CHAPTER 1

Recall on Bivariate Statistics

The objective of this statistical study is to analyze, within the same population of N individuals, two different variables and to investigate whether there exists a link or correlation between them.
For example: height and age, cholesterol and diet, irrigation and tomato yield, fertilizer and wheat production. The goal is to determine whether changes in one variable are systematically associated with changes in the other.

1 Bivariate Statistical Series

Definition

A **bivariate statistical series** is a dataset where two variables are observed simultaneously for the same individuals. This allows the study of the relationship between the two variables.

1.1.1 Scatter Plot

Definition

A **scatter plot** is a graphical representation of a bivariate dataset. Each individual is represented by a point $M_i(x_i, y_i)$ in a Cartesian coordinate system. The collection of all such points forms the *scatter plot*.

Example 01: Plant Production

Fertilizer (x, kg/ha) vs Wheat yield (y, q/ha). This example investigates whether increasing the amount of fertilizer leads to higher wheat yields.

x_i	0	50	100	150	200	250
y_i	15	20	32	40	43	44

Example 02: Agricultural Sciences

Irrigation (x, mm/week) vs Tomato yield (y, kg/m^2). This example examines whether more irrigation improves tomato production.

x_i	10	20	30	40	50	60
y_i	2.5	3.0	4.2	5.0	5.4	5.6

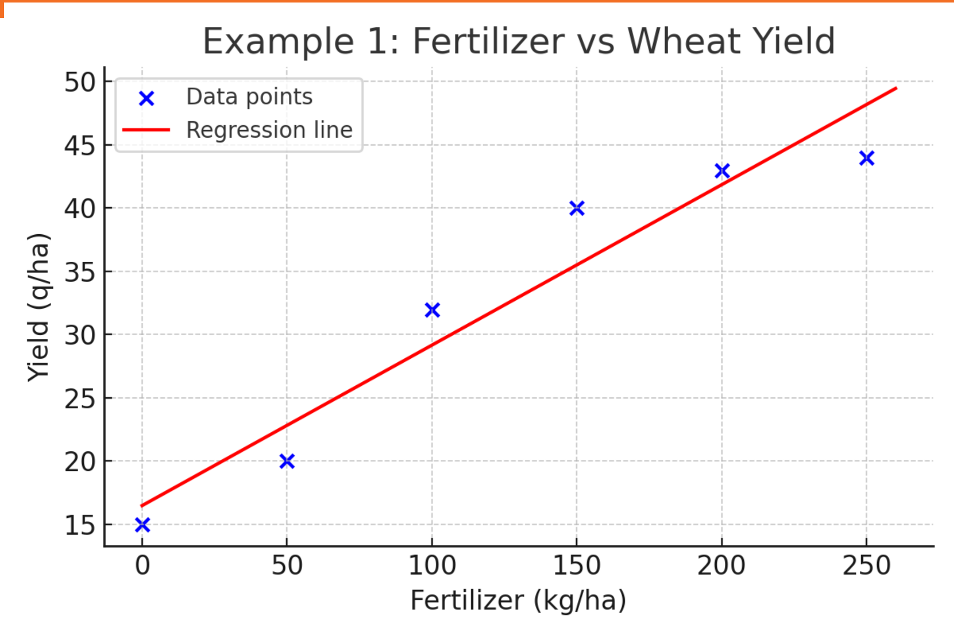

Figure 1.1: Scatter plot with regression line for Fertilizer vs Wheat Yield

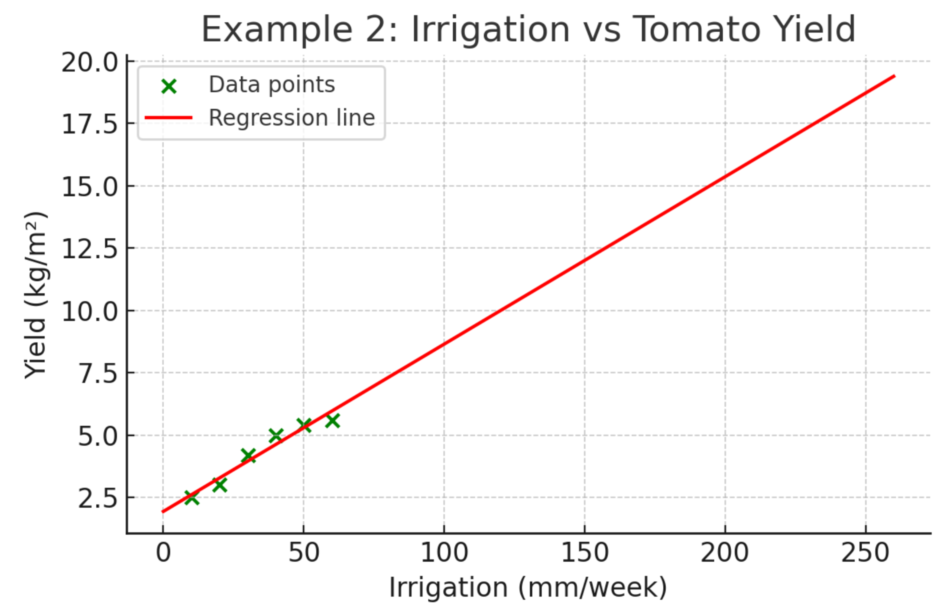

Figure 1.2: Scatter plot with regression line for Irrigation vs Tomato Yield

1.1.2 Marginal Means

Definition

The **marginal mean** is the average value of a variable:

$$\bar{x} = \frac{1}{N}\sum x_i, \qquad \bar{y} = \frac{1}{N}\sum y_i$$

Solutions – Means

Example 1 (Fertilizer vs Wheat yield):

$$N = 6,$$
$$\sum x_i = 0 + 50 + 100 + 150 + 200 + 250 = 750, \quad \bar{x} = \frac{750}{6} = 125.$$
$$\sum y_i = 15 + 20 + 32 + 40 + 43 + 44 = 194, \quad \bar{y} = \frac{194}{6} = 32.333333\ldots$$

Example 2 (Irrigation vs Tomato yield):

$$N = 6,$$
$$\sum x_i = 10 + 20 + 30 + 40 + 50 + 60 = 210, \quad \bar{x} = \frac{210}{6} = 35.$$
$$\sum y_i = 2.5 + 3.0 + 4.2 + 5.0 + 5.4 + 5.6 = 25.7, \quad \bar{y} = \frac{25.7}{6} = 4.283333\ldots$$

1.1.3 Covariance and Variance

Definition

The **covariance** measures whether two variables vary together:

$$\operatorname{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right) - \bar{x}\bar{y},$$

The **variance** measures the spread of a variable:

$$V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2,$$

$$V(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} y_i^2\right) - \bar{y}^2,$$

Standard deviations: $\sigma(x) = \sqrt{V(x)}$, $\sigma(y) = \sqrt{V(y)}$.

Solutions – Covariance, Variances, Std. dev.

Example 1 (Fertilizer vs Wheat yield):
First compute sums needed:

$$\sum x_i = 750, \qquad \sum y_i = 194,$$
$$\sum x_i y_i = 0 \cdot 15 + 50 \cdot 20 + 100 \cdot 32 + 150 \cdot 40 + 200 \cdot 43 + 250 \cdot 44 = 29800,$$
$$\sum x_i^2 = 0^2 + 50^2 + 100^2 + 150^2 + 200^2 + 250^2 = 137500,$$
$$\sum y_i^2 = 15^2 + 20^2 + 32^2 + 40^2 + 43^2 + 44^2 = 7034.$$

Now apply formulas (with $N = 6$):

$$\bar{x} = \frac{750}{6} = 125, \qquad \bar{y} = \frac{194}{6} = 32.333333\ldots$$
$$\frac{1}{N}\sum x_i y_i = \frac{29800}{6} = 4966.6666667.$$

$$\operatorname{cov}(x,y)=\frac{29800}{6}-\bar{x}\bar{y}=4966.6666667-125\times 32.3333333=4966.6666667-4041.6666667=925.$$

$$\frac{1}{N}\sum x_i^2=\frac{137500}{6}=22916.6666667,$$
$$V(x)=22916.6666667-125^2=22916.6666667-15625=7291.6666667.$$

$$\frac{1}{N}\sum y_i^2=\frac{7034}{6}=1172.3333333,$$

$$V(y)=1172.3333333-(32.3333333)^2=1172.3333333-1045.4444444=126.8888889.$$

$$\sigma(x)=\sqrt{7291.6666667}\approx 85.3913,\qquad \sigma(y)=\sqrt{126.8888889}\approx 11.2689.$$

Example 2 (Irrigation vs Tomato yield):
Compute sums:

$$\sum x_i=210,\qquad \sum y_i=25.7,$$
$$\sum x_iy_i=10\cdot 2.5+20\cdot 3.0+30\cdot 4.2+40\cdot 5.0+50\cdot 5.4+60\cdot 5.6=1017,$$
$$\sum x_i^2=10^2+20^2+30^2+40^2+50^2+60^2=9100,$$
$$\sum y_i^2=2.5^2+3.0^2+4.2^2+5.0^2+5.4^2+5.6^2=118.41.$$

With $N=6$:

$$\bar{x}=\frac{210}{6}=35,\qquad \bar{y}=\frac{25.7}{6}=4.283333\ldots$$
$$\frac{1}{N}\sum x_iy_i=\frac{1017}{6}=169.5.$$
$$\operatorname{cov}(x,y)=169.5-35\times 4.2833333=169.5-149.9166667=19.5833333.$$

$$\frac{1}{N}\sum x_i^2=\frac{9100}{6}=1516.6666667,\quad V(x)=1516.6666667-35^2=1516.6666667-1225=291.6666667.$$

$$\frac{1}{N}\sum y_i^2=\frac{118.41}{6}=19.735,\quad V(y)=19.735-(4.2833333)^2=19.735-18.3469444=1.3880556.$$

$$\sigma(x)=\sqrt{291.6666667}\approx 17.0783,\qquad \sigma(y)=\sqrt{1.3880556}\approx 1.17812.$$

1.1.4 Regression Lines

Theorem

The **regression line** of Y on X predicts Y from X: $Y=aX+b$ with

$$a=\frac{\operatorname{cov}(x,y)}{V(x)},\qquad b=\bar{y}-a\bar{x}.$$

Regression of X on Y: $X = a'Y + b'$ with

$$a' = \frac{\mathrm{cov}(x,y)}{V(y)}, \qquad b' = \bar{x} - a'\bar{y}.$$

Solutions – Regression

Example 1 (Fertilizer vs Wheat yield):

$$a = \frac{925}{7291.6666667} \approx 0.1268571429, \qquad b = 32.3333333 - 0.1268571429 \times 125 \approx 16.4761905.$$

So:

$$\boxed{Y \approx 0.12686\,X + 16.47619}$$

For X on Y:

$$a' = \frac{925}{126.8888889} \approx 7.289843, \qquad b' = 125 - 7.289843 \times 32.3333333 \approx -110.7049.$$

So:

$$\boxed{X \approx 7.28984\,Y - 110.7049}$$

Example 2 (Irrigation vs Tomato yield):

$$a = \frac{19.5833333}{291.6666667} \approx 0.0671428571, \qquad b = 4.2833333 - 0.0671428571 \times 35 \approx 1.9333333.$$

So:

$$\boxed{Y \approx 0.06714\,X + 1.93333}$$

For X on Y:

$$a' = \frac{19.5833333}{1.3880556} \approx 14.10847, \qquad b' = 35 - 14.10847 \times 4.2833333 \approx -25.43126.$$

So:

$$\boxed{X \approx 14.10847\,Y - 25.43126}$$

1.1.5 Linear Correlation Coefficient

Definition

The **linear correlation coefficient** r measures the strength of the linear relationship:

$$r = \frac{\mathrm{cov}(x,y)}{\sigma(x)\,\sigma(y)}.$$

Remark

1. $-1 \leq r \leq 1$.
2. If $r = 1$ or $r = -1$, then there is a perfect positive or negative correlation between X and

Y, and all points (x_i, y_i) lie exactly on the regression line.
A positive correlation means that an increase in X leads to an increase in Y.
A negative correlation means that an increase in X leads to a decrease in Y, or vice versa.
3. If $r = 0$, then there is no correlation between X and Y, and the points (x_i, y_i) are scattered randomly.
4. If $0 < r < 1$, then there is a weak, moderate, or strong positive correlation between X and Y.
5. If $-1 < r < 0$, then there is a weak, moderate, or strong negative correlation between X and Y.

Solutions – Correlation

Example 1:

$$r = \frac{925}{85.3913 \times 11.2689} \approx \frac{925}{962.05} \approx 0.9615 \approx 0.962.$$

Interpretation: strong positive linear correlation between fertilizer and wheat yield.

Example 2:

$$r = \frac{19.5833333}{17.0783 \times 1.17812} \approx \frac{19.5833}{20.1209} \approx 0.9730.$$

Interpretation: very strong positive linear correlation between irrigation and tomato yield.