Bivariate Data

So far almost all of our statistics has concerned a single variable at a time. Many of the most interesting questions in statistics are about *relationships*: does revising more improve your test score? Do taller parents have taller children? Does a new fertiliser increase crop yield?

Definition. Bivariate data consists of pairs of values (x_i, y_i) of two variables, measured on the same set of individuals.

Definition. In a bivariate setting:

- the independent variable (or explanatory variable) is the one we think of as 'causing' or 'explaining' the other;
- the dependent variable (or response variable) is the one we think of as responding;
- a controlled variable is an independent variable whose values are chosen by the experimenter (e.g. measuring a reaction at temperatures of exactly $10°, 20°, 30°$C).

By convention the independent variable is plotted on the horizontal axis and called x; the dependent variable goes on the vertical axis and is called y. In some situations (e.g. marks in two different exams) neither variable is naturally independent.

Example

For each situation, identify the independent and dependent variables, and state whether the independent variable is controlled.

(a) The mass of a chemical that dissolves in water at temperatures of $10, 20, 30, 40, 50$ degrees Celsius.

(b) The number of hours of sunshine and the daily takings of an ice-cream van.

(c) Students' marks in a Latin exam and in a Maths exam.

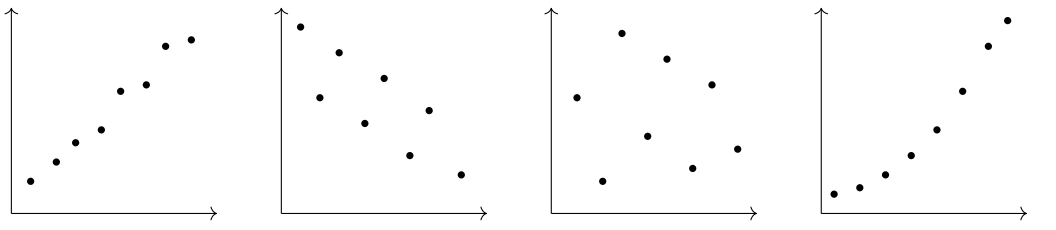
(a) Temperature is independent (and controlled, since the experimenter chooses the exact values); mass dissolved is dependent.

(b) Hours of sunshine is independent (but not controlled — nobody chooses the weather); takings are dependent.

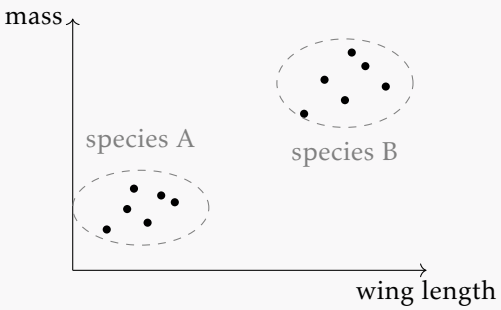
(c) Neither variable is naturally independent: it would be equally sensible to 'explain' either mark by the other. Here we simply have two response variables.

Scatter diagrams

We display bivariate data on a scatter diagram: one point per individual. The overall shape tells us about the relationship.



Remark (Distinct sections of the population)**.** A scatter diagram may show two (or more) distinct clusters — for example, plotting wing length against body mass for a sample containing two different species of bird. The clusters may each show strong correlation while the combined data shows weak correlation, or vice versa. Always look for, and comment on, distinct sections of the population.



Tip

In the exam you may be asked to add points to a given scatter diagram and interpret it, but you will not be asked to draw a complete scatter diagram from scratch.

Correlation Does Not Imply Causation

Fact — Correlation between two variables does **not** imply that one causes the other. There may be:

- a genuine causal relationship;
- a third variable (confounding variable) influencing both;
- no relationship at all — the correlation is coincidental.

Example

Daily ice-cream sales and daily incidents of drowning at beaches are positively correlated. Does ice cream cause drowning?

No. Hot weather is a third variable: on hot days more people buy ice cream and *more people swim, so more drownings occur. The correlation is real, but the causal arrows both come from temperature.*

Example

Students' results in two separate exams are usually positively correlated. Neither exam result causes the other — both are driven by underlying ability and preparation.

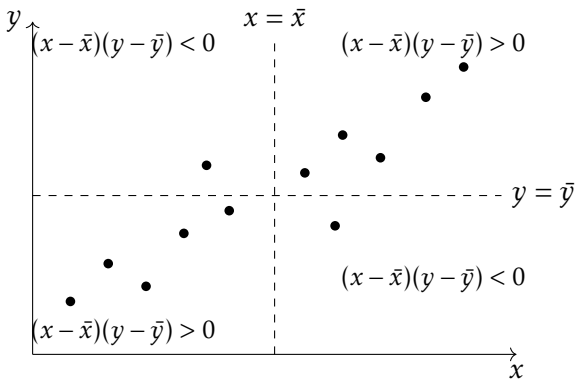
Tip

When asked to comment, be precise: "there is strong positive correlation between x and y, but this does not show that increasing x causes y to increase; both may be affected by [a named third variable in context]."

Textbook Exercises: [CUP.1] Ch 16 §4; [S1] Ch 9

Pearson's Product-Moment Correlation Coefficient

We want a *number* measuring how close the points on a scatter diagram lie to a straight line. The key idea: divide the diagram into four quadrants through the mean point $(\bar{x}, \bar{y})$.



For each point, the product $(x_i - \bar{x})(y_i - \bar{y})$ is positive in the top-right and bottom-left quadrants, and negative in the other two. So the sum

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

is large and positive for positive correlation, large and negative for negative correlation, and close to zero when the points are scattered evenly across all four quadrants. To turn this into a coefficient that does not depend on the units of measurement, we scale by the spread of each variable.

Definition (Summary statistics)**.**

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$
$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$
$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

Definition (Pearson's product-moment correlation coefficient)**.**

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}$$

Fact —

- $-1 \le r \le 1$ always.
- $r = 1$: perfect positive *linear* correlation (all points exactly on a line of positive gradient); $r = -1$: perfect negative linear correlation.
- r close to 0: little or no *linear* correlation. This does **not** rule out a non-linear relationship!

Remark. Note that $\frac{1}{n} S_{xx}$ is the variance of the x data, and $\frac{1}{n} S_{xy}$ is the *covariance* of x and y. So r is the covariance scaled by the two standard deviations — exactly the same normalisation that takes $\mathrm{Var}[X]$ to a unitless quantity. The fact that $|r| \le 1$ is the Cauchy–Schwarz inequality in disguise.

Tip

Your calculator computes r directly in its statistics/regression mode — use it to check any calculation by hand. In the exam you may be given raw data (small n) or summary statistics ($\sum x$, $\sum x^2$, $\sum xy$, etc.) — you will not be made to key in large data sets.

Example

Six students record their revision time x (hours) and test score y:

x	1	2	3	4	5	6
y	52	55	60	61	68	70

Calculate r and interpret it.

We compute: $\sum x = 21$, $\sum x^2 = 91$, $\sum y = 366$, $\sum y^2 = 22574$, $\sum xy = 1346$, $n = 6$.

$$S_{xx} = 91 - \frac{21^2}{6} = 17.5$$
$$S_{yy} = 22574 - \frac{366^2}{6} = 248$$
$$S_{xy} = 1346 - \frac{21 \times 366}{6} = 65$$
$$r = \frac{65}{\sqrt{17.5 \times 248}} = \frac{65}{\sqrt{4340}} = 0.987 \text{ (3 s.f.)}$$

There is very strong positive linear correlation between revision time and test score: students who revised longer tended to score higher.

Example

A sample of $n = 12$ pairs has summary statistics

$$\sum x = 66, \quad \sum y = 144, \quad \sum x^2 = 406, \quad \sum y^2 = 1796, \quad \sum xy = 825.$$

Calculate r.

$$S_{xx} = 406 - \frac{66^2}{12} = 43$$
$$S_{yy} = 1796 - \frac{144^2}{12} = 68$$
$$S_{xy} = 825 - \frac{66 \times 144}{12} = 33$$
$$r = \frac{33}{\sqrt{43 \times 68}} = \frac{33}{\sqrt{2924}} = 0.610 \text{ (3 s.f.)}$$

Effect of coding

Fact — Suppose $u = ax + b$ and $v = cy + d$ with $a, c \neq 0$. Then $r_{uv} = \pm r_{xy}$: linear coding leaves the correlation coefficient **unchanged**, unless exactly one of the variables is multiplied by a negative constant, in which case r changes sign (its magnitude is still unchanged).

The proof is a pleasant exercise in tracking constants through the summary statistics.

Adding a constant shifts every u_i and $\bar{u}$ equally, so the deviations are simply scaled: $u_i - \bar{u} = a(x_i - \bar{x})$ and $v_i - \bar{v} = c(y_i - \bar{y})$. Hence

$$S_{uu} = a^2 S_{xx}, \qquad S_{vv} = c^2 S_{yy}, \qquad S_{uv} = ac\, S_{xy},$$

and so, remembering that $\sqrt{a^2} = |a|$,

$$r_{uv} = \frac{ac\, S_{xy}}{|a||c|\sqrt{S_{xx}S_{yy}}} = \pm r_{xy},$$

with the + sign when $ac > 0$ and the − sign when $ac < 0$.

Example

Temperatures x (in °C) and energy use y (in kWh) give $r = -0.83$. What is the correlation coefficient between the temperature in °F, namely $u = 1.8x + 32$, and the energy use in MWh, $v = y/1000$?

Both codings multiply by positive constants (1.8 and 0.001), so the correlation coefficient is unchanged: $r_{uv} = -0.83$.

Textbook Exercises: [CUP.S] Ch 5 §1; [S1] Ch 9

Spearman's Rank Correlation Coefficient

Sometimes the data are not measurements but *rankings* — two judges ordering eight skaters, say. And sometimes a relationship is clearly increasing but not linear (e.g. exponential growth), so the pmcc undersells it. In both cases we measure agreement of *ranks*.

Definition (Spearman's rank correlation coefficient)**.** Rank each variable separately (1 to n), and let d_i be the difference between the ranks of the ith pair. Then

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}.$$

Remark. This formula is exactly the pmcc applied to the ranks — the special structure of the numbers $1,\dots,n$ collapses the pmcc formula to this simple form. So you can always **check your value of** r_s by typing the two lists of ranks into your calculator and computing the pmcc.

Fact —
- $r_s = 1$: the two rankings agree perfectly; $r_s = -1$: they are exact reverses.
- r_s measures association (any monotonic relationship), not just linear correlation.
- In practice: at most 10 pairs of values, and **no tied ranks**.

Example

Two judges rank eight bakers. Judge A's ranks are 1 to 8; Judge B ranks the same bakers $2,1,4,3,5,8,6,7$ respectively. Calculate r_s and comment.

A	1	2	3	4	5	6	7	8
B	2	1	4	3	5	8	6	7
d	−1	1	−1	1	0	−2	1	1
d^2	1	1	1	1	0	4	1	1

$\sum d^2 = 10$, *so*

$$r_s = 1 - \frac{6\times 10}{8(8^2-1)} = 1 - \frac{60}{504} = 0.881 \text{ (3 s.f.)}$$

The judges' rankings are in strong agreement.

Choosing between r and r_s

Fact —
- $r = 1 \implies r_s = 1$ (a perfect linear relationship preserves order), but **not conversely**: data lying exactly on $y = e^x$ has $r_s = 1$ but $r < 1$.
- Use the pmcc r when you are specifically interested in a *linear* relationship (and, for hypothesis tests, can assume bivariate normality).
- Use r_s when: the data are already ranks; the relationship appears monotonic but non-linear; or there are outliers / clearly non-normal data (r_s is robust, as it only sees order).
- If r and r_s are *both* close to zero, do not conclude there is no relationship: data lying on a parabola either side of its vertex can have both coefficients near zero, despite a perfect (non-monotonic)

relationship.

Example

The points $(-2,4),(-1,1),(0,0),(1,1),(2,4)$ lie exactly on $y = x^2$. Show that $S_{xy} = 0$, so that $r = 0$.

$\sum x = 0$, $\sum xy = (-2)(4)+(-1)(1)+0+(1)(1)+(2)(4) = -8-1+0+1+8 = 0$. *So* $S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 0-0 = 0$ *and hence* $r = 0$*: zero* linear *correlation, despite a perfect quadratic relationship. (Ranking the data would involve tied ranks here, but a similar non-monotonic example shows* r_s *can also be small.)*

Remark (Kendall's tau)**.** Another rank-based coefficient is Kendall's τ: of the $\binom{n}{2}$ pairs of data points, count how many are *concordant* (both coordinates in the same order) and how many *discordant*; then

$$\tau = \frac{(\text{concordant}) - (\text{discordant})}{\binom{n}{2}}.$$

It has a cleaner probabilistic interpretation than r_s and is widely used in practice.

Textbook Exercises: [CUP.S] Ch 5 §2; [S1] Ch 9

Hypothesis Tests Using the PMCC

The coefficient r computed from a sample is a *statistic*: a different sample would give a different value. The underlying *population* has a true correlation coefficient, denoted ρ (rho). Even when $\rho = 0$, a small sample will rarely give r exactly 0 — so how large does $|r|$ need to be before we believe there is genuine correlation?

Fact (Setting up the test) —

- $H_0 : \rho = 0$ (the null hypothesis is **always** $\rho = 0$).
- $H_1 : \rho > 0$, $\rho < 0$ or $\rho \neq 0$, according to whether we are testing for positive, negative or some linear correlation.
- You **must** use the symbol ρ in the hypotheses (not r), and define it in context: "where ρ is the population correlation coefficient between [variable 1] and [variable 2]".
- The test statistic is the sample value r, compared against a critical value from the table in the formula booklet (which depends on n and the significance level).

Remark (The bivariate normal assumption)**.** The critical-value tables are valid when the data come from a bivariate normal distribution — the exam allows you to assume this. Informally: a bivariate normal sample produces a roughly *elliptical* scatter of points, and slicing the joint distribution horizontally or vertically always gives a normal distribution. A scatter diagram that is roughly elliptical (no curvature, no funnel shapes, no outliers) suggests the assumption is reasonable.

Example

Ten students record their average nightly sleep x (hours) and their reaction time y (centiseconds). Summary statistics:

$$n = 10, \quad \sum x = 70, \quad \sum x^2 = 510, \quad \sum y = 50, \quad \sum y^2 = 270, \quad \sum xy = 336.$$

Test, at the 5% significance level, whether there is negative correlation between sleep and reaction time. (Critical value from the formula booklet: for $n = 10$, the one-tail 5% point is 0.5494.)

$H_0 : \rho = 0$, $H_1 : \rho < 0$, where ρ is the population correlation coefficient between nightly sleep and reaction time. Significance level 5%, one-tailed.

$$S_{xx} = 510 - \frac{70^2}{10} = 20$$

$$S_{yy} = 270 - \frac{50^2}{10} = 20$$

$$S_{xy} = 336 - \frac{70 \times 50}{10} = -14$$

$$r = \frac{-14}{\sqrt{20 \times 20}} = -0.7$$

Critical region: $r \leq -0.5494$. Since $-0.7 < -0.5494$, the result is significant: we reject H_0. There is sufficient evidence at the 5% level to suggest negative correlation between nightly sleep and reaction time in the population — students who sleep more tend to react faster.

Tip

Conclusions must never be over-assertive. Say "there is sufficient evidence to suggest…" or "there is insufficient evidence to suggest…" — never "accept H_1" or "this proves that…". Always finish with a sentence in context.

Example (OCR Further Statistics, June 2024)

A newspaper article claimed that "taller dog owners have taller dogs as pets". Alex investigated this claim and obtained data from a random sample of 16 fellow students who owned exactly one dog. The results are summarised as follows, where the height of the student, in cm, is denoted by h and the height of their dog, in cm, is denoted by d:

$$n = 16, \quad \sum h = 2880, \quad \sum d = 660, \quad \sum h^2 = 519276, \quad \sum d^2 = 30000, \quad \sum hd = 119425.$$

(a) Calculate the value of Pearson's product-moment correlation coefficient for the data.

(b) State what your answer tells you about a scatter diagram illustrating the data.

(c) Use the data to test, at the 5% significance level, the claim of the newspaper article. (For $n = 16$, the one-tail 5% point is 0.4259.)

(d) Explain whether the answer to part (a) would be likely to be different if the dogs' weights had been used instead of their heights.

(a)

$$S_{hh} = 519276 - \frac{2880^2}{16} = 876, \qquad S_{dd} = 30000 - \frac{660^2}{16} = 2775,$$
$$S_{hd} = 119425 - \frac{2880 \times 660}{16} = 625,$$
$$r = \frac{625}{\sqrt{876 \times 2775}} = 0.401 \text{ (3 s.f.)}$$

(b) The points would not lie very close to any straight line — only a weak upward linear trend would be visible. (The comment must be about the diagram*; "weak positive correlation" restates the number rather than interpreting it.)*

(c) $H_0: \rho = 0$, $H_1: \rho > 0$, *where* ρ *is the population correlation coefficient between owner's height and dog's height. Significance level* 5%, *one-tailed. Since* $0.401 < 0.4259$, *the result is not significant: we do not reject* H_0. *There is insufficient evidence at the* 5% *level to support the claim that taller dog owners have taller dogs.*

(d) The value would change, since weight is not a linear function of height (so this is not linear coding) — but as taller dogs are generally heavier, a weak positive correlation of broadly similar size would still be expected.

Using p-values and working backwards

You may instead be given (or asked for) a p-value: the probability, if H_0 is true, of obtaining a value of r at least as extreme as the one observed. We reject H_0 exactly when the p-value is less than the significance level. The tables let us trap a p-value between standard levels.

Example

A one-tailed test for positive correlation gives $r = 0.52$ from $n = 14$ data pairs. Extract from the critical-value table for $n = 14$:

one-tail level	5%	2.5%	1%	0.5%
critical value	0.4575	0.5324	0.6120	0.6614

(a) What can be said about the p-value of the test?

(b) At which of the levels 5%, 2.5%, 1% would H_0 be rejected?

(a) $r = 0.52$ lies between the 5% point (0.4575) and the 2.5% point (0.5324), so

$$0.025 < p\text{-value} < 0.05.$$

(b) The p-value is below 0.05 but above 0.025 (and hence above 0.01): reject H_0 at the 5% level; insufficient evidence to reject H_0 at the 2.5% or 1% levels.

Example (In class)

A researcher computes $r = 0.6$ from n pairs of observations and carries out a two-tailed test of $H_0 : \rho = 0$ at the 5% level. Using the table in the formula booklet, find the possible values of n for which H_0 is rejected.

Example (OCR Further Statistics, June 2022)

The directors of a large company believe that there are more computer failures in the Head Office when temperatures are higher. They obtain data for the Head Office for the maximum temperature, T °C, and the number of computer failures, X, on each of 12 randomly chosen days.

(a) State which of the following words can be applied to T: dependent, independent, controlled, response.

The data is summarised as follows:

$$n = 12, \quad \sum t = 261, \quad \sum x = 41, \quad \sum t^2 = 5869, \quad \sum x^2 = 311, \quad \sum tx = 1021.$$

(b) Calculate the value of the product-moment correlation coefficient r.

(c) The directors wish to investigate their belief using a significance test at the 1% level.

(i) Explain why a one-tailed test is appropriate in this situation.

(ii) Carry out the test. (For $n = 12$, the one-tail 1% point is 0.6581.)

(d) One of the directors prefers the temperatures to be given in Fahrenheit rather than Celsius, where $F = \frac{9}{5}C + 32$. State the value of r that would result from using temperatures in Fahrenheit in the calculation.

(a) Independent only. (Not controlled: nobody chooses the day's maximum temperature.)

(b)

$$S_{tt} = 5869 - \frac{261^2}{12} = 192.25, \qquad S_{xx} = 311 - \frac{41^2}{12} = 170.91\overline{6},$$
$$S_{tx} = 1021 - \frac{261 \times 41}{12} = 129.25,$$
$$r = \frac{129.25}{\sqrt{192.25 \times 170.91\overline{6}}} = 0.713 \text{ (3 s.f.)}$$

(c) (i) The directors' belief is directional — more *failures when temperatures are* higher — *so we test for positive correlation only.*

(ii) $H_0: \rho = 0$, $H_1: \rho > 0$, *where* ρ *is the population correlation coefficient between maximum temperature and number of computer failures. Since* $0.713 > 0.6581$, *the result is significant: we reject* H_0. *There is sufficient evidence at the* 1% *level to suggest positive correlation between temperature and the number of computer failures.*

(d) Converting to Fahrenheit is a linear coding with positive multiplier $\frac{9}{5}$, *so* r *is unchanged:* 0.713.

Hypothesis Tests Using Spearman's Coefficient

We can also test for *association* using r_s. This is a non-parametric test: it makes **no assumptions about the population** (no normality needed), because it only uses the ranks.

Fact (Setting up the test) — The hypotheses are written in words, naming the population:

- H_0: there is **no association** between [variable 1] and [variable 2] in the population.
- H_1: there is positive / negative / some association between [variable 1] and [variable 2] in the population.

The null hypothesis is always 'no association'. Critical values of r_s come from the formula booklet; they are computed by assuming that, under H_0, **all possible rankings are equally likely**.

Remark. Why words rather than symbols? There is no single population parameter that r_s estimates in the way that r estimates ρ, so hypotheses like "$\rho_s = 0$" are technically poor. Mark schemes expect the wording above, including the word *population*.

Example

Eight cyclists compete in two races. Their finishing positions are:

cyclist	A	B	C	D	E	F	G	H
race 1	1	2	3	4	5	6	7	8
race 2	1	3	2	4	6	5	8	7

Test, at the 5% significance level, whether the positions in the two races are positively associated. (Critical value: for $n = 8$, the one-tail 5% point of r_s is 0.6429.)

H_0: there is no association between the positions in race 1 and race 2 in the population of cyclists.
H_1: there is positive association between the positions in race 1 and race 2 in the population of cyclists.
Significance level 5%, one-tailed.
Differences d: $0, -1, 1, 0, -1, 1, -1, 1$, *so* $\sum d^2 = 6$.

$$r_s = 1 - \frac{6 \times 6}{8(8^2 - 1)} = 1 - \frac{36}{504} = 0.929 \text{ (3 s.f.)}$$

Since $0.929 > 0.6429$, the result is significant: we reject H_0. There is sufficient evidence at the 5% level to suggest positive association between the cyclists' positions in the two races.

Example

For a two-tailed test of association at the 5% level with $n = 10$ pairs, find the critical region for r_s. (For $n = 10$, the 2.5% one-tail point is 0.6485.)

A two-tailed test at 5% puts 2.5% in each tail, so the critical region is

$$r_s \geq 0.6485 \quad \text{or} \quad r_s \leq -0.6485.$$

Remark (Where the critical values come from)**.** Under H_0 all $n!$ rankings are equally likely, so for small n we can compute exact probabilities by counting. For example, with $n = 3$ there are $3! = 6$ equally likely rankings, and $r_s = 1$ only for the identity ranking, so $\mathbb{P}(r_s = 1) = \frac{1}{6}$ under H_0. The tables tabulate exactly these tail probabilities for larger n — a nice link to combinatorics, and a precursor to the non-parametric tests later in the course.

Example (OCR Further Statistics, March 2018)

At a wine-tasting competition, two judges give marks out of 100 to seven wines as follows.

wine	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Judge I	86.3	87.5	87.6	88.8	89.4	89.9	90.5
Judge II	85.3	88.1	82.7	87.7	89.0	89.4	91.5

(i) A spectator claims that there is a high level of agreement between the rank orders of the marks given by the two judges. Test the spectator's claim at the 1% significance level. (For $n = 7$, the one-tail 1% point of r_s is 0.8929.)

(ii) A competitor ranks the wines in a random order. The value of Spearman's rank correlation coefficient between the competitor's ranking and Judge I's ranking is r_s.

(a) Find the probability that $r_s = 1$.

(b) Show that r_s cannot take the value $\frac{55}{56}$.

(i) *H_0: there is no association between the two judges' marks in the population; H_1: there is positive association. Significance level 1%, one-tailed. Ranking each judge's marks (1 = lowest):*

Judge I	1	2	3	4	5	6	7
Judge II	2	4	1	3	5	6	7
d	−1	−2	2	1	0	0	0

$\sum d^2 = 10$, so

$$r_s = 1 - \frac{6 \times 10}{7(7^2 - 1)} = 1 - \frac{60}{336} = \frac{23}{28} = 0.821 \text{ (3 s.f.)}$$

Since $0.821 < 0.8929$, the result is not significant: we do not reject H_0. There is insufficient evidence at the 1% level of agreement between the judges' rank orders.

(ii) *(a) All 7! orderings are equally likely, and exactly one of them matches Judge I's ranking, so*

$$\mathbb{P}(r_s = 1) = \frac{1}{7!} = \frac{1}{5040}.$$

(b) With $n = 7$, $r_s = 1 - \frac{\sum d^2}{56}$, so $r_s = \frac{55}{56}$ would require $\sum d^2 = 1$. But this is impossible: a single misplaced rank forces at least one other rank to move too (the d_i sum to zero, so $\sum d_i^2$ has the same parity as $\sum d_i$ and is therefore even — it cannot equal 1).

Textbook Exercises: [CUP.S] Ch 5 §1–2; [CUP.2] Ch 18 §3