The parameter estimates for the intercept and the slope turn out to be (with the standard error in parentheses):

$$\hat{\alpha} = 2.605 \quad (\text{SE } 0.191), \qquad \hat{\beta} = -0.244 \quad (\text{SE } 0.022).$$

We conclude that the estimated effect of a doubling of the mixture proportion is a decrease of 0.244 in optical density. The 95% confidence interval for β is

$$\hat{\beta} \pm 2.306 \cdot \text{SE}(\hat{\beta}) = -0.244 \pm 2.306 \cdot 0.022 = (-0.295, -0.193),$$

where 2.306 is the 97.5% quantile in the t_8 distribution. This means that decreases between 19.3 and 29.5 when the mixture proportion is doubled are in agreement with the data on the 95% confidence level.

The regression model can be used to estimate (or predict) the optical density for a new mixture proportion, say 600. The expected optical density for such a mixture proportion is

$$\hat{\alpha} + \hat{\beta} \cdot \log_2(600) = 2.605 - 0.244 \cdot \log_2(600) = 0.353$$

and the corresponding confidence interval turns out to be (0.276, 0.430). □

5.4 Unpaired samples with different standard deviations

Throughout this chapter we have assumed that the standard deviation is the same for all observations, and the situation with two independent samples with the same standard deviation is a special case of the one-way ANOVA setup. The assumption of variance homogeneity is essential, in particular for the computation of standard errors and confidence intervals. However, in the situation with two unpaired samples it is possible to handle the situation with different standard deviations in the two groups as well.

Assume that the observations $y_1, \ldots, y_n$ are independent and come from two different groups, group 1 and group 2. Both the mean and standard deviation are allowed to vary between groups, so observations from group 1 are assumed to be $N(\mu_1, \sigma_1^2)$ distributed and observations from group 2 are assumed to be $N(\mu_2, \sigma_2^2)$ distributed.

The estimates of the means are unchanged and thus equal to the group sample means,

$$\hat{\mu}_1 = \bar{y}_1, \quad \hat{\mu}_1 = \bar{y}_2.$$

The variance of their difference is

$$\text{Var}(\hat{\mu}_2 - \hat{\mu}_1) = \text{Var}(\hat{\mu}_2) + \text{Var}(\hat{\mu}_1) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

where we have used Infobox 4.2. Replacing the true (population) variances, σ_1^2 and σ_2^2, with their sample estimates, s_1^2 and s_2^2, yields the estimated variance and hence the standard error,

$$\text{SE}(\hat{\mu}_2 - \hat{\mu}_1) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{5.25}$$

In order to construct a confidence interval from (5.22), all we need is a t quantile, but since we do not have a pooled standard deviation it is not obvious how many degrees of freedom to use. It turns out that it is appropriate to use

$$r = \frac{\left(\text{SE}_1^2 + \text{SE}_2^2\right)^2}{\frac{\text{SE}_1^4}{n_1-1} + \frac{\text{SE}_2^4}{n_2-1}} \tag{5.26}$$

degrees of freedom, where $\text{SE}_1 = s_1/\sqrt{n_1}$ and $\text{SE}_2 = s_2/\sqrt{n_2}$.

The number r is not necessarily an integer. The corresponding $1-\alpha$ confidence interval is

$$\hat{\mu}_2 - \hat{\mu}_1 \pm t_{1-\alpha/2,r} \cdot \text{SE}(\hat{\mu}_2 - \hat{\mu}_1).$$

Note that this confidence interval is only approximate, meaning that the coverage is only approximately (not exactly) 95%.

Example 5.11. Parasite counts for salmon (continued from p. 127). We already computed a 95% confidence interval for the difference between expected parasite counts for Ätran and Conon salmon under the assumption that the standard deviation is the same in both groups (Example 5.9, p. 127). If we are not willing to make this assumption, then we could compute the confidence interval based on (5.25) and (5.26) instead.

We get the standard error

$$\text{SE}(\hat{\mu}_2 - \hat{\mu}_1) = \sqrt{\frac{7.28^2}{13} + \frac{5.81^2}{13}} = 2.58,$$

exactly as in Example 5.9 because there are 13 fish in both samples. For the degrees of freedom we get $\text{SE}_1 = 7.28/\sqrt{13} = 2.019$, $\text{SE}_2 = 5.81/\sqrt{13} = 1.161$, and $r = 22.9$. The 97.5% quantile in t_{23} is 2.069, so the 95% confidence interval for $\alpha_{\text{Ätran}} - \alpha_{\text{Conon}}$ becomes

$$10.69 \pm 2.069 \cdot 2.58 = (5.35, 16.04),$$

almost the same as in Example 5.9, where the standard deviation was assumed to be the same for the two stocks. □

For the salmon data, the two confidence intervals (assuming equal standard deviations or not) were almost identical. This is so because there are the same number of observations in the two groups and because the sample standard deviations computed from the samples separately were close to each other. In other cases there may be a substantial difference between the two confidence intervals.

If the group standard deviations are close, we usually prefer to use the confidence interval based on the assumption of equal standard deviations, mainly because the estimate of the standard deviation is more precise, as it is based on all observations. The results are quite robust as long as the samples are roughly of the same size and not too small (Zar, 1999). Larger differences between the group standard deviations indicate that the assumption of equal standard deviations is not reasonable; hence we would rather use the confidence interval from the present section. In Case 3, Part II (p. 433) we will see an extreme example of such data.

Example 5.12. Vitamin A intake and BMR (continued from p. 84). Figure 4.9 showed histograms for men and women of the BMR variable, related to the basal metabolic rate. We concluded that the normal distribution was adequate to describe the data for each of the samples. The distribution for the sample of men seems to be slightly wider than for the sample of women, so if we want to estimate the difference in BMR between men and women we may want to allow for different standard deviations.

It turns out that the means of the BMR variable are 7.386 for men and 5.747 for women, respectively, whereas the standard deviations are 0.723 and 0.498. Hence the difference in expected BMR is estimated to

$$\hat{\alpha}_{\text{men}} - \hat{\alpha}_{\text{women}} = 7.386 - 5.747 = 1.639$$

and the corresponding standard error is

$$\text{SE}(\hat{\alpha}_{\text{men}} - \hat{\alpha}_{\text{women}}) = \sqrt{\frac{0.723^2}{1079} + \frac{0.498^2}{1145}} = 0.0265$$

because 1079 men and 1145 women participated in the study. Inserting into formula (5.26) yields $r = 1899.61$ so the relevant quantile becomes 1.961. This is of course close to the standard normal quantile because of the large number of observations. The 95% confidence interval becomes $(1.589, 1.692)$, so deviations in expected BMR between men and women in this interval are in accordance with the data. Notice that the confidence interval is very narrow. Again, this is due to the large samples, which imply that expected values are estimated with a large precision. □

5.5 R

As already mentioned in Sections 2.5 and 3.7.1, the `lm()` function is used to fit statistical models based on the normal distribution. Consider the situation with a response variable y and a single explanatory variable x, which