Analisi Multivariata dei Dati Sperimentali

FACOLTA' DI MEDICINA E CHIRURGIA

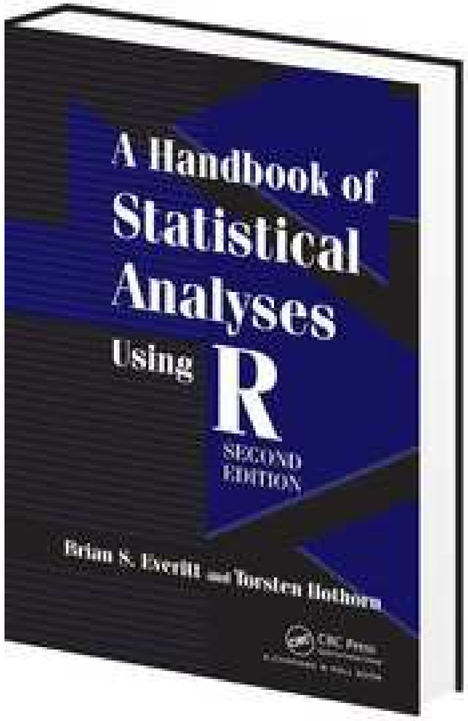
UNIVERSITA' DEGLI STUDI DI TRIESTE

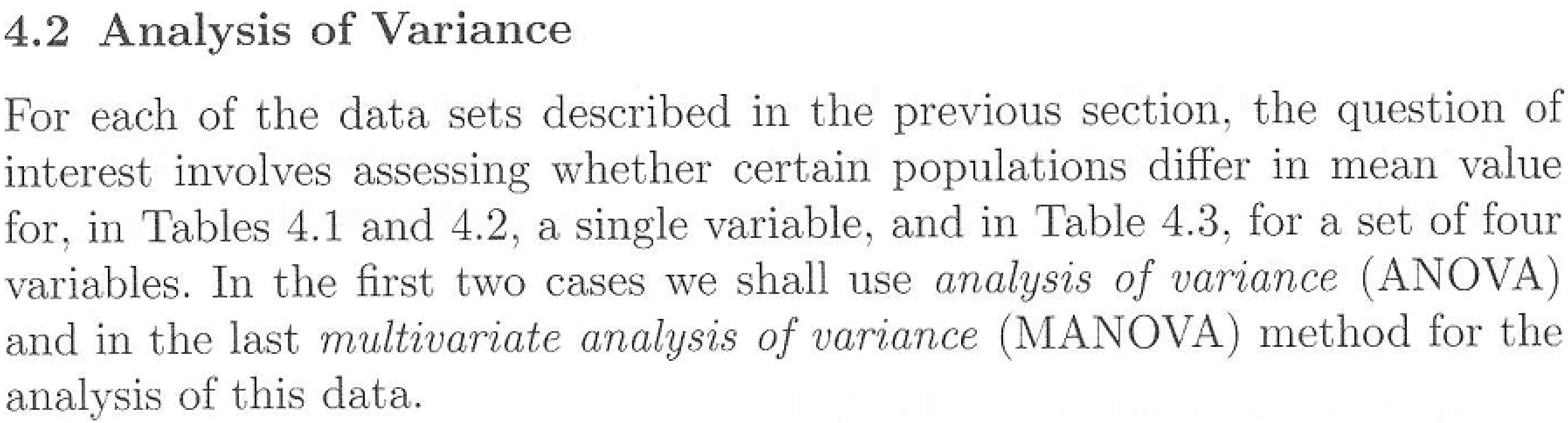
Massimo Borelli

DMI

la "anova"

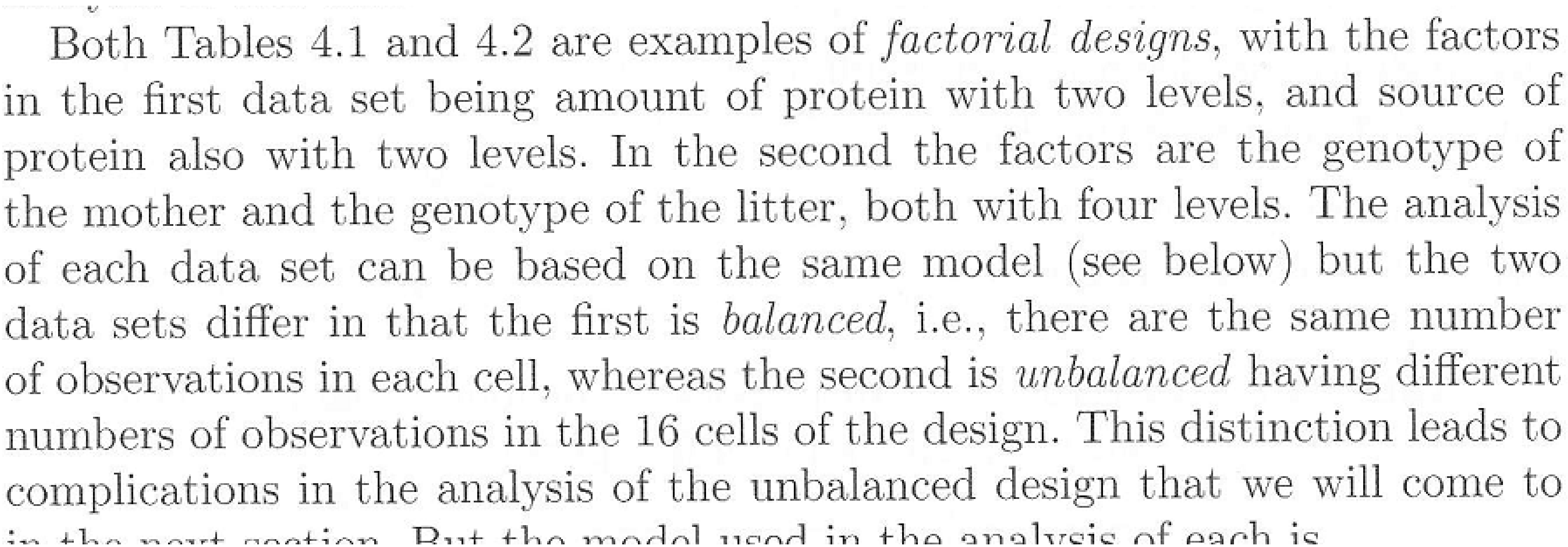
- intro teorica



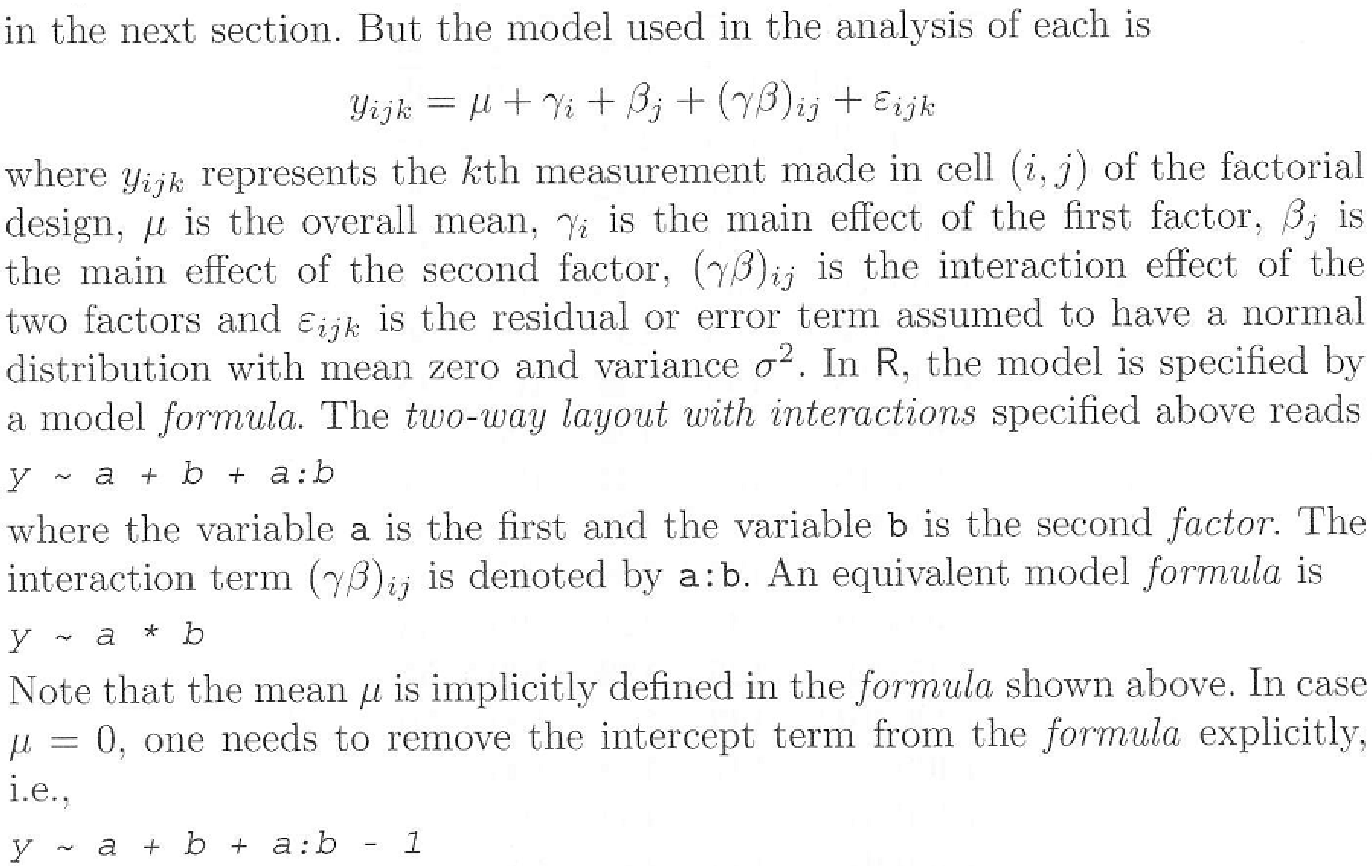


4.2 Analysis of Variance

For each of the data sets described in the previous section, the question of interest involves assessing whether certain populations differ in mean value for, in Tables 4.1 and 4.2, a single variable, and in Table 4.3, for a set of four variables. In the first two cases we shall use *analysis of variance* (ANOVA) and in the last *multivariate analysis of variance* (MANOVA) method for the analysis of this data.



Both Tables 4.1 and 4.2 are examples of *factorial designs*, with the factors in the first data set being amount of protein with two levels, and source of protein also with two levels. In the second the factors are the genotype of the mother and the genotype of the litter, both with four levels. The analysis of each data set can be based on the same model (see below) but the two data sets differ in that the first is *balanced*, i.e., there are the same number of observations in each cell, whereas the second is *unbalanced* having different numbers of observations in the 16 cells of the design. This distinction leads to complications in the analysis of the unbalanced design that we will come to in the next section. But the model used in the analysis of each is



in the next section. But the model used in the analysis of each is

$$y_{ijk} = \mu + \gamma_i + \beta_j + (\gamma\beta)_{ij} + \varepsilon_{ijk}$$

where y_{ijk} represents the kth measurement made in cell (i, j) of the factorial design, μ is the overall mean, γ_i is the main effect of the first factor, β_j is the main effect of the second factor, $(\gamma\beta)_{ij}$ is the interaction effect of the two factors and ε_{ijk} is the residual or error term assumed to have a normal distribution with mean zero and variance σ^2. In R, the model is specified by a model *formula*. The *two-way layout with interactions* specified above reads

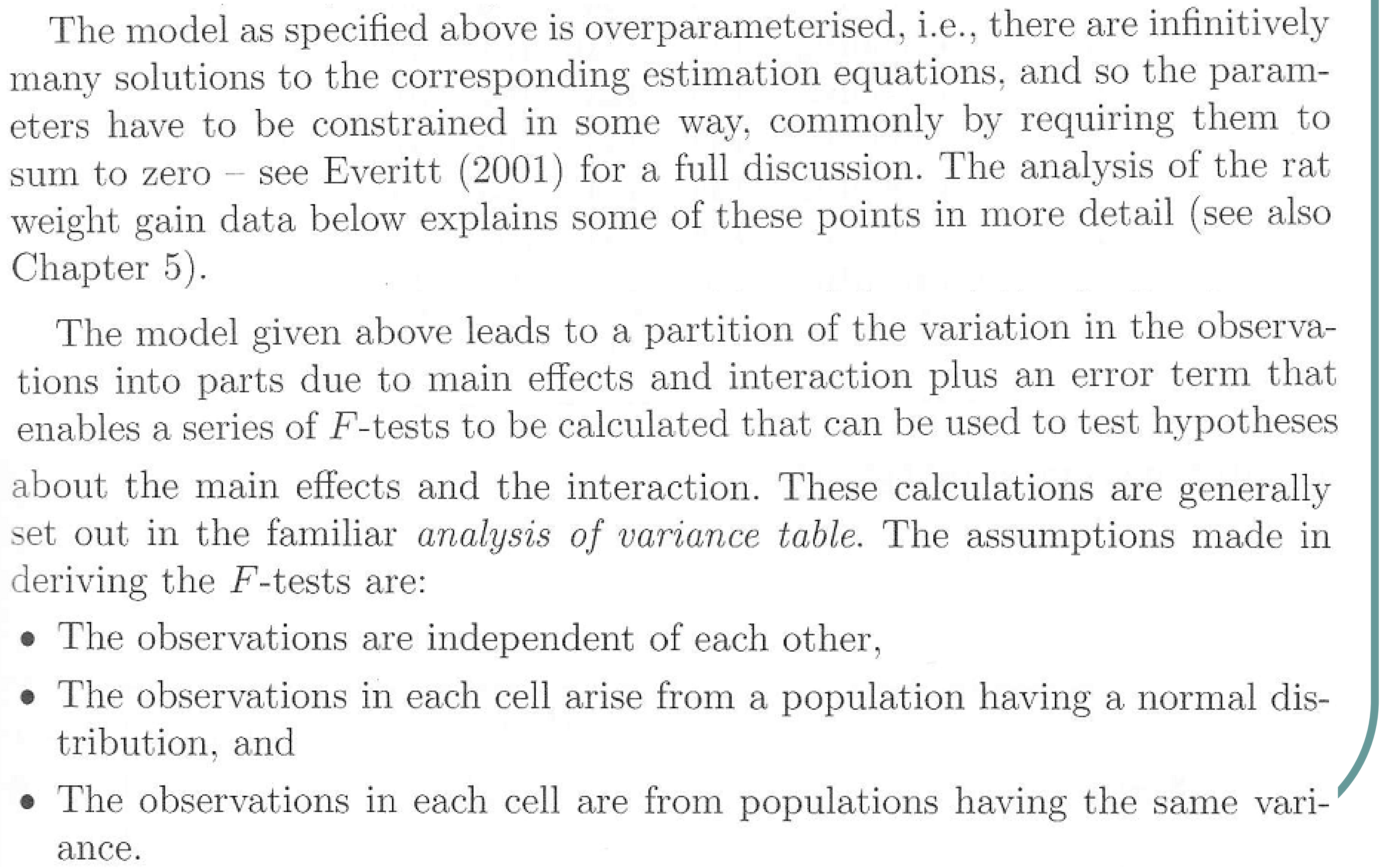
```
y ~ a + b + a:b
```

where the variable a is the first and the variable b is the second *factor*. The interaction term $(\gamma\beta)_{ij}$ is denoted by a:b. An equivalent model *formula* is

```
y ~ a * b
```

Note that the mean μ is implicitly defined in the *formula* shown above. In case $\mu = 0$, one needs to remove the intercept term from the *formula* explicitly, i.e.,

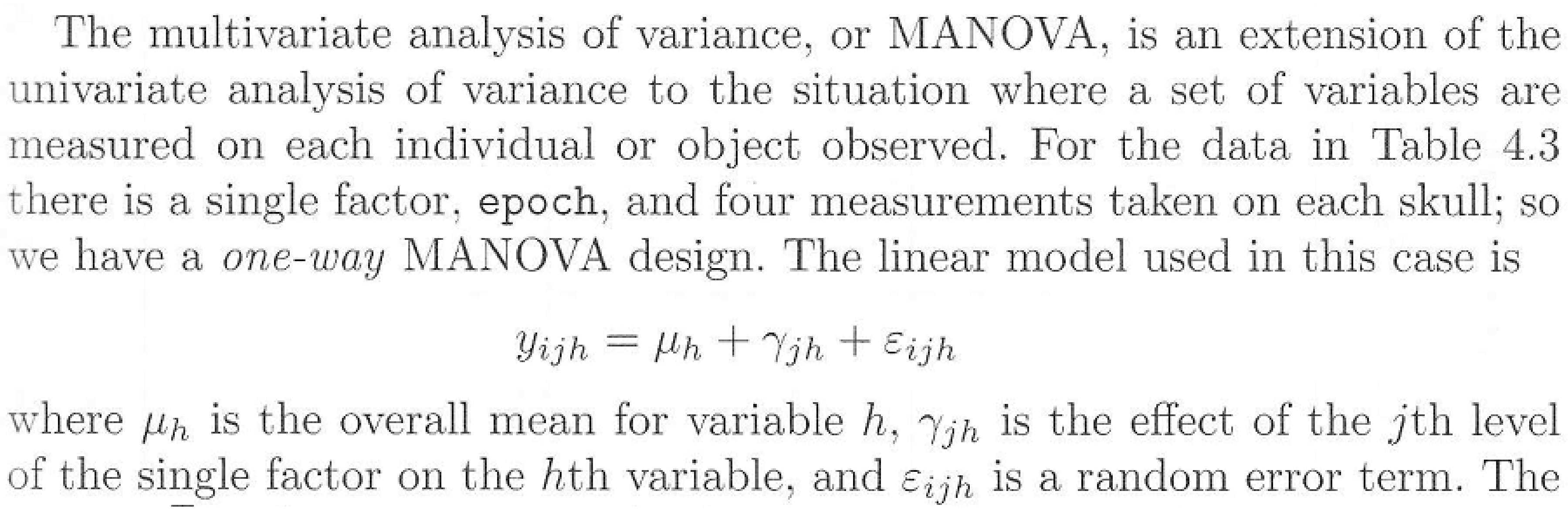
```
y ~ a + b + a:b - 1
```



The model as specified above is overparameterised, i.e., there are infinitively many solutions to the corresponding estimation equations, and so the parameters have to be constrained in some way, commonly by requiring them to sum to zero – see Everitt (2001) for a full discussion. The analysis of the rat weight gain data below explains some of these points in more detail (see also Chapter 5).

The model given above leads to a partition of the variation in the observations into parts due to main effects and interaction plus an error term that enables a series of F-tests to be calculated that can be used to test hypotheses about the main effects and the interaction. These calculations are generally set out in the familiar *analysis of variance table*. The assumptions made in deriving the F-tests are:

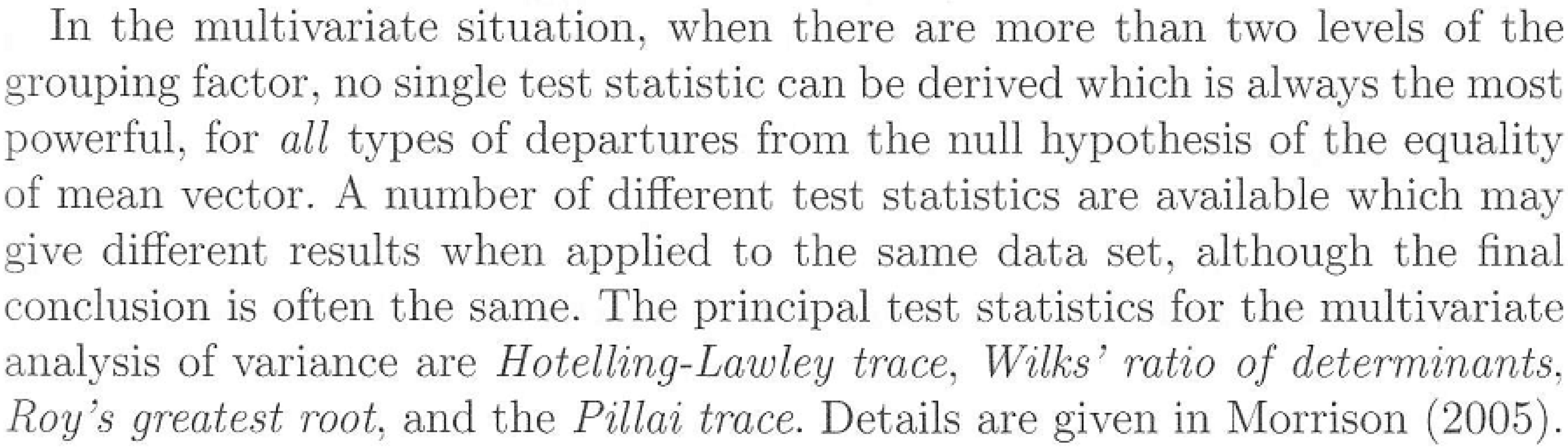
- The observations are independent of each other,
- The observations in each cell arise from a population having a normal distribution, and
- The observations in each cell are from populations having the same variance.



The multivariate analysis of variance, or MANOVA, is an extension of the univariate analysis of variance to the situation where a set of variables are measured on each individual or object observed. For the data in Table 4.3 there is a single factor, `epoch`, and four measurements taken on each skull; so we have a *one-way* MANOVA design. The linear model used in this case is

$$y_{ijh} = \mu_h + \gamma_{jh} + \varepsilon_{ijh}$$

where μ_h is the overall mean for variable h, γ_{jh} is the effect of the jth level of the single factor on the hth variable, and ε_{ijh} is a random error term. The



In the multivariate situation, when there are more than two levels of the grouping factor, no single test statistic can be derived which is always the most powerful, for *all* types of departures from the null hypothesis of the equality of mean vector. A number of different test statistics are available which may give different results when applied to the same data set, although the final conclusion is often the same. The principal test statistics for the multivariate analysis of variance are *Hotelling-Lawley trace*, *Wilks' ratio of determinants*, *Roy's greatest root*, and the *Pillai trace*. Details are given in Morrison (2005).