Day 7: Spearman's Rank Correlation Coefficient

Spearman's Rank Correlation Coefficient

We have two random variables, X and Y:

- $X = \{x_i, x_2, x_3, \ldots, x_n\}$
- $Y = \{y_i, y_2, y_3, \ldots, y_n\}$

If $\mathbf{Rank}_X$ and $\mathbf{Rank}_Y$ denote the respective ranks of each data point, then the Spearman's rank correlation coefficient, r_s, is the Pearson correlation coefficient of $\mathbf{Rank}_X$ and $\mathbf{Rank}_Y$.

Example

- $X = \{0.2, 1.3, 0.2, 1.1, 1.4, 1.5\}$
- $Y = \{1.9, 2.2, 3.1, 1.2, 2.2, 2.2\}$

$\mathbf{Rank}_X$:

$$\begin{bmatrix} X: & 0.2 & 1.3 & 0.2 & 1.1 & 1.4 & 1.5 \\ Rank: & 1 & 3 & 1 & 2 & 4 & 5 \end{bmatrix}$$

So, $\mathbf{Rank}_X = \{1, 3, 1, 2, 4, 5\}$

Similarly, $\mathbf{Rank}_Y = \{2, 3, 4, 1, 3, 3\}$

r_s equals the Pearson correlation coefficient of $\mathbf{Rank}_X$ and $\mathbf{Rank}_Y$, meaning that $r_s = 0.158114$.

Special Case: X and Y Don't Contain Duplicates

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

Here, d_i is the difference between the respective values of $\mathbf{Rank}_X$ and $\mathbf{Rank}_Y$.

Proof

Let's define P be the rank of X and Q be the rank of Y. Both P and Q are permutations of set $\{1, 2, 3, \ldots, n\}$, because data sets X and Y contain no duplicates in this special case.

Mean of P and Q:

$$\sum_i p_i = \sum_i q_i = \frac{n \cdot (n+1)}{2}$$

$$\Rightarrow \mu_P = \mu_Q = \mu = \frac{(n+1)}{2}$$

Standard Deviation of P and Q:

$$n \cdot (n^2 - 1)$$

$$\sum_i (p_i - \mu_p)^2 = \sum_i (p_i - \mu)^2 = \sum_i p_i^2 - 2\mu \sum_i p_i + \mu^2 \sum_i 1 = \frac{\quad}{12}$$

So,

$$\sigma_P = \sigma_Q = \sigma = \sqrt{\frac{\sum_i (p_i - \mu_p)^2}{n}} = \sqrt{\frac{n^2 - 1}{12}}$$

Calculating $\sum_i d_i^2$:

$$\sum_i d_i^2 = \sum_i (p_i - q_i)^2 = \sum_i p_i^2 - 2\sum_i (p_i q_i) + \sum_i q_i^2$$

We know that:

$$\sum_i p_i^2 = \sum_i q_i^2 = \frac{n \cdot (n+1) \cdot (2n+1)}{6}$$

So,

$$\sum_i (p_i q_i) = \frac{n \cdot (n+1)(n^2+1)}{6} - \frac{1}{2} \sum_i d_i^2$$

Covariance of P and Q:

$$\mathbf{cov}(P, Q) = \frac{\sum_i (p_i - \mu_p)(q_i - \mu_q)}{n} = \frac{\sum_i (p_i - \mu)(q_i - \mu)}{n}$$

$$\Rightarrow \mathbf{cov}(P, Q) = \frac{\sum_i (p_i q_i) - \mu \left(\sum_i p_i + \sum_i q_i\right) + \mu^2 \sum_i 1}{n}$$

$$\frac{n\cdot(n+1)\cdot(n^2+1)}{\ldots} - \frac{1}{\ldots} \sum d^2 - \mu\left(\sum p \ldots + \sum q \ldots\right) + \mu^2 \sum 1 \ldots$$

$$\Rightarrow \texttt{cov}(P, Q) = \frac{\frac{\cdots}{6} - \frac{1}{2}\sum_i d_i^2 - \mu\left(\sum_i p_i + \sum_i q_i\right) + \mu \sum_i 1}{n}$$

$$\Rightarrow \texttt{cov}(P, Q) = \frac{\frac{n\cdot(n^2-1)}{12} - \frac{1}{2}\sum_i d_i^2}{n}$$

Spearman's Rank Correlation Coefficient:

We know that the Spearman's rank correlation coefficient (r_s) of X and Y is equal to the Pearson correlation coefficient of P and Q. So,

$$r_s = \frac{\texttt{cov}(P, Q)}{\sigma_P \sigma_Q} = \frac{\texttt{cov}(P, Q)}{\sigma^2}$$

$$\Rightarrow r_s = \frac{\frac{\frac{n\cdot(n^2-1)}{12} - \frac{1}{2}\sum_i d_i^2}{n}}{\frac{(n^2-1)}{12}}$$

$$\Rightarrow r_s = \frac{\frac{n\cdot(n^2-1)}{12} - \frac{1}{2}\sum_i d_i^2}{\frac{n\cdot(n^2-1)}{12}}$$

$$\Rightarrow r_s = 1 - \frac{6\sum d_i^2}{n\cdot(n^2-1)}$$