PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE19CS203 – STATISTICS FOR DATA SCIENCE

Unit-5 - Power of Test and Simple Linear Regression

QUESTION BANK

Simple Linear Regression: Introduction to correlation

Exercises for section 7.1: [Text Book Exercise 7.1– Pg. No. [520 – 523]]

1. Compute the correlation coefficient for the following data set

X	1	2	3	4	5	6	7
y	2	1	4	3	7	5	6

2. For each of the following data sets, explain why the correlation coefficient is the same as for the data set in Exercise 1.

(a)

x	1	2	3	4	5	6	7
y	5	4	7	6	10	8	9

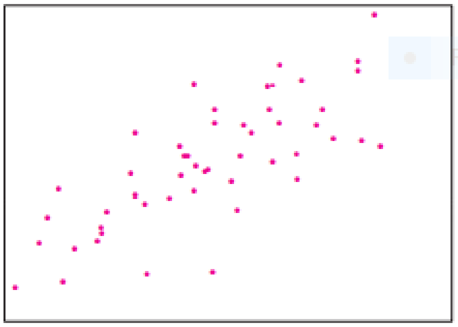
(b)

x	11	21	31	41	51	61	71
y	5	4	7	6	10	8	9

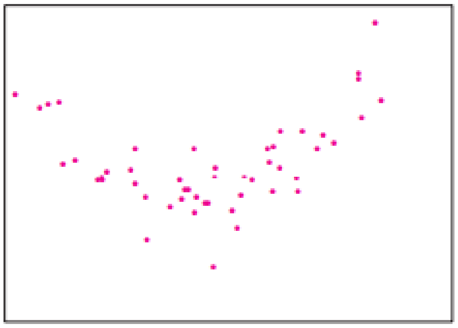
(c)

x	53	43	73	63	103	83	93
y	4	6	8	10	12	14	16

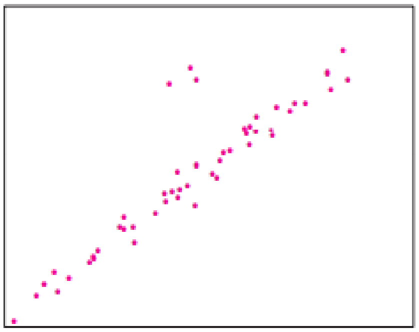
3. For each of the following scatterplots, state whether the correlation coefficient is an appropriate summary, and explain briefly.



(*a*)



(*b*)



(*c*)

4. True or false, and explain briefly:

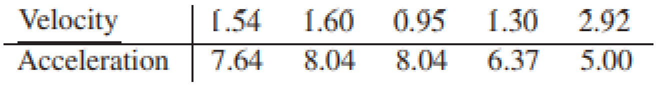
   a. If the correlation coefficient is positive, then above-average values of one variable are associated with above-average values of the other.

   b. If the correlation coefficient is negative, then below-average values of one variable are associated with below-average values of the other.

c. If y is usually less than x, then the correlation between y and x will be negative.

5. An investigator collected data on heights and weights of college students. The correlation between height and weight for men was about 0.6, and for women it was about the same. If men and women are taken together, will the correlation between height and weight be more than 0.6, less than 0.6, or about equal to 0.6? It might be helpful to make a rough scatterplot.

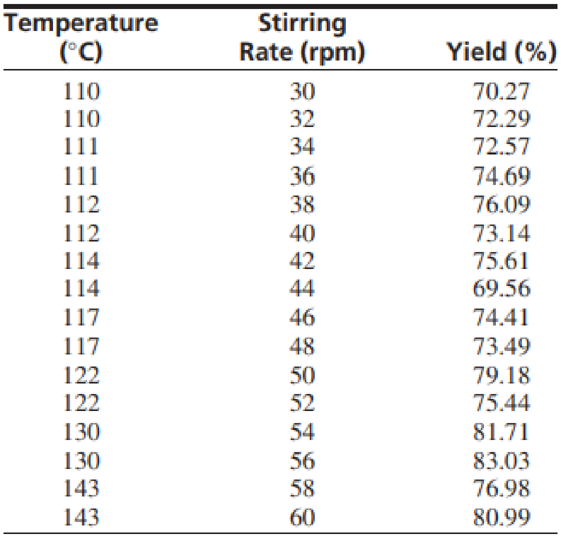
6. In a study of ground motion caused by earthquakes, the peak velocity (in m/s) and peak acceleration (in m/s^2) were recorded for five earthquakes. The results are presented in the following table.



Velocity	1.54	1.60	0.95	1.30	2.92
Acceleration	7.64	8.04	8.04	6.37	5.00

a. Compute the correlation coefficient between peak velocity and peak acceleration.
b. Construct a scatterplot for these data.
c. Is the correlation coefficient an appropriate summary for these data? Explain why or why not.
d. Someone suggests converting the units from meters to centimeters and from seconds to minutes. What effect would this have on the correlation?

7. A chemical engineer is studying the effect of temperature and stirring rate on the yield of a certain product. The process is run 16 times, at the settings indicated in the following table. The units for yield are percent of a theoretical maximum.



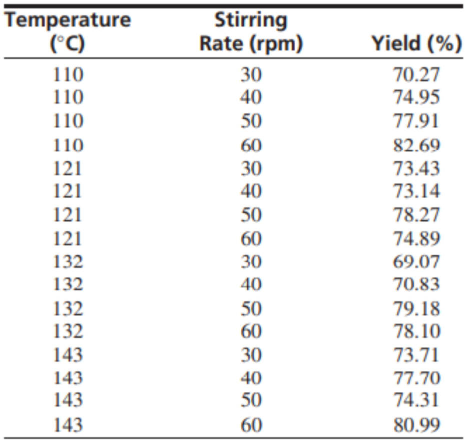
Temperature (°C)	Stirring Rate (rpm)	Yield (%)
110	30	70.27
110	32	72.29
111	34	72.57
111	36	74.69
112	38	76.09
112	40	73.14
114	42	75.61
114	44	69.56
117	46	74.41
117	48	73.49
122	50	79.18
122	52	75.44
130	54	81.71
130	56	83.03
143	58	76.98
143	60	80.99

a. Compute the correlation between temperature and yield, between stirring rate and yield, and between temperature and stirring rate.

b. Do these data provide good evidence that the yield is unaffected by temperature, within the range of the data? Or might the result be due to confounding? Explain.

c. Do these data provide good evidence that increasing the stirring rate causes the yield to increase, within the range of the data? Or might the result be due to confounding? Explain.

8. Another chemical engineer is studying the same process as in Exercise 7, and uses the following experimental matrix.



Temperature (°C)	Stirring Rate (rpm)	Yield (%)
110	30	70.27
110	40	74.95
110	50	77.91
110	60	82.69
121	30	73.43
121	40	73.14
121	50	78.27
121	60	74.89
132	30	69.07
132	40	70.83
132	50	79.18
132	60	78.10
143	30	73.71
143	40	77.70
143	50	74.31
143	60	80.99

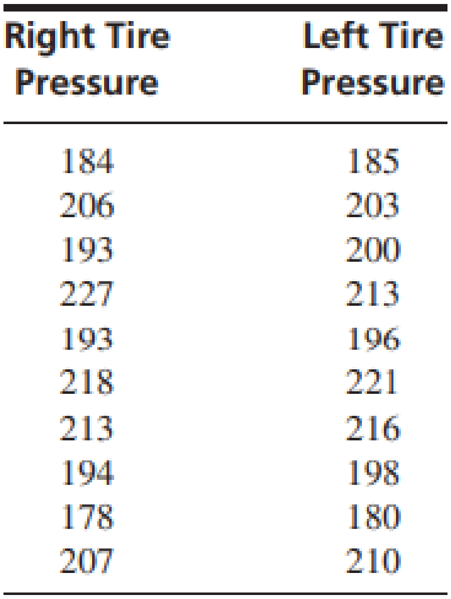
a. Compute the correlation between temperature and yield, between stirring rate and yield, and between temperature and stirring rate.

b. Do these data provide good evidence that the yield is unaffected by temperature, within the range of the data? Or might the result be due to confounding? Explain.

c. Do these data provide good evidence that increasing the stirring rate causes the yield to increase, within the range of the data? Or might the result be due to confounding? Explain.

d. Which experimental design is better, this one or the one in Exercise 7? Explain.

9. Tire pressure (in kPa) was measured for the right and left front tires on a sample of 10 automobiles. Assume that the tire pressures follow a bivariate normal distribution.



Right Tire Pressure	Left Tire Pressure
184	185
206	203
193	200
227	213
193	196
218	221
213	216
194	198
178	180
207	210

a. Find a 95% confidence interval for ρ, the population correlation between the pressure in the right tire and the pressure in the left tire.

b. Can you conclude that $\rho > 0.9$?

c. Can you conclude that $\rho > 0$?

10. In a sample of 300 steel rods, the correlation coefficient between diameter and length was r = 0.15.

a. Find the P-value for testing $H_0 : \rho \leq 0$ vs. $H_1 : \rho > 0$. Can you conclude that $\rho > 0$?

b. Does the result in part (a) allow you to conclude that there is a strong correlation between eccentricity and smoothness? Explain.

11. The article “Drift in Posturography Systems Equipped With a Piezoelectric Force Platform: Analysis and Numerical Compensation” (L. Quagliarella, N. Sasanelli, and V. Monaco, IEEE Transactions on Instrumentation and Measurement, 2008: 997–1004), reported the results of an experiment to determine the effect of load on the drift in signals derived from a piezoelectric force plates. The correlation coefficient y between output

and time under a load of 588 N was −0.9515. Measurements were taken 100 times per second for 300 seconds, for a total of 30,000 measurements. Find a 95% confidence interval for the population correlation ρ.

12. Phonics is an instructional method in which children are taught to connect sounds with letters or groups of letters. The article "Predictive Accuracy of Nonsense Word Fluency for English Language Learners" (M. Vanderwood, D. Linklater, and K. Healy, School Psychology Review 2008:5–17) reports that in a sample of 134 English-learning students, the correlation between the score on a phonics test given in first grade and a reading comprehension given in third grade was $r = 0.25$. Can you conclude that there is a positive correlation between phonics test score and the reading comprehension score?

13. The article "'Little Ice Age' Proxy Glacier Mall Balance Records Reconstructed from Tree Rings in the Mt. Waddington Area, British Columbia Coast Mountains, Canada" (S Larocque and D. Smith, The Holocene, 2005:748–757) evaluates the use of tree ring widths to estimate changes in the masses of glaciers. For the Sentinel glacier, the net mass balance (change in mass between the end of one summer and the end of the next summer) was measured for 23 years. During the same time period, the tree ring index for white bark pine trees was measured, and the sample correlation between net mass balance and tree ring index was $r = -0.509$. Can you conclude that the population correlation ρ differs from 0?
14. A scatterplot contains four points: (-2,-2), (-1,-1), (0,0), and $(1,1)$. A fifth point, $(2, y)$, is to be added to the plot. Let r represent the correlation between x and y
    a. Find the value of y so that $r = 1$.
    b. Find the value of y so that $r = 0$.
    c. Find the value of y so that $r = 0.5$.
    d. Find the value of y so that $r = -0.5$.
    e. Give a geometric argument to show that there is no value of y for which $r = -1$