Exercises chapter 2

2.5 Exercises

2.1 Genetically similar seeds are randomly assigned to be raised in either a nutritionally enriched environment (treatment group) or standard conditions (control group) using a completely randomized experimental design. After a predetermined time all plants are harvested, dried and weighed. The results, expressed in grams, for 20 plants in each group are shown in Table 2.7.

Table 2.7 *Dried weight of plants grown under two conditions.*

Treatment group		Control group	
4.81	5.36	4.17	4.66
4.17	3.48	3.05	5.58
4.41	4.69	5.18	3.66
3.59	4.44	4.01	4.50
5.87	4.89	6.11	3.90
3.83	4.71	4.10	4.61
6.03	5.48	5.17	5.62
4.98	4.32	3.57	4.53
4.90	5.15	5.33	6.05
5.75	6.34	5.59	5.14

We want to test whether there is any difference in yield between the two groups. Let Y_{jk} denote the kth observation in the jth group where $j = 1$ for the treatment group, $j = 2$ for the control group and $k = 1, \ldots, 20$ for both groups. Assume that the Y_{jk}'s are independent random variables with $Y_{jk} \sim \mathrm{N}(\mu_j, \sigma^2)$. The null hypothesis $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu$, that there is no difference, is to be compared with the alternative hypothesis $\mathrm{H}_1 : \mu_1 \neq \mu_2$.

(a) Conduct an exploratory analysis of the data looking at the distributions for each group (e.g., using dot plots, stem and leaf plots or Normal probability plots) and calculate summary statistics (e.g., means, medians, standard derivations, maxima and minima). What can you infer from these investigations?

(b) Perform an unpaired t-test on these data and calculate a 95% confidence interval for the difference between the group means. Interpret these results.

(c) The following models can be used to test the null hypothesis H_0 against the alternative hypothesis H_1, where

$$\begin{aligned} \mathrm{H}_0 : \mathrm{E}(Y_{jk}) &= \mu; \quad Y_{jk} \sim \mathrm{N}(\mu, \sigma^2), \\ \mathrm{H}_1 : \mathrm{E}(Y_{jk}) &= \mu_j; \quad Y_{jk} \sim \mathrm{N}(\mu_j, \sigma^2), \end{aligned}$$

for $j = 1, 2$ and $k = 1, \ldots, 20$. Find the maximum likelihood and least squares estimates of the parameters μ, μ_1 and μ_2, assuming σ^2 is a known constant.

(d) Show that the minimum values of the least squares criteria are

$$\text{for } \mathrm{H}_0,\ \widehat{S}_0 = \sum\sum(Y_{jk} - \overline{Y})^2, \text{ where } \overline{Y} = \sum_{j=1}^{2}\sum_{k=1}^{K} Y_{jk}/40;$$

$$\text{for } \mathrm{H}_1,\ \widehat{S}_1 = \sum\sum(Y_{jk} - \overline{Y}_j)^2, \text{ where } \overline{Y}_j = \sum_{k=1}^{K} Y_{jk}/20$$

for $j = 1, 2$.

(e) Using the results of Exercise 1.4 show that

$$\frac{1}{\sigma^2}\widehat{S}_1 = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{k=1}^{20}(Y_{jk} - \mu_j)^2 - \frac{20}{\sigma^2}\sum_{k=1}^{20}(\overline{Y}_j - \mu_j)^2,$$

and deduce that if H_1 is true

$$\frac{1}{\sigma^2}\widehat{S}_1 \sim \chi^2(38).$$

Similarly show that

$$\frac{1}{\sigma^2}\widehat{S}_0 = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{k=1}^{20}(Y_{jk} - \mu)^2 - \frac{40}{\sigma^2}\sum_{j=1}^{2}(\overline{Y} - \mu)^2$$

and if H_0 is true then

$$\frac{1}{\sigma^2}\widehat{S}_0 \sim \chi^2(39).$$

(f) Use an argument similar to the one in Example 2.2.2 and the results from (e) to deduce that the statistic

$$F = \frac{\widehat{S}_0 - \widehat{S}_1}{\widehat{S}_1/38}$$

has the central F-distribution $F(1, 38)$ if H_0 is true and a non-central distribution if H_0 is not true.

(g) Calculate the F-statistic from (f). and use it to test H_0 against H_1. What do you conclude?

(h) Compare the value of F-statistic from (g) with the t-statistic from (b), recalling the relationship between the t-distribution and the F-distribution (see Section 1.4.4) Also compare the conclusions from (b) and (g).

(i) Calculate residuals from the model for H_0 and use them to explore the distributional assumptions.

2.2 The weights, in kilograms, of twenty men before and after participation in a "waist loss" program are shown in Table 2.8 (Egger et al. 1999). We want to know if, on average, they retain a weight loss twelve months after the program.

Table 2.8 *Weights of twenty men before and after participation in a "waist loss" program.*

Man	Before	After	Man	Before	After
1	100.8	97.0	11	105.0	105.0
2	102.0	107.5	12	85.0	82.4
3	105.9	97.0	13	107.2	98.2
4	108.0	108.0	14	80.0	83.6
5	92.0	84.0	15	115.1	115.0
6	116.7	111.5	16	103.5	103.0
7	110.2	102.5	17	82.0	80.0
8	135.0	127.5	18	101.5	101.5
9	123.5	118.5	19	103.5	102.6
10	95.0	94.2	20	93.0	93.0

Let Y_{jk} denote the weight of the kth man at the jth time, where $j = 1$ before the program and $j = 2$ twelve months later. Assume the Y_{jk}'s are independent random variables with $Y_{jk} \sim \mathrm{N}(\mu_j, \sigma^2)$ for $j = 1, 2$ and $k = 1, \ldots, 20$.

(a) Use an unpaired t-test to test the hypothesis

$$\mathrm{H}_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad \mathrm{H}_1 : \mu_1 \neq \mu_2.$$

(b) Let $D_k = Y_{1k} - Y_{2k}$, for $k = 1, \ldots, 20$. Formulate models for testing H_0 against H_1 using the D_k's. Using analogous methods to Exercise 2.1 above, assuming σ^2 is a known constant, test H_0 against H_1.

(c) The analysis in (b) is a paired t-test which uses the natural relationship between weights of the *same* person before and after the program. Are the conclusions the same from (a) and (b)?

(d) List the assumptions made for (a) and (b). Which analysis is more appropriate for these data?

2.3 For model (2.7) for the data on birthweight and gestational age, using methods similar to those for Exercise 1.4, show

$$\begin{aligned}\widehat{S}_1 &= \sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{jk} - a_j - b_j x_{jk})^2 \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K}[(Y_{jk} - (\alpha_j + \beta_j x_{jk})]^2 - K\sum_{j=1}^{J}(\overline{Y}_j - \alpha_j - \beta_j \overline{x}_j)^2 \\ &- \sum_{j=1}^{J}(b_j - \beta_j)^2(\sum_{k=1}^{K} x_{jk}^2 - K\overline{x}_j^2)\end{aligned}$$

and that the random variables Y_{jk}, $\overline{Y}_j$ and b_j are all independent and have the following distributions

$$\begin{aligned}Y_{jk} &\sim \mathrm{N}(\alpha_j + \beta_j x_{jk}, \sigma^2), \\ \overline{Y}_j &\sim \mathrm{N}(\alpha_j + \beta_j \overline{x}_j, \sigma^2/K), \\ b_j &\sim \mathrm{N}(\beta_j, \sigma^2/(\sum_{k=1}^{K} x_{jk}^2 - K\overline{x}_j^2)).\end{aligned}$$

2.4 Suppose you have the following data

x:	1.0	1.2	1.4	1.6	1.8	2.0
y:	3.15	4.85	6.50	7.20	8.25	16.50

and you want to fit a model with

$$\mathrm{E}(Y) = \ln(\beta_0 + \beta_1 x + \beta_2 x^2).$$

Write this model in the form of (2.13) specifying the vectors $\mathbf{y}$ and $\boldsymbol{\beta}$ and the matrix $\mathbf{X}$.

2.5 The model for two-factor analysis of variance with two levels of one factor, three levels of the other and no replication is

$$\mathrm{E}(Y_{jk}) = \mu_{jk} = \mu + \alpha_j + \beta_k; \qquad Y_{jk} \sim \mathrm{N}(\mu_{jk}, \sigma^2),$$

where $j = 1, 2$; $k = 1, 2, 3$ and, using the sum-to-zero constraints, $\alpha_1 + \alpha_2 = 0, \beta_1 + \beta_2 + \beta_3 = 0$. Also the Y_{jk}'s are assumed to be independent.

Write the equation for $\mathrm{E}(Y_{jk})$ in matrix notation. (Hint: Let $\alpha_2 = -\alpha_1$, and $\beta_3 = -\beta_1 - \beta_2$).