

## Contrasts for qualitative factors

<http://marekrychlik.com/node/57>

A qualitative factor is a variate with a finite, discrete set of values. A factor can be unordered or ordered.

An example of an unordered factor is eye color in a population sample.

An example of an ordered factor is a student's grade on a standardized test.

An ordered factor does not have to be a quantitative factor, as in this last example. An example of a quantitative factor would be the amount of a fertilizer per acre in a crop yield experiment.

Statistical software, such as R, distinguishes between quantitative and qualitative (ordered and unordered) factors when performing analysis of variance. For example, when dealing with an unordered qualitative factor, R assigns the default set of contrasts to the factor, as generated by the function `contr.treatment()`. These factors are

$$\mu_i - \mu_1$$

where  $i = 2, 3, \dots, t$ , where  $t$  is the number of levels of the factor (typically equal to the number of treatments).

We note that the maximum number of linearly independent factors is  $t - 1$ , due to the requirement of orthogonality to  $\bar{x}$ , the sample mean.

Two contrasts are called orthogonal if they are not correlated as random variables. If we have two contrasts,  $c$  and  $d$ , and

$$c(X) = \sum_{i=1}^t c_i \mu_i$$
$$d(X) = \sum_{i=1}^t d_i \mu_i$$

then, under the assumption of independence of the sample,

$$Cov(c(X), d(X)) = \sum_{i=1}^t \sum_{j=1}^t c_i d_j Cov(\mu_i, \mu_j) = \sum_{i=1}^t c_i d_i Var(\mu_i, \mu_i).$$

If the  $i$ -th treatment group has  $r_i$  units then

$$\mu_i = \frac{1}{r_i} \sum_{j=1}^{r_i} X_{ij}$$

and it is easy to see that

$$Var(\mu_i, \mu_i) = \frac{\sigma^2}{r_i}$$

where  $\sigma^2$  is the population variance.

Thus, we have the following formula for the covariance of the contrasts:

$$Cov(c(X), d(X)) = \sigma^2 \sum_{i=1}^t \frac{c_i d_i}{r_i}.$$

We note that this is a bilinear form of the coefficients. It is positive definite. Moreover, when  $r = r_1 = r_2 = \dots = r_t$  then this form is

$$\frac{\langle c, d \rangle}{r} = \frac{1}{r} c^T d$$

and thus proportional to the standard dot product of vectors  $c^T d$ . Thus, we may identify contrasts with the vectors of their coefficients:

$$c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_t \end{pmatrix}, \quad d = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_t \end{pmatrix}$$

for the purpose of being able to apply linear algebra and geometric intuition in their study.

{The important fact about bilinear forms is that they define the notion of orthogonal. If the treatment groups are not even in size then the notion of orthogonality is non-standard, i.e. different from the one defined by the standard dot product.

{The case of  $t - 1$  mutually orthogonal contrasts is especially important. Under the assumption of normality, this implies independence of the contrasts as random variables. Moreover, in analysis of variance, in this case the sum of squares splits exactly between the contrasts, which is essentially the definition of a balanced design. We note that for unequal treatment groups the default contrasts are not orthogonal, and thus the design is not balanced. In R, an explicit warning is printed if the design is not balanced. There could be other reasons for this warning than the contrasts not being orthogonal.

{The Gram-Schmidt process may be used to fix a set of contrasts which are not orthogonal, but this typically leads to the loss of the clear intuitive form.

{Alternatively, one may evaluate the impact of non-orthogonality on analysis of variance. For small deviations from the equal treatment group condition, the impact will be small, resulting in somewhat higher P-values of the F-test.

{In addition to the default contrasts, R offers two other contrasts for qualitative factors:

- {Helmert contrasts (generated by `contr.helmert`)
- {Sum contrasts (generated by `contr.sum`)

### Helmert contrasts

{These contrasts are used to express the research hypothesis that the  $k + 1$ -th treatment is better than the mean of all treatments from 1 to  $k$ . Thus,

$$c_k(X) = \mu_{k+1} - \frac{1}{k} \sum_{i=1}^k \mu_i \quad \text{for } k = 1, 2, \dots, t - 1,$$

{In R, the  $k$ -th contrast is multiplied by  $k$  to make the coefficients integer.

{Helmert contrasts are orthogonal.

{The following R example yields Helmert contrasts for 4 treatment levels. Column  $k$  yields the coefficients of the  $k$ -th contrast. 

```
gt; contr.helmert(4) [,1] [,2] [,3] 1 -1 -1 -1 2 1 -1 -1 3 0 2 -1 4 0 0 3
```

{They can be used for comparison the  $k$ -th mean with the last one. They are orthogonal.

{The following R example gives the coefficients for the sum contrasts: 

```
gt; contr.sum(4) [,1] [,2] [,3] 1 1 0 0 2 0 1 0 3 0 0 1 4 -1 -1 -1
```

{The statistic

$$t = \frac{c(X)}{s \sqrt{\sum_{i=1}^t \frac{c_i^2}{r_i}}}$$

has the Student t-distribution with  $t - 1$  degrees of freedom (as long as  $c \neq 0$ ). Here  $s$  is the square root of the estimate of the variance:

{

$$s^2 = \frac{1}{N} \sum_{i=1}^t \sum_{j=1}^{r_i} (X_{ij} - \mu_i)^2.$$

h3; Treatment contrasts of R; h3;

{The default setting for R contrasts is somewhat puzzling: 

```
}
```

{nbsp;

{We observe that the sum of the coefficients in each column is not 0. What R calls contrasts, others may call index variables. The bottom line is what the model matrix (which may be called the design matrix by others) is, as this matrix determines which least squares problem is solved. The following R code (which should be put in a file `treatment.R`) illustrates this statement: 

```
}
```

{nbsp;

{We run the program with the following result: 

```
}
```

{This model matrix is consistent with the model:

$$\begin{aligned} y_{1j} &= \mu_1 + e_{ij}, & 1 \leq j \leq r_1, \\ y_{ij} &= \mu_1 + \tau_i + e_{ij}, & 2 \leq i \leq t, \quad 1 \leq j \leq r_i. \end{aligned}$$

{In other words, we use the first mean as the `base`.