

Data Preprocessing

Data Integration

Data Integration

- **Data integration** is the process of combining the data from multiple sources into a coherent data store
- These sources may include multiple databases or flat files
- **Example:**
 - Temperature sensor, pressure sensor and rain gauge records **temperature**, **atmospheric pressure** and **amount of rain** at different locations
 - **Each location has separate** temperature, pressure and amount of rain tables (database)

Data Integration

- **Data integration** is the process of combining the data from multiple sources into a coherent data store
- These sources may include multiple databases or flat files
- **Example:**

1	Dates	Station Id	Temperature	Humidity	Rain
2	08-07-2018	t10	25.46875	82.1875	6.75
3	09-07-2018	t10	26.19298	85.42	1762
4	10-07-2018	t10	25.17021	85.34043	652.5
5	11-07-2018	t10	25.1368	87.68657	963
6	08-07-2018	t11	23.53846	61.92308	3
7	09-07-2018	t11	26.8494	85.42	15
8	10-07-2018	t11	25.1368	75.07463	13583
9	11-07-2018	t11	27.35915	76.02113	19769
10	23-07-2018	t12	25.1368	94.4065	1071
11	24-07-2018	t12	24.16197	97.66901	438.8
12	25-07-2018	t12	25.29323	94.84211	13667
13	26-07-2018	t12	22.19718	99	864

Data Integration

- **Data integration** is the process of combining the data from multiple sources into a coherent data store
- These sources may include multiple databases or flat files
- **Example:**
 - Temperature sensor, pressure sensor and rain gauge records temperature, pressure and amount of rain at different locations
 - Each location has separate temperature, pressure and amount of rain tables (database)
- **Issues to consider during data integration:**
 - Schema integration (entity matching)
 - Data value conflict
 - Redundancy

Schema Integration (Entity Matching)

- **Database schema:** The organization of data as a blueprint of how the database is constructed
- **Entity:** Each entity in real-world problem is the attribute in the database
- Addresses the question of
 - *"how can equivalent real-world entities from multiple sources be matched up?"*
 - *"how can data analysts be sure that they are same?"*
- **Attribute name conflict** across the multiple sources of data
 - Example: `customer_id`, `customer_num`, `cust_num`
- **Entity identification problem:**
 - Metadata is associated with each attribute
 - Metadata include:
 - Name, Meaning, Data type, Range of values permitted

Data Value Conflict

- **Issue:** Detection and resolution of data value conflicts
- For the same real-world entity, attribute values from different sources may differ
- This may be due to difference in representation, scaling, or encoding
- Example:
 - “weight” attribute may be stored in metric unit (gram, kilogram, etc.) in one system, British imperial unit (pound, ounce, etc.) in another system
 - In a database for hotel chain in different countries:
 - “price of room” attribute may be stored with price value in different currencies
 - Categorical data: “gender” may be stored with male and female or M and F

Redundancy

- Major issue to be addressed
- Sources of redundancy:
 - An attribute may be redundant, if it can be derived from another attribute or set of attributes
 - **Example:** Attribute “Total Marks” derived from Marks from each courses
 - **Inconsistency in the attribute naming** can also cause redundancy in resulting data sets
 - **Example:** (1) `registration_id` and `roll_num`
(2) `customer_id` and `customer_num`
- Two types of redundancies:
 - **Redundancy between the attributes**
 - **Redundancy at the tuple level**
 - Duplication of tuples
 - Remove the duplicate tuples

Redundancy Between Attributes

- Two attributes may be related or dependent
- Detected by the correlation analysis
- Correlation analysis measures how strongly one attribute implies (related) to other, based on available data
- Correlation analysis for numerical attributes:
 - Compute correlation coefficient between two attributes A and B (e.g. Pearson's product moment coefficient i.e. Pearson's correlation coefficient)
- Correlation analysis for categorical attributes:
 - Correlation (relationship) between two categorical attributes A and B can be discovered by χ^2 (chi-square) test

Redundancy Between Numerical Attributes

- Pearson's correlation coefficient ($\rho_{A,B}$):

$$\rho_{A,B} = \frac{\frac{1}{N} \sum_{i=1}^N (a_i - \mu_A)(b_i - \mu_B)}{\sigma_A \sigma_B} = \frac{\text{Cov}(A, B)}{\sigma_A \sigma_B}$$

- N : number of tuples
 - a_i and b_i : respective values of attribute A and attribute B in tuple i
 - μ_A and μ_B : respective mean values of A and B
 - σ_A and σ_B : respective standard deviation of A and B
 - $\text{Cov}(A, B)$: Covariance between A and B
- Note: $-1 \leq \rho_{A,B} \leq +1$

Redundancy Between Numerical Attributes: Pearson's correlation coefficient

- If $\rho_{A,B}$ is greater than 0, then attributes A and B are positively correlated
 - The values of A increases as the values of B increases or vice versa
 - The higher the value, the stronger the correlation
- If $\rho_{A,B}$ is equal to 0, then attributes A and B have no correlation between them (may be independent)
- If $\rho_{A,B}$ is less than 0, then attributes A and B are negatively correlated
 - The values of A increases as the values of B decreases or vice versa
 - Each attribute discourages the other
 - The higher the value, the stronger the correlation
- A higher correlation value may indicate that A (or B) may be removed as a redundancy

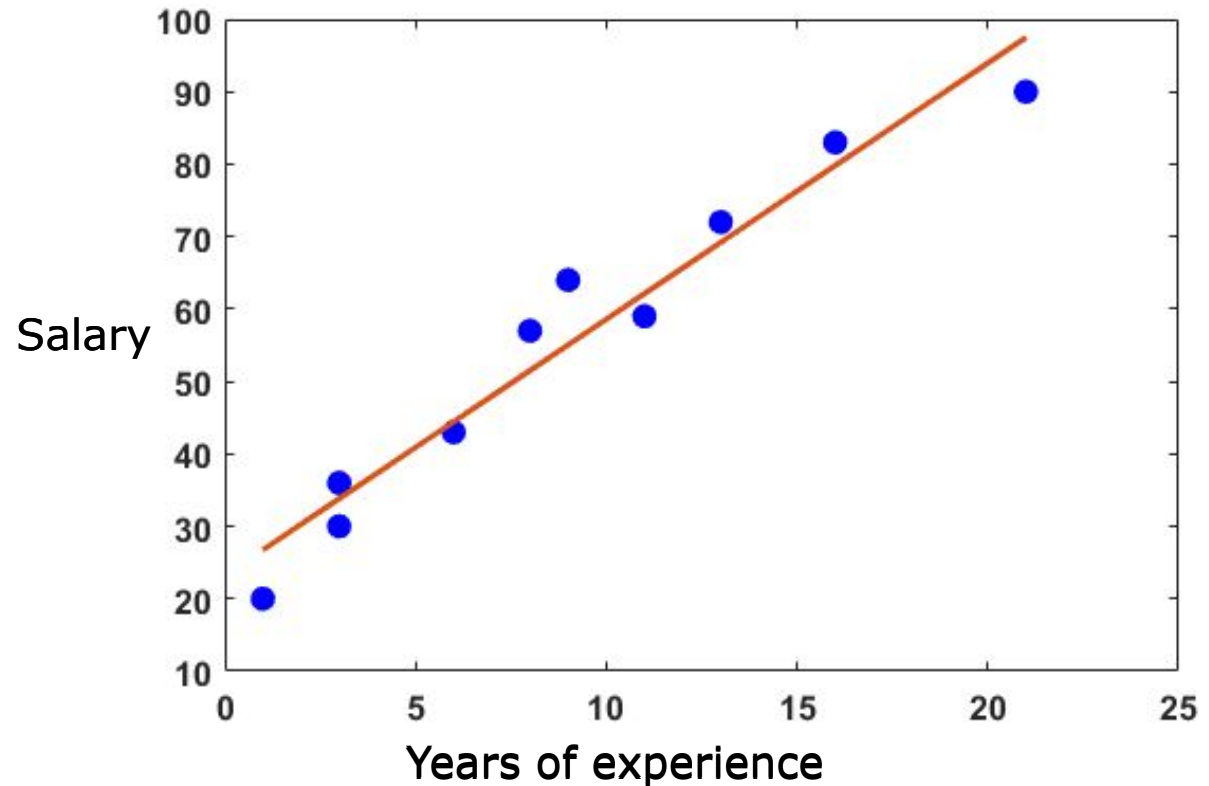
Redundancy Between Numerical Attributes: Pearson's correlation coefficient

- Assumption:
 - Both attributes (variables) should be **normally distributed** (normally distributed variables (normal distribution) have a bell-shaped curve)
 - **Linearity**: The two attributes have linear relationship
 - **Homoscedasticity**: Data is equally distributed about the regression line.
- **Scatter plots** can also be use to view correlation between the numerical attributes

Illustration of Pearson's Correlation Coefficient

Years of experience (x)	Salary (in Rs 1000) (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

- Scatter plot



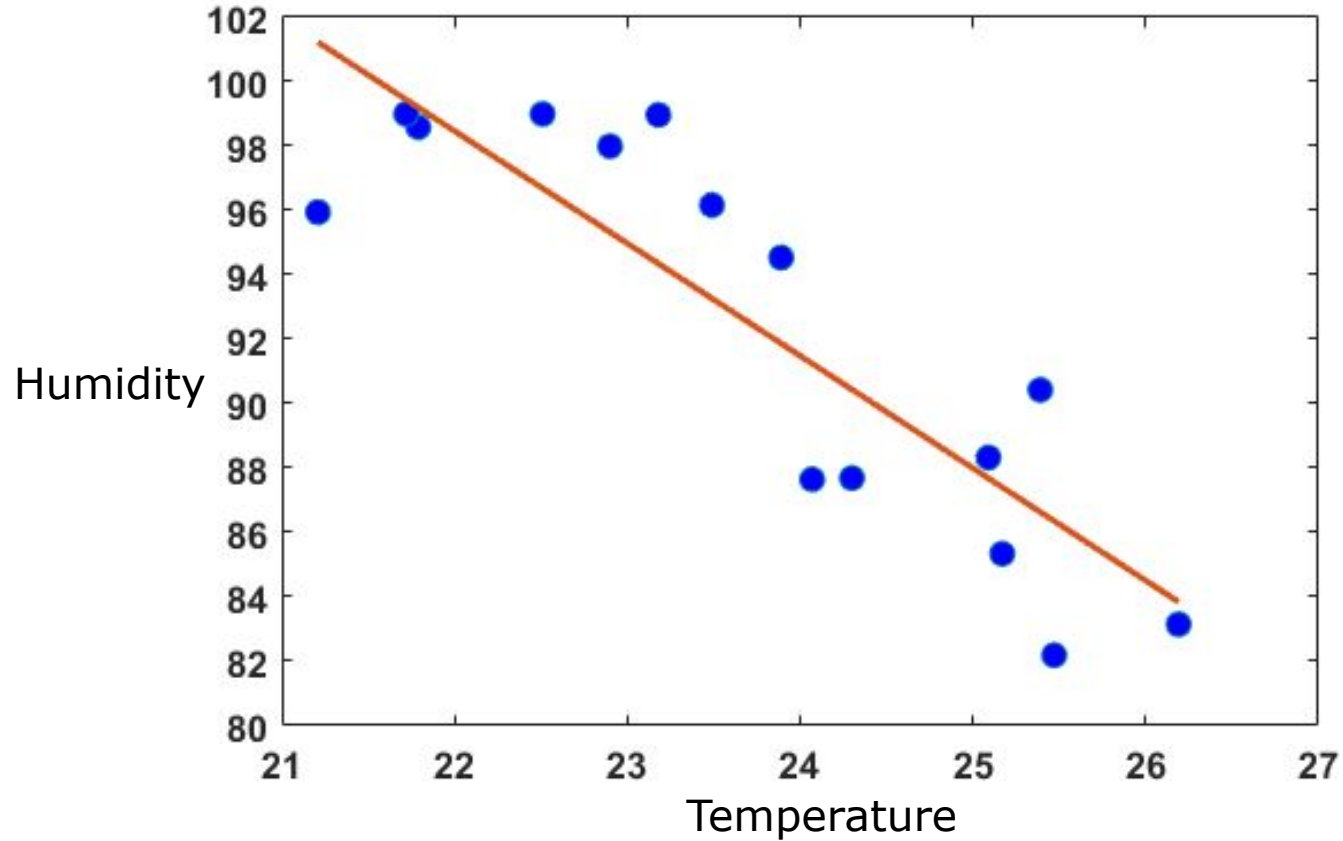
- The two attributes have linear relationship
- Data is equally distributed about the regression line (roughly)

$$\rho_{A,B} = 0.97$$

Illustration of Pearson's Correlation Coefficient

Temp (x)	Humidity (y)
25.47	82.19
26.19	83.15
25.17	85.34
24.30	87.69
24.07	87.65
21.21	95.95
23.49	96.17
21.79	98.59
25.09	88.33
25.39	90.43
23.89	94.54
22.51	99.00
22.90	98.00
21.72	99.00
23.18	98.97

- Scatter plot

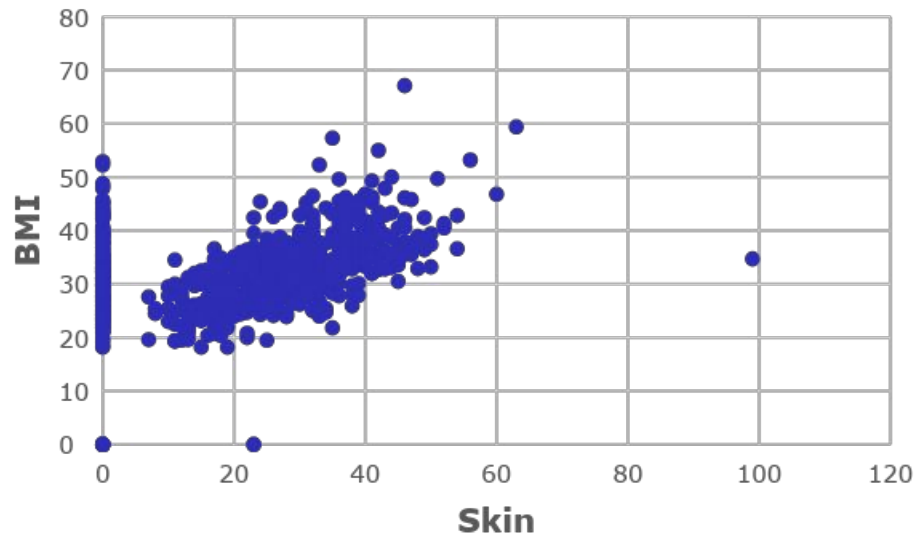


$$\rho_{A,B} = -0.8648$$

Illustration of Pearson's Correlation Coefficient: **Pima-Indians-Diabetes** Dataset

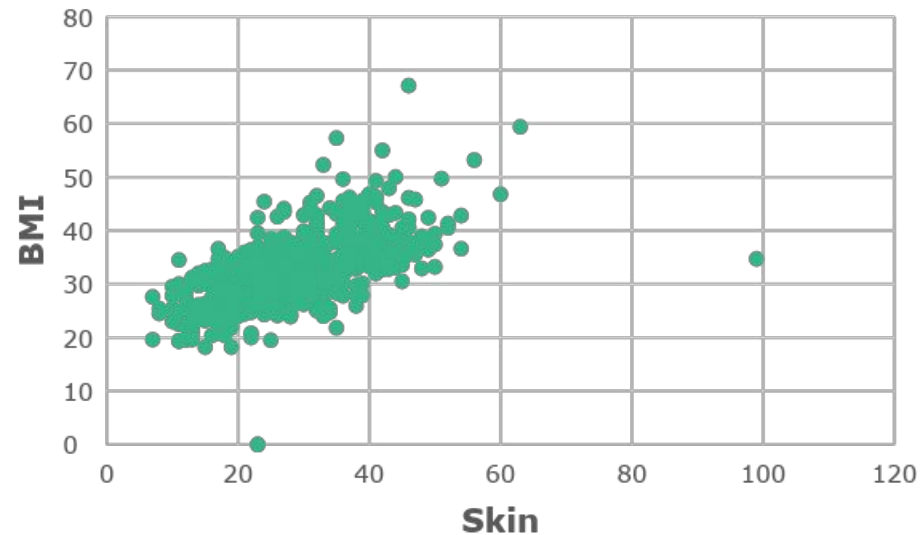
- Scatter plot between BMI (A) and Skin (B)

Before cleaning the data



$$\rho_{A,B} = 0.3926$$

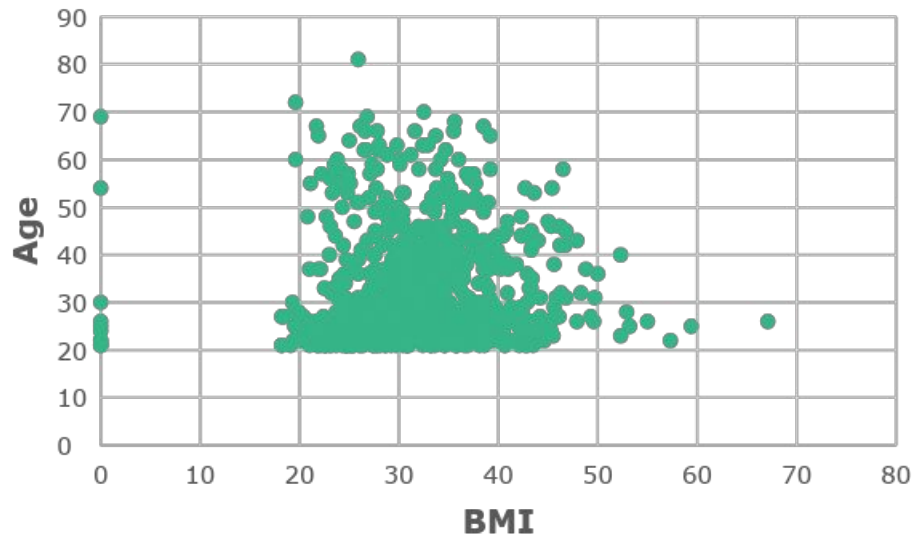
After cleaning the data



$$\rho_{A,B} = 0.6320$$

Illustration of Pearson's Correlation Coefficient: **Pima-Indians-Diabetes** Dataset

- Scatter plot between BMI (A) and Age (B)



$$\rho_{A,B} = 0.0564$$

Redundancy Between Numerical Attributes: Spearman Rank Correlation

- Rank correlation between variables: Statistical dependence between the rankings of two variables
 - The values the variables take should be at least ordinal
- The values in the attributes should be converted into ranks of the values (ordinal values), if the attribute is not ordinal
- Spearman rank correlation is a non-parametric measure of rank correlation between two attributes (variables)
- As it is non-parametric measure, it does not carry any assumptions about the distribution of the data
- The Spearman correlation coefficient is defined as the *Pearson correlation coefficient* between the rank variables

Redundancy Between Numerical Attributes: Spearman Rank Correlation

- Spearman correlation coefficient (ρ_{R_A, R_B}):

$$\rho_{R_A, R_B} = \frac{\text{Cov}(R_A, R_B)}{\sigma_{R_A} \sigma_{R_B}}$$

- R_A and R_B : ranks attribute A and attribute B
- σ_{R_A} and σ_{R_B} : respective standard deviation of ranks of A and B
- $\text{Cov}(R_A, R_B)$: Covariance between the ranks of A and B
- Only if all N ranks are *distinct integers*, then it can be computed using the popular formula

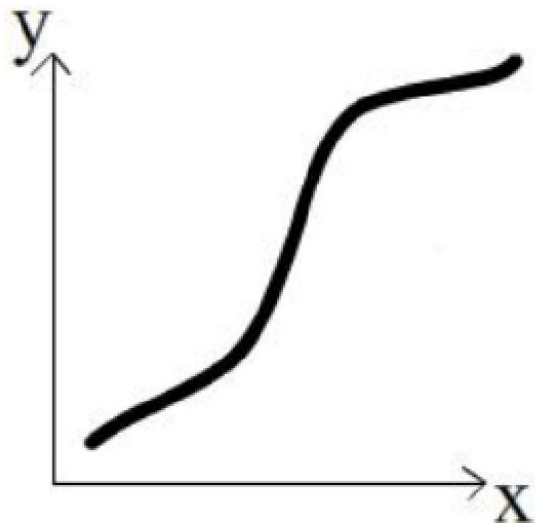
$$\rho_{R_A, R_B} = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)}$$

- N : number of tuples
- d_i : difference between the rank values of A and B in tuple i

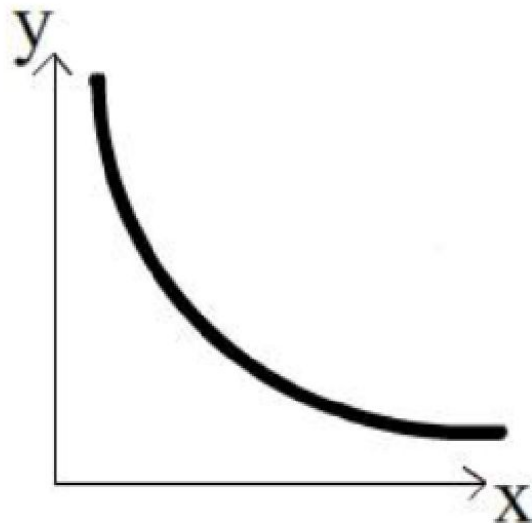
$$-1 \leq \rho_{R_A, R_B} \leq +1$$

Redundancy Between Numerical Attributes: Spearman Rank Correlation

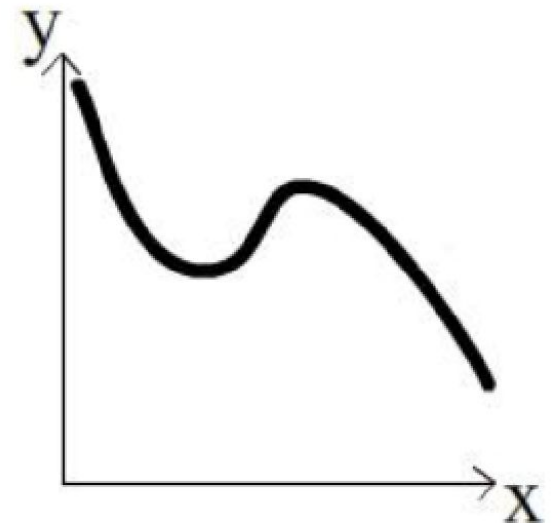
- Pearson's correlation assesses linear relationships
- Spearman's correlation assesses monotonic relationships (whether linear or not)



Monotonically increasing



Monotonically decreasing



Not monotonic

Illustration of Spearman's Correlation Coefficient

Years of experience (x)	Salary (in Rs 1000) (y)	R_x	R_y
3	30	2	2
8	57	4	5
9	64	5	7
13	72	7	8
3	36	2	3
6	43	3	4
11	59	6	6
21	90	9	10
1	20	1	1
16	83	8	9

$$\rho_{R_x, R_y} = 0.9806$$

- Convert the values of both attribute into rank values

Illustration of Spearman's Correlation Coefficient

Temp (x)	Humidity (y)	R _x	R _y
25.47	82.19	14	1
26.19	83.15	15	2
25.17	85.34	12	3
24.30	87.69	10	5
24.07	87.65	9	4
21.21	95.95	1	9
23.49	96.17	7	10
21.79	98.59	3	12
25.09	88.33	11	6
25.39	90.43	13	7
23.89	94.54	8	8
22.51	99.00	4	14
22.90	98.00	5	11
21.72	99.00	2	14
23.18	98.97	6	13

$$\rho_{R_x, R_y} = -0.8523$$

Correlation Between Numerical Attributes: Summary

- Pearson's correlation coefficient is applied on the two continuous valued attributes
- Spearman's correlation coefficient is applied on the two ranked valued (ordinal) discrete attributes

ρ range in positive correlation	ρ range in negative correlation	Correlation between A and B
0.0	0.0	None
(0.0 0.1]	(-0.0 -0.1]	Weak
(0.1 0.3]	(-0.1 -0.3]	Moderate
(0.3 0.5]	(-0.3 -0.5]	Strong
(0.5 1.0)	(-0.5 -1.0)	Very Strong
1.0	-1.0	Perfect

Redundancy Between Categorical (Discrete) Attributes

- Correlation relationship between two categorical attributes A and B can be discovered by χ^2 (chi-square) test
- Steps in χ^2 (chi-square) test :
 - Identify the two categorical attributes
 - **Null hypothesis**: Two attributes are independent (not related)
 - Complete the **contingency matrix** (table) with observed frequencies (count) and expected frequencies (probability)
 - Calculate the **observed χ^2 value** based on contingency matrix
 - Use the standard **χ^2 table** compare if the **observed χ^2 value** to **critical χ^2 value** for the problem's **degree of freedom** and **confidence (significance i.e. p-value) level**
 - If the **observed χ^2 value** < **critical χ^2 value** then the attributes are not related (null-hypothesis is true)

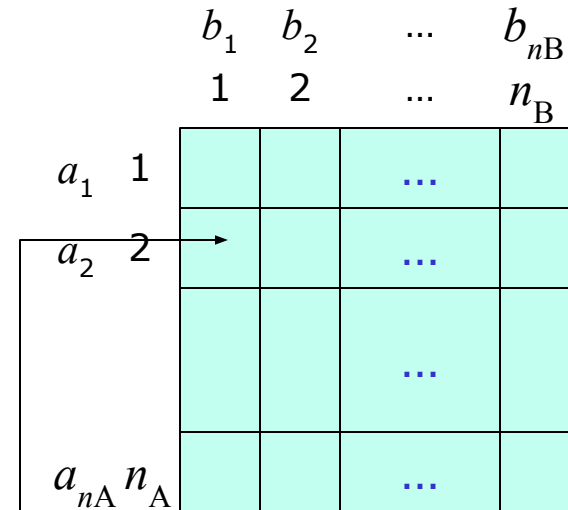
Redundancy Between Categorical (Discrete) Attributes

- Correlation relationship between two categorical attributes A and B can be discovered by χ^2 (chi-square) test
- Suppose attribute A has n_A distinct value $(a_1, a_2, \dots, a_i, \dots, a_{n_A})$
- **Example:** Let the attribute be `gender`
 - The distinct values `gender` can take are *male* and *female*
 - Number of distinct values: 2 i.e. $n_A = 2$

Redundancy Between Categorical (Discrete) Attributes

- Correlation relationship between two categorical attributes A and B can be discovered by χ^2 (chi-square) test
- Suppose attribute A has n_A distinct value ($a_1, a_2, \dots, a_i, \dots, a_{n_A}$)
- Suppose attribute B has n_B distinct value ($b_1, b_2, \dots, b_j, \dots, b_{n_B}$)
- The data tuples described by attributes A and B can be shown as a contingency table

- Contingency table has
 - n_A distinct values of A making up the rows
 - n_B distinct values of B making up the columns



		b_1	b_2	...	b_{n_B}
		1	2	...	n_B
a_1	1			...	
a_2	2			...	
				...	
a_{n_A}	n_A			...	

(a_i, b_j) denote event that i^{th} distinct value of A and j^{th} distinct value of B taken on jointly

Redundancy Between Categorical (Discrete) Attributes

- The observed χ^2 (chi-square) value (Pearson χ^2 statistics) is computed as

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

- o_{ij} : observed frequency (actual count) of joint event (A_i, B_j)
 - Number of times the i^{th} distinct value of attribute A is occurring jointly with j^{th} distinct value of attribute B
- e_{ij} : expected frequency (probability) of joint event (A_i, B_j)
- Expected frequency (e_{ij}), i.e., probability that i^{th} distinct value of attribute A is occurring jointly with j^{th} distinct value of attribute B, is computed as

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

- $\text{Count}(A = a_i)$: The number of tuple having distinct value a_i for A
- $\text{Count}(B = b_j)$: The number of tuple having distinct value b_j for B

- N : number of tuples

Redundancy Between Categorical (Discrete) Attributes

- The χ^2 statistic tests the hypothesis that A and B are independent or not related (Null hypothesis)
- The test is based on the significance level (p-value), with $(n_A - 1) * (n_B - 1)$ degree of freedom
 - p-value gives the evidence against the Null hypothesis
 - Smaller the p-value, stronger the evidence (confidence)
 - Example: p-value = 0.01 means 99% confidence that you can accept/reject the Null hypothesis
 - Degree of freedom (statistics): The number of values in the final calculation of a statistic that are free to vary
 - Usually one less than the number of items
- If the hypothesis can be rejected, then we say that A and B are statistically related or associated for the given data set

Redundancy Between Categorical Attributes: Illustration

- A group of 15 people are surveyed
- The gender of each person is noted
- Each person is polled as to whether their preferred type of reading material was fiction or nonfiction
- This leads to two attributes gender and preferred_reading
 - gender takes two distinct values male and female
 - preferred_reading takes two distinct values fiction and non-fiction

Redundancy Between Categorical Attributes: Illustration

Sl. No.	gender	preferred_reading
1	male	fiction
2	female	Non-fiction
3	female	Non-fiction
4	male	Non-fiction
5	female	fiction
6	female	Non-fiction
7	male	Non-fiction
8	male	fiction
9	female	Non-fiction
10	female	fiction
11	male	fiction
12	female	Non-fiction
13	female	Non-fiction
14	male	fiction
15	male	Non-fiction

- A group of 15 people are surveyed
- Size of the contingency matrix is 2 x 2

	fiction (b_1)	Non-fiction (b_2)	Total
male (a_1)	4 (o_{11})	3 (o_{12})	7
female (a_2)	2 (o_{21})	6 (o_{22})	8
Total	6	9	15

Redundancy Between Categorical Attributes: Illustration

- A group of 15 people are surveyed
- Size of the contingency matrix is 2 x 2

	fiction (b_1)	Non-fiction (b_2)	Total
male (a_1)	4 (o_{11}) 2.8 (e_{11})	3 (o_{12}) 4.2 (e_{12})	7
female (a_2)	2 (o_{21}) 3.2 (e_{21})	6 (o_{22}) 4.8 (e_{22})	8
Total	6	9	15

$$e_{11} = \frac{\text{count(male)} \times \text{count(fiction)}}{N} = \frac{7 \times 6}{15} = 2.8$$

$$e_{12} = \frac{\text{count(male)} \times \text{count(nonfiction)}}{N} = \frac{7 \times 9}{15} = 4.2$$

$$e_{21} = \frac{\text{count(female)} \times \text{count(fiction)}}{N} = \frac{8 \times 6}{15} = 3.2$$

$$e_{22} = \frac{\text{count(female)} \times \text{count(nonfiction)}}{N} = \frac{8 \times 9}{15} = 4.8$$

Sl. No.	gender	preferred _reading
1	male	fiction
2	female	Non-fiction
3	female	Non-fiction
4	male	Non-fiction
5	female	fiction
6	female	Non-fiction
7	male	Non-fiction
8	male	fiction
9	female	Non-fiction
10	female	fiction
11	male	Non-fiction
12	female	Non-fiction
13	female	Non-fiction
14	male	fiction
15	male	Non-fiction

Redundancy Between Categorical Attributes: Illustration

	fiction (b_1)	Non-fiction (b_2)	Total
male (a_1)	4 (o_{11}) 2.8 (e_{11})	3 (o_{12}) 4.2 (e_{12})	7
female (a_2)	2 (o_{21}) 3.2 (e_{21})	6 (o_{22}) 4.8 (e_{22})	8
Total	6	9	15

- The numbers in blue are the expected frequencies (probability)
- The χ^2 value is computed as

$$\chi^2 = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}}$$

$$\chi^2 = \frac{(4 - 2.8)^2}{2.8} + \frac{(3 - 4.2)^2}{4.2} + \frac{(2 - 3.2)^2}{3.2} + \frac{(6 - 4.8)^2}{4.8}$$

$$\chi^2 = 1.607$$

Redundancy Between Categorical Attributes: Illustration

- For 2 x 2 contingency table, the degree of freedom is $(2-1)*(2-1) = 1$
- Obtain the χ^2 value for 0.05 significance i.e. $p=0.05$ (95% chance or confidence) with 1 degree of freedom
 - χ^2 value is 3.841 (Taken from the table of χ^2 distribution)
- Computed χ^2 value for given population is 1.607
- The computed value is less than the 3.841
 - We accept the hypothesis that gender and preferred_reading are independent (not related)
- **Conclusion:** The two attributes (gender and preferred_reading) are not correlated for the given group of people