Knowledge Discovery & Data Mining

— Analyzing Feature Relationships —

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Outline

- Analyzing Feature Relationships
  - Introduction to Feature Analysis
  - Covariance (for numerical features)
  - Correlation Coefficient (for numerical features)
  - Spearman's Rank Correlation (Numeric & Ordinal Data)
  - Chi-Square Test (for categorical features)

Why Analyze Relationships Between Features?

- Purpose: Understanding the relationships between features is key to improving predictive models, detecting patterns, and identifying significant associations.
- Key Reasons:
  - Identify Correlations: Determine how one feature may influence or be related to another.
  - Improve Model Performance: Feature relationships can inform better feature selection and model building.
  - Detect Patterns: Recognize trends and patterns within the data.
  - Hypothesis Testing: Verify if observed relationships in the data are statistically significant.

Covariance

Covariance is measure assessing how much two attributes change together.

Consider two numeric attributes A and B and a set of *n* real valued observations {(a1, b1), . . . , (an, bn)}. The mean values (also known as the expected values) of A and B, that is,

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n} \quad \text{and} \quad E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}$$

Then the covariance between A and B is defined as

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

or

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

Covariance

- **Positive covariance**: If Cov(A,B) > 0, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If Cov(A,B) < 0 then if A is larger than its expected value, B is likely to be smaller than its expected value.

If A and B are independent, then $E(A \cdot B) = E(A) \cdot E(B)$. ➡ $Cov(A,B) = 0$

Covariance

Example. This table presents a simplified example of stock prices observed at five time points for AllElectronics and HighTech, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- E(AllElectronics) =
- E(HighTech) =
- Cov(AllElectroncis, HighTech) =

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Covariance

Example. This table presents a simplified example of stock prices observed at five time points for AllElectronics and HighTech, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- E(AllElectronics) = 4
- E(HighTech) = 10.8
- Cov(AllElectroncis, HighTech)

$$= \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80$$

$$= 50.2 - 43.2 = 7.$$

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Covariance

Example. This table presents a simplified example of stock prices observed at five time points for AllElectronics and HighTech, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

- E(AllElectronics) = 4
- E(HighTech) = 10.8
- Cov(AllElectroncis, HighTech)

$$= \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80$$

$$= 50.2 - 43.2 = 7.$$

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

stock prices for both companies rise together

Correlation Analysis (Numeric Data)

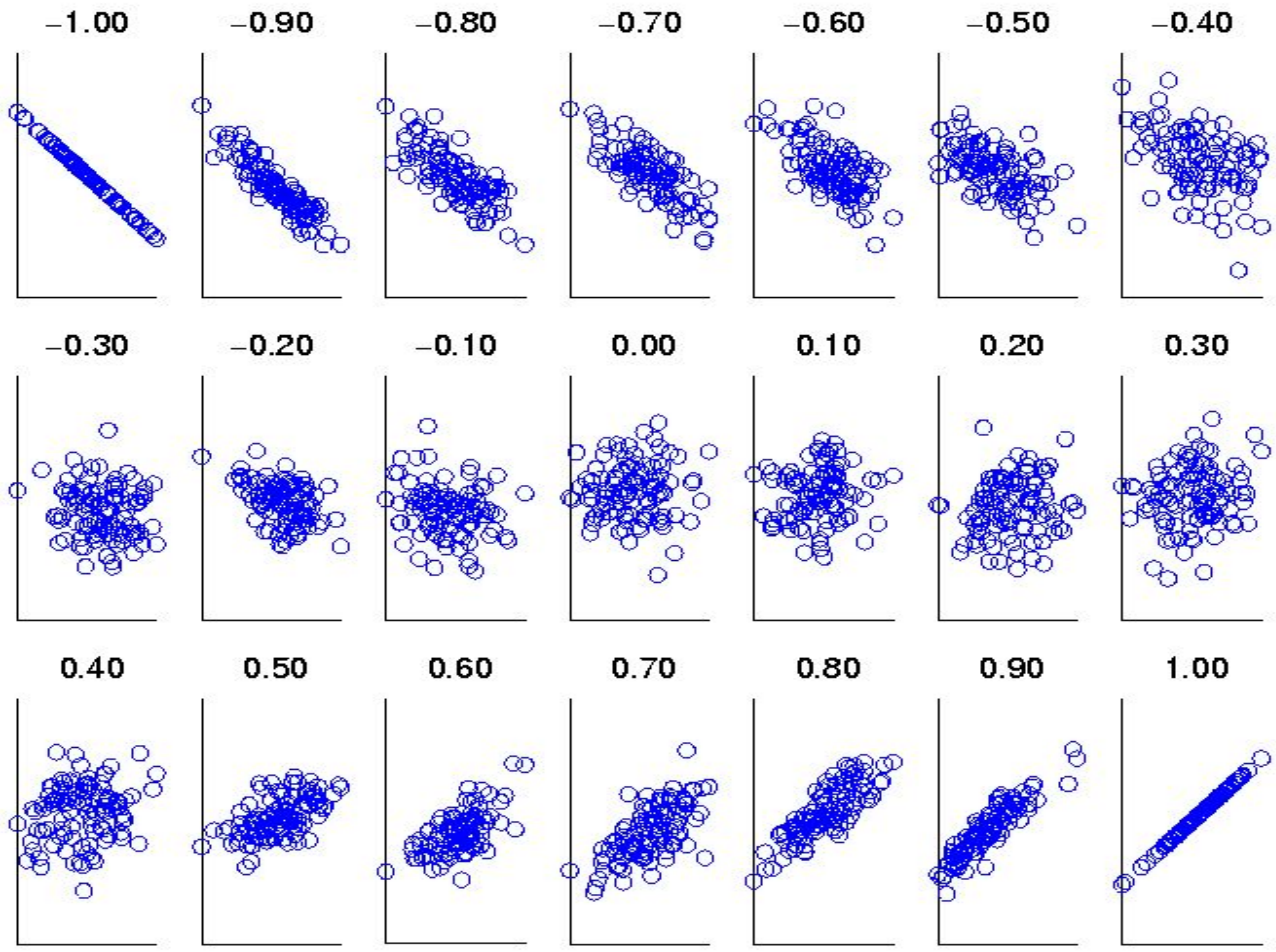
- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B} = \frac{Cov(A,B)}{\sigma_A\sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

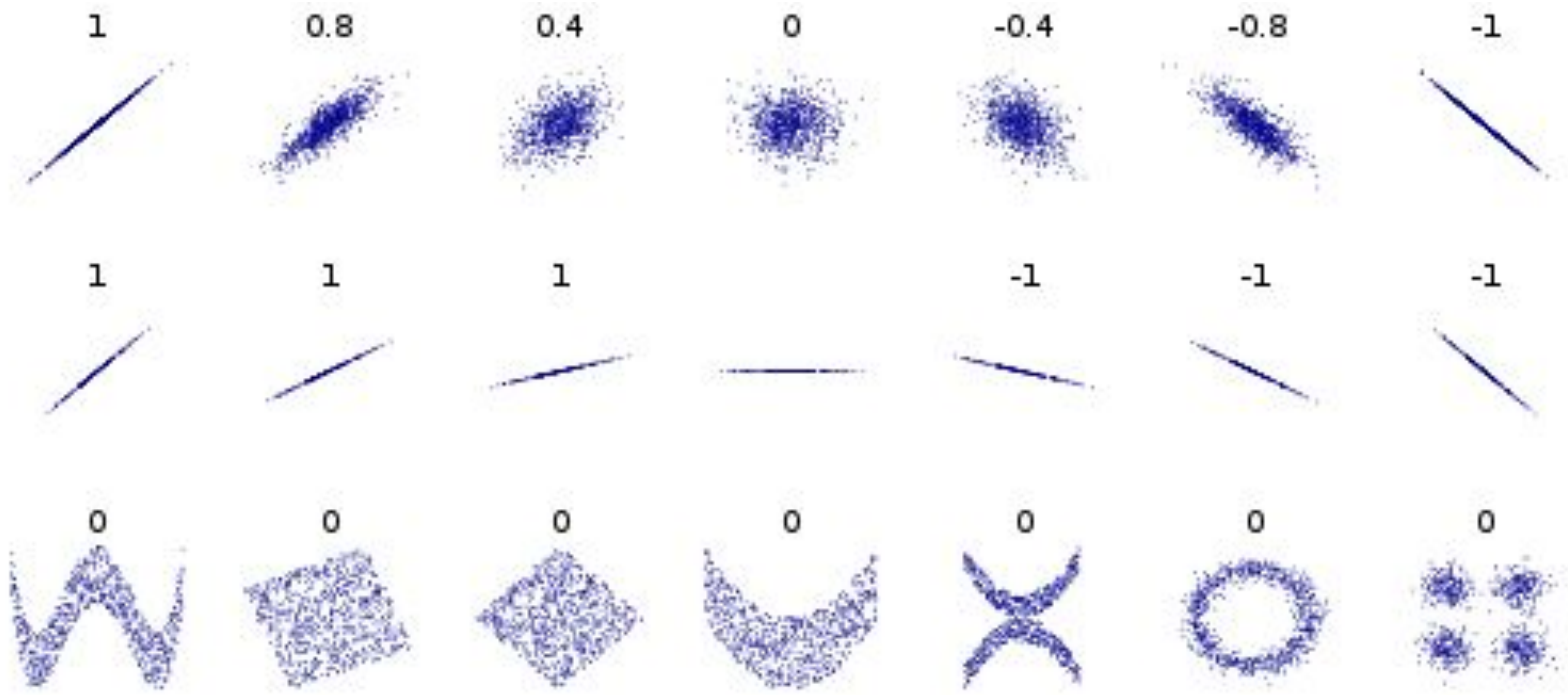
- $-1 \le r_{A,B} \le +1$, If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher the value, the stronger the correlation
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the correlation coefficient from –1 to 1.

Visually Evaluating Correlation



Correlation Coefficient (Numeric Data)

Example. $r_{AE,HT}$?

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

- Cov(AE, HT) = 7

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Correlation Coefficient (Numeric Data)

Example. $r_{AE,HT}$?

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

- Cov(AE, HT) = 7
- σAE = √2 ≈ 1.414
- σHT = √32.56 ≈ 5.706

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$r_{AE,HT}$ ≈ 7/(1.414*5.706) ≈ 0.868

Correlation Analysis (Nominal Data)

$r_{AE,HT} \approx 0.868$

t-test to assess the statistical significance of the correlation. $t = r \times \sqrt{\frac{n-2}{1-r^2}}$

$$t = 0.87 \times \sqrt{\frac{5-2}{1-0.87^2}} = 0.87 \times \sqrt{\frac{3}{1-0.7569}} = 0.87 \times \sqrt{\frac{3}{0.2431}}$$

$$\text{Degrees of freedom} = n - 2 = 5 - 2 = 3$$

$$t = 0.87 \times \sqrt{12.35} = 0.87 \times 3.515 = 3.06$$

For 3 degrees of freedom, the t-statistic required to reject the null hypothesis at the 0.05 significance level is 2.353. Based on our computed t-statistic, we can reject the null hypothesis that AllElectronics and HighTech are independent, indicating that the two features are correlated. **correlated**.

Table of Critical Values for Student's t-Distribution.

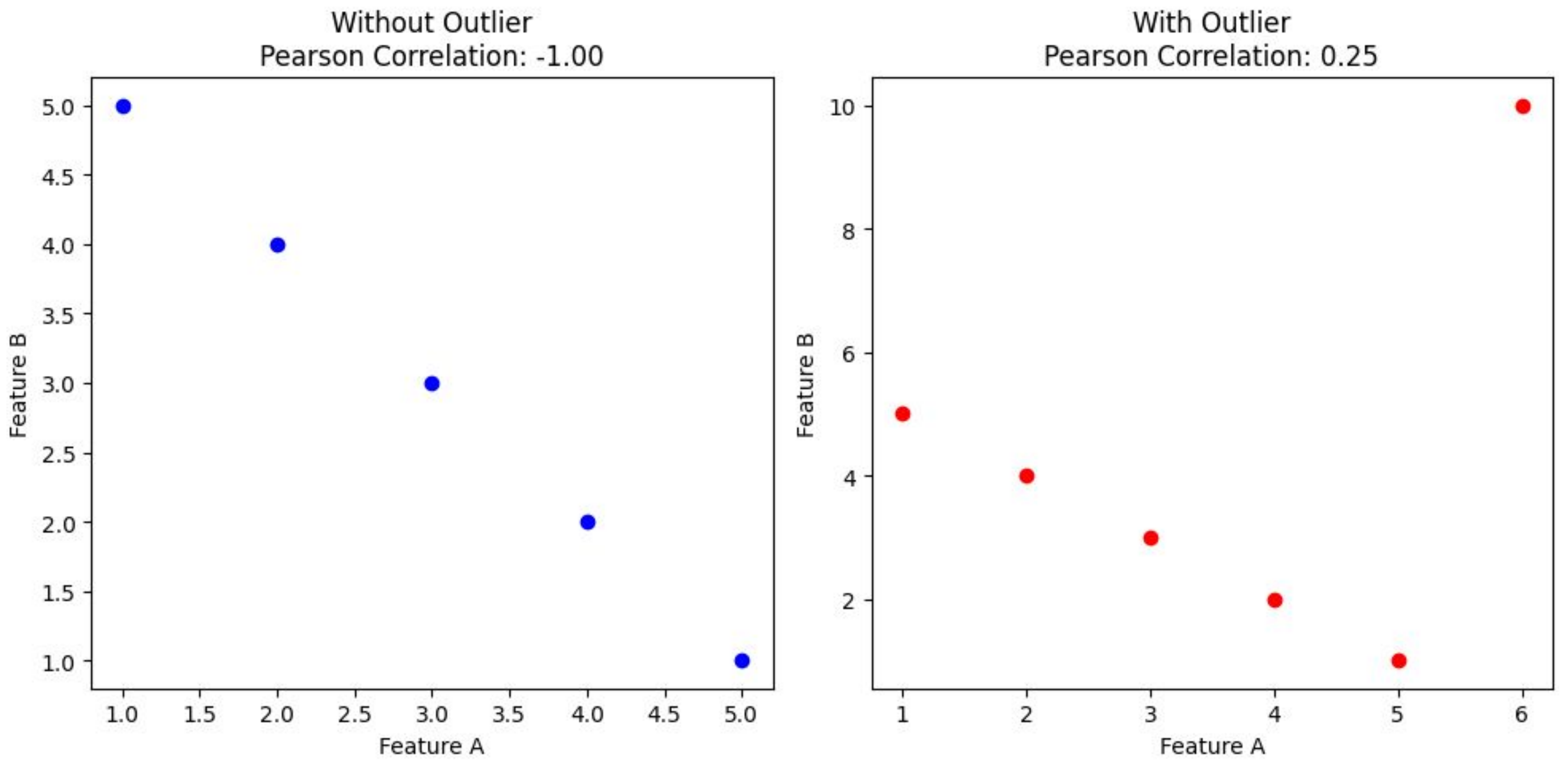
Correlation Coefficient (Numeric Data)

Assumptions of Pearson Correlation

- All objects (data points) should be **independent** of each other.
- The relationship between the two variables should be **linear**, not suitable for **nonlinear relationships**.
- Both attributes should be **continuous** and follow a **normal distribution**.

Limitations

- Pearson correlation is **extremely sensitive** to outliers.



Spearman's Rank Correlation (Numeric & Ordinal Data)

Spearman’s Rank Correlation Coefficient is a measure of the strength and direction of association between two ranked variables.

- It is a non-parametric test, meaning it does not assume a normal distribution of the data.
- It assesses how well the relationship between two variables can be described using a monotonic function (variables move in the same direction).

Spearman's Rank Correlation (Numeric & Ordinal Data)

Example.

Individual	Score in Test X	Score in Test Y
1	82	88
2	91	92
3	65	76
4	70	66
5	85	84

Spearman's Rank Correlation (Numeric & Ordinal Data)

Example.

Individual	Test X	Rank X	Test Y	Rank Y
1	82	3	88	3
2	91	1	92	1
3	65	5	76	5
4	70	4	66	4
5	85	2	84	2

Spearman's Rank Correlation (Numeric & Ordinal Data)

Example.

Individual	Rank X	Rank Y	di = Rank X - Rank Y
1	3	3	0
2	1	1	0
3	5	5	0
4	4	4	0
5	2	2	0

If no ties in the data

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 0}{5(5^2-1)} = 1 - 0 = 1$$

Perfect positive monotonic relationship between Test A and Test B scores

Otherwise

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

Where A is Rank X and B is Rank Y

Correlation Analysis (Nominal Data)

For nominal data, a correlation relationship between two attributes, A and B, can be discovered by a χ^2 (chi-square) test.

Correlation Analysis (Nominal Data)

Suppose A has **c** distinct values, namely a1,a2, …, ac. B has **r** distinct values, namely b1,b2, …, br. The data tuples described by A and B can be shown as a **contingency table**, with the **c values of A making up the columns** and the **r values of B making up the rows**. Let (A_i ,B_j) denote the joint event that attribute A takes on value ai and attribute B takes on value bj , that is, where (A = ai ,B = bj). Each and every possible (A_i ,B_j) joint event has its own cell (or slot) in the table.

Correlation Analysis (Nominal Data)

The χ^2 value (also known as the Pearson χ^2 statistic) is computed as

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250	200	
Nonfiction(nf)	50	1000	
Sum(col.)			

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i, B_j)

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250	200	?
Nonfiction(nf)	50	1000	?
Sum(col.)	?	?	?

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250	200	450
Nonfiction(nf)	50	1000	1050
Sum(col.)	300	1200	1500

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250 ($e_{f,g}$?)	200($e_{f,ng}$?)	450
Nonfiction(nf)	50($e_{nf,g}$?)	1000($e_{nf,ng}$?)	1050
Sum(col.)	300	1200	1500

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250 (90)	200($e_{f,ng}$?)	450
Nonfiction(nf)	50($e_{nf,g}$?)	1000($e_{nf,ng}$?)	1050
Sum(col.)	300	1200	1500

$$e_{f,g} = \frac{300 \times 450}{1500} = 90$$

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250 (90)	200(360)	450
Nonfiction(nf)	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250 (90)	200(360)	450
Nonfiction(nf)	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

Correlation Analysis (Nominal Data)

χ^2=507.93

The χ^2 statistic tests the hypothesis that A and B are independent, that is, there is no correlation between them. The test is based on a significance level, with $(r-1)\times(c-1)$ degrees of freedom. Since in this example, $r = 2$ and $c = 2$, the degrees of freedom are $(2-1)\times(2-1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the χ^2 distribution).

Based on our computed value, we can reject the hypothesis that play_game and preferred_reading are independent, so they are **correlated**.

Table of upper percentage points of the Chi-squared distribution

Partial correlation

Suppose we have the following data on three variables, *X*, *Y*, and *Z*:

X	Y	Z
2	1	0
6	3	0
10	2	1
20	4	1

These data have the feature that whenever $Z = 0$, *X* equals exactly twice *Y*, and whenever $Z = 1$, *X* is exactly 5 times *Y*. Thus, contingent on the value of *Z*, there is an exact relationship between *X* and *Y*; but the relationship cannot be said to be exact without reference to the value of *Z*.

In fact, if we compute the Pearson correlation coefficient between variables *X* and *Y*, the result is approximately 0.836, while if we compute the partial correlation between *X* and *Y*, using the formula given below, we find a partial correlation of 0.919, which is stronger than the full correlation.

Partial correlation

A simple way to compute the sample partial correlation for some data is to solve the two associated linear regression problems, get the residuals, and calculate the correlation between the residuals. Let X and Y be, as above, random variables taking real values, and let **Z** be the n-dimensional vector-valued random variable. We write x_i, y_i and $\mathbf{z}_i$ to denote the ith of N i.i.d. observations from some joint probability distribution over real random variables X, Y and **Z**, with $\mathbf{z}_i$ having been augmented with a 1 to allow for a constant term in the regression. Solving the linear regression problem amounts to finding (n+1)-dimensional regression coefficient vectors $\mathbf{w}_X^*$ and $\mathbf{w}_Y^*$ such that

$$\mathbf{w}_X^* = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (x_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\}$$

$$\mathbf{w}_Y^* = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\}$$

with N being the number of observations and $\langle \mathbf{w}, \mathbf{v} \rangle$ the scalar product between the vectors **w** and **v**.

The residuals are then

$$e_{X,i} = x_i - \langle \mathbf{w}_X^*, \mathbf{z}_i \rangle$$

$$e_{Y,i} = y_i - \langle \mathbf{w}_Y^*, \mathbf{z}_i \rangle$$

and the sample **partial** correlation is then given by the usual formula for sample correlation , but between these new *derived* values:

$$\hat{\rho}_{XY\cdot\mathbf{Z}} = \frac{N\sum_{i=1}^N e_{X,i}e_{Y,i} - \sum_{i=1}^N e_{X,i}\sum_{i=1}^N e_{Y,i}}{\sqrt{N\sum_{i=1}^N e_{X,i}^2 - \left(\sum_{i=1}^N e_{X,i}\right)^2}\ \sqrt{N\sum_{i=1}^N e_{Y,i}^2 - \left(\sum_{i=1}^N e_{Y,i}\right)^2}}.$$

Summary

- Feature Analysis: Relationships
  - Introduction to Feature Analysis
  - Covariance (for numerical features)
  - Correlation Coefficient (for numerical features)
  - Spearman's Rank Correlation (Numeric & Ordinal Data)
  - Chi-Square Test (for categorical features)