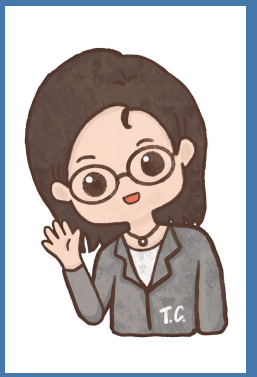
MTH 208 Exploratory Data Analysis

Lesson 06: Pattern Recognition and Association Analysis

Ying-Ju Tessa Chen, PhD

Associate Professor
Department of Mathematics
University of Dayton

@ying-ju
ying-ju
ychen4@udayton.edu



Learning Objectives

Overview

We study advanced concepts of identifying patterns, correlations, and associations within datasets in this lession. Building upon the foundation of reading scatterplots introduced earlier, we will explore techniques and methodologies to recognize underlying patterns and associations that are not immediately apparent.

Objectives

- Understand the difference between correlation and causation.
- Learn to identify and interpret various types of patterns in data.
- Explore methods to analyze association between categorical variables.
- Apply statistical measures to quantify relationships in data.

Correlation vs. Causation

- **Correlation:** A statistical measure that expresses the extent to which two variables change together. Correlation does not imply that one variable causes the change in another.
    - **Example:** Ice cream sales and drowning incidents are positively correlated, but one does not cause the other. Instead, they are both related to a third factor: warmer temperatures during summer months.
- **Causation (Causal Relationship):** A relationship where one variable directly affects another.
    - **Example:** A decrease in vaccination rates causes an increase in the spread of diseases that those vaccines prevent.

Correlation vs. Causation (Continued)

How to report correlation?

- `Bad:` Raising salaries increases productivity.
- `Good:` Employees with higher salaries tend to be more productive.
- `Bad:` $r = -0.99$. This proves that drinking more red wine lowers cholesterol.
- `Good:` There is a strong negative association between red wine consumption and cholesterol levels.
- `Bad:` A child that has two educated parents will graduate from college.
- `Good:` Children with educated parents are more likely to graduate from college.
- The vermiform appendix impacts the risk of developing Parkinson's disease
    - Appendix Removal Lowers Parkinson's Disease Risk by up to 25%
    - Appendix identified as a potential starting point for Parkinson's disease
- PARKINSON'S DISEASE IS MORE PREVALENT IN PATIENTS WITH APPENDECTOMIES: A NATIONAL POPULATION-BASED STUDY
    - Appendix Removal Associated with Development of Parkinson's Disease

Pattern Recognition in Scatter Plots

Scatter plots are a fundamental tool in exploratory data analysis, offering a visual representation of the relationship between two quantitative variables. Beyond simple linear correlations, scatter plots can reveal a variety of patterns that provide deeper insights into the data:

- **Linear Relationships:** A straight-line pattern indicating a positive or negative correlation.
- **Non-linear Relationships:** Curved patterns suggest a more complex relationship that might require transformation or different analytical approaches.
- **Clusters:** Groups of points that are closely bunched together, indicating subpopulations within the dataset.
- **Outliers:** Points that fall far from the main group of data points, which may indicate anomalies or errors in the data.

Association Analysis in Categorical Data

Introduction to Chi-Square Tests The Chi-square test of independence is a non-parametric statistical test used to determine if there is a significant association between two categorical variables from the same population. It's commonly applied in survey research, contingency table analysis, and various fields requiring statistical analysis of categorical data.

H0: The two variables are independent.

H1: The two variables relate to each other.

Key Concepts

- `Categorical Data:` Data that can be categorized into groups or categories that do not have a natural order or ranking. - Examples include gender, race, or a yes/no response.
- `Contingency Tables:` Also known as cross-tabulation tables or two-way tables, contingency tables display the frequency distribution of variables and are a key part of conducting a Chi-square test.
- `Expected Frequencies:` The frequencies we would expect in each category if there was no association between the variables.
- `Chi-square Statistic:` A measure that tells us how far the observed frequencies are from the expected frequencies. A higher value indicates a greater discrepancy and potentially a significant association.

Association Analysis in Categorical Data (Continued)

Example: Effectiveness of a Drug Treatment

Assume that there are 105 patients in the study and 50 of them were treated with the drug. In addition, the remaining 55 patients were in the control group. All patients' health condition was checked after a week. Here is an example using R code.

```r
# Install and load necessary package
if (!require("gmodels")) install.packages("gmodels")
library(gmodels)

# Read the dataset
drug_data <- read_csv("https://goo.gl/j6lRXD")

# Print the contingency table
(drug_table <- table(drug_data[,2:3]))
```

```
##              improvement
## treatment     improved not-improved
##   not-treated       26           29
##   treated           35           15
```

```r
# Perform the Chi-Square test
chi_square_result <- chisq.test(drug_table,
                                correct=TRUE)

# Print the results
print(chi_square_result)
```

```
##
##     Pearson's Chi-squared test with Yates' continuity correct
##
## data:  drug_table
## X-squared = 4.6626, df = 1, p-value = 0.03083
```

Since p-value is < 0.05, we reject the null hypothesis. We have sufficient evidence to conclude that the treatment and improvement are associated.

Note: Use the `correct=FALSE` option with reasonably large sample sizes, ie., if expected counts in any of the cells in the contingency table have more than 5 observations.

Association Analysis in Categorical Data (Continued)

Python code for the same example.

```python
import numpy as np
import pandas as pd
from scipy.stats import chi2_contingency

df = pd.read_csv("https://goo.gl/j6lRXD")

# Create a contingency table
contingency_table = pd.crosstab(df['treatment'], df['improvement

print("Contingency Table:")
print(contingency_table)

# Perform the Chi-Square test
chi2, p, dof, expected = chi2_contingency(contingency_table)

print(f"\nChi2 Statistic: {chi2}")
print(f"\nDegrees of Freedom: {dof}")
print(f"\np-value: {p}")
print("Expected Frequencies:")
print(expected)
```

```
## Contingency Table:
```

```
## improvement  improved  not-improved
## treatment
## not-treated        26            29
## treated            35            15
```

```
##
## Chi2 Statistic: 4.6625668947297125
```

```
##
## Degrees of Freedom: 1
```

```
##
## p-value: 0.030827072412198585
```

```
## Expected Frequencies:
```

```
## [[31.95238095 23.04761905]
##  [29.04761905 20.95238095]]
```

Quantifying Relationships

Three key statistical measures used to quantify these relationships are the Pearson correlation coefficient, the Spearman rank correlation coefficient, and the Kendall tau rank correlation coefficient.

Pearson Correlation Coefficient (r)

- **Definition:** The Pearson correlation coefficient measures the linear relationship between two continuous variables. It ranges from -1 to 1, where 1 means a perfect positive linear relationship, -1 means a perfect negative linear relationship, and 0 means no linear relationship.

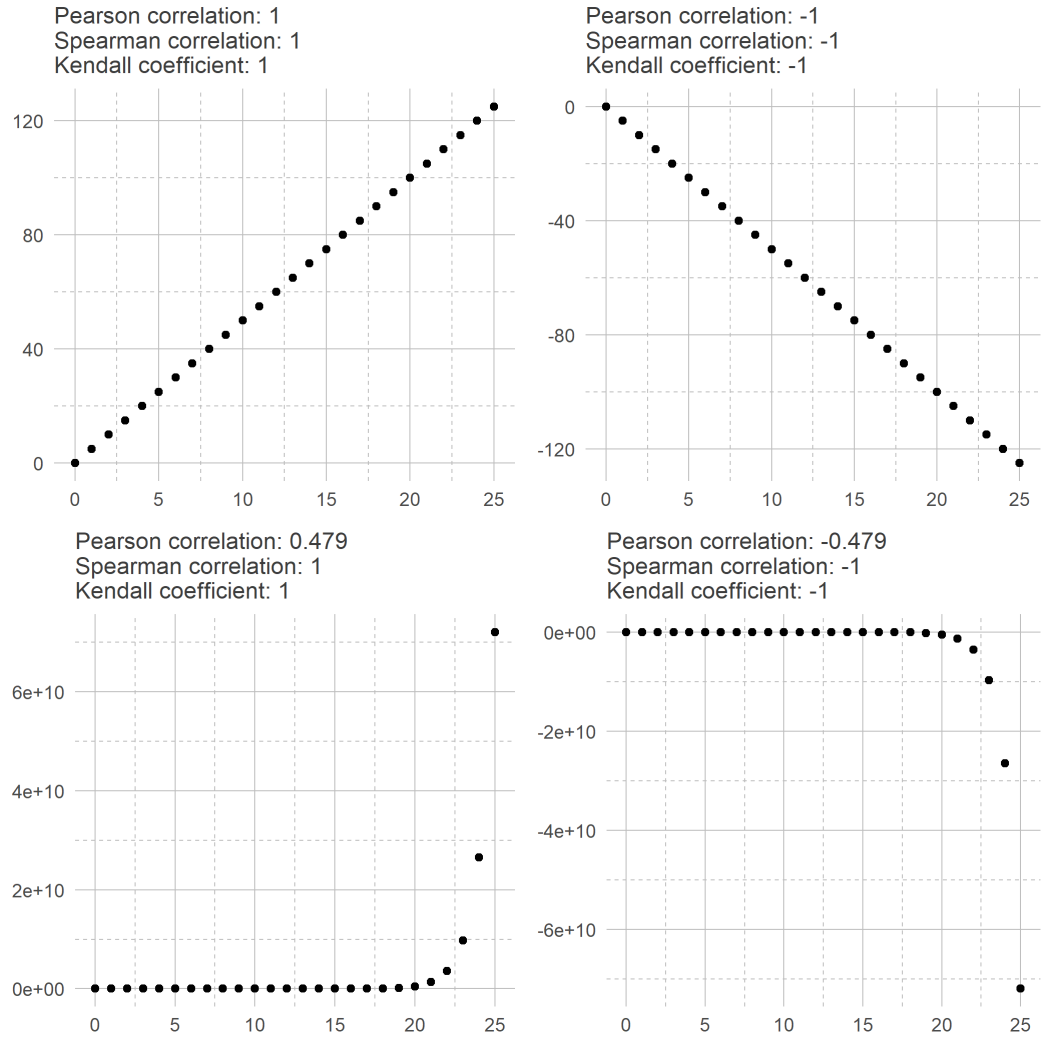
Spearman Rank Correlation Coefficient

- **Definition:** The Spearman correlation coefficient is a non-parametric measure of the strength and direction of the association that exists between two variables measured on at least an ordinal scale. It assesses how well the relationship between two variables can be described using a monotonic function.

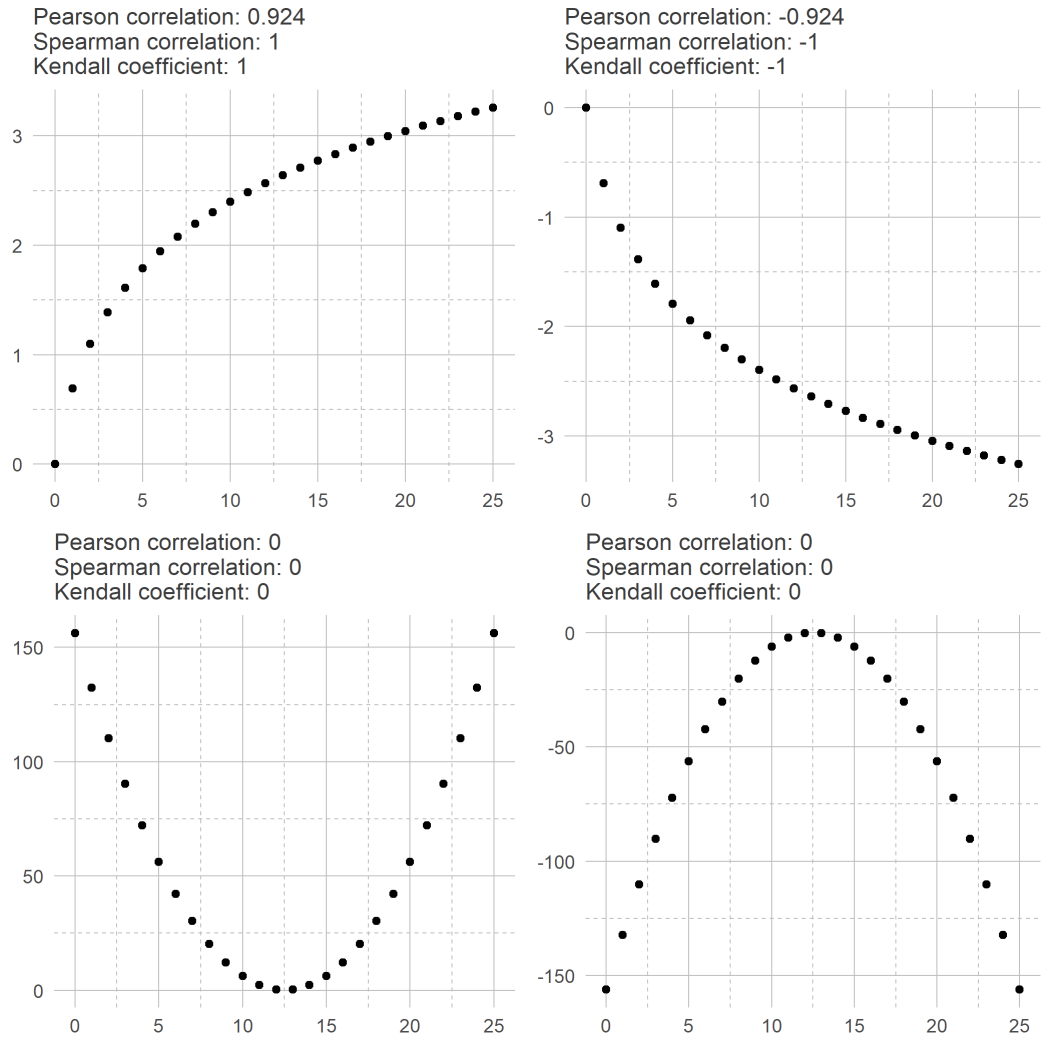
Kendall tau rank correlation coefficient

- **Definition:** The Kendall tau rank correlation coefficient, often referred to as Kendall's tau coefficient, is another non-parametric measure used to quantify the association between two measured quantities. It assesses the strength and direction of a relationship between two variables. Like Spearman's rho, it is useful for ordinal data or data that do not meet the assumptions of linearity and normal distribution required for Pearson's correlation coefficient. Kendall's tau is particularly well-suited for small datasets or datasets with a lot of tied ranks.

Quantifying Relationships (Continued)



Quantifying Relationships (Continued)



Quantifying Relationships (Continued)

We calculate the Pearson, Spearman, and Kendall's tau correlation coefficients between sepal length and sepal width using R code:

```
# Pearson correlation between sepal length and sepal width
pearson_sepal <- cor(iris$Sepal.Length, iris$Sepal.Width, method="pearson")

# Spearman correlation between sepal length and sepal width
spearman_sepal <- cor(iris$Sepal.Length, iris$Sepal.Width, method = "spearman")
cat("Pearson Correlation Coefficient (Sepal):", pearson_sepal, "\n")
```

```
## Pearson Correlation Coefficient (Sepal): -0.1175698
```

```
cat("Spearman Correlation Coefficient (Sepal):", spearman_sepal, "\n")
```

```
## Spearman Correlation Coefficient (Sepal): -0.1667777
```

```
cor.test(iris$Sepal.Length, iris$Sepal.Width, method="pearson")
```

```
##
##     Pearson's product-moment correlation
##
## data:  iris$Sepal.Length and iris$Sepal.Width
```

Quantifying Relationships (Continued)

The same example using Python code:

```
import seaborn as sns
import scipy.stats as stats
# Load the Iris dataset
iris = sns.load_dataset('iris')
pearson_coef, p_value = stats.pearsonr(iris['sepal_length'], iris['sepal_width'])
print(f"Pearson Correlation Coefficient (Sepal): {pearson_coef:.3f}, P-value: {p_value:.3f}")
```

```
## Pearson Correlation Coefficient (Sepal): -0.118, P-value: 0.152
```

```
spearman_coef, p_value = stats.spearmanr(iris['sepal_length'], iris['sepal_width'])
print(f"Spearman Correlation Coefficient (Sepal): {spearman_coef:.3f}, P-value: {p_value:.3f}")
```

```
## Spearman Correlation Coefficient (Sepal): -0.167, P-value: 0.041
```

Advanced Correlation Techniques

Partial Correlation

- `Definition:` Partial correlation measures the strength and direction of the relationship between two variables while controlling for the effect of one or more additional variables.
    - `Applicability:` Useful when you want to understand the direct relationship between two variables, independent of other variables that might affect their association.

Autocorrelation (Serial Correlation)

- `Definition:` Autocorrelation refers to the correlation of a variable with itself across different points in time. It's a measure of how related the values of a dataset are with its previous values.
    - `Applicability:` Particularly relevant in time-series analysis where the goal is to identify patterns or trends over time.

Advanced Correlation Techniques (Continued)

Here we use R code to find the Pearson partial correlation coefficition between Sepal.Length and Sepal.Width while controlling the effect of Petal.Letngh and Petal.Width.

```
library(ppcor)
result <- pcor.test(iris$Sepal.Length, iris$Sepal.Width,
                    iris[,c("Petal.Length", "Petal.Width")],
                    method="pearson")
print(result)
```

```
##    estimate      p.value statistic   n gp  Method
## 1 0.6285707 1.199846e-17   9.76538 150  2 pearson
```

```
model1 <- lm(iris$Sepal.Length~iris$Petal.Length+iris$Petal.Width)
model2 <- lm(iris$Sepal.Width~iris$Petal.Length+iris$Petal.Width)

cor(model1$residuals, model2$residuals)
```

```
## [1] 0.6285707
```

Advanced Correlation Techniques (Continued)

We use R code to calculate the autocorrelation in a vector by using the library tseries. We will use a function act() and this function has three parameters:

- data, an input vector
- number of lags (we take a look at some past event from some point in time t)
- plot the auto correlation

```
library(tseries)
mydata <- c(34, 56, 23, 45, 21, 64, 78, 90)
print(acf(mydata, pl=FALSE))
print(acf(mydata, lag=0, pl=FALSE))
print(acf(mydata, lag=1, pl=FALSE))
print(acf(mydata, lag=2, pl=FALSE))
print(acf(mydata, lag=6, pl=FALSE))
```

```
##
## Autocorrelations of series 'mydata', by lag
##
##      0      1      2      3      4      5      6
##  1.000  0.257  0.208 -0.389 -0.093 -0.268 -0.064 -0
```

```
##
## Autocorrelations of series 'mydata', by lag
##
## 0
## 1
```

```
##
## Autocorrelations of series 'mydata', by lag
##
##     0     1
## 1.000 0.257
```

```
##
## Autocorrelations of series 'mydata', by lag
##
##     0     1     2
## 1.000 0.257 0.208
```

```
##
## Autocorrelations of series 'mydata', by lag
##
##      0      1      2      3      4      5      6
##  1.000  0.257  0.208 -0.389 -0.093 -0.268 -0.064
```

References

The lectures of this course are based on the ideas from the following references.

- Exploratory Data Analysis by John W. Tukey
- A Course in Exploratory Data Analysis by Jim Albert
- The Visual Display of Quantitative Information by Edward R. Tufte
- Data Science for Business: what you need to know about data mining and data-analytic thinking by Foster Provost and Tom Fawcett
- Storytelling with Data: A Data Visualization Guide for Business Professionals by Cole Nussbaumer Knaflic