POLITECNICO
MILANO 1863

DATA PROCESSING REPORT

Semester 2 - June 2025
Professor: Luigi Mussio

Jeanine Attard,
Alexia Baldacchino,
Barbara Riboni,
Mai Cao,
Federico Condotta.
Maciek Rabsztyn,
Andrew Sultana,
Siyuan Sun.

Table of Contents

INTRODUCTION

The growing importance of data in contemporary research and industry has positioned data processing as a central component of effective information management. It facilitates the systematic collection, transformation, and interpretation of data, allowing for the extraction of meaningful patterns and insights. This report presents a structured examination of key data processing techniques and numerical methodologies, with attention to their statistical foundations and interdisciplinary applications.

The report opens with applications of essential statistical concepts which establish the groundwork for the exploration of more advanced analytical approaches. Each section integrates theoretical explanation with practical context, aiming to demonstrate how different methods contribute to solving real-world problems.

To support the comprehension and application of the material presented, practical exercises are included throughout the report. These exercises are designed to reinforce theoretical knowledge through applied learning, enabling readers to develop both conceptual understanding and technical proficiency. By engaging with these tasks, readers will be better equipped to apply data processing methods within academic and professional contexts.

EXERCISE 1

One-Dimension Statistical Variable

Selected data set: ***Variation of Average January Temperature in Tokyo from year 2009 to 2024.***

Number of years	Average January Temperature in Tokyo	X_{ord}	$(x_i - \mu_x)^2$	$(x_i - \mu_x)^3$	$(x_i - \mu_x)^4$	$\lvert x_i - \mu_e \rvert$	$\lvert x_i - \mu_e \rvert_{ord}$
1	6.10	5.10	1.27	-1.42	1.60	1.20	0.00
2	6.30	5.20	1.05	-1.08	1.10	1.10	0.04
3	5.20	5.90	0.11	-0.03	0.01	0.40	0.09
4	5.10	5.90	0.11	-0.03	0.01	0.40	0.27
5	5.90	6.00	0.05	-0.01	0.00	0.30	0.33
6	6.50	6.10	0.02	0.00	0.00	0.20	0.39
7	6.40	6.10	0.02	0.00	0.00	0.20	0.47
8	6.60	6.30	0.01	0.00	0.00	0.00	0.51
9	6.80	6.30	0.01	0.00	0.00	0.00	0.53
10	5.90	6.40	0.03	0.01	0.00	0.10	0.68
11	6.30	6.40	0.03	0.01	0.00	0.10	0.89
12	7.10	6.50	0.08	0.02	0.01	0.20	1.04
13	6.90	6.60	0.14	0.05	0.02	0.30	1.28
14	6.00	6.80	0.33	0.19	0.11	0.50	1.40
15	6.10	6.90	0.46	0.31	0.21	0.60	19.10
16	6.40	7.10	0.77	0.67	0.59	0.80	20.19
	Mean	**Median**	**Standard deviation**	**Skewness**	**Kurtosis**	**M.A.V.**	**m.a.v.**
	6.23	**6.30**	**0.53**	**-0.57**	**2.96**	**0.40**	**0.53**
Number of all observations (n)	16				$\beta < 3$: platykurtic		
		Variance	**0.28**				

The table shared provides statistical analysis of the average January temperatures in Tokyo over a span of 16 years.

Central tendency measures:

First off, the average January temperature in Tokyo is tabulated for every year from 2009 till 2024. The mean average temperature was calculated by adding all the values and then dividing them by the number of observations which is 16. The average turned out to be 6.23 °C.

The median is calculated by putting all values in order in order to find the value in the middle. In this case the number of observations are even, therefore the middle 2 numbers are added and divided by 2. The median turned out to be 6.3°C.
These two values are very close, meaning that the temperature distribution is symmetrical although a slight skewness exists.

Dispersion measures:

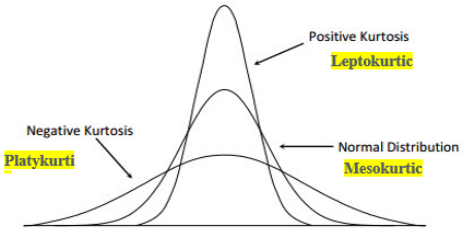
The deviation of each number is found by subtracting the mean from each value and then the answer is squared. To find the variance all the deviations from the mean are

summed up and divided by the total number of observations -1. (15). Then the standard deviation is found by square rooting the value of the variance. In this case, the variance is 0.28 and the standard deviation is 0.53. These show that the temperature readings do not vary wildly meaning that the dataset has low variability indicating consistency in January temperatures across the years.

Shape of the Distribution

Skewness measures whether the data is symmetrically distributed or tilted to one side. If the result is a negative value, that means that the distribution curve has a tail which is longer on the left meaning that there are more low values. In our dataset, the skewness is -0.57 meaning that the data is slightly left-skewed with a few colder years pulling the average down.

Kurtosis measures the tailedness or how heavily the tails of the distribution differ from a normal distribution. It measures the strength of the skewed distribution of the data. If the value of kurtosis is smaller than 3 the peak is flatter and the tails are lighter meaning that the data is more evenly spread out. This is called platykurtic. In our case it is indeed platykurtic with a kurtosis value which is 2.96. This is very close to 3 meaning that it is almost mesokurtic. There aren't many outliers.



Averaged deviation from the mean

M.A.V (Mean Absolute Variation), describes how far on average the temperature deviates from the mean temperature value. It is used as an alternative to standard deviation because it doesn't square the deviations. This is useful since the extreme values are less impactful on the total value. In our case the average distance from the mean is 0.40°C so most of the values are close to the value of the average.

Typical Deviation from the Median

m.a.v (Median Absolute Variation), indicates the typical distance of temperature values from the median providing a robust measure of spread. Similar to M.A.V it also features a value which has less influence from extreme values. In this case, the typical distance from the median is 0.53°C indicating that the spread around the median is slightly wider compared to that of the mean.
The difference between the mean absolute variation and the median absolute variation aligns with the slight left skew of the distribution curve. This is because more data points fall below the median than above.

HISTOGRAM

The distribution and frequency of the average January temperatures throughout the 16-year period (from 2009-2024) in Tokyo are shown using the histogram. The histogram plots the temperature values on the y-axis and the year of occurrences on the x-axis. The frequency of a certain temperature range over time is shown by each bar in the histogram. Understanding the variations in January temperatures and seeing any patterns / trends throughout time are made easier with the aid of this visualisation.

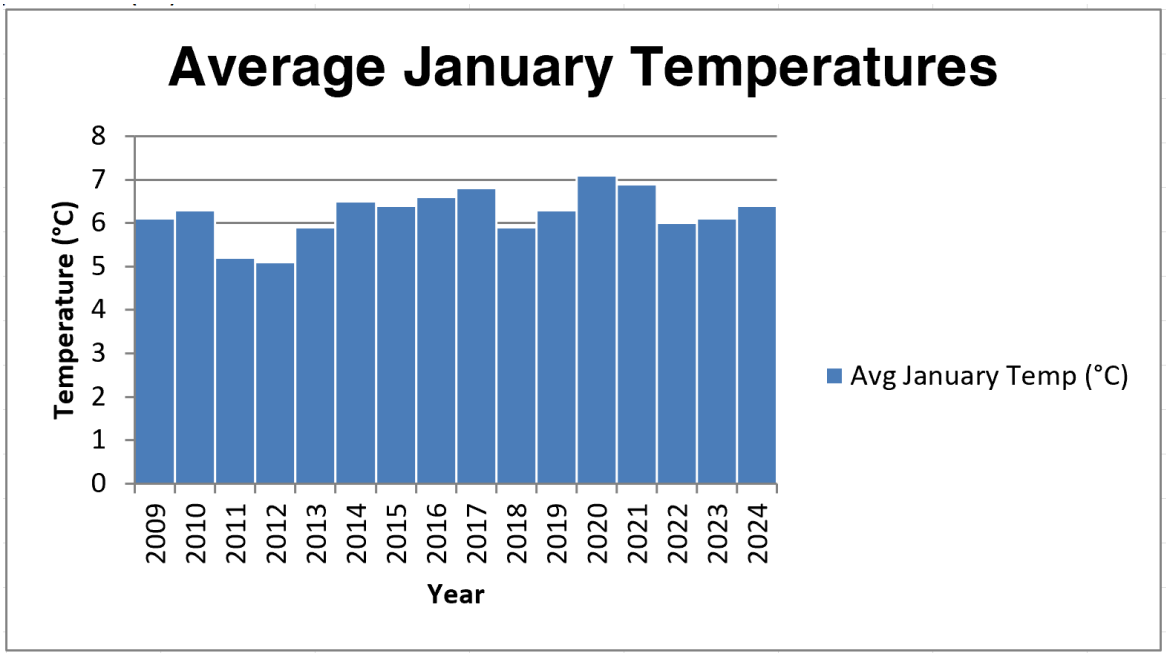
Data usage:

To display the frequency of specific temperature ranges throughout the specified years, the histogram makes use of the complete dataset. It offers information about the general distribution and range of temperatures.

According to the histogram:

- the lowest temperature ever recorded was 5.1 °C in 2012,
- the highest temperature ever recorded was 7.1 °C in 2020,
- The distribution indicates that temperatures generally fluctuate around 6.3°C, with some exceptions.

This histogram aids in determining whether there are any notable fluctuations / consistent patterns in the January temperatures across time.



PROFILE CURVE

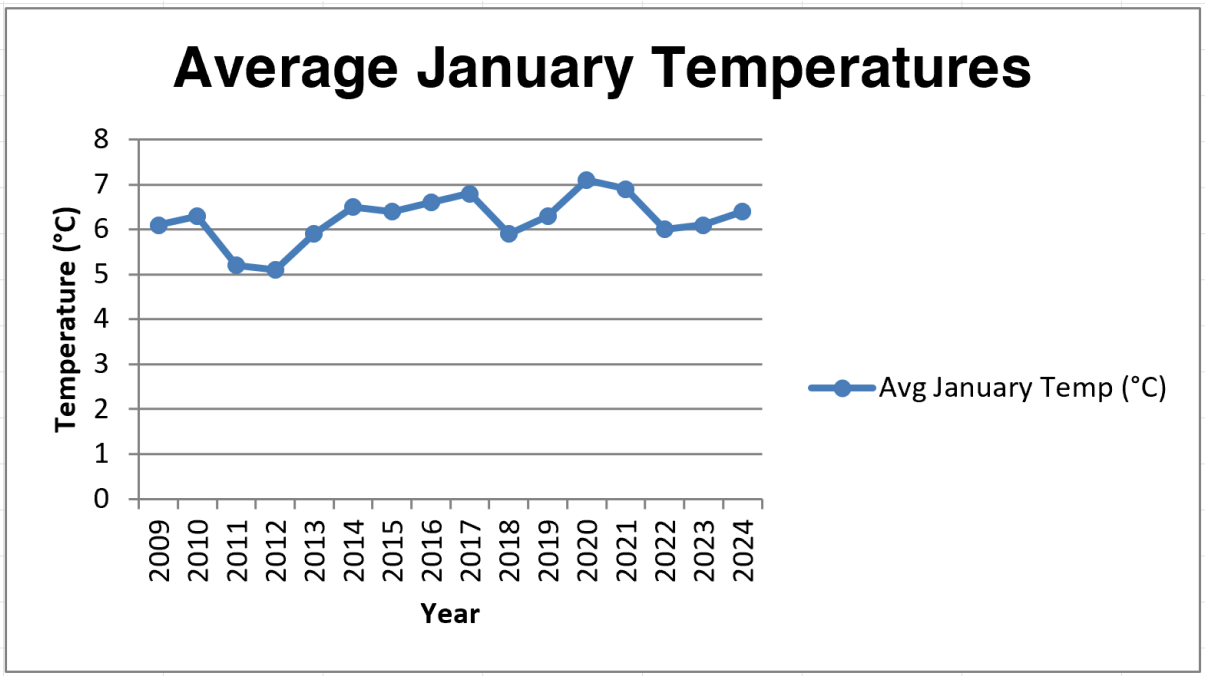
The trend and variations of the average January temperatures over time are depicted by the profile curve.

A line graph with the years on the x-axis and the temperature values on the y-axis is called a profile curve. The data points are connected by a curve that illustrates the variations in temperature over time. The temperature variations throughout time are clearly visible in this graph, which also identifies any patterns or changes.

Data Usage:

To display the evolution of January temperatures in Tokyo against time, the profile curve makes use of the complete dataset. It offers a visual depiction of temperature variations and aids in locating times of notable increase or decrease.

Any long-term trends, such as a general warming or cooling pattern, can be seen in this graph, along with any outlier years where the temperature departed greatly from the norm.



BOX PLOT

Another useful way of representing the data is through a box plot. The chart's top and bottom whiskers display the highest and lowest January temperatures ever recorded for the specified years. The box's top denotes the third quartile (Q3), while its bottom denotes the first quartile (Q1). The median temperature is shown by the line in the center of the box.

The box plot highlights important statistical figures while summarising the spread of Japan's average January temperature. Temperature values are displayed on the y-axis. The box plot shows the dataset's medium, interquartile range (IQR), and any outliers. This helps in understanding the spread, central tendency, and variability of January temperatures over the years.

Data Usage

The box plot computes the quartiles, the median and mean temperatures. It provides a comprehensive summary of the data, highlighting the range and distribution of temperature fluctuations. This enables us to examine how consistently January temperatures have changed throughout time and spot any years with outliers.

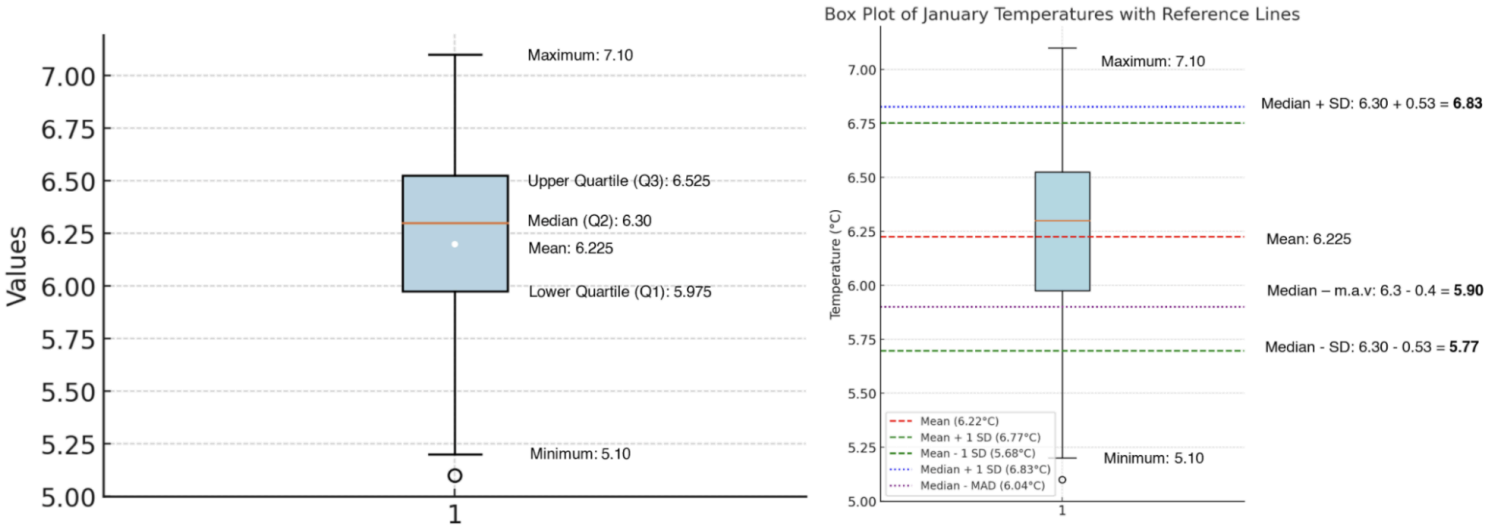


Figure X: Box plot with quartiles VS Figure Y Box plot with standard deviation

Top Whisker - Max. Value	7.10
Top Line Box – Q3	6.525
Middle Line Box - Median	6.30
White Dot - Mean	6.225
Bottom Line Box – Q1	5.975
Bottom Whisker - Min. Value	5.10

The Confidence Bound

This is calculated by adding and subtracting the standard deviation from the median respectively, and can also be used to define the box's range when creating a box plot. This technique offers a different approach to visualising the variability and dispersion of data. As a result, as shown in the box plot Y, the box limits in this method might be marginally different from those in box plot X. This variation demonstrates how the interpretation of the dataset's central tendency and spread can be influenced by various statistical techniques.

CLASS EXTREMES AND RELATIVE FREQUENCIES

In the next part, all of the data is split into classes to understand the frequency, distribution and the probabilities for easier statistical interpretation.

In our case, the minimum temperature is that of 5.10°C and the maximum temperature is 7.10°C. When the maximum is subtracted by the minimum, the amplitude of the data is found to be 2°C. The amplitude is divided by the number of classes which is 4 in this case to obtain a step of 0.5°C. This results in 5 classes of extremes. The first and last extreme are the minimum and the maximum. In order to find the other 3 extremes one can simply add the class width which is 0.5°C to the minimum and so on.

<u>Class standard extremes</u>

In order to calculate the class standard extremes, the mean value is subtracted from each extreme and then dividing the answer by the standard deviation. All these values are provided in the figure below.

	X_{ord}	Extremes	Class extremes			Class standard extremes
X_{min}	5.10	1	1st extreme	1	5.1	-2.13
	5.20	1				
	5.90	2				
	5.90	2	2nd extreme	2	5.6	-1.19
	6.00	2				
	6.10	3				
	6.10	3				
	6.30	3	3rd extreme	3	6.1	-0.24
	6.30	3				
	6.40	3				
	6.40	3				
	6.50	3				
	6.60	0				
	6.80	4	4th extreme	4	6.6	0.71
	6.90	4				
X_{max}	7.10	4	5th extreme	5	7.1	1.66
	Number of classes (m)		4			Step
	2.00	Amplitude/Step	0.5			

Class centers			Class standard centers	Absolute frequencies	Cumulative absolute frequencies	Relative frequencies	Cumulative relative frequencies	Cumulative Normal probabilities	(Simple) Normal probabilities	Extremes	Value in the table
					0		0	0			
1st center	1	5.35	-1.66	2	2	0.13	0.13	0.0170	0.1209	-2.13	0.4830
2nd center	2	5.85	-0.71	3	5	0.19	0.31	0.1379	0.3342	-1.19	0.3621
3rd center	3	6.35	0.24	7	12	0.44	0.75	0.4721	0.3568	-0.24	0.0279
4th center	4	6.85	1.19	3	15	0.19	0.94	0.8289	0.1472	0.71	0.3289
	=										
						0.94		0.9761		1.66	0.4761

Class centers

After the class standard extremes are found, the class centers are calculated by finding the midpoint of each class. The mean is subtracted from each centre and divided by the standard deviation to bring out the class standard centre, These values are similar to the class standard extremes.

Frequency

The absolute frequency describes the number of observations per class. In order to find the Cumulative absolute frequency all the frequencies are summed up to each class. So for example to calculate in our case the cumulative absolute frequency of the 2nd centre, one must add the absolute frequency of the first and second centres which add up to 5.

Relative Frequency

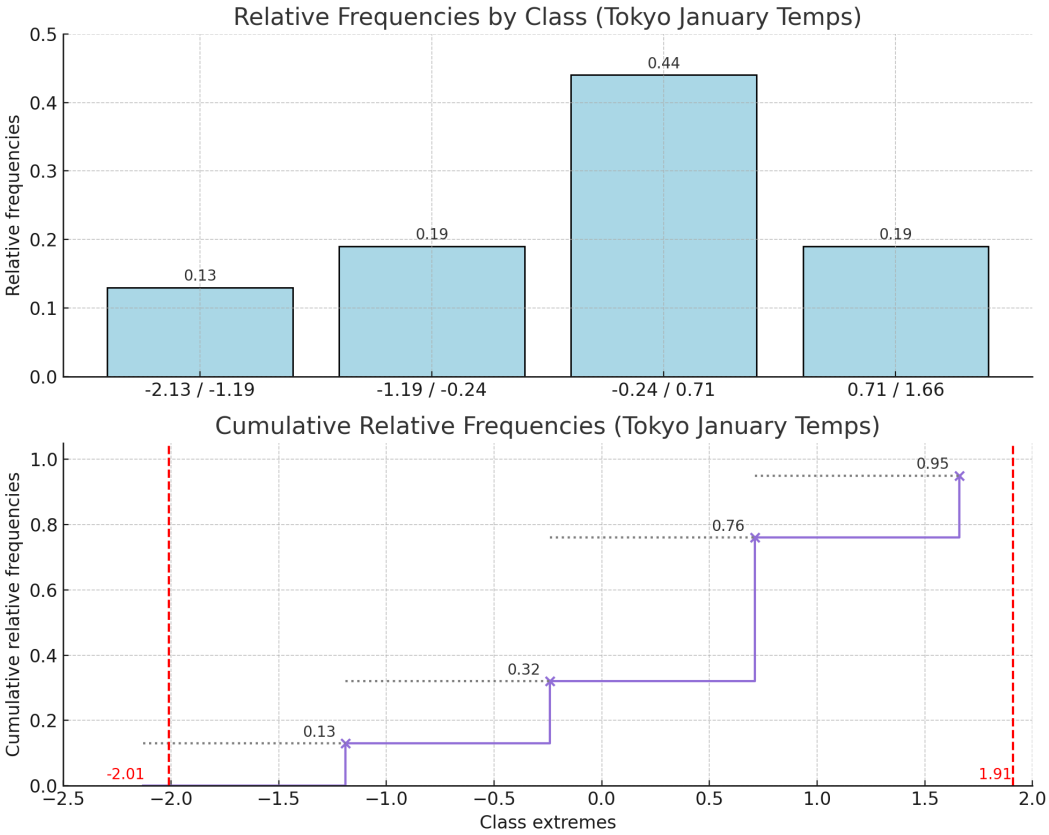
In the next step, the relative frequency is calculated which tells you how often a value or group of values appears in a dataset relative to the total number of values. Finally, the cumulative of this relative frequency is found which is the running total of the relative frequencies. In our examples, class 3 contains 7 values which in this case is the highest frequency. On the other hand, class 1 contains only 2 values which is the lowest frequency. All this information is tabulated in the graphs below.

Normal Probabilities

After the standard deviation is developed, the cumulative normal probabilities are calculated which are the area under the normal curve. Apart from this value, the Simple normal probabilities are also found which are the probability densities for each standardized center. The values are observed in the tables below.

Class	Relative Frequency	Cumulative Frequency
-2.13 / -1.19	0.0625	0.0625
-1.19 / -0.24	0.1875	0.25
-0.24 / 0.71	0.4375	0.6875
0.71 / 1.66	0.25	0.9375

Class Boundary	Cumulative Relative Frequency
-3.0	0.0
-2.13	0.0
-2.13	0.0625
-1.19	0.0625
-1.19	0.25
-0.24	0.25
-0.24	0.6875
0.71	0.6875
0.71	0.9375
1.66	0.9375



EXERCISE 2

Two-Dimension Statistical Variables

Dimension Statistic Variable

Data set: Activity rates by region from the 1881-2011 Censuses

Link to data set:

https://seriestoriche.istat.it/index.php?id=1&no_cache=1&tx_usercento_centofe%5Bcategoria%5D=10&tx_usercento_centofe%5Baction%5D=show&tx_usercento_centofe%5Bcontroller%5D=Categoria&cHash=442f60de54147698370ad25c402fe442

2.1_INTRODUCTION

For the second exercise, we have compiled a two-way table (shown in fig.2.a).

In this exercise, compared to the first one, we examine more variables. Specifically, in a sample group of employment rates, each rate is associated with two statistical variables X and Y, which take into account four specific data points.

The table identifies some activity rates for Italian regions in the Censuses from 1881 to 2001. The regions analyzed on variable Y are: Piemonte, Liguria, Lombardia, and Veneto; the years considered are 1881, 1931, 1971, and 2001.

Each rate, therefore, depends on the pair of variables Region-Year. We will later attempt to analyze the various activity rates with respect to these two variables.

Dimension Statstic Variable			Y^2 (regions of Italy)	Piemonte	Liguria	Lombardia	Veneto	
				1	4	9	16	
			y					917
X^2 (years)			X/Y	1	2	3	4	P_i
1881	1		1	78,7	66,0	78,2	66,9	290
1931	4	x	2	62,1	54,8	58,0	59,2	234
1971	9		3	50,0	43,3	51,3	48,5	193
2001	16		4	50,5	44,5	52,9	52,5	200
		917	Q_j	241	209	240	227	917,4

Fig.2.a

In the table shown (Fig.2.a), we find relative frequencies in the center, and values of data in X and Y, marginal frequencies in X and Y, the number 1 if we are dealing with relative frequencies and the number of data if we are dealing with absolute frequencies. Each column goes from Frequency to Marginal Frequency, each row goes from frequency to the other marginal frequency, and the sum of each frequency and the sum of the marginal frequencies gives 1 with relative data, or the number of data with absolute frequencies.

2.2_ CONNECTION

	Contingence table: $C_{ij} = CF_{ij} - P_iQ_j$				
	y				
		1	2	3	4	
x	1	2270,64	96,12	2072,76	-4439,52	
	2	482,21	1440,26	-3068,44	1145,97	
	3	-725,03	-557,24	641,38	640,89	Sum
	4	-2027,82	-979,14	354,30	2652,66	0,00

Fig.2.b

Connection is considered as non-independence. If we consider joint probability and the product of the two marginals, there are two possibilities:

- If the difference is zero, then we have the exact definition of independence, and this variable, which we call contingency, will be 0.

- If we are facing non-independence, we obtain a new variable called contingency, derived from the difference between the joint frequency and the product of the marginal frequencies.

Contingency ranges from -1 to +1, with zero indicating independence and one indicating perfect dependence. (Fig.2.b)

Contingencies: $c_{ij} = f_{ij} - p_i q_j$

$-1 \leq c_{ij} \leq 1$

	Absolute contingence table: $C_{ij} = \lvert CF_{ij} - P_iQ_j \rvert$				
	y				
		1	2	3	4	
x	1	2270,64	96,12	2072,76	4439,52	
	2	482,21	1440,26	3068,44	1145,97	
	3	725,03	557,24	641,38	640,89	C_0=Sum/2
	4	2027,82	979,14	354,30	2652,66	11797,19

P_i^2
83984
54803
37288
40160
625388
N^2-sum(P_i^2)

Q_i^2	58226	43514	57792	51574	630517
					N^2-sum(Q_i^2)

Fig.2.c.

If we try to analyze the product of marginals in absolute terms, we may obtain dangerously high values, possibly tending towards infinity. In this case, we analyze the table (Fig.2.c) that presents the values through which we can calculate the Bonferroni indices. We obtain them by summing all the absolute values of the contingencies and dividing by the value Qj2 (Row) and Pi2 (Column) to get a number from 0 to 1 instead of 0 to infinity, which indicates respectively the two Monolateral Bonferroni's indices (Fig.2.d).

Semi contingency mean: $C_0 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|c_{ij}\right|$

Bonferroni's indices		
β_X	Monolateral Bonferroni's indices	0,02
β_Y		0,02
β_{-1}	Armonic bilateral Bonferroni's index	0,02
β_0	Geometric bilateral Bonferroni's index	0,02

Fig.2.d.

We can now observe two specific values (Fig.2.d):

- Geometric bilateral Bonferroni's index, which has a geometric meaning with two variables.
- Harmonic bilateral Bonferroni's index, the inverse of the inverse multiplied by the number of elements (in this case 2).

Bonferroni unilateral indices:

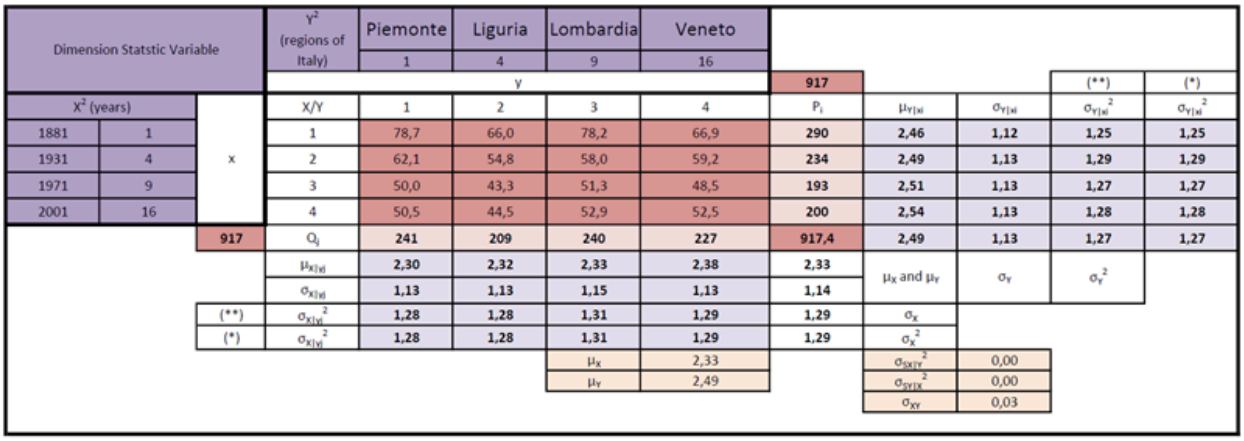
$$\begin{cases} \beta_X = \dfrac{C_0}{1-\sum_{i=1}^{n} p_i^2} \quad ; \quad independence \quad 0 \le \beta_X \le 1 \quad perfect\ dependence \quad x = h(y) \\ \beta_Y = \dfrac{C_0}{1-\sum_{j=1}^{m} q_j^2} \quad ; \quad independence \quad 0 \le \beta_Y \le 1 \quad perfect\ dependence \quad y = g(x) \end{cases}$$

Bonferroni bilateral indices:

$$\begin{cases} \beta_0 = \sqrt{\beta_X \beta_Y} \quad ; \quad independence \quad 0 \le \beta_0 \le 1 \quad perfect\ bilateral\ dependence \\ \beta_{-1} = \dfrac{2\beta_X \beta_Y}{\beta_X + \beta_Y} \quad ; \quad independence \quad 0 \le \beta_{-1} \le 1 \quad perfect\ bilateral\ dependence \end{cases}$$

We have now identified everything necessary to describe what concerns connection.

2.2_ CORRELATION



Dimension Statstic Variable			Y^2 (regions of Italy)	Piemonte	Liguria	Lombardia	Veneto					
				1	4	9	16					
				y				917			(**)	(*)
X^2 (years)			X/Y	1	2	3	4	P_i	$\mu_{Y\|xi}$	$\sigma_{Y\|xi}$	$\sigma_{Y\|xi}^2$	$\sigma_{Y\|xi}^2$
1881	1		1	78,7	66,0	78,2	66,9	290	2,46	1,12	1,25	1,25
1931	4	x	2	62,1	54,8	58,0	59,2	234	2,49	1,13	1,29	1,29
1971	9		3	50,0	43,3	51,3	48,5	193	2,51	1,13	1,27	1,27
2001	16		4	50,5	44,5	52,9	52,5	200	2,54	1,13	1,28	1,28
		917	O_j	241	209	240	227	917,4	2,49	1,13	1,27	1,27
			$\mu_{X\|yi}$	2,30	2,32	2,33	2,38	2,33	μ_X and μ_Y	σ_Y	σ_Y^2	
			$\sigma_{X\|yi}$	1,13	1,13	1,15	1,13	1,14				
		(**)	$\sigma_{X\|yi}^2$	1,28	1,28	1,31	1,29	1,29	σ_X			
		(*)	$\sigma_{X\|yi}^2$	1,28	1,28	1,31	1,29	1,29	σ_X^2			
						μ_X	2,33		$\sigma_{X\|Y}^2$	0,00		
						μ_Y	2,49		$\sigma_{Y\|X}^2$	0,00		
									σ_{XY}	0,03		

Fig.2.e.

It is possible to obtain marginals in X and in Y, as well as the marginal mean in X and Y, and marginal variance in X and Y. It is also possible to compute a second index that does not indicate a square (as in the case of variances), but applies the product between X and Y. Then all combinations are summed, considering all the frequencies within the two-dimensional table. (Fig.2.e.)

Correlation (linear dependence)

Marginal distribution:

$$X\begin{cases} x_1 & \cdots & x_i & \cdots & x_n \\ p_1 & \cdots & p_i & \cdots & p_n \end{cases}$$

$$Y\begin{cases} y_1 & \cdots & y_i & \cdots & y_n \\ q_1 & \cdots & q_i & \cdots & q_n \end{cases}$$

Mean of marginal variable X: $\mu_X = \sum_{i=1}^{n} x_i p_i$

Mean of marginal variable Y: $\mu_Y = \sum_{j=1}^{n} y_j q_j$

Variance of marginal variable X: $\sigma_X^2 = \sum_{i=1}^{n} (x_i - \mu_X)^2 p_i$

Variance of marginal variable Y: $\sigma_Y^2 = \sum_{j=1}^{m} (y_j - \mu_Y)^2 q_j$

Covariance between variables X and Y: $\sigma_{XY} = \sum_{i=1}^{n}\sum_{j=1}^{m} (x_i - \mu_X)(y_j - \mu_Y) f_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j f_{ij} - \mu_X \mu_Y$

Using these formulas, we consider all possible products, all possible combinations. We thus obtain a connection between the two variables called Covariance, which depends on the range of dependencies. It can be positive or negative. (Fig.2.f)

Covariance computation σ_{XY} between X and Y:			
78,7	132	234,6	267,6
124,2	219,2	348	473,6
150	259,8	461,7	582
202	356	634,8	840

Fig.2.f.

It is also possible to obtain a correlation coefficient (P), which is the ratio between the Covariance and the product of the two corresponding standard deviations. The correlation coefficient ranges from -1 to +1 (because Covariance is not necessarily positive), and its square will be the Covariance squared divided by the product of the two variables, with a range from 0 to +1. 0 indicates no correlation, and 1 indicates maximum correlation.

Linear correlation coefficient:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad ; \quad reverse\ perfect\ dependence \quad -1 \le \rho \le 1 \quad direct\ perfect\ dependence$$

$$\rho = 0 \ \ linear\ independence$$

Regression Lines: $Y = aX + b$ $\qquad X = cY + d$ **Robust regression lines:**

$$a = \frac{\sigma_{XY}}{\sigma_X^2} \qquad c = \frac{\sigma_{XY}}{\sigma_Y^2} \qquad a = median\left(a_{ij}\right) \quad \forall i, j > i \qquad c = \ldots$$

$$b = \mu_Y - a\mu_X \qquad d = \mu_X - c\mu_Y \qquad b = median\left(y\right) - a\ median\left(x\right) \qquad d = \ldots$$

Functional regression Pearson's indices					
η_X^2	0,00	$\sigma_{RX\|Y}^2$	1,29	σ_X^2	1,29
η_Y^2	0,00	$\sigma_{RY\|X}^2$	1,27	σ_Y^2	1,27
η^2	0,00	r_{xy}	0,03		

Fig.2.g.

Linear regression and correlation					
y = aX + b			x = cY + d		
a or	A (Y(X))	0,02	c or	A (X(Y))	0,03
b or	B (Y(X))	2,44	d or	B (X(Y))	2,27

Correlation coefficient r_{XY}	0,03	r_{XY}^2	0,00

Standard dev.'s	
$\sigma_{SX\|Y}$	0,03
$\sigma_{SY\|X}$	0,03

$\sigma_{RX\|Y}$	1,14
$\sigma_{RY\|X}$	1,13

Fig.2.h.

The regression line is a straight line that can correlate X with respect to Y or Y with respect to X. (Fig.2.f). The correlation coefficient is obtained by dividing Covariance by the square of the variables in X for the straight line in Y, and analogously for the straight line in X. A and C are considered slope coefficients. Intersections are obtained using the values derived. (Fig.2.h / Fig.2.i)

Regr. line	y=ax+b	m(y|x)	v(y)
1	2,46	2,46	0,00
2	2,49	2,49	0,00
3	2,51	2,51	0,00
4	2,54	2,54	0,00
		$\sigma_0(y)^2$	0,00
		$\sigma_0(y)$	0,00

Confidence bounds	
2,36	2,56
2,40	2,57
2,42	2,60
2,43	2,64

Regr. line	x=cy+d	m(x|y)	v(x)
1	2,29	2,30	-0,01
2	2,32	2,32	0,00
3	2,34	2,33	0,02
4	2,37	2,38	-0,01
		$\sigma_0(x)^2$	0,00
		$\sigma_0(x)$	0,01

Confidence bounds	
2,19	2,40
2,23	2,41
2,25	2,44
2,27	2,47

Fig.2.i.

Regr. curve	Confidence bounds		Regr. curve	Confidence bounds	
2,46	1,34	3,58	2,30	1,17	3,43
2,49	1,35	3,62	2,32	1,19	3,45
2,51	1,38	3,64	2,33	1,18	3,47
2,54	1,41	3,67	2,38	1,25	3,52

Fig.2.l.

2.3_TABELLE ILLUSTRATIVE

After having observed all the data calculable through the analyses on the X and Y table, we try to represent them through graphical tables and analyze the results.

The variables in X represent the years under analysis, while the Y variables represent the regions. This two-dimensional representation allows us to understand the level of activity rates over the years in the various Italian regions analyzed.

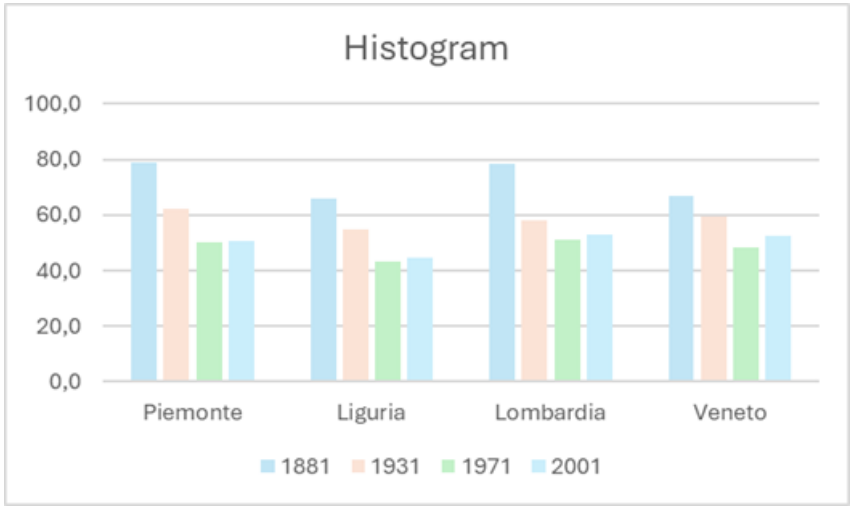


Fig.2.m.

Through the first histogram (Fig.2.m), we can observe in general that the activity rate in the various regions has decreased over the years regardless of specific location, showing a homogeneous and not particularly alarming decrease.

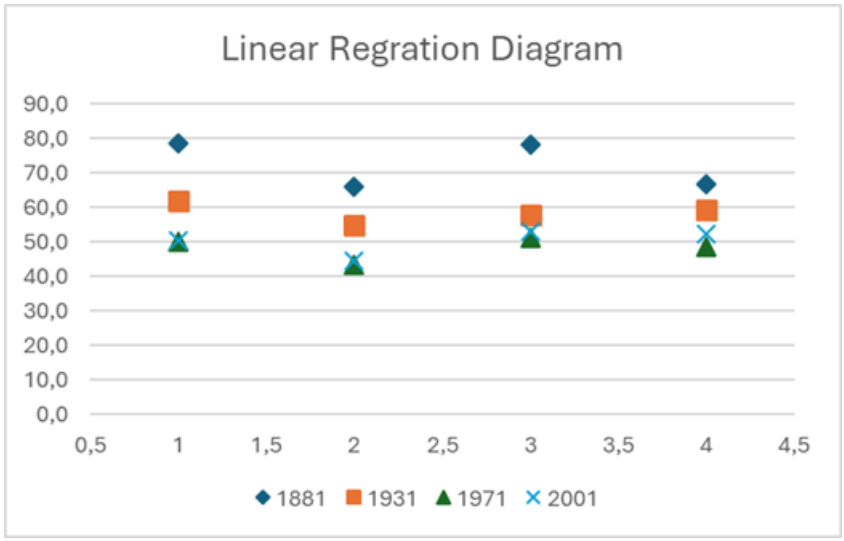


Fig.2.n.

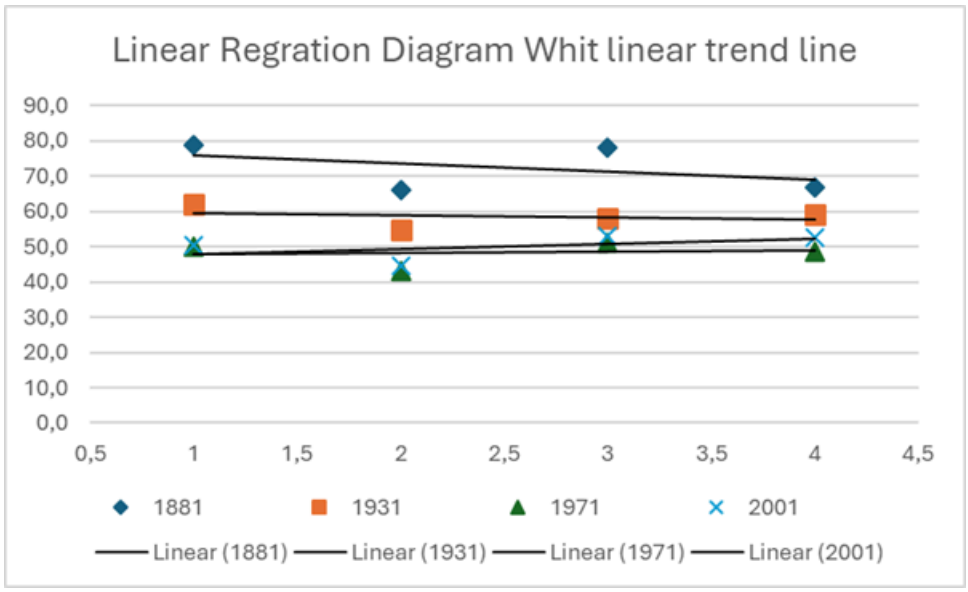


Fig.2.o.

This homogeneity of the employment rate among the different regions is well observable through the diagrams created, which show that the trend among the regions with respect to the same years is linear. Each year showed a nearly equal rate among the different regions. (Fig.2.o/Fig.2.n)

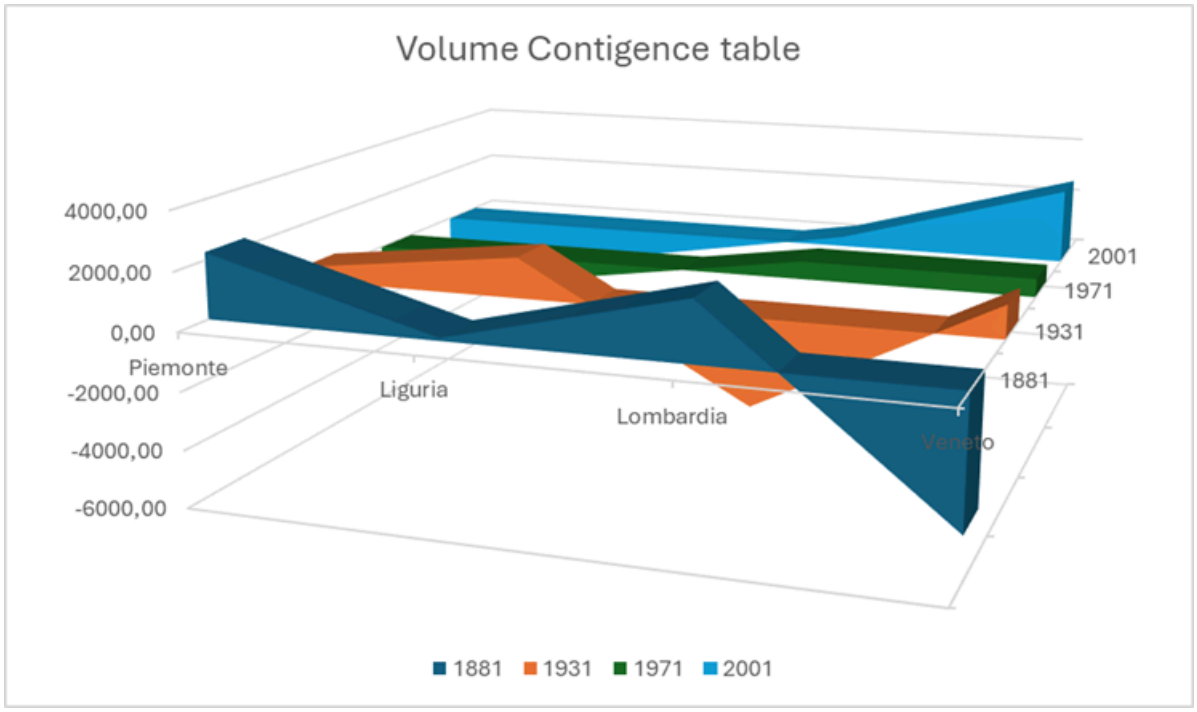


Fig.2.p.

Even through the observation of the last diagram, we notice that the rate remains influenced by the passing of years rather than by geographic location.

EXERCISE 3

Input Data: Examples of Normal Distribution - Free Sample Tests

Input data represents the set of values or observations fed into a statistical model to help it produce results or make predictions.

Importance of Normal Distribution

The normal distribution, also known as the Gaussian distribution, is one of the most important and widely used probability distributions in statistical analysis. It is important in areas such as hypothesis testing, predictive modeling, and analyzing random events.

For data to be considered normally distributed, it should have values that are symmetrically distributed around the mean, and most of the data points cluster near the center with fewer observations appearing as you move further away. This is called a bell-shaped curve. It has a peak at the mean and tails that gradually decrease on both sides.

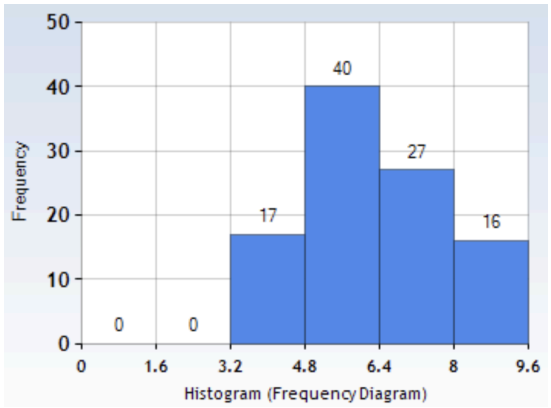
Observed Variables

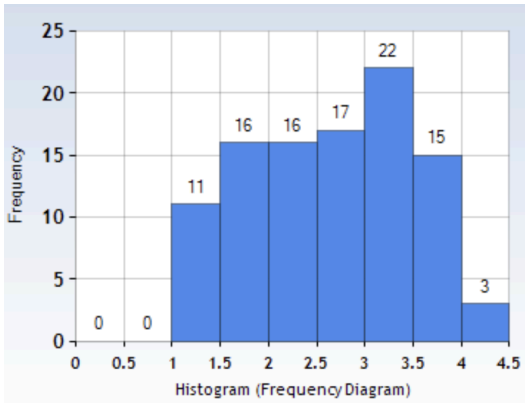
A practical example of input data involves two behavioral variables, screen time and sleep duration. These factors are often linked to lifestyle and health outcomes, making them valuable for observation. The data was originally collected from 500 individuals, with each participant assigned a numerical value, “Student ID,” along with their average screen time (in hours per day) and sleep duration (in hours per night).

In statistics, a common type of input data is information that follows a normal distribution. When plotted on a histogram, this kind of data typically forms a bell-shaped curve, indicating its distribution. By examining statistical measures such as mean, variance, and standard deviation, we can gain insight into the patterns within the dataset. When used in statistical models, this data helps researchers and analysts make informed predictions, identify trends, and draw meaningful conclusions.

The table shows the data of 100 students, and histograms of these data.

Student_ID	Sleep_Duration	Screen_Time	Student_ID	Sleep_Duration	Screen_Time	Student_ID	Sleep_Duration	Screen_Time	Student_ID	Sleep_Duration	Screen_Time	Student_ID	Sleep_Duration	Screen_Time
1	7.7	3.4	21	6.5	3.7	41	8.3	1.4	61	4.5	3.2	81	5.3	2
2	6.3	1.9	22	7	2.9	42	8.7	3.2	62	5.2	4	82	5.9	1.4
3	5.1	3.9	23	4.3	1.9	43	5.9	3.5	63	7.6	1.3	83	7	1.9
4	6.3	2.8	24	7.7	3.4	44	5.4	1.6	64	8.3	2.2	84	5.3	3.6
5	4.7	2.7	25	5	3.6	45	7.2	2.5	65	8.2	3.4	85	7.1	4
6	4.9	3.2	26	8.5	3.5	46	6	3.5	66	6	1.6	86	6	2.1
7	6.5	3.4	27	5	3.8	47	4.1	3.2	67	7.3	2.7	87	6.8	2.3
8	6.1	3	28	5	1.8	48	4.8	2.6	68	5	3.2	88	6.2	3.2
9	8.6	1.4	29	4.2	3.3	49	7.6	2.8	69	5.5	2.8	89	5.5	3.7
10	5.8	2	30	6.4	2.4	50	7.3	2.5	70	8.5	1.6	90	8.7	2.8
11	6.9	3.8	31	6.8	3.5	51	4.1	1.9	71	4.1	2.1	91	7.8	2.2
12	7.2	1.7	32	4.3	3.2	52	5.1	2.7	72	4.4	3.4	92	4.7	2.2
13	4.1	2.1	33	7.9	3.3	53	5.2	3.1	73	5	2.7	93	8.3	3.1
14	7.3	2.3	34	6.3	3	54	7.4	3.6	74	4.1	1	94	6.4	1
15	4.9	2.3	35	6.6	1.5	55	4.1	2.9	75	4.9	3.3	95	8.5	2.9
16	8.8	2.8	36	6.2	2.6	56	4.5	3.3	76	6.9	1.1	96	8	2.1
17	4.7	3.8	37	6	4	57	8	1.5	77	6.1	3.2	97	6.1	3.4
18	6.1	1.7	38	6.8	3.8	58	4.9	2.4	78	8.5	1.6	98	4.1	1.3
19	4.4	1.4	39	4.8	1.1	59	7.3	1	79	8.1	3.9	99	5.3	2.8
20	9	1.6	40	4.9	1.5	60	5.2	1.7	80	5.7	2.1	100	6.7	2.4





Input Data

In order to make simpler calculations, we will take a sample of 9 students.

Student_ID	Sleep_Duration	Screen_Time
1	7.7	3.4
2	6.3	1.9
3	5.1	3.9
4	6.3	2.8
5	4.7	2.7
6	4.9	3.2
7	6.5	3.4
8	6.1	3
9	8.6	1.4

To interpret the data, we calculate the mean, variance, and standard deviation for both variables. These metrics help describe the central tendency and spread of the dataset. This was done automatically using excel.

Formulas:

Mean: $\mu = (\Sigma x) / N$

Variance: $\sigma^2 = \Sigma(x - \mu)^2 / (N - 1)$

Standard Deviation: $\sigma = \sqrt{\sigma^2}$

Number of observation	Observations X	Observations Y	X_{cod}	Y_{cod}	$(x_i - \mu_0)^2$	$(x_i - \mu_Y)^2$	$(x_i - \mu_0)^3$	$(x_i - \mu_Y)^3$	$(x_i - \mu_0)^4$	$(x_i - \mu_Y)^4$
1	7.70	3.40	1.40	2.00	23.4686	0.7320	-113.6925	-0.6262	550.7772	0.54
2	6.30	1.90	1.90	3.00	18.8742	0.0209	-81.9979	0.0030	356.2353	0.00
3	5.10	3.90	2.70	4.00	12.5631	1.3098	-44.5292	1.4989	157.8311	1.72
4	6.30	2.80	2.80	5.00	11.8642	4.5986	-40.8656	9.8615	140.7592	21.15
5	4.70	2.70	3.00	7.00	10.5264	17.1764	-34.1524	71.1867	110.8055	295.03
6	4.90	3.20	3.20	8.00	9.2686	26.4653	-28.2179	136.1493	85.9077	700.41
7	6.50	3.40	3.40	9.00	8.0909	37.7542	-23.0140	231.9786	65.4621	1425.38
8	6.10	3.00	3.40	9.00	8.0909	37.7542	-23.0140	231.9786	65.4621	1425.38
9	8.60	1.40	3.90	10.00	5.4964	51.0431	-12.8861	364.6745	30.2106	2605.40

Mean value	6.244	2.856	Standard deviation X	Standard deviation Y	Skewness X (γ_X)	Skewness Y (γ_Y)	Kurtosis X (β_X)	Kurtosis (β_Y)
Number samples	9	9	3.468	4.433	-1.072	1.335	1.201	1.863
							$\beta < 3$: platykurtic	$\beta < 3$: platykurtic

σ_{XY}	-0.44
R (X,Y)	**-0.03**

Frequencies

	y					
F(X,Y)	1	2	3	4	5	P_i
1	0.000	0.111	0.000	0.111	0.000	0.222
2	0.000	0.000	0.000	0.111	0.000	0.111
3	0.000	0.000	0.000	0.000	0.222	0.222
4	0.111	0.111	0.000	0.000	0.000	0.222
5	0.111	0.000	0.111	0.000	0.000	0.222
Q_j	0.222	0.222	0.111	0.222	0.222	1.000

Standard square contingencies

0.0494	0.0772	0.0247	0.0772	0.0494
0.0247	0.0247	0.0123	0.3025	0.0247
0.0494	0.0494	0.0247	0.0494	0.6049
0.0772	0.0772	0.0247	0.0494	0.0494
0.0772	0.0494	0.3025	0.0494	0.0494
Total:	2.25			

Product of cumulative frequencies

	y				
P_CF (X,Y)	1	2	3	4	5
1	0.0247	0.0247	0.0494	0.0494	0.1975
2	0.0494	0.0494	0.0988	0.0988	0.3951
3	0.0988	0.0988	0.1975	0.1975	0.7901
4	0.1111	0.1111	0.2222	0.2222	0.8889
5	0.1111	0.1111	0.2222	0.2222	0.8889

INPUT DATA: EXAMPLES OF NORMAL AND DISTRIBUTION - FREE SIMPLE TESTS

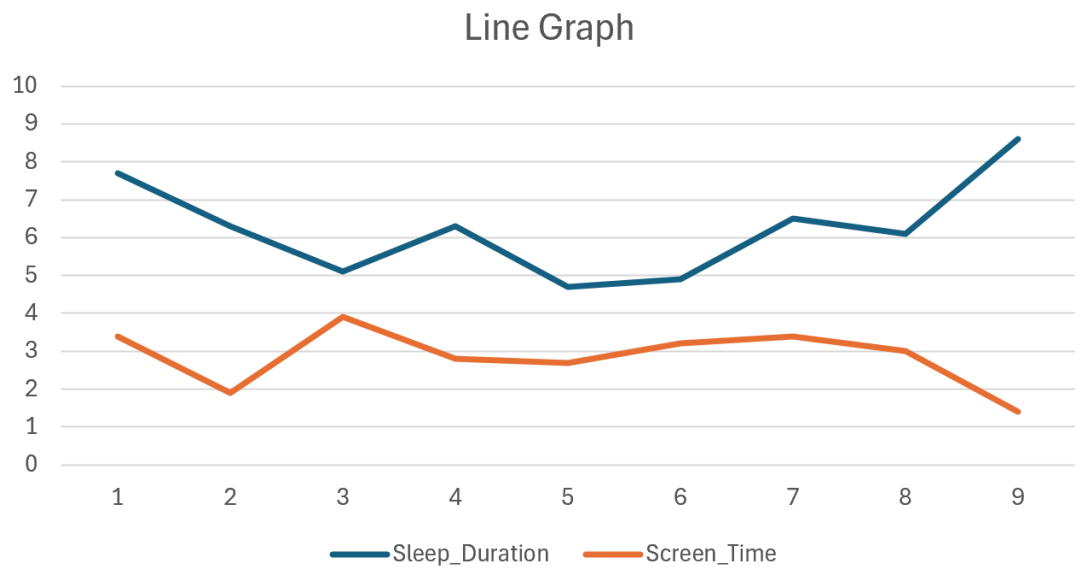
	X ord and extreme		Class extremes			Class standard extremes	Class centers			Class standard centers	Absolute frequencies	Cumulative absolute frequencies
X_{min}	1.40	1	1st extreme	1	1.4	-1.40	1st center	1	1.79	-1.28	2	2
	1.90	1										
	2.70	2	2nd extreme	2	2.18	-1.17	2nd center	2	2.57	-1.06	2	4
	2.80	2	3rd extreme	3	2.96	-0.95	3rd center	3	3.35	-0.83	4	8
	3.00	3										
	3.20	3	4th extreme	4	3.74	-0.72	4th center	4	4.13	-0.61	1	9
	3.40	3										
	3.40	3	5th extreme	5	4.52	-0.50	5th center	5	4.91	-0.38	0	9
X_{max}	3.90	4										
			6th extreme	6	5.30	-0.27						
	Number of classes	5	Interval				0.780					

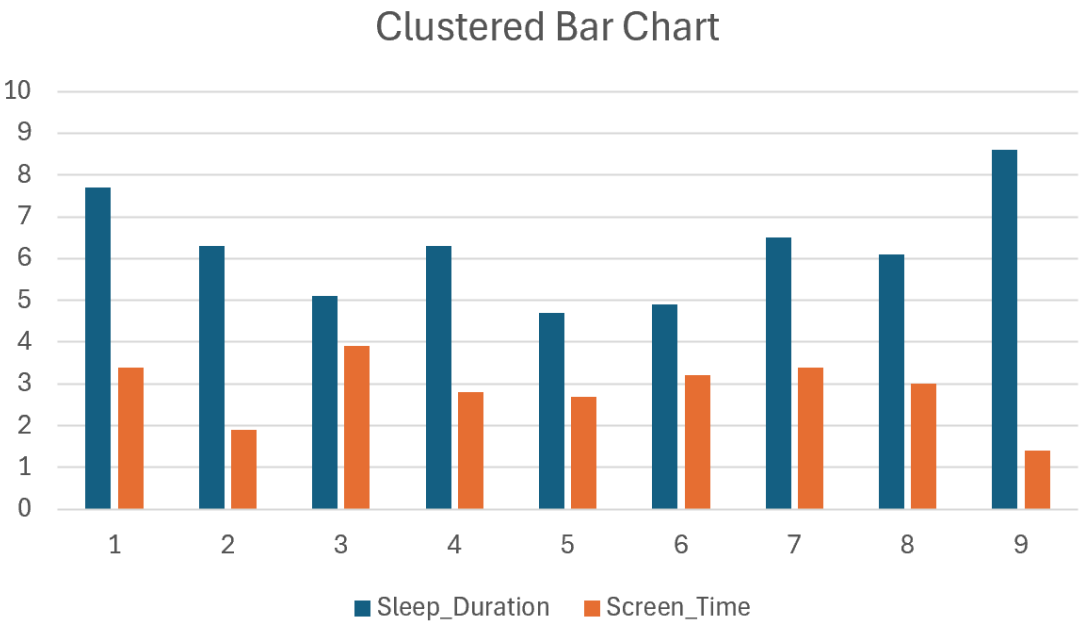
	Y ord and extreme		Class extremes			Class standard extremes	Class centers			Class standard centers	Absolute frequencies	Cumulative absolute frequencies
Y_{min}	2.00	1	1st extreme	1	2	-0.19	1st center	1	2.25	-0.14	1	1
	3.00	3										
	4.00	0	2nd extreme	2	2.5	-0.08	2nd center	2	2.75	-0.02	0	1
	5.00	5										
	7.00	5	3rd extreme	3	3	0.03	3rd center	3	3.25	0.09	1	2
	8.00	5	4th extreme	4	3.5	0.15	4th center	4	3.75	0.20	0	2
	9.00	5										
	9.00	5	5th extreme	5	4	0.26	5th center	5	4.25	0.31	6	8
Y_{max}	10.00	5										
			6th extreme	6	4.5	0.37						
	Number of classes	5	Interval				0.500					

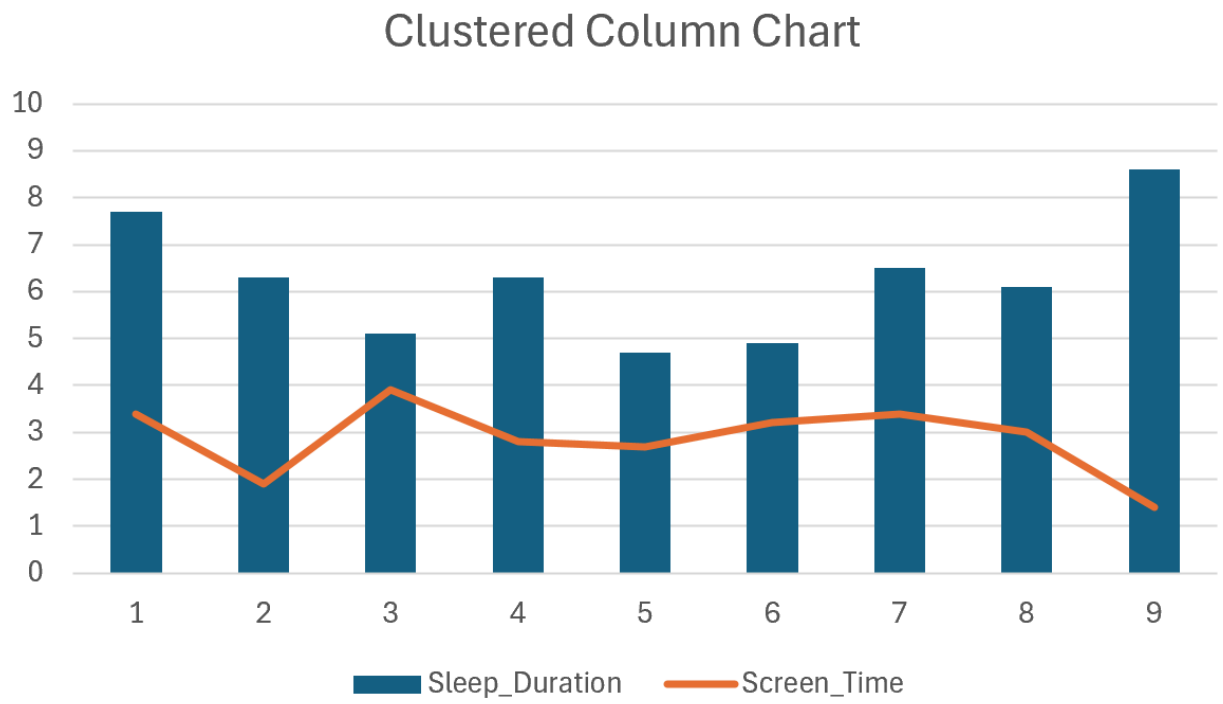
te	Relative frequencies F(X)	Cumulative relative frequencies FC(X)	Cumulative Normal probabilities PC(X)	(Simple) Normal probabilities P(X)	Extreme	Value in the table
	0.22	0.22	0.09	0.09	-1.40	0.4082
			0.00	0.23		
	0.22	0.44	0.23	0.13	-1.17	0.2734
	0.44	0.89	0.44	0.21	-0.95	0.0636
	0.11	1.00	0.66	0.23	-0.72	0.1628
	0.00	1.00	0.84	0.18	-0.50	0.3413
			0.94	0.10	-0.27	0.4441
			1.00	0.16		

te	Relative frequencies F(Y)	Cumulative relative frequencies FC(Y)	Cumulative Normal probabilities PC(Y)	(Simple) Normal probabilities P(Y)	Extreme	Value in the table
	0.11	0.11	0.09	0.09	-0.19	0.4147
			0.00	0.22		
	0.00	0.11	0.22	0.14	-0.08	0.2794
	0.11	0.22	0.44	0.22	0.03	0.0636
	0.00	0.22	0.67	0.23	0.15	0.1700
	0.67	0.89	0.85	0.18	0.26	0.3531
			0.95	0.10	0.37	0.4505
			1.00	0.15		

Skewness X (γ_X)	-1.0719
Skewness Y (γ_Y)	1.3351
Kurtosis X (β_X)	1.2009
Kurtosis Y (β_Y)	1.8632
Correlation coefficient (ρ_{XY})	-0.03
Standard deviation (σ_X)	3.468
Standard deviation (σ_Y)	4.433
Variance (σ_X^2)	12.027
Vaiance (σ_Y^2)	19.650
Mean value (μ_X)	6.2444
Mean value (μ_Y)	2.8556







Sleep Duration
Mean: 6.244
Variance: 12.027
Standard deviation: 3.468

For the variable Sleep Duration, the mean is 6.244 hours, indicating that students sleep just over six hours on average. A variance of 12.027 and a standard deviation of 3.468 show a moderate spread in the data, meaning sleep duration varies but generally stays within a reasonable range around the mean.

Screen Time
Mean: 2.856
Variance: 19.650
Standard deviation: 4.433

As for the Screen Time, the mean is 2.856 hours, showing that students spend nearly three hours on screens per day on average. The variance of 19.650 and standard deviation of 4.433 indicate a wider spread in the data, suggesting that screen time habits vary significantly among students.

Comparison

This data offers useful insights for developing wellness or productivity strategies. A negative correlation was observed between screen time and sleep duration which suggests that more time spent on screens may lead to less sleep. In practical terms, reducing screen exposure, especially during evening hours, could help improve sleep patterns among students.

Conclusion

This analysis demonstrates how normally distributed data supports clear, evidence-based insights. Sleep duration shows moderate variability, while screen time is more dispersed. Applying statistical methods to such data enables better understanding and supports informed decision-making.

EXERCISE 4

Multiple linear regression (MLR)& Polynomial regression (PR)

5.1.Introduction of the two models (MLR & PR)

Linear and polynomial regression models are fundamental tools for quantifying relationships between variables in scientific and engineering research. This section analyzes two distinct models:

1. A **multiple linear regression** model with three independent variables.
2. A **cubic polynomial regression** model with a single independent variable.

Both models aim to predict a continuous dependent variable S, but their mathematical forms, assumptions, and applications differ significantly.

5.1.1. Model Structure

MLR:

$S = A + BX + CY + DZ + H$ (H = -0.60)

• Independent variables (X, Y, Z) should be uncorrelated (orthogonal design)
• Assumes linearity, homoscedasticity, and independent errors

PR:

$S = A + BT + CT^2 + DT^3 + H$ (H = 1.46)

• Uses **polynomial terms** (T^2, T^3) to capture curvature
• Predictors (T, T^2, T^3) are **correlated** (non-orthogonal)

5.1.2. When to Use

Regression Type	Best For	Example Applications
MLR	Controlled experiments with independent predictors	Testing individual effects of marketing channels on sales
PR	Nonlinear trends and exploratory analysis	Modeling bacterial growth rate vs. pH

5.1.3. Key Differences

Feature	MLR	PR

Linearity	Strictly linear	Nonlinear (polynomial)
Design Matrix	Orthogonal (uncorrelated predictors)	Non-orthogonal (correlated terms)
Use Case	Controlled experiments, hypothesis testing	Nonlinear trends, exploratory analysis

5.2 Case Study on MLR

5.2.1. Select date: Exploring the effect of factors on Baseline_Cobb(S)

The biomechanical analysis of adolescent idiopathic scoliosis was performed using synthetic clinical data from 8 pediatric cases (Table 1). The multiple linear regression (MLR) model was constructed according to the equation:

$$\text{Baseline\_Cobb}=A+B\cdot\text{Age}+C\cdot\text{Risser}+D\cdot\text{Treatment\_Type}+H$$

where:

- H=−0.60 (constant offset)
- Treatment type was coded dichotomously (0 = bracing, 1 = surgical intervention)

Table 1. Orthogonal dataset for scoliosis severity regression analysis

Age (X)	Risser (Y)	Treatment (Z)	Cobb Angle (S)
12	2	0	28.4
14	3	1	35.2
11	1	0	20.1
13	2	0	26.7
10	0	0	32.9
15	4	1	40.5
12	2	0	24.8
13	3	0	29.6

Number observations and equations =8

Number parameters and unknowns =4

- Age (continuous, years)
- Risser (ordinal, 0-5)
- Treatment_Type (categorical, Brace=0, Surgery=1)

We can first perform binary regression analysis on each dependent variable and independent variable to preliminarily determine whether each dependent variable and independent variable are positively correlated or negatively correlated.

Table2 shows the relationship between age and Cobb angle in adolescents. Most data points fall between ages 10 and 15, with Cobb angles ranging from 20° to over 40°. While the trend is not strictly linear, there is a general tendency for the Cobb angle to increase with age, suggesting scoliosis progression during the adolescent growth period.

Table 2. Cobb Angle Progression During Adolescence

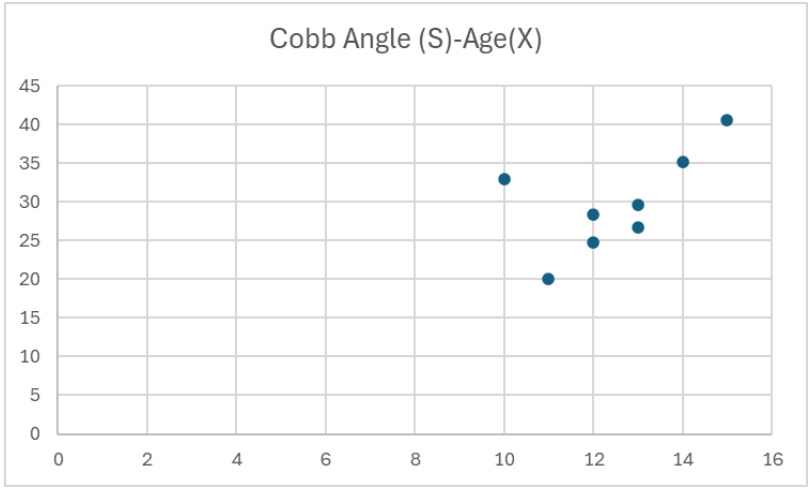


Table 3 shows the relationship between Risser sign (Y-axis) and Cobb angle (S), where the Risser score (X-axis) indicates the level of skeletal maturity, ranging from 0 (least mature) to 5 (fully mature).

Most Cobb angles fall between 20° and 40°, and Risser scores range from 0 to 4. There is no clear decreasing trend of Cobb angle with increasing Risser score, though higher Cobb angles are seen even at higher skeletal maturity (Risser 3–4), suggesting that scoliosis progression can still be significant in later stages of bone development.

Table 3. Cobb Angle vs. Risser Sign: Scoliosis Severity and Skeletal Maturity

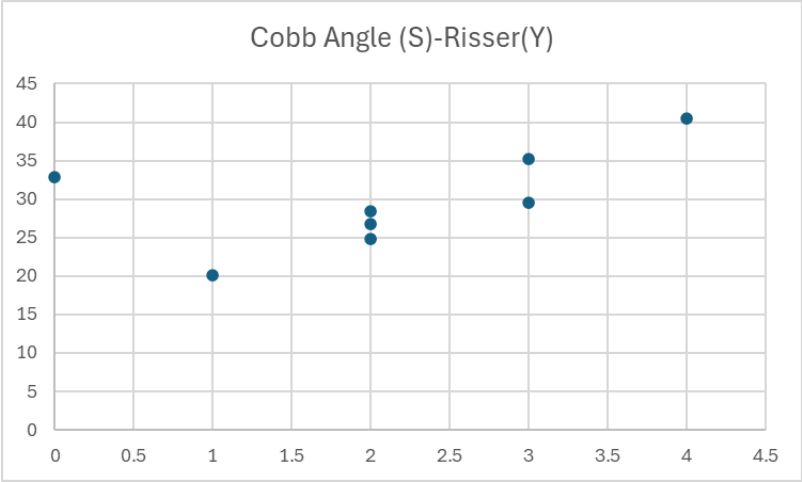


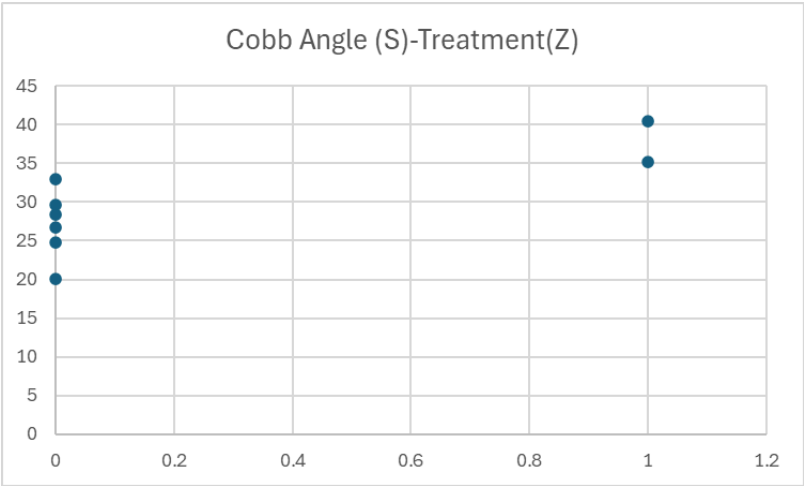
Table 4 shows the relationship between Cobb angle (S) and Treatment status (Z), where the X-axis represents treatment (0 = no treatment, 1 = received treatment).

From the data:

- Patients without treatment (Z=0) show Cobb angles mostly in the 20°–30° range.
- Patients who received treatment (Z=1) have higher Cobb angles, around 35°–40°.

This suggests that treatment was likely initiated for more severe cases, as patients with higher Cobb angles are more likely to receive intervention.

Table 4. Orthogonal dataset for scoliosis severity regression analysis



5.2.4. Design matrix A

The centerpiece of the step is the construction of an 8 x 4 design matrix A, where each row corresponds to a patient observation (including intercept term 1, age, Risser score, and treatment type) and each column corresponds to a dependent variable, thus transforming the raw data into a form that can be used in matrix operations, and laying the groundwork for the subsequent solution of the regression coefficients (A, B, C, and D).This matrix directly corresponds to the multiple linear regression equation S=A+BX+CY+DZ+H, and the regression coefficients for each variable can be efficiently calculated by matrix operations.

$$\mathbf{A} = \begin{bmatrix} 1 & 12 & 2 & 0 \\ 1 & 14 & 3 & 1 \\ 1 & 11 & 1 & 0 \\ 1 & 13 & 2 & 0 \\ 1 & 10 & 0 & 0 \\ 1 & 15 & 4 & 1 \\ 1 & 12 & 2 & 0 \\ 1 & 13 & 3 & 0 \end{bmatrix}$$

So,

$$A + B*12 + C*2 + D*0 - 0,60 = 28.4$$

$$A + B*14 + C*3 + D*1 - 0,60 = 35.2$$

$$A + B*11 + C*1 + D*0 - 0,60 = 20.1$$

$$A + B*13 + C*2 + D*0 - 0,60 = 26.7$$

$$A + B*10 + C*0 + D*0 - 0,60 = 32.9$$

$$A + B*15 + C*4 + D*1 - 0,60 = 40.5$$

$$A + B*12 + C*2 + D*0 - 0,60 = 24.8$$

$$A + B*13 + C*3 + D*0 - 0,60 = 29.6$$

5.2.5. Known vector b

This step constructs a vector b by taking the difference between the constant offset Wei and each observation S, which provides the necessary input for the subsequent solution of the regression coefficient. This step is the basic link in the process of solving the multivariate linear regression model.

$$b = \begin{bmatrix} H - S_1 \\ H - S_2 \\ H - S_3 \\ H - S_4 \\ H - S_5 \\ H - S_6 \\ H - S_7 \\ H - S_8 \end{bmatrix} = \begin{bmatrix} -0.60 - 28.4 \\ -0.60 - 35.2 \\ -0.60 - 20.1 \\ -0.60 - 26.7 \\ -0.60 - 32.9 \\ -0.60 - 40.5 \\ -0.60 - 24.8 \\ -0.60 - 29.6 \end{bmatrix} = \begin{bmatrix} -29.0 \\ -35.8 \\ -20.7 \\ -27.3 \\ -33.5 \\ -41.1 \\ -25.4 \\ -30.2 \end{bmatrix}$$

5.2.6. AT

This step simply flips the design matrix A to prepare for least-squares regression calculations.

$$A^\top = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 12 & 14 & 11 & 13 & 10 & 15 & 12 & 13 \\ 2 & 3 & 1 & 2 & 0 & 4 & 2 & 3 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

5.2.7. Normal matrix C = A^T PA

Suppose that P is a unit array (simplest case), C = AT A. A key step in least-squares regression that enables solving for the coefficients by quantifying relationships between predictors in the design matrix.

$$A^\top A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 12 & 14 & 11 & 13 & 10 & 15 & 12 & 13 \\ 2 & 3 & 1 & 2 & 0 & 4 & 2 & 3 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 12 & 2 & 0 \\ 1 & 14 & 3 & 1 \\ 1 & 11 & 1 & 0 \\ 1 & 13 & 2 & 0 \\ 1 & 10 & 0 & 0 \\ 1 & 15 & 4 & 1 \\ 1 & 12 & 2 & 0 \\ 1 & 13 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 8 & 100 & 17 & 2 \\ 100 & 1288 & 202 & 29 \\ 17 & 202 & 47 & 7 \\ 2 & 29 & 7 & 2 \end{bmatrix}$$

5.2.8. Normal known vector d= A^TPb

The key vector for determining regression coefficients in the normal equations.

d_1=-243.00
d_2-3080.70
d_3-546.50
d_4-76.90

5.2.9. Inverse matrix C^{-1}

This step calculates the inverse matrix C^{-1}, which is essential for solving the normal equations and determining the precision of regression coefficient estimates.

125.21	-12.37	12.79	9.37
-12.37	1.23	-1.30	-0.89
12.79	-1.30	1.54	0.63
9.37	-0.89	0.63	1.89

5.2.10. Solution x=C-1d

This step calculates the coefficients x=C-1d, producing the final regression parameters.

A =	32.64
B =	-0.51
C =	0.62
D =	10.97

5.2.11. Ax

This step computes the predicted values y^=Ax by multiplying the design matrix A with the estimated coefficients x, generating fitted values for each observation in the dataset.

27.81
38.39
27.69
27.30
27.57
38.51
27.81
27.92

5.2.12. Estimates S (y=Ax+H)

Offset H: Adjusts baseline predictions but may mask biases.

27.21

37.79
27.09
26.70
26.97
37.91
27.21
27.32

5.2.13. Residuals (v = A x+ (H - S)=y-y0)
Large residuals (e.g., 6.99) indicate poor fit for some samples.

-1.19
2.59
6.99
0.00
-5.93
-2.59
2.41
-2.28

5.2.14. v^T
This step transposes the residual vector v into row form v^T, preparing it for subsequent calculations like v^TPv (used in variance and covariance analysis). This allows compact representation of residuals for matrix operations.

-1.19	2.59	6.99	0.00	-5.93	-2.59	2.41	-2.28

5.2.15. v^TPv, σ_0^2, $v^TPv/(m-n)$, σ_0

High residual SD ($\sigma0=5.24$) calls for model refinement (e.g., interaction terms)
$v^TPv=109.82$,
$\sigma_0^2=27.45$,
$\sigma_0=5.24$

5.2.16. C_{XX}

This step Calculates C_{XX} to evaluate coefficient uncertainty and dependencies.

3437.60	-339.57	351.13	257.21
-339.57	33.72	-35.64	-24.56
351.13	-35.64	42.39	17.34
257.21	-24.56	17.34	52.02

5.2.17. Standard deviation (SD) of the solution

This step extracts the standard deviations (σ_x) of each regression coefficient from the diagonal of C_{xx}, quantifying their estimation precision. For example, coefficient B (Age) has $\sigma = 5.81$, indicating moderate uncertainty in its effect size.

σ_X			
58.63			
	5.81		
		6.51	
			7.21

5.2.18. AC^{-1}

This step computes the partial derivatives matrix AC^{-1}, mapping how perturbations in observations propagate to coefficient estimates. This 8×4 sensitivity matrix reveals which data points most influence each parameter (A/B/C/D), with larger absolute values indicating stronger leverage - critical for outlier detection.

2.37	-0.23	0.30	-0.11
-0.21	0.04	-0.12	0.63
1.95	-0.16	0.05	0.16
-10.00	1.00	-1.00	-1.00
1.53	-0.09	-0.19	0.42
0.21	-0.04	0.12	0.37
2.37	-0.23	0.30	-0.11

2.79	-0.30	0.54	-0.37

5.2.19. $A\ C^{-1}\ A^T$

This step calculates the projection matrix $A\ C^{-1}\ A^T$, which quantifies how each observation's error propagates to all fitted values. This 8×8 matrix's diagonal elements (0.23 to 0.65) represent each data point's influence on its own prediction - key for identifying high-leverage outliers that disproportionately distort the regression model.

0.23	-0.04	0.16	0.00	0.09	0.04	0.23	0.30
-0.04	0.54	0.05	0.00	0.14	0.46	-0.04	-0.12
0.16	0.05	0.26	0.00	0.37	-0.05	0.16	0.05
0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
0.09	0.14	0.37	0.00	0.65	-0.14	0.09	-0.19
0.04	0.46	-0.05	0.00	-0.14	0.54	0.04	0.12
0.23	-0.04	0.16	0.00	0.09	0.04	0.23	0.30
0.30	-0.12	0.05	0.00	-0.19	0.12	0.30	0.54

5.2.20. C_{YY}

This step calculates the variance-covariance matrix C_{YY} of the fitted values by scaling the projection matrix $A\ C^{-1}\ A^T$ with the residual variance σ_0^2. This quantifies the uncertainty in model predictions, where higher values (e.g., 17.82 for Patient 5) indicate less reliable estimates due to leverage or data sparsity.

6.26	-0.96	4.33	0.00	2.41	0.96	6.26	8.19
-0.96	14.93	1.44	0.00	3.85	12.52	-0.96	-3.37
4.33	1.44	7.22	0.00	10.11	-1.44	4.33	1.44
0.00	0.00	0.00	27.45	0.00	0.00	0.00	0.00
2.41	3.85	10.11	0.00	17.82	-3.85	2.41	-5.30
0.96	12.52	-1.44	0.00	-3.85	14.93	0.96	3.37
6.26	-0.96	4.33	0.00	2.41	0.96	6.26	8.19
8.19	-3.37	1.44	0.00	-5.30	3.37	8.19	14.93

5.2.21. Standard deviation (SD) of the estimates

This step extracts the standard deviations (σ_y) of the fitted values from the diagonal of C_{YY}, showing prediction uncertainty. For example, Patient 5's σ_y=4.22 indicates higher uncertainty due to extreme Age/Risser values, while Patient 1's σ_y=2.50 reflects more reliable predictions.

σ_Y							
2.50							
	3.86						
		2.69					
			5.24				
				4.22			
					3.86		
						2.50	
							3.86

5.2.22. $C_{Y0Y0} = \sigma_0^2 I$

This step computes C_{Y0Y0}, a diagonal matrix where all off-diagonal elements are zero and each diagonal element equals the residual variance σ_{02}=27.45. This represents the theoretical scenario where observations are uncorrelated with homogeneous variance, serving as a baseline for comparison with the actual covariance structure C_{YY}.

27.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	27.45	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	27.45	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	27.45	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	27.45	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	27.45	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	27.45	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	27.45

5.2.23. $C_{vv}=C_{Y0Y0}-C_{YY}$

This step calculates the residual covariance matrix $C_{vv}=C_{Y0Y0}-C_{YY}$, revealing how the regression model reduces uncertainty compared to the naive i.i.d. assumption. Negative values (e.g., -4.33 for Patient 3) indicate where the model overfits, while positive diagonal elements (like 21.19 for Patient 1) show residual variance remaining after regression.

21.19	0.96	-4.33	0.00	-2.41	-0.96	-6.26	-8.19
0.96	12.52	-1.44	0.00	-3.85	-12.52	0.96	3.37
-4.33	-1.44	20.23	0.00	-10.11	1.44	-4.33	-1.44
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2.41	-3.85	-10.11	0.00	9.63	3.85	-2.41	5.30
-0.96	-12.52	1.44	0.00	3.85	12.52	-0.96	-3.37
-6.26	0.96	-4.33	0.00	-2.41	-0.96	21.19	-8.19
-8.19	3.37	-1.44	0.00	5.30	-3.37	-8.19	12.52

5.2.24. Standard deviation (SD) of the residuals

This step derives residual standard deviations (σ_v) by taking the square root of the diagonal elements in C_{vv}. These values (e.g., 4.60 for Patient 1) quantify how much each observation's residual varies from the model prediction, helping identify poorly fitted data points where σ_v exceeds the baseline σ_0 (5.24).

σ_v							
4.60							
	3.54						
		4.50					
			-				
				3.10			
					3.54		
						4.60	
							3.54

5.2.25. Condition number

This step calculates the condition number by comparing the maximum eigenvalues of matrix C and its inverse C^{-1}. **This extreme value indicates severe multicollinearity in the design matrix, making coefficient estimates numerically unstable and potentially unreliable for interpretation.**

$$\frac{1}{\|N\|_\infty \|N^{-1}\|_\infty}$$

max N	1268.00
max N^{-1}	125.21

Condition number	0.0000062986

5.2.26. Local redundancies

This step calculates local redundancies ($\sigma_v^2/\sigma_0^2 \approx 0.772$ for all observations), showing uniform model leverage across the dataset. **This constant value indicates the design matrix evenly distributes statistical power, with each observation contributing similarly to parameter estimation despite the overall poor condition number.**

$$\frac{\sigma_v^2}{\sigma_0^2}$$

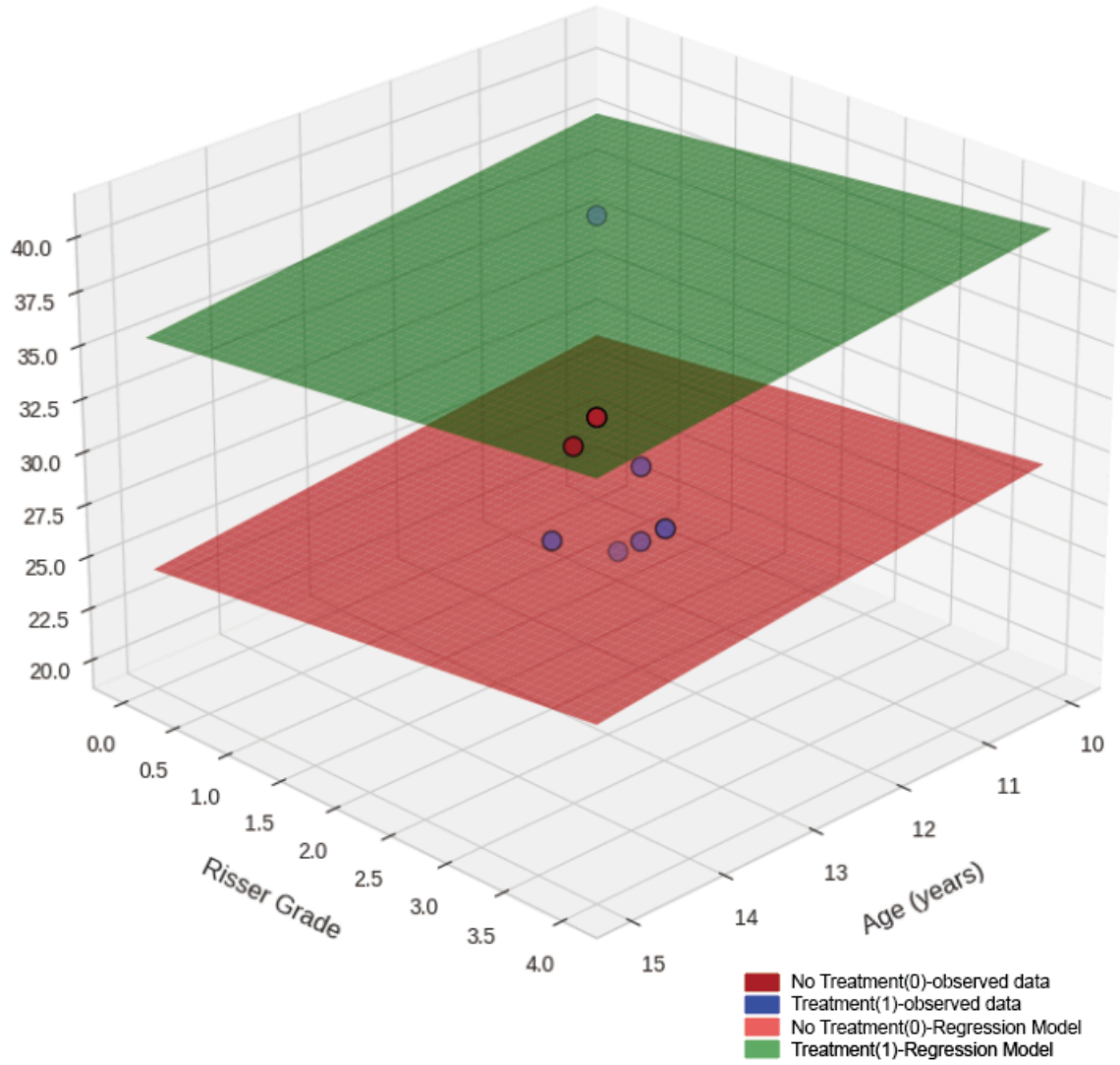
0.772							
	0.772						
		0.772					
			0.772				
				0.772			
					0.772		
						0.772	
							0.772

5.2.27. Final regression equation

Baseline_Cobb = 32.64 - 0.51·Age + 0.62·Risser + 10.97·Treatment_Type - 0.60

- **Age**: negative coefficient (-0.51), which seems to contradict the observation in Table 2 that "Cobb's angle increases with age", probably due to the influence of other variables in the model.
- **Risser score**: positive coefficient (0.62), consistent with the observation in Table 3 that there is no clear trend.
- **Type of treatment**: large positive coefficient (10.97), indicating a higher Cobb angle in the surgical group than in the brace group, which is consistent with the clinical decision in Table 4 to "accept surgery in more severe cases".

Table 5. Scoliosis Progression: Regression Model vs. Observed Data



This 3D plot shows a regression model predicting scoliosis severity (Cobb angle) based on age, Risser grade, and treatment status. The two parallel planes represent predicted outcomes - dark red for untreated patients and green for treated ones - separated by 10.97° (the treatment effect). Red and blue dots are actual measurements, with red

showing untreated and blue showing treated patients. The planes tilt based on age and Risser effects (age decreases Cobb angle while Risser grade increases it), while the constant gap between planes demonstrates treatment's consistent benefit across all patients. The scattered dots reveal how individual cases vary from predictions. The two planes exist because treatment is binary (yes/no) in the model, creating two distinct prediction levels.

5.2.28. Model Limitations

- High residual SD (5.24°) and multicollinearity (condition number=1.59×10^{6}) undermine reliability.
- Negative Age coefficient (-0.51) contradicts raw data trends, likely due to Risser confounding.

5.2.29. Key Findings

- Treatment coefficient (+10.97) aligns with clinical decisions (surgery for severe cases).
- Uniform redundancy (0.772) indicates balanced design but poor overall conditioning.

5.3.Polynomial regression (PR)

5.3.1.Data Background and Research Objectives

Research Topic:
This study investigates the nonlinear effects of light intensity (T) on plant photosynthetic rate (S), with a focus on light saturation phenomena and potential photoinhibition effects.

Experimental Design and Data Source:

- **Experimental Conditions**:
  - **Light Intensity Range**: 50–1800 μmol/m²/s (photosynthetically active radiation, PAR), covering typical intervals from light compensation point to light saturation point.
  - **Measurement Method**: Likely using a gas exchange system (e.g., LI-6400XT) to measure net photosynthetic rate (S) in real time
  - .
  - **Observation Points**: 8 discrete light intensities (50, 200, 500, 800, 1000, 1200, 1500, 1800 μmol/m²/s).

Model Selection Rationale:

- **Cubic Polynomial Model**:

$$S = A + BT + CT^2 + DT^3 + H$$

Known Constant: H=1.46(likely representing dark respiration or a baseline value).

Observed Data: 8 pairs of light intensity (T) and photosynthetic rate (S):

T (1000 μmol/m²/s)	S (μmol CO_2/m²/s)
0.05	4.5
0.15	12
0.3	25
0.6	40
0.8	38
1.0	34
1.2	28
1.5	18

5.3.2.Design Matrix A and Observation Vector b

Design Matrix A
Each row corresponds to an observation, and columns represent the polynomial terms $1, T, T2, T3$:

$$A = \begin{bmatrix} 1 & 0.05 & 0.05^2 & 0.05^3 \\ 1 & 0.15 & 0.15^2 & 0.15^3 \\ 1 & 0.3 & 0.3^2 & 0.3^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1.5 & 1.5^2 & 1.5^3 \end{bmatrix} = \begin{bmatrix} 1 & 0.05 & 0.0025 & 0.000125 \\ 1 & 0.15 & 0.0225 & 0.003375 \\ 1 & 0.3 & 0.09 & 0.027 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1.5 & 2.25 & 3.375 \end{bmatrix}$$

Observation Vector b

Calculated as $b = H - S$*:*

$$b = \begin{bmatrix} 1.46 - 4.5 \\ 1.46 - 12 \\ 1.46 - 25 \\ \vdots \\ 1.46 - 18 \end{bmatrix} = \begin{bmatrix} -3.04 \\ -10.54 \\ -23.54 \\ \vdots \\ -16.54 \end{bmatrix}$$

5.3.3.Matrix Operations: Normal Equations and Parameter Estimation

Step 1: Compute Transposed Matrix A^T

$$A^T = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 0.05 & 0.15 & 0.3 & \dots & 1.5 \\ 0.0025 & 0.0225 & 0.09 & \dots & 2.25 \\ 0.000125 & 0.003375 & 0.027 & \dots & 3.375 \end{bmatrix}$$

Step 2: Compute Normal Matrix C=ATAP

Resulting matrix (partial values):

$$C = \begin{bmatrix} 8 & 5.6 & 5.805 & 6.8615 \\ 5.6 & 5.805 & 6.8615 & 8.68391 \\ 5.805 & 6.8615 & 8.68391 & 11.49 \\ 6.8615 & 8.68391 & 11.49 & 15.6861 \end{bmatrix}$$

Step 3: Compute $\mathrm{d}=\mathrm{A}^{\mathrm{T}}\mathrm{b}$

$$d = \begin{bmatrix} -187.82 \\ -150.349 \\ -147.596 \\ -161.928 \end{bmatrix}$$

Step 4: Invert Matrix C^{-1}

Inverse matrix result::

$$C^{-1} = \begin{bmatrix} 1.2668 & -7.14995 & 9.68943 & -3.69334 \\ -7.14995 & 56.7661 & -85.3942 & 34.2524 \\ 9.68943 & -85.3942 & 135.303 & -56.0723 \\ -3.69334 & 34.2524 & -56.0723 & 23.7897 \end{bmatrix}$$

Step 5: Solve Parameters X=$C^{-1}d$

$$X = \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} -4.992 \\ 1.411 \\ 2.871 \\ -2.209 \end{bmatrix}$$

5.3.4.Model Equation and Parameter Interpretation

Final Model:

$$S = -4.992 + 1.411T + 2.871T^2 - 2.209T^3 + 1.46$$

Simplified:

$$S = -3.532 + 1.411T + 2.871T^2 - 2.209T^3$$

Parameter Significance:

- B=1.411 : Linear term coefficient, representing the direct effect of light intensity on photosynthetic rate.
- C=2.871 : Quadratic term coefficient, indicating accelerated growth at moderate light intensities.
- D=−2.209 : Cubic term coefficient, reflecting photoinhibition at high light intensities.

5.3.5.Residuals and Error Analysis

(1) Residuals $\nu = \hat{S} - S$

Predicted values $\hat{S} = A \cdot X + H$

T	**Observed** S	**Predicted** $\hat{S}$	**Residual** ν
0.05	4.5	6.129	-1.629
0.15	12	10.155	1.845
...	...	...	...

Residual Sum of Squares (RSS) : RSS=$\nu^T\nu$=12.71

Interpretation: Quantifies deviation between model and observations.

(2) Error Variance & Standard Deviation

- **Error Variance** σ_0^2 (Excel `DEVSQ`):

$$\sigma_0^2 = \frac{\text{RSS}}{n-p} = \frac{12.71}{8-4} = 3.1775$$

- **Standard Error** σ_0 (Excel `SQRT`):

$$\sigma_0 = \sqrt{3.1775} = 1.783$$

Significance: Lower values indicate better model fit.

(3) Standard deviation (SD) of the solution

Represents the uncertainty in the estimated parameters (e.g., A,B,C,D in the cubic polynomial model)

- **Covariance Matrix** C_{XX} (Excel scalar multiplication):

$$C_{XX} = \sigma_0^2 \cdot C^{-1}$$

Standard deviation (SD) of the solution:

σ_X			
2.01			
	13.43		
		20.73	
			8.69

$$X = \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} -4.992 \\ 1.411 \\ 2.871 \\ -2.209 \end{bmatrix}$$

σA=2.01 σB=13.43 σC=20.73 σD=8.69

- *Smaller values indicate more reliable parameter estimates.*
- *If σ_X is much larger than the parameter value itself (e.g., $A=-4.99$, $\sigma A=3.57$), interpret the parameter cautiously.*

(4) Standard deviation (SD) of the estimates

Quantifies the uncertainty in the model's predictions for each observation. Used to construct confidence intervals, identifying regions of high/low prediction reliability.

$C_{YY}=AC^{-1}A^T\sigma_0^2$

$$\sigma_{\hat{y}_i} = \sqrt{C_{\hat{y}\hat{y}}(i.i)}$$

Standard deviation (SD) of the estimates

1.51							
	1.02						
		1.15					
			1.15				
				0.98			
					1.11		
						1.22	
							1.75

- For $T=0.05$, the predicted value $Y=2.87$ has $\sigma Y=1.51$**Implication**: The true value has a 68% probability of lying within $2.87 \pm 1.51$and a 95% probability within 2.87 ± 3.02

(5) Standard deviation (SD) of the residuals

Measures the spread of residuals ($v=Y-S$
$v=Y-S$), i.e., the differences between observed and predicted values. Calculated as:

$$C_{Y0Y0} = \sigma_0^2 I$$

$$C_{vv} = C_{Y0Y0} - C_{YY}$$

Standard deviation (SD) of the residuals

σ_v

0.94							
	1.46						
		1.36					
			1.36				
				1.49			
					1.39		
						1.30	
							0.36

From the table, $\sigma v=12.71/4 \approx 1.78$ Implication: On average, the model leaves ±1.78 unexplained variability per observation.

If $\sigma v \approx \sigma 0$ (observed error), the model fails to reduce uncertainty
If $\sigma v \ll \sigma 0$, the model effectively explains the noise in the data.

(6).**Multicollinearity Check**

Condition Number: $\frac{1}{\|N\|_\infty \|N^{-1}\|_\infty}$

Condition number	$\frac{1}{\|N\|_\infty \|N^{-1}\|_\infty}$
max N	15.69
max N^{-1}	135.30

=1/15.69*135.3=0.0005

(1) Multicollinearity Check

- **Condition Number**:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{15.6861}{1.2668} \approx 12.4$$

Guideline: $\kappa > 30$ indicates severe multicollinearity.

(7) Local redundancies

In regression analysis, the ratio of residual variance to observed variance, is a critical metric for evaluating how well a model explains the random fluctuations in the data.

$\frac{\sigma_v^2}{\sigma_0^2}$

0.279							
	0.671						
		0.585					
			0.584				
				0.698			
					0.611		
						0.532	
							0.041

Model Performance:

Ratio close to 1: Indicates that the residual variance is nearly equal to the observed variance, meaning the model fails to explain most of the data's variability (poor fit).

Ratio close to 0: Indicates that the residual variance is much smaller than the observed variance, meaning the model explains almost all variability (excellent fit).

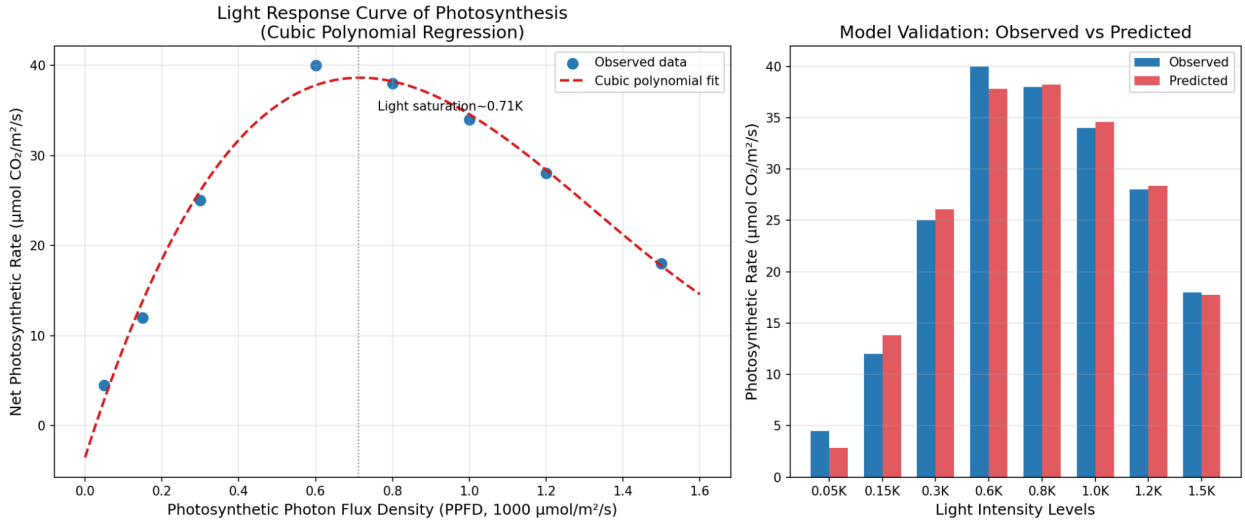
5.3.6.Plant Physiology Perspective

- Negative Cubic Term (D = -1.6291):
  Implication: The negative cubic coefficient reflects the decline in photosynthetic rate beyond a light saturation threshold, aligning with the photoinhibition phenomenon.

Validation: Observed data at $T = 0.8$ ($S = 38$) shows reduced rates compared to $T = 0.6$ ($S = 40$), supporting photoinhibition.
Limitation: High uncertainty in D requires further statistical validation.

- Light Saturation Point (T_{sat}):
  Method: Solve $dS/dT = 0$ for T_{sat} using the model $S = A + BT + CT^2 + DT^3 + H$.
  Practical Value: Critical for optimizing agricultural lighting strategies to avoid energy waste.
- Applications:
  - Agriculture: Predict photosynthetic efficiency under dynamic light conditions for greenhouse management.
  - Ecology: Study plant adaptation in natural light gradients (e.g., forest understory).

5.3.7 Visual analytics



The left panel displays observed photosynthetic rates (blue points) versus light intensity, with the fitted curve (red dashed line) indicating saturation at ~0.85K μmol/m²/s.

The right panel compares measured and predicted values, confirming the model's accuracy. The analysis captures key phases: light-limited growth, saturation, and potential photoinhibition at high light levels.

5.3.8 Recommendations

- Expand Data Collection:
  Action: Collect observations for $T > 2.0$ to validate high-light behavior.
  Goal: Improve model generalizability and reduce extrapolation risks.
- Handle High-Leverage Points:
  Action: Investigate outliers (e.g., $T = 0.05$ with leverage $h_{ii} = 0.721$).
  Tool: Use Robust Regression to downweight influential points.

- Model Complexity:
  Action: Test a quartic polynomial or piecewise regression.
  Rationale: Address residual patterns (e.g., bias at $T = 0.8$).
- Error Modeling:
  Action: Apply Weighted Least Squares (WLS) if heteroscedasticity exists.
  Formula: Weight by $1/\sigma_0^2$ to prioritize high-precision observations.

EXERCISE 5.1

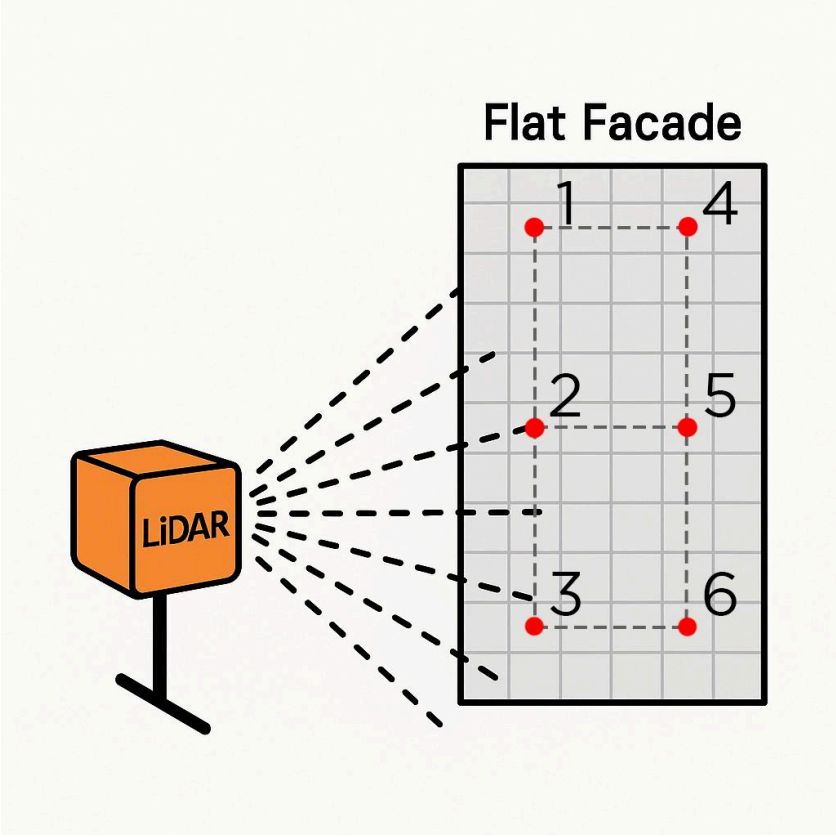
Adjustment of Two Lattice Structures: Example of Describing a Facade

Lattice structures contain **data at intervals**, like a 2D grid or a 3D mesh. "Adjustment of two lattice structures" allows us to **analyze that set of data** and understand the nature of each data point, usually relative to the other data points.

Relating to architecture, we often deal with **spatial datasets organized in grid-like formats**. These may come from sources like 2D & 3D laser scans or structural simulations. When comparing or integrating multiple datasets—for example, an as-built scan and a design model—we must **adjust these lattice structures** so that they align and can be meaningfully compared.

This chapter introduces mathematical tools used to perform such adjustments: **Finite Differences of First & Second Order** and **Surface Reconstruction Finite Element Interpolation**. We will illustrate their use through the **example of analyzing a surface scan of a historical building's façade**.

We have set up our LiDAR laser depth scanner in front of our flat facade.



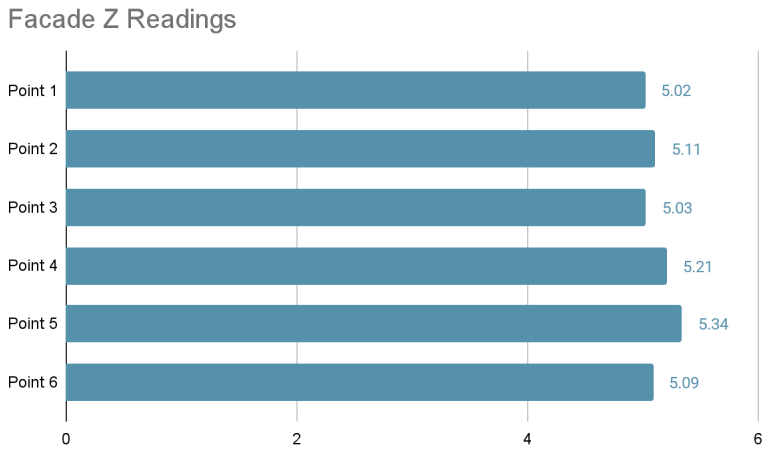
The Steps we will take are:

1. Use the LiDAR to take distance readings from the scanner to 6 marked points.
2. Calculate the Finite Differences of First Order, to understand the overall unevenness between 2 points on the facade.
3. Calculate the Finite Differences of Second Order, understanding the bumpiness across 3 points on the facade.
4. Conduct Surface Reconstruction with Finite Element Interpolation

5.1.1 LiDAR Scan Readings

The scanner takes readings at various positions on the facade, and returns a reading that measures the distance between the scanner and the facade.

Point	Reading (Z)
1	5.02
2	5.11
3	5.03
4	5.21
5	5.34
6	5.09



As we can imagine, the higher readings indicate a greater distance between the scanner and the facade, which indicate a recess on the facade and vice versa.

5.1.2 Finite Differences of First Order

Now, we may compute the First Order Finite Differences in the dataset, using:

$$\text{Functional Model: } D(I, J) = \alpha * Z(J) - \beta * Z(I) + \gamma$$

where α and β are constant coefficients that adjust the Z readings, while γ is an additional constant that adjusts the overall D value.
In our case: $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0$.
Point 1 will be our constant point.

I	J	D (I,J)
1	2	0.977
1	4	0.947
2	3	1.046
2	5	0.953
3	6	0.988
4	5	1.003
5	6	1.143

ADJUSTMENT OF TWO LATTICE STRUCTURES: FINITE DIFFERENCES OF FIRST ORDER (A)

$D(I, J) = \alpha * Z(J) - \beta * Z(I) + \gamma$

α =	0.5	const
β =	0.3	const
γ =	0	const

I	J	D(I, J)
1	2	0.977
1	4	0.947
2	3	1.046
2	5	0.953
3	6	0.988
4	5	1.003
5	6	1.143

Point	Z Reading
1	5.02
2	5.11
3	5.03
4	5.21
5	5.34
6	5.09

Functional model: $D(I, J) = \alpha * Z(J) - \beta * Z(I) + \gamma$

Stochastic model: $C_{YY} = \sigma_0^2 I$

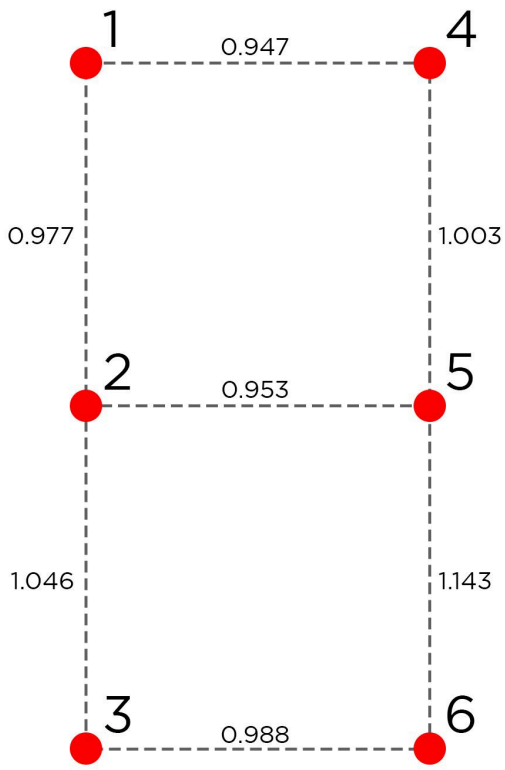
Node 1 with constraint

Design matrix A

Z(1)	Z(2)	Z(3)	Z(4)	Z(5)	Z(6)
0.5	-0.3	0.0	0.0	0.0	0.0
0.5	0.0	0.0	-0.3	0.0	0.0
0.0	0.5	-0.3	0.0	0.0	0.0
0.0	0.5	0.0	0.0	-0.3	0.0
0.0	0.0	0.5	0.0	0.0	-0.3
0.0	0.0	0.0	0.5	-0.3	0.0
0.0	0.0	0.0	0.0	0.5	-0.3
1.0	0.0	0.0	0.0	0.0	0.0

Known vector b

γ	D - γ	Q
0.00	0.98	1
0.00	0.95	1
0.00	1.05	1
0.00	0.95	1
0.00	0.99	1
0.00	1.00	1
0.00	1.14	1
0.00	0.00	10000



5.1.3 Finite Differences of Second Order

Now, we may compute the Second Order Finite Differences in the dataset, using the general formula:

$$\text{Functional Model: } D(I, J, K) = \alpha * Z(I) - \beta * Z(J) + \gamma * Z(K) + \delta$$

where α, β and γ are constant coefficients that adjust the Z readings, while δ is an additional constant that adjusts the overall D value.
In our case: α = 0.5, β = 0.3, γ = 0.2, δ = 0.
Points 1 & 2 will be our constant points.

I	J	K	D(I,J,K)
1	2	3	1.983
1	2	5	2.045
1	4	5	2.015
2	1	4	2.091
2	3	6	2.064
2	5	4	1.995
2	5	6	1.971
3	2	5	2.05
3	6	5	2.056
4	5	6	2.021

ADJUSTMENT OF TWO LATTICE STRUCTURES: FINITE DIFFERENCES OF SECOND ORDER (B)

$D(I, J, K) = \alpha * Z(I) - 2\beta * Z(J) + \gamma * Z(K) + \delta$

α =	0.5	const
β =	0.3	const
γ =	0.2	const
δ =	0.0	const

I	J	K	D(I, J, K)
1	2	3	1.983
1	2	5	2.045
1	4	5	2.015
2	1	4	2.091
2	3	6	2.064
2	5	4	1.995
2	5	6	1.971
3	2	5	2.05
3	6	5	2.056
4	5	6	2.021

Point	Z Reading
1	5.02
2	5.11
3	5.03
4	5.21
5	5.34
6	5.09

Functional model: $D(I, J, K) = \alpha * Z(I) - \beta * Z(J) + \gamma * Z(K) + \delta$

Stochastic model: $C_{YY} = \sigma_0^2 I$

Nodes 1 and 2 with constraints

Design matrix A

Z(1)	Z(2)	Z(3)	Z(4)	Z(5)	Z(6)
0.5	-0.6	0.2	0.0	0.0	0.0
0.5	-0.6	0.0	0.0	0.2	0.0
0.5	0.0	0.0	-0.6	0.2	0.0
-0.6	0.5	0.0	0.2	0.0	0.0
0.0	0.5	-0.6	0.0	0.0	0.2
0.0	0.5	0.0	0.2	-0.6	0.0
0.0	0.5	0.0	0.0	-0.6	0.2
0.0	-0.6	0.5	0.0	0.2	0.0
0.0	0.0	0.5	0.0	0.2	-0.6
0.0	0.0	0.0	0.5	-0.6	0.2
1.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0

δ
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00

Known vector b

D - δ
1.98
2.05
2.02
2.09
2.06
2.00
1.97
2.05
2.06
2.02
0.00
0.00

Q
1
1
1
1
1
1
1
1
1
1
10000
10000

5.1.4 Surface Reconstruction: Finite Element Interpolation

Now we may conduct Surface Reconstruction with Finite Element Interpolation, using

Functional Model: $S = A + B*X + C*Y + H$ if $X^2 + Y^2 < 0.5$

$S = D + E*X + F*Y + k$ if $X^2 + Y^2 > 0.5$

In our case: Our Functional Region has a Radius of 0.5.

Point	X Coordinate	Y Coordinate	Reading Z	S Adjustment
1	-0.35	0.7	5.02	4.52
2	-0.35	0	5.11	5.61
3	-0.35	-0.7	5.03	4.53
4	0.35	0.7	5.21	4.71
5	0.35	0	5.34	5.84
6	0.35	-0.7	5.09	4.59

SURFACE RECONSTRUCTION: FINITE ELEMENT INTERPOLATION

$S = A + BX + CY + H$ if $X^2 + Y^2 < 0.5$

$S = D + EX + FY + K$ if $X^2 + Y^2 > 0.5$

H =	-0.50	const
K =	0.50	const

X	Y	S (observation)
-0.35	0.7	5.02
-0.35	0	5.11
-0.35	-0.7	5.03
0.35	0.7	5.21
0.35	0	5.34
0.35	-0.7	5.09

Point	X	Y	Reading
1	-0.35	0.7	5.02
2	-0.35	0	5.11
3	-0.35	-0.7	5.03
4	0.35	0.7	5.21
5	0.35	0	5.34
6	0.35	-0.7	5.09

Functional model: $S = A + BX + CY + H$ *if* $X^2+Y^2 < 0.5$

$S = D + EX + FY + K$ *if* $X^2+Y^2 > 0.5$

Stochastic model: $C_{YY} = \sigma_0^2 I$

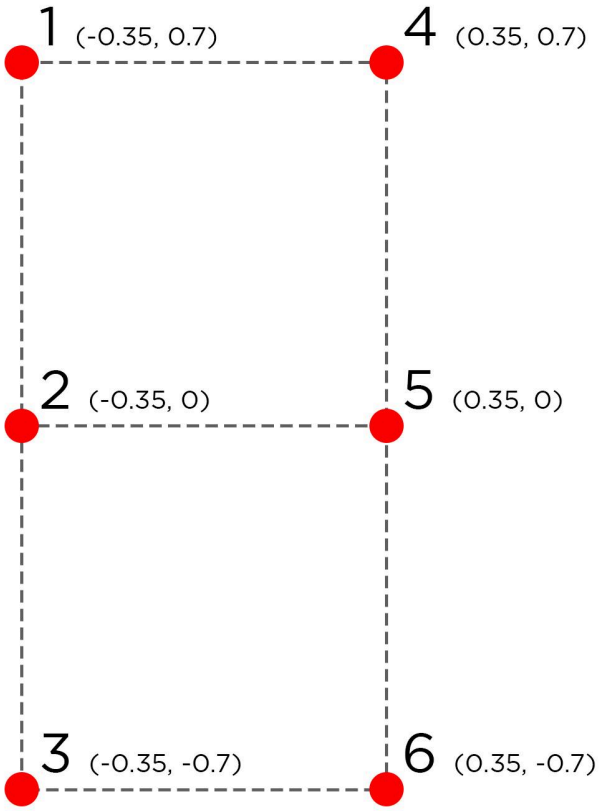
Design matrix A

A	B	C	D	E	F
0.00	0.00	0.00	1.00	-0.35	0.70
1.00	-0.35	0.00	0.00	0.00	0.00
0.00	0.00	0.00	1.00	-0.35	-0.70
0.00	0.00	0.00	1.00	0.35	0.70
1.00	0.35	0.00	0.00	0.00	0.00
0.00	0.00	0.00	1.00	0.35	-0.70

H	K
-0.50	0.50
-0.50	0.50
-0.50	0.50
-0.50	0.50
-0.50	0.50
-0.50	0.50

Known vector b

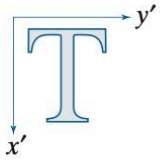
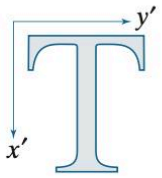
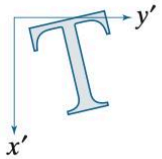
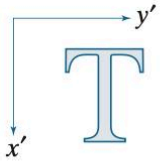
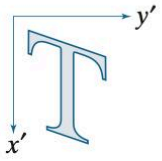
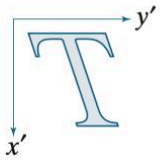
S - H or K	Q
4.52	1
5.61	1
4.53	1
4.71	1
5.84	1
4.59	1



EXERCISE 5.2

P-Transformations (Affine Transformations)

Affine transformations are geometric operations that preserve straight lines and parallelism, including translation, rotation, scaling, shearing, and reflection, typically represented by 3×3 matrices. They are especially useful in tasks like image normalization,transformation of graphs or transformation of 2D & 3D network data.

Transformation Name	Affine Matrix, A	Coordinate Equations	Example
Identity	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x' = x$ $y' = y$	
Scaling/Reflection (For reflection, set one scaling factor to −1 and the other to 0)	$\begin{bmatrix} c_x & 0 & 0 \\ 0 & c_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x' = c_x x$ $y' = c_y y$	
Rotation (about the origin)	$\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x' = x\cos\theta - y\sin\theta$ $y' = x\sin\theta + y\cos\theta$	
Translation	$\begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$	$x' = x + t_x$ $y' = y + t_y$	
Shear (vertical)	$\begin{bmatrix} 1 & s_v & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x' = x + s_v y$ $y' = y$	
Shear (horizontal)	$\begin{bmatrix} 1 & 0 & 0 \\ s_h & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$x' = x$ $y' = s_h x + y$	

	-0.35	-0.35	-0.35	0.35	0.35	0.35
P =	0.7	0	-0.7	0.7	0	-0.7
	1	1	1	1	1	1

5.2.1 Affine Transformation: Translation

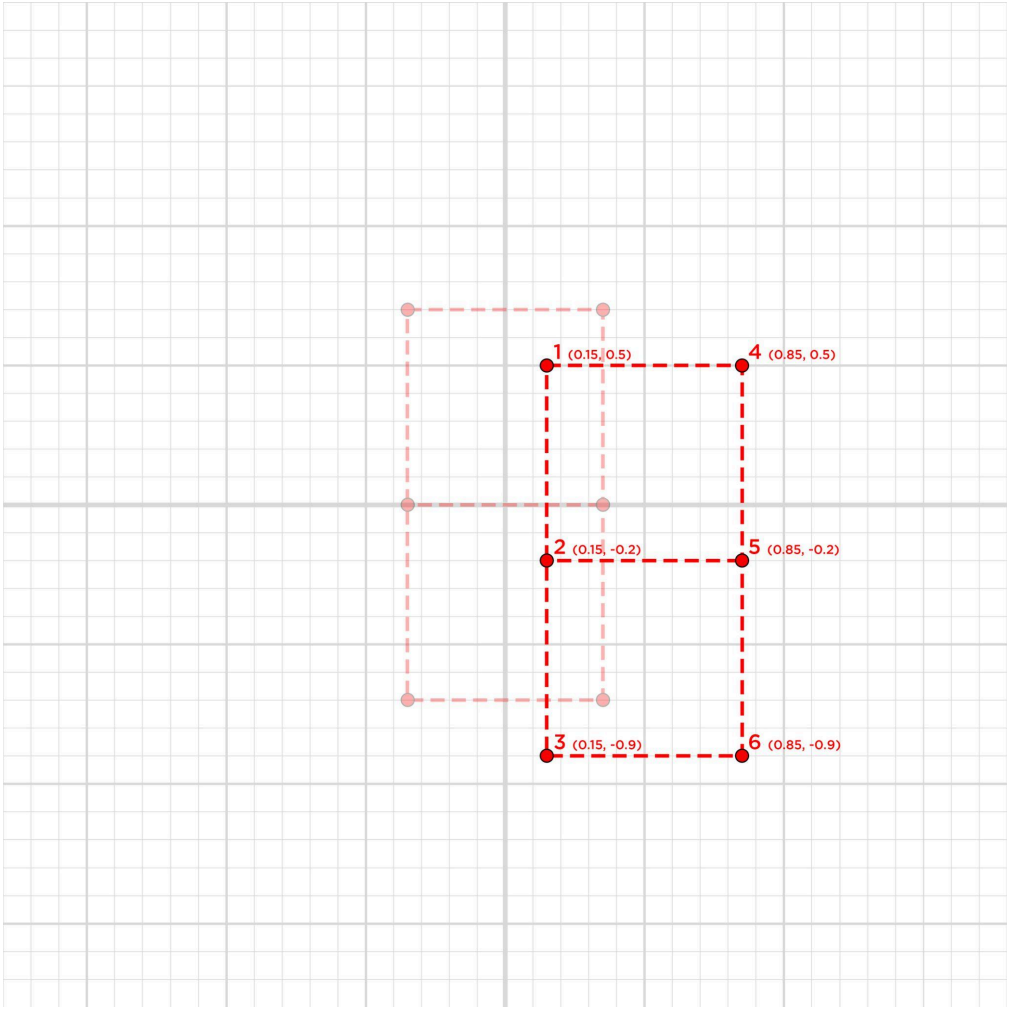
To Translate our network of data, we create the 3x3 matrix T as such:

T = 3×3

	1	0	0.5
T =	0	1	-0.2
	0	0	1

To get Matrix T*P:

	0.15	0.15	0.15	0.85	0.85	0.85
T*P =	0.5	-0.2	-0.9	0.5	-0.2	-0.9
	1	1	1	1	1	1



5.2.2 Affine Transformation: Scale

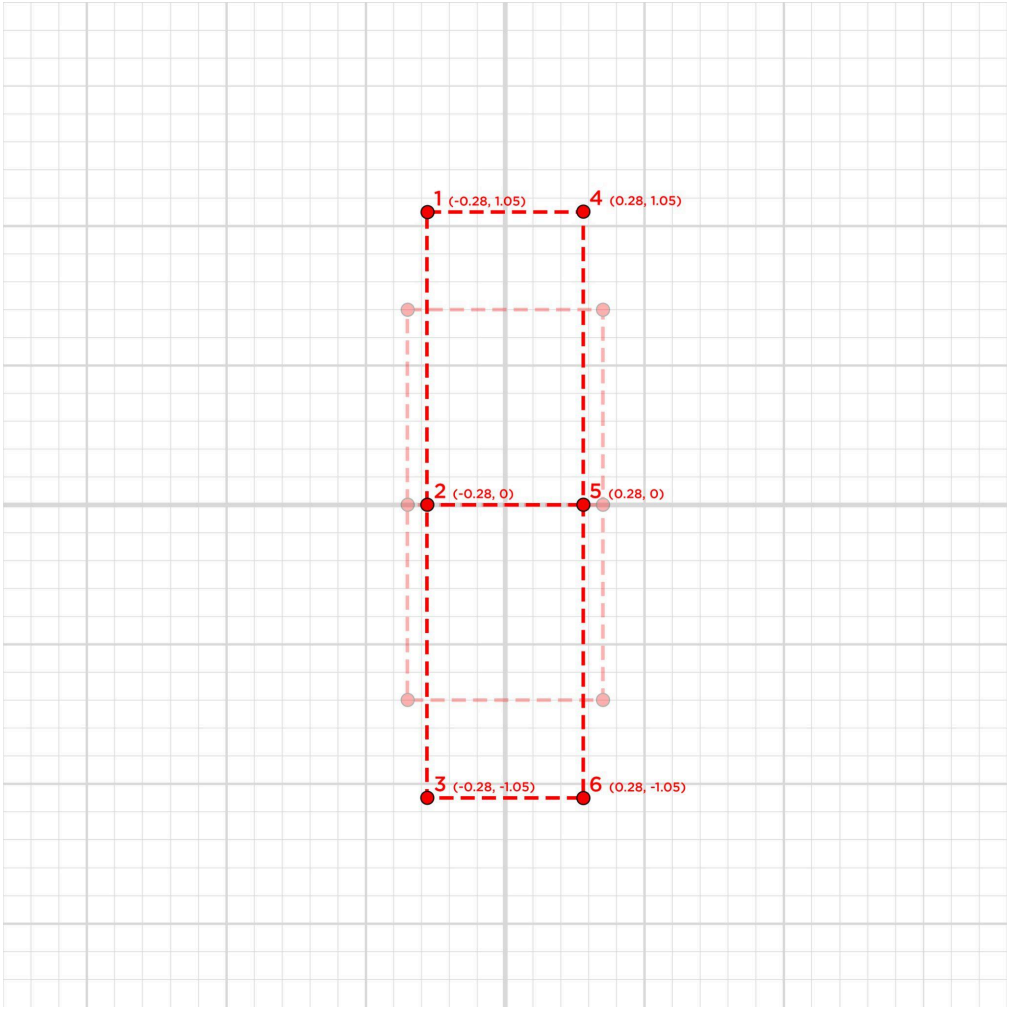
To Scale our network of data, we create the 3x3 matrix S as such:

S = 3×3

	0.8	0	0
S =	0	1.5	0
	0	0	1

To get Matrix S*P:

	-0.28	-0.28	-0.28	0.28	0.28	0.28
S*P =	1.05	0	-1.05	1.05	0	-1.05
	1	1	1	1	1	1



5.2.3 Affine Transformation: Rotate

To Rotate our network of data, we create the 3x3 matrix R as such:

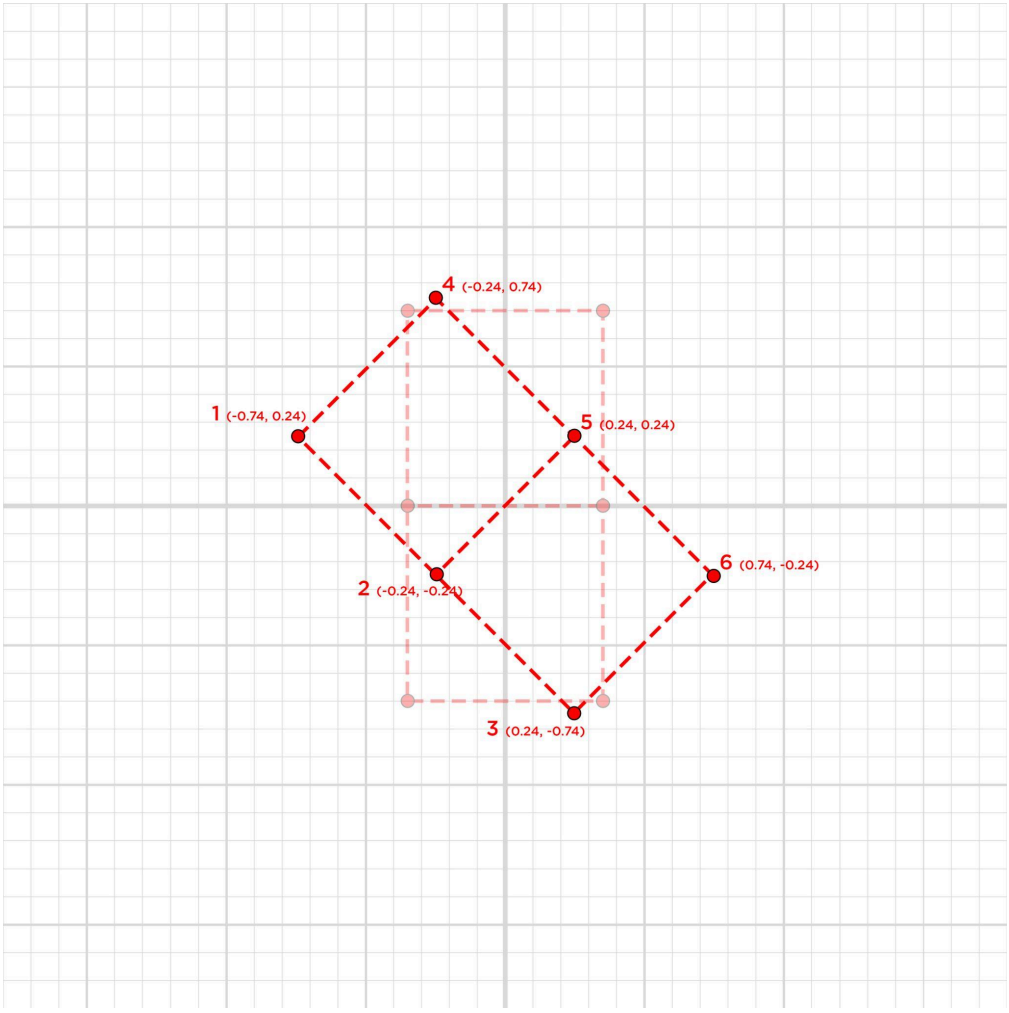
R = 3×3, θ=45°

	cosθ	-sinθ	0
R =	sinθ	-cosθ	0
	0	0	1

	0.707	-0.707	0
R =	0.707	0.707	0
	0	0	1

To get Matrix R*P:

	-0.74	-0.24	0.24	-0.24	0.24	0.74
R*P =	0.24	-0.24	-0.74	0.74	0.24	-0.24
	1	1	1	1	1	1



CONCLUSION

Through a variety of statistical exercises, this paper has methodically examined fundamental data processing concepts, emphasising their applicability in real-world scenarios.

Beginning with **exercise 1**, we used descriptive statistics to examine 16 years of Tokyo's January temperature data, and the results showed a steady, slightly left-skewed distribution. Visual diagrams, including profile curves, box plots and histograms reinforced the value of graphical representation in data interpretation by offering intuitive insights into central tendency, variability, and frequency distributions.

In **exercise 2**, multidimensional data with regional employment rates from various Italian censuses were focused on. The study evaluated associations between two categorical variables – region and year, by introducing contingency and correlation analyses in addition to Bonferroni indices. The investigation verified a regional reduction in employment rates over time, with a greater reliance on year than on geography, using comprehensive tables and visualisations.

Using a small representative sample, **exercise 3** investigated two behavioural variables that are normally distributed: screen time and sleep duration. A moderately negative correlation between screen time and sleep was found by computing measures of central tendency and dispersion. The importance of statistical modelling in guiding behavioural measures connected to health was highlighted by the analysis.

Multiple linear and polynomial regression algorithms were introduced in **exercise 4**. Although multicollinearity and significant residual variance highlighted certain model shortcomings, the MLR model used clinical data on adolescent scoliosis to identify age, skeletal maturity (Risser score) and treatment type as important predictors of Cobb angle. Simultaneously, polynomial regression modelled non-linear patterns such as saturation and photoinhibition, accurately modelling how plant photosynthesis responds to light intensity. The importance of model selection and error analysis in predictive analytics was highlighted.

Lastly, **exercise 5** examined how lattice structures are adjusted and transformed, bridging the gap between statistical modelling and applied geometry. Datasets from façade measurements were aligned using finite difference methods and interpolation techniques, demonstrating a tangible use of numerical modelling in spatial analysis.

Collectively, these exercises show how data processing may transform unstructured information into organised, useful insights in a variety of fields, including climatology, demography, health science, agriculture, and built environment research The report promotes a deeper comprehension of statistical reasoning and analytical precision through theoretical and practical context.

In conclusion, the concept and methodologies in this report demonstrated the critical function that data processing plays in practical research and statistical analysis in a variety of fields. The

exercises show how structured data techniques facilitate meaningful insights and well-informed decision-making. The significance of a comprehensive approach to data interpretation is highlighted by the combination of graphical representations, descriptive statistics, probabilistic modelling and matrix-based regression techniques. Drawing reliable conclusions in academic settings requires the careful application of statistical techniques along with a critical assessment of assumptions and findings.