Solution to Exercise 2.4 (Version 1, 16/09/14)

from Statistical Methods in Biology: Design & Analysis of Experiments and Regression (2014) S.J. Welham, S.A. Gezan, S.J. Clark & A. Mead. Chapman & Hall/CRC Press, Boca Raton, Florida. ISBN: 978-1-4398-0878-8

Exercise 2.4*

A soil scientist sampled two fields to get background measurements of carbon biomass (measured as mg C per kg of soil) prior to a field experiment. Six samples were taken from each field: the samples from the first field gave 910, 1058, 929, 1103, 1056, 1022 mg C kg^{-1}; the samples from the second field gave 1255, 1121, 1111, 1192, 1074, 1415 mg C kg^{-1}. File CARBON.DAT contains the unit number (*Sample*), field number (factor Field) and carbon biomass measurement (variate *Carbon*) for each sample. Use a two-sided two-sample t-test to test whether there is any difference in average biomass between the two fields, and calculate a 95% confidence interval for the difference.

Solution 2.4

This data set consists of $N = 12$ samples, with $n = 6$ samples per field. We will denote the k^{th} measure of carbon biomass in the first and second fields as y_{1k} and y_{2k}, respectively, where $k = 1 \ldots 6$. The null hypothesis, of no evidence for a difference between population mean carbon biomass contents in each field, for the two-sample t-test is H_0: $\mu_1 = \mu_2$, and this is to be tested against a two-sided alternative hypothesis, i.e. H_1: $\mu_1 \neq \mu_2$.

We use Table S2.4.1 to help with the calculations.

Table S.2.4.1. Carbon measurements for each field, with deviations from field sample means (Deviation) and squared deviations.

	Field 1			Field 2		
Sample	Carbon	Deviation	Deviation^2	Carbon	Deviation	Deviation^2
1	910	-103.0	10609.00	1255	60.3	3640.11
2	1058	45.0	2025.00	1121	-73.7	5426.78
3	929	-84.0	7056.00	1111	-83.7	7000.11
4	1103	90.0	8100.00	1192	-2.7	7.11
5	1056	43.0	1849.00	1074	-120.7	14560.44
6	1022	9.0	81.00	1415	220.3	48546.78
Sum	6078	0.0	29720.00	7168	0.0	79181.33

The sample means for first and second fields, $\bar{y}_{1\bullet}$ and $\bar{y}_{2\bullet}$, respectively, are

$$\bar{y}_{1\bullet} = \frac{1}{n}\sum_{k=1}^{n} y_{1k} = \frac{6078}{6} = 1013.0, \text{ and}$$

$$\bar{y}_{2\bullet} = \frac{1}{n}\sum_{k=1}^{n} y_{2k} = \frac{7168}{6} = 1194.7\,.$$

The unbiased sample variances for first and second fields, s_1^2 and s_2^2, respectively, are calculated from the sums of the squared deviations about the field sample means as

$$s_1^2 = \frac{1}{(n-1)}\sum_{k=1}^{n_1}\left(y_{1k} - \bar{y}_{1\bullet}\right)^2 = \frac{29720.00}{5} = 5944.00,$$

and

$$s_2^2 = \frac{1}{(n-1)}\sum_{k=1}^{n_2}\left(y_{2k} - \bar{y}_{2\bullet}\right)^2 = \frac{79181.33}{5} = 15836.27.$$

The pooled estimate of variance is

$$s_p^2 = \frac{(n-1)\times s_1^2 + (n-1)\times s_2^2}{n+n-2} = \frac{(6-1)\times 5944.00 + (6-1)\times 15836.27}{6+6-2} = 10890.13\,.$$

The t statistic is then calculated as

$$\mathrm{t} = \frac{\bar{y}_{1\bullet} - \bar{y}_{2\bullet}}{\sqrt{s_p^2 \times \left(\frac{1}{n}+\frac{1}{n}\right)}} = \frac{1013.0 - 1194.7}{\sqrt{10890 \times \left(\frac{1}{6}+\frac{1}{6}\right)}} = \frac{1013.0 - 1194.7}{60.25} = -3.015\,.$$

This statistic has $n + n - 2 = 10$ df. The 97.5th percentile value for the t-distribution with 10 df is $\mathrm{t}_{10}^{[0.025]} = 2.228$. The absolute value of the observed test statistic (3.015) is larger than this critical value. Hence, we reject the null hypothesis at the 5% ($\alpha = 0.05$) significance level, and conclude that there is evidence to indicate that mean carbon biomass differs between the two fields, with more carbon biomass present in the second field. The observed significance level is $P = 0.013$ (calculated as $P = 2*\mathrm{Prob}(|t_{10}| \geq 3.015)$), which leads us to the same conclusion.

The denominator of the t-statistic is the standard error of the difference (SED) between the two field means, i.e. SED = 60.25. A 95% confidence interval for the difference in population means ($\mu_1 - \mu_2$) is computed as

$$\left(\left(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}\right) - \left(\mathrm{t}_{10}^{[0.025]} \times \mathrm{SED}\right),\ \left(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}\right) + \left(\mathrm{t}_{10}^{[0.025]} \times \mathrm{SED}\right) \right)$$

$$= \left(\left(1013 - 1194.7\right) - \left(2.228 \times 60.25\right),\ \left(1013 - 1194.7\right) + \left(2.228 \times 60.25\right) \right)$$

$$= \left(-315.9\,,\ -47.4 \right).$$

This confidence interval does not contain zero, which concurs with the result of the t-test.