Descriptive Statistics
Part -2

Sourav Karmakar

souravkarmakar29@gmail.com

Relationships among Variables

While it is important to know how to describe the distribution of a single variable, most studies pose questions that involve exploring the relationship between **two** variables using the collected data. Here are a few examples of such questions with the two variables highlighted:

1. How is the **number of calories** (Y) in a sandwich related to the **type of sandwich** (X) (veg, egg or chicken)
2. Are the **smoking habits** (Y) of a person (yes/no) related to the person's **gender** (X) (male/female)?
3. What is the relationship between a person's **age** (X) and **farsightedness?** (Y)
4. Is there a relationship between **gender** (X) and **test scores** (Y) on a particular standardized test?
5. Can you predict a person's **favorite type of music** (Y) (classical, rock, jazz) based on his/her **Age** (X)?

In most studies involving two variables, each of the variables has a role. We distinguish between:

- The **explanatory** variable (also commonly referred to as the **independent variable**) — the variable that claims to explain, predict or affect the response. Typically, it is denoted by **X.**
- The **response** variable (also commonly referred to as the **dependent variable**) — the outcome of the study. Denoted by **Y.**

Can you predict the **X** and **Y** of above examples?

Relationships among Variables

If we further classify each of the two relevant variables according to **type** (categorical or quantitative), we get the following 4 possibilities for **"role-type classification"**

- Categorical explanatory and quantitative response (C → Q)
- Categorical explanatory and categorical response (C → C)
- Quantitative explanatory and quantitative response (Q → Q)
- Quantitative explanatory and categorical response (Q → C)

		Response: Categorical	Response: Quantitative
Explanatory	Categorical	C → C	C → Q
	Quantitative	Q → C	Q → Q

1. How is the **number of calories** in a sandwich related to (or affected by) the **type of sandwich** (veg, egg or chicken) **Ans: C → Q**

2. Are the **smoking habits** of a person (yes/no) related to the person's **gender** (male/female)? **Ans: C → C**

3. What is the relationship between a person's **age** and **farsightedness?** **Ans: Q → Q**

4. Is there a relationship between **gender** and **test scores** on a particular standardized test? **Ans: C → Q**

5. Can you predict a person's **favorite type of music** (classical, rock, jazz) based on his/her **Age**? **Ans: Q → C**

C → C Relationships

A survey is done among 1200 college students and they were asked how do they feel about their **body-image** (over-weight / under-weight/ about right) . The table of the data looks like following:

Student	Gender	Body-Image
1	M	over-weight
2	M	about-right
3	F	over-weight
.	.	.
.	.	.
.	.	.
1198	F	under-weight
1199	M	about-right
1200	F	over-weight

Explanatory

Response

Suppose we ask questions like:

- Are men and women just as likely to think their weight is about-right?
- Is there a difference between the genders in feelings about body image?

To answer these type of questions we need to examine the relationship between two categorical variables, **gender** and **body image**.

Here **Gender** is **Explanatory** variable and **Body Image** is the **Response** variable

In order to summarize the relationship between two categorical variables, we create a display called a **two-way table**.

C → C Relationships

Two Way Table:

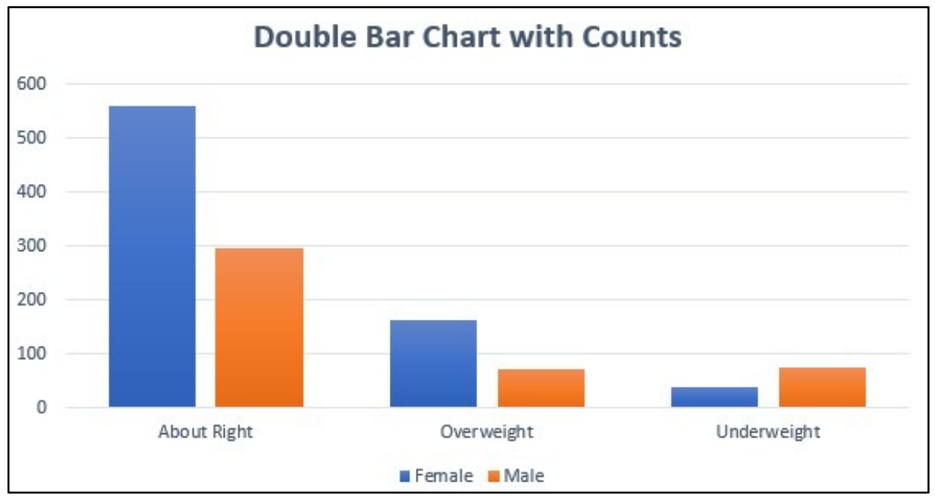
		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

This **Total** gives the summary of the categorical variable **Body Image**

This **Total** gives the summary of the categorical variable **Gender**

A way to visualize the distribution of categorical variables and relationship among them is to plot a **Double Bar Chart**

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

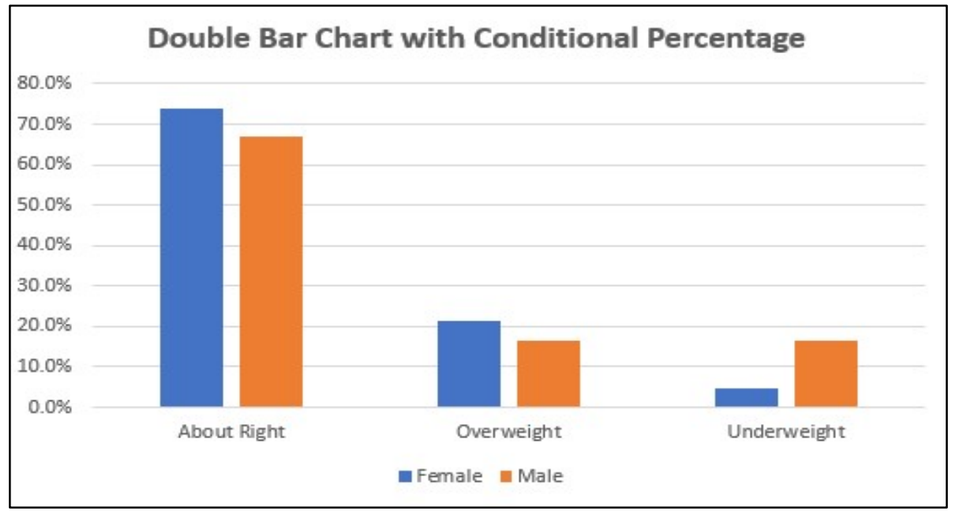


C → C Relationships

As we are examining the relationship among gender and body image, hence we create a conditional percentage table and plot corresponding Double Bar Chart

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	73.7%	21.4%	4.9%	100%
	Male	67.0%	16.4%	16.6%	100%



C → Q Relationships

The **Survey of Study Habits and Attitudes** (SSHA) is a psychological test designed to measure the motivation, study habits, and attitudes toward learning of college students. Is there a relationship between **gender** and **SSHA scores**? In other words, is there a "gender effect" on SSHA scores? Data were collected from 40 randomly selected college students, and here is what the raw data look like:

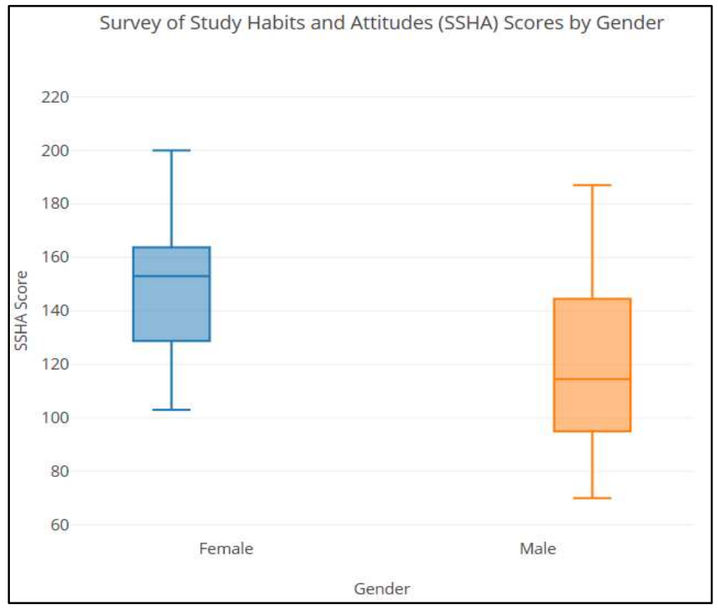
Explanatory ↖ Response ↗

	Gender	SSHA score
Student 1	Female	154
Student 2	Female	109
Student 3	Male	108
Student 4	Female	115
.	.	.
.	.	.
.	.	.
Student 40	Male	140

Following is the *Five-number summary* of SSHA score separated Gender-wise

Side-by-side boxplots supplemented by descriptive statistics allow us to compare the distribution of SSHA scores within each category of the explanatory variable **gender**:

Statistic	Female	Male
min	103	70
Q1	128.75	95
Median	153	114.5
Q3	163.75	144.5
Max	200	187



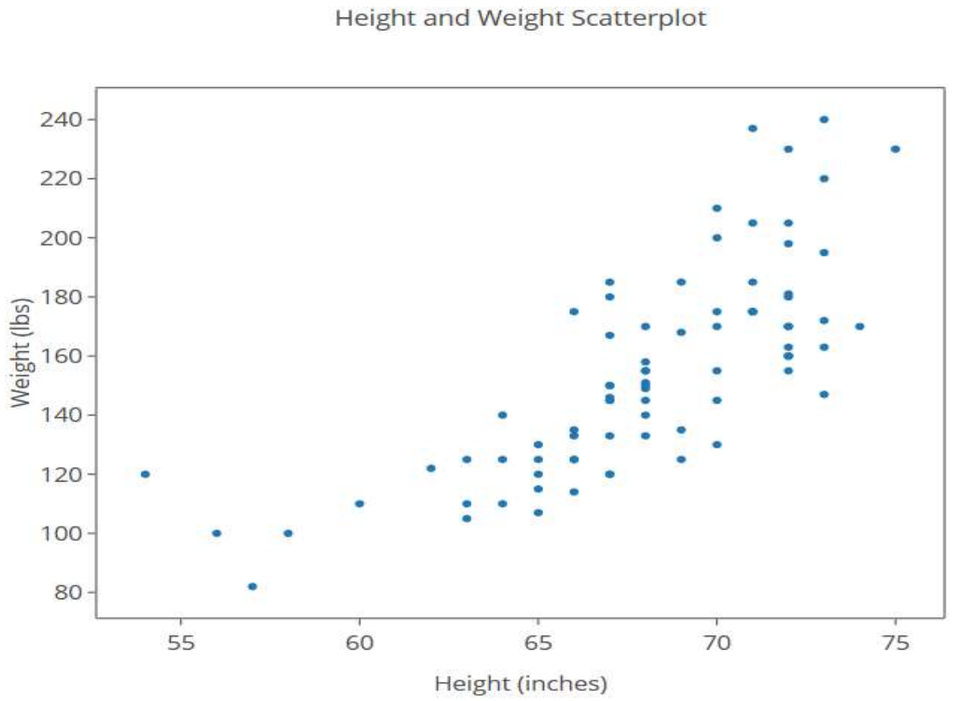
Q → Q Relationships

We look at height and weight data that were collected from 57 males and 24 females and use the data to explore how the weight of a person is related to (or affected by) his or her height. This implies that **height** will be our **explanatory variable** and **weight** will be our **response variable**. The dataset look something like following:

Sl. No.	height (inches)	weight (lbs)	male
1	67	115.3	1
2	62	120	0
3	73.5	142	1
.	.	.	.
.	.	.	.
.	.	.	.
79	71	175	1
80	63	125	0
81	58	100	0

We create scatter plots from the given data which looks like the one shown beside.

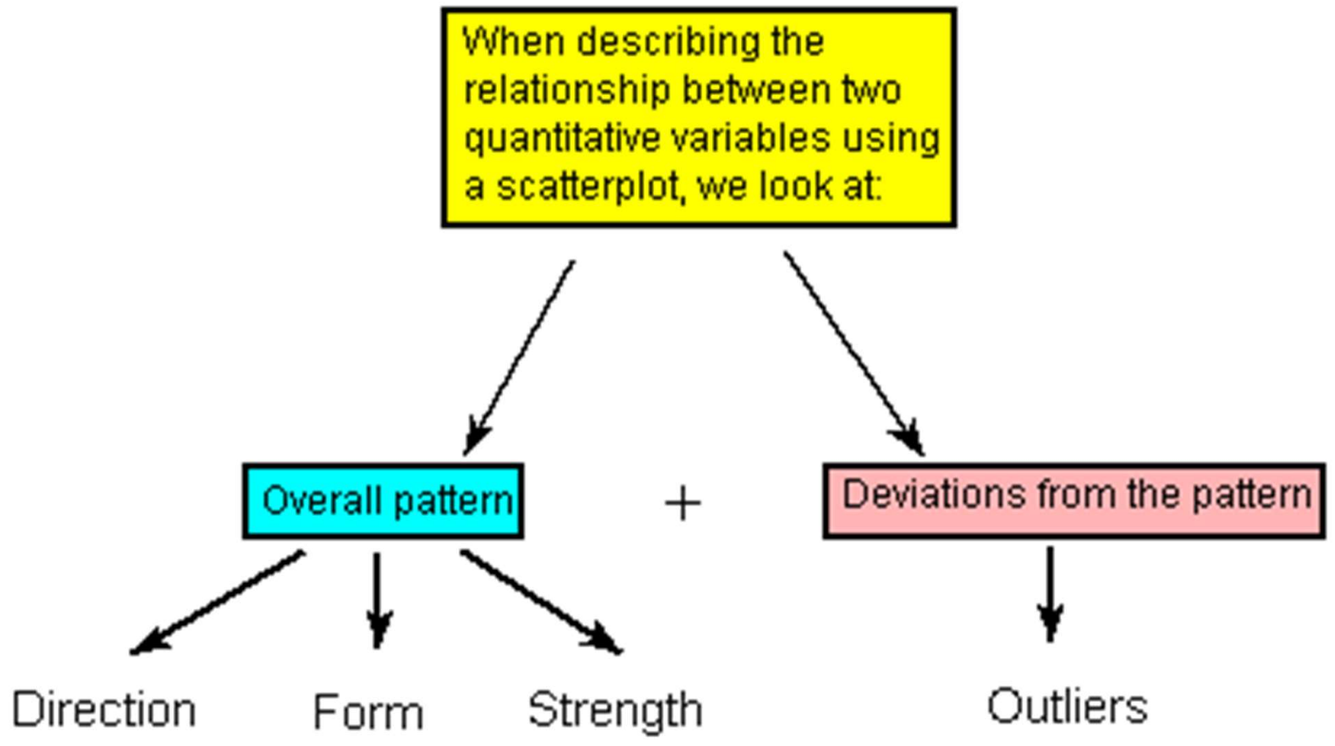
Drawing scatter plot is simple. You just have to plot each points specified by the co-ordinates. Here the co-ordinates are **(height, weight)**



Q → Q Relationships

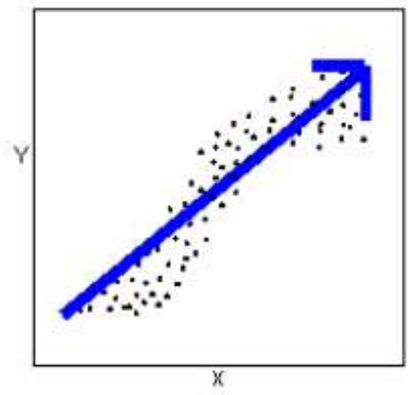
Interpreting the Scatterplot: How do we explore the relationship between two quantitative variables using the scatterplot?

Recall that when we described the distribution of a single quantitative variable with a histogram, we described the overall pattern of the distribution (shape, center, spread) and any deviations from that pattern (outliers). **We do the same thing with the scatterplot.**

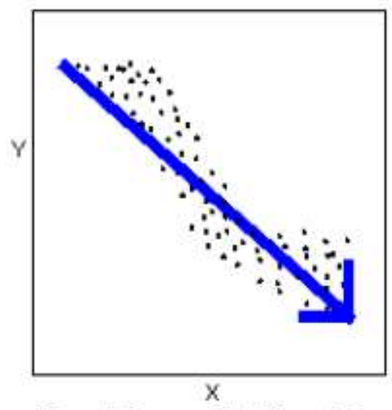


Q → Q Relationships

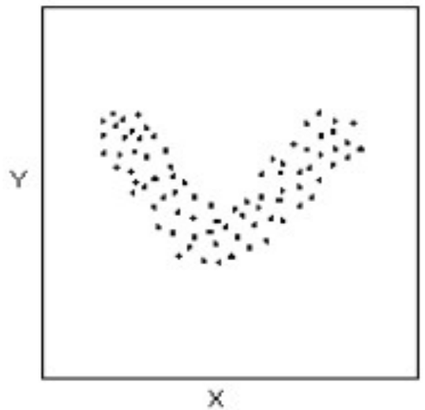
The **direction** of the relationship can be positive, negative, or neither:



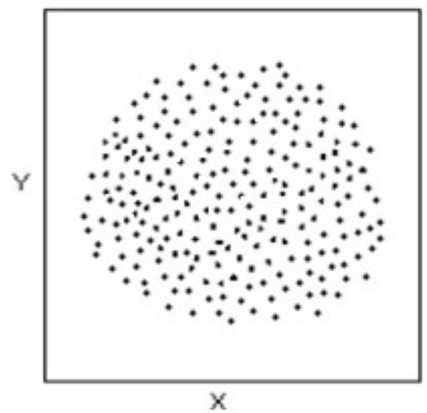
Positive relationship



Negative relationship

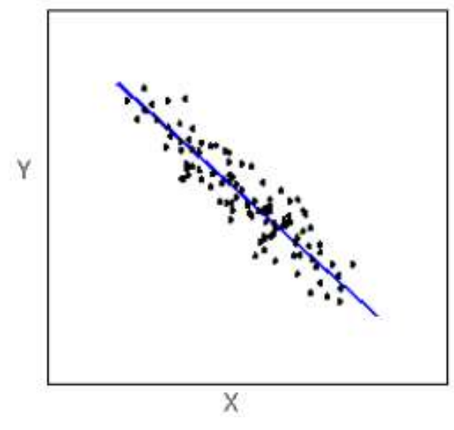


Neither positive nor negative

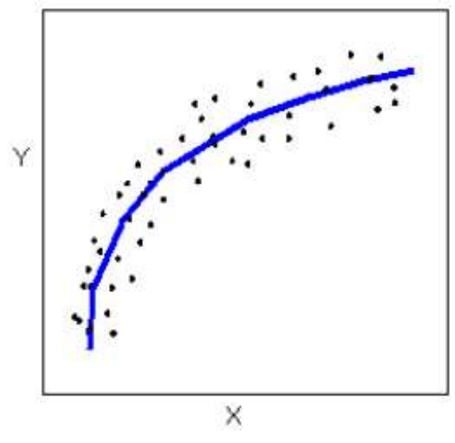


no relationship

The **form** of the relationship is its general shape. When identifying the form, we try to find the simplest way to describe the shape of the scatterplot. There are many possible forms.



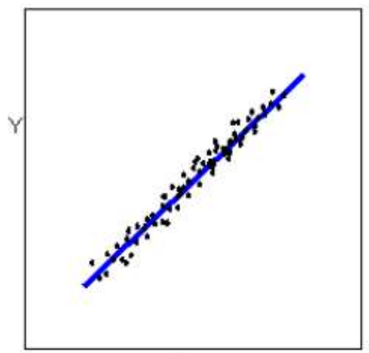
Linear



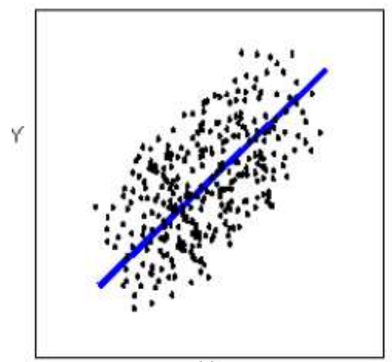
Curvilinear

Q → Q Relationships

The **strength** of the relationship is determined by how closely the data follow the form of the relationship. Let's look, for example, at the following two scatterplots displaying positive, linear relationships:

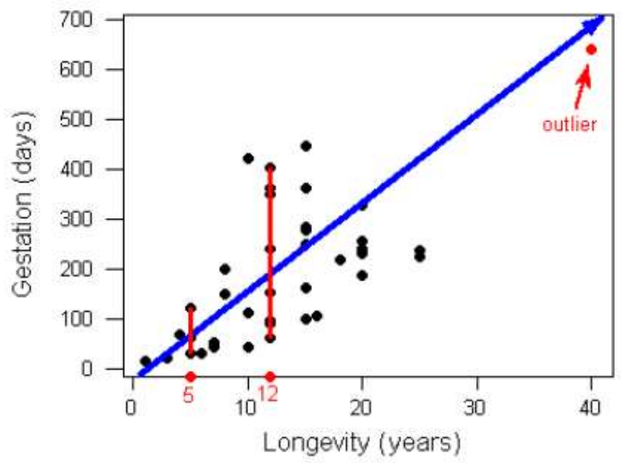


strong relationship



weaker relationship

Consider the following dataset. There is one **outlier** spotted in the dataset.



Can you guess the **direction** of the relationship? **Positive**

Can you guess the **form** of the relationship? **Linear**

Can you guess the **strength** of the relationship? **Weak**

Covariance

X	Y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n

① Total number of observations = n.

② Average of X : $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

③ Average of Y : $\bar{y} = \dfrac{y_1 + y_2 + \cdots + y_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$

④ $\dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \dfrac{(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \cdots + (x_n-\bar{x})(y_n-\bar{y})}{n}$

Example:-

X	Y
1	3
2	4
3	5
4	7
5	11

$\bar{x} = \dfrac{1+2+3+4+5}{5} = \dfrac{15}{5} = 3$, $\bar{y} = \dfrac{3+4+5+7+11}{5} = \dfrac{30}{5} = 6$

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
-2	-3	6
-1	-2	2
0	-1	0
1	1	1
2	5	10

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{6+2+0+1+10}{5} = \frac{19}{5} = 3.8$$

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$Cov(X,X) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = Var(X)$$

X, Y, Z

$$Cov(X,Y) = Cov(Y,X)$$

	X	Y	Z
X	Var(X)	Cov(X,Y) ✓	Cov(X,Z) ✓
Y	Cov(Y,X) ✓	Var(Y)	Cov(Y,Z) ✓
Z	Cov(Z,X) ✓	Cov(Z,Y) ✓	Var(Z)

Variance - Covariance matrix is a symmetric matrix

or simply Covariance matrix.

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$$

$$= Cov(Y,X)$$

$$\begin{bmatrix} 2 & 1.5 & 2.3 \\ 1.5 & 1 & 1.7 \\ 2.3 & 1.7 & 3 \end{bmatrix}$$

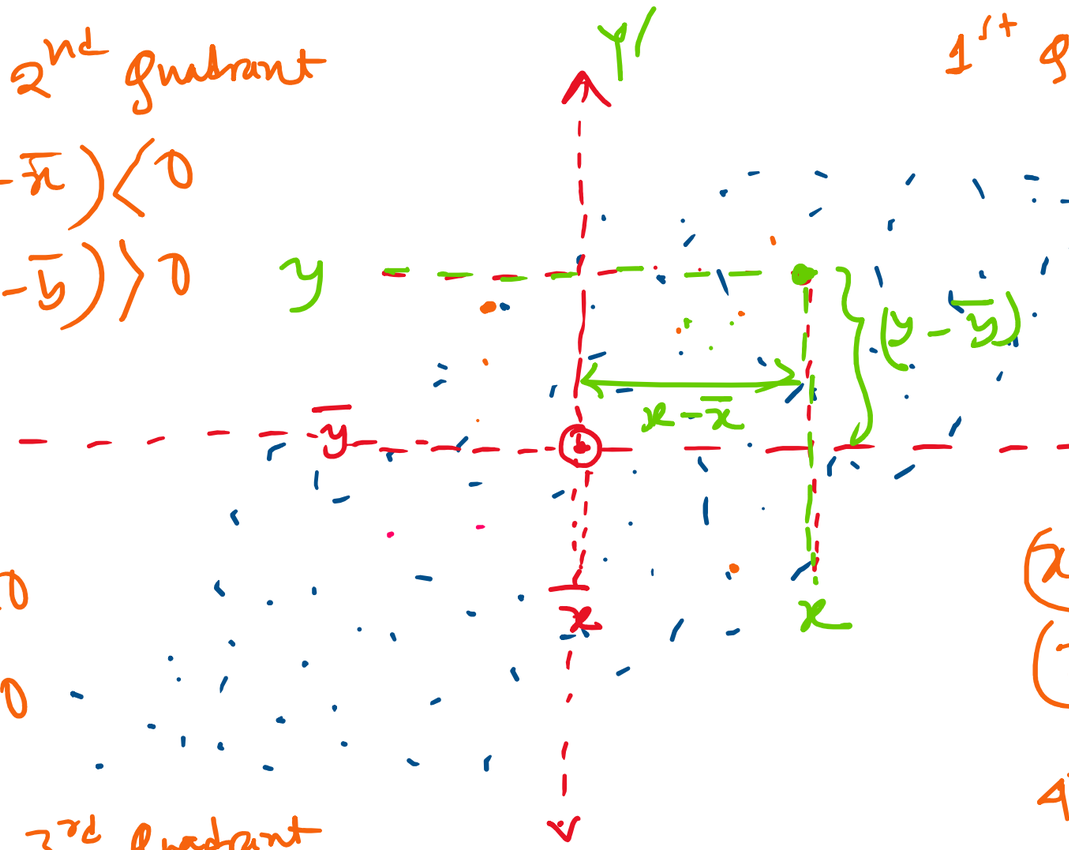
Symmetric matrix

$(x-\bar{x})(y-\bar{y})<0$

2nd quadrant

$(x-\bar{x})<0$

$(y-\bar{y})>0$



1st quadrant

$(x-\bar{x})>0$

$(y-\bar{y})>0$

$(x-\bar{x})(y-\bar{y})>0$

$(x-\bar{x})<0$

$(y-\bar{y})<0$

3rd quadrant

$(x-\bar{x})(y-\bar{y})>0$

$(x-\bar{x})>0$

$(y-\bar{y})<0$

$(x-\bar{x})(y-\bar{y})<0$

4th quadrant.

$$\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y}) > 0$$

$$Cov(X,Y) > 0$$

Q → Q Relationships

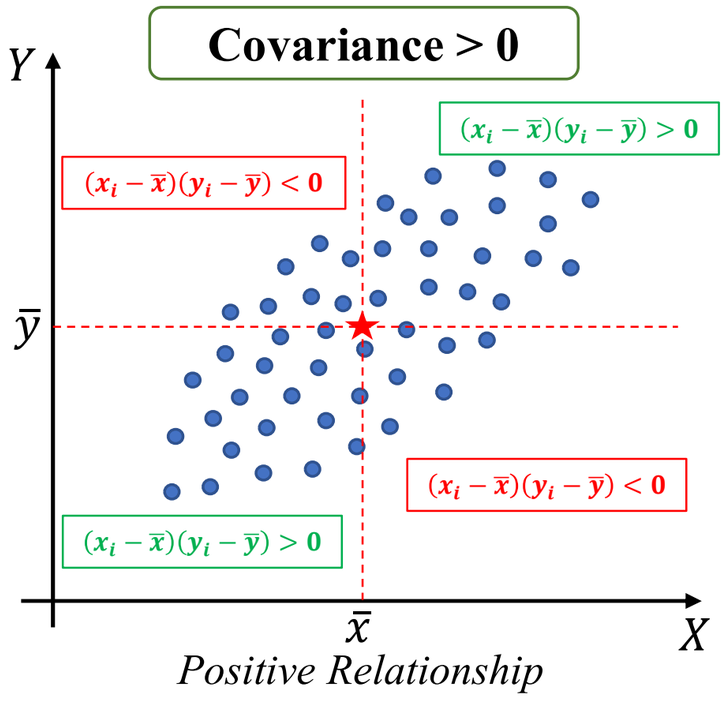
The Covariance (Cov):
The covariance between two quant. variable measures how variation in one variable affects the variation in another.

Calculation: The covariance between two quantitative variable, X and Y is calculated as following:

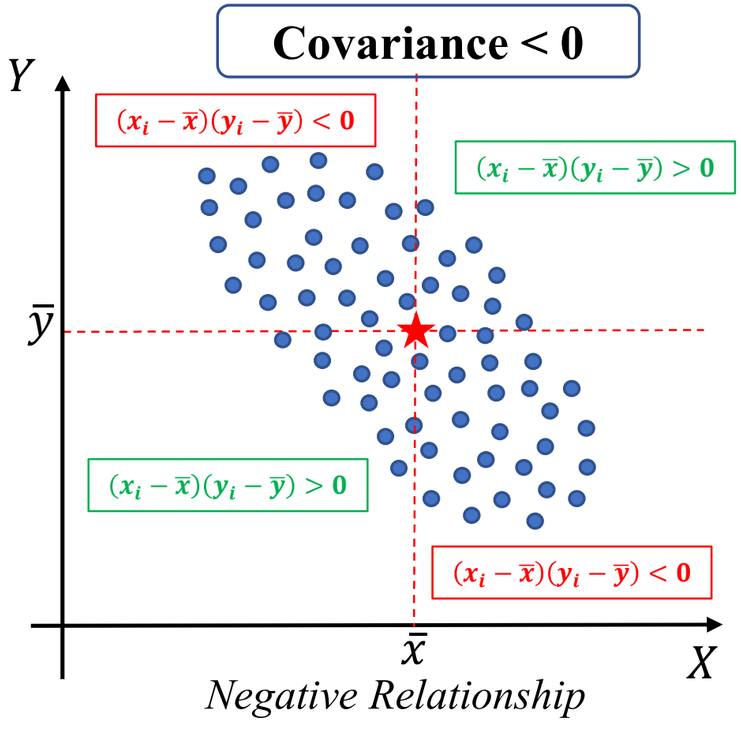
X	x_1	x_2	x_3	. . .	x_{n-2}	x_{n-1}	x_n
Y	y_1	y_2	y_3	. . .	y_{n-2}	y_{n-1}	y_n

$$Covariance(X,Y) = Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

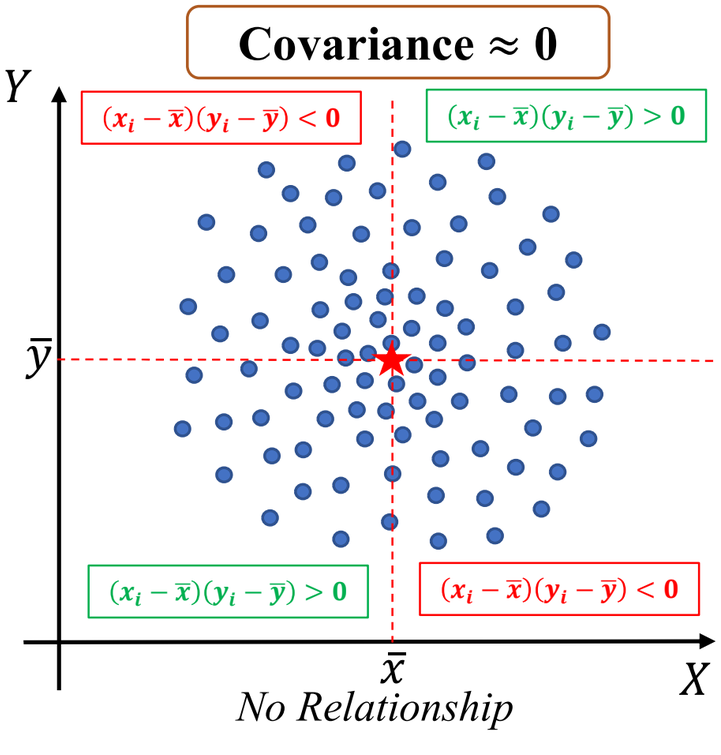
Intuition of Covariance:



Positive Relationship



Negative Relationship



No Relationship

Q → Q Relationships

The Correlation Coefficient (r):
The numerical measure that assesses the strength of a linear relationship is called the **correlation coefficient.**

Definition: The **correlation coefficient (r)** is a numerical measure that measures the **strength** and **direction** of a *linear relationship* between two quantitative variables.

Calculation: The **correlation coefficient (r)** between two quantitative variable X and Y is calculated as

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
.	.
.	.
.	.
x_{n-2}	y_{n-2}
x_{n-1}	y_{n-1}
x_n	y_n

Let σ_x and σ_y be the standard deviations of X and Y respectively and there are n data-points. Then the correlation coefficient between X and Y (denoted by r_{xy} is):

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

Pearson's Correlation Coeff.

Expanding σ_x and σ_y and noting that $\bar{x}$ and $\bar{y}$ are the mean of X and Y respectively. Then the correlation coefficient between X and Y (denoted by r_{xy} is) can be written as:

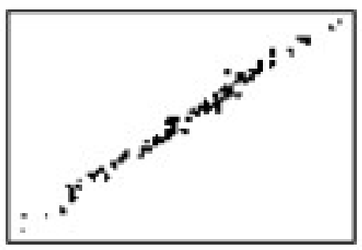
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Q → Q Relationships

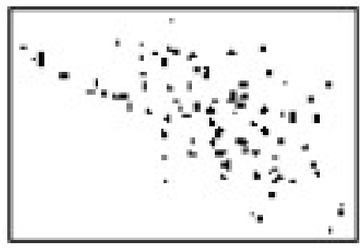
Interpreting the value of Correlation Coefficient (r):

The **correlation coefficient** lies between -1 to +1, i.e., $|r| \leq 1 \quad or \quad -1 \leq r \leq 1$

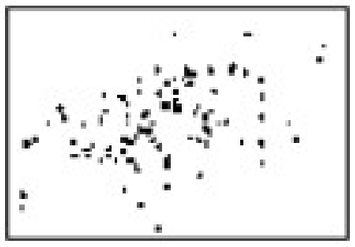
- Positive value of *r* indicates positive relationship.
- Negative value of *r* indicates negative relationship.
- *r* value close to 1 indicates strong positive linear relationship.
- *r* value close to -1 indicates strong negative linear relationship.
- *r* value close to 0 indicates the relationship is neither positive nor negative.



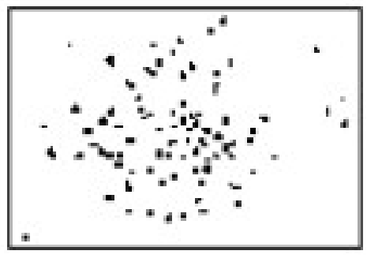
r = 0.995



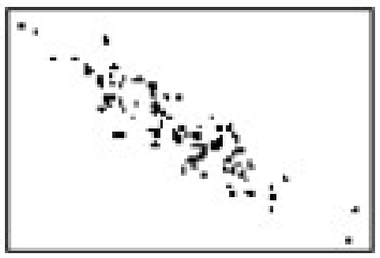
r = -0.575



r = 0.436



r = 0.100



r = -0.897

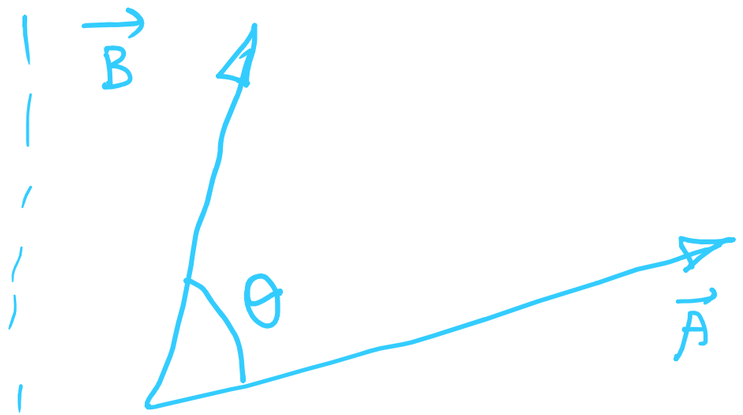
Correlation Coefficient lies between -1 to $+1$?

$$r_{xy} = \frac{COV(X,Y)}{\sqrt{VAR(X)} \cdot \sqrt{VAR(Y)}}$$

X	Y	$X-\bar{x}$	$Y-\bar{y}$
x_1	y_1	$x_1-\bar{x}$	$y_1-\bar{y}$
x_2	y_2	$x_2-\bar{x}$	$y_2-\bar{y}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$x_n-\bar{x}$	$y_n-\bar{y}$
		$\vec{V_1}$	$\vec{V_2}$

$$\bar{x} = \frac{1}{n}\sum_i x_i$$

$$\bar{y} = \frac{1}{n}\sum_i y_i$$



$$\vec{A} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad \vec{B} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$\vec{A} \cdot \vec{B} = \vec{A}^T B$$

$$= \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$= a_1 b_1 + a_2 b_2$$

$$\frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \, \|\vec{B}\|} = \cos\theta$$

$$-1 \leq \cos\theta \leq 1$$

$$\vec{V_1} \cdot \vec{V_2} = (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})$$

$$\|\vec{V_1}\| = \sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}$$

$$\|\vec{V_2}\| = \sqrt{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}$$

$$\vec{V_1} \cdot \vec{V_2} = (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\|\vec{V_1}\| = \sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2} = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\|\vec{V_2}\| = \sqrt{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2} = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\vec{V_1} \cdot \vec{V_2}}{\|\vec{V_1}\| \, \|\vec{V_2}\|} = \cos\theta$$

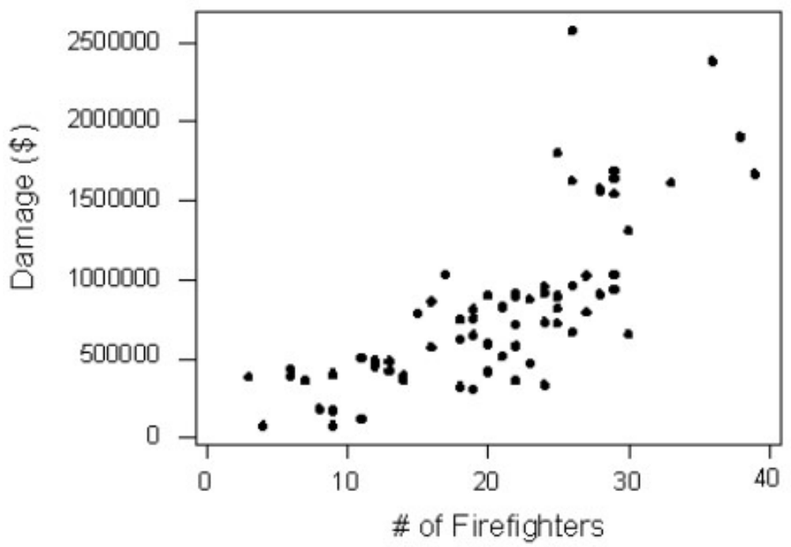
(θ is the angle between V_1 & V_2)

$$-1 \leq \cos\theta \leq 1$$

$$\therefore \boxed{-1 \leq r_{xy} \leq 1}$$

Causation

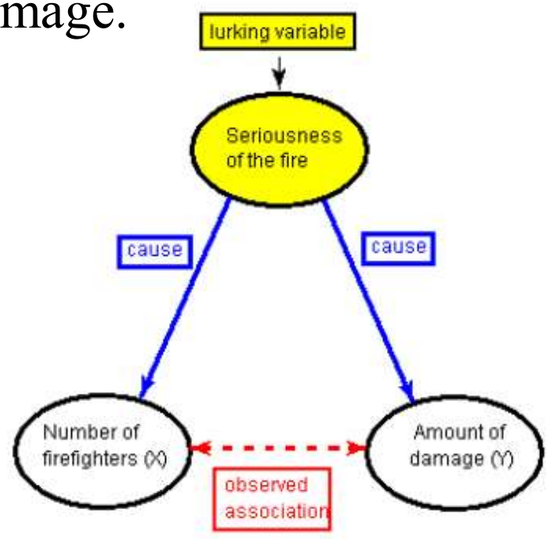
The scatterplot below illustrates how the number of firefighters sent to fires (X) is related to the amount of damage caused by fires (Y) in a certain city.



The scatterplot clearly displays a strong (slightly curved) **positive** relationship between the two variables. Would it, then, be reasonable to conclude that sending more firefighters to a fire causes more damage, or that the city should send fewer firefighters to a fire, in order to decrease the amount of damage done by the fire?

Of course not! So, what is going on here?

There is a **third variable in the background**—the seriousness of the fire—that is responsible for the observed relationship. More serious fires require more firefighters and cause more damage.



Here, the seriousness of the fire is a **lurking variable.** A **lurking variable** is a variable that is not among the explanatory or response variables in a study but could substantially affect your interpretation of the relationship among those variables.

Hence, Correlation doesn't imply Causation

Thank You