```
## Full coefficients are
##
## (Intercept):     5.032
## group:            ctrl   trt1   trt2
##                  0.000 -0.371  0.494
```

We can get the estimated cell means $\widehat{\mu}_i$ with *predict(fit, newdata = data.frame(group = c("ctrl", "trt1", "trt2")))*:

```
##     1     2     3
## 5.032 4.661 5.526
```

The output with *contr.sum* looks as follows:

```
options(contrasts = c("contr.sum", "contr.poly"))
fit2 <- aov(weight ~ group, data = PlantGrowth)
coef(fit2)
```

```
## (Intercept)      group1      group2
##       5.073      -0.041      -0.412
```

Now, *(Intercept)* is the global mean, *group1* the difference of the first (control) group and *group2* the difference of the second group. With *dummy.coef*, the full picture can be retrieved again:

```
## Full coefficients are
##
## (Intercept):     5.073
## group:            ctrl   trt1   trt2
##                  0.041  0.412  0.453
```

Tests

Our null hypothesis is that all groups share the same mean, i.e.:

$$Y_{ij} = \mu + \epsilon_{ij},\ \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$
$$H_0 : \mu_1 = \ldots = \mu_g$$

This is the single mean model and is a special case of the cell means model with $\alpha_1 = \ldots = \alpha_g = 0$. The alternative is therefore:

$$H_A : \mu_k \neq \mu_l \text{ for at least one pair } k \neq l.$$

The total variation of the response around the overall mean can be decomposed into variation "between groups" and variation "within groups":

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{\cdot\cdot})^2}_{SS_T} = \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2}_{SS_{\text{Trt}}} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2}_{SS_E},$$

Where SS_T is the total sum of squares, SS_{Trt} the treatment sum of squares (between groups) and SS_E the error sum of squares (within groups). SS_{Trt} can also be interpreted as the reduction in residual sum of squares when comparing the cell means with the single mean model. This information can be summarized in the ANOVA table (see Appendix). If all the groups share the same (theoretical) mean, we expect the treatment sum of squares to be small. The idea is now to compare the variation between groups with the variation within groups. Under H_0, it holds that:

$$F = \frac{MS_{\text{Trt}}}{MS_E} \sim F_{g-1,\, N-g}$$

Where:

$$MS_{\text{Trt}} = \frac{SS_{Trt}}{g-1}$$

Under H_0, MS_{Trt} is also an estimator for σ^2 and therefore $F = \frac{MS_{Trt}}{MS_E} \approx 1$

We reject the null hypothesis if the observed F value lies in a "extreme" region of the corresponding distribution, more precise we reject H_0 in favor of H_A if F is larger than the 95% quantile. The F-test is a omnibus test because it compares all group means simultaneously. Increasing the denominator degrees of freedom will decrease the corresponding quantile (which gives more power). In R, *summary(fit)* gives the ANOVA table and p-value. As the global test can also be interpreted as a test for comparing two different models, namely the cell means and the single means model, there's also another approach. *anova* can be used to compare the two models:

```
## Fit single mean model (1 means global mean):
fit.single <- aov(weight ~ 1, data = PlantGrowth)
## Compare with cell means model:
anova(fit.single, fit)
```

To perform statistical inference for the individual α_i's, *summary.lm(fit)* (for tests; retrieves all the parameters including standard errors) and *confint(fit)* (for confidence intervals) can be used. Interpretation depends on the side-constraint, an example output of *summary.lm(fit)* looks like this:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.3600     0.1965   17.10 1.39e-07
treatmentCommercial   4.1200     0.2779   14.82 4.22e-07
```

Checking Model Assumptions

Statistical inference (p-values, confidence intervals, ...) is only valid if the model assumptions are fulfilled. So far, this means

- are the errors independent?
- are the errors normally distributed?
- is the error variance constant?
- do the errors have mean zero?

The errors ϵ_{ij} can't be observed, but the residuals r_{ij} can be used as estimates:

$$r_{ij} = y_{ij} - \widehat{\mu}_i.$$

In a QQ-plot, the empirical quantiles are plotted against the theoretical quantiles (of a standard normal distribution). We should more or less see a straight line if the distribution assumption is correct. This is done in R with *plot(fit, which = 2)*. If the QQ-plot suggests non-normality, we can try to use a transformation of the response to accommodate this problem.

The Tukey-Anscombe plot plots the residuals r_{ij} vs. the fitted values $\widehat{\mu}_i$. It allows us to check whether the residuals have constant variance and whether the residuals have mean zero. For the one-way ANOVA situation we could also read off the same information from the plot of the data itself and the residuals always have mean zero (per group). In R, the plot is generated by *plot(fit, which = 1)*.

Transformations Affect Interpretation

Whenever we transform the response we implicitly also change the interpretation of the model parameters. Therefore, while it is conceptually attractive to model the problem on an appropriate scale of the response, this typically has the side effect of making interpretation (much) more difficult. For example, if we use the logarithm,

$$\log(Y_{ij}) = \mu + \alpha_i + \epsilon_{ij}$$

all the α_i's (and their estimates) have to be interpreted on the log-scale. For example, if we us *contr.treatment* and we have $\widehat{\alpha}_2 = 1.5$. This means: on the log-scale we estimate that the average value of group 2 is 1.5 larger than the average value of group 1 (additive shift). What about the original scale? We know that $\mathbb{E}[\log(Y_{ij})] = \mu + \alpha_i$, but the expected value on the original scale does (in general) not directly follow the transformation, i.e. $\mathbb{E}[Y_{ij}] \neq e^{\mu+\alpha_i}$. However, we can make a statement about the median. On the log-scale the median is equal to the mean (because we have a symmetric distribution around $\mu + \alpha_i$) Hence,

$$\text{median}(\log(Y_{ij})) = \mu + \alpha_i$$

In contrast to the mean, any quantile directly transforms with a strictly monotone increasing function. As the median is nothing else than the 50%-quantile, we have

$$\text{median}(Y_{ij}) = e^{\mu+\alpha_i}$$

Similarly, for the ratio

$$\frac{\text{median}(Y_{2j})}{\text{median}(Y_{1j})} = \frac{e^{\mu+\alpha_2}}{e^{\mu}} = e^{\alpha_2}$$

Hence, we can make a statement that on the original scale the median of group 2 is $e^{\widehat{\alpha}_2} = e^{1.5} = 4.48$ as large as the median of group 1. This means that additive effects on the log-scale become multiplicative effects on the original scale. Unfortunately, the statement is only about the median and not the mean on the original scale.

If we also consider a confidence intervals for α_2, e.g. $[1.2, 1.8]$, the transformed version $[e^{1.2}, e^{1.8}]$ is a confidence interval for e^{α_2} which is the ratio of medians on the original scale.