Homework 9

蒋翌坤 20307100013

1. It is because Spearman correlation is based on the rank of the data. Outliers are transformed such that they are not far away from the other data points. However, Pearson correlation is based on the actual value of the data, so outliers will have a large impact on the correlation coefficient.

For two data sets, Spearman correlation coefficient is $r_s = 1$, Pearson correlation coefficient is $r_p = 0.96$. It is obvious Spearman correlation coefficient reflects the association better since the data is all square related and the correlation should be 1.

2. Since there are no ties, a_i, b_i are both some arrangements of $1, \dots, n$, $\bar{a} = \bar{b} = \frac{n+1}{2}$

$$\sum_{i=1}^{n}(a_i - \bar{a})^2 = \sum_{i=1}^{n}(b_i - \bar{b})^2 = \sum_{i=1}^{n}(i - \frac{n+1}{2})^2 = \frac{n(n^2-1)}{12}$$

$$\begin{aligned}
r_s &= \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2 \sum_{i=1}^{n}(b_i - \bar{b})^2}} = \frac{12}{n(n^2-1)} \sum_{i=1}^{n}(a_i - \frac{n+1}{2})(b_i - \frac{n+1}{2}) \\
&= \frac{12}{n(n^2-1)} \sum_{i=1}^{n}\Big(a_i b_i - \frac{n+1}{2}(a_i + b_i) + \frac{(n+1)^2}{4}\Big) \\
&= \frac{12}{n(n^2-1)} \Big(\sum_{i=1}^{n} a_i b_i - \frac{n+1}{2}\frac{2n(n+1)}{2} + \frac{n(n+1)^2}{4}\Big) \\
&= \frac{12}{n(n^2-1)} \Big(\sum_{i=1}^{n} a_i b_i - \frac{n(n+1)^2}{4}\Big) \\
&= -\frac{6}{n(n^2-1)} \Big(\frac{n(n+1)^2}{2} - \frac{2n(n+1)(2n+1)}{6} + \sum_{i=1}^{n}(a_i^2 - 2a_i b_i + b_i^2)\Big) \\
&= 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n}(a_i - b_i)^2
\end{aligned}$$

3. There are $n!$ distinguishable sets of pair (a_i, b_i). Under null hypothesis that $r_s = 0$, it means x_i and $|e_i|$ are independent. So the $n!$ distinguishable sets of pair (a_i, b_i) are equally likely to appear. The observed r_s is therefore distribution-free regardless of how x_i and $|e_i|$ are distributed. $f_{r_s}(r) = \frac{u_r}{n!}$ for some u_r representing the number of pairings which lead to a value r for the statistic. The bigger the absolute value of r is, the smaller u_r is.

The p-value formula $K^{-1}\sum_{k=1}^{K}\mathbb{I}(|r_s| \leq |r_s^{(k)}|)$ is like the probability of $|r_s|$ being smaller than the Spearman coefficient of other permutations. Under null hypothesis, the p-value formula converges to $1 - \int_{-r_s}^{r_s} f_{r_s}(r)dr$, it shows that when observed $|r_s|$ is small, it is more likely to accept the null hypothesis.