Solution to Exercise 2.5 (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 2.5

In Example 12.1 (Table 12.1 and A.1) we describe an experiment in which several morphological traits were measured on 190 seeds from a line of diploid wheat. Two of the traits measured on each seed were length (mm) and weight (mg). The unit numbers (*DSeed*) and length and weight measurements (variates *Length* and *Weight*) can be found in file TRITICUM.DAT. Produce a scatter plot of these two traits and calculate the unbiased sample variances and covariance between them. Derive their sample correlation coefficient, *r*. Is there evidence of association between these two variables?

Solution 2.5

First, we plot the data (Figure S2.5.1). The relationship between the two variables appears approximately linear so a correlation coefficient should give a meaningful summary of the relationship between them.

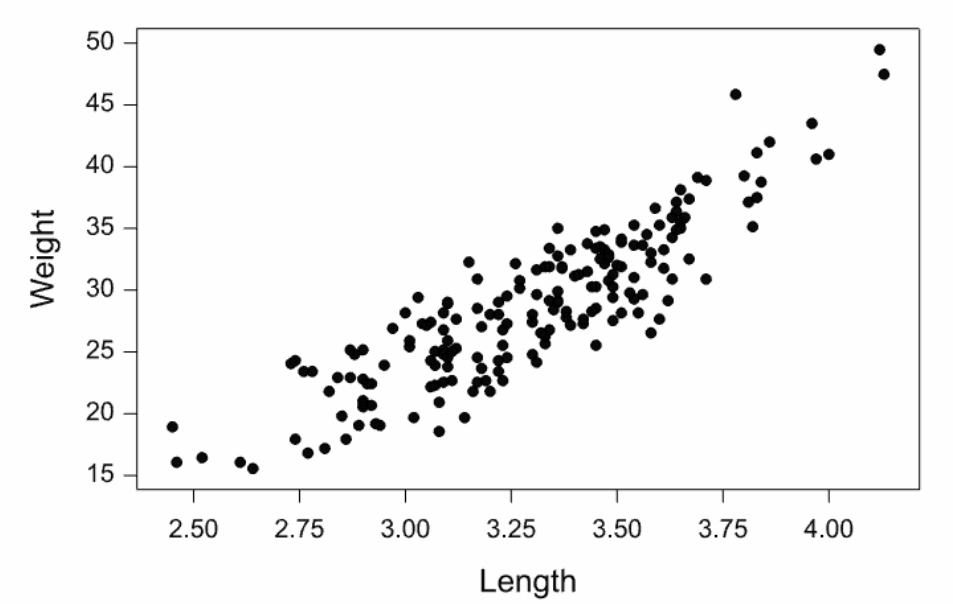

Figure S2.5.1. Scatter plot of seed weights (mg) against seed lengths (mm).

The quantities required for calculation of the sample variances and covariances are illustrated for the first five and last five seeds in Table S2.5.1. We denote the weights as response variable *y* and the lengths as response variable *x*.

Using the totals in Table S2.5.1 we calculate the sample variances for weights and lengths as

$$s_y^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} \left(y_i - \bar{y} \right)^2 = \frac{7294.4090}{189} = 38.5948$$

$$s_x^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} \left(x_i - \bar{x} \right)^2 = \frac{19.2699}{189} = 0.1020.$$

The sample covariance is calculated as

$$s_{xy} = \frac{1}{(N-1)} \sum_{i=1}^{N} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) = \frac{330.9297}{189} = 1.7510.$$

Table S2.5.1 Calculations of quantities required to obtain the sample variances, covariance and correlation coefficient for the seed weights and lengths.

Seed, i	Weight, y_i	Length, x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	30.15	3.27	1.492	−0.025	2.2265	0.0006	−0.0375
2	35.51	3.65	6.852	0.355	46.9521	0.1259	2.4314
3	29.16	3.36	0.502	0.065	0.2522	0.0042	0.0326
4	16.82	2.77	−11.838	−0.525	140.1345	0.2758	6.2167
5	23.42	2.78	−5.238	−0.515	27.4350	0.2654	2.6983
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
186	27.29	3.04	−1.368	−0.255	1.8710	0.0651	0.3490
187	27.66	3.60	−0.998	0.305	0.9957	0.0929	−0.3042
188	26.54	3.58	−2.118	0.285	4.4853	0.0811	−0.6033
189	30.90	3.17	2.242	−0.125	5.0273	0.0157	−0.2806
190	18.94	2.45	−9.718	−0.845	94.4365	0.7143	8.2131
Total	5444.99	626.08	0.000	0.000	7294.4090	19.2699	330.9297

Finally, the sample correlation coefficient is calculated as

$$r = \frac{s_{xy}}{s_x \times s_y} = \frac{1.7510}{\sqrt{38.5948} \times \sqrt{0.1020}} = 0.883.$$

In order to test whether there is statistical evidence of association between the two variables, we compute the test statistic

$$\mathrm{t} = r\sqrt{\frac{N-2}{1-r^2}} = 0.883\sqrt{\frac{188}{1-0.883^2}} = 25.751.$$

The 97.5$^{\text{th}}$ percentile value for the t-distribution with 188 df is $t_{188}^{[0.025]} = 1.973$. As the absolute value of the observed test statistic is (much) larger than this critical value, we reject the null hypothesis at the 5% ($\alpha = 0.05$) significance level and conclude that there is evidence of correlation between the two variables. In fact the observed significance level is $P<0.001$, indicating very strong evidence of correlation between seed weight and length.