统计学三大相关系数

统计学三大相关性系数反应的是两个变量之间变化趋势的方向和程度，取值范围为$[-1,+1]$，0表示两个变量不相关，正值表示正相关，负值表示负相关，值越大表示相关性越强。

Kendall Correlation Coefficient(肯德尔相关性系数)

肯德尔相关性系数也成为肯德尔秩相关系数，其计算的是分类变量，也就是离散变量。对于肯德尔相关性系数中的两个序列，将其对应位置的元素组成Pair对，因此组成Pair对集合。为了便于描述以X和Y表示两个序列，则将(X_i, Y_i)组成一个Pair对，对于两个Pair对，如果满足$X_i > X_j AND Y_i > Y_j$或者$X_i < X_j AND Y_i < Y_j$，则认为这两个Pair对是一致的；如果满足$X_i > X_j AND Y_i < Y_j$或者$X_i < X_j AND Y_i > Y_j$，则认为这两个Pair对是不一致的；如果满足$X_i = X_j$或者$Y_i = Y_j$则认为这两个Pair对既不是一致的也不是不一致的。

肯德尔相关系数有三个公式

公式1:

$$Tau - a = \frac{C - D}{\frac{1}{2}N(N-1)}$$

C为一致性的元素对数，D表示不一致的元素对数，N表示序列中的元素数。因为X和Y共组成N个Pair，因此这N个Pair可以构建出$\frac{1}{2}N(N-1)$个Pair对。该公式适用于两个序列均不存在相同元素的情况。

公式2:

$$Tau - b = \frac{C - D}{\sqrt{(N3 - N1)(N3 - N2)}}$$

C为一致性的元素对数，D为不一致的元素对数，其中$N3 = \frac{1}{2}N(N-1)$即总的Pair对数，$N1 = \sum_{i=1}^{s} \frac{1}{2} U_i(U_i - 1)$ 其中s为序列X中有重复值的元素的个数，比如元素X为(1, 2, 3, 3, 4, 2)则s的值为2，因为其中2和3是序列中重复的值，U_i是以重复的元素组成的小集合可以组成的Pair对的数目，比如以上面的例子，有两个2，则以2作为一个小集

合(2,2)可以组成的Pair对的个数为1个。$N2 = \sum_{i=1}^{t} \frac{1}{2} V_i(Vi - 1)$，理解方式与上线N2的理解是一致的。

公式3:

$$Tau - c = \frac{C - D}{\frac{1}{2} N^2 \frac{M-1}{M}}$$

其中C、D和N的解释与上面的公式1、2一致。

Pearson Correlation Coefficient(皮尔逊相关性系数)

皮尔逊相关性系数的计算公式如下所示:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma X \sigma Y}$$

即皮尔逊相关系数等于两个变量之间的协方差除以两个变量的标准差的乘积。皮尔逊相关系数就对应着概率论书籍上提到的相关系数。协方差的计算公式为:

$$cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

其中E表示期望，即协方差表示的X的偏差与Y的偏差乘积的期望。因为相关系数是有量纲的，因此直接使用有量纲的指标不容易直接进行评估，因此为了消除量纲，对协方差除以一个相同的量纲从而得到相关系数，也即皮尔逊相关系数。

Spearman Correlation Coefficient(斯皮尔曼相关性系数)

斯皮尔曼相关性系数，也叫斯皮尔曼秩相关系数，其通常被认为是排列后的变量之间的皮尔逊线性相关系数，其计算公式为:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

其中d为秩次差，n为元素的个数。
计算斯皮尔曼相关系数的方法:
1) 首先对两个序列进行排序，假设两个序列为X，Y，对两者进行排序后，分别顺序记录两个序列中各个元素在排序后所在的位置值，该值即为秩，得到X'和Y'

2) 计算出$X^{'}$和$Y^{'}$对应位置的秩的差值，此值即对应着上面公式中的d_i

3) 带入公式即可计算出相关系数