Week 10 Correlation Analysis

Group 22

May 10, 2021

Outline

1. Intro
2. The Three
   - Pearson
   - Spearman
   - Kendall
3. Range & Testing
   - Pearson & Spearman
   - Kendall
4. Comparison
5. Plan

Intro

- many methods can use in this area, but three domain
- Pearson correlation coefficient: linear
- Spearman's rank correlation coefficient: monotonic
- Kendall rank correlation coefficient: monotonic

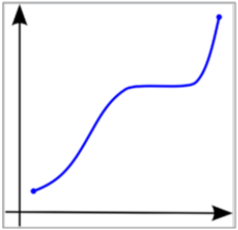

Figure 1 - A monotonically increasing function

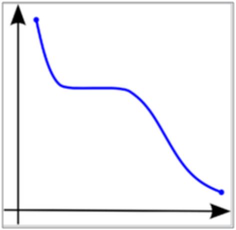

Figure 2 - A monotonically decreasing function

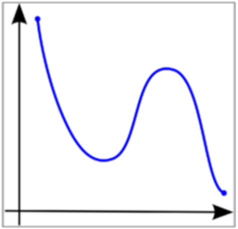

Figure 3 - A function that is not monotonic [1]

[1] wiki:https://en.wikipedia.org/wiki/Monotonic_function

Outline

1. Intro
2. The Three
   - Pearson
   - Spearman
   - Kendall
3. Range & Testing
   - Pearson & Spearman
   - Kendall
4. Comparison
5. Plan

Pearson

- Definition: $\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$
- Assumptions:
  1. variables should be continuous;
  2. without outliers or without significant outliers;
  3. variable should be normally distributed;
  4. linearity (semi) and homoscedasticity

Outline

1. Intro
2. The Three
   - Pearson
   - Spearman
   - Kendall
3. Range & Testing
   - Pearson & Spearman
   - Kendall
4. Comparison
5. Plan

Spearman

- Definition: $r_s = \rho_{\mathrm{rg}_X,\mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X,\mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}}$
- Assumptions:
  1. variables should be ordinal or continuous;
  2. monotonic relationship (semi)
- without normality assumption, nonparametric statistic

Outline

1. Intro
2. The Three
   - Pearson
   - Spearman
   - Kendall
3. Range & Testing
   - Pearson & Spearman
   - Kendall
4. Comparison
5. Plan

Kendall

- Definition:

$$\tau = \frac{(\text{ number of concordant pairs }) - (\text{ number of discordant pairs })}{\binom{n}{2}}$$

concordant pairs $= (x_i > x_j \text{ and } y_i > y_j)$ or $(x_i < x_j \text{ and } y_i < y_j)$

- Assumptions: same as Spearman:
  1. ordinal or continuous;
  2. monotonic relationship (semi)
  3. without normality assumption

Outline

1. Intro
2. The Three
   - Pearson
   - Spearman
   - Kendall
3. Range & Testing
   - Pearson & Spearman
   - Kendall
4. Comparison
5. Plan

Pearson & Spearman

- Range: $[-1, 1]$
- Student's t-distribution: how significant it is?

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- Fisher transformation: the confidence interval

$$F(r) \equiv \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

Outline

1. Intro
2. The Three
   - Pearson
   - Spearman
   - Kendall
3. Range & Testing
   - Pearson & Spearman
   - Kendall
4. Comparison
5. Plan

Pearson & Spearman

- Range: $[-1, 1]$
- z-score: $z = \frac{S+\delta}{\sigma_S} \sim N(0,1)$, where

$$
\begin{aligned}
S &= n_C - n_D \\
\sigma_S^2 &= \frac{\left(N^2-N\right)(2N+5)-T_X''-T_Y''}{18} + \frac{T_X' T_Y'}{9(N^2-N)(N-2)} + \frac{T_X T_Y}{2(N^2-N)} \\
T_X' &= \sum_{i=1}^{S_X} \left(t_{(X)i}^2 - t_{(X)i}\right)\left(t_{(X)i} - 2\right) \\
T_X'' &= \sum_{i=1}^{S_X} \left(t_{(X)i}^2 - t_{(X)i}\right)\left(2t_{(X)i} + 5\right) \\
T_Y' &= \sum_{i=1}^{S_X} \left(t_{(Y)i}^2 - t_{(Y)i}\right)\left(t_{(Y)i} - 2\right) \\
T_Y'' &= \sum_{i=1}^{S_X} \left(t_{(Y)i}^2 - t_{(Y)i}\right)\left(2t_{(Y)i} + 5\right) \\
\delta &= \begin{cases} -1 \text{ if } S > 0 \\ 1 \text{ if } S < 0 \end{cases}
\end{aligned}
$$

Comparison

- Non-parametric correlations contain less information than parametric correlations, such as the mean and deviation of the data, thus they are less powerful but more general (ordinal or continuous).
- Pearson: linear; Spearman and Kendall: monotonic.
- In practice, Kendall correlation is more robust and efficient than Spearman correlation, that is Kendall prefers small data or outliers situations.

Plan

- test normality, whether Pearson or others;
- test linear or monotonic, whether Pearson or others;
- if cannot find a relationship, do data preprocessing.