CORRELATION & REGRESSION Part 1

Question 1 (**)

The annual car sales of a small car manufacturer, c, and the annual advertising expenditure, £a, has product moment correlation coefficient r_{ac}.

The data is coded as

$$x = c - 7000 \quad \text{and} \quad y = \frac{a}{1000},$$

and the summary is shown in the table below.

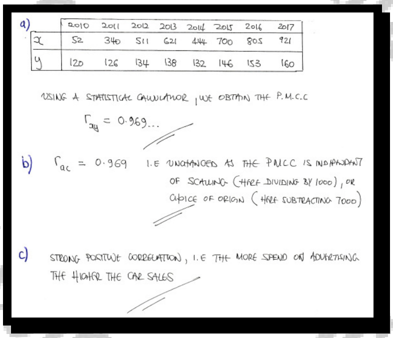
Year	2010	2011	2012	2013	2014	2015	2016	2017
x	52	340	511	621	444	700	805	921
y	120	126	134	138	132	146	153	160

a) Find, by a statistical calculator, the value of the product moment correlation coefficient between x and y, denoted by r_{xy}.

b) State with full justification the value of r_{ac}.

c) Interpret the value of r_{ac}.

$\boxed{r_{xy} \approx 0.969}$, $\boxed{r_{ac} \approx 0.969}$


a)

	2010	2011	2012	2013	2014	2015	2016	2017
x	52	340	511	621	444	700	805	921
y	120	126	134	138	132	146	153	160

USING A STATISTICAL CALCULATOR, WE OBTAIN THE P.M.C.C

$r_{xy} = 0.969...$

b) $r_{ac} = 0.969$ I.E UNCHANGED AS THE PMCC IS INDEPENDENT OF SCALING (HERE DIVIDING BY 1000), OR CHOICE OF ORIGIN (HERE SUBTRACTING 7000)

c) STRONG POSITIVE CORRELATION, I.E THE MORE SPEND ON ADVERTISING THE HIGHER THE CAR SALES

Question 2 **(**)**

The percentage mock exam marks, of a random sample of 8 G.C.S.E. students, in Geography and History are recorded in the table below.

Student	A	B	C	D	E	F	G	H
Geography	80	29	56	56	58	45	67	72
History	78	49	65	50	75	50	60	47

Test, at the 10% level of significance, whether there is evidence of positive correlation between the percentage mock exam marks in Geography and History.

$\boxed{\quad}$, $\boxed{\text{not significant evidence as } 0.4897 < 0.5067}$

USING A CALCULATOR IN STATS MODE TO OBTAIN THE P.M.C.C

(x) GEOGRAPHY %	80	29	56	56	58	45	67	72
(y) HISTORY %	78	49	65	50	75	50	60	47

$r = 0.4896746\ldots \simeq 0.4897$

NEXT SETTING HYPOTHESES

- $H_0: \rho = 0$
- $H_1: \rho > 0$, WHERE ρ IS THE P.M.C.C OF THE ENTIRE POPULATION, NOT THAT OF THE SAMPLE OF 8

THE CRITICAL VALUE FOR $n=8$, AT 10% SIGNIFICANCE IS 0.5067

AS $0.4897 < 0.5067$ THERE IS NO SIGNIFICANT EVIDENCE OF POSITIVE CORRELATION BETWEEN THE PERCENTAGE MARKS IN GEOGRAPHY & HISTORY.

THERE IS NO SUFFICIENT EVIDENCE TO REJECT H_0

Question 3 (**)

The table below shows the number of Maths teachers x, working in 8 different towns and the number of burglaries y, committed in a given month in the same 8 towns.

Town	A	B	C	D	E	F	G	H
x	35	42	21	55	33	29	39	40
y	30	28	21	38	35	27	30	k

a) Use a statistical calculator to find the product moment correlation coefficient between the number of maths teachers and the number of burglaries, for the towns A to G.

b) Interpret the value of the product moment correlation coefficient in the context of this question.

c) Test, at the 5% level of significance, whether there is evidence of positive correlation between the number of maths teachers and the number of burglaries, for the towns A to G.

d) Comment on the statement

"… the Maths teachers are likely to be responsible for the burglaries …"

e) Use linear regression to estimate the value of k, for town H.

$\boxed{}$, $\boxed{r \approx 0.792}$, $\boxed{\text{significant evidence as } 0.792 > 0.6694}$, $\boxed{k \approx 31}$

a) INPUTTING THE FIRST 7 PAIRS OF DATA INTO A STATISTICAL CALCULATOR WE OBTAIN

$r = 0.792$

b) AS THE NUMBER OF MATHS TEACHERS INCREASES, SO DO THE NUMBER OF BURGLARIES, IE POSITIVE CORRELATION

c) SETTING HYPOTHESES

$H_0 : \rho = 0$

$H_1 : \rho > 0$, WHERE ρ DENOTES THE P.M.C.C OF ALL, IE THE POPULATION, NOT JUST THE SAMPLE OF 7

THE CRITICAL VALUE FOR $n=7$, AT 5% SIGNIFICANCE IS 0.6694

AS $0.792 > 0.6694$, THERE SUFFICIENT EVIDENCE OF POSITIVE CORRELATION, I.E SUFFICIENT EVIDENCE TO REJECT H_0

d) CORRELATION DOES NOT IMPLY CAUSE, THERE MIGHT BE A CONNECTION TO A THIRD VARIABLE HERE THE TOWNS' POPULATIONS

THE STATEMENT IS NOT LIKELY TO BE TRUE

e) USING A STATISTICAL CALCULATOR TO OBTAIN A REGRESSION LINE

$y = a + bx$

$y = 0.408x + 15.1$ (TO 3 s.f.)

WHEN $x = 40$

$y = 0.408 \times 40 + 15.1 \simeq 31.42\ldots \simeq 31$

Question 4 (**)

The table below shows the marks obtained by a group of students, in two separate tests.

Student	A	B	C	D	E	F	G	H
Test 1	28	39	18	30	42	43	33	10
Test 2	12	23	16	16	28	18	24	7

The first test is out of 50 marks while the second test is out of 30 marks.

Let x and y represent the marks obtained in Test 1 and Test 2, respectively.

a) Use a statistical calculator to find the value of the product moment correlation coefficient between x and y.

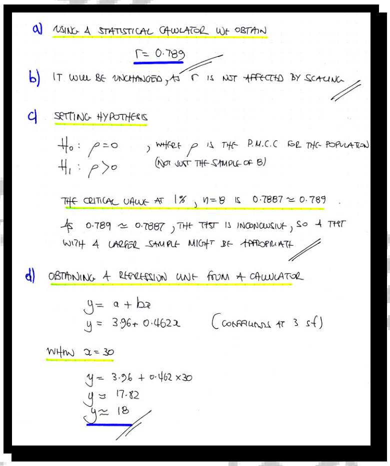
b) Explain how the value of the product moment correlation coefficient between x and y will be affected if the individual test marks were converted into percentage marks.

c) Test, at the 1% level of significance, whether there is evidence of positive correlation between x and y.

A student was absent from the second test but he obtained 30 marks in the first test.

d) Use linear regression to estimate this student's mark in the second test.

$\boxed{r \approx 0.789}$, $\boxed{\text{unchanged}}$, $\boxed{\text{inconclusive test as } 0.789 \approx 0.7889}$, $\boxed{\approx 18}$



a) USING A STATISTICAL CALCULATOR WE OBTAIN

$r = 0.789$

b) IT WILL BE UNCHANGED, AS r IS NOT AFFECTED BY SCALING

c) SETTING HYPOTHESIS

$H_0: \rho = 0$, WHERE ρ IS THE P.M.C.C FOR THE POPULATION (NOT JUST THE SAMPLE OF 8)

$H_1: \rho > 0$

THE CRITICAL VALUE AT 1%, $n = 8$ IS $0.7887 \simeq 0.789$.

AS $0.789 \simeq 0.7887$, THE TEST IS INCONCLUSIVE, SO A TEST WITH A LARGER SAMPLE MIGHT BE APPROPRIATE

d) OBTAINING A REGRESSION LINE FROM A CALCULATOR

$y = a + bx$

$y = 3.96 + 0.462x$ (COEFFICIENTS AT 3 s.f.)

WHEN $x = 30$

$y = 3.96 + 0.462 \times 30$

$y \simeq 17.82$

$y \simeq 18$

Question 5 (**)

The table below shows the maximum daytime temperature, in °C, at a certain city centre, and the amount of a certain pollutant in mg per litre.

Maximum Temperature	10	12	14	16	18	20	22	24
Amount of Pollutant	513	475	525	530	516	520	507	521

a) Find, using a statistical calculator, the value of the product moment correlation coefficient for the above data.

b) State, with justification, the value of the product moment correlation coefficient, if the maximum daily temperatures were to be measured in degrees Fahrenheit.

c) Test, at the 10% level of significance, whether there is evidence of positive correlation in these bivariate data.

$\boxed{}$, $\boxed{r = 0.320}$, $\boxed{\text{unchanged}}$, $\boxed{\text{no significant evidence as } 0.320 < 0.5067}$

MAX TEMPERATURE °C	10	12	14	16	18	20	22	24
AMOUNT OF POLLUTANT (in mg / litre)	513	475	525	530	516	520	507	521

a) FROM CALCULATOR IN STAT MODE

$r = 0.320$

b) UNCHANGED AT 0.320, AS THE P.M.C.C IS INDEPENDENT OF SCALING (OR CHANGE OF ORIGIN)

c) SETTING HYPOTHESES

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

WHERE ρ IS THE P.M.C.C OF ALL PAIRINGS OF TEMPERATURES & AMOUNT OF POLLUTANT (POPULATION)

THE CRITICAL VALUE AT 10% SIGNIFICANCE & $n = 8$ IS 0.5067

AS $0.320 < 0.5067$ IT APPEARS THERE IS NO POSITIVE CORRELATION BETWEEN THE MAX DAILY TEMPERATURE & THE AMOUNT OF POLLUTANT

INSUFFICIENT EVIDENCE TO REJECT H_0

Question 6 (**)

The percentage test exam marks, of a random sample of 8 students, in Physics and Chemistry are recorded in the table below.

Student	A	B	C	D	E	F	G	H
Physics	70	36	56	56	58	45	67	72
Chemistry	78	49	55	50	75	50	60	57

Test, at the 5% level of significance, whether there is evidence of positive correlation between the percentage test marks in Physics and Chemistry.

$\boxed{}$, $\boxed{\text{not significant evidence as } 0.5814 < 0.6215}$

USING THE CALCULATOR IN STATISTICAL MODE

$P.M.C.C = r = 0.5814...$

SETTING HYPOTHESES

$H_0 : \rho = 0$

$H_1 : \rho > 0$

WHERE ρ IS THE P.M.C.C. OF THE ENTIRE POPULATION

THE CRITICAL VALUE AT $n=8$ AT 5% SIGNIFICANCE IS 0.6215

AS $0.5814 < 0.6215$ THERE IS NO SIGNIFICANT EVIDENCE OF POSITIVE CORRELATION BETWEEN THE TEST MARKS IN PHYSICS & CHEMISTRY

INSUFFICIENT EVIDENCE TO REJECT H_0

Question 7 (**)

The table below shows the daily number of shoplifting incidents in a shopping mall, for a given seven day week and the number of the security guards employed in each of these seven days.

Number of Shoplifting Incidents	17	20	23	11	35	32	21
Number of Security Guards Employed	6	6	5	7	4	3	5

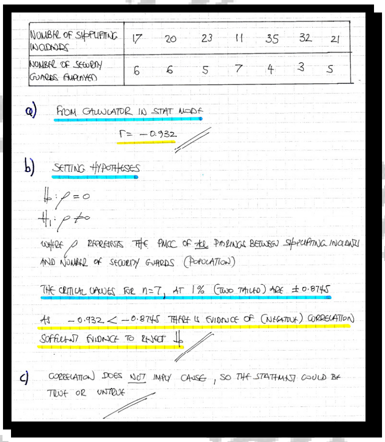
a) Find, using a statistical calculator, the value of the product moment correlation coefficient for these data.

b) Test, at the 1% level of significance, whether there is evidence of correlation in these bivariate data.

c) Briefly comment on the statement:

"… Increasing the number of security guards will result in a decrease in the shoplifting incidents …"

$\boxed{}$, $\boxed{r = -0.932}$, $\boxed{\text{significant evidence as } -0.932 < -0.8745}$



NUMBER OF SHOPLIFTING INCIDENTS	17	20	23	11	35	32	21
NUMBER OF SECURITY GUARDS EMPLOYED	6	6	5	7	4	3	5

a) FROM CALCULATOR IN STAT MODE

$r = -0.932$

b) SETTING HYPOTHESES

$H_0: \rho = 0$

$H_1: \rho \neq 0$

WHERE ρ REPRESENTS THE PMCC OF ALL PAIRINGS BETWEEN SHOPLIFTING INCIDENTS AND NUMBER OF SECURITY GUARDS (POPULATION)

THE CRITICAL VALUES FOR $n=7$, AT 1% (TWO TAILED) ARE ± 0.8745

AS $-0.932 < -0.8745$ THERE IS EVIDENCE OF (NEGATIVE) CORRELATION

SUFFICIENT EVIDENCE TO REJECT H_0

c) CORRELATION DOES NOT IMPLY CAUSE, SO THE STATEMENT COULD BE TRUE OR UNTRUE

Question 89 (**)

An electrical appliances supplier wishes to investigate the impact of advertising on the sales of his washing machines.

He records the number of monthly advertisements placed on the local radio station and the number of washing machines sold.

This is a table of his results.

Number of Advertisements (x)	52	37	66	45	77	27	80	19	47	40
Number of Washing Machines Sold (y)	180	115	171	166	177	99	174	100	143	164

Test, at the 10% level of significance, whether there is evidence of correlation between x and y, and explain what conclusions the electrical appliances supplier should make from this value.

$\boxed{}$, $\boxed{\text{significant evidence as } 0.817 > 0.5494}$

USING THE CALCULATOR IN STATISTIC MODE, THE P.M.C.C IS

$r = 0.816579... \simeq 0.817$

SETTING HYPOTHESES

- $H_0 : \rho = 0$
- $H_1 : \rho \neq 0$

WHERE ρ IS THE P.M.C.C FOR "ALL DATA", I.E. POPULATION

THE CRITICAL VALUES AT 10%, TWO TAILED, FOR $n = 10$ ARE ± 0.5494

AS $0.817 > 0.5494$ THERE IS EVIDENCE OF (POSITIVE) CORRELATION, SO SUFFICIENT EVIDENCE TO REJECT H_0

CONCLUSION DRAWN

AS THERE IS EVIDENCE OF POSITIVE CORRELATION, SO PLACING ADS IN THE LOCAL RADIO APPEARS TO HAVE THE DESIRED EFFECT.

Question 9 (**)

The table below shows the number of Maths teachers x, working in 8 different schools and the number of students y, in each of these 8 schools.

School	A	B	C	D	E	F	G	H
$\boldsymbol{x}$	5	9	11	17	12	10	9	8
$\boldsymbol{y}$	225	247	334	811	382	340	285	k

a) Use a statistical calculator to find the product moment correlation coefficient between the number of maths teachers and the number of students, for the schools A to G.

b) Use linear regression to estimate the value of k, for school H.

Justify the reliability of the estimate.

$\boxed{r \approx 0.913}$, $\boxed{k \approx 252-253}$

a) USING A CALCULATOR IN STATISTICS MODE, THE P.M.C.C BASED ON 7 OBS

$r = 0.913$

b) AGAIN FROM A STATISTICAL CALCULATOR

$a = -147.6$
$b = 50.1$

$\Rightarrow \quad y = 50.1x - 147.6$

$\Rightarrow \quad y = 50.1 \times 8 - 147.6$

$\Rightarrow \quad y \simeq 253 \quad \leftarrow \text{"}k\text{"}$

AS 8 LIES IN THE RANGE $5 \leq x \leq 17$, WHICH WAS USED TO PRODUCE THE REGRESSION LINE AND r INDICATES STRONG CORRELATION, THE ESTIMATE SHOULD BE VERY RELIABLE

Question 10 (**)

The table below shows the times obtained by a group of students, in two separate runs of a lap of the school's stadium.

Student	A	B	C	D	E	F	G	H
Run 1 (sec)	65	76	71	73	76	69	60	66
Run 2 (sec)	71	78	68	68	74	75	64	66

Let x and y represent the times obtained in Run 1 and Run 2, respectively.

a) Use a statistical calculator to find the value of the product moment correlation coefficient between x and y.

b) Explain how the value of the product moment correlation coefficient between x and y will be affected if the individual times were converted into minutes.

c) Test, at the 1% level of significance, whether there is evidence of positive correlation between x and y.

d) Repeat the test of part **(c)** at the 5% level of significance,

A student was absent from Run 1 but he ran Run 2 in 80 seconds.

e) Use linear regression to estimate this student's time in Run 2.

$\boxed{}$, $\boxed{r \approx 0.692}$, $\boxed{\text{unchanged}}$, $\boxed{\text{not significant at 1\% as } 0.692 < 0.789}$, $\boxed{\text{significant at 5\% as } 0.692 > 0.6215}$, $\boxed{\approx 77}$

a) FROM CALCULATOR IN STATISTICAL MODE

P.M.C.C. $= r = 0.692$

b) UNCHANGED AT 0.692

(INDEPENDENT OF SCALING/UNITS)

c) SETTING UP HYPOTHESES

$H_0 : \rho = 0$
$H_1 : \rho > 0$ } WHERE ρ IS THE P.M.C.C OF THE ENTIRE POPULATION

THE CRITICAL VALUE FOR $n=8$, AT 1% SIGNIFICANCE IS 0.7887 (TABLES)

AS $0.692 < 0.7887$ THERE IS NO SIGNIFICANT EVIDENCE OF POSITIVE CORRELATION – INSUFFICIENT EVIDENCE TO REJECT H_0

d) THE CRITICAL VALUE FOR $n=8$ AT 5% IS NOW 0.6215 (SAME HYPOTHESES)

AS $0.692 > 0.6215$ THERE IS SIGNIFICANT EVIDENCE OF POSITIVE CORRELATION AT 5% – SUFFICIENT EVIDENCE TO REJECT H_0

e) USING A STATISTICAL CALCULATOR WE OBTAIN THE REGRESSION LINE

$y = 29.2 + 0.595x$

USING $x = 80$

$y = 29.2 + 0.595 \times 80$

$y = 76.8\ldots$

$\therefore \simeq 77$ seconds

Question 11 (**)

The table below shows the number of priests x, working in 8 different towns and the number of shoplifting incidents y, committed in a given month in the same 8 towns.

Town	A	B	C	D	E	F	G	H
$\boldsymbol{x}$	15	12	11	25	23	19	19	22
$\boldsymbol{y}$	310	281	215	328	305	277	300	k

a) Use a statistical calculator to find the product moment correlation coefficient between the number of priests and the number of shoplifting incidents, for the towns A to G. **(1)**

b) Interpret the value of the product moment correlation coefficient in the context of this question. **(1)**

c) Explain how the value of the product moment correlation coefficient between x and y will be affected if the burglaries were recorded in **hundreds**. **(1)**

d) Test, at the 5% level of significance, whether there is evidence of positive correlation between the number of priests and the number of shoplifting incidents, for the towns A to G. **(4)**

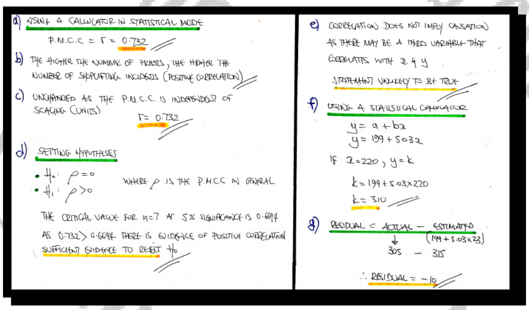
e) Comment on the statement

"… the priests are likely to be responsible for the shopliftings …" **(1)**

f) Use linear regression to estimate the value of k, for town H. **(3)**

g) Calculate the residual for town E. **(2)**

$\boxed{}$, $\boxed{r \approx 0.732}$, $\boxed{\text{significant evidence as } 0.732 > 0.6694}$, $\boxed{k \approx 310}$, $\boxed{\approx -10}$



a) USING A CALCULATOR IN STATISTICAL MODE

P.M.C.C $= r = 0.732$

b) THE HIGHER THE NUMBER OF PRIESTS, THE HIGHER THE NUMBER OF SHOPLIFTING INCIDENTS (POSITIVE CORRELATION)

c) UNCHANGED AS THE P.M.C.C. IS INDEPENDENT OF SCALING (UNITS)

$r = 0.732$

d) SETTING HYPOTHESES

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

WHERE ρ IS THE P.M.C.C IN GENERAL

THE CRITICAL VALUE FOR $n=7$ AT 5% SIGNIFICANCE IS 0.6694

AS $0.732 > 0.6694$ THERE IS EVIDENCE OF POSITIVE CORRELATION

SUFFICIENT EVIDENCE TO REJECT H_0

e) CORRELATION DOES NOT IMPLY CAUSATION AS THERE MAY BE A THIRD VARIABLE THAT CORRELATES WITH x & y

STATEMENT UNLIKELY TO BE TRUE

f) USING A STATISTICAL CALCULATOR

$y = a + bx$

$y = 199 + 5.03x$

IF $x = 220$, $y = k$

$k = 199 + 5.03 \times 220$

$k = 310$

g) RESIDUAL = ACTUAL − ESTIMATED

$305 - 315$ ($199 + 5.03 \times 23$)

$\therefore$ RESIDUAL $= -10$

Question 12 (***)

The table below shows the number of revolutions of a drill bit N, and the maximum temperature T °C reached by this drill bit after revolving for one minute.

N	600	700	900	1050	1300
T	49	47	52	52	53

a) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

b) Use a statistical calculator to determine …

i. … the value of the product moment correlation coefficient between N and T.

ii. … the equation of the regression line between T and N, giving the answer in the form

$$T = a + bN,$$

where a and b are constants.

c) Interpret in the context of this question the physical meaning of a and b.

Comment further on the likely value of a in a real life scenario.

d) Use the equation of the regression line to estimate the value of T when …

i. … $N = 1600$.

ii. … $N = 825$.

Comment further on the reliability of each of these estimates.

$\boxed{r = 0.845}$, $\boxed{T = 43.7 + 0.0076N}$, $\boxed{N_{1600} \approx 56}$, $\boxed{N_{825} \approx 50}$

a) EXPLANATORY (INDEPENDENT) IS THE "N" AS IT IS "N" THAT AFFECTS THE TEMPERATURE AND NOT THE OTHER WAY ROUND

RESPONSE VARIABLE (DEPENDENT) IS THE "T"

b) I) FROM A STATISTICAL CALCULATOR

P.M.C.C $= r = 0.845$

II) ALSO USING A CALCULATOR

$T = 43.7 + 0.0076N$

c) a — "y INTERCEPT"

a IS THE TEMPERATURE OF THE DRILL BIT BEFORE IT IS REVOLVED

b — "GRADIENT"

b IS THE EXTRA TEMPERATURE RISE PER REVOLUTION OF THE DRILL BIT

a IS UNLIKELY TO BE 43.7°C AS THIS WOULD REPRESENT THE "ROOM TEMPERATURE"

d) USING THE REGRESSION LINE

$T = 43.7 + 0.0076N$

- $T_{1600} = 43.7 + 0.0076 \times 1600 \approx 55.9°$

UNRELIABLE AS "N" IS WAY ABOVE THE LARGEST VALUE OF N THAT WAS USED TO CREATE THE EQUATION OF THE REGRESSION LINE

- $T_{825} = 43.7 + 0.0076 \times 825 \approx 50.0$

POSSIBLY RELIABLE AS IT LIES WITHIN THE VALUES OF "N" WHICH WAS USED TO CREATE THE REGRESSION LINE

Question 13 (***)

The table below shows the amount spent per month by a local radio on marketing M, in £1000, and the number of listeners L, in 1000, in that month.

M	6.5	8	11.5	14.25	15
L	87	139	119	127	147

a) Use a statistical calculator to find …

i. … the value of the product moment correlation coefficient between M and L.

ii. … the equation of the regression line between M and L, giving the answer in the form

$$L = a + bM,$$

where a and b are constants.

b) Use the equation of the regression line to estimate the number of cars that are expected to be sold in a month where the amount spent on marketing and advertising is …

i. … £9,800.

ii. … £20,000.

Comment further on the reliability of each of these two estimates.

c) Interpret in the context of this question the physical meaning of a and b.

d) Calculate the residual of the month where £14,250 was spent.

$\boxed{}$, $\boxed{r = 0.635}$, $\boxed{L = 80.3 + 3.94M}$, $\boxed{M_{9.8} \approx 119}$, $\boxed{M_{20} \approx 159}$, $\boxed{\approx -11}$

a) USING A CALCULATOR IN STATISTICAL MODE

I) P.M.C.C = r = 0.635

II) REGRESSION LINE ⟹ L = a + bM

L = 80.3 + 3.94M

b) USING THE REGRESSION LINE

I) IF M = 9.8

L = 80.3 + 3.94 × 9.8 = 119

SHOULD BE RELIABLE (WITHIN REASON AS THE PMCC IS ONLY 0.635) AS THE VALUE OF M LIES WITHIN THE VALUES OF M THAT WAS USED TO CREATE THE REGRESSION LINE

II) IF M = 20

L = 80.3 + 3.94 × 20 = 159

NOT LIKELY TO BE RELIABLE AS THIS VALUE OF M IS "WAY ABOVE" IS (EXTRAPOLATION)

c) a = "Y INTERCEPT"

a IS THE NUMBER OF LISTENERS IF NO MONEY WAS SPEND IN MARKETING/ADVERTISING

b = "GRADIENT"

b IS EXTRA LISTENERS PER 1000 SPENT ON MARKETING/ADVERTISING

d) "RESIDUAL = ACTUAL − PREDICTED"

(14.25) 125

80.3 + 3.94 × 14.25 = 136

∴ RESIDUAL = −11

Question 14 (***)

The table below shows the amount spent per month by a car dealership on marketing and advertising m, in £1000, and the number of cars c sold that month.

m	6	7	8	9	10
c	8	13	11	12	14

a) Use a statistical calculator to find …

i. … the value of the product moment correlation coefficient between m and c.

ii. … the equation of the regression line between m and c, giving the answer in the form

$$c = a + bm,$$

where a and b are constants.

b) Use the equation of the regression line to estimate the number of cars that are expected to be sold in a month where the amount spent on marketing and advertising is …

i. … £8,800.

ii. … £20,000.

Comment further on the reliability of each of these two estimates.

c) Interpret in the context of this question the physical meaning of a and b.

$\boxed{}$, $\boxed{r = 0.755}$, $\boxed{c = 2.8 + 1.1m}$, $\boxed{c_{8.8} \approx 12}$, $\boxed{c_{20} \approx 25}$

a) USING A STATISTICAL CALCULATOR WE OBTAIN

$r = 0.755$

b) FROM THE CALCULATOR

$y = a + bx \implies c = a + bm$

$\implies c = 2.8 + 1.1m$

c) I) $C_{8.8} = 2.8 + 1.1 \times 8.8 \approx 12.48$ IE AROUND 12 CARS

AS THE P.M.C.C IS FAIRLY HIGH, AND 8.8 (£8,800) LIES WITHIN THE RANGE OF VALUES OF m WHICH WAS USED FOR THE REGRESSION LINE, THE ESTIMATE SHOULD BE RELIABLE (INTERPOLATION)

II) $C_{20} = 2.8 + 1.1 \times 20 \approx 24.8$ IE AROUND 25 CARS

AS $m = 20$ IS "WAY ABOVE" THE LARGEST VALUE OF m, WHICH WAS USED TO PRODUCE THE REGRESSION LINE, THE ESTIMATE COULD NOT BE RELIABLE (EXTRAPOLATION)

d) • $a = 2.8$ ("y INTERCEPT")

THE NUMBER OF CARS EXPECTED TO BE SOLD IF NO MONEY WAS SPENT ON ADVERTISING

• $b = 1.1$ ("GRADIENT")

THE NUMBER OF EXTRA CARS EXPECTED TO BE SOLD PER £1000 SPENT ON ADVERTISING

Question 15 (***)

The table below shows the maximum temperature T °C on five different days and the corresponding ice cream sales, N, of a certain shop on those days.

T	15	20	25	30	35
N	69	165	172	200	232

a) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

b) Use a statistical calculator to determine …

i. … the value of the product moment correlation coefficient between T and N.

ii. … the equation of the regression line between N and T, giving the answer in the form

$$N = a + bT,$$

where a and b are constants.

c) Interpret in the context of this question the physical meaning of a and b.

d) Use the equation of the regression line to estimate the value of N when …

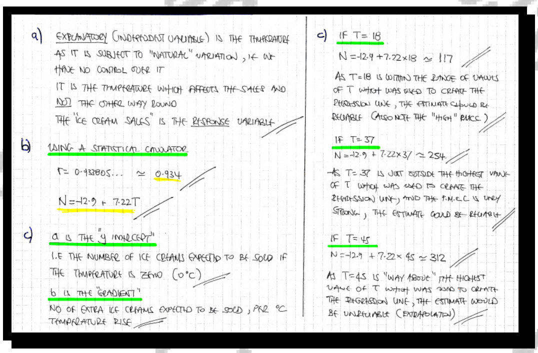
i. … $T = 18$°C.

ii. … $T = 37$°C.

iii. … $T = 45$°C

Comment further on the reliability of each of these estimates.

$\boxed{}$, $\boxed{r = 0.934}$, $\boxed{N = 7.22T - 12.9}$, $\boxed{N_{18} \approx 117}$, $\boxed{N_{37} \approx 254}$, $\boxed{T_{45} \approx 312}$


a) EXPLANATORY (INDEPENDENT VARIABLE) IS THE TEMPERATURE AS IT IS SUBJECT TO "NATURAL" VARIATION, I.E. WE HAVE NO CONTROL OVER IT
IT IS THE TEMPERATURE WHICH AFFECTS THE SALES AND NOT THE OTHER WAY ROUND
THE "ICE CREAM SALES" IS THE RESPONSE VARIABLE

b) USING A STATISTICAL CALCULATOR
$r = 0.933805\ldots \simeq 0.934$
$N = -12.9 + 7.22T$

c) a IS THE "y INTERCEPT"
I.E THE NUMBER OF ICE CREAMS EXPECTED TO BE SOLD IF THE TEMPERATURE IS ZERO (0°C)
b IS THE "GRADIENT"
NO OF EXTRA ICE CREAMS EXPECTED TO BE SOLD, PER °C TEMPERATURE RISE

c) IF $T = 18$
$N = -12.9 + 7.22 \times 18 \simeq 117$
AS $T = 18$ IS WITHIN THE RANGE OF VALUES OF T WHICH WAS USED TO CREATE THE REGRESSION LINE, THE ESTIMATE SHOULD BE RELIABLE (ALSO NOTE THE "HIGH" PMCC)

IF $T = 37$
$N = -12.9 + 7.22 \times 37 \simeq 254$
AS $T = 37$ IS JUST OUTSIDE THE HIGHEST VALUE OF T WHICH WAS USED TO CREATE THE REGRESSION LINE, AND THE P.M.C.C IS VERY STRONG, THE ESTIMATE COULD BE RELIABLE

IF $T = 45$
$N = -12.9 + 7.22 \times 45 \simeq 312$
AS $T = 45$ IS "WAY ABOVE" THE HIGHEST VALUE OF T WHICH WAS USED TO CREATE THE REGRESSION LINE, THE ESTIMATE WOULD BE UNRELIABLE (EXTRAPOLATION)

Question 16 **(***)**

It is an **actual fact** that “sleeping with your clothes and shoes on is strongly correlated with waking up with a headache”.

Evidently the conclusion is that “sleeping with your clothes and shoes on causes a headache”.

Discuss the validity of the above conclusion indicating how a strong correlation is possible in the above scenario.

$\boxed{\checkmark}$, $\boxed{\text{explanation as appropriate}}$

THE CORRELATION MIGHT BE POSSIBLE BUT THE CONCLUSION NOT LIKELY TO BE CORRECT

CORRELATION $\not\Rightarrow$ CAUSE

THERE MAY BE 'ANOTHER VARIABLE' WHICH CONNECTS THE TWO

E.g "SLEEPING WITH YOUR CLOTHES ON"

$\Downarrow$

MAYBE YOU DRANK HEAVILY THE NIGHT BEFORE?
MAYBE YOU TOOK DRUGS/MEDICINES?

$\Downarrow$

AS A RESULT YOU WOKE UP WITH A HEADACHE

CORRELATION
&
REGRESSION
Part 2

Question 1 (**)

The table below shows the marks obtained by a group of students, in two separate tests.

Student	A	B	C	D	E	F	G	H
Test 1	27	38	17	29	41	42	32	9
Test 2	13	24	17	17	29	19	25	8

The first test is out of 50 marks while the second test is out of 30 marks.

Let x and y represent the marks obtained in Test 1 and Test 2, respectively.

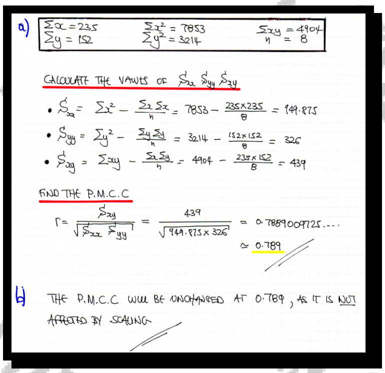
The following summary statistics are given.

$$\sum x = 235, \ \sum x^2 = 7853, \ \sum y = 152, \ \sum y^2 = 3214, \ \sum xy = 4904.$$

a) Find the value of the product moment correlation coefficient between x and y.

b) Explain how the value of the product moment correlation coefficient between x and y will be affected if the individual test marks were converted into percentage marks.

$\boxed{}$, $\boxed{r \approx 0.789}$, $\boxed{\text{unchanged}}$



a) $\sum x = 235$, $\sum y = 152$, $\sum x^2 = 7853$, $\sum y^2 = 3214$, $\sum xy = 4904$, $n = 8$

CALCULATE THE VALUES OF S_{xx} S_{yy} S_{xy}

- $S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = 7853 - \frac{235 \times 235}{8} = 949.875$
- $S_{yy} = \sum y^2 - \frac{\sum y \sum y}{n} = 3214 - \frac{152 \times 152}{8} = 326$
- $S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 4904 - \frac{235 \times 152}{8} = 439$

FIND THE P.M.C.C

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{439}{\sqrt{949.875 \times 326}} = 0.7889009725\ldots \simeq 0.789$$

b) THE P.M.C.C WILL BE UNCHANGED AT 0.789, AS IT IS NOT AFFECTED BY SCALING

Question 2 (**)

The table below shows the number of Maths teachers x, working in 8 different towns and the number of burglaries y, committed in a given month in the same 8 towns.

Town	A	B	C	D	E	F	G	H
x	37	40	21	50	32	27	39	40
y	30	28	20	35	34	27	31	26

a) Calculate the product moment correlation coefficient between the number of maths teachers and the number of burglaries.

b) Interpret the value of the product moment correlation coefficient in the context of this question.

c) Comment on the statement

"… the Maths teachers are likely to be responsible for the burglaries …"

$\boxed{}$, $\boxed{r \approx 0.692}$

a) OBTAIN SUMMARY STATISTICS FROM A CALCULATOR

$\Sigma x = 286$ $\quad \Sigma x^2 = 10784$ $\quad \Sigma xy = 8466$

$\Sigma y = 231$ $\quad \Sigma y^2 = 6831$ $\quad n = 8$

OBTAIN THE VALUES OF S_{xx}, S_{yy}, S_{xy}

- $S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 10784 - \frac{286 \times 286}{8} = 559.5$
- $S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 6831 - \frac{231 \times 231}{8} = 160.875$
- $S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 8466 - \frac{286 \times 231}{8} = 207.75$

CALCULATE THE P.M.C.C.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{207.75}{\sqrt{559.5 \times 160.875}} = 0.6924632\ldots \simeq 0.692$$

b) POSITIVE CORRELATION, I.E. THE HIGHER THE NUMBER OF TEACHERS (x) THE HIGHER THE NUMBER OF BURGLARIES (y), AND VISE VERSA

c) CORRELATION DOES NOT IMPLY CAUSATION — THEY MAY BE ANOTHER VARIABLE (OR VARIABLES) THAT x & y ARE BOTH CONNECTED TO

$\therefore$ STATEMENT IS NOT LIKELY TO BE VALID

Question 3 (**)

An electrical appliances supplier wishes to investigate the impact of advertising on the sales of his washing machines.

He records the number of monthly advertisements placed on the local radio station and the number of washing machines sold.

This is a table of his results.

Number of Advertisements (x)	52	37	66	45	77	27	80	19	47	40
Number of Washing Machines Sold (y)	80	75	81	76	77	49	84	50	63	64

Find, by detailed calculations, the value of the product moment correlation coefficient between x and y, and explain what conclusions the electrical appliances supplier should make from this value.

$\boxed{}$, $\boxed{r = 0.820}$

OBTAIN SUMMARY STATISTICS WITH A CALCULATOR

- $\sum x = 490$ • $\sum y = 699$ • $\sum xy = 36144$
- $\sum x^2 = 27682$ • $\sum y^2 = 50313$ • $n = 10$

FIND S_{xx}, S_{yy} AND S_{xy}

$$S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = 27682 - \frac{490 \times 490}{10} = 3672$$

$$S_{yy} = \sum y^2 - \frac{\sum y \sum y}{n} = 50313 - \frac{699 \times 699}{10} = 1452.9$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 36144 - \frac{490 \times 699}{10} = 1893$$

FINALLY WE HAVE THE P.M.C.C.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{1893}{\sqrt{3672 \times 1452.9}} = 0.81956\ldots \approx 0.820$$

AS r IS WELL ABOVE 0.5 AND CLOSE TO 1, THERE IS A GOOD SUGGESTION THAT PLACING ADS TO THE LOCAL RADIO STATION HAS THE DESIRED EFFECT

Question 4 (**+)

An electrical tester wishes to test the accuracy of a voltmeter used in a lab.

He uses a carefully calibrated voltage source and takes readings with the voltmeter he wishes to be tested.

This is a table of his results. x

Actual Voltage (x)	10	20	30	40	50	60	70	80	90	100
Voltmeter Reading (y)	9	19	34	39	54	61	68	80	92	99

a) Show, by detailed calculations, that the product moment correlation coefficient between x and y is approximately 1.

b) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

Full workings must be shown for this part of the question.

c) Calculate the residual for $x = 50$.

FS2-N, $\boxed{y \approx 0.667 + 0.997x}$

a) FROM CALCULATOR IN STAT MODE

$$\boxed{\begin{array}{lll} \Sigma x = 550 & \Sigma y = 555 & \Sigma xy = 38750 \\ \Sigma x^2 = 38500 & \Sigma y^2 = 39045 & n = 10 \end{array}}$$

- $S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 38500 - \frac{550 \times 550}{10} = 8250$
- $S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 39045 - \frac{555 \times 555}{10} = 8242.5$
- $S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 38750 - \frac{550 \times 555}{10} = 8225$

NOW THE P.M.C.C

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{8225}{\sqrt{8250 \times 8242.5}} = 0.997\ldots \simeq 1$$

AS REQUIRED

b) STARTING WITH THE GRADIENT

$$b = \frac{S_{xy}}{S_{xx}} = \frac{8225}{8250} = \frac{329}{330} = 0.99696\ldots$$

- $\bar{x} = \frac{\Sigma x}{n} = \frac{550}{10} = 55$ • $\bar{y} = \frac{\Sigma y}{n} = \frac{555}{10} = 55.5$

THE REGRESSION LINE IS GIVEN BY

$$y - \bar{y} = b(x - \bar{x})$$
$$y - \bar{y} = bx - b\bar{x}$$
$$y = bx + (\bar{y} - b\bar{x})$$
$$y = 0.997x + 0.667$$

COEFFICIENTS TO 3 s.f.

c) FIRSTLY ESTIMATE THE VALUE $x = 50$

$$y = 0.997 \times 50 + 0.667$$
$$y \approx 50.5$$

RESIDUAL = OBSERVED − PREDICTED
$= 54 - 50.5$
$= 3.5$

Question 5 (**+)

The table below shows the marks obtained by a group of students, in two separate tests.

Student	A	B	C	D	E	F	G	H	I	J
Test 1	17	11	16	9	12	12	11	4	7	15
Test 2	24	21	24	20	22	18	18	9	15	21

Let x and y represent the marks obtained in Test 1 and Test 2, respectively.

d) Find the value of S_{xx}, S_{yy} and S_{xy}.

e) Show that the product moment correlation coefficient between x and y is approximately 0.9.

f) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

$\boxed{}$, $\boxed{S_{xx} = 146.4}$, $\boxed{S_{yy} = 185.6}$, $\boxed{S_{xy} = 148.2}$, $\boxed{y = 7.660 + 1.012x}$

a) START BY GETTING SUMMARY STATISTICS FROM A STAT CALCULATOR

$\Sigma x = 114$ $\quad \Sigma x^2 = 1446$ $\quad \Sigma xy = 2337$

$\Sigma y = 192$ $\quad \Sigma y^2 = 3872$ $\quad n = 10$

OBTAIN THE VALUES OF S_{xx}, S_{yy} & S_{xy}

- $S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 1446 - \frac{114 \times 114}{10} = 146.4$
- $S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 3872 - \frac{192 \times 192}{10} = 185.6$
- $S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 2337 - \frac{114 \times 192}{10} = 148.2$

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

$r = \frac{148.2}{\sqrt{146.4 \times 185.6}} = 0.899060007\ldots \simeq 0.9$ APPROX

AS REQUIRED

c) CALCULATE ALL THE AUXILIARIES

- $\bar{x} = \frac{\Sigma x}{n} = \frac{114}{10} = 11.4$ & $\bar{y} = \frac{\Sigma y}{n} = \frac{192}{10} = 19.2$
- $b = \frac{S_{xy}}{S_{xx}} = \frac{148.2}{146.4} = 1.01229508\ldots \simeq 1.012$
- $a = \bar{y} - b\bar{x} = 19.2 - (1.01229\ldots)(11.4) = 7.659836\ldots \simeq 7.660$

$\therefore \; y = 7.660 + 1.012x$

Question 6 (**+)

The table below shows 10 pairs of bivariate data.

x	10	30	50	60	70	80	90	100	110	140
y	15	8	3	6	11	8	6	2	3	1

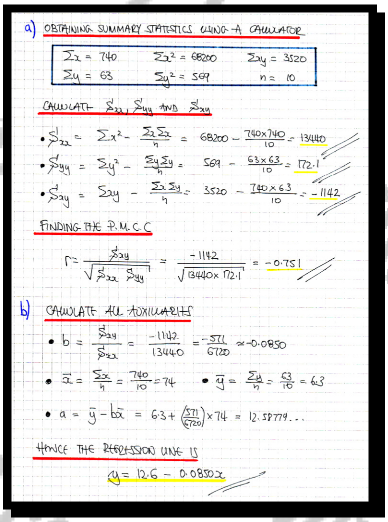
a) Determine the value of S_{xx}, S_{yy} and S_{xy}, and hence calculate the value of the product moment correlation coefficient between x and y.

b) Find the equation of the least squares regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

$\boxed{S_{xx} = 13440}$, $\boxed{S_{yy} = 172.1}$, $\boxed{S_{xy} = -1142}$, $\boxed{r = -0.751}$, $\boxed{y = 12.6 - 0.0850x}$



a) OBTAINING SUMMARY STATISTICS USING A CALCULATOR

$\Sigma x = 740 \quad \Sigma x^2 = 68200 \quad \Sigma xy = 3520$

$\Sigma y = 63 \quad \Sigma y^2 = 569 \quad n = 10$

CALCULATE S_{xx}, S_{yy} AND S_{xy}

- $S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 68200 - \frac{740 \times 740}{10} = 13440$
- $S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 569 - \frac{63 \times 63}{10} = 172.1$
- $S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 3520 - \frac{740 \times 63}{10} = -1142$

FINDING THE P.M.C.C

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-1142}{\sqrt{13440 \times 172.1}} = -0.751$$

b) CALCULATE ALL AUXILIARIES

- $b = \frac{S_{xy}}{S_{xx}} = \frac{-1142}{13440} = -\frac{571}{6720} \approx -0.0850$
- $\bar{x} = \frac{\Sigma x}{n} = \frac{740}{10} = 74$ • $\bar{y} = \frac{\Sigma y}{n} = \frac{63}{10} = 6.3$
- $a = \bar{y} - b\bar{x} = 6.3 + \left(\frac{571}{6720}\right) \times 74 = 12.58779\ldots$

HENCE THE REGRESSION LINE IS

$y = 12.6 - 0.0850x$

Question 7 (+)**

The table below shows the heights and weights of a random sample of 10 pupils, where the heights are given to the nearest cm and the weights to the nearest $5\,\text{kg}$.

Pupil	A	B	C	D	E	F	G	H	I	J
Height (cm)	148	164	156	172	147	184	162	155	182	165
Weight (kg)	40	60	55	75	40	80	65	50	80	70

Let x and y represent the respective heights and weights of these pupils and r the product moment correlation coefficient between x and y.

a) Determine the value of S_{xx}, S_{yy} and S_{xy}, and hence calculate the value of r, correct to three decimal places.

b) Interpret in context the value of r.

c) State the value of r if the heights were measured in metres instead of cm.

d) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = bx + a,$$

where a and b are constants.

$\boxed{}$, $\boxed{S_{xx} = 1480.5}$, $\boxed{S_{yy} = 2052.5}$, $\boxed{S_{xy} = 1677.5}$, $\boxed{r \approx 0.962}$

$\boxed{r \approx 0.962 \text{ regardless of units}}$, $\boxed{y = 1.13x - 124}$

a) OBTAINING SUMMARY STATISTICS FOR THE DATA USING A CALCULATOR

$$\boxed{\begin{array}{lll} \sum x^2 = 268803 & \sum x = 1635 & \sum xy = 102230 \\ \sum y^2 = 39875 & \sum y = 615 & n = 10 \end{array}}$$

CALCULATING S_{xx}, S_{yy}, S_{xy}

- $S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = 268803 - \frac{1635 \times 1635}{10} = 1480.5$
- $S_{yy} = \sum y^2 - \frac{\sum y \sum y}{n} = 39875 - \frac{615 \times 615}{10} = 2052.5$
- $S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 102230 - \frac{1635 \times 615}{10} = 1677.5$

FINALLY THE P.M.C.C. CAN BE FOUND

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{1677.5}{\sqrt{1480.5 \times 2052.5}} = 0.962$$

b) STRONG POSITIVE CORRELATION, I.E. THE TALLER THE PUPIL THE HEAVIER AND VICE VERSA

c) r WOULD BE UNCHANGED AT 0.962, AS THE P.M.C.C IS INDEPENDENT OF SCALE (UNITS)

d) PREPARE AUXILIARIES

- $b = \frac{S_{xy}}{S_{xx}} = \frac{1677.5}{1480.5} = \frac{3355}{2961} = 1.133$
- $\bar{x} = \frac{\sum x}{n} = \frac{1635}{10} = 163.5$
- $\bar{y} = \frac{\sum y}{n} = \frac{615}{10} = 61.5$
- $a = \bar{y} - b\bar{x} = 61.5 - \frac{3355}{2961} \times 163.5$

$\qquad = -123.7558257\ldots$

HENCE THE REGRESSION LINE IS GIVEN

$y = -124 + 1.13x$

$y = 1.13x - 124$

Question 8 (**+)

The table below shows the midday daily temperature x, in °C, and the number of cups of tea y, sold in a small café.

x	20	25	26	27	29	29	32	36
y	100	80	72	74	65	69	63	60

a) Find the value of S_{xx}, S_{yy} and S_{xy}, and hence calculate the product moment correlation coefficient between x and y.

b) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

c) Use the equation of the regression line to estimate the value of y when …

i. … $x = 40$.

ii. … $x = 50$.

Comment further on the reliability of these two estimates.

$\boxed{}$, $\boxed{S_{xx} = 160}$, $\boxed{S_{yy} = 1128.875}$, $\boxed{S_{xy} = -392}$, $\boxed{r = -0.922}$,
$\boxed{y = 141.475 - 2.45x}$, $\boxed{y_{40} \approx 43}$, $\boxed{y_{50} \approx 19}$

a) OBTAINING SUMMARY STATISTICS FROM A CALCULATOR

$$\boxed{\begin{array}{lll} \Sigma x = 224 & \Sigma x^2 = 6432 & \Sigma xy = 15932 \\ \Sigma y = 583 & \Sigma y^2 = 43615 & n = 8 \end{array}}$$

- $S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 6432 - \frac{224 \times 224}{8} = 160$
- $S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 43615 - \frac{583 \times 583}{8} = 1128.875$
- $S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 15932 - \frac{224 \times 583}{8} = -392$

THE P.M.C.C CAN NOW BE CALCULATED

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-392}{\sqrt{160 \times 1128.875}} = -0.922$$

b) PREPARING ALL THE AUXILIARIES

- $b = \frac{S_{xy}}{S_{xx}} = \frac{-392}{160} = -2.45$
- $\bar{x} = \frac{\Sigma x}{n} = \frac{224}{8} = 28$
- $\bar{y} = \frac{\Sigma y}{n} = 72.875$
- $a = \bar{y} - b\bar{x} = 72.875 - (-2.45)(28) = 141.475$

HENCE THE REGRESSION LINE IS GIVEN BY

$$y = 142 - 2.45x$$

c) I) WHEN $x = 40$

$y = 142 - 2.45 \times 40 \simeq 44$ (43 IS O.K TOO)

AS 40 IS JUST OUTSIDE THE RANGE OF TEMPERATURES WHICH WAS USED TO PRODUCE THE REGRESSION LINE AND r IS FAIRLY STRONG, THE ESTIMATE COULD BE RELIABLE

II) WHEN $x = 50$

$y = 142 - 2.45 \times 50 \approx 19-20$

AS 50 IS "WAY ABOVE" THE LARGEST VALUE OF x (TEMPERATURE) WHICH WAS USED TO CREATE THE REGRESSION LINE, THE ESTIMATE IS NOT LIKELY TO BE RELIABLE

Question 9 (***)

The table below shows the average midday temperature x of a seaside town, in °C, and the number of people y, that used a certain restaurant in that town.

x	17	20	25	29	27	21	20	24
y	40	42	42	43	44	39	41	45

a) Find the value of S_{xx}, S_{yy} and S_{xy}, and hence calculate the product moment correlation coefficient between x and y.

b) State the value of the product moment correlation coefficient between x and y if the temperature was measured in degrees Fahrenheit instead of Centigrade.

c) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

d) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

e) Interpret in the context of this question the physical meaning of b.

f) Use the equation of the regression line to estimate the value of y when …

i. … $x = 16$.

ii. … $x = 35$.

Comment further on the reliability of each of these two estimates.

$\boxed{\text{FS2L}}$, $\boxed{r \approx 0.670 \text{ regardless of units}}$, $\boxed{y = 34.4 + 0.331x}$, $\boxed{y_{16} \approx 40}$, $\boxed{y_{35} \approx 46}$

[solution overleaf]

a) USING THE CALCULATOR IN STAT MODE

$\Sigma x = 183$	$\Sigma x^2 = 4301$	$\Sigma xy = 7724$
$\Sigma y = 336$	$\Sigma y^2 = 14140$	$n = 8$

FIND THE VALUES OF S_{xx}, S_{yy}, S_{xy}

$$S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 4301 - \frac{183 \times 183}{8} = 114.875$$

$$S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 14140 - \frac{336 \times 336}{8} = 28$$

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 7724 - \frac{183 \times 336}{8} \; 38$$

CALCULATE THE P.M.C.C

$$r = \frac{S_{xy}}{S_{xx} S_{yy}} = \frac{38}{\sqrt{28 \times 114.875}} \approx 0.670$$

b) THE P.M.C.C WILL BE UNCHANGED, AS IT REMAINS UNAFFECTED BY SCALING OR CHOICE OF ORIGIN

c) COLLECT | CALCULATE ALL THE AUXILIARIES

- $b = \frac{S_{xy}}{S_{xx}} = \frac{38}{114.875} = \frac{304}{919} \approx 0.331$
- $\bar{x} = \frac{\Sigma x}{n} = \frac{183}{8} = 22.875$ • $\bar{y} = \frac{\Sigma y}{n} = \frac{336}{8} = 42$
- $a = \bar{y} - b\bar{x} = 42 - \left(\frac{304}{919}\right)(22.875) \approx 34.43307943\ldots$

$\therefore \; y = 34.4 + 0.331x$

d) THE EXPLANATORY VARIABLE (INDEPENDENT VARIABLE) IS THE TEMPERATURE x, AS IT IS THE TEMPERATURE THAT AFFECTS THE DINNERS & NOT THE OTHER WAY ROUND

y (NO OF DINNERS) IS CALLED THE RESPONSE VARIABLE

e) $b = 0.331$ IS THE GRADIENT OF LINE

IT REPRESENTS THE EXTRA NUMBER OF DINNERS PER DEGREE OF TEMPERATURE RISE

f) I) IF $x = 16$ $\quad y = 34.4 + 0.331 \times 16 \simeq 40$

ALTHOUGH THE P.M.C.C IS NOT VERY STRONG, 16 IS ONLY 1 DEGREE LESS THAN THE SMALLEST VALUE OF x WHICH WAS USED TO CREATE THE REGRESSION LINE, SO IT COULD BE RELIABLE

II) IF $x = 35$ $\quad y = 34.4 + 0.331 \times 35 \simeq 46$

GIVEN THAT THE P.M.C.C IS NOT STRONG, AND 35 IS "WAY ABOVE" THE HIGHEST TEMPERATURE WHICH WAS USED TO CREATE THE REGRESSION LINE, THE ESTIMATE WILL BE UNRELIABLE

Question 10 (***)

The table below shows the maximum temperature T °C on five different days and the corresponding ice cream sales, N, of a certain shop on those days.

T	15	20	25	30	35
N	79	145	182	255	302

a) Find the value of S_{TT}, S_{NN} and S_{TN}, and hence, determine the value of the product moment correlation coefficient between T and N.

b) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

c) Determine the equation of the regression line between N and T, giving the answer in the form

$$N = a + bT ,$$

where a and b are constants.

d) Interpret in the context of this question the physical meaning of b.

e) Use the equation of the regression line to estimate the value of N when …

i. … $T = 18$°C.

ii. … $T = 37$°C.

iii. … $T = 45$°C

Comment further on the reliability of each of these estimates.

$\boxed{\text{SZ-L}}$, $\boxed{S_{TT} = 250}$, $\boxed{S_{NN} = 31145.2}$, $\boxed{S_{TN} = 2780}$, $\boxed{r = 0.996}$,
$\boxed{N = 11.12T - 85.4}$, $\boxed{N_{18} \approx 115}$, $\boxed{N_{37} \approx 326}$, $\boxed{T_{45} \approx 415}$

[solution overleaf]

a) OBTAIN SUMMARY STATISTICS

$\sum T = 125$	$\sum T^2 = 3375$	$\sum TN = 26855$
$\sum N = 963$	$\sum T^2 = 216619$	$n = 5$

- $S_{TT} = \sum T^2 - \frac{\sum T \sum T}{n} = 3375 - \frac{125 \times 125}{5} = 250$
- $S_{NN} = \sum N^2 - \frac{\sum N \sum N}{n} = 216619 - \frac{963 \times 963}{5} = 31145.2$
- $S_{TN} = \sum TN - \frac{\sum T \sum N}{n} = 26855 - \frac{125 \times 963}{5} \; 2780$

CALCULATING THE P.M.C.C

$$r = \frac{S_{TN}}{\sqrt{S_{TT}\, S_{NN}}} = \frac{2780}{\sqrt{250 \times 31145.2}} \simeq 0.99627\ldots \simeq 0.996$$

b) T (TEMPERATURE) IS THE EXPLANATORY VARIABLE (INDEPENDENT) AS IT IS THE TEMPERATURE WHICH AFFECTS THE ICE CREAM SALES AND NOT THE OTHER WAY ROUND.

THE ICE CREAM SALES (N) IS THE RESPONSE VARIABLE.

c) OBTAIN ALL THE AUXILIARIES

- $b = \frac{S_{TN}}{S_{TT}} = \frac{2780}{250} = 11.12$
- $\bar{T} = \frac{\sum T}{n} = \frac{125}{5} = 25$
- $\bar{N} = \frac{\sum N}{n} = \frac{963}{5} = 192.6$
- $a = \text{"}\bar{y} - b\bar{x}\text{"} = \bar{N} - b\bar{T} = 192.6 - 11.12 \times 25 = -85.4$

$$\therefore N = 11.12T - 85.4$$

d) $b = 11.12$ IS THE GRADIENT

IT REPRESENTS THE NUMBER OF EXTRA ICE CREAMS TO BE SOLD, PER DEGREE RISE

e) I) IF $T = 18$, $N = 11.12 \times 18 - 85.4 \approx 115$

IT SHOULD BE RELIABLE AS THIS TEMPERATURE IS BETWEEN 15 & 35 DEGREES AND THE P.M.C.C IS VERY STRONG

IF $T = 37$, $N = 11.12 \times 37 - 85.4 \approx 326$

IT SHOULD BE RELIABLE AS 37 IS ONLY JUST ABOUT 35°C AND THE P.M.C.C IS VERY STRONG

IF $T = 45$, $N = 11.12 \times 45 - 85.4 \approx 415$

NOT LIKELY TO BE RELIABLY AS THIS TEMPERATURE IS WAY ABOUT 35° AND THERE IS NO EVIDENCE THAT THIS LINEAR PATTERN CONTINUES

Question 11 (***+)

The table below shows the marks obtained by a group of students, in two separate tests.

Student	A	B	C	D	E	F	G	H
Test 1	35	42	21	55	33	29	39	40
Test 2	30	28	21	38	35	27	30	k

Use linear regression for the test marks of the students A – G , to estimate the value of k , for student H.

Detailed workings are expected.

$\boxed{k \approx 31}$

OBTAIN SUMMARY STATISTICS FOR A–G

$\sum x = 254$ $\quad \sum x^2 = 9906$ $\quad n = 7$

$\sum y = 209$ $\quad \sum xy = 7865$

TEST 1 → x
TEST 2 → y

CALCULATE S_{xx} & S_{xy}

$$S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = 9906 - \frac{254 \times 254}{7} = \frac{4826}{7} \approx 689.43$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 7865 - \frac{254 \times 209}{7} = \frac{1969}{7} \approx 281.29$$

OBTAIN THE GRADIENT

$$b = \frac{S_{xy}}{S_{xx}} = \frac{1969/7}{4826/7} = 0.407998\ldots$$

OBTAIN THE EQUATION

$$\left.\begin{aligned} \bar{x} &= \frac{\sum x}{n} = \frac{254}{7} \\ \bar{y} &= \frac{\sum y}{n} = \frac{209}{7} \end{aligned}\right\} \Rightarrow a = \bar{y} - b\bar{x}$$

$$\Rightarrow a = \frac{209}{7} - 0.407998\ldots \times \frac{254}{7}$$

$$\Rightarrow a = \frac{286}{19} \approx 15.0526\ldots$$

$$\therefore \ y = 0.408x + 15.05$$

FINALLY USING THE ABOVE EQUATION WITH $x = 40$

$$y = 0.408 \times 40 + 15.05$$

$$y \approx 31.3726\ldots$$

$$\therefore k = 31$$

Question 12 (***+)

The table below shows the amount spent per month by a car dealership on marketing and advertising m, in £1000, and the number of cars c sold that month.

m	7	8	9	10	11
c	7	12	10	11	13

a) Find the value of the product moment correlation coefficient between m and c.

b) Determine the equation of the regression line between m and c, giving the answer in the form

$$c = a + bm,$$

where a and b are constants.

c) Use the equation of the regression line to estimate the number of cars that are expected to be sold in a month where the amount spent on marketing and advertising is …

i. … £8,800.

ii. … £20,000.

Comment further on the reliability of each of these two estimates.

d) Interpret in the context of this question the physical meaning of a and b.

$\boxed{}$, $\boxed{r = 0.755}$, $\boxed{c = 0.7 + 1.1m}$, $\boxed{c_{8.8} \approx 10}$, $\boxed{c_{20} \approx 23}$

a) START BY OBTAINING THE SUMMARY STATISTICS

- $\sum m = 45$ • $\sum m^2 = 415$ • $\sum mc = 488$
- $\sum c = 53$ • $\sum c^2 = 583$ • $n = 5$

CALCULATING S_{mm}, S_{cc} & S_{mc}

$S_{mm} = \sum m^2 - \frac{\sum m \sum m}{n} = 415 - \frac{45 \times 45}{5} = 10$

$S_{cc} = \sum c^2 - \frac{\sum c \sum c}{n} = 583 - \frac{53 \times 53}{5} = 21.2$

$S_{mc} = \sum mc - \frac{\sum m \sum c}{n} = 488 - \frac{45 \times 53}{5} = 11$

FIND THE P.M.C.C

$r = \frac{S_{mc}}{\sqrt{S_{mm} S_{cc}}} = \frac{11}{\sqrt{10 \times 21.2}} = 0.755$ 3 d.p.

b) FIND ALL AUXILIARIES FIRST

$\bar{m} = \frac{\sum m}{n} = \frac{45}{5} = 9$

$\bar{c} = \frac{\sum c}{n} = \frac{53}{5} = 10.6$

$b = \frac{S_{mc}}{S_{mm}} = \frac{11}{10} = 1.1$

$a = \bar{c} - b\bar{m} = 10.6 - 1.1(9) = 0.7$

$\therefore \; c = 0.7 + 1.1m$

c) I) $c_{8.8} = 0.7 + 1.1 \times 8.8 = 10.38$ I.E AROUND 10 CARS

AS THE CORRELATION COEFFICIENT IS FAIRLY HIGH, AND 8.8 (£8800) LIES WITHIN THE RANGE OF VALUES OF m WHICH WAS USED FOR THE REGRESSION LINE, THE ESTIMATE SHOULD BE RELIABLE (IT WOULD HAVE BEEN MORE RELIABLE IF MORE POINTS WERE USED)

(INTERPOLATION)

II) $c_{20} = 0.7 + 1.1 \times 20 = 22.7$ IE AROUND 23 CARS

AS $m = 20$, IS "WAY ABOVE" THE LARGEST VALUE OF m, WHICH WAS USED TO PRODUCE THE REGRESSION LINE, THE ESTIMATE COULD NOT BE RELIABLE

(EXTRAPOLATION)

d) • $a = 0.7$ ("y INTERCEPT")

NO OF CAR EXPECTED TO BE SOLD WHEN NO MONEY IS SPENT ON ADVERTISING

• $b = 1.1$ ("GRADIENT")

NO OF EXTRA CARS EXPECTED TO BE SOLD PER £1000 SPENT ON ADVERTISING

Question 13 (***+)

The table below shows the tomato yield obtained by a group of ten plants that were given different amounts of fertilizer and allowed to grow in otherwise identical conditions.

Plant	A	B	C	D	E	F	G	H	I	J
Amount of Fertilizer (grams)	0	10	20	30	40	50	60	70	80	90
Tomato Yield (kilograms)	1.2	1.9	2.1	2.4	2.5	2.7	3.0	k	3.2	3.1

a) Find an equation of the line of least squares using the plants A to G, I and J, and hence estimate the value of k, for the plant H.

Detailed workings are expected in this part

b) Interpret in context the gradient of the line of least squares.

c) Calculate the residual of plant J.

d) The residual of the plant A is -0.42. Find the …

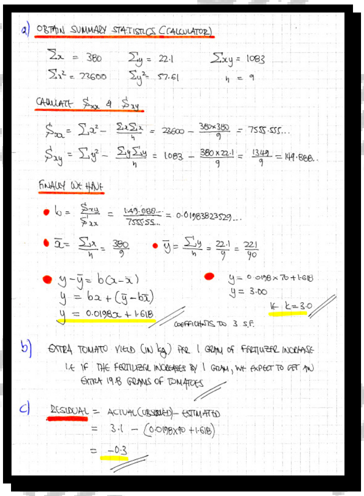
i. … sum of the residuals for the plants B to I.

ii. … mean of the residuals for the plants B to I.

Another plant N, not included in the table, was given 200 grams of fertilizer.

e) Discuss briefly, mathematically and in context, whether it is appropriate to use the line of least squares to predict its yield.

$\boxed{}$, $\boxed{Y = 0.0198F + 1.618}$, $\boxed{k = 3.0}$, $\boxed{R_H = -0.3}$, $\boxed{0.72}$, $\boxed{0.09}$


a) OBTAIN SUMMARY STATISTICS (CALCULATOR)

$\sum x = 380$ $\quad \sum y = 22.1$ $\quad \sum xy = 1083$

$\sum x^2 = 23600$ $\quad \sum y^2 = 57.61$ $\quad n = 9$

CALCULATE S_{xx} & S_{xy}

$S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = 23600 - \frac{380 \times 380}{9} = 7555.555\ldots$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 1083 - \frac{380 \times 22.1}{9} = \frac{1349}{9} = 149.888\ldots$

FINALLY WE HAVE

• $b = \frac{S_{xy}}{S_{xx}} = \frac{149.888\ldots}{7555.55\ldots} = 0.01983823529\ldots$

• $\bar{x} = \frac{\sum x}{n} = \frac{380}{9}$ • $\bar{y} = \frac{\sum y}{n} = \frac{22.1}{9} = \frac{221}{90}$

• $y - \bar{y} = b(x - \bar{x})$

$y = bx + (\bar{y} - b\bar{x})$

$y = 0.0198x + 1.618$

COEFFICIENTS TO 3 S.F.

• $y = 0.0198 \times 70 + 1.618$

$y = 3.00$

$k = 3.0$

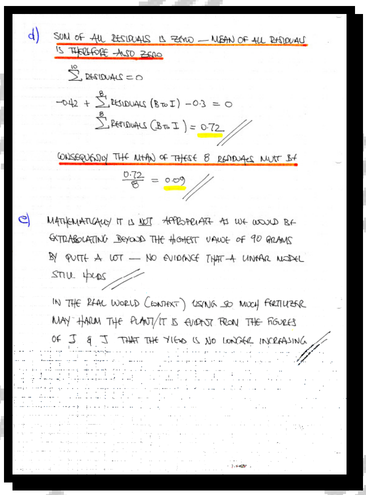
b) EXTRA TOMATO YIELD (IN kg) PER 1 GRAM OF FERTILIZER INCREASE

I.E. IF THE FERTILIZER INCREASES BY 1 GRAM, WE EXPECT TO GET AN EXTRA 19.8 GRAMS OF TOMATOES

c) RESIDUAL = ACTUAL (OBSERVED) − ESTIMATED

$= 3.1 - (0.0198 \times 90 + 1.618)$

$= -0.3$


d) SUM OF ALL RESIDUALS IS ZERO — MEAN OF ALL RESIDUALS IS THEREFORE ALSO ZERO

$\sum^{10} \text{RESIDUALS} = 0$

$-0.42 + \sum^{8} \text{RESIDUALS (B TO I)} - 0.3 = 0$

$\sum^{8} \text{RESIDUALS (B TO I)} = 0.72$

CONSEQUENTLY THE MEAN OF THESE 8 RESIDUALS MUST BE

$\frac{0.72}{8} = 0.09$

e) MATHEMATICALLY IT IS NOT APPROPRIATE AS WE WOULD BE EXTRAPOLATING BEYOND THE HIGHEST VALUE OF 90 GRAMS BY QUITE A LOT — NO EVIDENCE THAT A LINEAR MODEL STILL HOLDS

IN THE REAL WORLD (CONTEXT) USING SO MUCH FERTILIZER MAY HARM THE PLANT/IT IS EVIDENT FROM THE FIGURES OF I & J THAT THE YIELD IS NO LONGER INCREASING

Question 14 (***+)

The table below shows a set of bivariate data involving two variables x and y.

x	1003	1006	1012	1015	1021
y	0.0017	0.0027	0.0045	0.0056	0.0077

a) Use the coding equations

$$X = \frac{x-1012}{3} \quad \text{and} \quad Y = 10000y - 27$$

to find the value of S_{XX}, S_{YY} and S_{XY}.

b) Show that the product moment correlation coefficient between X and Y is approximately 0.9993.

c) State with justification the value of the product moment correlation coefficient between x and y.

d) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

$\boxed{}$, $\boxed{S_{XX} = 22.8}$, $\boxed{S_{YY} = 2251.2}$, $\boxed{S_{XY} = 226.4}$, $\boxed{r = 0.9993}$, $\boxed{y = 0.00033x - 0.33}$

a)

x	1003	1006	1012	1015	1021
y	0.0017	0.0027	0.0045	0.0056	0.0077

$X = \frac{x-1012}{3} \qquad Y = 10000y - 27$

REWRITE THE TABLE

X	−3	−2	0	1	3
Y	−10	0	18	29	50

OBTAIN SUMMARY STATISTICS

- $\sum X = -1$
- $\sum X^2 = 23$
- $\sum XY = 209$
- $\sum Y = 87$
- $\sum Y^2 = 3765$
- $n = 5$

FIND S_{XX}, S_{YY} & S_{XY}

- $S_{XX} = \sum X^2 - \frac{\sum X \sum X}{n} = 23 - \frac{(-1)^2}{5} = 22.8$
- $S_{YY} = \sum Y^2 - \frac{\sum Y \sum Y}{n} = 3765 - \frac{87^2}{5} = 2251.2$
- $S_{XY} = \sum XY - \frac{\sum X \sum Y}{n} = 209 - \frac{87(-1)}{5} = 226.4$

b) COMPUTE THE P.M.C.C

$r = \frac{S_{XY}}{S_{XX} S_{YY}} = \frac{226.4}{\sqrt{22.8 \times 2251.2}} = 0.9993139\ldots \approx 0.9993$

c) THE P.M.C.C WILL BE UNCHANGED AT 0.9993, AS THE P.M.C.C IS INDEPENDENT OF SCALING (MULTIPLICATION/DIVISION) AND CHOICE OF ORIGIN (ADDING/SUBTRACTING)

d) CALCULATE AUXILIARIES

$b = \frac{S_{XY}}{S_{XX}} = \frac{226.4}{22.8} = \frac{567}{57} \simeq 9.93$

$a = \bar{Y} - b\bar{X} = \left(\frac{87}{5}\right) - \frac{567}{57}\left(\frac{-1}{5}\right) = \frac{1105}{57} \simeq 19.385\ldots$

UNCODING FROM "CAPITALS" TO "LOWER CASE"

$Y = bX + a$

$10000y - 27 = \frac{567}{57}\left(\frac{x-1012}{3}\right) + \frac{1105}{57}$

$\times 171$

$1710000y - 4617 = 567(x-1012) + 3315$

$1710000y = 567x - 565872$

$y = 0.00033x - 0.331$

Question 15 (***+)

The table below shows a set of bivariate data involving two variables t and v.

t	151	154	157	163	169
v	8800	7800	7400	6500	3100

a) Use the coding equations

$$x = \frac{t-157}{3} \quad \text{and} \quad y = \frac{v}{100}$$

to find the value of S_{xx}, S_{yy} and S_{xy}.

b) Show that the product moment correlation coefficient between x and y is approximately -0.958.

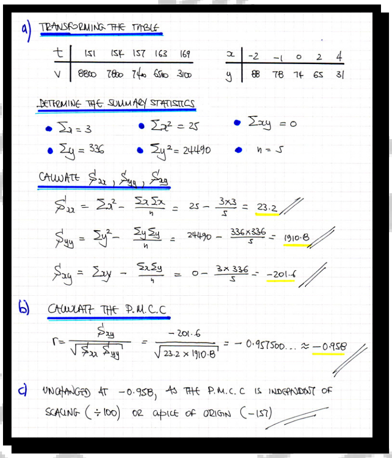
c) State with justification the value of the product moment correlation coefficient between t and v.

d) Determine the equation of the regression line between t and v, giving the answer in the form

$$v = A + Bt,$$

where A and B are constants.

$\boxed{S_{xx} = 23.2}$, $\boxed{S_{yy} = 1910.8}$, $\boxed{S_{xy} = -201.6}$, $\boxed{r = -0.958}$, $\boxed{v = 52717 - 290t}$



a) TRANSFORMING THE TABLE

t	151	154	157	163	169
v	8800	7800	7400	6500	3100

x	-2	-1	0	2	4
y	88	78	74	65	31

DETERMINE THE SUMMARY STATISTICS

- $\Sigma x = 3$
- $\Sigma x^2 = 25$
- $\Sigma xy = 0$
- $\Sigma y = 336$
- $\Sigma y^2 = 24490$
- $n = 5$

CALCULATE S_{xx}, S_{yy}, S_{xy}

$S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 25 - \frac{3 \times 3}{5} = 23.2$

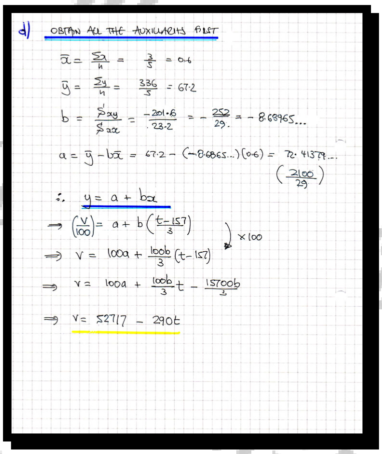
$S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 24490 - \frac{336 \times 336}{5} = 1910.8$

$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 0 - \frac{3 \times 336}{5} = -201.6$

b) CALCULATE THE P.M.C.C

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-201.6}{\sqrt{23.2 \times 1910.8}} = -0.957500\ldots \approx -0.958$

c) UNCHANGED AT -0.958, AS THE P.M.C.C IS INDEPENDENT OF SCALING $(\div 100)$ OR CHOICE OF ORIGIN (-157)



d) OBTAIN ALL THE AUXILIARIES FIRST

$\bar{x} = \frac{\Sigma x}{n} = \frac{3}{5} = 0.6$

$\bar{y} = \frac{\Sigma y}{n} = \frac{336}{5} = 67.2$

$b = \frac{S_{xy}}{S_{xx}} = \frac{-201.6}{23.2} = -\frac{252}{29} = -8.68965\ldots$

$a = \bar{y} - b\bar{x} = 67.2 - (-8.6865\ldots)(0.6) = 72.41379\ldots \left(\frac{2100}{29}\right)$

$\therefore \; y = a + bx$

$\Rightarrow \left(\frac{v}{100}\right) = a + b\left(\frac{t-157}{3}\right)$ $\times 100$

$\Rightarrow v = 100a + \frac{100b}{3}(t-157)$

$\Rightarrow v = 100a + \frac{100b}{3}t - \frac{15700b}{3}$

$\Rightarrow v = 52717 - 290t$

Question 16 (***+)

Clinical trials are carried out to determine the effect of a stimulant.

Ten volunteers were given different amounts of the stimulant, X milligrams, and the amount of their nightly sleep, Y hours, were recorded in the following night.

The following summary statistics were obtained.

$$\sum X = 900, \quad \sum Y = 78.4, \quad \sum X^2 = 114\,000, \quad \sum Y^2 = 616.18, \quad \sum XY = 6834$$

The following claims are made.

- Claim 1

  For every additional 60 milligrams of the stimulant, the nightly sleep typically reduces by 40 minutes.

- Claim 2

  The expected nightly sleep would have been 8 hours if no stimulant was taken.

Comment briefly on these two claims, fully supported by appropriate calculations.

, claims not justified supported by the regression line equation

TO INVESTIGATE THESE CLAIMS WE NEED THE REGRESSION LINE

- $S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = 114000 - \frac{900 \times 900}{10} = 33000$
- $S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 6834 - \frac{900 \times 78.4}{10} = -222$

TO INVESTIGATE THE FIRST CLAIM WE NEED THE GRADIENT

$b = \frac{S_{xy}}{S_{xx}} = \frac{-222}{33000} = -\frac{37}{5000} = -0.0067272\ldots$

REDUCTION IN HOURS PER 1 MILLIGRAM

$\Rightarrow \quad -0.00672\ldots \times 60 = -0.4036\ldots$

REDUCTION IN HOURS PER 60 mg

$= -0.4036\ldots \times 60$

$= -24.218\ldots$

CLAIM NOT JUSTIFIED AS EVERY ADDITIONAL 60 mg REDUCE THE NIGHTLY SLEEP BY APPROXIMATELY 24 MINUTES

NEXT FIND THE Y INTERCEPT

- $\bar{x} = \frac{\sum x}{n} = \frac{900}{10} = 90$
- $\bar{y} = \frac{\sum y}{n} = \frac{78.4}{10} = 7.84$

$a = \bar{y} - b\bar{x} = 7.84 - (-0.0067272\ldots) \times 90 = 7.84 + 90 \times 0.0067272\ldots$

$= \frac{929}{110} \simeq 8.45$ HOURS

CLAIM NOT JUSTIFIED AS $8.45 > 8$

Question 17 (***+)

Dolphins are thought to communicate with each other by high pitch noises they produce. The frequency, v kHz, of the noise made by a dolphin is recorded at 15 different sea depths, d m. These data are summarized below.

$$\sum d = 385.5, \ \sum d^2 = 11543.25, \ \sum v = 22.5, \ \sum v^2 = 38.25, \ \sum dv = 650.25$$

a) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

b) Find the value of S_{dd}, S_{vv} and S_{dv} for this data.

c) Calculate the product moment correlation coefficient between d and v.

d) Interpret the value of the product moment correlation coefficient in the context of this question.

e) Give a reason to support the fitting of a regression line of the form

$$v = a + bd,$$

where a and b are constants.

f) Determine the value of a and b, correct to three significant figures.

g) Interpret in the context of this question the physical meaning of a and b.

$\boxed{}$, $\boxed{S_{dd} = 1635.9}$, $\boxed{S_{vv} = 4.5}$, $\boxed{S_{dv} = 72}$, $\boxed{r \approx 0.839}$, $\boxed{a \approx 0.369}$, $\boxed{b \approx 0.0440}$

$\boxed{\begin{array}{lll} \sum d = 385.5 & \sum v = 22.5 & \sum dv = 650.25 \\ \sum d^2 = 11543.25 & \sum v^2 = 38.25 & n = 15 \end{array}}$

a) DEPTH IS THE EXPLANATORY VARIABLE, I.E INDEPENDENT VARIABLE AND FREQUENCY IS THE RESPONSE VARIABLE (DEPENDENT VARIABLE)
THIS IS BECAUSE IT IS THE DEPTH WHICH AFFECTS FREQUENCY AND NOT THE OTHER WAY ROUND

b) CALCULATE S_{dd} S_{vv} & S_{dv}

$S_{dd} = \sum d^2 - \frac{\sum d \sum d}{n} = 11543.25 - \frac{385.5 \times 385.5}{15} = 1635.9$

$S_{vv} = \sum v^2 - \frac{\sum v \sum v}{n} = 38.25 - \frac{22.5 \times 22.5}{15} = 4.5$

$S_{dv} = \sum dv - \frac{\sum d \sum v}{n} = 650.25 - \frac{385.5 \times 22.5}{15} = 72$

c) FIND THE PMCC. $r = \frac{S_{dv}}{\sqrt{S_{dd} S_{vv}}} = \frac{72}{\sqrt{1635.9 \times 4.5}} \simeq 0.839$

d) POSITIVE CORRELATION, I.E THE GREATER THE DEPTH, THE HIGHER THE FREQUENCY AND VISE VERSA

e) THE P.M.C.C IS REASONABLY HIGH TO SUGGEST A GOOD LINEAR MODEL MIGHT BE APPROPRIATE

f) $b = \frac{S_{dv}}{S_{dd}}$

$b = \frac{72}{1635.9}$

$d = \frac{240}{5453} \approx 0.0440$ 3 s.f.

"$a = \bar{y} - b\bar{x}$", HERE $a = \bar{v} - b\bar{d}$ ← $\frac{\sum d}{n} = \frac{385.5}{15} = 25.7$

↑ $\frac{\sum v}{n} = \frac{22.5}{15} = 1.5$

$a = 1.5 - 0.0440\ldots \times 25.7$

$a = 0.369$ 3 sf

g) $a =$ "y INTERCEPT"
THIS REPRESENTS THE "FREQUENCY OF DOLPHIN" WHEN IT IS AT THE SURFACE (ZERO DEPTH)

$b =$ "gradient"
INCREASE IN THE FREQUENCY PER METRE DEPTH, FOR EVERY METRE FOR WHICH THE DOLPHIN DIVES, THE FREQUENCY INCREASES BY 0.044 KHz

Question 18 (****)

The mean and variance of 10 independent observations of a random variable x, are 66.5 and 85.8, respectively.

Based on of a random sample of 10 independent observations of another variable y, the regression line of y on x is

$$y = 0.0949x - 0.0130.$$

Determine the product moment correlation coefficient between x and y. assuming further that $S_{yy} = 8.1$.

$\boxed{\text{FS2}}$, $\boxed{r = 0.977}$

WE NEED TO OBTAIN THE SUMS "BACKWARDS"

$\bar{x} = \frac{\sum x}{n}$

$66.5 = \frac{\sum x}{10}$

$\sum x = 665$

$\sigma_x^2 = \frac{\sum x^2}{n} - \bar{x}^2$

$85.8 = \frac{\sum x^2}{10} - 66.5^2$

$85.8 = \frac{\sum x^2}{10} - 4422.25$

$\sum x^2 = 45080.5$

HENCE WE HAVE

$$S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = 45080.5 - \frac{665 \times 665}{10} = 858$$

NOW FROM THE "GRADIENT" OF THE REGRESSION LINE

$$b = \frac{S_{xy}}{S_{xx}} \Rightarrow 0.0949 = \frac{S_{xy}}{858}$$

$$\Rightarrow S_{xy} = 81.4242$$

FINALLY THE P.M.C.C CAN BE FOUND

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{81.4242}{\sqrt{858 \times 8.1}} = 0.9767\ldots \approx 0.977$$

Question 19 (****)

A gym opened on the first day of January of a given year.

The months of that year were numbered as 1, 2, 3, … , 12.

The number of **new** members, N, at the end of each month m, was recorded for those 12 months.

The regression line of N on m was found to be

$$N = 34 + 35m.$$

Use the regression line to find the total number of members which joined that gym during that year.

No credit will be given for adding 69, 104, 139, … , 419, 454.

$\boxed{}$, $\boxed{3138}$

LOOKING AT THE REGRESSION LINE & NOTING THAT WE ARE REQUIRING ΣN

$$\Rightarrow N = \underset{a}{34} + \underset{b}{35}m$$

$$\Rightarrow a = \bar{N} - b\bar{m}$$

$$\Rightarrow 34 = \bar{N} - 35\bar{m}$$

$$\Rightarrow 34 = \frac{\Sigma N}{12} - 35\frac{\Sigma m}{12}$$

$$\Rightarrow 408 = \Sigma N - 35\Sigma m$$

NOW $\Sigma m = 1 + 2 + 3 + \dots + 12 = \frac{1}{2} \times 12 \times 13 = 78$

$$\Rightarrow 408 = \Sigma N - 35 \times 78$$

$$\Rightarrow 408 = \Sigma N - 2730$$

$$\Rightarrow \Sigma N = 3138$$

NOTE THAT WE GET THE SAME ALTHOUGH WE CANNOT GET CREDIT FOR

Jan $34 + 35 \times 1 = 69$
Feb $34 + 35 \times 2 = 104$
Mar $34 + 35 \times 3 = 139$
⋮
Dec $34 + 35 \times 12 = 454$

A.P WITH $a = 69$, $d = 35$, $L = 454$, $n = 12$

$\therefore S_n = \frac{n}{2}[a + L]$

$S_{12} = \frac{12}{2}(69 + 454)$

$S_{12} = 3138$

Question 20 (****+)

Some summary statistics for a set of bivariate data, based two variables x and y, are given below.

$$n = 10, \quad \bar{x} = 15, \quad \bar{y} = 48, \quad \sigma_x^2 = 186, \quad \sigma_y^2 = 172, \quad \sum xy = 8850.$$

a) Find the value of each of the following sums.

$$\sum x, \quad \sum y, \quad \sum x^2, \quad \sum y^2.$$

b) Calculate the product moment correlation coefficient between x and y.

c) Describe briefly the effect on the product moment correlation coefficient if another piece of data, $x = 10$ with $y = 70$, is added to the other 10 bivariate observations.

$\boxed{}$, $\boxed{\sum x = 150}$, $\boxed{\sum y = 480}$, $\boxed{\sum x^2 = 4110}$, $\boxed{\sum y^2 = 24760}$, $\boxed{r \approx 0.922}$

$\bar{x} = 15$ $\quad \sigma_x^2 = 186$ $\quad \Sigma xy = 8850$

$\bar{y} = 48$ $\quad \sigma_y^2 = 172$ $\quad n = 10$

a)
- $\bar{x} = \frac{\Sigma x}{n}$, $\quad 15 = \frac{\Sigma x}{10}$, $\quad \Sigma x = 150$
- $\sigma_x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$, $\quad 186 = \frac{\Sigma x^2}{10} - 15^2$, $\quad 1860 = \Sigma x^2 - 2250$, $\quad \Sigma x^2 = 4110$
- $\bar{y} = \frac{\Sigma y}{n}$, $\quad 48 = \frac{\Sigma y}{10}$, $\quad \Sigma y = 480$
- $\sigma_y^2 = \frac{\Sigma y^2}{n} - \bar{y}^2$, $\quad 172 = \frac{\Sigma y^2}{10} - 48^2$, $\quad 1720 = \Sigma y^2 - 23040$, $\quad \Sigma y^2 = 24760$

b) COMPUTE S_{xx}, S_{yy}, S_{xy}

$$S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 4110 - \frac{150 \times 150}{10} = 1860$$

$$S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 24760 - \frac{480 \times 480}{10} = 1720$$

$$S_{xy} = \sum xy - \frac{\Sigma x \Sigma y}{n} = 8850 - \frac{150 \times 480}{10} = 1650$$

THE P.M.C.C CAN NOW BE CALCULATED

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{1650}{\sqrt{1860 \times 1720}} = 0.922$$

c) THE P.M.M.C, BEING 0.922 SUGGEST FAIRLY STRONG POSITIVE CORRELATION

THE POINT $(10, 70)$ COMPARED WITH $(\bar{x}, \bar{y}) = (15, 48)$ DOES NOT FIT THE STRONG POSITIVE CORRELATION AS x IS GETTING SMALLER COMPARED WITH $\bar{x}$ BUT y IS GETTING LARGER COMPARED WITH $\bar{y}$.

HENCE THE P.M.M.C WILL DECREASE

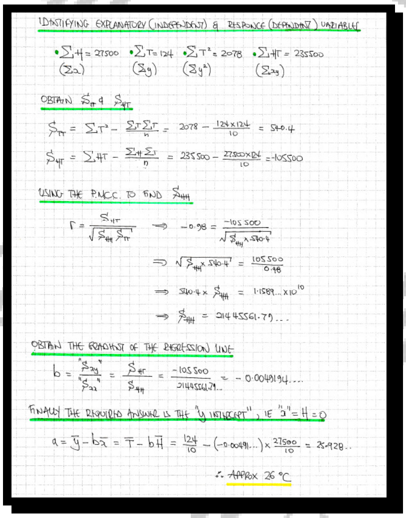
Question 21 (****+)

On a certain mountain climb, a scientist recorded the temperature, T °C, at ten different heights, H m above sea level, and some of his results are summarized below.

$$\sum T = 124, \quad \sum T^2 = 2078, \quad \sum H = 27\,500, \quad \sum HT = 235\,500$$

If the product moment correlation coefficient for this data is -0.98, determine an estimate for the temperature at sea level on the day of the climb.

$\boxed{}$, $\boxed{\approx 25.9\ °\text{C}}$


IDENTIFYING EXPLANATORY (INDEPENDENT) & RESPONCE (DEPENDENT) VARIABLES

• $\sum H = 27500$ $(\sum x)$ • $\sum T = 124$ $(\sum y)$ • $\sum T^2 = 2078$ $(\sum y^2)$ • $\sum HT = 235500$ $(\sum xy)$

OBTAIN S_{TT} & S_{HT}

$$S_{TT} = \sum T^2 - \frac{\sum T \sum T}{n} = 2078 - \frac{124 \times 124}{10} = 540.4$$

$$S_{HT} = \sum HT - \frac{\sum H \sum T}{n} = 235500 - \frac{27500 \times 124}{10} = -105500$$

USING THE P.M.C.C. TO FIND S_{HH}

$$r = \frac{S_{HT}}{\sqrt{S_{HH} S_{TT}}} \implies -0.98 = \frac{-105\,500}{\sqrt{S_{HH} \times 540.4}}$$

$$\implies \sqrt{S_{HH} \times 540.4} = \frac{105500}{0.98}$$

$$\implies 540.4 \times S_{HH} = 1.1589\ldots \times 10^{10}$$

$$\implies S_{HH} = 21445561.79\ldots$$

OBTAIN THE GRADIENT OF THE REGRESSION LINE

$$b = \frac{\text{"}S_{xy}\text{"}}{\text{"}S_{xx}\text{"}} = \frac{S_{HT}}{S_{HH}} = \frac{-105\,500}{21445561.79\ldots} = -0.0049194\ldots$$

FINALLY THE REQUIRED ANSWER IS THE "y INTERCEPT", IE "x" = H = 0

$$a = \bar{y} - b\bar{x} = \bar{T} - b\bar{H} = \frac{124}{10} - (-0.00491\ldots) \times \frac{27500}{10} = 25.928\ldots$$

$\therefore$ APPROX 26 °C

Question 22 (****+)

The number of letters x in people's first names and number of letters y in people's surnames is researched.

The summary data of the number of letters in the first names and the surnames of a random sample of 20 individuals is shown below.

$$\sum x = 125, \quad \sum x^2 = 796, \quad \sum y = 140, \quad \sum y^2 = 1032, \quad \sum xy = 882$$

a) Calculate the product moment correlation coefficient between x and y.

The name "Richard Edwards" is added to the sample, making the total number of people in the sample, 21.

b) Without a direct recalculation , ...

i. ... show that S_{xx} of the 21 first names is likely to have a different value to the original value of S_{xx} of the original 20 first names.

ii. ... determine the effect of adding "Richard Edwards" to S_{yy} and S_{xy}.

c) Given further that adding "Richard Edwards" increases the value of S_{xx} explain with justification whether the product moment correlation coefficient between x and y, increases or decreases.

$\boxed{}$, $\boxed{r \approx 0.253}$

a) $\sum x = 125$, $\sum y = 140$, $\sum xy = 882$
$\sum x^2 = 796$, $\sum y^2 = 1032$, $n = 20$

- $S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = 796 - \frac{125^2}{20} = 14.75$
- $S_{yy} = \sum y^2 - \frac{\sum y \sum y}{n} = 1032 - \frac{140^2}{20} = 52$
- $S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 882 - \frac{125 \times 140}{20} = 7$

$r = \frac{7}{\sqrt{14.75 \times 52}} = 0.253$

b) LOOKING AT THE NAME

RICHARD = 7 LETTERS
EDWARDS = 7 LETTERS

- $S_{xx} = \sum x^2 - \frac{\sum x \sum x}{n} = n\left[\frac{\sum x^2}{n} - \frac{\sum x}{n}\frac{\sum x}{n}\right]$
  $= n\left[\frac{\sum x^2}{n} - \bar{x}^2\right] = n\sigma^2$
  $= \cancel{n} \times \frac{\sum (x-\bar{x})^2}{\cancel{n}} = \sum (x-\bar{x})^2$
- $S_{yy} = \sum (y-\bar{y})^2$
- $S_{xy} = \sum (x-\bar{x})(y-\bar{y})$

NOW LOOKING AT THE MEANS BEFORE ADDING THE EXTRA NAME

$\bar{x}_{20} = \frac{\sum x}{n} = \frac{125}{20} = 6.25$ $\quad$ $\bar{x}_{21} = \frac{132}{2} = 6.2857\ldots$

$\bar{y}_{20} = \frac{\sum y}{n} = \frac{140}{20} = 7$ $\quad$ $\bar{y}_{21} = \frac{147}{21} = 7$

ADDING "RICHARD" INTO THE x's & EDWARDS TO THE ys

$S_{xx} = \sum_{i=1}^{21} (x-\bar{x})^2 = \left[\sum_{i=1}^{20} (x-6.28)^2\right] + (7-6.2857.)^2 =$ CHANGES

$S_{yy} = \sum_{i=1}^{21} (y-\bar{y}) = \left[\sum_{i=1}^{20} (y-\bar{y})^2\right] + \cancel{(7-7)^2} =$ UNCHANGED AT 14.75

$S_{xy} = \sum_{i=1}^{21} (x-\bar{x})(y-\bar{y}) = \left[\sum_{i=1}^{20} (x-6.28)^2 (y-7)^2\right] + \cancel{(7-0)(7-6.285..)}$

= UNCHANGED AT 7

c) IF S_{xx} INCREASES THEN

$r = \frac{S_{xy} \text{ ← UNCHANGED}}{\sqrt{S_{xx} S_{yy}}}$

CHANGES (INCREASE) $\quad$ UNCHANGED

$\therefore$ r WILL DECREASE

Question 23 (****+)

Two variables, x and y, have the following regression equations, based on 5 observations.

$$y \text{ on } x : \ y = 18.5 + 0.1x$$

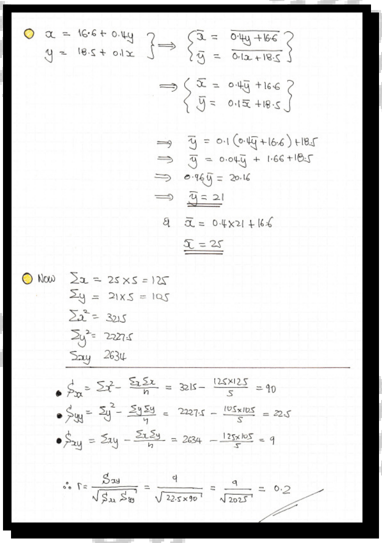
$$x \text{ on } y : \ x = 16.6 + 0.4y$$

The following summary statistics are also given.

$$\sum x^2 = 3215, \quad \sum y^2 = 2227.5, \quad \sum xy = 2634$$

Show that the product moment correlation coefficient between x and y is 0.2.

proof



$$\left.\begin{array}{l} x = 16.6 + 0.4y \\ y = 18.5 + 0.1x \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \bar{x} = \overline{0.4y + 16.6} \\ \bar{y} = \overline{0.1x + 18.5} \end{array}\right\}$$

$$\Rightarrow \left\{\begin{array}{l} \bar{x} = 0.4\bar{y} + 16.6 \\ \bar{y} = 0.1\bar{x} + 18.5 \end{array}\right\}$$

$$\Rightarrow \bar{y} = 0.1(0.4\bar{y} + 16.6) + 18.5$$

$$\Rightarrow \bar{y} = 0.04\bar{y} + 1.66 + 18.5$$

$$\Rightarrow 0.96\bar{y} = 20.16$$

$$\Rightarrow \bar{y} = 21$$

& $\bar{x} = 0.4 \times 21 + 16.6$

$$\bar{x} = 25$$

Now $\Sigma x = 25 \times 5 = 125$

$\Sigma y = 21 \times 5 = 105$

$\Sigma x^2 = 3215$

$\Sigma y^2 = 2227.5$

$\Sigma xy \ 2634$

- $S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 3215 - \frac{125 \times 125}{5} = 90$
- $S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 2227.5 - \frac{105 \times 105}{5} = 22.5$
- $S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 2634 - \frac{125 \times 105}{5} = 9$

$$\therefore r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{9}{\sqrt{22.5 \times 90}} = \frac{9}{\sqrt{2025}} = 0.2$$

SPEARMAN'S RANK

Question 1 **(**)**

Nine gymnasts performed in a gymnastics competition.

Their names were Arnold (A), Brian (B), Christian (C), Damon (D), Eli (E), Fabian (F), Gordon (G), Harry (H) and Ian (I).

Rank	1	2	3	4	5	6	7	8	9
Judge 1	D	C	E	B	F	A	I	H	G
Judge 2	D	E	F	C	I	B	A	G	H

a) Calculate Spearman's rank correlation coefficient for this data.

b) Test whether or not the judges are generally in agreement, at the 1% level of significance, stating your hypotheses clearly.

$\boxed{r_s = \frac{5}{6} \approx 0.833}$, $\boxed{\text{evidence of agreement, } 0.8333 > 0.7833}$

a) REWRITE THE TABLE IN MORE USER FRIENDLY FORM

GYMNAST	A	B	C	D	E	F	G	H	I
JUDGE 1 RANK	6	4	2	1	3	5	9	8	7
JUDGE 2 RANK	7	6	4	1	2	3	8	9	5
d^2	1	4	4	0	1	4	1	1	4

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 20}{9 \times 80} = \frac{5}{6} = 0.8333$$

b) $H_0 : \rho_s = 0$ (JUDGES ARE NOT IN GENERAL AGREEMENT)

$H_1 : \rho_s > 0$ (JUDGES ARE IN GENERAL AGREEMENT)

THE CRITICAL VALUE FOR $n=9$, AT 1% SIGNIFICANCE IS 0.7833

AS $0.8333 > 0.7833$, THERE IS EVIDENCE THAT THE JUDGES ARE IN GENERAL AGREEMENT — REJECT H_0

Question 2 (**)

The data in the table below shows the time, in seconds, for the fastest qualifying lap for 8 different Formula One racing drivers, and their finishing order in the actual race.

Fastest Qualifying Lap	49.12	49.34	49.07	48.55	49.40	49.27	49.77	48.87
Finishing Position	5	6	1	3	7	4	8	2

a) Calculate Spearman's rank correlation coefficient for this data.

b) Test whether or not there is any association between the fastest qualifying lap time and the finishing position for Formula One racing drivers, at the 5% level of significance, stating your hypotheses clearly.

$\boxed{r_s = \frac{37}{42} \approx 0.8810}$, $\boxed{\text{evidence of association, } 0.8810 > 0.7381}$

a) START BY REWRITING THE TABLE IN RANKS

LAP TIME RANK	4	6	3	1	7	5	8	2
FINISH ORDER RANK	5	6	1	3	7	4	8	2
d^2	1	0	4	4	0	1	0	0

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 10}{8 \times 63} = 1 - \frac{5}{42} = \frac{37}{42} \simeq 0.8810$$

b)
$H_0 : \rho_s = 0$ (NO ASSOCIATION, "POSITIVE" OR "NEGATIVE")
$H_1 : \rho_s \neq 0$ (ASSOCIATION EXISTS)

THE CRITICAL VALUE FOR $n=8$, AT 5%, TWO TAILED, IS ± 0.7381

AS $0.8810 > 0.7381$, THERE IS SIGNIFICANT EVIDENCE OF (POSITIVE) ASSOCIATION BETWEEN THE FASTEST QUALIFYING LAP TIME AND THE RACE FINISHING POSITION. — REJECT H_0

Question 3 (**)

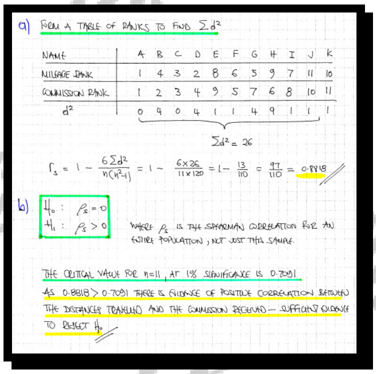
The table below shows the mileages travelled by eleven salesmen and the commission they got paid during a given month.

Name	Monthly mileage	Monthly commission
Alan	734	£800
Brian	650	£660
Christian	668	£620
Dominic	709	£610
Ethan	437	£450
Finlay	551	£560
Graham	580	£510
Hamish	387	£520
Ian	450	£460
James	298	£430
Kevin	325	£390

c) Calculate Spearman's rank correlation coefficient for this data.

d) Test whether or not there is evidence of positive correlation between the mileages travelled and the amount of commission received, at the 1% level of significance, stating your hypotheses clearly.

$\boxed{}$, $\boxed{r_s = \frac{97}{110} \approx 0.8818}$, $\boxed{\text{positive correlation, } 0.8818 > 0.7091}$



a) FORM A TABLE OF RANKS TO FIND $\sum d^2$

NAME	A	B	C	D	E	F	G	H	I	J	K
MILEAGE RANK	1	4	3	2	8	6	5	9	7	11	10
COMMISSION RANK	1	2	3	4	9	5	7	6	8	10	11
d^2	0	4	0	4	1	1	4	9	1	1	1

$\sum d^2 = 26$

$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 26}{11 \times 120} = 1 - \frac{13}{110} = \frac{97}{110} = 0.8818$

b) $\boxed{\begin{array}{l} H_0 : \rho_s = 0 \\ H_1 : \rho_s > 0 \end{array}}$ WHERE ρ_s IS THE SPEARMAN CORRELATION FOR AN ENTIRE POPULATION, NOT JUST THIS SAMPLE.

THE CRITICAL VALUE FOR $n=11$, AT 1% SIGNIFICANCE IS 0.7091

AS $0.8818 > 0.7091$ THERE IS EVIDENCE OF POSITIVE CORRELATION BETWEEN THE DISTANCES TRAVELLED AND THE COMMISSION RECEIVED — SUFFICIENT EVIDENCE TO REJECT H_0

Question 4 (**)

The actual ages, in complete years, of seven cats is shown below.

Cat Name	Riri	Loulou	Ginge	Puss	Ollie	Rex	Mog
Age in years	3	4	18	21	5	11	9

These seven cats were seen by a vet, during a day's surgery, and the vet was asked to order them according to their age by examination only.

He ordered the cats' ages, older first, as follows.

Ginge, Puss, Mog, Rex, Loulou, Riri, Ollie.

c) Calculate Spearman's rank correlation coefficient between the actual age of the cats and the vet's order.

d) Test whether or not the vet has the ability to identify the age of cats, at the 1% level of significance, stating your hypotheses clearly.

$\boxed{}$, $\boxed{r_s = \frac{23}{28} \approx 0.8214}$, $\boxed{\text{no evidence of association, } 0.8214 < 0.8929}$

a) FORM A TABLE OF RANKS TO FIND $\sum d^2$

NAME	GINGE	PUSS	MOG	REX	LOULOU	RIRI	OLLIE
VET'S RANK	1	2	3	4	5	6	7
"ACTUAL" RANK	2	1	4	3	6	7	5
d^2	1	1	1	1	1	1	4

$\sum d^2 = 10$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 10}{7 \times 48} = \frac{23}{28} = 0.8214$$

b) SETTING HYPOTHESES, WHERE ρ_s IS THE SPEARMAN CORRELATION FOR AN ENTIRE POPULATION, NOT OF THESE 7 CATS

$H_0 : \rho_s = 0$
$H_1 : \rho_s > 0$

THE CRITICAL VALUE FOR $n=7$ AT 1% SIGNIFICANCE IS 0.8929

AS $0.8214 < 0.8929$ IT APPEARS THAT THE VET DOES NOT HAVE THE ABILITY TO IDENTIFY THE AGES OF CATS – SUFFICIENT EVIDENCE TO REJECT H_0

Question 5 (**+)

Six ordered pairs (x, y), of bivariate data, are shown in the following set of axes.

Determine the Spearman's rank correlation coefficient for this data.

$\boxed{}$, $\boxed{r_s = \dfrac{31}{35} \approx 0.886}$

LABELLING THE POINTS AS "A – F" FROM LEFT TO RIGHT

POINT	A	B	C	D	E	F
Rank in x	6	5	4	3	2	1
Rank in y	6	5	3	4	1	2
d^2	0	0	1	1	1	1

USING THE STANDARD FORMULA WITH $\Sigma d^2 = 4$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{\cancel{6} \times 4}{\cancel{6} \times 35} = 1 - \frac{4}{35} = \frac{31}{35} \approx 0.886$$

Question 6 (**+)

The table below shows, for a group of students in a recent mock exam, the number of marks lost, y, and the corresponding number of papers, x, they practiced leading up to that exam.

Student	A	B	C	D	E	F	G	H	I	J
Number of Papers (x)	17	39	24	26	11	22	25	10	8	6
Number of Marks Lost (y)	12	5	11	14	10	9	8	15	19	17

a) Find the value of S_{xx}, S_{yy} and S_{xy}, and hence determine the value of the product moment correlation coefficient between x and y.

b) Comment briefly on the result of part (**a**).

c) Obtain the Spearman's rank correlation coefficient between x and y.

d) Test, at the 1% level of significance, whether there is evidence of negative association between the ranks of x and y.

$\boxed{}$, $\boxed{S_{xx} = 957.6}$, $\boxed{S_{yy} = 166}$, $\boxed{S_{xy} = -317}$, $\boxed{r = -0.795}$, $\boxed{r_s = -0.745}$

STUDENT	A	B	C	D	E	F	G	H	I	J
NO OF PAPERS (x)	17	39	24	26	11	22	25	10	8	6
NO OF MARKS LOST (y)	12	5	11	14	10	9	8	15	19	17

$\Sigma x = 188$, $\Sigma y = 120$, $\Sigma x^2 = 4492$, $\Sigma y^2 = 1606$, $\Sigma xy = 1939$

a) OBTAIN S_{xx}, S_{xy}, S_{yy}

$S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 4492 - \frac{188 \times 188}{10} = 957.6$

$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 1939 - \frac{188 \times 120}{10} = -317$

$S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 1606 - \frac{120 \times 120}{10} = 166$

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-317}{\sqrt{957.6 \times 166}} = -0.795$

b) NEGATIVE CORRELATION, IE AS THE NUMBER OF PAPERS PRACTICED INCREASES THE NUMBER OF MARKS LOST DECREASES & VICE VERSA

c) RANKING THE DATA

STUDENT	A	B	C	D	E	F	G	H	I	J
"PAPER" RANK	6	1	4	2	7	5	3	8	9	10
"ERROR" RANK	5	10	6	4	7	8	9	3	1	2
d^2	1	81	4	4	0	9	36	25	64	64

$\Sigma d^2 = 288$

$r_s = 1 - \frac{6 \Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 288}{10 \times 99} = -0.745$

d) SETTING THE HYPOTHESES

H_0 : THERE IS NO ASSOCIATION BETWEEN THE RANKS, $\rho_s = 0$

H_1 : THERE IS NEGATIVE ASSOCIATION BETWEEN THE RANKS, IE $\rho_s < 0$

THE CRITICAL VALUE AT 1%, FOR $n=10$, IS -0.7455

STRICTLY SPEAKING AS $-0.7455 < -0.7454545...$ THERE IS NO SIGNIFICANT EVIDENCE OF NEGATIVE ASSOCIATION BETWEEN THE RANKS (AT 1% SIGNIFICANCE)

HOWEVER AS THE r_s IS SO CLOSE TO THE CRITICAL VALUE RETESTING WITH A LARGER SAMPLE IS ADVISABLE OR AN ALTERNATIVE TEST USING THE P.M.C.C