CORRELATION ANALYSIS

Major Points - Correlation

- Questions answered by correlation
- Scatterplots
- An example
- The correlation coefficient
- Other kinds of correlations
- Factors affecting correlations
- Testing for significance

The Question

- Are two variables related?
  - Does one increase as the other increases?
    - e. g. skills and income
  - Does one decrease as the other increases?
    - e. g. health problems and nutrition
- How can we get a numerical measure of the degree of relationship?

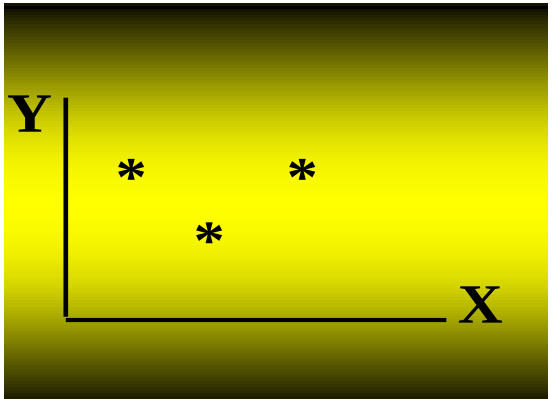
Correlation

Finding the relationship between two quantitative variables without being able to infer causal relationships

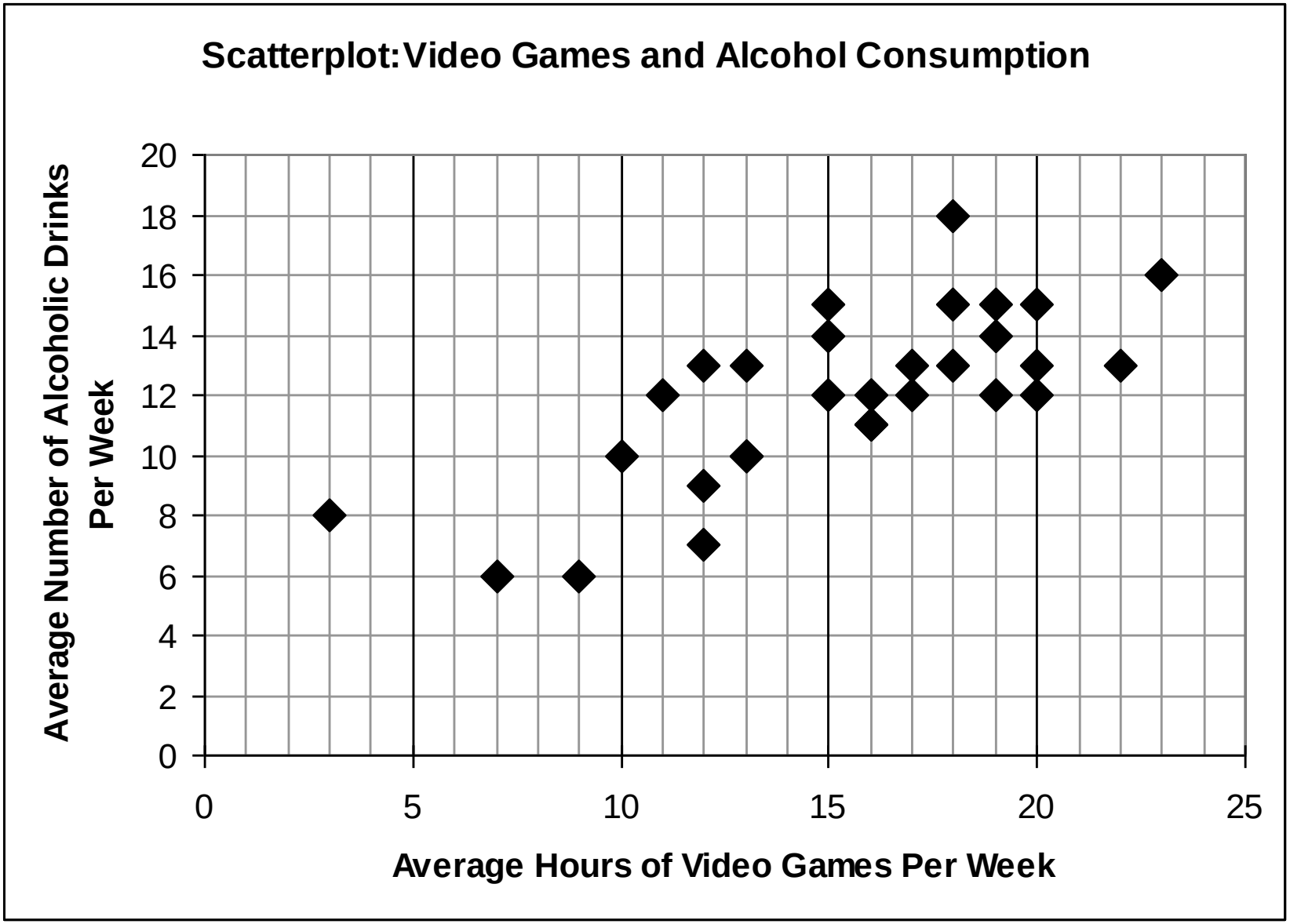
Correlation is a statistical technique used to determine the degree to which two variables are related

Scatterplots

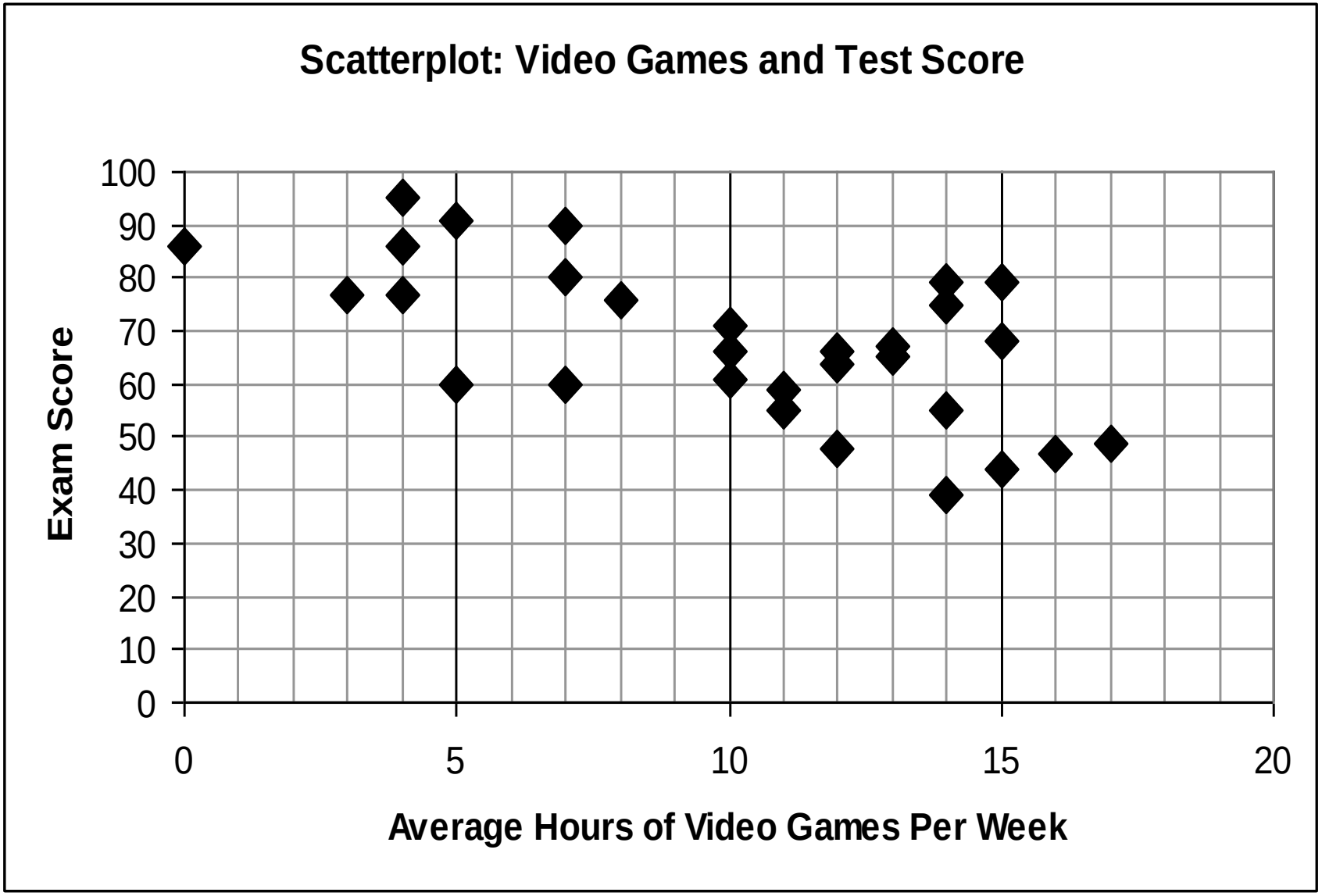
- Graphically depicts the relationship between two variables in two dimensional space. Rectangular coordinate
- Two quantitative variables
- One variable is called independent (X) and the second is called dependent (Y)
- Points are not joined



Direct Relationship



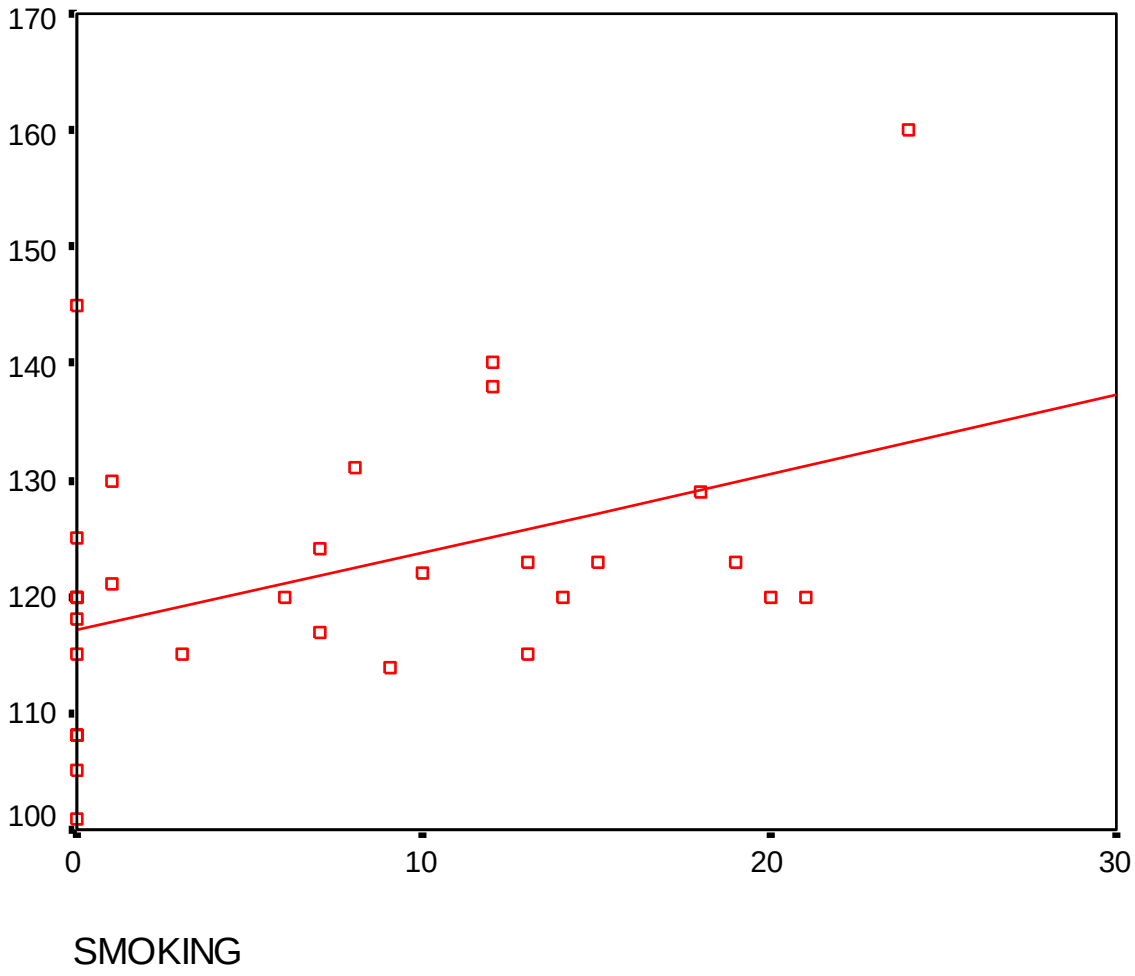
Inverse Relationship



An Example

- Does smoking cigarettes increase systolic blood pressure?
- Plotting number of cigarettes smoked per day against systolic blood pressure
  - Fairly moderate relationship
  - Relationship is positive

Trend?



Correlation

- Co-relation
- The relationship between two variables
- Measured with a correlation coefficient
- Most popularly seen correlation coefficient: Pearson Product-Moment Correlation

Types of Correlation

- Positive correlation
  - High values of X tend to be associated with high values of Y.
  - As X increases, Y increases
- Negative correlation
  - High values of X tend to be associated with low values of Y.
  - As X increases, Y decreases
- No correlation
- No consistent tendency for values on Y to increase or decrease as X increases

Correlation Coefficient

- A measure of degree of relationship between two variables.
- Between 1 and -1
- Sign refers to direction.
- Based on covariance
  - Measure of degree to which large scores on X go with large scores on Y, and small scores on X go with small scores on Y
  - Think of it as variance, but with 2 variables instead of 1 (What does that mean??)

Correlation

High positive correlation

Zero correlation

High negative correlation

stronger

+1.00

perfect positive

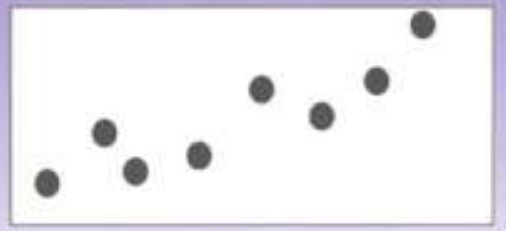
as one event increases, the second exactly increases

weaker

+.50

positive

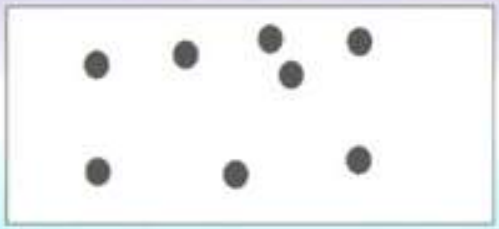
as one event increases, the second sometimes increases



0

zero correlation

no relationship between the events

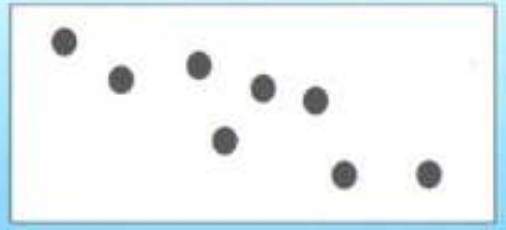


weaker

-.50

negative

as one event increases, the second sometimes decreases



stronger

-1.00

perfect negative

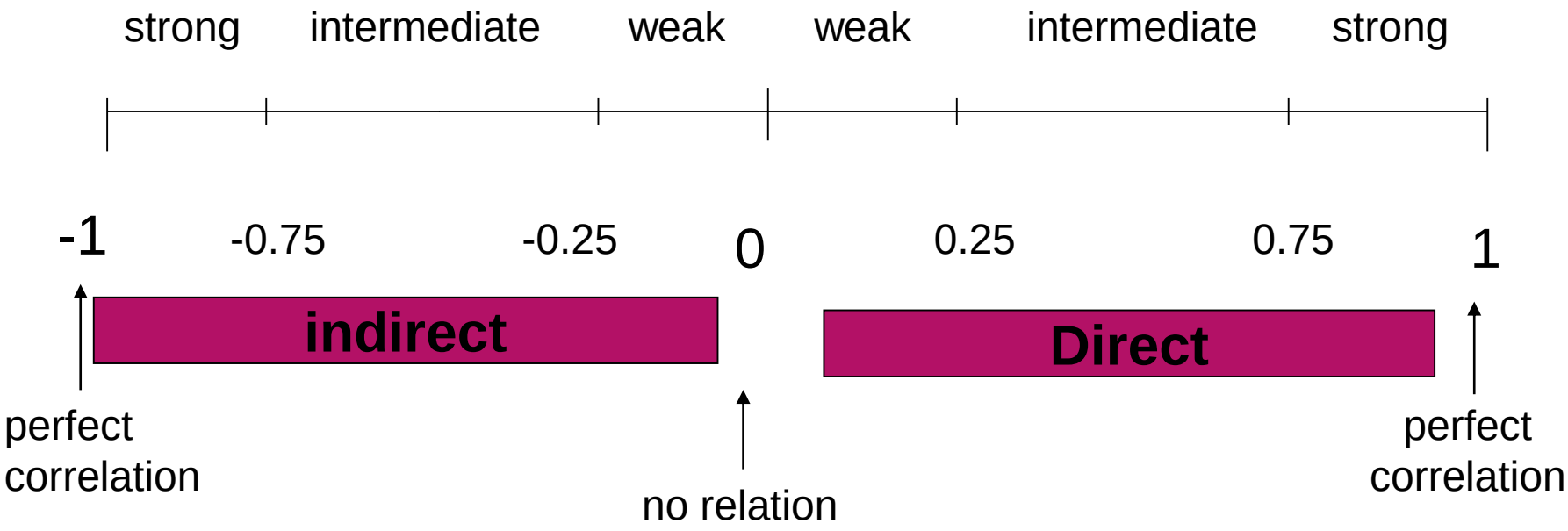
as one event increases, the second exactly decreases

Simple Correlation coefficient (r)

- It is also called Pearson's correlation or product moment correlation coefficient.
- It measures the nature and strength between two variables of the quantitative type.

- The sign of r denotes the nature of association
- while the value of r denotes the strength of association.

- The value of r ranges between (-1) and (+1)
- The value of r denotes the strength of the association as illustrated by the following diagram.



- If r = Zero this means no association or correlation between the two variables.
- If $0 < r < 0.25$ = weak correlation.
- If $0.25 \leq r < 0.75$ = intermediate correlation.
- If $0.75 \leq r < 1$ = strong correlation.
- If r = l = perfect correlation.

Significance test

- **Performing the Hypothesis Test**
- Null hypothesis: H_0: $\rho = 0$
- Alternate hypothesis: H_a: $\rho \neq 0$
- **What the Hypothesis Means in Words:**
- **Null hypothesis H_0:** The population correlation coefficient *is not* significantly different from zero. There *is not*a significant linear relationship (correlation) between *x* and *y* in the population.
- **Alternate hypothesis H_a:** The population correlation coefficient *is* significantly different from zero. There *is* a significant linear relationship (correlation) between *x* and *y* in the population.

Significance test

- **Drawing a Conclusion:**There are two methods to make a conclusion. The two methods are equivalent and give the same result.
- **Method 1: Use the *p*-value.**
- **Method 2: Use a table of critical values.**

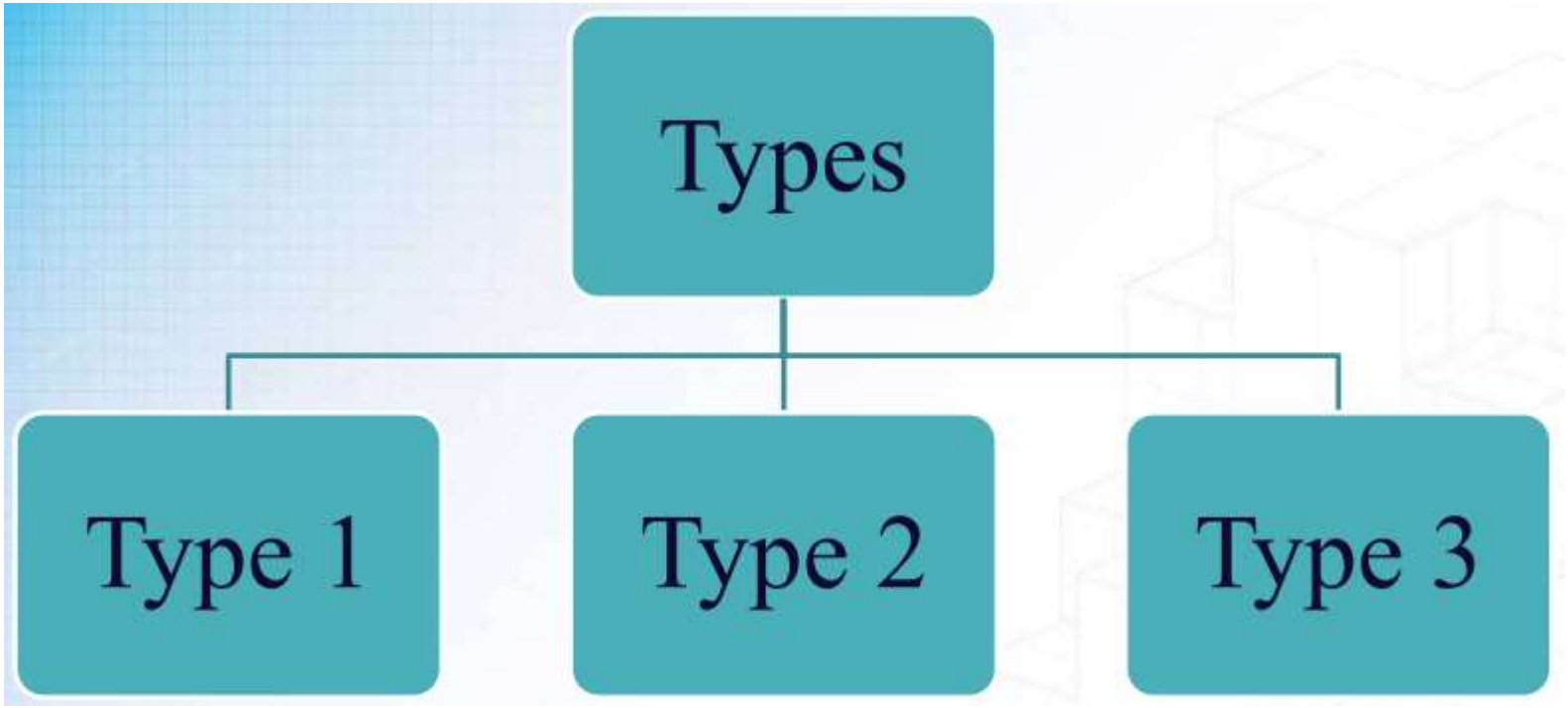
METHOD 1: Using a *p*-Value to Make a Decision

- If the *p*-value is less than the significance level ($\alpha = 0.05$),
- **Decision:** Reject the null hypothesis.
- **Conclusion:** There is sufficient evidence to conclude there is a significant linear relationship between *x* and *y*because the correlation coefficient is significantly different from zero.

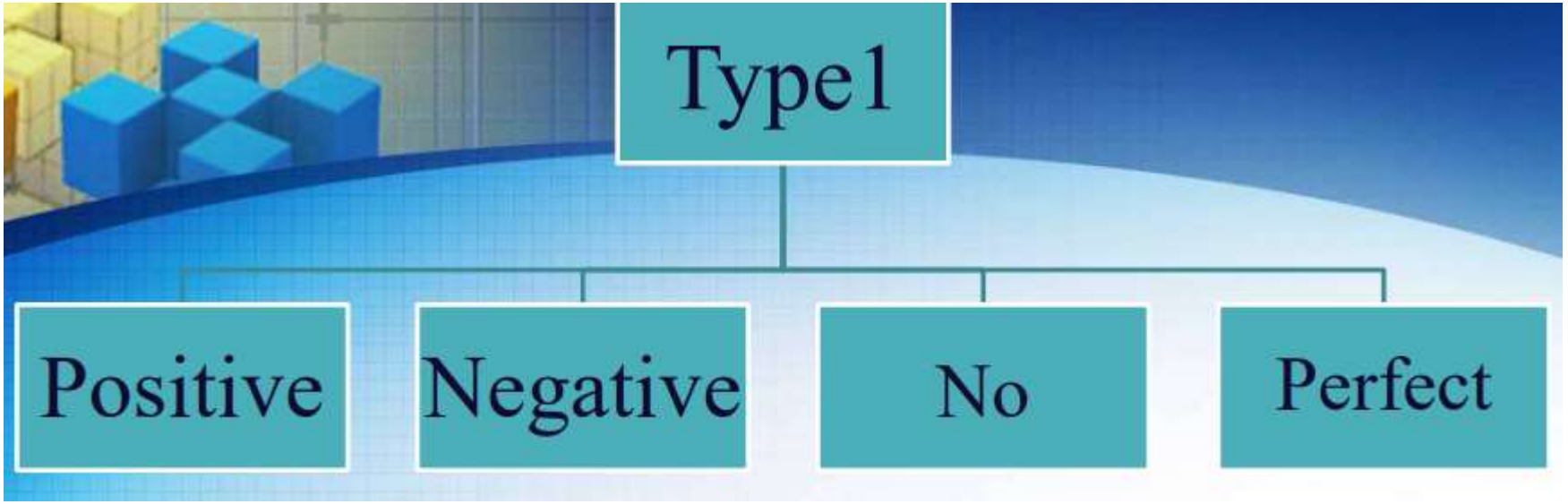
- If the *p*-value is *not* less than the significance level ($\alpha = 0.05$),
- **Decision:** Do not reject the null hypothesis.
- **Conclusion:** There is insufficient evidence to conclude there is a significant linear relationship between *x* and *y*because the correlation coefficient is not significantly different from zero.

Types of Correlation

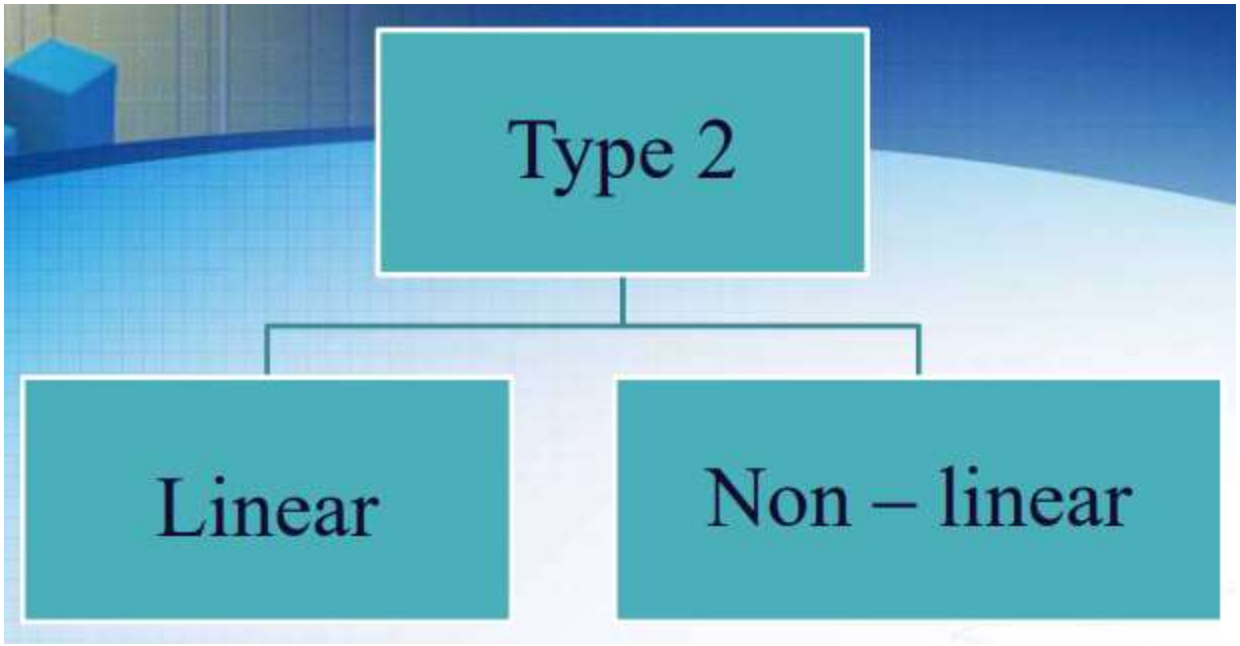
There are three types of correlation



Types of Correlation

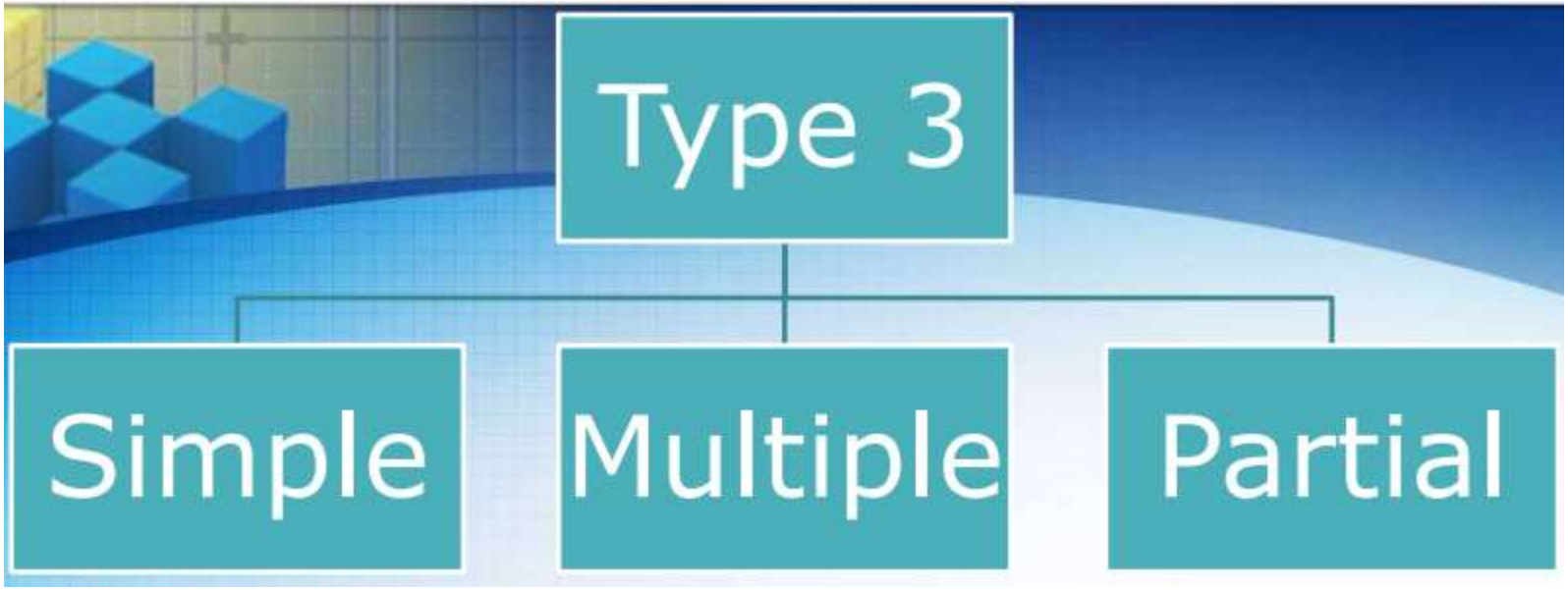


Types of Correlation



- When plotted on a graph it tends to be a perfect line
- When plotted on a graph it is not a straight line

Types of Correlation



Simple Correlation

- Simple correlation is a measure used to determine the strength and the direction of the relationship between two variables, *X* and *Y*.

Correlations

		Age	Cholesterol
Age	Pearson Correlation	1	.882**
	Sig. (2-tailed)		.001
	N	10	10
Cholesterol	Pearson Correlation	.882**	1
	Sig. (2-tailed)	.001	
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Multiple Correlation

- correlation between more than two variables

Correlations

		reading score	writing score	math score	science score	female
reading score	Pearson Correlation[a]	1	.597**	.662**	.630**	-.053
	Sig. (2-tailed)[b]	.	.000	.000	.000	.455
	N[c]	200	200	200	200	200
writing score	Pearson Correlation	.597**	1	.617**	.570**	.256**
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	200	200	200	200	200
math score	Pearson Correlation	.662**	.617**	1	.631**	-.029
	Sig. (2-tailed)	.000	.000	.	.000	.680
	N	200	200	200	200	200
science score	Pearson Correlation	.630**	.570**	.631**	1	-.128
	Sig. (2-tailed)	.000	.000	.000	.	.071
	N	200	200	200	200	200
female	Pearson Correlation	-.053	.256**	-.029	-.128	1
	Sig. (2-tailed)	.455	.000	.680	.071	.
	N	200	200	200	200	200

**. Correlation is significant at the 0.01 level (2-tailed).

Partial Correlation

- Partial correlation **measures the strength of a relationship between two variables, while controlling for the effect of one or more other variables.**

- For example, you might want to see if there is a correlation between amount of food eaten and blood pressure, while controlling for weight or amount of exercise.
- What is the relationship between test scores and GPA scores after controlling for hours spent studying?

Partial Correlation

Correlations

Control Variables			Weight	VO2max	Age
-none-[a]	Weight	Correlation	1.000	-.307	-.004
		Significance (2-tailed)	.	.002	.972
		df	0	98	98
	VO2max	Correlation	-.307	1.000	-.191
		Significance (2-tailed)	.002	.	.057
		df	98	0	98
	Age	Correlation	-.004	-.191	1.000
		Significance (2-tailed)	.972	.057	.
		df	98	98	0
Age	Weight	Correlation	1.000	-.314	
		Significance (2-tailed)	.	.002	
		df	0	97	
	VO2max	Correlation	-.314	1.000	
		Significance (2-tailed)	.002	.	
		df	97	0	

a. Cells contain zero-order (Pearson) correlations.

Autocorrelation

- Autocorrelation is a characteristic of data in which the correlation between the values of the same variables is based on related objects.
- It violates the assumption of instance independence, which underlies most of the conventional models.
- It generally exists in those types of data-sets in which the data, instead of being randomly selected, is from the same source.
- It occurs mostly due to dependencies within the data.

Autocorrelation

- cross sectional and time series data

- In cross sectional data, if the change in the income of a person A affects the savings of person B (a person other than person A), then autocorrelation is present.

- In the case of time series data, if the observations show inter-correlation, specifically in those cases where the time intervals are small, then these inter-correlations are given the term of autocorrelation.

INTERPRET

- **Strength**
- **Direction**

Correlations

	Hydrogen	Porosity
Porosity	0.624783	
	0.0169	
Strength	-0.790146	-0.527459
	0.0008	0.0526

Cell Contents: Pearson correlation
P-Value

In these results, the Pearson correlation between porosity and hydrogen is about 0.624783, which indicates that there is a moderate positive relationship between the variables. The Pearson correlation between strength and hydrogen is about -0.790146, and between strength and porosity is about -0.527459. The relationship between these variables is negative, which indicates that, as hydrogen and porosity increase, strength decreases.

INTERPRET

- Significance

Correlations

	Hydrogen	Porosity
Porosity	0.624783	
	0.0169	
Strength	-0.790146	-0.527459
	0.0008	0.0526

Cell Contents: Pearson correlation
P-Value

In these results, the p-values for the correlation between porosity and hydrogen and between strength and hydrogen are both less than the significance level of 0.05, which indicates that the correlation coefficients are significant. The p-value between strength and porosity is 0.0526. Because the p-value is greater than the significance level of 0.05, there is inconclusive evidence about the significance of the association between the variables.