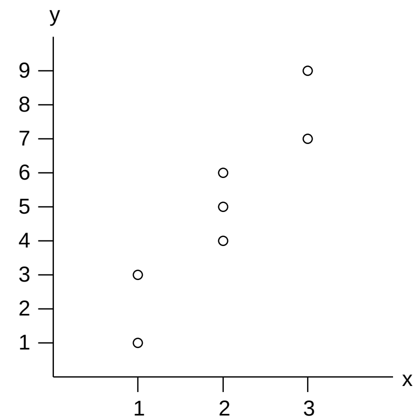

Figure 8.7 (Exercise 8.15) Three treatments, 1, 2 and 3; seven observations.

8.15. Consider the scatter plot of Figure 8.7, where we have seven data points and we have three treatments whose effect on y we wish to study using one-way ANOVA. Show that the F-statistic for testing the hypothesis $H\colon \mu_1 = \mu_2 = \mu_3$ has value $F = 12$ with $g - 1 = 2$ and $n - g = 4$ degrees of freedom, and

$$\mathrm{SST} = \mathrm{SS}_{\mathrm{Between}} + \mathrm{SS}_{\mathrm{Within}}\colon \quad 42 = 36 + 6\,.$$

• Solution to Ex. 8.15:

We have the following linear model:

$$\mathscr{M}_*\colon \quad \mathbf{y} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,,$$

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_2 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,.$$

Moreover,

$$\mathbf{H} = \begin{pmatrix} \mathbf{J}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_2 \end{pmatrix}, \quad \mathbf{M} = \mathbf{I}_7 - \mathbf{H} = \begin{pmatrix} \mathbf{C}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_2 \end{pmatrix},$$

$$\mathbf{Hy} = \begin{pmatrix} \bar{y}_1 \mathbf{1}_2 \\ \bar{y}_2 \mathbf{1}_3 \\ \bar{y}_3 \mathbf{1}_2 \end{pmatrix}, \quad \mathbf{My} = \begin{pmatrix} \mathbf{C}_2 \mathbf{y}_1 \\ \mathbf{C}_3 \mathbf{y}_2 \\ \mathbf{C}_2 \mathbf{y}_3 \end{pmatrix}, \quad (\mathbf{H} - \mathbf{J}_7)\mathbf{y} = \begin{pmatrix} (\bar{y}_1 - \bar{y})\mathbf{1}_2 \\ (\bar{y}_2 - \bar{y})\mathbf{1}_3 \\ (\bar{y}_3 - \bar{y})\mathbf{1}_2 \end{pmatrix},$$

$$\mathrm{SST} = \mathbf{y}'\mathbf{C}_7\mathbf{y} = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2, \quad \mathrm{SSE} = \mathbf{y}'\mathbf{My} = \mathrm{SS}_1 + \mathrm{SS}_2 + \mathrm{SS}_g\,,$$

$$\mathrm{SSR} = \mathbf{y}'(\mathbf{H} - \mathbf{J}_7)\mathbf{y} = \sum_{i=1}^{3} n_i(\bar{y}_i - \bar{y})^2 := \mathrm{SSB}\,, \quad \mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}\,,$$

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{3} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2,$$
$$\mathrm{SST} = \mathrm{SS}_{\mathrm{Between}} + \mathrm{SS}_{\mathrm{Within}}\,,$$
$$\mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} = \mathbf{y}'(\mathbf{H} - \mathbf{J})\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}\,.$$

Under the hypothesis $H\colon \mu_1 = \mu_2 = \mu_3$ we have the model

$$\text{model under } H\colon \quad \mathbf{y} = \mathbf{1}\mu + \boldsymbol{\varepsilon}\,,$$

and hence $\mathrm{SSE}_H = \mathrm{SST}$ and $\mathrm{SSE}_H - \mathrm{SSE} = \mathrm{SSB}$, and, under normality, the F-test statistics for testing H is

$$\begin{aligned} F &= \frac{(\mathrm{SSE}_H - \mathrm{SSE})/(3-1)}{\mathrm{SSE}\,/(7-3)} \\ &= \frac{\mathbf{y}'(\mathbf{H} - \mathbf{J}_7)\mathbf{y}/2}{\mathbf{y}'\mathbf{My}/4} = \frac{\mathrm{SSB}\,/2}{\mathrm{SSE}\,/4} \sim \mathrm{F}(3-1, 7-3)\,. \end{aligned}$$

Notice:

$$\bar{y}_1 = 2, \quad \bar{y}_2 = 5, \quad \bar{y}_3 = 8, \quad \bar{y} = 5,$$

$$\begin{aligned} \mathrm{SSE} &= \mathrm{SS}_{\mathrm{Within}} = 2 + 2 + 2 = 6, \\ \mathrm{SSR} &= \mathrm{SSB} = \mathrm{SS}_{\mathrm{Between}} = 2(2-5)^2 + 3(5-5)^2 + 2(8-5)^2 = 36, \\ \mathrm{SST} &= (1-5)^2 + (3-5)^5 + (4-5)^2 + (6-5)^5 + (7-5)^2 + (9-5)^2 \\ &= \mathrm{SSR} + \mathrm{SSE} = 6 + 36 = 42, \\ F &= 12. \end{aligned}$$

□

8.16 (Continued ...). Suppose that the scatter plot of Figure 8.7 represents a two-dimensional discrete uniform distribution of random variables x

and y. Let $\mathrm{E}(y \mid x)$ denote a random variable whose values are conditional means $\mathrm{E}(y \mid x = 1)$, $\mathrm{E}(y \mid x = 2)$, $\mathrm{E}(y \mid x = 3)$ with probabilities $\mathrm{P}(x = 1)$, $\mathrm{P}(x = 2)$, $\mathrm{P}(x = 3)$, respectively. Define the random variable $\mathrm{var}(y \mid x)$ in the corresponding way and show that

$$\mathrm{var}(y) = \mathrm{var}[\mathrm{E}(y \mid x)] + \mathrm{E}[\mathrm{var}(y \mid x)]\,. \tag{8.196}$$

Casella (2008, p. 9) calls (8.196) as "a most important equality" in his book on *Statistical Design*. Confirm the connection between (8.196) and the ANOVA decomposition $\mathrm{SST} = \mathrm{SSB} + \mathrm{SSE}$.

NOTICE: O'Hagan (2012) refers to (8.196) as "my thing of beauty".
O'Hagan, T. (2012). A thing of beauty. *Significance*, 9, 26–28.

• SOLUTION TO EX. 8.16:

Let us first show that

$$\mathrm{E}(y) = \mathrm{E}[\mathrm{E}(y \mid x)]\,.$$

Now we consider $\mathrm{E}(y \mid x)$ as a random variable such that

$$\begin{aligned} \text{values of } \mathrm{E}(y \mid x)\text{:}\quad & 2,\ 5,\ 8, \\ \text{probabilities:}\quad & \frac{2}{7}\,,\ \frac{3}{7}\,,\ \frac{2}{7}\,, \\ \mathrm{E}[\mathrm{E}(y \mid x)] &= \frac{2}{7}\,2 + \frac{3}{7}\,5 + \frac{2}{7}\,8 = 5, \\ \mathrm{E}(y) &= \frac{1}{7}\,(1+3+4+5+6+7+9) = 5. \end{aligned}$$

Consider then the variance:

$$\begin{aligned} \mathrm{var}(y) &= \frac{1}{7}\,[(1-5)^2 + (3-5)^2 + (4-5)^2 + (6-5)^2 + (7-5)^2 + (9-5)^2] \\ &= \frac{1}{7}\,42 = \frac{1}{7}\,\mathrm{SST}\,. \end{aligned}$$

Because the values of $\mathrm{E}(y \mid x)$ are $2,\ 5,\ 8$, with probabilities $\frac{2}{7}\,,\ \frac{3}{7}\,,\ \frac{2}{7}\,$, the variance of random variable $\mathrm{E}(y \mid x)$ is

$$\begin{aligned} \mathrm{var}[\mathrm{E}(y \mid x)] &= \frac{2}{7}\,(2-5)^2 + \frac{3}{7}\,(5-5)^2 + \frac{2}{7}\,(8-5)^2 \\ &= \frac{1}{7}\,[2(2-5)^2 + 3(5-5)^2 + 2(8-5)^2] \\ &= \frac{1}{7}\,36 = \frac{1}{7}\,\mathrm{SS}_{\mathrm{Between}}\,. \end{aligned}$$

Similarly we have for the random variable $\mathrm{var}(y \mid x)$:

$$\begin{aligned}
\text{value of } \operatorname{var}(y \mid x = 1)\colon &\quad \frac{1}{2}\,[(1-2)^2 + (3-2)^2] = \frac{1}{2}\,2 = \frac{1}{2}\,\mathrm{SS}_1 \\
\text{value of } \operatorname{var}(y \mid x = 2)\colon &\quad \frac{1}{3}\,[(4-5)^2 + (5-5)^2 + (6-5)^2] = \frac{1}{3}\,2 = \frac{1}{3}\,\mathrm{SS}_2 \\
\text{value of } \operatorname{var}(y \mid x = 3)\colon &\quad \frac{1}{2}\,[(7-8)^2 + (9-8)^2] = \frac{1}{2}\,2 = \frac{1}{2}\,\mathrm{SS}_3 \\
\text{probabilities:} &\quad \frac{2}{7}\,,\ \frac{3}{7}\,,\ \frac{2}{7}\,,
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[\operatorname{var}(y \mid x)] &= \frac{2}{7} \cdot \frac{1}{2}\,\mathrm{SS}_1 + \frac{3}{7} \cdot \frac{1}{3}\,\mathrm{SS}_2 + \frac{2}{7} \cdot \frac{1}{2}\,\mathrm{SS}_3 \\
&= \frac{1}{7}\,(\mathrm{SS}_1 + \mathrm{SS}_2 + \mathrm{SS}_3) = \frac{1}{7}\ \mathrm{SS}_{\text{Within}}
\end{aligned}$$

Hence we have

$$\begin{aligned}
\operatorname{var}(y) &= \operatorname{var}[\mathrm{E}(y \mid x)] + \mathrm{E}[\operatorname{var}(y \mid x)]\,, \\
\frac{1}{7}\ \mathrm{SST} &= \frac{1}{7}\ \mathrm{SS}_{\text{Between}} + \frac{1}{7}\ \mathrm{SS}_{\text{Within}}\,.
\end{aligned}$$

□
