1 - \frac{\sum_{i=1}^T t_i(t_i - 1)(t_i + 1)}{n.(n. - 1)(n. + 1)},$$

where T is the total number of values at which there are ties, t_i is the number of ties at the i th tied value.

Under null hypothesis that there is no difference among the groups,

$$\chi_{\text{mww}} \sim \chi_1^2, \quad \chi_{\text{kw}} \sim \chi_{g-1}^2.$$

§2.5. Further notes

- **Computation issues**

The data analyses discussed in this chapter can be implemented in R by the following functions:

```
t.test, lm, ks.test,  
wilcox.test, kruskal.test.
```

- **Multivariate responses**

Univariate t -test and ANOVA can be extended to the multivariate responses straightforwardly. For MANOVA (multivariate analysis of variance), see, e.g., R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis, Chapter 6. (Prentice Hall).

- **Responses are categorical**

Generalized linear models with categorical responses are applicable, see, e.g., L. Fahrmeir and G. Tutz, *Multivariate Statistical Modelling Based on Generalized Linear Models*, Chapter 3. (Springer)