Exercises: Some Solutions (October 3, 2012)

These exercises are supposed to be warm-up exercises—to get rid of the possible rust in the reader's matrix engine.

0.1 (Contingency table, 2×2). Consider three frequency tables (contingency tables) below. In each table the row variable is x.

(a) Write up the original data matrices and calculate the correlation coefficients r_{xy}, r_{xz} and r_{xu}.

(b) What happens if the location (cell) of the zero frequency changes while other frequencies remain mutually equal?

(c) Explain why $r_{xy} = r_{xu}$ even if the u-values 2 and 5 are replaced with arbitrary a and b such that $a < b$.

		y: 0	y: 1		z: 0	z: 1		u: 2	u: 5
x	0	1	1	0	2	2	0	1	1
	1	0	1	1	0	2	1	0	1

• Solution to Ex. 0.1:

$$\mathbf{A} = (\mathbf{x} : \mathbf{y}) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{B} = (\mathbf{x} : \mathbf{z}) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{C} = (\mathbf{x} : \mathbf{u}) = \begin{pmatrix} 0 & 2 \\ 0 & 5 \\ 5 & 5 \end{pmatrix}.$$

In all cases the correlation coefficient is 0.5. Recall that

$$\begin{aligned} r_{xy} &= \frac{\mathrm{SP}_{xy}}{\sqrt{\mathrm{SS}_x \, \mathrm{SS}_y}} = \frac{s_{xy}}{s_x s_y}\,, \\ \mathrm{SS}_x &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \tfrac{1}{n}\Big(\sum_{i=1}^n x_i\Big)^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2, \\ \mathrm{SP}_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \tfrac{1}{n}\Big(\sum_{i=1}^n x_i\Big)\Big(\sum_{i=1}^n y_i\Big) \\ &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\,. \end{aligned}$$

NOTE: In **A** the variables x and y have the same variances, in which case the correlation coefficient and the slope $\hat{\beta}_1$ are identical:

$$\hat{\beta}_1 = \frac{\mathrm{SP}_{xy}}{\mathrm{SS}_x} = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x} = r_{xy}, \text{ if } s_x = s_y\,.$$

If the cell of the zero-frequency changes, then $|r_{xy}|$ remains the same but the sign may change. In the following cases $r_{xy} = -0.5$:

$$x\ \begin{matrix} & y \\ & \begin{matrix} 0 & 1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \end{matrix} & \boxed{\begin{matrix} 0 & 1 \\ 1 & 1 \end{matrix}} \end{matrix} \qquad x\ \begin{matrix} & y \\ & \begin{matrix} 0 & 1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \end{matrix} & \boxed{\begin{matrix} 1 & 1 \\ 1 & 0 \end{matrix}} \end{matrix}$$

(c) We try to find to find real numbers α and β which have the property $u_i = \alpha + \beta y_i, i = 1, 2, 3$, i.e.,

$$a = \alpha + \beta \cdot 0, \quad b = \alpha + \beta \cdot 1 \implies \alpha = a,\ \beta = b - a \implies u_i = a + (b - a)y_i.$$

Because the u-values are obtained by a linear transformation $u_i = a+(b-a)y_i$, where $b - a > 0$, we necessarily have $r_{xu} = r_{xy}$.

If the transformation were $u_i = c + dy_i$, where $d < 0$, then $r_{xu} = -r_{xy}$. □

0.2. Prove the following results concerning two dichotomous variables whose observed frequency table is given below.

$$\mathrm{var_s}(y) = \frac{1}{n-1}\frac{\gamma\delta}{n} = \frac{n}{n-1}\cdot\frac{\delta}{n}\Big(1 - \frac{\delta}{n}\Big),$$

$$\mathrm{cov_s}(x, y) = \frac{1}{n-1}\frac{ad - bc}{n}, \quad \mathrm{cor_s}(x, y) = \frac{ad - bc}{\sqrt{\alpha\beta\gamma\delta}} = r\,,$$

$$\chi^2 = \frac{n(ad - bc)^2}{\alpha\beta\gamma\delta} = nr^2.$$

		y = 0	y = 1	total
x	0	a	b	α
	1	c	d	β
total		γ	δ	n

• SOLUTION TO EX. 0.2:

$$\bar{y} = \frac{\delta}{n}\,, \quad \bar{x} = \frac{\beta}{n}\,,$$

$$\begin{aligned}
\mathrm{var_s}(y) &= \frac{1}{n-1}\Big(\sum_{i=1}^n y_i^2 - n\bar{y}^2\Big) = \frac{1}{n-1}\Big(\delta - n\,\frac{\delta^2}{n^2}\Big) \\
&= \frac{1}{n-1}\,\delta\Big(1 - \frac{\delta}{n}\Big) \\
&= \frac{n}{n-1}\,\frac{\delta}{n}\,\Big(1 - \frac{\delta}{n}\Big) = \frac{n}{n-1}\,\frac{\delta\gamma}{n^2}\,, \\
\mathrm{cov_s}(x, y) &= \frac{1}{n-1}\Big(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\Big) = \frac{1}{n-1}\Big(d - n\,\frac{\beta}{n}\,\frac{\delta}{n}\Big)
\end{aligned}$$

$$= \frac{n}{n-1}\left(\frac{d}{n} - \frac{\beta}{n}\frac{\delta}{n}\right).$$

In view of

$$\begin{aligned}\frac{d}{n} - \frac{\beta}{n}\cdot\frac{\delta}{n} &= \frac{1}{n^2}\left[nd - (c+d)(b+d)\right] \\ &= \frac{1}{n^2}\left[(a+b+c+d)d - (cb+cd+db+d^2)\right] \\ &= \frac{1}{n^2}(ad-bc),\end{aligned}$$

we get

$$\begin{aligned}\operatorname{cov}_{\mathrm{s}}(x,y) &= \frac{n}{n-1}\,\frac{ad-bc}{n^2} \\ &= \frac{1}{n-1}\,\frac{ad-bc}{n}.\end{aligned}$$

NOTE: According to (0.131) and (0.133), we can consider a data matrix $\mathbf{U} = (\mathbf{u}_{(1)} : \ldots : \mathbf{u}_{(n)})'$ and define a discrete random vector $\mathbf{u}_*$ with probability function

$$\mathrm{P}(\mathbf{u}_* = \mathbf{u}_{(i)}) = \tfrac{1}{n}, \quad i = 1, \ldots, n,$$

i.e., every data point has the same probability to be the value of the random vector $\mathbf{u}_*$. Then

$$\mathrm{E}(\mathbf{u}_*) = \bar{\mathbf{u}}, \quad \operatorname{cov}(\mathbf{u}_*) = \tfrac{1}{n}\mathbf{U}'\mathbf{C}\mathbf{U} = \tfrac{n-1}{n}\mathbf{S}.$$

Below in Exercise 0.3 the considerations are done for a 2-dimensional random vector which is obtained from the frequency table above so that each observation has the same probability $1/n$. □

0.3. Let $\mathbf{z} = \binom{x}{y}$ be a discrete 2-dimensional random vector which is obtained from the frequency table in Exercise 0.2 so that each observation has the same probability $1/n$. Prove that then

$$\mathrm{E}(y) = \frac{\delta}{n},\ \operatorname{var}(y) = \frac{\delta}{n}\left(1-\frac{\delta}{n}\right),\ \operatorname{cov}(x,y) = \frac{ad-bc}{n^2},\ \operatorname{cor}(x,y) = \frac{ad-bc}{\sqrt{\alpha\beta\gamma\delta}}.$$

• SOLUTION TO EX. 0.3:

$$\begin{aligned}\mathrm{E}(x) = \mu_x &= \frac{\alpha}{n}0 + \frac{\beta}{n}1 = \frac{\beta}{n}, \quad \mathrm{E}(y) = \frac{\delta}{n}, \\ \operatorname{var}(x) = \sigma_x^2 &= \mathrm{E}(x^2) - \mu_x^2 \\ &= \frac{\alpha}{n}0^2 + \frac{\beta}{n}1^2 - \frac{\beta^2}{n^2} = \frac{\beta}{n}\left(1-\frac{\beta}{n}\right) = \frac{\beta}{n}\frac{\alpha}{n},\end{aligned}$$

$$\operatorname{var}(y) = \frac{\delta}{n}\left(1 - \frac{\delta}{n}\right) = \frac{\gamma}{n}\frac{\delta}{n} = \frac{\gamma}{n}\left(1 - \frac{\gamma}{n}\right),$$

$$\begin{aligned}\operatorname{cov}(x, y) = \sigma_{xy} &= \mathrm{E}(xy) - \mu_x \mu_y \\ &= \frac{d}{n} - \frac{\beta}{n} \cdot \frac{\delta}{n} \\ &= \frac{1}{n^2}\,[nd - (c+d)(b+d)] \\ &= \frac{1}{n^2}\,[(a+b+c+d)d - (cb + cd + db + d^2)] \\ &= \frac{1}{n^2}\,(ad - bc)\,.\end{aligned}$$

□

0.4 (Continued ...). Show that in terms of the probabilities:

$$\begin{aligned}\operatorname{var}(y) &= p_{\cdot 1} p_{\cdot 2}\,, \\ \operatorname{cov}(x, y) &= p_{11} p_{22} - p_{12} p_{21}\,, \\ \operatorname{cor}(x, y) &= \frac{p_{11} p_{22} - p_{12} p_{21}}{\sqrt{p_{\cdot 1} p_{\cdot 2} p_{1\cdot} p_{2\cdot}}} = \varrho_{xy}\,.\end{aligned}$$

		y: 0	y: 1	total
x	0	p_{11}	p_{12}	$p_{1\cdot}$
	1	p_{21}	p_{22}	$p_{2\cdot}$
	total	$p_{\cdot 1}$	$p_{\cdot 2}$	1

• Solution to Ex. 0.4:

All probabilities are obtained from the table of Exercise 0.2 by dividing each figure by n. Hence

$$\mathrm{E}(x) = \frac{\beta}{n} = p_{2\cdot}\,, \quad \mathrm{E}(y) = \frac{\delta}{n} = p_{\cdot 2}\,,$$

$$\operatorname{var}(x) = \frac{\beta}{n}\left(1 - \frac{\beta}{n}\right) = \frac{\alpha}{n}\left(1 - \frac{\alpha}{n}\right) = p_{1\cdot} p_{2\cdot}\,,$$

$$\operatorname{var}(y) = \frac{\delta}{n}\left(1 - \frac{\delta}{n}\right) = \frac{\gamma}{n}\left(1 - \frac{\gamma}{n}\right) = p_{\cdot 1} p_{\cdot 2}\,,$$

$$\begin{aligned}\operatorname{cov}(x, y) &= \frac{d}{n} - \frac{\beta}{n} \cdot \frac{\delta}{n} \\ &= p_{22} - p_{2\cdot} p_{\cdot 2} \\ &= \frac{ad - bc}{n^2} \\ &= p_{11} p_{22} - p_{12} p_{21}\,.\end{aligned}$$

□

0.5 (Continued ...). Confirm:

$$\varrho_{xy} = 0 \Leftrightarrow \det\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = 0 \Leftrightarrow \frac{p_{11}}{p_{21}} = \frac{p_{12}}{p_{22}} \Leftrightarrow \frac{a}{c} = \frac{b}{d}\,.$$

0.6 (Continued ...). Show, using (0.85) (p. 19), that the dichotomous random variables x and y are statistically independent if and only if $\varrho_{xy} = 0$. By the way, for interesting comments on 2×2 tables, see Speed (2008b).

• SOLUTION TO EX. 0.6:

The random variables x and y are statistically independent if and only if

$$\mathrm{P}(x = i,\, y = j) = \mathrm{P}(x = i)\,\mathrm{P}(y = j) \quad \text{for all } i = 0, 1,\; j = 0, 1\,, \qquad (*)$$

i.e.,

$$p_{ij} = p_{i\cdot}p_{\cdot j} \quad \text{for all } i = 1, 2,\; j = 1, 2\,. \qquad (1)$$

while x and y are uncorrelated if and only if $\operatorname{cov}(x, y) = p_{22} - p_{2\cdot}p_{\cdot 2} = 0$, i.e.,

$$p_{22} = p_{2\cdot}p_{\cdot 2}\,. \qquad (2)$$

Suppose that (2) holds, i.e., x and y are uncorrelated. Then (2) implies that (1) holds for $i = j = 2$. Moreover, (2) implies

$$\begin{aligned} p_{21} &= p_{2\cdot} - p_{22} = p_{2\cdot} - p_{2\cdot}p_{\cdot 2} = p_{2\cdot}(1 - p_{\cdot 2}) = p_{2\cdot}p_{\cdot 1}\,, \\ p_{12} &= p_{\cdot 2} - p_{22} = p_{\cdot 2} - p_{2\cdot}p_{\cdot 2} = p_{\cdot 2}(1 - p_{2\cdot}) = p_{\cdot 2}p_{1\cdot}\,, \\ p_{11} &= p_{1\cdot} - p_{12} = p_{1\cdot} - p_{1\cdot}p_{\cdot 2} = p_{1\cdot}(1 - p_{\cdot 2}) = p_{1\cdot}p_{\cdot 1}\,. \end{aligned}$$

Thus we have shown that (2) implies (1). Recall that statistical independence implies that $\mathrm{E}(xy) = \mathrm{E}(x)\,\mathrm{E}(y) = \mu_x\mu_y$ and hence $\operatorname{cov}(x, y) = \mathrm{E}(xy) - \mu_x\mu_y = 0$. □

0.7 (Continued, in a way ...). Consider dichotomous variables x and y whose values are A_1, A_2 and B_1, B_2, respectively, and suppose we have n observations from these variables. Let us define new variables in the following way:

$$\begin{aligned} x_1 &= 1 \text{ if } x \text{ has value } A_1, \text{ and } x_1 = 0 \text{ otherwise,} \\ x_2 &= 1 \text{ if } x \text{ has value } A_2, \text{ and } x_2 = 0 \text{ otherwise,} \end{aligned}$$

and let y_1 and y_2 be defined in the corresponding way with respect to the values B_1 and B_2. Denote the observed $n \times 4$ data matrix as $\mathbf{U} = (\mathbf{x}_1 : \mathbf{x}_2 : \mathbf{y}_1 : \mathbf{y}_2) = (\mathbf{X} : \mathbf{Y})$. We are interested in the statistical dependence of the variables x and y and hence we prepare the following frequency table (contingency table):

		y		
		B_1	B_2	total
x	A_1	f_{11}	f_{12}	r_1
	A_2	f_{21}	f_{22}	r_2
	total	c_1	c_2	n

Let e_{ij} denote the expected frequency (for the usual χ^2-statistic for testing the independence) of the cell (i,j) and

$$e_{ij} = \frac{r_i c_j}{n}\,, \quad \mathbf{E} = (\mathbf{e}_1 : \mathbf{e}_2)\,, \quad \mathbf{F} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} = (\mathbf{f}_1 : \mathbf{f}_2)\,,$$

and $\mathbf{c} = \binom{c_1}{c_2}$, $\mathbf{r} = \binom{r_1}{r_2}$. We may assume that all elements of $\mathbf{c}$ and $\mathbf{r}$ are nonzero. Confirm the following:

(a) $\operatorname{cor_d}(\mathbf{x}_1, \mathbf{x}_2) = -1$, $\operatorname{rank}(\mathbf{X} : \mathbf{Y}) \leq 3$, $\operatorname{rank}[\operatorname{cor_d}(\mathbf{X} : \mathbf{Y})] \leq 2$,

(b) $\mathbf{X}'\mathbf{1}_n = \mathbf{r}$, $\mathbf{Y}'\mathbf{1}_n = \mathbf{c}$, $\mathbf{X}'\mathbf{X} = \operatorname{diag}(\mathbf{r}) = \mathbf{D}_\mathbf{r}$, $\mathbf{Y}'\mathbf{Y} = \operatorname{diag}(\mathbf{c}) = \mathbf{D}_\mathbf{c}$,

(c) $\mathbf{X}'\mathbf{Y} = \mathbf{F}$, $\mathbf{E} = \mathbf{r}\mathbf{c}'/n = \mathbf{X}'\mathbf{1}_n\mathbf{1}_n'\mathbf{Y}/n = \mathbf{X}'\mathbf{J}\mathbf{Y}$.

(d) The columns of $\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}$ represent the conditional relative frequencies (distributions) of x.

(e) $\mathbf{F} - \mathbf{E} = \mathbf{X}'\mathbf{C}\mathbf{Y}$, where $\mathbf{C}$ is the centering matrix, and hence $\frac{1}{n-1}(\mathbf{F} - \mathbf{E})$ is the sample (cross)covariance matrix between the x- and y-variables.

• Solution to Ex. 0.7:

(a) Suppose that the observations are arranged so that

$$\mathbf{X} = (\mathbf{x}_1 : \mathbf{x}_2) = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{r_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{r_2} \end{pmatrix} \in \mathbb{R}^{n \times 2}.$$

Then obviously $\operatorname{cor_d}(\mathbf{x}_1, \mathbf{x}_2) = -1$ and similarly $\operatorname{cor_d}(\mathbf{y}_1, \mathbf{y}_2) = -1$. Moreover,

$$\operatorname{rk}(\mathbf{X} : \mathbf{Y}) = \operatorname{rk}(\mathbf{X}) + \operatorname{rk}(\mathbf{Y}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y}) \leq 2 + 2 - 1 = 3\,,$$

because $\mathbf{X}\mathbf{1}_2 = \mathbf{1}_n = \mathbf{Y}\mathbf{1}_2$ and hence $\dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y}) \geq 1$.
Denote $\operatorname{cor_d}(\mathbf{x}_1, \mathbf{y}_1) = a$. Then, in view of $y_2 = -y_1 + 1$, we have $\operatorname{cor_d}(\mathbf{x}_1, \mathbf{y}_2) = -a$, and in view of $x_2 = -x_1 + 1$, we have $\operatorname{cor_d}(\mathbf{x}_2, \mathbf{y}_1) = -a$, and similarly $\operatorname{cor_d}(\mathbf{x}_2, \mathbf{y}_2) = a$, and so we can conclude that

$$\operatorname{cor_d}(\mathbf{X}:\mathbf{Y})] = \begin{pmatrix} 1 & -1 & a & -a \\ -1 & 1 & -a & a \\ a & -a & 1 & -1 \\ -a & a & -1 & 1 \end{pmatrix} := \mathbf{R}\,,$$

$$\operatorname{rank}(\mathbf{R}) = \operatorname{rank}\begin{pmatrix} 1 & a \\ -1 & -a \\ a & 1 \\ -a & -1 \end{pmatrix} \leq 2\,.$$

(b) $\mathbf{X}'\mathbf{1}_n = \begin{pmatrix} \mathbf{1}'_{r_1} & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}'_{r_2} \end{pmatrix}\mathbf{1}_n = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \mathbf{r}\,, \quad \mathbf{Y}'\mathbf{1}_n = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \mathbf{c}\,,$

$\mathbf{X}'\mathbf{X} = \begin{pmatrix} r_1 & 0 \\ 0 & r_2 \end{pmatrix} = \operatorname{diag}(\mathbf{r}) = \mathbf{D}_{\mathbf{r}}, \quad \mathbf{Y}'\mathbf{Y} = \operatorname{diag}(\mathbf{c}) = \mathbf{D}_{\mathbf{c}}\,,$

(c) frequency table: $\mathbf{X}'\mathbf{Y} = \mathbf{F}$,
theoretical frequencies: $\mathbf{E} = \mathbf{r}\mathbf{c}'/n = \mathbf{X}'\mathbf{1}_n\mathbf{1}_n'\mathbf{Y}/n = \mathbf{X}'\mathbf{J}\mathbf{Y}$.

(d) The columns of

$$\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1} = \mathbf{F}\mathbf{D}_{\mathbf{r}}^{-1} = \begin{pmatrix} f_{11}/c_1 & f_{12}/c_2 \\ f_{21}/c_1 & f_{22}/c_2 \end{pmatrix}, \text{ where } c_i = \#(y = B_i),$$

represent the conditional relative frequencies (distributions) of x.

(e) $\mathbf{F} - \mathbf{E} = \mathbf{X}'\mathbf{C}\mathbf{Y}$, where $\mathbf{C}$ is the centering matrix, and hence $\frac{1}{n-1}(\mathbf{F} - \mathbf{E})$ is the sample (cross)covariance matrix between the x- and y-variables.

□

0.8 (Continued ...).

(a) Show that $\mathbf{X}'\mathbf{C}\mathbf{Y}$, $\mathbf{X}'\mathbf{C}\mathbf{X}$, and $\mathbf{Y}'\mathbf{C}\mathbf{Y}$ are double-centered; $\mathbf{A}_{n\times p}$ is said to be double-centered if $\mathbf{A}\mathbf{1}_p = \mathbf{0}_n$ and $\mathbf{A}'\mathbf{1}_n = \mathbf{0}_p$.

(b) Prove that $\mathbf{1}_n \in \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y})$ and that it is possible that $\dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y}) > 1$.

(c) Show, using the rule $\operatorname{rk}(\mathbf{C}\mathbf{Y}) = \operatorname{rk}(\mathbf{Y}) - \dim \mathscr{C}(\mathbf{Y}) \cap \mathscr{C}(\mathbf{C})^{\perp}$, see Theorem 5 (p. 145), that

$$\operatorname{rk}(\mathbf{Y}'\mathbf{C}\mathbf{Y}) = \operatorname{rk}(\mathbf{C}\mathbf{Y}) = c - 1 \quad \text{and} \quad \operatorname{rk}(\mathbf{X}'\mathbf{C}\mathbf{X}) = \operatorname{rk}(\mathbf{C}\mathbf{X}) = r - 1\,,$$

where c and r refer to the number of categories of y and x, respectively; see Exercise 19.12 (p. 435) In this situation of course $c = r = 2$.

(d) Confirm that $(\mathbf{Y}'\mathbf{Y})^{-1}$ is a generalized inverse of $\mathbf{Y}'\mathbf{C}\mathbf{Y}$, i.e.,

$$\mathbf{Y}'\mathbf{C}\mathbf{Y} \cdot (\mathbf{Y}'\mathbf{Y})^{-1} \cdot \mathbf{Y}'\mathbf{C}\mathbf{Y} = \mathbf{Y}'\mathbf{C}\mathbf{Y}\,,$$
$$(\mathbf{D}_{\mathbf{c}} - \tfrac{1}{n}\mathbf{c}\mathbf{c}') \cdot \mathbf{D}_{\mathbf{c}}^{-1} \cdot (\mathbf{D}_{\mathbf{c}} - \tfrac{1}{n}\mathbf{c}\mathbf{c}') = \mathbf{D}_{\mathbf{c}} - \tfrac{1}{n}\mathbf{c}\mathbf{c}'.$$

See also part (b) of Exercise 0.11 (p. 56), Exercise 4.9 (p. 145), and Exercise 19.12 (p. 435).

• Solution to Ex. 0.8:

(a) $\mathbf{1}_2'\mathbf{X}'\mathbf{C}\mathbf{Y}\mathbf{1}_2 = \mathbf{1}_n'\mathbf{C}\mathbf{1}_n = 0\,.$

(b) $\mathbf{X}\mathbf{1}_2 = \mathbf{1}_n = \mathbf{Y}\mathbf{1}_2 \implies \mathbf{1}_n \in \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y})$ and hence $\dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y}) \geq 1$. For example, if $\mathbf{X} = \mathbf{Y}$, then $\dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y}) = 2 > 1$.

(c) $\operatorname{rk}(\mathbf{Y}'\mathbf{C}\mathbf{Y}) = \operatorname{rk}[(\mathbf{C}\mathbf{Y})'\mathbf{C}\mathbf{Y}] = \operatorname{rk}(\mathbf{Y}'\mathbf{C}) = \operatorname{rk}(\mathbf{Y}) - \dim \mathscr{C}(\mathbf{Y}) \cap \mathscr{C}(\mathbf{1}_n)$
$= \operatorname{rk}(\mathbf{Y}) - 1 = c - 1.$

(d) Because

$$(\mathbf{D_c} - \tfrac{1}{n}\mathbf{c}\mathbf{c}') \cdot \mathbf{D_c}^{-1} \cdot (\mathbf{D_c} - \tfrac{1}{n}\mathbf{c}\mathbf{c}') = (\mathbf{I}_n - \tfrac{1}{n}\mathbf{c}\mathbf{c}'\mathbf{D_c}^{-1})(\mathbf{D_c} - \tfrac{1}{n}\mathbf{c}\mathbf{c}')$$
$$= \mathbf{D_c} - \tfrac{1}{n}\mathbf{c}\mathbf{c}' - \tfrac{1}{n}\mathbf{c}\mathbf{c}' + \tfrac{1}{n^2}\mathbf{c}\,\underbrace{\mathbf{c}'\mathbf{D_c}^{-1}\mathbf{c}}_{=n}\,\mathbf{c}'\,,$$

and

$$\mathbf{c}'\mathbf{D_c}^{-1}\mathbf{c} = (c_1, \dots, c_c)\operatorname{diag}(1/c_1, \dots, 1/c_c)(c_1, \dots, c_c)' = c_1 + \dots + c_c = n,$$

the claim follows. □

0.9 (Continued . . .).

(a) What is the interpretation of the matrix

$$\mathbf{G} = \sqrt{n}\,(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{C}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2} = \sqrt{n}\,\mathbf{D_r}^{-1/2}(\mathbf{F} - \mathbf{E})\mathbf{D_c}^{-1/2}\,?$$

• Solution to (a):

$$\mathbf{G} = \sqrt{n}\,(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{C}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2} = \sqrt{n}\,\mathbf{D_r}^{-1/2}\underbrace{(\mathbf{F} - \mathbf{E})}_{:=\mathbf{Z}}\mathbf{D_c}^{-1/2}$$
$$= \sqrt{n}\,\operatorname{diag}(\tfrac{1}{\sqrt{r_1}}, \dots, \tfrac{1}{\sqrt{r_r}})\mathbf{Z}\operatorname{diag}(\tfrac{1}{\sqrt{c_1}}, \dots, \tfrac{1}{\sqrt{c_c}})$$
$$= \left\{\sqrt{n}\,\frac{z_{ij}}{\sqrt{r_i c_j}}\right\} = \left\{\frac{f_{ij} - e_{ij}}{\sqrt{r_i c_j/n}}\right\} = \left\{\frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}}\right\}.$$

(b) Convince yourself that the matrix

$$\mathbf{G}_* = \mathbf{D_r}^{-1/2}(\mathbf{F} - \mathbf{E})\mathbf{D_c}^{-1/2}$$

remains invariant if instead of frequencies we consider proportions so that the matrix $\mathbf{F}$ is replaced with $\frac{1}{n}\mathbf{F}$ and the matrices $\mathbf{E}$, $\mathbf{D_r}$ and $\mathbf{D_c}$ are calculated accordingly.

(c) Show that the χ^2-statistic for testing the independence of x and y can be written as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(f_{ij}-e_{ij})^2}{e_{ij}} = \|\mathbf{G}\|_F^2 = \operatorname{tr}(\mathbf{G}'\mathbf{G}) = n \operatorname{tr}(\mathbf{P_X C P_Y C})\,.$$

See also Exercise 19.13.

• **Solution to (c):**

$$\begin{aligned}
\|\mathbf{G}\|_F^2 &= \operatorname{tr}(\mathbf{G}\mathbf{G}') \\
&= n \operatorname{tr}[(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{C}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2} \cdot (\mathbf{Y}'\mathbf{Y})^{-1/2}\mathbf{Y}'\mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}] \\
&= n \operatorname{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{C}\mathbf{X}] \\
&= n \operatorname{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{C}] = n \operatorname{tr}(\mathbf{P_X C P_Y C})\,. \qquad \square
\end{aligned}$$

(d) Show that the contribution of the ith column of $\mathbf{F}$ on the χ^2, $\chi^2(\mathbf{f}_i)$, say, can be expressed as (a kind of squared Mahalanobis distance)

$$\chi^2(\mathbf{f}_i) = (\mathbf{f}_i - \mathbf{e}_i)'\mathbf{D}^{-1}(\mathbf{f}_i - \mathbf{e}_i)\,,$$

where

$$\mathbf{D} = \operatorname{diag}(\mathbf{e}_i) = \begin{pmatrix} e_{i1} & 0 \\ 0 & e_{i2} \end{pmatrix} = \begin{pmatrix} r_1 c_i/n & 0 \\ 0 & r_2 c_i/n \end{pmatrix} = c_i \begin{pmatrix} r_1/n & 0 \\ 0 & r_2/n \end{pmatrix}.$$

0.10 (Multinomial distribution). Consider the random vectors (for simplicity only three-dimensional)

$$\mathbf{z}_1 = \begin{pmatrix} z_{11} \\ z_{21} \\ z_{31} \end{pmatrix}, \dots, \mathbf{z}_m = \begin{pmatrix} z_{1m} \\ z_{2m} \\ z_{3m} \end{pmatrix}, \quad \mathbf{x} = \mathbf{z}_1 + \cdots + \mathbf{z}_m\,,$$

where $\mathbf{z}_i$ are identically and independently distributed random vectors so that each $\mathbf{z}_i$ is defined so that only one element gets value 1 the rest being 0. Let $\mathrm{P}(z_{i1} = 1) = p_1$, $\mathrm{P}(z_{i2} = 1) = p_2$, and $\mathrm{P}(z_{i3} = 1) = p_3$ for $i = 1, \dots, m$; $p_1 + p_2 + p_3 = 1$, each $p_i > 0$, and denote $\mathbf{p} = (p_1, p_2, p_3)'$. Show that

$$\mathrm{E}(\mathbf{z}_i) = (p_1, p_2, p_3)' = \mathbf{p}\,, \quad \mathrm{E}(\mathbf{x}) = m\mathbf{p}\,,$$

and

$$\begin{aligned}
\operatorname{cov}(\mathbf{z}_i) &= \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{pmatrix} = \begin{pmatrix} p_1 & 0 & 0 \\ 0 & p_2 & 0 \\ 0 & 0 & p_3 \end{pmatrix} - \mathbf{p}\mathbf{p}' \\
&:= \mathbf{D_p} - \mathbf{p}\mathbf{p}' := \mathbf{\Sigma}\,, \quad \operatorname{cov}(\mathbf{x}) = m\mathbf{\Sigma}\,.
\end{aligned}$$

Then $\mathbf{x}$ follows a multinomial distribution with parameters m and $\mathbf{p}$: $\mathbf{x} \sim \mathrm{Mult}(m, \mathbf{p})$.

0.11 (Continued ...). Confirm:

(a) $\boldsymbol{\Sigma}$ is double-centered (row and column sums are zero), singular and has rank 2.

(b) $\boldsymbol{\Sigma}\mathbf{D}_{\mathbf{p}}^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$, i.e., $\mathbf{D}_{\mathbf{p}}^{-1}$ is a generalized inverse of $\boldsymbol{\Sigma}$. Confirm that $\mathbf{D}_{\mathbf{p}}^{-1}$ does not necessarily satisfy any other Moore–Penrose conditions. See also Exercises 0.8 (p. 53) and 4.9 (p. 145).

(c) We can think (confirm ...) that the columns (or rows if we wish) of a contingency table are realizations of a multinomial random variable. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ represent two columns (two observations) from such a random variable and assume that instead of the frequencies we consider proportions $\mathbf{y}_1 = \mathbf{x}_1/c_1$ and $\mathbf{y}_2 = \mathbf{x}_2/c_2$. Then $\mathbf{x}_i \sim \mathrm{Mult}(c_i, \mathbf{p})$ and $\mathrm{cov}(\mathbf{y}_i) = \frac{1}{c_i}\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = \mathbf{D}_{\mathbf{p}} - \mathbf{p}\mathbf{p}'$, and the squared Mahalanobis distance, say M, between the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ can be defined as follows:

$$\begin{aligned} M &= c_1c_2(c_1+c_2)^{-1}(\mathbf{x}_1-\mathbf{x}_2)'\boldsymbol{\Sigma}^{-}(\mathbf{x}_1-\mathbf{x}_2) \\ &= c_1c_2(c_1+c_2)^{-1}(\mathbf{x}_1-\mathbf{x}_2)'\mathbf{D}_{\mathbf{p}}^{-1}(\mathbf{x}_1-\mathbf{x}_2)\,. \end{aligned}$$

Neudecker (1997), Puntanen, Styan & Subak-Sharpe (1998), Greenacre (2007, p. 270).

0.12. Let $\mathbf{P}_{n\times n}$ be an idempotent matrix. Show that

$$\mathscr{C}(\mathbf{P}) \cap \mathscr{C}(\mathbf{I}_n - \mathbf{P}) = \{\mathbf{0}\} \quad \text{and} \quad \mathscr{C}(\mathbf{I}_n - \mathbf{P}) = \mathscr{N}(\mathbf{P})\,.$$

• SOLUTION TO EX. 0.12:

(a) $\mathbf{u} \in \mathscr{C}(\mathbf{P}) \cap \mathscr{C}(\mathbf{I}-\mathbf{P}) \implies \exists\, \boldsymbol{\alpha}, \boldsymbol{\beta}:$

$$\mathbf{u} = \mathbf{P}\boldsymbol{\alpha} = (\mathbf{I}-\mathbf{P})\boldsymbol{\beta}.$$

Premultiplying the above equation by $\mathbf{P}$ yields

$$\mathbf{P}\mathbf{u} = \mathbf{P}^2\boldsymbol{\alpha} = \mathbf{P}(\mathbf{I}-\mathbf{P})\boldsymbol{\beta} = (\mathbf{P}-\mathbf{P}^2)\boldsymbol{\beta} = \mathbf{0}, \quad \text{because } \mathbf{P}^2 = \mathbf{P}\,.$$

Hence also $\mathbf{P}^2\boldsymbol{\alpha} = \mathbf{0}$, i.e., $\mathbf{P}^2\boldsymbol{\alpha} = \mathbf{P}\boldsymbol{\alpha} = \mathbf{u} = \mathbf{0}$.

(b) Let's first show that $\mathscr{C}(\mathbf{I}-\mathbf{P}) \subset \mathscr{N}(\mathbf{P})$. Now $\mathbf{u} \in \mathscr{C}(\mathbf{I}-\mathbf{P}) \implies \exists\, \boldsymbol{\alpha}$:

$$\mathbf{u} = (\mathbf{I}-\mathbf{P})\boldsymbol{\alpha}. \tag{$*$}$$

Premultiplying $(*)$ by $\mathbf{P}$:

$$\mathbf{P}\mathbf{u} = \mathbf{P}(\mathbf{I}-\mathbf{P})\boldsymbol{\alpha} = (\mathbf{P}-\mathbf{P}^2)\boldsymbol{\alpha} = \mathbf{0} \implies \mathbf{u} \in \mathscr{N}(\mathbf{P}),$$

and thereby $\mathscr{C}(\mathbf{I}-\mathbf{P}) \subset \mathscr{N}(\mathbf{P})$. It remains to show that $\mathscr{N}(\mathbf{P}) \subset \mathscr{C}(\mathbf{I}-\mathbf{P})$:

$$\begin{aligned}\mathbf{u} \in \mathscr{N}(\mathbf{P}) &\implies \mathbf{Pu} = \mathbf{0} \implies \mathbf{u} - \mathbf{Pu} = \mathbf{u} \\ &\implies (\mathbf{I}-\mathbf{P})\mathbf{u} = \mathbf{u} \implies \mathbf{u} \in \mathscr{C}(\mathbf{I}-\mathbf{P}).\end{aligned}$$

□

0.13. Confirm: $\mathbf{A} \geq_{\mathsf{L}} \mathbf{B}$ and $\mathbf{B} \geq_{\mathsf{L}} \mathbf{C} \implies \mathbf{A} \geq_{\mathsf{L}} \mathbf{C}$.

• Solution to Ex. 0.13:

$$\begin{aligned}\mathbf{A} \geq_{\mathsf{L}} \mathbf{B} \text{ and } \mathbf{B} \geq_{\mathsf{L}} \mathbf{C} &\implies \mathbf{A}-\mathbf{B} = \mathbf{KK}',\ \mathbf{B}-\mathbf{C} = \mathbf{LL}' \\ &\implies \mathbf{A} = \mathbf{KK}' + \mathbf{B},\ \mathbf{C} = -\mathbf{LL}' + \mathbf{B} \\ &\implies \mathbf{A}-\mathbf{C} = \mathbf{KK}' + \mathbf{LL}' \geq_{\mathsf{L}} \mathbf{0}.\end{aligned}$$

□

0.14. Suppose that $\mathbf{x}$ and $\mathbf{y}$ are p-dimensional random vectors. Confirm:

(a) $\operatorname{cov}(\mathbf{x}+\mathbf{y}) = \operatorname{cov}(\mathbf{x}) + \operatorname{cov}(\mathbf{y}) \iff \operatorname{cov}(\mathbf{x},\mathbf{y}) = -\operatorname{cov}(\mathbf{y},\mathbf{x})$;
if $\mathbf{A} = -\mathbf{A}'$, $\mathbf{A}$ is said to be skew-symmetric.

(b) $\operatorname{cov}(\mathbf{x}-\mathbf{y}) = \operatorname{cov}(\mathbf{x}) - \operatorname{cov}(\mathbf{y}) \iff \operatorname{cov}(\mathbf{x},\mathbf{y}) + \operatorname{cov}(\mathbf{y},\mathbf{x}) = 2\operatorname{cov}(\mathbf{y})$.

• Solution to Ex. 0.14:

(a) $\operatorname{cov}(\mathbf{x}+\mathbf{y}) = \operatorname{cov}(\mathbf{x}) + \operatorname{cov}(\mathbf{y}) + \operatorname{cov}(\mathbf{x},\mathbf{y}) + \operatorname{cov}(\mathbf{y},\mathbf{x}) = \operatorname{cov}(\mathbf{x}) + \operatorname{cov}(\mathbf{y})$
$\iff$
$\operatorname{cov}(\mathbf{x},\mathbf{y}) = -\operatorname{cov}(\mathbf{y},\mathbf{x}) \iff \mathbf{\Sigma}_{\mathbf{xy}} = -\mathbf{\Sigma}'_{\mathbf{xy}}$.

(b) $\operatorname{cov}(\mathbf{x}-\mathbf{y}) = \operatorname{cov}(\mathbf{x}) + \operatorname{cov}(\mathbf{y}) - \operatorname{cov}(\mathbf{x},\mathbf{y}) - \operatorname{cov}(\mathbf{y},\mathbf{x}) = \operatorname{cov}(\mathbf{x}) - \operatorname{cov}(\mathbf{y})$
$\iff$
$-\operatorname{cov}(\mathbf{y}) = \operatorname{cov}(\mathbf{y}) - \operatorname{cov}(\mathbf{x},\mathbf{y}) - \operatorname{cov}(\mathbf{y},\mathbf{x})$
$\iff$
$\operatorname{cov}(\mathbf{x},\mathbf{y}) + \operatorname{cov}(\mathbf{y},\mathbf{x}) = 2\operatorname{cov}(\mathbf{y})$.

□

0.24. Consider the set of numbers $\mathcal{A} = \{1, 2, \dots, N\}$ and let $x_1, x_2, \dots, x_p$ denote a random sample selected without a replacement from $\mathcal{A}$. Denote $y = x_1 + x_2 + \dots + x_p = \mathbf{1}_p' \mathbf{x}$. Confirm the following:

(a) $\operatorname{var}(x_i) = \frac{N^2-1}{12}$, $\quad \operatorname{cor}(x_i, x_j) = -\frac{1}{N-1} = \varrho$, $\quad i, j = 1, \dots, p$,

(b) $\operatorname{cor}^2(x_1, y) = \operatorname{cor}^2(x_1, x_1 + \dots + x_p) = \frac{1}{p} + \big(1 - \frac{1}{p}\big)\varrho$.

See also Section 10.6 (p. 234).

• Solution to Ex. 0.24:

Clearly each x_i follows a discrete uniform distribution, $\operatorname{Unif}(1, \dots, N)$, so that

$$\mathrm{E}(x_i) = \frac{1}{N}(1 + 2 + \dots + N) = \frac{N+1}{2} := \mu\,,$$

and the variance is

$$\operatorname{var}(x_i) = \frac{1}{N}\sum_{i=1}^{N}(i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N} i^2 - \mu^2 = \frac{N^2-1}{12} := \sigma^2,$$

where we have used the fact

$$\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}\,.$$

We will next show that

$$\operatorname{cov}(x_i, x_j) = -\frac{1}{N-1}\frac{N^2-1}{12} = -\frac{1}{N-1}\sigma^2 = -\frac{N+1}{12}\,,$$

$$\operatorname{cor}(x_i, x_j) = \frac{-\frac{1}{N-1}\sigma^2}{\sigma \cdot \sigma} = -\frac{1}{N-1} := \varrho\,, \quad i \neq j\,.$$

For convenience, let $z_1, z_2, \dots, z_p$ denote a random sample selected *with* a replacement from $\{1, 2, \dots, N\}$. Because z_i and z_j $(i \neq j)$ are uncorrelated we trivially have

$$\operatorname{cov}(z_i, z_j) = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{\ell=1}^{N}(k-\mu_y)(\ell-\mu_y) := \frac{1}{N^2}\mathrm{SP}_{z_i z_j} = 0\,.$$

In view of

$$\operatorname{cov}(x_i, x_j) = \frac{1}{N(N-1)}\sum_{k=1}^{N}\sum_{\substack{\ell=1\\k\neq\ell}}^{N}(k-\mu_y)(\ell-\mu_y) := \frac{1}{N(N-1)}\mathrm{SP}_{x_i x_j}\,,$$

we get

$$
\begin{aligned}
\mathrm{SP}_{x_ix_j} = \mathrm{SP}_{z_iz_j} - \sum_{k=1}^{N}(k-\mu_y)^2 = 0 - N\sigma^2 = -N\,\frac{N^2-1}{12}\,,\\
\operatorname{cov}(x_i,x_j) = -\frac{1}{N(N-1)}\,N\,\frac{N^2-1}{12} = -\frac{N+1}{12}\,,\\
\operatorname{cor}(x_i,x_j) = -\frac{(N+1)/12}{(N^2-1)/12} = -\frac{1}{N-1} = \varrho\,.
\end{aligned}
$$

If $y = x_1 + \cdots + x_p = \mathbf{1}'\mathbf{x}$, then

$$
\operatorname{cor}^2(x_1,y) = \operatorname{cor}^2(x_1, x_1 + \cdots + x_p) = \frac{\operatorname{cov}^2(x_1,y)}{\operatorname{var}(x_1)\operatorname{var}(y)}\,.
$$

Now

$$
\begin{aligned}
\operatorname{var}(x_i) &= \frac{N^2-1}{12} = \sigma^2, \quad i = 1,\dots,p\,,\\
\operatorname{var}(y) &= \operatorname{var}(\mathbf{1}'\mathbf{x}) = \mathbf{1}'\operatorname{cov}(\mathbf{x})\mathbf{1} = \mathbf{1}'\mathbf{\Sigma}\mathbf{1}\,,
\end{aligned}
$$

where

$$
\begin{aligned}
\operatorname{cov}(\mathbf{x}) = \mathbf{\Sigma} = \sigma^2\begin{pmatrix} 1 & \varrho & \dots & \varrho \\ \varrho & 1 & \dots & \varrho \\ \vdots & \vdots & \ddots & \vdots \\ \varrho & \varrho & \dots & 1 \end{pmatrix} \in \mathbb{R}^{p\times p},\\
\operatorname{cor}(x_i,x_j) = \varrho = \frac{\sigma_{ij}}{\sigma_i\sigma_j} = \frac{\sigma_{ij}}{\sigma^2} \implies \sigma_{ij} = \sigma^2\varrho\,.
\end{aligned}
$$

Hence

$$
\operatorname{var}(y) = \mathbf{1}'\mathbf{\Sigma}\mathbf{1} = p\sigma^2[1+(p-1)\varrho]\,.
$$

Moreover,

$$
\begin{aligned}
\operatorname{cov}(x_1,y) &= \operatorname{cov}(x_1, x_1 + \cdots + x_p)\\
&= \operatorname{cov}(x_1,x_1) + \operatorname{cov}(x_1,x_2) + \cdots + \operatorname{cov}(x_1,x_p)\\
&= \sigma^2 + \sigma_{12} + \cdots + \sigma_{1p}\\
&= \sigma^2[1+(p-1)\varrho]\,,
\end{aligned}
$$

and thereby

$$
\begin{aligned}
\operatorname{cor}^2(x_1,y) &= \frac{\sigma^4[1+(p-1)\varrho)]^2}{\sigma^2\cdot p\sigma^2[1+(p-1)\varrho]}\\
&= \frac{1}{p} + \frac{p-1}{p}\,\varrho\\
&= \frac{1}{p} + \big(1-\frac{1}{p}\big)\varrho\,.
\end{aligned}
$$

□

0.25 (Hotelling's T^2). Let $\mathbf{U}_1'$ and $\mathbf{U}_2'$ be independent random samples from $\mathrm{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathrm{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively. Denote $\mathbf{T}_i = \mathbf{U}_i'(\mathbf{I}_{n_i} - \mathbf{J}_{n_i})\mathbf{U}_i$, and $\mathbf{S}_* = \frac{1}{f}(\mathbf{T}_1 + \mathbf{T}_2)$, where $f = n_1 + n_2 - 2$. Confirm that

$$T^2 = \tfrac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)'\mathbf{S}_*^{-1}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2) \sim \mathrm{T}^2(p, n_1 + n_2 - 2)\,, \tag{a}$$

where $\mathrm{T}^2(a, b)$ refers to the Hotelling's T^2 distribution; see (0.128) (p. 26). It can be shown that if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, then

$$\tfrac{n_1+n_2-p-1}{(n_1+n_2-2)p}\, T^2 \sim \mathrm{F}(p, n_1 + n_2 - p - 1)\,. \tag{b}$$

Notice: In (a) we should actually assume that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

• **Solution to Ex. 0.25:**

Recall the Wishart distribution and Hotelling's T^2 distribution:

- Let $\mathbf{U}' = (\mathbf{u}_{(1)} : \ldots : \mathbf{u}_{(n)})$ be a random sample from $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, i.e., $\mathbf{u}_{(i)}$'s are independent and each $\mathbf{u}_{(i)} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Then $\mathbf{W} = \mathbf{U}'\mathbf{U} = \sum_{i=1}^n \mathbf{u}_{(i)}\mathbf{u}_{(i)}'$ is said to a have a Wishart distribution with n degrees of freedom and scale matrix $\boldsymbol{\Sigma}$, and we write $\mathbf{W} \sim \mathrm{W}_p(n, \boldsymbol{\Sigma})$.
- Let $\mathbf{U}'$ be a random sample from $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\bar{\mathbf{u}} = \frac{1}{n}\mathbf{U}'\mathbf{1}_n$ and $\mathbf{T} = \mathbf{U}'(\mathbf{I} - \mathbf{J})\mathbf{U}$ are independent and $\mathbf{T} \sim \mathrm{W}_p(n - 1, \boldsymbol{\Sigma})$.
- Suppose $\mathbf{v} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{W} \sim \mathrm{W}_p(m, \boldsymbol{\Sigma})$, $\mathbf{v}$ and $\mathbf{W}$ are independent, and that $\boldsymbol{\Sigma}$[6] is positive definite. Hotelling's T^2 distribution is the distribution of
$$T^2 = m \cdot \mathbf{v}'\mathbf{W}^{-1}\mathbf{v} = \mathbf{v}'\big(\tfrac{1}{m}\mathbf{W}\big)^{-1}\mathbf{v}\,, \tag{c}$$
and is denoted as $T^2 \sim \mathrm{T}^2(p, m)$.

In the situation of Ex. 0.25, $\mathbf{T}_1 \sim \mathrm{W}_p(n_1 - 1, \boldsymbol{\Sigma})$ and $\mathbf{T}_2 \sim \mathrm{W}_p(n_2 - 1, \boldsymbol{\Sigma})$ and $\mathbf{T}_1$ and $\mathbf{T}_2$ are independent. Hence it is easy to conclude (at least easy to believe ...) that their sum has property

$$\mathbf{T}_1 + \mathbf{T}_2 \sim \mathrm{W}_p(n_1 + n_2 - 2, \boldsymbol{\Sigma})\,.$$

Suppose that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Then the difference $\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2$ obviously has the distribution

$$\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2 \sim \mathrm{N}_p(\mathbf{0}, \tfrac{1}{n_1}\boldsymbol{\Sigma} + \tfrac{1}{n_2}\boldsymbol{\Sigma}) = \mathrm{N}_p(\mathbf{0}, \tfrac{n_1+n_2}{n_1 n_2}\boldsymbol{\Sigma})\,,$$

and thereby

$$\sqrt{\tfrac{n_1 n_2}{n_1+n_2}}\,(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2) \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})\,.$$

---

[6] In the *Tricks Book* here is erroneously $\mathbf{W}$.

Substituting

$$\mathbf{v} = \sqrt{\tfrac{n_1 n_2}{n_1+n_2}}\,(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)\,,$$
$$\mathbf{W} = \mathbf{T}_1 + \mathbf{T}_2\,, \quad m = n_1 + n_2 - 2\,,$$

into (c) yields (a).

For further details concerning the testing of hypothesis $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, see pages 233–234. □

0.26 (Continued ...). Show that if $n_1 = 1$, then the Hotelling's T^2 becomes

$$T^2 = \tfrac{n_2}{n_2+1}(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2)'\mathbf{S}_2^{-1}(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2)\,.$$

• Solution to Ex. 0.26:

Now we have only one observation $\mathbf{u}_{(1)}$ from population 1 and n_2 observations from population 2 and

$$\mathbf{S}_2 = \tfrac{1}{n_2-1}\,\mathbf{T}_2\,.$$

Hotelling's T^2 becomes

$$\begin{aligned} T^2 &= \tfrac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)'\mathbf{S}_*^{-1}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2) \\ &= \tfrac{n_2}{n_2+1}(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2)'\mathbf{S}_2^{-1}(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2) \sim \mathrm{T}^2(p, n_2 - 1)\,. \end{aligned}$$

If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, then

$$\tfrac{n_1+n_2-p-1}{(n_1+n_2-2)p}\,T^2 \sim \mathrm{F}(p, n_1 + n_2 - p - 1)\,,$$

which in this case becomes

$$\tfrac{n_2-p}{(n_2-1)p}\,T^2 \sim \mathrm{F}(p, n_2 - p)\,,$$
$$\tfrac{n_2(n_2-p)}{(n_2^2-1)p}\,(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2)'\mathbf{S}_2^{-1}(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2) \sim \mathrm{F}(p, n_2 - p)\,. \tag{a}$$

Notice: We can denote

$$(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2)'\mathbf{S}_2^{-1}(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2) = \mathrm{MHLN}^2(\mathbf{u}_{(1)}, \bar{\mathbf{u}}_2, \mathbf{S}_2)\,. \tag{b}$$

Above $\bar{\mathbf{u}}_2$ and $\mathbf{S}_2$ are being calculated from the sample $\mathbf{U}_2$ while the single observation $\mathbf{u}_{(1)}$ does not belong to this sample. The resulting Mahalanobis distance in (b) differs from the "usual" Mahalanobis distance (squared)

$$(\mathbf{u}_{(i)} - \bar{\mathbf{u}})'\mathbf{S}^{-1}(\mathbf{u}_{(i)} - \bar{\mathbf{u}}) = \mathrm{MHLN}^2(\mathbf{u}_{(i)}, \bar{\mathbf{u}}, \mathbf{S})\,, \tag{c}$$

where $\mathbf{u}_{(i)}$ is one observation in the data matrix $\mathbf{U}$, $\mathbf{S} = \text{cov}_\text{d}(\mathbf{U})$, and $\bar{\mathbf{u}} = \mathbf{U}'\mathbf{1}_n/n$.

Problem: Try to compare (b) and (c).

• • If $n_1 = 1$ and also $p = 1$, then (a) becomes

$$\frac{n_2(n_2-1)}{n_2^2-1}(u_1-\bar{u}_2)s_2^{-2}(u_1-\bar{u}_2) = \frac{n_2}{n_2+1}\,\frac{(u_1-\bar{u}_2)^2}{s_2^2} \sim \text{F}(1, n_2-1)\,,$$

where $\bar{u}_2$ and s_2^2 are calculated from the "second" sample. A clearer notation can be obtained from Exercise 8.9 (p. 186) which expresses the square root of the above test statistics as

$$\begin{aligned} t &= \frac{y_n - \bar{y}_{(n)}}{s_{(n)}/\sqrt{1-\frac{1}{n}}} = \sqrt{\frac{n-1}{n}}\;\frac{y_n - \bar{y}_{(n)}}{s_{(n)}} \\ &= \frac{y_n - \bar{y}}{s_{(n)}\sqrt{1-\frac{1}{n}}} = \sqrt{\frac{n}{n-1}}\;\frac{y_n-\bar{y}}{s_{(n)}}\,, \end{aligned}$$

where $\bar{y}_{(n)}$ is the mean of $y_1, \dots, y_{n-1}$ and $s_{(n)}$ their standard deviation; $\bar{y}$ is the mean of all y_i's. This t-test statistic is the externally Studentized residual.

•• If $p = 1$ then, using the notation of Exercise 8.12 (p. 187), Hotelling's T^2 becomes

$$\begin{aligned} T^2 &= \tfrac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)'\mathbf{S}_*^{-1}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2) \\ &= \tfrac{n_1 n_2}{n_1+n_2} \cdot (\bar{y}_1 - \bar{y}_2) \cdot \left(\frac{\text{SS}_1 + \text{SS}_2}{n_1+n_2-2}\right)^{-1} \cdot (\bar{y}_1 - \bar{y}_2) \sim \text{T}^2(1, n_1+n_2-2)\,, \end{aligned}$$

and the Hotelling's T^2 is precisely the F-test statistics for the hypothesis $\mu_1 = \mu_2$:

$$\begin{aligned} T^2 &= F \\ &= \frac{n_1 n_2}{n_1+n_2} \cdot (\bar{y}_1 - \bar{y}_2) \cdot \left(\frac{\text{SS}_1 + \text{SS}_2}{n_1+n_2-2}\right)^{-1} \cdot (\bar{y}_1 - \bar{y}_2) \\ &= \frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{\text{SS}_1 + \text{SS}_2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2}{\frac{\text{SS}_1 + \text{SS}_2}{n_1+n_2-2}} \\ &= \frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{\text{SSE}}{n-2}\,\frac{n_1+n_2}{n_1 n_2}} \sim \text{F}(1, n_1+n_2-2) = \text{t}^2(n_1+n_2-2)\,. \end{aligned}$$

•• If $\mathbf{U}'$ is a random sample from $\text{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

• Hotelling's T^2: $T^2 = n(\bar{\mathbf{u}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{u}} - \boldsymbol{\mu}_0) = n \cdot \text{MHLN}^2(\bar{\mathbf{u}}, \boldsymbol{\mu}_0, \mathbf{S})$,

$$\tfrac{n-p}{(n-1)p}T^2 \sim \mathrm{F}(p, n-p, \theta), \quad \theta = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\,.$$

- Hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is rejected at risk level α, if

$$n(\bar{\mathbf{u}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{u}} - \boldsymbol{\mu}_0) > \tfrac{p(n-1)}{n-p}F_{\alpha;p,n-p}\,.$$

- A $100(1-\alpha)\%$ confidence region for the mean of the $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the ellipsoid determined by all $\boldsymbol{\mu}$ such that

$$n(\bar{\mathbf{u}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{u}} - \boldsymbol{\mu}) \leq \tfrac{p(n-1)}{n-p}F_{\alpha;p,n-p}\,.$$

□