WIKIPEDIA
The Free Encyclopedia

Kendall rank correlation coefficient

In statistics, the **Kendall rank correlation coefficient**, commonly referred to as **Kendall's τ coefficient** (after the Greek letter τ, tau), is a statistic used to measure the ordinal association between two measured quantities. A **τ test** is a non-parametric hypothesis test for statistical dependence based on the τ coefficient.

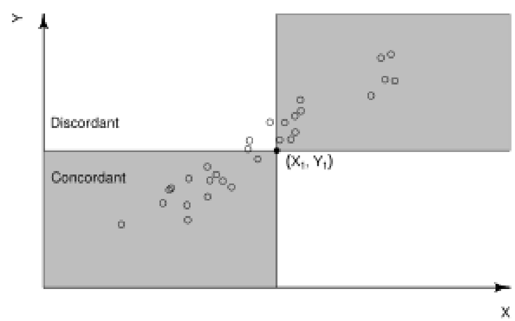
It is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities. It is named after Maurice Kendall, who developed it in 1938,[1] though Gustav Fechner had proposed a similar measure in the context of time series in 1897.[2]

Intuitively, the Kendall correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully different for a correlation of −1) rank between the two variables.

Both Kendall's τ and Spearman's ρ can be formulated as special cases of a more general correlation coefficient.

Definition

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of observations of the joint random variables *X* and *Y*, such that all the values of (x_i) and (y_i) are unique (ties are neglected for simplicity). Any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be *concordant* if the sort order of (x_i, x_j) and (y_i, y_j) agrees: that is, if either both $x_i > x_j$ and $y_i > y_j$ holds or both $x_i < x_j$ and $y_i < y_j$; otherwise they are said to be *discordant*.



All points in the gray area are concordant and all points in the white area are discordant with respect to point (X_1, Y_1). With $n = 30$ points, there are a total of $\binom{30}{2} = 435$ possible point pairs. In this example there are 395 concordant point pairs and 40 discordant point pairs, leading to a Kendall rank correlation coefficient of 0.816.

The Kendall τ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})} = 1 - \frac{2(\text{number of discordant pairs})}{\binom{n}{2}}.$$

[3]

Where $\binom{n}{2} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of ways to choose two items from n items.

Properties

The denominator is the total number of pair combinations, so the coefficient must be in the range $-1 \leq \tau \leq 1$.

- If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1.
- If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other) the coefficient has value −1.
- If *X* and *Y* are independent and not constant, then the expectation of the coefficient is zero.
- An explicit expression for Kendall's rank coefficient is $\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$.

Hypothesis test

The Kendall rank coefficient is often used as a test statistic in a statistical hypothesis test to establish whether two variables may be regarded as statistically dependent. This test is non-parametric, as it does not rely on any assumptions on the distributions of *X* or *Y* or the distribution of (*X*,*Y*).

Under the null hypothesis of independence of *X* and *Y*, the sampling distribution of *τ* has an expected value of zero. The precise distribution cannot be characterized in terms of common distributions, but may be calculated exactly for small samples; for larger samples, it is common to use an approximation to the normal distribution, with mean zero and variance

$$\frac{2(2n+5)}{9n(n-1)}.$$ [4]

Accounting for ties

A pair $\{(x_i, x_j), (y_i, y_j)\}$ is said to be *tied* if $x_i = x_j$ or $y_i = y_j$; a tied pair is neither concordant nor discordant. When tied pairs arise in the data, the coefficient may be modified in a number of ways to keep it in the range [−1, 1]:

Tau-a

The Tau-a statistic tests the strength of association of the cross tabulations. Both variables have to be ordinal. Tau-a will not make any adjustment for ties. It is defined as:

$$\tau_A = \frac{n_c - n_d}{n_0}$$

where n_c, n_d and n_0 are defined as in the next section.

Tau-b

The Tau-b statistic, unlike Tau-a, makes adjustments for ties.[5] Values of Tau-b range from −1 (100% negative association, or perfect inversion) to +1 (100% positive association, or perfect agreement). A value of zero indicates the absence of association.

The Kendall Tau-b coefficient is defined as:

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

where

$$\begin{aligned} n_0 &= n(n-1)/2 \\ n_1 &= \sum_i t_i(t_i-1)/2 \\ n_2 &= \sum_j u_j(u_j-1)/2 \\ n_c &= \text{Number of concordant pairs} \\ n_d &= \text{Number of discordant pairs} \\ t_i &= \text{Number of tied values in the } i^{\text{th}} \text{ group of ties for the first quantity} \\ u_j &= \text{Number of tied values in the } j^{\text{th}} \text{ group of ties for the second quantity} \end{aligned}$$

A simple algorithm developed in BASIC computes Tau-b coefficient using an alternative formula. [6]

Be aware that some statistical packages, e.g. SPSS, use alternative formulas for computational efficiency, with double the 'usual' number of concordant and discordant pairs.[7]

Tau-c

Tau-c (also called Stuart-Kendall Tau-c)[8] is more suitable than Tau-b for the analysis of data based on non-square (i.e. rectangular) contingency tables.[8][9] So use Tau-b if the underlying scale of both variables has the same number of possible values (before ranking) and Tau-c if they differ. For instance, one variable might be scored on a 5-point scale (very good, good, average, bad, very bad), whereas the other might be based on a finer 10-point scale.

The Kendall Tau-c coefficient is defined as:[9]

$$\tau_C = \frac{2(n_c - n_d)}{n^2 \frac{(m-1)}{m}}$$

where

$$\begin{aligned} n_c &= \text{Number of concordant pairs} \\ n_d &= \text{Number of discordant pairs} \\ r &= \text{Number of rows} \\ c &= \text{Number of columns} \\ m &= \min(r, c) \end{aligned}$$

Significance tests

When two quantities are statistically independent, the distribution of τ is not easily characterizable in terms of known distributions. However, for τ_A the following statistic, z_A, is approximately distributed as a standard normal when the variables are statistically independent:

$$z_A = \frac{3(n_c - n_d)}{\sqrt{n(n-1)(2n+5)/2}}$$

Thus, to test whether two variables are statistically dependent, one computes z_A, and finds the cumulative probability for a standard normal distribution at $-|z_A|$. For a 2-tailed test, multiply that number by two to obtain the p-value. If the p-value is below a given significance level, one rejects the null hypothesis (at that significance level) that the quantities are statistically independent.

Numerous adjustments should be added to z_A when accounting for ties. The following statistic, z_B, has the same distribution as the τ_B distribution, and is again approximately equal to a standard normal distribution when the quantities are statistically independent:

$$z_B = \frac{n_c - n_d}{\sqrt{v}}$$

where

$$\begin{aligned}
v &= (v_0 - v_t - v_u)/18 + v_1 + v_2 \\
v_0 &= n(n-1)(2n+5) \\
v_t &= \textstyle\sum_i t_i(t_i-1)(2t_i+5) \\
v_u &= \textstyle\sum_j u_j(u_j-1)(2u_j+5) \\
v_1 &= \textstyle\sum_i t_i(t_i-1)\sum_j u_j(u_j-1)/(2n(n-1)) \\
v_2 &= \textstyle\sum_i t_i(t_i-1)(t_i-2)\sum_j u_j(u_j-1)(u_j-2)/(9n(n-1)(n-2))
\end{aligned}$$

This is sometimes referred to as the Mann-Kendall test.[10]

Algorithms

The direct computation of the numerator $n_c - n_d$, involves two nested iterations, as characterized by the following pseudocode:

```
numer := 0
for i := 2..N do
    for j := 1..(i − 1) do
        numer := numer + sign(x[i] − x[j]) × sign(y[i] − y[j])
return numer
```

Although quick to implement, this algorithm is $O(n^2)$ in complexity and becomes very slow on large samples. A more sophisticated algorithm[11] built upon the Merge Sort algorithm can be used to compute the numerator in $O(n \cdot \log n)$ time.

Begin by ordering your data points sorting by the first quantity, x, and secondarily (among ties in x) by the second quantity, y. With this initial ordering, y is not sorted, and the core of the algorithm consists of computing how many steps a Bubble Sort would take to sort this initial y. An enhanced Merge Sort algorithm, with $O(n \log n)$ complexity, can be applied to compute the number of swaps, $S(y)$, that would be required by a Bubble Sort to sort y_i. Then the numerator for τ is computed as:

$$n_c - n_d = n_0 - n_1 - n_2 + n_3 - 2S(y),$$

where n_3 is computed like n_1 and n_2, but with respect to the joint ties in x and y.

A Merge Sort partitions the data to be sorted, y into two roughly equal halves, y_{left} and y_{right}, then sorts each half recursive, and then merges the two sorted halves into a fully sorted vector. The number of Bubble Sort swaps is equal to:

$$S(y) = S(y_{\mathrm{left}}) + S(y_{\mathrm{right}}) + M(Y_{\mathrm{left}}, Y_{\mathrm{right}})$$

where Y_{left} and Y_{right} are the sorted versions of y_{left} and y_{right}, and $M(\cdot,\cdot)$ characterizes the Bubble Sort swap-equivalent for a merge operation. $M(\cdot,\cdot)$ is computed as depicted in the following pseudo-code:

```
function M(L[1..n], R[1..m]) is
    i := 1
    j := 1
    nSwaps := 0
    while i ≤ n and j ≤ m do
        if R[j] < L[i] then
            nSwaps := nSwaps + n − i + 1
            j := j + 1
        else
```

```
        i := i + 1
    return nSwaps
```

A side effect of the above steps is that you end up with both a sorted version of x and a sorted version of y. With these, the factors t_i and u_j used to compute τ_B are easily obtained in a single linear-time pass through the sorted arrays.

Software Implementations

- R's statistics base-package implements the test `cor.test(x, y, method = "kendall")` (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/cor.test.html) in its "stats" package (also `cor(x, y, method = "kendall")` will work, but the latter does not return the p-value).
- For Python, the SciPy library implements the computation of τ in `scipy.stats.kendalltau` (https://web.archive.org/web/20181008171919/https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html)

See also

Mathematics portal

- Correlation
- Kendall tau distance
- Kendall's W
- Spearman's rank correlation coefficient
- Goodman and Kruskal's gamma
- Theil–Sen estimator
- Mann–Whitney U test - it is equivalent to Kendall's tau correlation coefficient if one of the variables is binary.

References

1. Kendall, M. (1938). "A New Measure of Rank Correlation". *Biometrika*. **30** (1–2): 81–89. doi:10.1093/biomet/30.1-2.81 (https://doi.org/10.1093%2Fbiomet%2F30.1-2.81). JSTOR 2332226 (https://www.jstor.org/stable/2332226).
2. Kruskal, W. H. (1958). "Ordinal Measures of Association". *Journal of the American Statistical Association*. **53** (284): 814–861. doi:10.2307/2281954 (https://doi.org/10.2307%2F2281954). JSTOR 2281954 (https://www.jstor.org/stable/2281954). MR 0100941 (https://mathscinet.ams.org/mathscinet-getitem?mr=0100941).
3. Nelsen, R.B. (2001) [1994], "Kendall tau metric" (https://www.encyclopediaofmath.org/index.php?title=Kendall_tau_metric), *Encyclopedia of Mathematics*, EMS Press
4. Prokhorov, A.V. (2001) [1994], "Kendall coefficient of rank correlation" (https://www.encyclopediaofmath.org/index.php?title=Kendall_coefficient_of_rank_correlation), *Encyclopedia of Mathematics*, EMS Press
5. Agresti, A. (2010). *Analysis of Ordinal Categorical Data* (Second ed.). New York: John Wiley & Sons. ISBN 978-0-470-08289-8.
6. Alfred Brophy (1986). "An algorithm and program for calculation of Kendall's rank correlation coefficient" (https://link.springer.com/content/pdf/10.3758/BF03200993.pdf) (PDF). *Behavior Research Methods, Instruments, & Computers*. **18**: 45–46. doi:10.3758/BF03200993 (https://doi.org/10.3758%2FBF03200993). S2CID 62601552 (https://api.semanticscholar.org/CorpusID:62601552).
7. IBM (2016). *IBM SPSS Statistics 24 Algorithms* (http://www-01.ibm.com/support/docview.wss?uid=swg27047033#en). IBM. p. 168. Retrieved 31 August 2017.
8. Berry, K. J.; Johnston, J. E.; Zahran, S.; Mielke, P. W. (2009). "Stuart's tau measure of effect size for ordinal variables: Some methodological considerations" (https://doi.org/10.3758%2Fbrm.41.4.1144). *Behavior Research Methods*. **41** (4): 1144–1148. doi:10.3758/brm.41.4.1144 (https://doi.org/10.3758%2Fbrm.41.4.1144). PMID 19897822 (https://pubmed.ncbi.nlm.nih.gov/19897822).
9. Stuart, A. (1953). "The Estimation and Comparison of Strengths of Association in Contingency Tables". *Biometrika*. **40** (1–2): 105–110. doi:10.2307/2333101 (https://doi.org/10.2307%2F2333101). JSTOR 2333101 (https://www.jstor.org/stable/2333101).
10. Glen_b. "Relationship between Mann-Kendall and Kendall Tau-b" (https://stats.stackexchange.com/q/414038).

11. Knight, W. (1966). "A Computer Method for Calculating Kendall's Tau with Ungrouped Data". *Journal of the American Statistical Association*. **61** (314): 436–439. doi:10.2307/2282833 (https://doi.org/10.2307%2F2282833). JSTOR 2282833 (https://www.jstor.org/stable/2282833).

Further reading

- Abdi, H. (2007). "Kendall rank correlation" (http://www.utdallas.edu/~herve/Abdi-KendallCorrelation2007-pretty.pdf) (PDF). In Salkind, N.J. (ed.). *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage.
- Daniel, Wayne W. (1990). "Kendall's tau" (https://books.google.com/books?id=0hPvAAAAMAAJ&pg=PA365). *Applied Nonparametric Statistics* (2nd ed.). Boston: PWS-Kent. pp. 365–377. ISBN 978-0-534-91976-4.
- Kendall, Maurice; Gibbons, Jean Dickinson (1990) [First published 1948]. *Rank Correlation Methods* (https://archive.org/details/rankcorrelationm0000kend). Charles Griffin Book Series (5th ed.). Oxford: Oxford University Press. ISBN 978-0195208375.
- Bonett, Douglas G.; Wright, Thomas A. (2000). "Sample size requirements for estimating Pearson, Kendall, and Spearman correlations". *Psychometrika*. **65** (1): 23–28. doi:10.1007/BF02294183 (https://doi.org/10.1007%2FBF02294183). S2CID 120558581 (https://api.semanticscholar.org/CorpusID:120558581).

External links

- Tied rank calculation (http://www.statsdirect.com/help/nonparametric_methods/kend.htm)
- Software for computing Kendall's tau on very large datasets (http://law.di.unimi.it/software/law-docs/it/unimi/dsi/law/stat/KendallTau.html)
- Online software: computes Kendall's tau rank correlation (http://www.wessa.net/rwasp_kendall.wasp)