Correlation Mixed exercise

1 a $S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 465 - \frac{67 \times 67}{10} = 16.1$

$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 429 - \frac{65 \times 65}{10} = 6.5$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 434 - \frac{67 \times 65}{10} = -1.5$

$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-1.5}{\sqrt{16.1 \times 6.5}} = \frac{-1.5}{10.2298\ldots} = -0.1466\ldots = -0.147$ (3 s.f.)

b The coding is linear, so the product moment correlation coefficient will be unaffected by the coding. So the product moment correlation coefficient between s and a is –0.147.

c This is a weak negative correlation that is close to 0. There is little evidence to suggest that students in the group who are good at science will also be good at art.

2 a $S_{jj} = \sum j^2 - \frac{\left(\sum j\right)^2}{n} = 52\,335 - \frac{979 \times 979}{20} = 4412.95$

$S_{pp} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = 32\,156 - \frac{735 \times 735}{20} = 5144.75$

$S_{jp} = \sum jp - \frac{\sum j \sum p}{n} = 39\,950 - \frac{979 \times 735}{20} = 3971.75$

b $r = \frac{S_{jp}}{\sqrt{S_{jj}S_{pp}}} = \frac{3971.75}{\sqrt{4412.95 \times 5144.75}} = \frac{3971.75}{4764.8215} = 0.8335\ldots = 0.834$ (3 s.f.)

c There is a strong positive correlation between the amount of juice and the cost, as the product moment correlation coefficient is close to 1. So Nimer is correct.

3 a $S_{pp} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = \sum (x-10)^2 - \frac{\left(\sum (x-10)\right)^2}{n}$

$= \sum x^2 - 20\sum x + 100n - \frac{\left(\left(\sum x\right) - 10n\right)^2}{n}$

$= \sum x^2 - 20\sum x + 100n - \frac{\left(\sum x\right)^2 - 20n\sum x + 100n^2}{n}$

$= \sum x^2 - 20\sum x + 100n - \frac{\left(\sum x\right)^2}{n} + 20\sum x - 100n$

$= \sum x^2 - \frac{\left(\sum x\right)^2}{n} = S_{xx}$

3 b $S_{qq} = \sum q^2 - \frac{\left(\sum q\right)^2}{n} = 77.0375 - \frac{\left(\sum \frac{1}{20} y\right)^2}{n} = 77.0375 - \frac{\left(\sum y\right)^2}{400n}$

$= 77.0375 - \frac{491^2}{400 \times 8} = 1.69968\ldots = 1.70$ (3 s.f.)

$r = \frac{S_{pq}}{\sqrt{S_{pp} S_{qq}}} = \frac{-11.625}{\sqrt{85.5 \times 1.69968\ldots}} = -0.964$ (3.s.f).

c The coding is linear, so the product moment correlation coefficient will be unaffected by the coding. So the product moment correlation coefficient between x and y is –0.964.

d The correlation coefficient suggests a strong negative linear correlation, but the scatter diagram shows a non-linear fit.

4 a Spearman's rank correlation coefficient is appropriate as the data is given in ranks rather than raw scores.

b The table shows d and d^2 for each pair of ranks:

Cat	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***
First judge	4	6	1	2	5	3	10	9	8	7
Second judge	2	9	3	1	7	4	6	8	5	10
d	2	–3	–2	1	–2	–1	4	1	3	–3
d²	4	9	4	1	4	1	16	1	9	9

$\sum d^2 = 58$

$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 58}{10(10^2-1)} = 0.648$ (3 s.f.)

c The null hypothesis is assumed to be correct; the alternative hypothesis is what can be concluded about the parameter if the assumption about the null hypothesis is shown to be wrong. The null hypothesis is only rejected if the probability of it being correct is less than or equal to the significance level of the test.

d Test for a positive association between the judges' rankings.

$H_0: \rho = 0,\ H_1: \rho > 0$

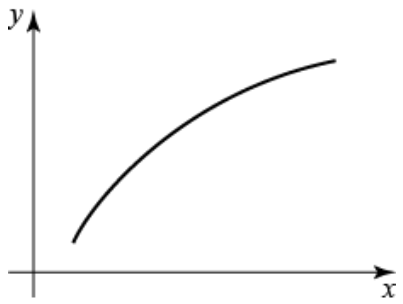
Sample size = 10

Significance level = 0.05

The critical value for r_s for a 0.05 significance level with a sample size of 10 is $r_s = 0.5636$.

As $0.648 > 0.5636$, r_s lies within the critical region, so reject H_0. There is sufficient evidence at the 5% significance level that there is a correlation between the rankings of the two judges.

5 a There are several reasons for using Spearman's rank correlation coefficient. (1) It can be used where the relationship between the two variables isn't linear. For example, such as this relationship:



(2) It can be used where the results are in the form of rankings already, such as where items have been put in order of preference of judgements, or where alphabetical grades have been awarded.

(3) It can be used where one or both sets of data are not from a normally distributed population.

b The table shows d and d^2 for each pair of ranks:

Applicant	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***
Tutor 1	1	2	3	4	5	6	7	8	9
Tutor 2	1	3	5	4	2	7	9	8	6
d	0	–1	–2	0	3	–1	–2	0	3
$\boldsymbol{d^2}$	0	1	4	0	9	1	4	0	9

$$\sum d^2 = 28$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 28}{9(9^2-1)} = 0.767 \text{ (3 s.f.)}$$

H_0: $\rho = 0$ There is no correlation between the ranks awarded by the two tutors.

H_1: $\rho > 0$ There is positive correlation between the ranks awarded by the two tutors.

Sample size = 9

The critical value for r_s for a 0.025 significance level with a sample size of 9 is $r_s = 0.7000$.

As $0.767 > 0.7000$, r_s lies within the critical region, so reject H_0. There is sufficient evidence at the 2.5% significance level that there is a correlation between the rankings of the two tutors.
The critical value for r_s for a 0.01 significance level with a sample size of 9 is $r_s = 0.7833$.

As $0.767 < 0.7833$, accept H_0. There is insufficient evidence of positive correlation at the 1% significance level of a correlation between the rankings of the two tutors.

6 a The table shows d and d^2 for each pair of ranks:

Ski Jumper	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***
First jump	2	9	7	4	10	8	6	5	1	3
Second jump	4	10	5	1	8	9	2	7	3	6
d	–2	–1	2	3	2	–1	4	–2	–2	–3
$\boldsymbol{d^2}$	4	1	4	9	4	1	16	4	4	9

$\sum d^2 = 56$

$r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)} = 1 - \dfrac{6\times 56}{10(10^2-1)} = 0.66$ (2 d.p.)

b H_0: $\rho = 0$ There is no correlation between the order of merit for the two jumps.

H_1: $\rho > 0$ There is a positive correlation between the order of merit for the two jumps.

Sample size = 10

Significance level = 0.05

The critical value for r_s for a 0.05 significance level with a sample size of 10 is $r_s = 0.5636$.

As $0.66 > 0.5636$, r_s lies within the critical region, so reject H_0. There is sufficient evidence at the 5% significance level to conclude there is a positive correlation between the order of merit for the two jumps – jumpers who did well in the first jumps are also likely to do well in the second jump.

7 Data from the expert is given in ranks rather than raw scores, therefore use the Spearman's rank correlation coefficient. The table shows the ranks and d and d^2 for each pair of ranks:

Bowl	***A***	***B***	***C***	***D***	***E***	***F***	***G***
Date of manufacture	1920	1857	1710	1896	1810	1690	1780
Rank, date	7	5	2	6	4	1	3
Rank, expert's order	7	3	4	6	2	1	5
d	0	2	–2	0	2	0	–2
***d*²**	0	4	4	0	4	0	4

$\sum d^2 = 16$

$r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)} = 1 - \dfrac{6\times 16}{7(7^2-1)} = 0.7143$ (4 s.f.)

Test for a positive association between the expert's order and the actual date order.

$H_0: \rho = 0,\ H_1: \rho > 0$

Sample size = 10

The choice of significance level is optional. However, the significance level used must be stated.

The critical value for r_s for a 0.01 significance level with a sample size of 7 is $r_s = 0.8929$.

As $0.7143 < 0.8929$, do not reject H_0. There is insufficient evidence at the 1% significance level that the expert is able to judge relative age accurately.

The critical value for r_s for a 0.025 significance level with a sample size of 7 is $r_s = 0.7857$.

As $0.7143 < 0.7857$, do not reject H_0. There is insufficient evidence at the 2.5% significance level that the expert is able to judge relative age accurately.

The critical value for r_s for a 0.05 significance level with a sample size of 7 is $r_s = 0.7143$.

In this case, the test statistic r_s equals the critical value, so there is not sufficient evidence to reject H_0. There is insufficient evidence at the 5% significance level that the expert is able to judge relative age accurately.

8 a $\bar{x} = \dfrac{\sum x}{20} = 4.535 \Rightarrow \sum x = 4.535 \times 20 = 90.7$

$\bar{t} = \dfrac{\sum t}{20} = 15.15 \Rightarrow \sum t = 15.15 \times 20 = 303$

$$r = \frac{S_{xt}}{\sqrt{S_{xx}S_{tt}}} = \frac{\sum xt - \dfrac{\sum x \sum t}{n}}{\sqrt{\left(\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}\right)\left(\sum t^2 - \dfrac{\left(\sum t\right)^2}{n}\right)}}$$

$$= \frac{1433.8 - \dfrac{(90.7)(303)}{20}}{\sqrt{\left(493.77 - \dfrac{90.7^2}{20}\right)\left(4897 - \dfrac{303^2}{20}\right)}} = 0.375 \text{ (3 s.f.)}$$

b $H_0: \rho = 0,\ H_1: \rho > 0$

Sample size = 20

Significance level = 0.05

The critical value for r for a 0.05 significance level with a sample size of 20 is $r = 0.3783$, so the critical region is $r > 0.3783$.

As $0.375 < 0.3783$, accept H_0. There is insufficient evidence of positive correlation between distance and time at the 5% level of significance.

c Both distance and journey time are normally distributed.

9 a The table shows the ranks for statistics and geography and d and d^2 for each pair of ranks:

Student	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***
Statistics	64	71	49	38	72	55	54	68
Geography	55	50	51	47	65	45	39	82
Rank, statistics	4	2	7	8	1	5	6	3
Rank, geography	3	5	4	6	2	7	8	1
d	1	–3	3	2	–1	–2	–2	2
***d*²**	1	9	9	4	1	4	4	4

$\sum d^2 = 36$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 36}{8(8^2-1)} = 0.571 \text{ (3 s.f.)}$$

9 b H_0: $\rho = 0$ There is no association between the marks in statistics and geography.

H_1: $\rho \neq 0$ There is an association between the marks in statistics and geography.

Sample size = 8

Significance level in each tail = 0.025

Critical values for Spearman's rank correlation coefficient r_s for a 0.025 significance level with a sample size of 8 are $r_s = \pm 0.7381$.

As $0.571 < 0.7381$, accept H_0. There is no evidence at the 5% significance level that there is an association between the results in statistics and geography. Students who are good at one of these subjects aren't necessarily going to do well in the other subject.

10 a The table shows the ranks for life expectancy and literacy and d and d^2 for each pair of ranks:

Life expectancy	49	76	69	71	50	64	78	74
Literacy	25	88	80	62	37	86	89	67
Rank, life expectancy	8	2	5	4	7	6	1	3
Rank, literacy	8	2	4	6	7	3	1	5
$\boldsymbol{d}$	0	0	1	–2	0	3	0	–2
$\boldsymbol{d^2}$	0	0	1	4	0	9	0	4

$\sum d^2 = 18$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 18}{8(8^2-1)} = 0.786 \text{ (3 s.f.)}$$

b H_0: $\rho = 0$ There is no correlation between the rankings of life expectancy and literacy.

H_1: $\rho \neq 0$ There is a correlation between the rankings of life expectancy and literacy.

Sample size = 8

Significance level in each tail = 0.025

Critical values for Spearman's rank correlation coefficient r_s for a 0.025 significance level with a sample size of 8 are $r_s = \pm 0.7381$.

As $0.7876 > 0.7381$, r_s lies within the critical region, so reject H_0. There is evidence at the 5% significance level of correlation between the rankings of life expectancy and literacy for women. In countries where a higher percentage of women are literate, they appear to have a higher life expectancy.

c It cannot be assumed that both female life expectancy and the percentage of women who are literate are both normally distributed.

d i The rank would still be the same, as the next highest percentage is 80. Therefore the coefficient would not change.

ii Both quantities would get the highest rank, thus $d = 0$. However, as n increases, the coefficient increases.

e Change each of the tied ranks to the corresponding mean of the tied ranks and calculate the product correlation coefficient directly from the ranked data.

11 a The table shows d and d^2 for each pair of ranks:

Official judge	1	2	3	4	5	6
Student vet	1	5	4	2	6	3
$\boldsymbol{d}$	0	–3	–1	2	–1	3
$\boldsymbol{d^2}$	0	9	1	4	1	9

$$\sum d^2 = 24$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 24}{6(6^2-1)} = 0.314 \text{ (3 s.f.)}$$

b H_0: $\rho = 0$ There is no correlation between the rankings of the official judge and the student vet.

H_1: $\rho > 0$ There is a positive correlation between the rankings of the judge and the student vet.

Sample size = 6

Significance level = 0.05

The critical value for r_s for a 0.05 significance level with a sample size of 6 is $r_s = 0.8286$.

As $0.66 > 0.5636$, accept H_0. There is insufficient evidence at the 5% significance level that there is an association between the rankings of the official judge and the student vet. They appear to be ranking using different criteria.

12 a $$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}} = \frac{13\,014 - \frac{(393)(326)}{10}}{\sqrt{\left(16\,483 - \frac{393^2}{10}\right)\left(10\,968 - \frac{326^2}{10}\right)}} = 0.340 \text{ (3 d.p.)}$$

b Use Spearman's rank correlation coefficient when one or both sets of data aren't from normal distributions; when at least one set of data is given as grades (letters) *or* ranking of preference or size; when the relationship between the data sets is non-linear.

c The table shows the ranks for x and y and d and d^2 for each pair of ranks:

$\boldsymbol{x}$	30	52	38	48	56	44	41	25	32	27
$\boldsymbol{y}$	22	38	40	34	35	32	28	27	29	41
$\boldsymbol{r_x}$	8	2	6	3	1	4	5	10	7	9
$\boldsymbol{r_y}$	10	3	2	5	4	6	8	9	7	1
$\boldsymbol{d}$	–2	–1	4	–2	–3	–2	–3	1	0	8
$\boldsymbol{d^2}$	4	1	16	4	9	4	9	1	0	64

$$\sum d^2 = 112$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 112}{10(10^2-1)} = 0.321 \text{ (3 s.f.)}$$

12 d The hypothesis test should use Spearman's rank correlation coefficient because it is unlikely that salary and age are both normally distributed.

H_0: $\rho = 0$ There is no correlation between age and salary.

H_1: $\rho \neq 0$ There is a correlation between the age and salary.

Sample size = 10

Significance level in each tail = 0.025

Critical values for Spearman's rank correlation coefficient r_s for a 0.025 significance level with a sample size of 10 are $r_s = \pm 0.6485$.

As $0.321 < 0.5636$, accept H_0. There is no evidence at the 5% significance level of a correlation between the rankings of salary and age. This means that in this profession, the older a person is doesn't mean they will necessarily earn more than younger colleagues.

13 a $\sum x = 487 \qquad \sum y = 923$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}} = \frac{62\,412 - \frac{(487)(923)}{10}}{\sqrt{\left(30\,625 - \frac{3487^2}{10}\right)\left(135\,481 - \frac{923^2}{10}\right)}} = 0.937 \text{ (3 d.p.)}$$

b The table shows the ranks for x and y and d and d^2 for each pair of ranks:

Machine	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***
x	63	12	34	81	51	14	45	74	24	89
y	111	25	41	181	64	21	51	145	43	241
r_x	4	10	7	2	5	9	6	3	8	1
r_y	4	9	8	2	5	10	6	3	7	1
d	0	1	–1	0	0	–1	0	0	1	0
d^2	0	1	1	0	0	1	0	0	1	0

$\sum d^2 = 4$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 4}{10(10^2-1)} = 0.976 \text{ (3 d.p.)}$$

c H_0: $\rho = 0$ There is no association between age of the machine and maintenance costs.

H_1: $\rho \neq 0$ There is an association between age of the machine and maintenance costs.

Sample size = 10

Significance level in each tail = 0.025

Critical values for Spearman's rank correlation coefficient r_s for a 0.025 significance level with a sample size of 10 are $r_s = \pm 0.6485$.

As $0.782 > 0.6485$, r_s lies within the critical region, so reject H_0. There is evidence at the 5% significance level of an association between the age of this type of machine and maintenance costs. It appears that older machines cost more to maintain.

14 a Using a calculator gives $r = -0.975$ (3 s.f.)

The summary statistics, which are not needed if calculating the product moment correlation coefficient by inputting the raw data into a calculator, are:

$\sum x = 6320 \quad \sum x^2 = 5\,639\,200 \quad \sum y = 112 \quad \sum y^2 = 1524 \quad \sum xy = 66\,450$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = -12\,198.9 \text{ (1 d.p.)}$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 130.2 \text{ (1 d.p.)}$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 1\,201155.6 \text{ (1 d.p.)}$$

b $H_0: \rho = 0, \ H_1: \rho < 0$

Sample size = 9

Significance level = 0.05

The critical value for r for a 0.05 significance level with a sample size of 9 is $r = -0.5822$.

As $-0.975 < -0.5822$, r_s lies within the critical region, so reject H_0. There is sufficient evidence at the 5% significance level that height above sea level and temperature are negatively correlated.

c H_0: $\rho = 0$ There is no association between hours of sunshine and temperature.

H_1: $\rho \neq 0$ There is an association between hours of sunshine and temperature.

Sample size = 9

Significance level in each tail = 0.025

Critical values for Spearman's rank correlation coefficient r_s for a 0.025 significance level with a sample size of 10 are $r_s = \pm 0.7000$.

As $0.767 > 0.7000$, r_s lies within the critical region, so reject H_0. There is evidence at the 5% significance level of a positive association between hours of sunshine and temperature. The more hours of sunshine the warmer the temperature.

15 a Spearman's rank correlation coefficient should be used if at least one of the sets of data isn't from a normal distribution, or if at least one of the sets of data is a letter grading or an order of preference (ranking). It should also be used if there is a non-linear association between the variables.

b The table shows d and d^2 for each pair of ranks:

Qualified judge	1	2	3	4	5	6	7	8	9	10
Trainee judge	1	2	5	6	7	8	10	4	3	9
$\boldsymbol{d}$	0	0	–2	–2	–2	–2	–3	4	6	1
$\boldsymbol{d^2}$	0	0	4	4	4	4	9	16	36	1

$\sum d^2 = 78$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 78}{10(10^2-1)} = 0.527 \text{ (3 s.f.)}$$

15 c H_0: $\rho = 0$ There is no correlation between the rankings of the qualified judge and trainee judge.

H_1: $\rho \neq 0$ There is a correlation between the rankings of the qualified judge and trainee judge.

Sample size = 10

Significance level = 0.05

The critical value for Spearman's rank correlation coefficient r_s for a 0.05 significance level with a sample size of 10 is $r_s = 0.5636$.

As $0.527 < 0.5636$, accept H_0. There is insufficient evidence at the 5% significance level of agreement between the rankings awarded by the qualified judge and trainee judge.

16 The table shows the respective ranks of each team for position and attendance (there are no tied ranks) and d and d^2 for each pair of ranks:

Club	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***
Position	1	2	3	4	5	6	7	8
Average attendance	30	32	12	19	27	18	15	25
r_{position}	1	2	3	4	5	6	7	8
$r_{\text{attendance}}$	2	1	8	5	3	6	7	4
d	–1	1	–5	–1	2	0	0	4
d^2	1	1	25	1	4	0	0	16

$\sum d^2 = 48$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 48}{8(8^2-1)} = 0.429 \text{ (3 s.f.)}$$

There is weak positive correlation between the ranking of attendance and position in the league. Being higher in the league doesn't necessarily mean the attendance will be higher.

17 a H_0: $\rho = 0$, H_1: $\rho > 0$

Sample size = 9

Significance level = 0.05

The critical value for r for a 0.05 significance level with a sample size of 9 is $r = 0.5822$.

As $0.972 > 0.5822$, r_s lies within the critical region, so reject H_0. There is sufficient evidence at the 5% significance level of a positive correlation between the age of a baby and its weight. This means the older a baby is, the heavier it is likely to be.

17 b The table shows the respective ranks for true weight and the boy's guesses (there are no tied ranks) and d and d^2 for each pair of ranks:

Baby	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***
Weight	4.4	5.2	5.8	6.4	6.7	7.2	7.6	7.9	8.4
r_{weight}	1	2	3	4	5	6	7	8	9
r_{boy}	1	4	2	6	3	8	5	9	7
d	0	–2	1	–2	2	–2	2	–1	2
d^2	0	4	1	4	4	4	4	1	4

$$\sum d^2 = 26$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 26}{9(9^2-1)} = 0.783 \text{ (3 s.f.)}$$

c $H_0: \rho = 0,\ H_1: \rho > 0$

Sample size = 9

Significance level = 0.05

The critical value for Spearman's rank correlation coefficient r_s for a 0.05 significance level with a sample size of 9 is $r_s = 0.6000$.

As $0.783 > 0.6000$, r_s lies within the critical region, so reject H_0. There is sufficient evidence at the 5% significance level that actual weight and the boys' guesses are positively correlated.

Challenge

a Since both x_i and y_i are integers from 1 to n with no ties:

$$\sum x_i^2 = \sum y_i^2 = \sum_{r=1}^{r=n} r^2 = \frac{1}{6}n(n+1)(1+2n)$$

And $\sum x_i = \sum y_i = \sum_{r=1}^{r=n} r = \frac{1}{2}n(n+1)$.

b First notice that $\bar{x} = \bar{y}$, as the sets of numbers are identical. Then:

$$\sum (y_i - \bar{y})^2 = \sum y_i^2 - 2\bar{y}\sum y_i + n\bar{y}^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 = \sum (x_i - \bar{x})^2$$

Therefore:

$$\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} = \sqrt{\sum (x_i - \bar{x})^2 \sum (x_i - \bar{x})^2} = \sum (x_i - \bar{x})^2$$

So:

$$\begin{aligned}\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} &= \sum (x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 \\ &= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2 \\ &= \frac{1}{6}n(n+1)(1+2n) - \frac{1}{4}n(n+1)^2 = n(n+1)\left(\frac{1}{6} + \frac{1}{3}n - \frac{1}{4}n - \frac{1}{4}\right) \\ &= \frac{1}{12}n(n+1)(n-1) = \frac{n(n^2-1)}{12}\end{aligned}$$

Challenge

c $\sum(y_i - x_i)^2 = \sum(y_i^2 - 2x_iy_i + x_i^2) = 2\sum x_i^2 - 2\sum x_iy_i$

$\Rightarrow \sum x_iy_i = \sum x_i^2 - \dfrac{\sum d_i^2}{2}$

d $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_iy_i - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y})$

$= \sum x_iy_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}$

$= \sum x_i^2 - \dfrac{1}{2}\sum d_i^2 - 2n\bar{x}^2 + n\bar{x}^2$ using the result of part **c** for $\sum x_iy_i$

$= \sum x_i^2 - \dfrac{1}{2}\sum d_i^2 - n\bar{x}^2$

$= \dfrac{1}{6}n(n+1)(1+2n) - \dfrac{1}{4}n(n+1)^2 - \sum d_i^2$

$= \dfrac{1}{12}n(n+1)(2+4n-3n-3) - \dfrac{1}{2}\sum d_i^2$

$= \dfrac{1}{12}n(n+1)(n-1) - \dfrac{1}{2}\sum d_i^2$

$= \dfrac{1}{12}n(n^2-1) - \dfrac{1}{2}\sum d_i^2$

e Using the results from parts **b** and **d**

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2\sum(y_i - \bar{y})^2}} = \frac{\frac{1}{12}n(n^2-1) - \frac{1}{2}\Sigma d_i^2}{\frac{1}{12}n(n^2-1)} = 1 - 6\frac{\sum d_i^2}{n(n^2-1)}$$