Solutions Manual to

CLINICAL STATISTICS:

Introducing Clinical Trials,

Survival Analysis,

and

Longitudinal Data Analysis

OLGA KOROSTELEVA

Department of Mathematics and Statistics

California State University, Long Beach

Chapter 2

Section 2.1

EXERCISE 2.1 Equations (2.1) and (2.2) imply the system

$$\begin{cases} k = \Phi^{-1}(1 - \alpha) \\ k - \dfrac{\delta}{\sigma\sqrt{2/n}} = \Phi^{-1}(\beta)\,. \end{cases}$$

Therefore,

$$\frac{\delta}{\sigma\sqrt{2/n}} = \Phi^{-1}(1 - \alpha) - \Phi^{-1}(\beta)\,.$$

Expressing n, arrive at (2.3),

$$n = 2\,(\sigma/\delta)^2 \Big(\Phi^{-1}(1 - \alpha) - \Phi^{-1}(\beta)\Big)^2\,.$$

EXERCISE 2.2 In this exercise, $H_0 : \mu_A = \mu_B$ is tested against $H_1 : \mu_A \neq \mu_B$, $\alpha = 0.05$, $\beta = 0.15$, $\delta = 7$, and $\sigma = 16$. Denote by $\bar{x}_A$ and $\bar{x}_B$ the sample mean responses for group A and group B, respectively. Under H_0, the distribution of $\bar{x}_A - \bar{x}_B$ is $\mathcal{N}\big(0,\, 2\sigma^2/n\big)$, and under a specific alternative $H_1 : \mu_A - \mu_B = \delta$, the distribution is $\mathcal{N}\big(\delta, 2\sigma^2/n\big)$. The acceptance region for the test is

$$\Big\{ -k < \frac{\bar{x}_A - \bar{x}_B}{\sigma\sqrt{2/n}} < k \Big\} = \Big\{ -k\,\sigma\sqrt{2/n} < \bar{x}_A - \bar{x}_B < k\,\sigma\sqrt{2/n} \Big\}\,,$$

where the critical value $k > 0$. The equations for α and β are of the form

$$1 - \alpha = \mathbb{P}\left(-k < \frac{\bar{x}_A - \bar{x}_B}{\sigma\sqrt{2/n}} < k \;\middle|\; \frac{\bar{x}_A - \bar{x}_B}{\sigma\sqrt{2/n}} \sim \mathcal{N}(0, 1) \right)$$

$$= \mathbb{P}\big(\,\big|\,Z\,\big| \,<\, k\big), \;\text{ where }\; Z \sim \mathcal{N}(0,1)\,,$$

and

$$\beta \,=\, \mathbb{P}\left(-\,k\,\sigma\,\sqrt{2/n} \,<\, \bar{x}_A \,-\, \bar{x}_B \,<\, k\,\sigma\,\sqrt{2/n}\,\middle|\, \bar{x}_A \,-\, \bar{x}_B \sim \mathcal{N}(\delta,\, 2\sigma^2/n)\right)$$

$$= \mathbb{P}\left(\left|\,Z \,+\, \frac{\delta}{\sigma\sqrt{2/n}}\right| \,<\, k\right), \;\text{ where }\; Z \,\sim\, \mathcal{N}(0,1)\,.$$

From here,

$$k \,=\, \Phi^{-1}\big(1 - \alpha/2\big)\,,$$

and

$$\beta \,=\, \Phi\Big(k \,-\, \frac{\delta}{\sigma\,\sqrt{2/n}}\Big) \,-\, \Phi\Big(\,-\,k \,-\, \frac{\delta}{\sigma\,\sqrt{2/n}}\Big)\,.$$

For $\alpha \,=\, 0.05$ and $\beta \,=\, 0.15$, $\delta \,=\, 7$, and $\sigma \,=\, 16$, the numerical solution is $k \,=\, 1.96$ and $n \,\geq\, 93.82$. Thus, in practice, $n \,=\, 94$, which corresponds to $\beta \,=\, 0.1493$.

Exercise 2.3 Take $X \,\sim\, \mathit{Poisson}(\lambda)$, and assume that $H_0 \,:\, \lambda \,\geq\, \lambda_0$ is tested against $H_1 \,:\, \lambda \,<\, \lambda_0$ for some λ_0. To compute the likelihood ratio

$$\Lambda(x) \,=\, \frac{\displaystyle\max_{\lambda\,\geq\,\lambda_0} \lambda^x\, e^{-\lambda}/x!}{\displaystyle\max_{\lambda\,>\,0} \lambda^x\, e^{-\lambda}/x!}$$

consider the case $x \,\geq\, \lambda_0$. The maximum likelihood estimator of λ is x, therefore,

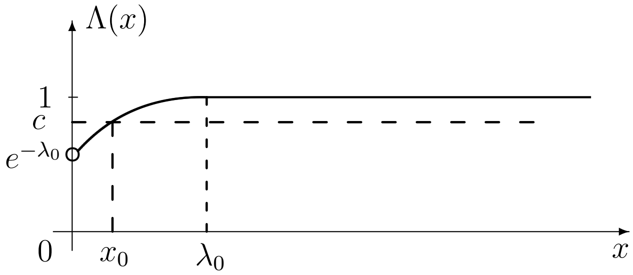
$$\Lambda(x) \,=\, \frac{x^x\, e^{-x}/x!}{x^x\, e^{-x}/x!} \,=\, 1\,.$$

Consider the case $x \,<\, \lambda_0$. Since when $\lambda \,\geq\, x$ the function $\lambda^x\, e^{-\lambda}/x!$ is strictly decreasing, the maximum for $\lambda \,\geq\, \lambda_0 \,>\, x$ is achieved at $\lambda \,=\, \lambda_0$.

Hence, the likelihood ratio is

$$\Lambda(x) \;=\; \frac{\lambda_0^x\, e^{-\lambda_0}/x!}{x^x\, e^{-x}/x!} \;=\; (\lambda_0/x)^x\, e^{-(\lambda_0 - x)}\,.$$

The acceptance region for the likelihood ratio is given by $\big\{x : \Lambda(x) > c\big\}$ for some constant c. From the graph of $\Lambda(x)$ below, this region is equivalent to $\{x : x > x_0\}$ for some $x_0 > 0$. Since the distribution of X is discrete, x_0 can be assumed integer.



(b) By definition, the probability of type I error equals

$$\alpha \;=\; \max_{\lambda \geq \lambda_0} \mathbb{P}(X \,\leq\, x_0) \;=\; \max_{\lambda \geq \lambda_0} \sum_{i=0}^{x_0} \frac{\lambda^i\, e^{-\lambda}}{i\,!}\,.$$

Consider the function

$$g(\lambda) \;=\; \sum_{i=0}^{x_0} \frac{\lambda^{\,i}\, e^{-\lambda}}{i\,!}\,.$$

Taking the derivative of $g(\lambda)$, get

$$g\,'(\lambda) \;=\; \sum_{i=1}^{x_0} \frac{i\,\lambda^{\,i\,-\,1}\, e^{-\lambda}}{i\,!} \;-\; \sum_{i=0}^{x_0} \frac{\lambda^{\,i}\, e^{-\lambda}}{i\,!}$$

$$=\; \sum_{i=0}^{x_0-1} \frac{\lambda^{\,i}\, e^{-\lambda}}{i\,!} \;-\; \sum_{i=0}^{x_0} \frac{\lambda^{\,i}\, e^{-\lambda}}{i\,!} \;=\; -\,\frac{\lambda^{\,x_0}\, e^{-\lambda}}{x_0\,!} \;<\; 0\,.$$

Thus, $g(\lambda)$ is a strictly decreasing function, and therefore reaches its maximum at the left-most point $\lambda = \lambda_0$.

EXERCISE 2.4 Suppose N_t is a random number of events in the interval $[\,0\,,\,t\,]$. Then $N_t \sim \mathit{Poisson}(\lambda\, t)$. The interarrival times between two events are $\mathit{Exponential}\big(1/\lambda\big)$ random variables. Let T_n be the waiting time for the n-th event. Then T_n is a sum of n independent interarrival times, and therefore has a $\mathit{Gamma}\big(n\,,\,1/\lambda\big)$ distribution with the density

$$f_{T_n}(y) \;=\; \frac{\lambda^n\, y^{n-1}}{\Gamma(n)}\, e^{-\lambda\, y}\,, \;\; y \,>\, 0\,,\; \lambda \,>\, 0\,,$$

where $\Gamma(n) \;=\; \int_0^\infty x^{n-1}\, e^{-x}\, dx$ is the gamma function. Thus,

$$\mathbb{P}(N_t \,>\, n) \;=\; \mathbb{P}(N_t \,\geq\, n+1) \;=\; \mathbb{P}\big(\,(n+1)\text{st arrival occurred before time } t\,\big)$$

$$=\; \mathbb{P}(T_{n+1} < t) \;=\; \int_0^t \frac{\lambda^{n+1}\, y^n}{\Gamma(n+1)}\, e^{-\lambda\, y}\, dy \;=\; \int_0^{\lambda\, t} \frac{u^n}{\Gamma(n+1)}\, e^{-u}\, du\,.$$

Finally, note that n in the definition of the gamma function does not need to be an integer. This proves (2.6).

Also, note that (2.6) is in agreement with the usual definition of the Poisson probability mass function. Indeed,

$$\mathbb{P}(N_t \,=\, n) \;=\; \mathbb{P}(N_t \,>\, n-1) \;-\; \mathbb{P}(N_t \,>\, n)$$

$$=\; \frac{1}{n\,!}\int_0^{\lambda} \big(n\, u^{\,n\,-\,1} \;-\; u^{\,n}\big)\, e^{-u}\, du \;=\; \frac{u^{\,n}}{n\,!}\, e^{-u}\,\Big|_0^{\lambda} \;=\; \frac{\lambda^{\,n}}{n\,!}\, e^{-\lambda}\,.$$

Section 2.2

Subsection 2.2.1

EXERCISE 2.5 In (2.11), writing the probabilities as integrals, obtain

$$0.95 \;=\; \int_{-\infty}^{k} \int_{-\infty}^{\sqrt{2}\,k-x} (2\pi)^{-1}\, e^{-(x^2+y^2)/2}\, dy\, dx\,,$$

and

$$0.25 \;=\; \int_{-\infty}^{k-\sqrt{n^*}} \int_{-\infty}^{\sqrt{2}\,k-2\sqrt{n^*}-x} (2\pi)^{-1}\, e^{-(x^2+y^2)/2}\, dy\, dx\,.$$

The solution of this system is $k \;=\; 1.875$ and $n^* \;=\; 3.029\,$.

EXERCISE 2.6 Consider the m-th test, $m \;=\; 1, \ldots, N$. Denote by $\bar{x}_{tr}^{(i)}$ and $\bar{x}_{c}^{(i)}$ the respective group sample means in the i-th set of $2n$ subjects, $i \;=\; 1,\, \ldots,\, m\,$. Let

$$\bar{x}_{tr} \;=\; \frac{\bar{x}_{tr}^{(1)} \,+\, \ldots \,+\, \bar{x}_{tr}^{(m)}}{m} \quad \text{and} \quad \bar{x}_{c} \;=\; \frac{\bar{x}_{c}^{(1)} \,+\, \ldots \,+\, \bar{x}_{c}^{(m)}}{m}$$

be the respective group sample means in the combined set of $2nm$ subjects. The difference

$$\bar{x}_{tr} \,-\, \bar{x}_{c} \,=\, \frac{\bar{x}_{tr}^{(1)} \,-\, \bar{x}_{c}^{(1)}}{m} \,+\, \ldots \,+\, \frac{\bar{x}_{tr}^{(m)} \,-\, \bar{x}_{c}^{(m)}}{m}$$

is the sum of m independent random variables, which under H_0 have distribution $\mathcal{N}\big(0,\, 2\sigma^2/(m^2\, n)\big)$, and under a specific alternative $H_1 \,:\, \mu_{tr} \,-\, \mu_c \,=\, \delta$, have distribution $\mathcal{N}\big(\delta,\, 2\sigma^2/(m^2\, n)\big)$. Thus, under H_0, $\bar{x}_{tr} - \bar{x}_c \,\sim$

$\mathcal{N}\big(0,\, 2\sigma^2/(m\,n)\big)$, and the acceptance region is

$$\left\{ \bar{x}_{tr} \,-\, \bar{x}_c \,<\, k\,\sigma\sqrt{\frac{2}{m\,n}} \right\} \,=\, \left\{ \big(\bar{x}_{tr}^{(1)} \,-\, \bar{x}_c^{(1)}\big) \,+\, \ldots \,+\, \big(\bar{x}_{tr}^{(m)} \,-\, \bar{x}_c^{(m)}\big) \,<\, \sqrt{m}\,k\,\sigma\sqrt{\frac{2}{n}} \right\}.$$

The equation for α is

$$1-\alpha \,=\, \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \left\{ \big(\bar{x}_{tr}^{(1)} \,-\, \bar{x}_c^{(1)}\big) \,+\, \ldots \,+\, \big(\bar{x}_{tr}^{(m)} \,-\, \bar{x}_c^{(m)}\big) \,<\, \sqrt{m}\,k\,\sigma\sqrt{\frac{2}{n}} \right\} \right),$$

where $\bar{x}_{tr}^{(i)} \,-\, \bar{x}_c^{(i)}$, $i \,=\, 1, \ldots, N$, are independent $\mathcal{N}\big(0,\, 2\sigma^2/n\big)$ random variables,

$$=\, \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \Big\{ Z_1 \,+\, \ldots \,+\, Z_m \,<\, \sqrt{m}\,k \Big\} \right),$$

where

$$Z_i \,=\, \frac{\bar{x}_{tr}^{(i)} \,-\, \bar{x}_c^{(i)}}{\sigma\,\sqrt{2/n}},\; i \,=\, 1, \ldots, N,$$

are independent $\mathcal{N}(0,1)$ random variables.

The equation for β is

$$\beta \,=\, \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \left\{ \big(\bar{x}_{tr}^{(1)} \,-\, \bar{x}_c^{(1)}\big) \,+\, \ldots \,+\, \big(\bar{x}_{tr}^{(m)} \,-\, \bar{x}_c^{(m)}\big) \,<\, \sqrt{m}\,k\,\sigma\sqrt{\frac{2}{n}} \right\} \right),$$

where $\bar{x}_{tr}^{(i)} \,-\, \bar{x}_c^{(i)}$, $i \,=\, 1, \ldots, N$, are independent $\mathcal{N}\big(\delta,\, 2\sigma^2/n\big)$ random variables,

$$=\, \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \left\{ Z_1 \,+\, \ldots \,+\, Z_m \,+\, \frac{m\,\delta}{\sigma\,\sqrt{2/n}} \,<\, \sqrt{m}\,k \right\} \right),$$

where

$$Z_i = \frac{\bar{x}_{tr}^{(i)} - \bar{x}_c^{(i)} - \delta}{\sigma\sqrt{2/n}}\,,\; i = 1, \ldots, N\,,$$

are independent $\mathcal{N}(0,1)$ random variables. These equations are equivalent to (2.12) with $n^* = (1/2)(\delta/\sigma)^2\, n$.

Exercise 2.7 For $N = 3$, the system (2.12) has the form

$$1 - \alpha = \mathbb{P}\Big(Z_1 < k,\; Z_1 + Z_2 < \sqrt{2}\,k,\; Z_1 + Z_2 + Z_3 < \sqrt{3}\,k \Big)\,,$$

and

$$\beta = \mathbb{P}\Big(Z_1 + \sqrt{n^*} < k,\; Z_1 + Z_2 + 2\sqrt{n^*} < \sqrt{2}k,\; Z_1 + Z_2 + Z_3 + 3\sqrt{n^*} < \sqrt{3}k \Big)\,,$$

where Z_1, Z_2, and Z_3 are independent $\mathcal{N}(0,1)$ random variables. These equations can be written as a system of integral equations

$$1 - \alpha = \int_{-\infty}^{k} \int_{-\infty}^{\sqrt{2}\,k - x} \int_{-\infty}^{\sqrt{3}\,k - x - y} (2\pi)^{-3/2}\, e^{-(x^2+y^2+z^2)/2}\, dz\, dy\, dx\,,$$

and

$$\beta = \int_{-\infty}^{k-\sqrt{n^*}} \int_{-\infty}^{\sqrt{2}\,k - 2\sqrt{n^*} - x} \int_{-\infty}^{\sqrt{3}\,k - 3\sqrt{n^*} - x - y} (2\pi)^{-3/2}\, e^{-(x^2+y^2+z^2)/2}\, dz\, dy\, dx\,.$$

For $\alpha = 0.05$ and $\beta = 0.25$, the numerical solution is $k = 1.992$ $\big($equivalently, $\alpha' = 1 - \Phi(k) = 0.023\big)$, and $n^* = 2.137$. Thus, the interim group size $n \geq 2(\sigma/\delta)^2\, n^* = 38.47$, or $n = 39$. The actual probability of type II error that corresponds to this group size is 0.245.

In this group sequential testing, the first test is conducted at 2.3% significance level with the group size $n = 39$. If the null is rejected, the trial stops. Otherwise, the trial continues until $(39)(2) = 78$ subjects in each group are accrued. The second test is carried out at 2.3% significance level. If the null is rejected, the trial is discontinued. If H_0 is accepted, then more subjects are enrolled. The trial is terminated and the third test at 2.3% significance level is done when the group size reaches $(39)(3) = 117$ subjects.

When $N = 1$, the required group size is 97. For $N = 2$, the maximum size is 110. Thus, for $N = 3$ there is a possibility that the trial continues longer than when $N = 1$ or $N = 2$, but as a trade-off, there are two chances to stop the trial earlier.

Exercise 2.8 In Exercise 2.2, $H_0 : \mu_A = \mu_B$ is tested against $H_1 : \mu_A \neq \mu_B$, $\alpha = 0.05$, $\beta = 0.15$, $\delta = 7$, and $\sigma = 16$. The non-sequential test $(N = 1)$ is conducted with the group size $n = 94$.

Fix any $N \geq 1$, and consider the m-th interim test, $m = 1, \ldots, N$. As in the solution to Exercise 2.6, denote by $\bar{x}_{tr}^{(i)}$ and $\bar{x}_c^{(i)}$ the respective group sample means in the i-th set of $2\,n$ subjects, $i = 1, \ldots, m$. Let

$$\bar{x}_{tr} = \frac{\bar{x}_{tr}^{(1)} + \ldots + \bar{x}_{tr}^{(m)}}{m} \quad \text{and} \quad \bar{x}_c = \frac{\bar{x}_c^{(1)} + \ldots + \bar{x}_c^{(m)}}{m}$$

be the respective group sample means in the combined set of $2nm$ subjects. The difference

$$\bar{x}_{tr} - \bar{x}_c = \frac{\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}}{m} + \ldots + \frac{\bar{x}_{tr}^{(m)} - \bar{x}_c^{(m)}}{m}$$

is the sum of m independent random variables, which under H_0 have distribution $\mathcal{N}\big(0,\, 2\sigma^2/(m^2\, n)\big)$, and under a specific alternative $H_1 \,:\, \mu_{tr} - \mu_c \,=\, \delta$, have distribution $\mathcal{N}\big(\delta,\, 2\sigma^2/(m^2\, n)\big)$. Thus, under H_0, $\bar{x}_{tr} \,-\, \bar{x}_c \sim \mathcal{N}\big(0,\, 2\sigma^2/(m\, n)\big)$, and the acceptance region is

$$\left\{ -\, k\,\sigma\sqrt{\frac{2}{m\,n}} \,<\, \bar{x}_{tr} - \bar{x}_c \,<\, k\,\sigma\sqrt{\frac{2}{m\,n}} \right\}$$

$$= \left\{ -\sqrt{m}\, k\,\sigma\sqrt{\frac{2}{n}} \,<\, \big(\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}\big) \,+\, \ldots \,+\, \big(\bar{x}_{tr}^{(m)} - \bar{x}_c^{(m)}\big) \,<\, \sqrt{m}\, k\,\sigma\sqrt{\frac{2}{n}} \right\}$$

$$= \left\{ \Big|\, \big(\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}\big) \,+\, \ldots \,+\, \big(\bar{x}_{tr}^{(m)} - \bar{x}_c^{(m)}\big) \Big| \,<\, \sqrt{m}\, k\,\sigma\sqrt{\frac{2}{n}} \right\}.$$

The relation between the significance level α' and the critical value of the acceptance region k is given by the formula $k \,=\, \Phi^{-1}\big(1 \,-\, \alpha'/2\big)$, or, equivalently, $\alpha' \,=\, 2\big(1 \,-\, \Phi(k)\big)$.

The equation for α is

$$1 - \alpha \,=\, \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \left\{ \Big|\, \big(\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}\big) \,+\, \ldots \,+\, \big(\bar{x}_{tr}^{(m)} - \bar{x}_c^{(m)}\big) \Big| \,<\, \sqrt{m}\, k\,\sigma\sqrt{\frac{2}{n}} \right\} \right),$$

where $\bar{x}_{tr}^{(i)} \,-\, \bar{x}_c^{(i)}$, $i \,=\, 1, \ldots, N$, are independent $\mathcal{N}\big(0,\, 2\sigma^2/n\big)$ random variables,

$$= \mathbb{P}\Big(\bigcap_{m\,=\,1}^{N} \Big\{ \Big|\, Z_1 \,+\, \ldots \,+\, Z_m \Big| \,<\, \sqrt{m}\, k \Big\} \Big),$$

where

$$Z_i \,=\, \frac{\bar{x}_{tr}^{(i)} \,-\, \bar{x}_c^{(i)}}{\sigma\,\sqrt{2/n}}\,,\; i \,= 1, \ldots, N\,,$$

are independent $\mathcal{N}(0, 1)$ random variables.

The equation for β is

$$\beta \;=\; \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \left\{ \left| \, \left(\bar{x}_{tr}^{(1)} - \bar{x}_{c}^{(1)}\right) \;+\; \ldots \;+\; \left(\bar{x}_{tr}^{(m)} - \bar{x}_{c}^{(m)}\right) \right| \;<\; \sqrt{m}\, k\, \sigma \sqrt{\frac{2}{n}} \right\} \right) \, ,$$

where $\bar{x}_{tr}^{(i)} \;-\; \bar{x}_{c}^{(i)}$, $i \;=\; 1, \ldots, N$, are independent $\mathcal{N}\big(\delta,\, 2\sigma^2/n\big)$ random variables,

$$= \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \left\{ \left| Z_1 + \cdots + Z_m \;+\; \frac{m\,\delta}{\sigma\,\sqrt{2/n}} \right| \;<\; \sqrt{m}\, k \right\} \right) \, ,$$

where

$$Z_i \;=\; \frac{\bar{x}_{tr}^{(i)} \;-\; \bar{x}_{c}^{(i)} \;-\; \delta}{\sigma\,\sqrt{2/n}} \, , \; i \;=\; 1, \ldots, N \, ,$$

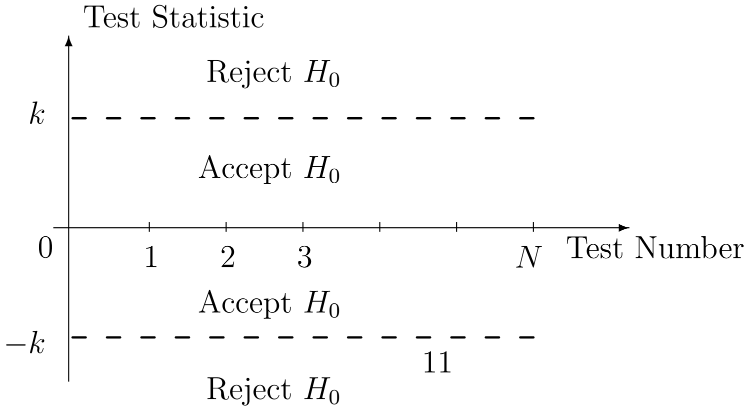
are independent $\mathcal{N}(0,1)$ random variables. Let $n^* \;=\; (1/2)(\delta/\sigma)^2\, n$. Then these equations become

$$1 \;-\; \alpha \;=\; \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \left\{ \left| Z_1 \;+\; \ldots \;+\; Z_m \right| \;<\; \sqrt{m}\, k \right\} \right) \, ,$$

and

$$\beta \;=\; \mathbb{P}\left(\bigcap_{m\,=\,1}^{N} \left\{ \left| Z_1 \;+\; \ldots \;+\; Z_m \;+\; m\,\sqrt{n^*} \right| \;<\; \sqrt{m}\, k \right\} \right) \, ,$$

where $Z_1, \ldots, Z_N$ are independent $\mathcal{N}(0,1)$ random variables. A schematic plot of the acceptance region for the m-th test is given below.



Consider the case $N = 2$. The equations for α and β are

$$1 - \alpha = \mathbb{P}\Big(\big| Z_1 \big| < k, \ \big| Z_1 + Z_2 \big| < \sqrt{2}\, k \Big) ,$$

and

$$\beta = \mathbb{P}\Big(\big| Z_1 + \sqrt{n^*} \big| < k, \ \big| Z_1 + Z_2 + 2\sqrt{n^*} \big| < \sqrt{2}\, k \Big) ,$$

where Z_1 and Z_2 are independent $\mathcal{N}(0,1)$ random variables. In the integral form these equations are

$$1 - \alpha = \int_{-k}^{k} \int_{-\sqrt{2}\,k - x}^{\sqrt{2}\,k - x} (2\pi)^{-1} \, e^{-(x^2+y^2)/2} \, dy \, dx \, ,$$

and

$$\beta = \int_{-k - \sqrt{n^*}}^{k - \sqrt{n^*}} \int_{-\sqrt{2}\,k - 2\sqrt{n^*} - x}^{\sqrt{2}\,k - 2\sqrt{n^*} - x} (2\pi)^{-1} \, e^{-(x^2+y^2)/2} \, dy \, dx \, .$$

The solution of this system with $\alpha = 0.05$ and $\beta = 0.15$ is $k = 2.178$ $\big($equivalently, $\alpha' = 2\big(1 - \Phi(k)\big) = 0.029\big)$ and $n^* = 4.963$. The interim group sample size is $n \geq 2(16/7)^2(4.963) = 51.86$, hence $n = 52$. The corresponding $\beta = 0.149$.

Consider the case $N = 3$. The equations for α and β are

$$1 - \alpha = \mathbb{P}\Big(\big| Z_1 \big| < k, \ \big| Z_1 + Z_2 \big| < \sqrt{2}\, k, \ \big| Z_1 + Z_2 + Z_3 \big| < \sqrt{3}\, k \Big) ,$$

$$= \int_{-k}^{k} \int_{-\sqrt{2}\,k - x}^{\sqrt{2}\,k - x} \int_{-\sqrt{3}\,k - x - y}^{\sqrt{3}\,k - x - y} (2\pi)^{-3/2}\, e^{-(x^2+y^2+z^2)/2}\, dz\, dy\, dx\,,$$

and

$$\beta = \mathbb{P}\Big(\big| Z_1 + \sqrt{n^*} \big| < k,\ \big| Z_1 + Z_2 + 2\sqrt{n^*}\,\big| < \sqrt{2}\,k\,,$$

$$\big| Z_1 + Z_2 + Z_3 + 3\sqrt{n^*}\,\big| < \sqrt{3}\,k \Big)\,,$$

$$= \int_{-k - \sqrt{n^*}}^{k - \sqrt{n^*}} \int_{-\sqrt{2}\,k - 2\sqrt{n^*} - x}^{\sqrt{2}\,k - 2\sqrt{n^*} - x} \int_{-\sqrt{3}\,k - 3\sqrt{n^*} - x - y}^{\sqrt{3}\,k - 3\sqrt{n^*} - x - y} (2\pi)^{-3/2}\, e^{-(x^2+y^2+z^2)/2}\, dz\, dy\, dx\,.$$

For $\alpha = 0.05$ and $\beta = 0.15$, the solution is $k = 2.289$ (equivalently, $\alpha' = 2\big(1 - \Phi(k)\big) = 0.022$), and $n^* = 3.467$. The interim sample size $n \geq 2(16/7)^2\,(3.467) = 36.23$, hence $n = 37$. The corresponding $\beta = 0.1425$.

For $N = 1$, the group size is 94. For $N = 2$, the maximum group size is $(2)(52) = 104$, and for $N = 3$, it is $(3)(37) = 111$.

Exercise 2.9 (a) A random variable $X_{mt} \sim \mathit{Poisson}(\lambda\, m\, t)$ is a sum of m independent random variables $X_t^{(1)}, \dots, X_t^{(m)} \sim \mathit{Poisson}(\lambda\, t)$. Therefore, the equation for the overall probability of type I error α is

$$\alpha = \mathbb{P}\Big(\bigcap_{m=1}^{N} \Big\{ \frac{X_{mt} - 0.024\, m\, t}{\sqrt{0.024\, m\, t}} \leq k \Big\} \Big)\,, \quad X_{mt} \sim \mathit{Poisson}(0.024\, m\, t)\,,$$

$$= \mathbb{P}\Big(\bigcap_{m=1}^{N} \Big\{ \frac{X_t^{(1)} - 0.024\, t}{\sqrt{0.024\, t}} + \dots + \frac{X_t^{(m)} - 0.024\, t}{\sqrt{0.024\, t}} \leq \sqrt{m}\, k \Big\} \Big)\,.$$

By the Central Limit Theorem, for large t, $Z_i = (X_t^{(i)} - 0.024\, t)/\sqrt{0.024\, t}$ is approximately $\mathcal{N}(0,1)$ random variable, $i = 1, \dots, N$. Hence, an approximate expression for α is

$$\alpha = \mathbb{P}\Big(\bigcap_{m=1}^{N} \Big\{ Z_1 + \ldots + Z_m \le \sqrt{m}\, k \Big\} \Big),$$

where $Z_1, \ldots, Z_N$ are independent $\mathcal{N}(0,1)$ random variables.

(b) The equation for the overall probability of type II error β is

$$1-\beta = \mathbb{P}\left(\bigcap_{m=1}^{N} \left\{ \frac{X_{mt} - 0.024\, m\, t}{\sqrt{0.024\, m\, t}} \le k \right\} \right), \quad X_{mt} \sim Poisson(0.012\, m\, t)\,,$$

$$= \mathbb{P}\left(\bigcap_{m=1}^{N} \left\{ \frac{1}{\sqrt{2m}} \big(Z_1 + \ldots + Z_m \big) - \frac{0.012\; mt}{\sqrt{\;0.024\, m\, t}} \le k \right\} \right)$$

$$= \mathbb{P}\left(\bigcap_{m=1}^{N} \left\{ Z_1 + \ldots + Z_m \le \sqrt{2\, m}\, k + m \sqrt{0.012\, t} \right\} \right),$$

where $Z_i = (X_t^{(i)} - 0.012\, t)/\sqrt{0.012\, t}$ are independent approximately $\mathcal{N}(0,1)$ random variables, $X_t^{(i)} \sim Poisson(0.012\, t)$, $i = 1, \ldots, N$.

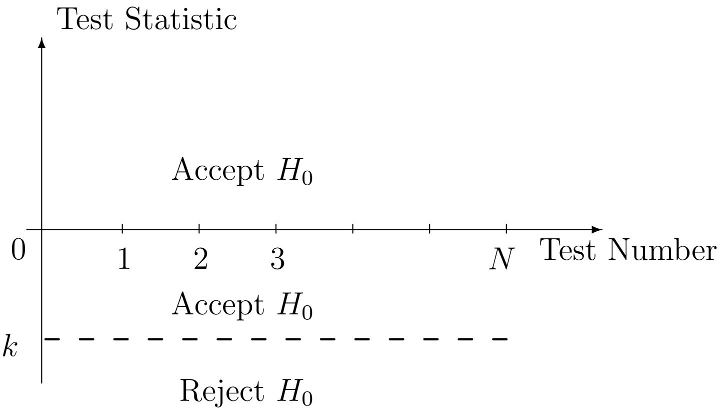
(c) For $N = 2$, $\alpha = 0.05$, and $\beta = 0.2$, the numerical solution is $k = -1.2571$ and $t = 577.67$. Thus, the first test is conducted at $t = 577.67$ (roughly 580) patient-years. If the observed number of endocarditis complications exceeds 9 (since $0.024t + k\sqrt{0.024t} = 9.18$), then the alternative is accepted and the trial is stopped. If the number of complications is 9 or less, the trial continues until $2t = 1155.34$ (roughly 1160) patient-years are accumulated. At this point the trial is stopped. If the number of complications exceeds 21 (since $(0.024)\,(2\,t) + k\sqrt{(0.024)\,(2\,t)} = 21.11$), the null is accepted. Otherwise, H_1 is accepted.

(d) For the m-th test,

$$\alpha' = \mathbb{P}\Big(Z_1 + \ldots Z_m < \sqrt{m}\,k \,\Big|\, Z_1, \ldots, Z_m \overset{iid}{\sim} \mathcal{N}(0,1) \Big) = \mathbb{P}(Z < k) = \Phi(k)\,.$$

For $N = 2$, $\alpha' = \Phi(-1.2571) = 0.1044$.

(e) The acceptance region for the m-th test, $m = 1, \ldots, N$, is drawn below.



Exercise 2.10 (a) The acceptance region for the first test is $\{ X_t > t\,K \}$, and for the second one, $\{ X_{2t} > 2\,t\,K \}$. The random variable $X_{2t} \sim$ *Poisson*$(2\,\lambda\,t)$ can be written as the sum of two independent *Poisson*$(\lambda\,t)$ random variables, X_t and Y_t. Therefore, the equations for α and β, the overall probabilities of type I and II errors, are

$$\alpha = \mathbb{P}\Big(X_t \le K\,t,\; X_t + Y_t \le 2\,K\,t \Big)\,,$$

where X_t and Y_t are independent *Poisson*$(0.024\,t)$ random variables,

$$= \left[\sum_{i=0}^{K\,t} \frac{(0.024\,t)^i}{i\,!}\, e^{-0.024\,t} \right]^2 + \sum_{i=0}^{K\,t} \sum_{j=K\,t+1}^{2\,K\,t-i} \frac{(0.024\,t)^{i+j}}{i\,!j\,!}\, e^{-0.048\,t}\,,$$

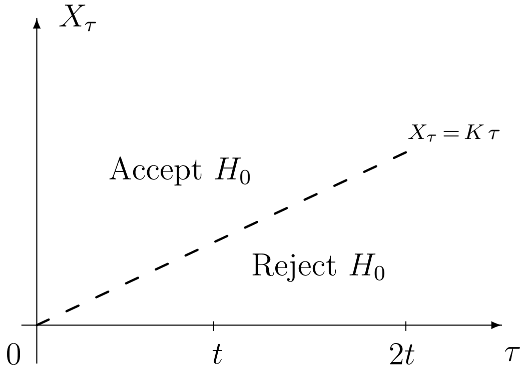
and

$$1-\beta=\mathbb{P}\Big(X_t\le K\,t,\;X_t+Y_t\le 2\,K\,t\Big),$$

where X_t and Y_t are independent $Poisson(0.012\,t)$ random variables,

$$=\left[\sum_{i=0}^{K\,t}\frac{(0.012\,t)^{i}}{i\,!}\,e^{-0.012\,t}\right]^2+\sum_{i=0}^{K\,t}\sum_{j=K\,t+1}^{2\,K\,t-i}\frac{(0.012\,t)^{i+j}}{i\,!j\,!}\,e^{-0.024\,t}.$$

(b) The acceptance regions are drawn below.



(c) The closest values to $\alpha=0.05$ and $\beta=0.2$ are achieved when $t=500$, and $K=0.016$. The actual values of α and β are 0.0401 and 0.1917, respectively.

(d) This sequential testing is carried out the following way. The first test is conducted at $t=500$ patient-years. If $K\,t+1=9$ or more endocarditis cases are observed, then H_0 is accepted and the trial is stopped. If $K\,t=8$ or less cases are observed, the trial continues until a total of $2\,t=1000$ patient-years are accumulated. Then the trial is stopped and the second test is conducted. If $2\,K\,t+1=17$ or more events are recorded, then H_0 is accepted, otherwise, H_1 is accepted.

(e) $\alpha' = \mathbb{P}\big(\text{reject } H_0 \,\big|\, H_0 \text{ is true}\big) = \mathbb{P}\big(X_t \leq Kt\big)$, where $X_t \sim \mathit{Poisson}(0.024t)$. For the first test $\alpha' = \mathbb{P}(X \leq 8)$, where $X \sim \mathit{Poisson}(12)$. Thus, $\alpha' = 0.1550$. For the second test $\alpha' = \mathbb{P}(X \leq 16)$, where $X \sim \mathit{Poisson}(24)$. Hence, $\alpha' = 0.0563$.

Subsection 2.2.2

Exercise 2.11 The mode solves the maximization problem

$$x^{a-1} e^{-x/b} \to \max_x .$$

Setting the derivative equal to zero, obtain

$$(a-1)\, x^{a-2} - \frac{x^{a-1}}{b} = 0,$$

hence, $x = (a-1)\, b$.

Exercise 2.12 By Bayes' formula, the posterior distribution of R equals

$$f_R(x \,|\, n,\, t) = C\, f(n \,|\, x,\, t)\, \pi(x)$$

$$= C_1\, (x\, t)^n\, e^{-x t}\, x^{a-1}\, e^{-x/b} = C_2\, x^{n+a-1}\, e^{-x\,(t+1/b)},$$

where C, C_1 and C_2 are the normalizing constants. Therefore, the posterior distribution of R is $\mathit{Gamma}\big(n + a,\, 1/(t + 1/b)\big)$.

Exercise 2.13 The posterior probability of H_1 is computed according to (2.17), where a and b satisfy (2.14) and (2.16) with $\mathbb{P}(R < 0.024) = 0.7$.

The stopping rules for $t = 400$ and $t = 600$ patient-years are given in the table below.

t	n	$\mathbb{P}(H_1 \mid n,t)$	t	n	$\mathbb{P}(H_1 \mid n,t)$
400	**5**	**0.9582**	400	15	0.0678
	6	0.9109		**16**	**0.0388**
600	**8**	**0.9726**	600	21	0.0627
	9	0.9463		**22**	**0.0391**

To stop the trial at 400 patient-years, 5 (or fewer) or 16 (or more) endocarditis cases should be observed. If between 6 and 15 cases occur, then the trial continues until 600 patient-years are accumulated. At this time, if between 6 and 8, or 22 or more events have been recorded, then the trial is terminated.

To compare these rules with the ones given in Table 2.1, notice that H_1 is accepted even with a larger number of observed complications. However, at 400 patient-years, the alternative may be rejected with only 16 cases. There is no difference in the rules for rejecting H_1 at 600 patient-years.

Exercise 2.14 (a) The prior distribution of μ is $\mathcal{N}(0, \sigma^2)$. Therefore,

$$\mathbb{P}(H_1) = \mathbb{P}(\mu > 0) = \mathbb{P}(Z > 0) = 0.5 .$$

(b) The posterior density of μ, given $\bar{x}$, is

$$f_\mu(y \mid \bar{x}) = C \exp\left\{ -\frac{(\bar{x} - y)^2}{4\sigma^2/n} - \frac{y^2}{2\sigma^2} \right\}$$

$$= C \exp\left\{ -\frac{1}{4\sigma^2/n} \left[(\bar{x} - y)^2 + \frac{2}{n} y^2 \right] \right\}$$

$$= C_1 \exp\left\{ -\frac{1}{4\sigma^2/n} \left(1 + 2/n\right) \left(y - \frac{\bar{x}}{1 + 2/n}\right)^2 \right\},$$

where C and C_1 are the normalizing constants. This is the normal density with mean $\bar{x}/(1 + 2/n)$ and variance $(2\sigma^2/n)/(1 + 2/n)$.

(c) The alternative is accepted if $\mathbb{P}(H_1) = \mathbb{P}\big(\mu > 0 \,|\, \bar{x}\big) \geq 0.95$. Hence,

$$0.95 \leq \mathbb{P}\left(Z > \frac{-\bar{x}/(1+2/n)}{\sqrt{(2\sigma^2/n)/(1+2/n)}} \right) = \mathbb{P}\left(Z > -\frac{\bar{x}}{\sigma\sqrt{2/n(1+2/n)}} \right).$$

From here, $\bar{x} \geq 1.645\,\sigma\sqrt{2/n(1+2/n)} = 5.0327$. The alternative is rejected if $\mathbb{P}(H_1) = \mathbb{P}\big(\mu > 0 \,|\, \bar{x}\big) \leq 0.05$. Hence,

$$\mathbb{P}\left(Z > -\frac{\bar{x}}{\sigma\sqrt{2/n(1+2/n)}} \right) \leq 0.05,$$

and, therefore, $\bar{x} \leq -1.645\,\sigma\sqrt{2/n(1+2/n)} = -5.0327$. Thus, the stopping rule for the interim Bayesian analysis at $n = 50$ is to stop and reject H_1 if $\bar{x} \leq -5.0327$, to stop and accept H_1 if $\bar{x} \geq 5.0327$, or to continue the trial if $-5.0327 < \bar{x} < 5.0327$.

Section 2.3

Exercise 2.15 Define the function

$$v(n) = \frac{1}{n} + \frac{1}{N - n}.$$

Then $\mathbb{V}ar(\bar{x}_1 - \bar{x}_2) = \sigma^2 v(n)$. To minimize $v(n)$ with respect to n, set the first derivative equal to zero,

$$v'(n) = -\frac{1}{n^2} + \frac{1}{(N - n)^2} = 0\,.$$

Thus $n^2 = (N - n)^2$, or $n = N - n$. Hence $n = N/2$.

Exercise 2.16 In the first method, a subject is allocated to group 1 if digit 1 is seen in the table of random digits, given that only 1, 2, or 3 are accepted. Therefore,

$$\mathbb{P}\big(\text{allocation to group 1}\big) = \mathbb{P}\big(\text{see 1}\,\big|\,\text{see 1, 2, or 3}\big) = \frac{1/10}{3/10} = \frac{1}{3}\,.$$

In the second method, a subject is allocated to group 1 if digits 1,2, or 3 are seen, provided that zero is not accepted. Hence,

$$\mathbb{P}\big(\text{allocation to group 1}\big) = \mathbb{P}\big(\text{see 1, 2, or 3}\,\big|\,\text{do not see 0}\big) = \frac{3/10}{9/10} = \frac{1}{3}\,.$$

Section 2.4

Exercise 2.17 By the CLT,

$$\hat{\lambda}_i = \frac{\widehat{\mathbb{E}(X_i)}}{T_i} = \frac{n_i}{T_i} \overset{\textit{approx.}}{\sim} \mathcal{N}\big(\lambda_i\,,\, \frac{\lambda_i}{T_i}\big)\,,\ i = 1 \text{ or } 2.$$

Under the null hypothesis, $\lambda_1 = \lambda_2 = \lambda$, say. The pooled estimator of λ is $\hat{\lambda}_{\text{pooled}} = \frac{n_1 + n_2}{T_1 + T_2}$, and therefore, the test statistic is

$$z = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\hat{\lambda}_{\text{pooled}}\left(\frac{1}{T_1} + \frac{1}{T_2}\right)}} = \frac{\frac{n_1}{T_1} - \frac{n_2}{T_2}}{\sqrt{\frac{n_1+n_2}{T_1+T_2}\left(\frac{1}{T_1} + \frac{1}{T_2}\right)}}\,.$$

Exercise 2.18 The derivation of (2.19) is identical to the proof in the previous exercise, with T_1 and T_2 replaced by N_1 and N_2, respectively.

Chapter 3

Section 3.1

EXERCISE 3.1 (a) Since $f(t) = F'(t)$ and $F(t) = 1 - S(t)$ by (3.1), obtain

$$f(t) = \big(1 - S(t)\big)' = -S'(t),$$

and, therefore, (3.2) can be written as

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}.$$

(b) Substituting x for t in the above expression, get the differential equation

$$h(x) = -\frac{S'(x)}{S(x)}.$$

Integrating both sides of this equation from 0 to t and applying (3.3) produces

$$H(t) = \int_0^t h(x)\,dx = -\int_0^t \frac{S'(x)}{S(x)}\,dx = -\ln S(t).$$

(c) Expressing $S(t)$ in the above formula yields

$$S(t) = \exp\big\{-H(t)\big\} = \exp\Big\{-\int_0^t h(x)dx\Big\}.$$

(d) From (3.2) and part (c), $f(t) = h(t)S(t) = h(t)\exp\big\{-H(t)\big\}$.

EXERCISE 3.2 The cdf is $F(t) = \int_0^t f(x)\,dx = 1 - \exp\big\{-\lambda t^{\alpha}\big\}$, $t \geq 0$, thus, by definition, $S(t) = 1 - F(t) = \exp\big\{-\lambda t^{\alpha}\big\}$, $h(t) = f(t)/S(t) = \alpha\lambda t^{\alpha-1}$, and $H(t) = \int_0^t h(x)\,dx = \lambda t^{\alpha}$.

Section 3.2

Exercise 3.3

Time	At risk	Died	Number	Survival Rate	Estimator
t_i	n_i	d_i	Censored	$\left(1-\frac{d_i}{n_i}\right)$	$\hat{S}(t)$, $t_i \leq t < t_{i+1}$
0	10	0	0	$1-0=1.00$	1.00
2.1	10	1	0	$1-\frac{1}{10}=0.90$	(1.00)(0.90)=0.90
2.9	9	1	0	$1-\frac{1}{9}=0.89$	(0.90)(0.89)=0.80
3.0	8	0	1	$1-0=1.00$	(0.80)(1.00)=0.80
3.6	7	1	1	$1-\frac{1}{7}=0.86$	(0.80)(0.86)=0.69
4.5	5	1	0	$1-\frac{1}{5}=0.80$	(0.69)(0.80)=0.55
5.6	4	1	0	$1-\frac{1}{4}=0.75$	(0.55)(0.75)=0.41
6.9	3	1	1	$1-\frac{1}{3}=0.67$	(0.41)(0.67)=0.27
9.1	1	0	1	$1-0=1.00$	(0.27)(1.00)=0.27

The SAS code for this exercise is

```
data exercise3_3;

input duration status @@;

datalines;
 2.1  1    2.9  1    3.6  1    4.5  1

 5.6  1    6.9  1    3.0  0    3.6  0

 6.9  0    9.1  0
;


proc lifetest;
```

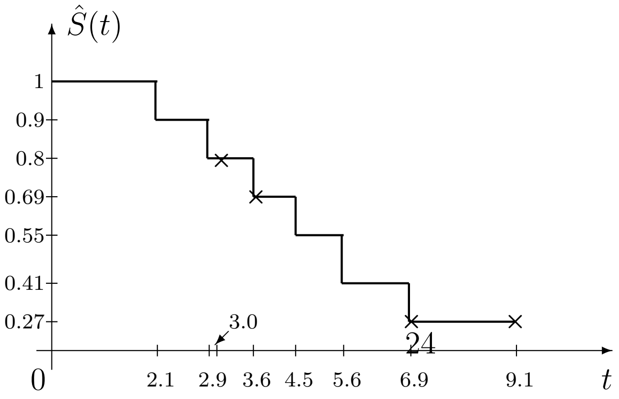
```
   time duration * status(0);
run;
```

The SAS output for this example includes the following columns.

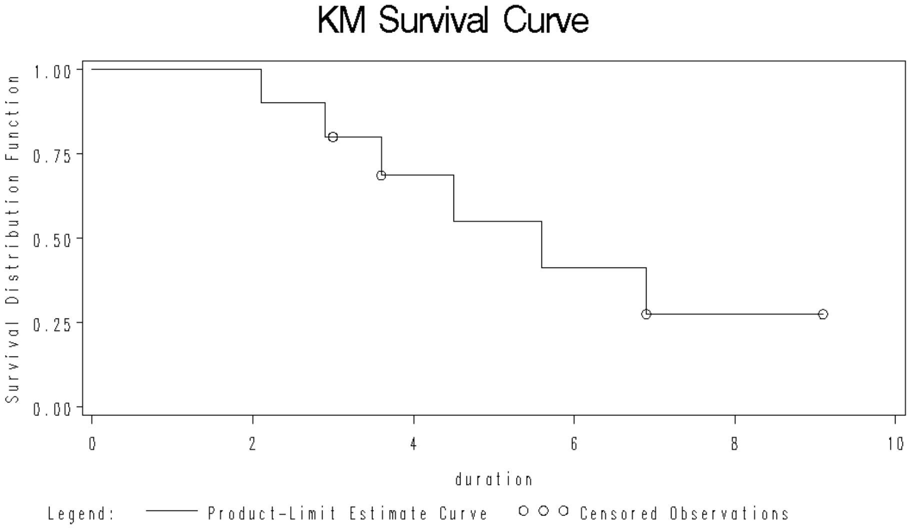
```
duration  Survival
 0.0000     1.0000
 2.1000     0.9000
 2.9000     0.8000
 3.0000*      .
 3.6000     0.6857
 3.6000*      .
 4.5000     0.5486
 5.6000     0.4114
 6.9000     0.2743
 6.9000*      .
 9.1000*      .
```

Section 3.3

Exercise 3.4 The KM survival curve is given in the figure below.



Adding the statement `plots=(survival)` to the `lifetest` procedure in the SAS code for Exercise 3.3 produces the KM survival curve below.



Section 3.4

Exercise 3.5 (a) To see whether the new treatment is effective, use the log-rank test to compare the survival functions for the treatment and control groups. The times until deaths in the two groups combined are 1.2 1.6, 2.3, 3.1, 3.2, and 3.6. The 2×2 tables corresponding to each of these times are

$$t_1 = 1.2$$

	Status of Subject		
Group	Died	Survived	*Total*
Treatment	0	5	5
Control	1	3	4
Total	1	8	9

$$d_{11} = 0, \ \mathbb{E}(d_{11}) = \frac{(5)(1)}{9} = \frac{5}{9}, \ \mathbb{V}ar(d_{11}) = \frac{(5)(4)(8)(1)}{(9)^2(8)} = \frac{20}{81}.$$

$$t_2 = 1.6$$

	Status of Subject		
Group	Died	Survived	*Total*
Treatment	0	5	5
Control	1	2	3
Total	1	7	8

$$d_{12} = 0, \ \mathbb{E}(d_{12}) = \frac{(5)(1)}{8} = \frac{5}{8}, \ \mathbb{V}ar(d_{12}) = \frac{(5)(3)(7)(1)}{(8)^2(7)} = \frac{15}{64}.$$

$$t_3 = 2.3$$

	Status of Subject		
Group	Died	Survived	*Total*
Treatment	1	4	5
Control	1	1	2
Total	2	5	7

$$d_{13} = 1, \ \mathbb{E}(d_{13}) = \frac{(5)(2)}{7} = \frac{10}{7}, \ \mathbb{V}ar(d_{13}) = \frac{(5)(2)(5)(2)}{(7)^2(6)} = \frac{50}{147}.$$

$$t_4 = 3.1$$

	Status of Subject		
Group	Died	Survived	*Total*
Treatment	1	3	4
Control	1	0	1
Total	2	3	5

$$d_{14} = 1, \; \mathbb{E}(d_{14}) = \frac{(4)(2)}{5} = \frac{8}{5}, \; \mathbb{V}ar(d_{14}) = \frac{(4)(1)(3)(2)}{(5)^2(4)} = \frac{6}{25}.$$

$$t_5 = 3.2$$

	Status of Subject		
Group	Died	Survived	*Total*
Treatment	1	2	3
Control	0	0	0
Total	1	2	3

$$d_{15} = 1, \; \mathbb{E}(d_{15}) = \frac{(3)(1)}{3} = 1, \; \mathbb{V}ar(d_{15}) = \frac{(3)(0)(2)(1)}{(3)^2(2)} = 0.$$

$$t_6 = 3.6$$

	Status of Subject		
Group	Died	Survived	*Total*
Treatment	2	0	2
Control	0	0	0
Total	2	0	2

$$d_{16} = 2, \; \mathbb{E}(d_{16}) = \frac{(2)(2)}{2} = 2, \; \mathbb{V}ar(d_{16}) = \frac{(2)(0)(0)(2)}{(2)^2(1)} = 0.$$

Consequently,

$$U \;=\; \Big(0-\frac{5}{9}\Big)+\Big(0-\frac{5}{8}\Big)+\Big(1-\frac{10}{7}\Big)+\Big(1-\frac{8}{5}\Big)+(1-1)+(2-2) \;=\; -\,2.2091\,,$$

and

$$\mathbb{V}ar(U) \;=\; \frac{20}{81}+\frac{15}{64}+\frac{50}{147}+\frac{6}{25}+0+0 \;=\; 1.0614\,.$$

The log-rank test statistic is $z \;=\; -2.2091/\sqrt{1.0614} \;=\; -2.1443$. The approximate P-value for the two-sided test is $2\,\mathbb{P}(Z \;>\; 2.1443) \;=\; 0.032 \;<\; 0.05$. Hence, the null hypothesis of equal survival functions is rejected at 5% significance level, and the conclusion is that the new treatment is effective.

(b) The calculations summarized in the tables below produce the Kaplan-Meier estimator of the survival curves for the two groups.

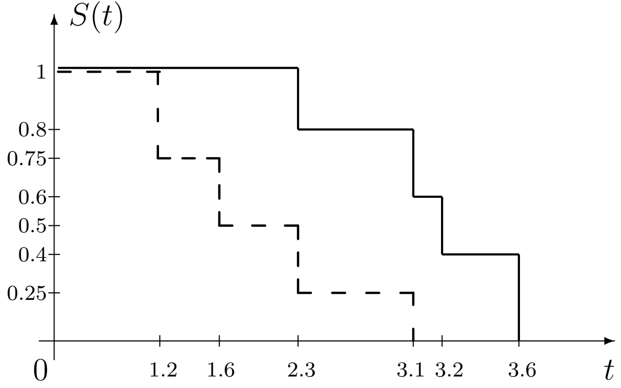
Treatment group:

Time	At risk	Died	Number	Survival Rate	Estimator
t_i	n_i	d_i	Censored	$\left(1-\frac{d_i}{n_i}\right)$	$\hat{S}(t),\, t_i \le t < t_{i+1}$
0	5	0	0	$1-0=1.00$	1.00
2.3	5	1	0	$1-\frac{1}{5}=0.80$	(1.00)(0.80)=0.80
3.1	4	1	0	$1-\frac{1}{4}=0.75$	(0.80)(0.75)=0.60
3.2	3	1	0	$1-\frac{1}{3}=0.67$	(0.60)(0.67)=0.40
3.6	2	2	0	$1-\frac{2}{2}=0.00$	(0.40)(0.00)=0.00

Control group:

Time	At risk	Died	Number	Survival Rate	Estimator
t_i	n_i	d_i	Censored	$\left(1-\frac{d_i}{n_i}\right)$	$\hat{S}(t)$, $t_i \le t < t_{i+1}$
0	4	0	0	1-0=1.00	1.00
1.2	4	1	0	$1-\frac{1}{4}=0.75$	(1.00)(0.75)=0.75
1.6	3	1	0	$1-\frac{1}{3}=0.67$	(0.75)(0.67)=0.50
2.3	2	1	0	$1-\frac{1}{2}=0.50$	(0.50)(0.50)=0.25
3.1	1	1	0	$1-\frac{1}{1}=0.00$	(0.25)(0.00)=0.00

The figure below displays the two survival curves. The control group curve (depicted by the dashed line) lies clearly underneath the one for the treatment group (the solid-lined curve), thus, supporting the statement that the treatment is effective.



(c) The SAS code for this exercise is

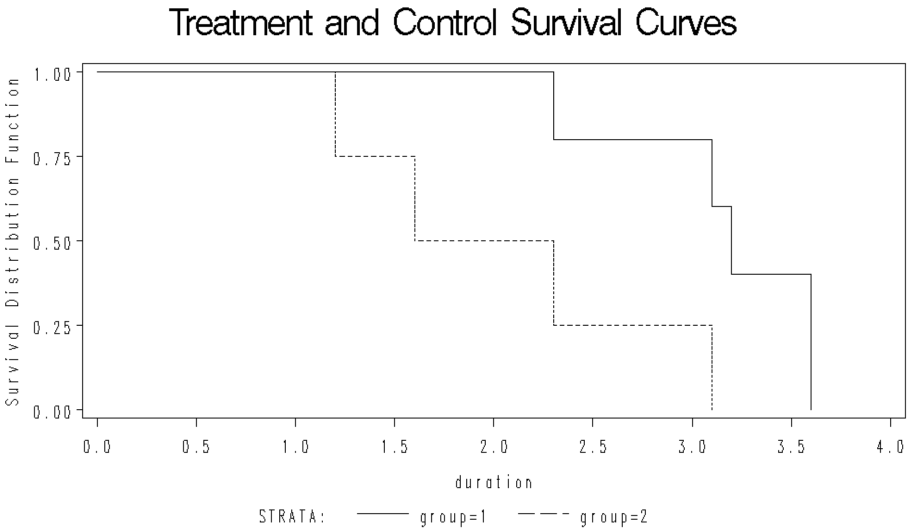
```
data exercise3_5;

input duration status group @@;
```

```
datalines;
 2.3  1  1    3.1  1  1    3.2  1  1

 3.6  1  1    3.6  1  1    1.2  1  2

 1.6  1  2    2.3  1  2    3.1  1  2


;


title’Treatment and Control Survival Curves’;

proc lifetest plots = (survival);

   time duration * status(0);

   strata group;

symbol1 value = none color = black line = 1; /*solid line*/

symbol2 value = none color = black line = 2; /*dashed line*/

run;
```

The log-rank statistic and the P-value are

```
                                 Pr >

  Test     Chi-Square  DF  Chi-Square

Log-Rank    4.5978      1    0.0320
```

The P-value is the same as the one computed by hand. The graph of the respective survival curves plotted by SAS is presented below.

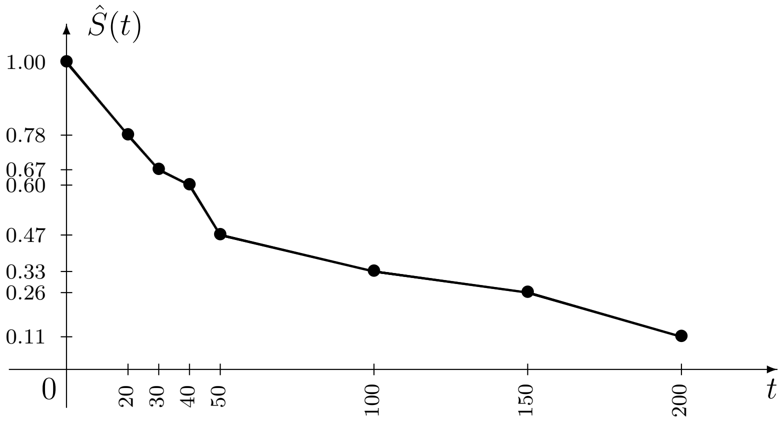


Section 3.5

Exercise 3.6 The calculations of the actuarial estimator of the survival function are summarized in the following table.

Interval	Died	Censored	At Risk	Interval Survival Rate	Survival Function
$[t_i,\ t_{i+1})$	d_i	c_i	$\tilde{n}_i$	$1 - d_i/\tilde{n}_i$	$\hat{S}(t_i)$
[0, 20)	10	0	45.0	$1 - \frac{10}{45} = 0.78$	1.00
[20, 30)	5	0	35.0	$1 - \frac{5}{35} = 0.86$	(1.00)(0.78)=0.78
[30, 40)	3	0	30.0	$1 - \frac{3}{30} = 0.90$	(0.78)(0.86)=0.67
[40, 50)	6	0	27.0	$1 - \frac{6}{27} = 0.78$	(0.67)(0.90)=0.60
[50, 100)	6	1	20.5	$1 - \frac{6}{20.5} = 0.71$	(0.60)(0.78)=0.47
[100, 150)	2	8	10.0	$1 - \frac{2}{10} = 0.80$	(0.47)(0.71)=0.33
[150, 200)	2	1	3.5	$1 - \frac{2}{3.5} = 0.43$	(0.33)(0.80)=0.26
[200, 300)	0	1	0.5	$1 - \frac{0}{0.5} = 1.00$	(0.26)(0.43)=0.11

The actuarial curve is given in the following figure.



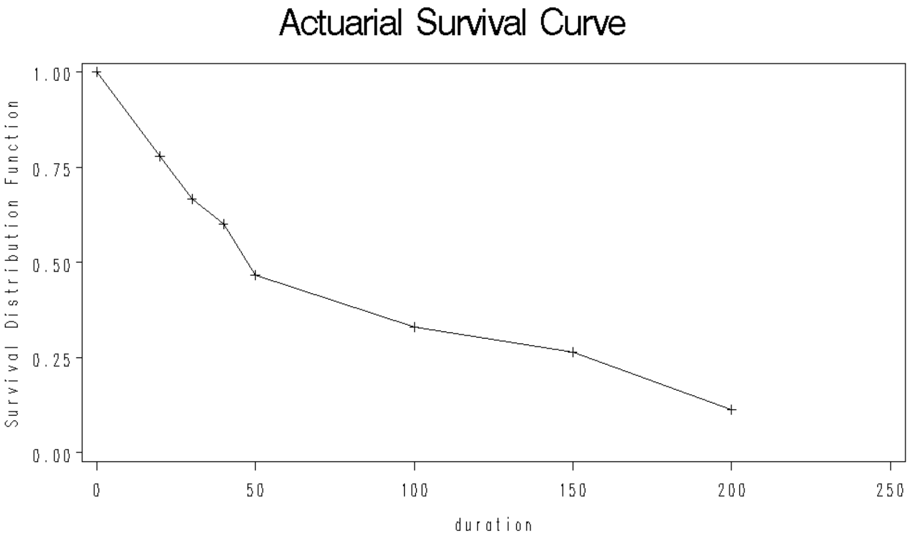
The SAS code for this exercise is as follows.

```
data exercise3_6;

input duration status @@;

datalines;
 15   1     11   1     22   1    121   0     38   1

 45   1     76   1     18   1    139   0    105   0

 51   1     44   1     10   1    111   0    137   0

 11   1    132   1     43   1     10   1    271   0

 77   0     56   1     44   1     28   1     27   1

 36   1     11   1     76   1    115   0    148   0

 43   1     56   1    179   0    182   1    123   1

 27   1    174   1     16   1     24   1     18   1

 95   1    128   0     40   1     36   1     13   1
;


title‘Actuarial Survival Curve’;
```

```
proc lifetest method = act plots = (survival)
intervals = 0, 20, 30, 40, 50, 100, 150, 200;
   time duration * status(0);
run;
```

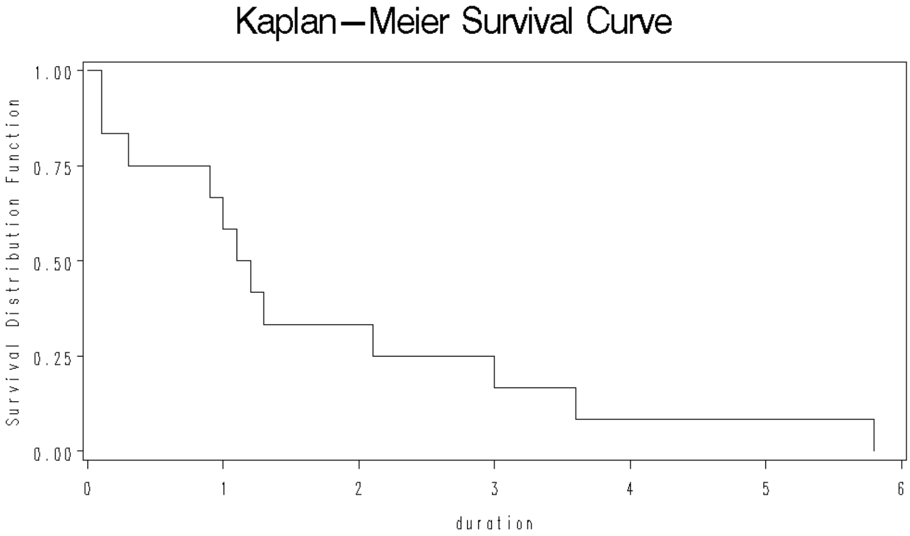
The relevant SAS output is presented below.

```
                                        Effective
       Interval      Number   Number     Sample
    [Lower,  Upper)  Failed  Censored     Size      Survival
       0       20      10        0        45.0       1.0000
      20       30       5        0        35.0       0.7778
      30       40       3        0        30.0       0.6667
      40       50       6        0        27.0       0.6000
      50      100       6        1        20.5       0.4667
     100      150       2        8        10.0       0.3301
     150      200       2        1         3.5       0.2641
     200        .       0        1         0.5       0.1132
```



Section 3.6

EXERCISE 3.7 The KM survival curve is constructed using SAS software. The graph is given below.



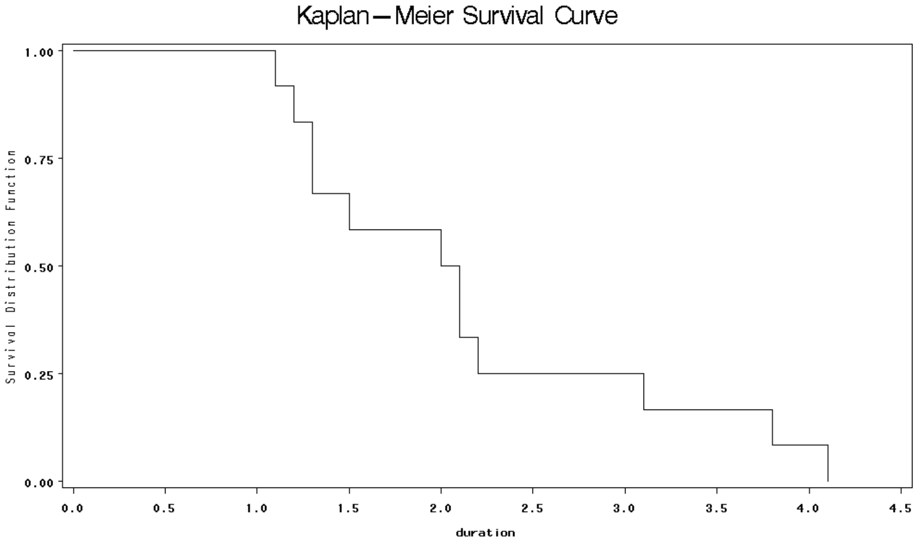
As seen on the picture, the estimator of the survival function decays exponentially starting right at the baseline. This implies the appropriateness of the exponential distribution model. Using (3.10) with $\delta_i = 1$ for all $i = 1, \dots, 12$, obtain the maximum likelihood estimator of the parameter λ

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i} = \frac{12}{0.1 + \ldots + 5.8} = \frac{12}{20.5} = 0.5854 \, .$$

Thus, according to (3.11), the maximum likelihood estimator of the survival function is

$$\hat{S}(t) = \exp\Big\{ -0.5854\, t \Big\}, \, t \geq 0 \, .$$

EXERCISE 3.8 The Kaplan-Meier survival curve plotted in SAS is given in the figure below.



The graph shows the one hundred percent survival for roughly the first month after entering the trial, and then the percentage of surviving subjects decreases exponentially. This indicates that the Weibull distribution model may be appropriate.

From (3.12), the parameter estimators $\hat{\alpha}$ and $\hat{\lambda}$ solve the system of normal equations

$$\begin{cases} 12/\hat{\alpha} \,+\, 8.0507 \,-\, \hat{\lambda}\big[(1.1)^{\hat{\alpha}}\,\ln(1.1) \,+\, \ldots \,+\, (4.1)^{\hat{\alpha}}\,\ln(4.1)\big] \,=\, 0 \\ 12/\hat{\lambda} \,-\, \big[(1.1)^{\hat{\alpha}} \,+\, \ldots \,+\, (4.1)^{\hat{\alpha}}\big] \,=\, 0\,. \end{cases}$$

The numerical solution is $\hat{\alpha} \,=\, 2.3812$ and $\hat{\lambda} \,=\, 0.1198$. Thus, according to (3.13), the maximum likelihood estimator of the survival function is

$$\hat{S}(t) = \exp\,\big\{-\,0.1198\,t^{\,2.3812}\big\},\; t \geq 0\,.$$

Section 3.7

EXERCISE 3.9 It is given that $\ln T = \beta_0 \,+\, \ldots \,+\, \beta_m\,x_m \,+\, \varepsilon$, and $f_\varepsilon(x) \,=\, \exp\big\{\,x \,-\, \exp\{x\}\,\big\}$. Denote the regression term by $R \,=\, \beta_0 \,+\, \ldots \,+\, \beta_m\,x_m$. Hence,

$$F_T(t) \,=\, \mathbb{P}(T \,\leq\, t) \,=\, \mathbb{P}\big(\,\ln T \,\leq\, \ln t\,\big) = \mathbb{P}\big(\varepsilon \,\leq\, \ln t - R\big) \,=\, F_\varepsilon\,\big(\ln t - R\big)\,,$$

and, therefore,

$$f_T(t) \,=\, F'_T(t) \,=\, F'_\varepsilon\big(\ln t \,-\, R\big)$$

$$= \frac{1}{t}\,f_\varepsilon\,\big(\ln t \,-\, R\big) \,=\, \frac{1}{t}\,\exp\Big\{\ln t \,-\, R \,-\, \exp\big\{\ln t \,-\, R\big\}\Big\}$$

$$= \frac{1}{t}\,t\,\exp\big\{-\,R\big\}\,\exp\Big\{-\,t\,\exp\big\{\,-\,R\,\big\}\Big\} = \lambda\,\exp\{-\,\lambda\,t\}\,,$$

where

$$\lambda = \exp\big\{-R\big\} = \exp\big\{-(\beta_0 + \ldots + \beta_m x_m)\big\}.$$

EXERCISE 3.10 Write the parametric regression in the form $\ln T = R + \sigma\,\varepsilon$, where $R = \beta_0 + \ldots + +\beta_m x_m$ is the regression part. It is given that the error term ε has density $f_\varepsilon(x) = \exp\big\{x - \exp\{x\}\big\}$. The cdf of T can be written as

$$F_T(t) = \mathbb{P}(T \le t) = \mathbb{P}\big(\ln T \le \ln t\big)$$

$$= \mathbb{P}\Big(\varepsilon \le \frac{1}{\sigma}\big(\ln t - R\big)\Big) = F_\varepsilon\Big(\frac{1}{\sigma}\big(\ln t - R\big)\Big),$$

yielding the expression for the density

$$f_T(t) = F'_T(t) = F'_\varepsilon\Big(\frac{1}{\sigma}\big(\ln t - R\big)\Big) = \frac{1}{\sigma\,t}\,f_\varepsilon\Big(\frac{1}{\sigma}\big(\ln t - R\big)\Big)$$

$$= \frac{1}{\sigma\,t}\,\exp\Big\{\frac{1}{\sigma}\big(\ln t - R\big) - \exp\big\{\frac{1}{\sigma}\big(\ln t - R\big)\big\}\Big\}$$

$$= \frac{1}{\sigma}\,t^{\,1/\sigma-1}\,\exp\big\{-R/\sigma\big\}\,\exp\Big\{-t^{1/\sigma}\,\exp\big\{-R/\sigma\big\}\Big\}$$

$$= \alpha\,\lambda\,t^{\,\alpha-1}\exp\{-\lambda\,t^{\alpha}\},\ \text{ where }\ \alpha = 1/\sigma,$$

$$\text{and }\ \lambda = \exp\big\{-R/\sigma\big\} = \exp\big\{-(\beta_0 + \ldots + \beta_m x_m)/\sigma\big\}.$$

Note that, as expected, the proof in Exercise 3.9 is a special case of this one with $\sigma = 1$.

EXERCISE 3.11 The relevant SAS code is

```
data fromExercise3_7;
input duration status @@;
```

```
datalines;
 0.1  1    0.1  1    0.3  1    0.9  1
 1.0  1    1.1  1    1.2  1    1.3  1
 2.1  1    3.0  1    3.6  1    5.8  1
;

proc lifereg;
model duration * status(0) = / dist = exponential;

proc lifereg;
model duration * status(0) = / dist = weibull;
run;
```

From the SAS output, the values of the log-likelihood functions are $\ln L(\lambda) = -18.9214$ and $\ln L(\alpha, \lambda) = -18.9208$. Therefore, the goodness-of-fit test statistic equals to $-2\left(\ln L(\lambda) - \ln L(\alpha, \lambda)\right) = 0.0012$, with the approximate P-value$=\mathbb{P}\left(\chi^2(1) > 0.0012\right) = 0.9724 > 0.05$. The conclusion is that the exponential model is more adequate for these data (the same conclusion as in Exercise 3.7).

SAS gives the estimator of the `Intercept` $\hat{\beta}_0 = 0.5355$. Hence, the estimator of the parameter λ of the exponential distribution is $\hat{\lambda} = \exp\{-\beta_0\} = 0.5854$, as was obtained by hand in Exercise 3.7.

Exercise 3.12 The SAS code for obtaining the goodness-of-fit test statistic and the estimates of the model parameters is

```
data fromExercise3_8;
input duration status @@;
datalines;
 1.1  1    1.2  1    1.3  1    1.3  1
 1.5  1    2.0  1    2.1  1    2.1  1
 2.2  1    3.1  1    3.8  1    4.1  1
;

proc lifereg;
model duration * status(0) = / dist = exponential;

proc lifereg;
model duration * status(0) = / dist = weibull;
run;
```

The SAS output contains the values of the log-likelihood functions under the exponential model, $\ln L(\lambda) = -13.1349$, and under the Weibull model, $\ln L(\alpha, \lambda) = -7.8864$. The test statistic is computed as $-2\big(\ln L(\lambda) - \ln L(\alpha, \lambda)\big) = 10.4970$. The approximate P-value is $\mathbb{P}\big(\chi^2(1) > 10.4970\big) = 0.0012 < 0.05$, which confirms that the Weibull distribution is more appropriate to model the data (the same conclusion as in Exercise 3.8).

The estimates of the model parameters α and λ can be calculated from the values of the `Intercept` $\hat{\beta}_0 = 0.8913$, and the `Scale` $\hat{\sigma} = 0.4200$. The estimates are $\hat{\alpha} = 1/\hat{\sigma} = 1/0.4200 = 2.3810$ and $\hat{\lambda} = \exp\big\{-\hat{\beta}_0/\hat{\sigma}\big\} =$

$\exp\big\{-0.8913/0.4200\big\} = 0.1198$. Note that the discrepancy between this estimate of α and the one computed in Exercise 3.8 is due to the round-off error.

EXERCISE 3.13 To accommodate the four levels of the categorical variable `number of infections`, introduce a new variable `infections`:

`infections`	`number of infections`
1	1
2	2
3	3
4	4, 5, or 6

The SAS code for this exercise is

```
data ear_infections;
input age infections duration censored @@;
datalines;
  2.0  1  10.1  0     2.1  1  10.9  0     3.0  4  1.6  0
  3.1  1  10.1  1     3.8  4   0.3  0     4.2  3  7.3  1
  5.1  3   8.2  0     5.4  2   8.0  0     6.0  1  5.7  0
  7.0  1   4.9  0     7.6  3   2.5  0     7.7  3  1.0  0
  7.8  3   2.8  0     8.1  4   1.4  1     8.2  2  6.3  0
  8.5  2   4.0  0     9.4  4   1.8  0    11.0  2  1.9  1
 12.5  2   2.5  1    13.1  3   1.9  1
;
```

```
proc lifereg;
class infections;
model duration*censored(1) = age infections/dist = exponential;
run;

proc lifereg;
class infections;
model duration*censored(1) = age infections/dist = weibull;
run;
```

(a) From the SAS output, the values of the log-likelihood functions are $\ln L(\lambda) = -20.9059$ (for the exponential model), and $\ln L(\alpha, \lambda) = -15.3939$ (for the Weibull model). The test statistic is $-2\left(\ln L(\lambda) - \ln L(\alpha, \lambda)\right) = 11.0240$, which approximate P-value equals to $\mathbb{P}\left(\chi^2(1) > 11.0240\right) = 0.0009 < 0.05$. Therefore, the Weibull model should be used in this problem.

(b) For the Weibull model, SAS gives the following estimates of the regression coefficients and their P-values:

```
Parameter      Estimate  Pr > ChiSq
Intercept        1.1060     0.0279
age             -0.0815     0.1281
infections 1     1.4156     0.0006
infections 2     1.3611    <0.0001
infections 3     1.1295     0.0008
```

The P-value for the covariate `age` is larger than 0.05, therefore, it may be reasonable to remove it from the model, and re-run SAS to obtain parameter estimates for the reduced model. The new parameter estimates and P-values are:

```
Parameter      Estimate  Pr > ChiSq
Intercept        0.4800      0.0697
infections 1     1.7936    < 0.0001
infections 2     1.3988      0.0002
infections 3     1.3060      0.0002
Scale            0.4583
```

From here, the estimators of the Weibull regression model parameters are $\hat{\alpha} = 1/0.4583 = 2.1820$, and

$$\hat{\lambda} = \exp\Big\{ -\big(0.4800 + 1.7936\,x_1 + 1.3988\,x_2 + 1.3060\,x_3\big)/0.4583 \Big\}$$

$$= \exp\Big\{ -1.0473 - 3.9136\,x_1 - 3.0521\,x_2 - 2.8497\,x_3 \Big\},$$

where $x_1 = 1$ if the number of previous infections $= 1$, and 0 otherwise; $x_2 = 1$ if the number of previous infections $= 2$, and 0 otherwise; and $x_3 = 1$ if the number of previous infections $= 3$, and 0 otherwise.

Section 3.8

Exercise 3.14 When the partial likelihood estimates of $\beta_1, \dots, \beta_m$ are plugged in, the log-likelihood function (3.35) takes the form

$$\ln L\big(\pi_1, \ldots, \pi_n, \hat{\beta}_1, \ldots, \hat{\beta}_m\big) = \sum_{i=1}^{n}\left[\sum_{j\,\in\, D(t_i)} \ln\big(1 - \pi_i^{\hat{r}_j}\big) + \sum_{j\,\in\, R(t_i)\setminus D(t_i)} \hat{r}_j \ln \pi_i\right],$$

where $\hat{r}_j = \exp\big\{\hat{\beta}_1 x_{j1} + \ldots + \hat{\beta}_m x_{j\,m}\big\}$. Equating to zero the partial derivatives of $\ln L$ with respect to $\pi_i\,, i = 1, \ldots n$, obtain the system of normal equations

$$\frac{\partial}{\partial \pi_i} \ln L\big(\hat{\pi}_1, \ldots, \hat{\pi}_n, \hat{\beta}_1, \ldots, \hat{\beta}_m\big) = \sum_{j\,\in\, D(t_i)} \frac{-\,\hat{r}_j\,\hat{\pi}_i^{\hat{r}_j - 1}}{1 - \hat{\pi}_i^{\hat{r}_j}} + \sum_{j\,\in\, R(t_i)\setminus D(t_i)} \frac{\hat{r}_j}{\hat{\pi}_i} = 0\,.$$

The algebraic manipulations presented below simplify these equations to the form given in (3.36),

$$\sum_{j\,\in\, D(t_i)}\left[\frac{-\,\hat{r}_j\,\hat{\pi}_i^{\hat{r}_j - 1}}{1 - \hat{\pi}_i^{\hat{r}_j}} - \frac{\hat{r}_j}{\hat{\pi}_i}\right] + \sum_{j\,\in\, R(t_i)} \frac{\hat{r}_j}{\hat{\pi}_i} = 0\,,$$

$$\sum_{j\,\in\, D(t_i)} \frac{-\,\hat{r}_j\,\hat{\pi}_i^{\hat{r}_j} - \hat{r}_j\,(1 - \hat{\pi}_i^{\hat{r}_j})}{(1 - \hat{\pi}_i^{\hat{r}_j})\,\hat{\pi}_i} + \sum_{j\,\in\, R(t_i)} \frac{\hat{r}_j}{\hat{\pi}_i} = 0\,,$$

$$\sum_{j\,\in\, D(t_i)} \frac{\hat{r}_j}{1 - \hat{\pi}_i^{\hat{r}_j}} = \sum_{j\,\in\, R(t_i)} \hat{r}_j\,.$$

Exercise 3.15 Let `x1, x2` and `x3` denote the indicators of 1, 2, and 3 previous ear infections, as defined in the solution of Exercise 3.13. To estimate the parameters in the Cox proportional hazards model, use the SAS code below:

```
data fromExercise3_13;
```

```
input x1 x2 x3 duration censored @@;
datalines;
 1  0  0  10.1  0    1  0  0  10.9  0
 0  0  0   1.6  0    1  0  0  10.1  1
 0  0  0   0.3  0    0  0  1   7.3  1
 0  0  1   8.2  0    0  1  0   8.0  0
 1  0  0   5.7  0    1  0  0   4.9  0
 0  0  1   2.5  0    0  0  1   1.0  0
 0  0  1   2.8  0    0  0  0   1.4  1
 0  1  0   6.3  0    0  1  0   4.0  0
 0  0  0   1.8  0    0  1  0   1.9  1
 0  1  0   2.5  1    0  0  1   1.9  1
;

proc phreg outest = betas;
   model duration * censored(1) = x1 x2 x3;
   baseline out = outdata survival = s;
run;

proc print data = betas;
run;

proc print data = outdata;
run;
```

The estimates of β's are:

```
                Parameter
    Variable    Estimate    Pr > ChiSq
       x1       -4.14885        0.0028
       x2       -2.92502        0.0209
       x3       -2.75137        0.0257
```

Note that these estimators are similar to the ones obtained in Exercise 3.13, but are not very close.

To estimate nonparametrically the baseline survival function $S_0(t)$, study the following output:

```
 x1    x2    x3   duration     s

0.25  0.25  0.3     0.0     1.00000

0.25  0.25  0.3     0.3     0.98241

0.25  0.25  0.3     1.0     0.96276

0.25  0.25  0.3     1.6     0.92952

0.25  0.25  0.3     1.8     0.86799

0.25  0.25  0.3     2.5     0.75104

0.25  0.25  0.3     2.8     0.62266

0.25  0.25  0.3     4.0     0.49982

0.25  0.25  0.3     4.9     0.39150

0.25  0.25  0.3     5.7     0.30261

0.25  0.25  0.3     6.3     0.22564

0.25  0.25  0.3     8.0     0.13039

0.25  0.25  0.3     8.2     0.04800

0.25  0.25  0.3    10.1     0.00704

0.25  0.25  0.3    10.9     0.00000
```

The estimate `s` given in the last column is

$$S_{est}(t) \;=\; \big[\hat{S}_0(t)\big]^{\exp\big\{-\,(4.14885)\,(0.25)\,-\,(2.92502)\,(0.25)\,-\,(2.75137)\,(0.3)\big\}}$$

$$=\; \big[\hat{S}_0(t)\big]^{\exp\{-\,2.59388\}}\,,\ \text{and therefore } \hat{S}_0(t) \;=\; \big[S_{est}(t)\big]^{\exp\{2.59388\}}\,.$$

The estimator of the baseline survival function for the Weibull model is

$$\hat{S}_0^{wei}(t) \;=\; \exp\big\{\,-\,\exp\{-1.0473\}\,t^{\,2.1820}\big\}\,.$$

To see how similar these functions are, in SAS type

```
data new;
set outdata;
s_null = s**exp(2.59388);
s_wei = exp(-exp(-1.0473) * duration**2.1820);
run;

proc print data = new;
run;
```

The values that the two functions assume at the times of death are

```
duration   s_null    s_wei

   0.0     1.00000  1.00000

   0.3     0.78860  0.97495

   1.0     0.60176  0.70407

   1.6     0.37608  0.37588

   1.8     0.15040  0.28218

   2.5     0.02169  0.07494

   2.8     0.00177  0.03623

   4.0     0.00009  0.00073

   4.9     0.00000  0.00001

   5.7     0.00000  0.00000

   6.3     0.00000  0.00000

   8.0     0.00000  0.00000

   8.2     0.00000  0.00000

  10.1     0.00000  0.00000

  10.9     0.00000  0.00000
```

The step-function `s_null` decreases a bit faster than the other function, however, there is no huge difference in their behavior.

The estimate of the survival function in this model is

$$\hat{S}(t) \;=\; \left[S_{est}(t)\right]^{\exp\left\{\,2.59388-4.14885\,x_1-2.92502\,x_2-2.75137\,x_3\,\right\}}.$$

The fitted regression coefficients provide the estimate of the ratio of hazard functions for subjects with one, two, three, and four or more previous ear infections. The results are:

- The ratio of the hazard functions for subjects with one and two ear infections is $100\exp\{-4.14885 + 2.92502\}\% = 29.41\%$ (implying that the hazard of getting a recurrence for the subjects with one previous ear infection is only 29.41 percent of the hazard function for subjects with two ear infections).
- The ratio of the hazard functions for subjects with one and three ear infections is $100\exp\{-4.14885 + 2.75137\}\% = 24.72\%$.
- The ratio of the hazard functions for subjects with one and four or more ear infections is $100\exp\{-4.14885\}\% = 1.58\%$.
- The ratio of the hazard functions for subjects with two and three ear infections is $100\exp\{-2.92502 + 2.75137\}\% = 84.06\%$.
- The ratio of the hazard functions for subjects with two and four or more ear infections is $100\exp\{-2.92502\}\% = 5.37\%$.
- The ratio of the hazard functions for subjects with three and four or more ear infections is $100\exp\{-2.75137\}\% = 6.38\%$.

Exercise 3.16 To fit the Cox model to these data, write in SAS

```
data cirrhosis;
input age alcohol duration censored @@;
datalines;
```

```
 42  1  0.2  0    45  0  1.7  0    47  0  1.6  0
 49  1  1.4  0    51  0  2.4  0    53  0  3.5  0
 54  1  2.8  0    55  1  2.2  1    57  0  4.5  0
 57  0  3.6  0    58  0  5.1  0    61  1  3.4  0
 62  0  2.4  0    67  1  5.3  0    68  1  2.6  1
 68  1  3.8  0    69  1  5.8  0

;

proc phreg outest = betas;
   model duration * censored(1) = age alcohol;
   baseline out = outdata survival = s;
run;

proc print data = betas;
run;

proc print data=outdata;
run;
```

The SAS output is

```
                    Parameter
        Variable    Estimate    Pr > ChiSq
          age       -0.33933      0.0005
        alcohol      1.81807      0.0395
```

```
  age     alcohol  duration     s

56.6471  0.52941     0.0     1.00000

56.6471  0.52941     0.2     0.99504

56.6471  0.52941     1.4     0.97855

56.6471  0.52941     1.6     0.95453

56.6471  0.52941     1.7     0.91576

56.6471  0.52941     2.4     0.74933

56.6471  0.52941     2.8     0.62053

56.6471  0.52941     3.4     0.42553

56.6471  0.52941     3.5     0.23179

56.6471  0.52941     3.6     0.07579

56.6471  0.52941     3.8     0.01855

56.6471  0.52941     4.5     0.00248

56.6471  0.52941     5.1     0.00002

56.6471  0.52941     5.3     0.00000

56.6471  0.52941     5.8     0.00000
```

From the output, the estimator of the baseline survival function is

$$\hat{S}_0(t) \;=\; \big[S_{est}(t)\big]^{\exp\big\{(0.33933)(56.6471)\,-\,(1.81807)(0.52941)\big\}} \;=\; \big[S_{est}(t)\big]^{\exp\{18.25956\}},$$

where $S_{est}(t)$ is a step-function, which values at the remission times are given in column `s` in the above table.

The estimator of the survival function in the Cox model is

$$\hat{S}(t) \;=\; \big[S_{est}(t)\big]^{\exp\big\{\,18.25956 - 0.33933\,\mathrm{age} + 1.81807\,\mathrm{alcohol}\,\big\}}.$$

The regression coefficients are interpreted as follows:

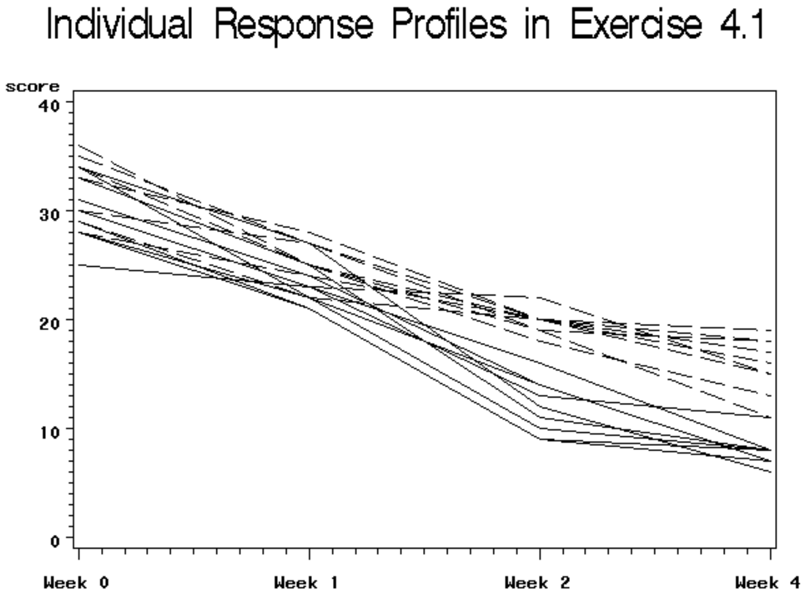
• With one-year age increase, the subject's hazard function changes by $100\big(\exp\{-0.33933\} - 1\big)\,\% = -28.78\,\%$ that is, the "hazard" of going into remission decreases by 28.78 percent with one-year increase in age.

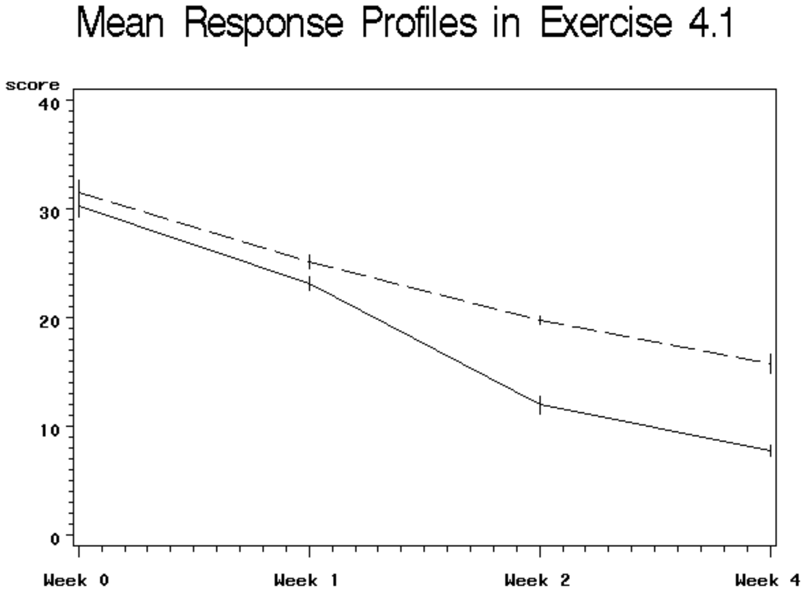
• The ratio of the hazard functions for two same-age subjects, one of whom does not abuse alcohol and the other one does, is $100\exp\{-1.81807\}\,\% = 16.23\,\%$ which means that the "hazard" of going into remission for subjects who do not abuse alcohol is only 16.23 percent of that for alcohol-abusive subjects. So, yes, the subjects who abuse alcohol have a larger "hazard" of going into remission.

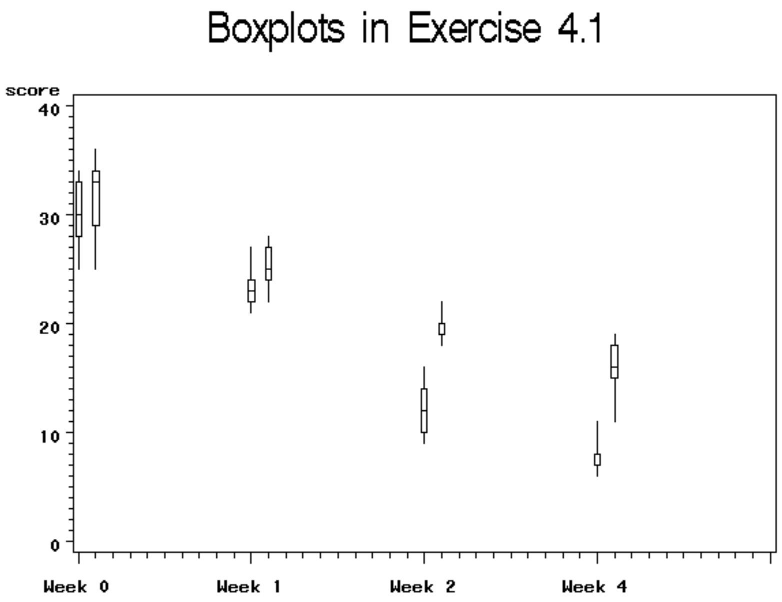
Chapter 4

Section 4.2

Exercise 4.1 Below are the individual response profiles, the mean response profiles, and the boxplots for the treatment and the control groups.







The solid lines correspond to the treatment group, and the dashed ones, to the control group. Also, of the two boxplots at each visit, the treatment group boxplots are on the left, and the control group ones, on the right.

As seen on these graphs, the individual response profiles for the subjects in the treatment group tend to lie below those for the control group subjects. The mean response profile for the control group is above that for the treatment group at every point. Finally, the boxplots for the treatment group tend to be lower than their counterparts for the control group. These observations support the conclusion that the mental distress for the subjects in the treatment group is lower than that for the subjects in the control group, and therefore, the new drug is effective as a reducer of mental distress.

The SAS code for this exercise is

```
data mental_distress;
input individual group visit0 visit1 visit2 visit4 @@;
datalines;
  1  1  25  23  16   8     2  1  34  22   4   7
  3  1  31  24  14   7     4  1  34  27  12   6
  5  1  33  25  11   8     6  1  30  23  13  11
  7  1  28  22  10   8     8  1  29  21   9   8
  9  1  28  21   9   7    10  2  33  25  18  13
 11  2  30  27  19  11    12  2  29  22  20  15
 13  2  35  27  20  18    14  2  25  23  22  15
 15  2  36  25  20  16    16  2  34  25  19  18
 17  2  33  28  20  17    18  2  28  24  20  19
;

data new;
set metal_distress;
array x{4} visit0 visit1 visit2 visit4;
do visits = 1 to 4;
score = x{visits};
if group = 1 then boxposition = visits;
else boxposition = visits + 0.1;
output;
end;
keep individual group visits score boxposition;
run;
```

```
axis1 label = none
value = (t=1 ‘Week 0’ t=2 ‘Week 1’ t=3 ‘Week 2’ t=4 ‘Week 4’ t=5 ‘’);


title’Individual Response Profiles in Exercise 4.1’;
proc gplot data = new;
plot score * visits = individual / nolegend haxis = axis1;
symbol1 interpol = join value = none color = black
line = 1 repeat = 9;
symbol2 interpol = join value = none color = black
line = 2 repeat = 9;
run;

goptions reset = symbol;

title’Mean Response Profiles in Exercise 4.1’;
proc gplot data = new;
plot score * visits = group / nolegend haxis = axis1;
symbol1 interpol = stdm1j color = black line = 1;
symbol2 interpol = stdm1j color = black line = 2;
run;

goptions reset = symbol;

title’Boxplots in Exercise 4.1’;
proc gplot data = new;
```

```
plot score * boxposition = group / nolegend haxis = axis1;
symbol1 interpol = box00 value = star color = black;
symbol2 interpol = box00 value = x color = black;
run;
```

Section 4.3

EXERCISE 4.2 Let $a = \sigma_u^2$, and $b = \sigma^2$. To find the determinant of the matrix

$$\begin{bmatrix} a+b & a & a & \dots & a \\ a & a+b & a & \dots & a \\ & & \dots & & \\ a & a & a & \dots & a+b \end{bmatrix},$$

first replace each row by the difference between this row and the next one, except for the last row. The determinant does not change under this elementary row operation. The resulting matrix is

$$\begin{bmatrix} b & -b & 0 & \dots & 0 \\ 0 & b & -b & \dots & 0 \\ & & \dots & & \\ a & a & a & \dots & a+b \end{bmatrix}.$$

Denote by D_k the determinant of a $k \times k$ matrix of the above form. Then expanding the determinant along the first column, and noting that the determinant of the $(k-1)\times(k-1)$ lower triangular matrix $\begin{bmatrix} -b & 0 & \dots & 0 \\ b & -b & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & -b \end{bmatrix}$ is equal to $(-b)^{k-1}$, produces the recursive formula

$$D_k = b\,D_{k-1} + (-1)^{k+1}\,a\,(-b)^{k-1} = b\,D_{k-1} + a\,b^{k-1}$$

$$= b\big(b\,D_{k-2} + a\,b^{k-2}\big) + a\,b^{k-1} = b^2\,D_{k-2} + 2\,a\,b^{k-1}$$

$$= \dots = b^{k-1}\,D_1 + (k-1)\,a\,b^{k-1}$$

$$= b^{k-1}\,(a+b) + (k-1)\,a\,b^{k-1} = b^k + k\,a\,b^{k-1}\,.$$

This shows (4.3). To verify (4.4), it suffices to show that $\mathbf{V}_0\,\mathbf{V}_0^{-1} = \mathbf{I}_k$. In terms of a and b,

$$\mathbf{V}_0\,\mathbf{V}_0^{-1} = \frac{1}{b^2 + abk}\big(b\,\mathbf{I}_k + a\,\mathbf{J}_k\big)\big((b+ak)\,\mathbf{I}_k - a\,\mathbf{J}_k\big)$$

$$= \frac{1}{b^2 + abk}\big(b^2\,\mathbf{I}_k + \cancel{ab\mathbf{J}_k} + abk\,\mathbf{I}_k + \cancel{a^2k\,\mathbf{J}_k} - \cancel{ab\mathbf{J}_k} - \cancel{a^2k\,\mathbf{J}_k}\big) = \mathbf{I}_k\,.$$

Exercise 4.3 By the properties of the matrix $\mathbf{A}$ and the definition of the matrix $\mathbf{B}$,

$$\big(\mathbf{A}\,\mathbf{A}'\big)\,\mathbf{B}\,\mathbf{B}' - \big(\mathbf{A}\,\mathbf{A}'\big)\,\mathbf{B}\,\mathbf{A}'\big(\mathbf{A}\,\mathbf{A}'\big)^{-\mathbf{1}}\,\mathbf{A}\,\mathbf{B}'$$

$$= \mathbf{I}_{n\,k-p-2}\,\mathbf{B}\,\mathbf{B}' - \mathbf{I}_{n\,k-p-2}\,\mathbf{B}\,\mathbf{A}'\big(\mathbf{I}_{n\,k-p-2}\big)^{-1}\,\mathbf{A}\,\mathbf{B}'$$

$$= \mathbf{B}\mathbf{B}' - \mathbf{B}\mathbf{A}'\mathbf{A}\mathbf{B}' = \mathbf{B}\mathbf{B}' - \mathbf{B}\left(\mathbf{I}_{nk} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\right)\mathbf{B}'$$

$$= \mathbf{B}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\mathbf{B}' = \left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

since $\mathbf{B}\mathbf{X} = \left(\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X} = \mathbf{I}_{p+2}$.

EXERCISE 4.4 (a) Refer to the solution of Exercise 4.1. As seen on the individual response profiles graph, the variance of the responses in each group is approximately constant. Also, it may be assumed that the covariance between the scores on the health questionnaire for any two visits is roughly the same. Thus the data meet the assumptions of the random intercept model.

(b) Let the covariate `group` be $x_i = 1$, if the i-th subject is from group 1 (the treatment group), and $x_i = 2$, if the i-th subject belongs to group 2 (the control group). Denote by y_{ij} the general health questionnaire score for the i-th subject at time t_j, where $t_1 = 0$, $t_2 = 1$, $t_3 = 2$, and $t_4 = 4$ weeks. Then the random intercept model is

$$y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 t_j + u_i + \varepsilon_{ij}, \quad i = 1, \ldots, 18, \ j = 1, \ldots, 4,$$

where the random intercepts $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2)$ are independent of the random errors $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

The required SAS code for the dataset `mental_distress` is

```
data new;
set mental_distress;
array x{4} visit0 visit1 visit2 visit4;
array t{4} t1-t4 (0 1 2 4);
do visits = 1 to 4;
score = x{visits};
time = t{visits};
output;
end;
keep individual group score time;
run;

proc mixed data = new method = ml;
   model score = group time / solution;
   random intercept / subject = individual;
run;

proc mixed data = new method = reml;
   model score = group time / solution;
   random intercept / subject = individual;
run;
```

This code produces the following estimators of the regression parameters

Both ML and REML Methods

$$\begin{array}{llrr}
 & \texttt{Effect} & \texttt{Estimate} & \texttt{Pr > |t|} \\
\hat{\beta}_0 \rightarrow & \texttt{Intercept} & \texttt{21.8417} & \texttt{<.0001} \\
\hat{\beta}_1 \rightarrow & \texttt{group} & \texttt{4.7500} & \texttt{<.0001} \\
\hat{\beta}_2 \rightarrow & \texttt{time} & \texttt{-4.7508} & \texttt{<.0001}
\end{array}$$

$$\begin{array}{rlrl}
 & \texttt{Covariance} & & \\
 & \texttt{Parameter} & \texttt{Estimate} & \\
 & \texttt{Intercept} & \texttt{0} & \leftarrow \hat{\sigma}_u^2 \\
\text{ML Method} & \texttt{Residual} & \texttt{12.5196} & \leftarrow \hat{\sigma}^2 \\
\text{REML Method} & \texttt{Residual} & \texttt{13.0640} & \leftarrow \hat{\sigma}^2
\end{array}$$

Note that the random intercept term u_i is not present in the fitted model since its variance is equal to zero.

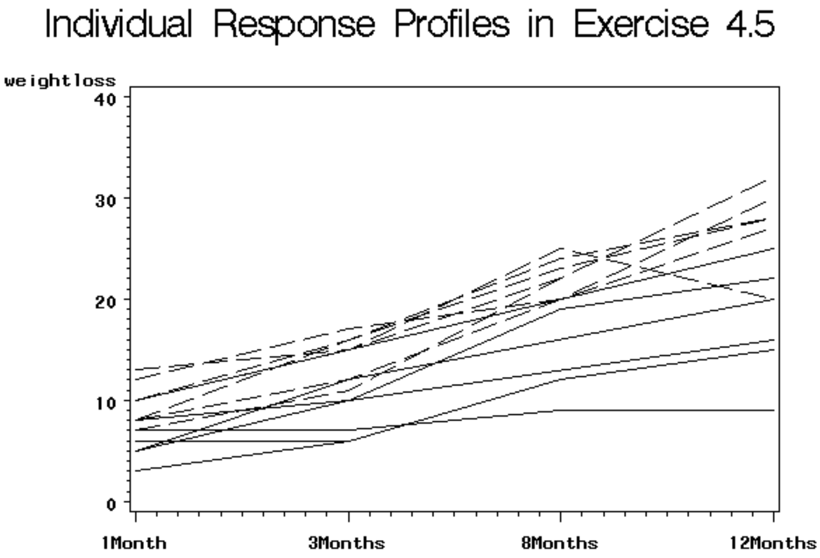
(c) The estimated values of the coefficients reveal that at every given week, an average score in the treatment group is about 4.75 points lower than that for the treatment group, and that the average score in either group decreases by roughly 4.75 points every week.

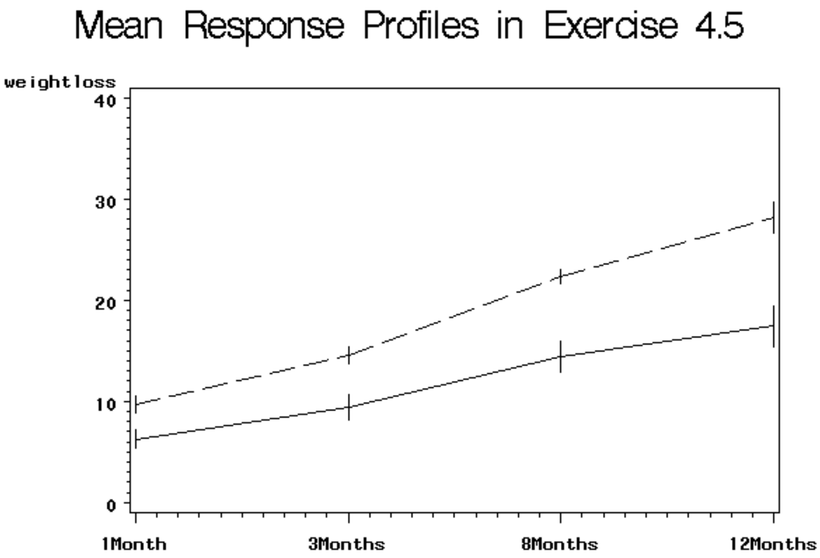
Since the subjects in the treatment group have a lower level of mental distress than those in the control group, the new drug is effective.

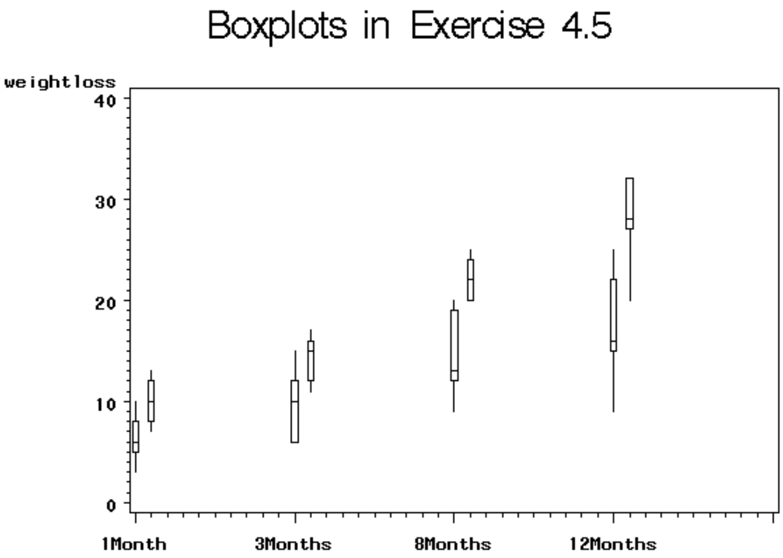
Section 4.4

Exercise 4.5 (a) The individual and mean response profiles, as well as the boxplots are presented below. The solid lines correspond to group 1 (open surgery), and the dashed ones, to group 2 (laparoscopic surgery). Also, the left-hand boxplots are for group 1, and the right-hand ones, for group 2.

As seen in the graphs, the subjects who undergo the laparoscopic surgery tend to lose more weight than those with the open surgery procedure. In addition, the observed responses show increasing variation as time progresses, hence the random slope and intercept model may be appropriate in this case.







(b) Let y_{ij} be the percentage loss of excess weight for the i-th subject at time t_j , $i = 1, \ldots, 14$, $j = 1, \ldots, 4$, where $t_1 = 1$, $t_2 = 3$, $t_3 = 8$, and $t_4 = 12$ months. Denote by $x_{1\,i}$ the group number (1 or 2) of the i-th subject. The random slope and intercept model has the form

$$y_{ij} = \beta_0 + \beta_1 \, x_{1\,i} + \beta_2 \, t_j + u_{i1} + u_{i2} \, t_j + \varepsilon_{ij} ,$$

where u_{i1} are the random intercepts, u_{i2} are the random slopes, and ε_{ij} are the random errors. The unknowns in this model are β_0, β_1, β_2, $\mathbb{V}ar(u_{i1}) = \sigma^2_{u_1}$, $\mathbb{V}ar(u_{i2}) = \sigma^2_{u_2}$, $\mathbb{V}ar(\varepsilon_{ij}) = \sigma^2$, and $\mathbb{C}ov\big(u_{i1} \, , \, u_{i2}\big) = \sigma_{u_1 \, u_2}$.

The SAS code that produces the graphs in part (a) and the parameter estimators is

```
data gastric_bypass;
input individual group visit1 visit3 visit8 visit12 @@;
```

```
datalines;
  1  1   5  12  16  20     2  1   7   7   9   9
  3  1   3   6  12  15     4  1  10  15  20  25
  5  1   8  10  13  16     6  1   5  10  19  22
  7  1   6   6  12  15     8  2   8  12  20  27
  9  2  10  15  22  32    10  2  12  17  20  30
 11  2   8  16  23  28    12  2   7  11  22  32
 13  2  10  16  24  28    14  2  13  15  25  20
;

data new;
set gastric_bypass;
array x{4} visit1 visit3 visit8 visit12;
array t{4} t1-t4 (1 3 8 12);
do visits = 1 to 4;
time = t{visits};
weightloss = x{visits};
if group = 1 then boxposition = visits;
else boxposition = visits + 0.1;
output;
end;
keep individual group visits time boxposition weightloss;
run;

axis1 label = none
value =(t=1 ‘1Month’ t=2 ‘3Months’ t=3 ‘8Months’ t=4 ‘12Months’
t=5 ‘’);

title’Individual Response Profiles in Exercise 4.5’;
proc gplot data = new;
plot weightloss * visits = individual / nolegend haxis = axis1;
symbol1 interpol = join value = none color = black
line = 1 repeat = 7;
symbol2 interpol = join value = none color = black
line = 2 repeat = 7;
run;

goptions reset = symbol;

title’Mean Response Profiles in Exercise 4.5’;
proc gplot data = new;
plot weightloss * visits = group / nolegend haxis = axis1;
symbol1 interpol = stdm1j color = black line = 1;
symbol2 interpol = stdm1j color = black line = 2;
run;

goptions reset = symbol;

title’Boxplots in Exercise 4.5’;
proc gplot data = new;
plot weightloss * boxposition = group / nolegend haxis = axis1;
symbol1 interpol = box value = star color = black;
```

```
symbol2 interpol = box value = x color = black;

run;


proc mixed data = new method = ml;

model weightloss = group time / solution;

random intercept time / subject = individual type = un;

run;


proc mixed data = new method = reml;

model weightloss = group time / solution;

random intercept time / subject = individual type = un;

run;
```

The parameter estimators are

	Effect	Estimate	Pr > |t|	
$\hat{\beta}_0 \rightarrow$	Intercept	-0.34440	0.8504	ML method
$\hat{\beta}_0 \rightarrow$	Intercept	-0.34439	0.8610	REML method
$\hat{\beta}_1 \rightarrow$	group	5.1534	<.0001	ML method
$\hat{\beta}_1 \rightarrow$	group	5.1533	0.0001	REML method
$\hat{\beta}_2 \rightarrow$	time	1.3166	<.0001	

Covariance Parameter	Estimate ML method	Estimate REML method	
UN(1,1)	4.7547	5.6920	$\leftarrow \hat{\sigma}^2_{u_1}$
UN(2,1)	-0.6481	-0.7201	$\leftarrow \hat{\sigma}_{u_1\,u_2}$
UN(2,2)	0.2231	0.2439	$\leftarrow \hat{\sigma}^2_{u_2}$
Residual	3.5492	3.5493	$\leftarrow \hat{\sigma}^2$

(c) The P-value for the intercept for both estimation methods is larger than 0.05, meaning that the intercept is indistinguishable from zero. The covariates `group` and `time` have a significant effect on weight loss, since the respective P-values are smaller than 0.05. As the estimated regression coefficients indicate, at every fixed visit time, subjects in group 2 (laparoscopic surgery) lose about 5.15 percent more excessive weight than the group 1 subjects (open surgery). This implies that the laparoscopic surgery is more effective than the open surgery. Also, the average monthly excessive weight loss is about 1.32 percent.

Section 4.5

Exercise 4.6 It is assumed that w_i has constant mean and variance. From (4.14),

$$\mathbb{E}\big(w_i(t_j)\big) \;=\; \rho\,\mathbb{E}\big(w_i(t_{j-1})\big) \;+\; \cancelto{0}{\mathbb{E}\big(z_i(t_j)\big)}.$$

Hence, $\mathbb{E}\big(w_i(t_j)\big) \;=\; 0$. From (4.14) again, and by independence of $w_i(t_j - 1)$

and $z_i(t_j)$,

$$\mathbb{V}ar\big(w_i(t_j)\big) \;=\; \rho^2\,\mathbb{V}ar\big(w_i(t_j\,-\,1)\big) \,+\, \mathbb{V}ar\big(z_i(t_j)\big)$$

$$=\; \rho^2\,\mathbb{V}ar(w_i(t_j\,-\,1)\big) \,+\, \big(1\,-\,\rho^2\big)\,\sigma^2\,,$$

therefore, $\mathbb{V}ar\big(w_i(t_j)\big) \;=\; \sigma^2$.

Exercise 4.7 (a) The mixed-effects model with spatial power covariance structure for the error terms is

$$y_{ij} \;=\; \beta_0 \,+\, \beta_1\, x_{1i} \,+\, \beta_2\, t_j \,+\, u_{i1} \,+\, u_{i2}\, t_j \,+\, w_i(t_j)\,,$$

where y_{ij} denotes the percentage of excess weight loss for the i-th subject at time t_j, x_{1i} is the group (1 or 2) of the i-th subject, $t_1 \;=\; 1$, $t_2 \;=\; 3$, $t_3 \;=\; 8$, and $t_4 \;=\; 12$ months, $i \;=\; 1\,\ldots,\, 14$, and $j \;=\; 1,\,\ldots,\,4$. The variables u_{i1} and u_{i2} are the random intercept and slope, respectively. The error terms $w_i(t_j)$ have mean zero and the $56\,\times\,56$ block-diagonal covariance matrix with $4\,\times\,4$ blocks of the form

$$\sigma^2\begin{bmatrix} 1 & \rho^2 & \rho^7 & \rho^{11} \\ \rho^2 & 1 & \rho^5 & \rho^9 \\ \rho^7 & \rho^5 & 1 & \rho^4 \\ \rho^{11} & \rho^9 & \rho^4 & 1 \end{bmatrix}.$$

The parameters of this model are β_0, β_1, β_2, $\sigma^2_{u_1}$, $\sigma^2_{u_2}$, $\sigma_{u_1\,u_2}$, σ^2, and ρ. The SAS program that estimates these parameters using the ML and REML methods is as follows. It utilizes the dataset `new` defined in Exercise 4.5.

```
proc mixed data = new method = ml;

model weightloss = group time / solution;

random intercept time / subject = individual type = un;

repeated / subject = individual type = sp(pow)(time);

run;


proc mixed data = new method = reml;

model weightloss = group time / solution;

random intercept time / subject = individual type = un;

repeated/ subject = individual type = sp(pow)(time);

run;
```

The results are

ML method

```
                Effect        Estimate  Pr > |t|

$\hat{\beta}_0 \rightarrow$  Intercept      -0.7006    0.6948

$\hat{\beta}_1 \rightarrow$  group           5.2282    <.0001

$\hat{\beta}_2 \rightarrow$  time            1.3342    <.0001
```

```
Covariance

Parameter   Estimate

UN(1,1)            0  $\leftarrow \hat{\sigma}^2_{u_1}$

UN(2,1)      -0.4367  $\leftarrow \hat{\sigma}_{u_1 u_2}$

UN(2,2)       0.2034  $\leftarrow \hat{\sigma}^2_{u_2}$

SP(POW)       0.8054  $\leftarrow \hat{\rho}$

Residual      7.2830  $\leftarrow \hat{\sigma}^2$
```

REML method

```
                Effect         Estimate  Pr > |t|

$\hat{\beta}_0 \rightarrow$  Intercept       -0.6937    0.7163

$\hat{\beta}_1 \rightarrow$  group            5.2058    <.0001

$\hat{\beta}_2 \rightarrow$  time             1.3354    <.0001
```

```
Covariance
Parameter   Estimate
UN(1,1)            0  $\leftarrow \hat{\sigma}^2_{u_1}$
UN(2,1)      -0.4677  $\leftarrow \hat{\sigma}_{u_1 u_2}$
UN(2,2)       0.2202  $\leftarrow \hat{\sigma}^2_{u_2}$
SP(POW)       0.8238  $\leftarrow \hat{\rho}$
Residual      7.9828  $\leftarrow \hat{\sigma}^2$
```

(b) This model and the random slope and intercept model have very close estimates of β_1 and β_2, and the coefficient β_0 is insignificantly small. The striking difference between these two models is that in the present model the variance of the random intercept $\sigma^2_{u_1}$ is estimated as zero, and the variation in the data that was explained by $\sigma^2_{u_1}$ and σ^2 in the other model is now explained by σ^2 and ρ.

Section 4.6

EXERCISE 4.8 (a) The random intercept logistic regression model for these data has the form

$$\pi_{ij}\,(u_i) \;=\; \frac{\exp\left\{\beta_0 \,+\, \beta_1\, x_{1\,i} \,+\, \beta_2\, t_j \,+\, u_i\right\}}{1 \,+\, \exp\left\{\beta_0 \,+\, \beta_1\, x_{1\,i} \,+\, \beta_2\, t_j \,+\, u_i\right\}}\,,$$

where $\pi_{ij}\,(u_i)$ is the probability of the i-th subject having toenail fungus present on the j-th visit, conditioned on the value of the random intercept u_i, $i = 1, \ldots, 11$, $j = 1, \ldots, 4$. The covariate $x_{1\,i}$ is the group (1 or 2) of the i-th subject, and $t_1 = 3$, $t_2 = 6$, $t_3 = 12$, and $t_4 = 16$ weeks.

The parameters of the model are β_0, β_1, β_2, and σ^2_u. To compute the maximum likelihood estimators of the parameters, use the SAS code

```
data toenail_fungus;

input individual group week3 week6 week12 week16 @@;

datalines;
  1  1  1  1  0  1        2  1  1  1  1  0

  3  1  1  1  0  0        4  1  1  0  0  0

                    ...

 21  2  1  0  0  0       22  2  1  1  1  1

;


data new;

set toenail_fungus;

array x{4} week3 week6 week12 week16;

array t{4} (3 6 12 16);

do visits = 1 to 4;

fungus = x{visits};

time = t{visits};

output;

end;

keep individual group fungus time;

run;


proc glimmix data = new;

model fungus(event = "1") = group time / solution

dist = binary link = logit;
```

```
random intercept / subject = individual type = un;
run;
```

The result is

```
                   Effect      Estimate  Pr > |t|
$\hat{\beta}_0 \to$  Intercept     4.3131    0.0103
$\hat{\beta}_1 \to$  group        -1.3210    0.1260
$\hat{\beta}_2 \to$  time         -0.2770    <.0001

                 Covariance
                 Parameter   Estimate
                 UN(1,1)       2.2111  $\leftarrow \hat{\sigma}_u^2$
```

Note that since the P-value for $\hat{\beta}_1$ is larger than 0.05, the variable `group` is an insignificant covariate, implying that the tested treatment is NOT effective.

(b) After removing the variable `group`, the estimate of the regression coefficient for the variable `time` becomes -0.2649, indicating that the weekly percentage change in odds of having fungus is $100\big(\exp\{-0.2649\} - 1\big)\,\% = -\,23.27\,\%$, a decrease of 23.27%.

Section 4.7

Exercise 4.9 (a) The data code for this exercise is

```
data recorded;
input individual group size3 size6 size12
size18 size24 score3 score6 score12 score18 score24 @@;
datalines;
  1  1  3.0  2.7  2.3  2.1  1.8  90  85  70  67  63
  2  1  2.9  2.4  1.8  1.7  0.2  87  90  90  90  90
  3  1  2.4  2.3   .    .    .   78  67   .   .   .
                         ...
 22  2  2.4  2.1  2.1  1.7  1.7  85  75  41  37  33
 23  2  3.4  2.3  1.4  0.9  0.0  74  70  64  63  42
 24  2  2.4  2.0  1.0  0.3  0.0  96  67  78  74  55
;

data unbalanced;
set recorded;
array x{5} size3 size6 size12 size18 size24;
array z{5} score3 score6 score12 score18 score24;
array t{5} (3 6 12 18 24);
do visits = 1 to 5;
tumorsize = x{visits};
time = t{visits};
score = z{visits};
output;
end;
keep individual group tumorsize score time;
run;
```

```
data complete;
set unbalanced;
if individual = 3 then delete;
if individual = 8 then delete;
if individual = 13 then delete;
if individual = 16 then delete;
if individual = 19 then delete;
run;

data imputed;
set unbalanced;
     if individual = 3 then do;
       if tumorsize = .  then tumorsize = 2.3;
       if score = .  then score = 67;
     end;
       if individual = 8 then do;
         if tumorsize = .  then tumorsize = 2.575;
         if score = .  then score = 90.5;
       end;
     if individual = 13 then do;
       if tumorsize = .  then tumorsize = 2.0;
       if score = .  then score = 63;
     end;
       if individual = 16 then do;
         if tumorsize = .  then tumorsize = 3.1;
         if score = .  then score = 85;
       end;
     if individual = 19 then do;
       if time = 6 then do;
         tumorsize = 2.59;
         score = 75.59;
       end;
       if time = 12 then do;
         tumorsize = 2.165;
         score = 65.7;
       end;
     end;

run;

proc mixed data = unbalanced method = reml;
model score = group tumorsize time / solution;
random intercept time / subject = individual type = un;
run;

proc mixed data = complete method = reml;
model score = group tumorsize time / solution;
random intercept time / subject = individual type = un;
run;

proc mixed data = imputed method = reml;
```

```
model score = group tumorsize time / solution;

random intercept time / subject = individual type = un;

run;
```

The estimated regression coefficients in the full model described in Example 4.4 are

UNBALANCED DATASET

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	95.8035	<.0001
$\hat{\beta}_1 \rightarrow$	group	-2.6659	0.2983
$\hat{\beta}_2 \rightarrow$	tumorsize	-1.6992	0.2347
$\hat{\beta}_3 \rightarrow$	time	-1.5779	<.0001

COMPLETE DATASET

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	98.4185	<.0001
$\hat{\beta}_1 \rightarrow$	group	-3.7938	0.1400
$\hat{\beta}_2 \rightarrow$	tumorsize	-2.1314	0.1597
$\hat{\beta}_3 \rightarrow$	time	-1.5967	<.0001

IMPUTED DATASET

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	91.8824	<.0001
$\hat{\beta}_1 \rightarrow$	group	-3.4728	0.1984
$\hat{\beta}_2 \rightarrow$	tumorsize	-0.3206	0.8176
$\hat{\beta}_3 \rightarrow$	time	-1.2976	<.0001

(b) One striking difference between the above models and the full-data model in Example 4.4 is that in these three models the variable `group` is insignificant, whereas in the latter model it is a significant covariate.

(c) The parameter estimates in the reduced model with insignificant variables `group` and `tumorsize` removed are

```
                         UNBALANCED   DATASET

                         Effect         Estimate   Pr > |t|
$\hat{\beta}_0 \rightarrow$   Intercept       86.4128     <.0001
$\hat{\beta}_3 \rightarrow$   time            -1.4451     <.0001

                      Covariance
                      Parameter    Estimate
                      UN(1,1)        8.4446   $\leftarrow \hat{\sigma}^2_{u_1}$
                      UN(2,1)        0.1419   $\leftarrow \hat{\sigma}_{u_1\,u_2}$
                      UN(2,2)        0.6425   $\leftarrow \hat{\sigma}^2_{u_2}$
                      Residual      43.8996   $\leftarrow \hat{\sigma}^2$
```

```
                         COMPLETE   DATASET

                         Effect      Estimate   Pr > |t|
$\hat{\beta}_0 \rightarrow$   Intercept    86.0504     <.0001
$\hat{\beta}_3 \rightarrow$   time         -1.4267     <.0001
```

```
      Covariance
      Parameter   Estimate
      UN(1,1)       3.9146   ← $\hat{\sigma}^2_{u_1}$
      UN(2,1)      -0.3043   ← $\hat{\sigma}_{u_1 u_2}$
      UN(2,2)       0.6667   ← $\hat{\sigma}^2_{u_2}$
      Residual     47.1345   ← $\hat{\sigma}^2$
```

```
         IMPUTED       DATASET
         Effect      Estimate   Pr > |t|
$\hat{\beta}_0$ →   Intercept    85.6606     <.0001
$\hat{\beta}_3$ →   time         -1.2748     <.0001

      Covariance
      Parameter   Estimate
      UN(1,1)      20.1029   ← $\hat{\sigma}^2_{u_1}$
      UN(2,1)      -0.8331   ← $\hat{\sigma}_{u_1 u_2}$
      UN(2,2)       0.7089   ← $\hat{\sigma}^2_{u_2}$
      Residual     39.7188   ← $\hat{\sigma}^2$
```

The estimators of the coefficients, $\hat{\sigma}^2_{u_2}$, and $\hat{\sigma}^2$ do not variate much in these models, whereas the values of $\hat{\sigma}^2_{u_1}$ and $\hat{\sigma}_{u_1 u_2}$ change drastically.

Exercise 4.10 (a) The SAS code for this exercise is

```
data recorded;
input individual age calcium history mos3 mos9 mos12 mos18 @@;
datalines;
  1  76  0  1  1  1  1  1     2  57  1  1  1  1  0  0
  3  58  0  1  1  1  1  0     4  62  1  0  1  0  0  0
  5  60  1  0  0  0  .  .     6  58  0  1  1  0  1  1
  7  52  1  0  0  1  0  0     8  74  0  1  1  0  1  0
  9  51  0  0  0  1  .  .    10  56  1  0  0  1  0  0
 11  75  0  1  1  1  1  1    12  63  1  0  1  0  .  0
 13  67  1  1  0  1  0  0    14  68  0  0  1  1  0  0
 15  56  1  0  1  0  .  .    16  62  1  0  1  0  1  1
 17  60  1  1  0  0  0  0    18  61  1  1  1  .  0  0
 19  54  1  0  1  0  0  0    20  53  0  0  1  1  0  0
;

data unbalanced;
set recorded;
array x{4} mos3 mos9 mos12 mos18;
array t{4} (3 9 12 18);
do visits=1 to 4;
disease=x{visits};
time=t{visits};
output;
end;
keep individual age calcium history disease time;
run;
```

```
data complete;
set unbalanced;
if individual = 5 then delete;
if individual = 9 then delete;
if individual = 12 then delete;
if individual = 15 then delete;
if individual = 18 then delete;
run;

data imputed;
set unbalanced;
     if individual = 5 then do;
       if disease = .  then disease = 0;
     end;
       if individual = 9 then do;
         if disease = .  then disease = 1;
       end;
     if individual = 12 then do;
       if time = 12 then disease = 0;
     end;
       if individual = 15 then do;
         if disease = .  then disease = 0;
       end;
     if individual = 18 then do;
       if time = 9 then disease = 1;
     end;
run;


proc glimmix data = unbalanced;
model disease(event = "1") = calcium time / solution
dist = binary link = logit;
random intercept / subject = individual type = un;
run;


proc glimmix data = complete;
model disease(event = "1") = calcium time / solution
dist = binary link = logit;
random intercept / subject = individual type = un;
run;


proc glimmix data = imputed;
model disease(event = "1") = calcium time / solution
dist = binary link = logit;
random intercept / subject = individual type = un;
run;
```

The estimated regression coefficients in the full model described in Example 4.6 are

UNBALANCED DATASET

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	-0.6031	0.8395
$\hat{\beta}_1 \rightarrow$	age	0.05124	0.2887
$\hat{\beta}_2 \rightarrow$	calcium	-1.6821	0.0129
$\hat{\beta}_3 \rightarrow$	history	0.3494	0.5998
$\hat{\beta}_4 \rightarrow$	time	-0.1859	0.0028

COMPLETE DATASET

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	-0.2605	0.9422
$\hat{\beta}_1 \rightarrow$	age	0.04888	0.3843
$\hat{\beta}_2 \rightarrow$	calcium	-1.7259	0.0371
$\hat{\beta}_3 \rightarrow$	history	0.1492	0.8511
$\hat{\beta}_4 \rightarrow$	time	-0.1823	0.0072

IMPUTED DATASET

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	1.2780	0.6569
$\hat{\beta}_1 \rightarrow$	age	0.02255	0.6227
$\hat{\beta}_2 \rightarrow$	calcium	-1.9835	0.0024
$\hat{\beta}_3 \rightarrow$	history	0.4704	0.4630
$\hat{\beta}_4 \rightarrow$	time	-0.1775	0.0024

(b) As in the full model of Example 4.6, the covariates `age` and `history` are insignificant in all of the above models.

(c) The parameter estimates in the reduced model for the three datasets are

```
                         UNBALANCED    DATASET

                         Effect        Estimate  Pr > |t|

$\hat{\beta}_0 \rightarrow$   Intercept       2.7712    0.0029

$\hat{\beta}_2 \rightarrow$   calcium        -2.0010    0.0017

$\hat{\beta}_4 \rightarrow$   time           -0.1728    0.0037


                        Covariance

                        Parameter   Estimate

                        UN(1,1)      0.01407  $\leftarrow \hat{\sigma}_u^2$
```

```
                         COMPLETE    DATASET

                         Effect      Estimate  Pr > |t|

$\hat{\beta}_0 \rightarrow$   Intercept     2.9552    0.0069

$\hat{\beta}_2 \rightarrow$   calcium      -2.0542    0.0048

$\hat{\beta}_4 \rightarrow$   time         -0.1764    0.0077


                        Covariance

                        Parameter   Estimate

                        UN(1,1)       0.1718  $\leftarrow \hat{\sigma}_u^2$
```

IMPUTED DATASET

```
                    Effect      Estimate   Pr > |t|
$\hat{\beta}_0 \rightarrow$   Intercept     2.9458     0.0019
$\hat{\beta}_2 \rightarrow$   calcium      -2.1819     0.0004
$\hat{\beta}_4 \rightarrow$   time         -0.1738     0.0025

                  Covariance
                  Parameter   Estimate
                  UN(1,1)            0   $\leftarrow \hat{\sigma}_u^2$
```

The estimators of the coefficients and the variance of the random intercept do not differ too much in these three models as well as in the reduced model in Example 4.6. The only point worth mentioning is that $\hat{\sigma}_u^2$ for the imputed dataset happened to be zero.