UPPSALA
UNIVERSITET

Working Paper 2013:10

Department of Statistics

# Confidence Intervals for Ranks

Sture Holm

Working Paper 2013:10
June 2013
Department of Statistics
Uppsala University
Box 513
SE-751 20 UPPSALA
SWEDEN

Working papers can be downloaded from www.statistics.uu.se

Title: Confidence Intervals for Ranks

Author: Sture Holm
E-mail: holm@chalmers.se

# Confidence intervals for ranks

**Sture Holm**[1]

Department of Mathematical statistics, Chalmers and University of Gothenburg, SE-412 96 Göteborg, Sweden

It has become common to try to rank medical units in different ways. Hospitals may for instance be ranked with respect to success rates for some kind of operation and local medical care centers may be ranked with respect to service at the center by letting patients judge the service of the unit by a graded verbal scale with ordered categories. These types of investigations often present the obtained ranking without any quality measure. In the present paper I suggest a simple general procedure which gives confidence interval estimates for the rank of units. An essential part of the construction is to use multiple test principles for finding the bounds of the confidence interval. It is shown how the general method can be used with binomial data and with ordered categorical data.

*Key words:* Confidence interval, multinomial distribution, rank, ordered categorical data.

## 1 A general principle for constructing confidence intervals for ranks

A confidence interval is an interval with some kind of random bounds, which has a big probability, the confidence degree, of hitting some kind of parameter. The parameter may be a basic parameter in the probabilistic model for the data, or a secondary parameter, i. e. a function of the basic parameters. When we want to make a confidence interval for the theoretical rank of a unit in some respect, it is typically a secondary parameter.

Think for instance of the case where we want to rank m hospitals with respect to success rate for some kind of operation. Then the theoretical model may be that the number of successes for unit number i is binomial with parameters $(n_i, p_i)$, where $n_i$ is known and $p_i$ is unknown. Results for different units are supposed to be independent. The theoretical unknown ranks of the unit i are given by the numbers greater than $\#\left(j, 1 \le j \le m : p_j < p_i\right)$ and smaller than $m - \#\left(j, 1 \le j \le m : p_j > p_i\right)$. Normally there is only one such number, but occasionally some $p_j$ coincide with $p_i$ and then the rank is defined as a whole set of consecutive values.

Many ranking problems with other types of basic statistical data have the same or similar type of structure.

**Definition 1**

A ranking problem is of type 1 if the theoretical ranking of a unit i is determined by the rank of an individual position parameter $\mu_i$ for unit i in the set of the position parameters $\mu_j$ for all units.

[1] Author e-mail: holm@chalmers.se

Since ranking is in fact a property of differences of units only, the positions themselves are not important. In some problems it is most relevant to use difference parameters instead of position parameters. Let $\lambda_{i,j}$ denote a parameter describing the difference between unit i and unit j in such a way that positive values of $\lambda_{i,j}$ indicate that unit i is superior to unit j and that negative values indicate the opposite. Then the theoretical rank of unit i is given by the numbers greater than $\#\left( j,1 \le j \le m : \lambda_{i,j} > 0 \right)$ and smaller than $m - \#\left( j,1 \le j \le m : \lambda_{i,j} < 0 \right)$

**Definition 2**

A ranking problem is of type 2 if the theoretical ranking of a unit i is determined by the signs of pair-wise difference parameters $\lambda_{i,j}$ between unit i and all other units j.

The problem of possible non-transitivity, i. e. a case of the type $\lambda_{i,j} > 0$, $\lambda_{j,l} > 0$ and $\lambda_{l,i} > 0$, would be a problem in an attempt to make a ranking for all units simultaneously. Our definition of the rank(s) of a fixed unit i will however give a non-ambiguous result, since we compare all other units j with that fixed unit i only. It is not influenced by any ordering between these other units. Since the method presented in the following will also work without problem for such possible situation, we postpone further discussion on non-transitivity to a special section at the end of the paper.

Now we will describe the general method to create confidence intervals for the theoretical rank of units, which is the essence of this paper.

**Suggested method for construction of confidence intervals for rank**

Suppose that we want to make a symmetric confidence interval with confidence degree $1-\alpha$ for the rank of unit i. Suppose further that for each $j \neq i$ there is a test statistic $T_{i,j}^{(-)}$ for the hypothesis $\mu_i - \mu_j \le 0$ against $\mu_i - \mu_j > 0$ in a ranking problem of type 1 or for the hypothesis $\lambda_{i,j} \le 0$ against $\lambda_{i,j} > 0$ in a ranking problem of type 2. Rejection is supposed to be for big positive values of $T_{i,j}^{(-)}$. Next make a multiple test of the m-1 hypotheses $\mu_i - \mu_j \le 0 : \ 1 \le j \le m,\ j \neq i$ for a ranking problem of type 1 or of the m-1 hypotheses $\lambda_{i,j} \le 0 : \ 1 \le j \le m,\ j \neq i$ for a ranking problem of type 2, at the multiple level of significance $\alpha / 2$. Let $N^{(-)}$ be the number of rejected hypotheses.

In the same way suppose that for each $j \neq i$ there is a test statistic $T_{i,j}^{(+)}$ for the hypothesis $\mu_i - \mu_j \ge 0$ against $\mu_i - \mu_j < 0$ in a ranking problem of type 1 or for the hypothesis $\lambda_{i,j} \ge 0$ against $\lambda_{i,j} < 0$ in a ranking problem of type 2. Rejection is supposed to be for big positive values of $T_{i,j}^{(+)}$. Make a multiple test of the m-1 hypotheses $\mu_i - \mu_j \ge 0 : \ 1 \le j \le m,\ j \neq i$ for a ranking problem of type 1 or of the m-1 hypotheses $\lambda_{i,j} \ge 0 : \ 1 \le j \le m,\ j \neq i$ for a ranking problem of type 2, at the multiple level of

significance $\alpha/2$. Let $N^{(+)}$ be the number of rejected hypotheses. Then the suggested confidence interval for a theoretical rank of unit i (a single rank or a set of ranks) is

$$\left[N^{(-)}+1; m-N^{(+)}\right].$$

**Comments**

Observe that the confidence interval is closed so the end points are included. The interval is aimed to be symmetric in the sense of having the same risk $\alpha/2$ to miss on lower and upper side. They may, however, have tails of different lengths on both sides of some type of point estimate of the rank.

In some cases, for instance when ranking process incidence intensity for different units, it is natural to use the ratio of some positive parameter for comparing units. Such cases are easily transformed to the above problem of type 2 by using logarithms of parameters, and they will not be considered separately.

**Theorem 1.**

The above suggested type of confidence interval for a single theoretical rank or a set of theoretical ranks for a unit has confidence level at least $1-\alpha$.

**Proof.** Consider one of the sides of the suggested confidence interval, e.g. the lower one. If the theoretical single rank of unit i, or the smallest theoretical rank in a set of ranks of unit i, equals $r$, there are $r-1$ specific units on the "smaller side". When we make a level $\alpha/2$ multiple test of $\mu_i-\mu_j\le 0:\ 1\le j\le m,\ j\ne i$ (or $\lambda_{i,j}\le 0:\ 1\le j\le m,\ j\ne i$), there are now $m-r$ true hypotheses. But a multiple test with multiple level of significance $\alpha/2$ has this same small probability of rejecting <u>any</u> true hypothesis. So with a probability at most $\alpha/2$, none of the $m-r$ true hypotheses will falsely be declared to be on the "smaller side". When this event has occurred the number $N^{(-)}$ will be at most $r-1$ and the confidence bound on the lower side will be correct.

The upper side is treated in the same way for the greatest rank in the set of ranks of unit i and according to the Boole inequality the total risk is bounded by $\alpha$.

Q. E. D.

The units in an investigation often have different sample sizes. Most standard multiple tests, however, require a more regular structure. Thus one usually has to relay on a more flexible multiple test like the Bonferroni method or the sequentially rejective (step-down) method in Holm (1979).

Using this latter method in the multiple test on multiple level $\alpha/2$ in one of the sides, means to start first to use the Bonferroni method for m-1 one-sided hypothesis and reject those who have an individual p value below $\alpha/\left(2(m-1)\right)$. If some hypotheses are rejected the procedure goes on with the Bonferroni method used on the remaining non-rejected hypotheses, where the original m-1 is replaced by the number of these non-rejected hypotheses. This is repeated until (for the first time !) no further rejection can be done.

When the set of rejected hypotheses is determined we only count the number of rejected hypotheses, which gives $N^{(-)}$ for the lower side test and $N^{(+)}$ for the upper side test.

**Comments**

It is well known that multiple tests using the Boole inequality, like Holm (1979), are powerful for negatively correlated or independent test statistics. For test statistics with high positive correlation they are not so powerful. Since all other units are compared with the investigated unit, and the sample sizes are probably of the same order, we have here a correlation of approximately 50 %. Is there a big loss in power in such a case?

Usually the test statistics are asymptotically normal, and we can get a slight indication of the efficiency by comparing the one-sided rejection limits for an exact multiple t test with infinite degrees of freedom (the normal distribution case) with the rejection limits given by the Boole inequality. It turns out that for instance 1 % multiple level and 8 cases has the exact limit 2.97 and the approximate Boolean limit 3.02. For 20 cases the same limits are 3.21 and 3.29. Thus there is some loss, but it is not very big. And a more exact multiple test with different sample sizes and different correlations would require very much computational work.

## 2 Binomial data

Suppose that the data from m units consist of sample sizes $N_1, N_2, \ldots, N_m$ and observations $M_1, M_2, \ldots, M_m$ of occurrence of a studied event for instance success of a type of operation. As a mathematical model we suppose that $M_i$ has a binomial distribution with parameters $N_i$ and $p_i$, where $p_i$ is unknown, and that all $M_i : s$ are independent.

The theoretical rank of unit i is determined by the signs of the differences $p_i - p_j$ for all $j : 1 \le j \le m, j \ne \grave{\imath}$. A suitable test statistic for testing a hypothesis $p_i - p_j \le 0$ against $p_i - p_j > 0$ is

$$T_{i,j}^{(-)} = \frac{\frac{M_j}{N_j} - \frac{M_i}{N_i}}{\sqrt{\frac{M_j\left(N_j - M_j\right)}{N_j^3} + \frac{M_i\left(N_i - M_i\right)}{N_i^3}}},$$

which has an approximate normal (0;1) distribution at the boundary of the hypothesis. For the upper side the test statistic $T_{i,j}^{(+)}$ is the same one with sign changed. The tests in the two directions are easily performed. Let us take a numerical example.

Suppose that there are 10 units with the following data, which are in fact artificially generated with random sample sizes in the interval [100;500] and random probability parameter in the interval [0.4;0.8].

**Table 1.** Data for a binomial example

| Unit | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|

| Sample size | 157 | 100 | 245 | 199 | 107 | 299 | 479 | 305 | 442 | 207 |
|---|---|---|---|---|---|---|---|---|---|---|
| Occurrences | 78 | 61 | 182 | 146 | 70 | 210 | 327 | 158 | 214 | 150 |
| Estimated p | 0.497 | 0.610 | 0.743 | 0.734 | 0.654 | 0.702 | 0.683 | 0.518 | 0.484 | 0.725 |

In order to get confidence intervals for the rank of all units, there are first performed 180 one-sided pair-wise tests to get one-sided p values. Next there are performed 20 multiple tests. As an example take the multiple test on the upper side of unit A. The nine p values became (rounded to four decimals)

0.0362, 0.0000, 0.0000, 0.0049, 0.0000, 0.0000, 0.3328, 0.6073, 0.0000.

In the first Bonferroni step the comparison bound equals $\alpha/18 = 0.0028$ for confidence degree 95 %. Then 5 hypotheses are immediately rejected (corresponding to comparison with units C, D, F, G, J). In the next step the Bonferroni comparison bound is $\alpha/2\cdot 4 = 0.0063$ since there remain 4 non-rejected hypotheses. Thus there is one further rejection corresponding to unit E, but in the next step there are no further rejections and the procedure stops. We get 6 rejections in all, the outcome of $N^{(+)}$ is 6 and the upper confidence bound for the rank of unit A is 10-6=4.

The results for all units are the following:

**Table 2.** Rank confidence intervals for 10 units with an individual confidence degree 95 %.

| Unit | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Interval | [1;4] | [1;10] | [4;10] | [4;10] | [2;10] | [4;10] | [4;10] | [1;4] | [1;4] | [4;10] |

It may be observed that even if the sample sizes are reasonably big (100-479), the lengths of the confidence intervals are big in most cases. The cases A, H, I in the low register with estimates in the neighborhood of 0.5 are estimated to have among the 4 lowest ranks, while the cases C, D, F, G, J in the high register with estimates around 0.7 are estimated to have among the 7 highest ranks. For the middle cases B, E there is almost no information on their ranks. All this indicates that one can not expect short confidence intervals for ranks unless one has very big sample sizes.

## 3 Ordered categorical data

A common type of investigation on the service of medical care units is based on questionnaires, where patients are asked to evaluate the service of the visited unit in a categorical scale, with a number of ordered alternatives. In this situation there is often also an interest in ranking each unit within a set of units. Now we will show how one can construct confidence intervals for separate units.

Suppose that there are m units in the investigation and that the scale has K ordered alternatives. The different centers (units) may have different numbers of respondents, which are denoted by $N_1, N_2, \ldots, N_m$. In the general model for the data the probability for category k in unit i equals $q_{ik}$, i.e. the result for unit i is

multinomial with parameters $N_i$ and $(q_{i1}, q_{i2}, \ldots, q_{iK})$. We denote the number of results in category k for unit i by $N_{ik}$. The number of unknown parameters in this general model equals $m(K-1)$.

Since the data is obtained from ordered verbal expressions it is not really numeric, but has only an ordered categorical character. Cox and Hinkley (1974) in the section 6.3 on rank tests on page 187 say: 'Statistics that involve the observations only through (the rank) r are therefore particularly appropriate if the scale of measurement is so arbitrary defined that it does little more than define the ordering of the different values.' See also Altman (1991) section 2 for a discussion of different data types. On page 16 he comments on the continuous visual analogue scale (VAS) that there is no absolute meaning of the score and one might prefer a method of analysis based on the rank ordering of the scores. This is true for the discrete verbal scales as well. Statistical methods in this situation ought preferably to be invariant under monotone transformation, which means that the methods should be based on ranks or inversions only.

For comparison between two units, e.g. units number i and j, a suitable comparison parameter is the probability of the result in unit i being more favorable than the result in unit j minus the probability of the result in unit j being more favorable than the result in unit i. Thus in this ranking problem of type 2 we have a theoretical comparison parameter

$$\lambda_{i,j} = \sum_{k=1}^{K} q_{ik} \left( \sum_{l=1}^{k-1} q_{jl} - \sum_{l=k+1}^{K} q_{jl} \right).$$

If the two units have the same distribution on categories, the parameter will be 0, a positive value indicates that unit i has a tendency of getting higher values than unit j and a negative value indicates the opposite. This is a non-parametric position parameter, which is symmetric in the two categories and adapted for cases where coinciding values are important. It is related to the main parameter in the Wilcoxon-Mann-Whitney test.

These comparison parameters may simply be estimated by using the relative frequencies for the corresponding probabilities. In formulas the estimate of the probability difference equals

$$Z_{ij} = \frac{1}{N_i N_j} \sum_{k=1}^{K} N_{ik} \left( \sum_{l=1}^{k-1} N_{jl} - \sum_{l=k+1}^{K} N_{jl} \right)$$

where

$$N_i = \sum_{j=1}^{k} N_{ij} .$$

With notations from the general description we can now use $Z_{i,j}$ for $T_{i,j}^{(-)}$ and $-Z_{i,j}$ for $T_{i,j}^{(+)}$.

We suppose that the number of individuals in the units are big enough to make a normal approximation of the multinomial distribution applicable. Using the delta method we can get an approximate variance and asymptotic normality for our estimate. For the delta method, which means linearization to differential elements, see e.g. Serfling (1980) or Lehmann (1999). Straightforward calculation shows that this variance is estimated by

$$V_{ij} = \frac{1}{N_i}\left(W_{ij} - Z_{ij}^2\right) + \frac{1}{N_j}\left(W_{ji} - Z_{ji}^2\right)$$

where

$$W_{ij} = \frac{1}{N_i N_j^2} \sum_{k=1}^{K} N_{ik} \left( \sum_{l=1}^{k-1} N_{jl} - \sum_{l=k+1}^{K} N_{jl} \right)^2 .$$

Now consider a fixed unit, for instance unit number i. For each one of the other units, exemplified by unit j, we can calculate the two one-sided p values for test of the probability parameter having value 0. In the normal approximation gives the one-sided p values

$$1 - \Phi\left( \frac{Z_{ij}}{\sqrt{V_{ij}}} \right)$$

or

$$\Phi\left( \frac{Z_{ij}}{\sqrt{V_{ij}}} \right)$$

for the tests on the two sides. The lay out of the calculation is the same as in the binomial example. The procedure may seem to include a very heavy calculation with all these multiple tests. In a statistics program package with possibility to program repetition of the same type of calculation it is in fact quite easy. For example the statistics package R, which is a free software available at www.r-project.org works well enough. Observe that in a sense we are working with ranks at two levels here, since the test for each pair is a variant of Wilcoxon-Mann-Whitney test, which is a rank test within that pair, and it is used for getting a confidence interval for the rank of a unit among m units. The two rank methods should not be confused.

In order to give insight in details of the procedure for the case of ordered categorical data I present a very small numerical example consisting of 10 units and a scale consisting of 5 ordered categories. The example is an artificial one, generated by computer. I have used a basic distribution for the response alternatives with probabilities 0.1, 0.15, 0.2, 0.35 and 0.2. This is transformed randomly by using a Lehmann alternative with an exponent of type $e^U$ where $U$ is normally distributed with mean 0 and standard deviation 0.2 for each unit in order to make some suitable difference between the cases. For Lehmann alternative see e.g. Maritz (1981) section 4.4. The sample sizes for the units are randomly chosen from a uniform distribution in the interval [100;500].

The units marked with letters A - J have the following results of number of respondents in category order from "worst" to "best":

A: 35, 44, 51, 88, 46 B: 19, 29, 34, 41, 30 C: 43, 52, 54, 99, 56

D: 42, 50, 70, 117, 61 E: 13, 37, 49, 85, 45 F: 15, 21, 28, 71, 35

G: 16, 40, 46, 86, 34 H: 19, 36, 39, 73, 31 I: 17, 37, 55, 126, 65

J: 16, 15, 18, 50, 20

Running now the described calculations we get the following 95 % intervals for the ranks of the different units. Observe again that the intervals are closed, so the ranks at the interval ends are estimated as 'possible'.

**Table 3**. Estimated rank intervals and numbers of ranks outside the confidence interval.

| Unit | Rank interval 95 % | Number of rejected ranks |
|---|---|---|
| A | [1;8] | 2 |
| B | [1;7] | 3 |
| C | [1;7] | 3 |
| D | [1;8] | 2 |
| E | [6;9] | 6 |
| F | [7;10] | 6 |
| G | [1;9] | 1 |
| H | [1;8] | 2 |
| I | [8;10] | 7 |
| J | [1;9] | 1 |

The intervals have an individual confidence of 95 %. The lengths of the intervals depend on the numbers of respondents in the units as well as the difference of the true probability distributions between units.

Most intervals are very long even if the sample sizes in the example are reasonably big and there are reasonable differences between the distributions on categories. The only cases with shorter intervals are three units E, F and I in the upper region.

All this indicates that the interval estimation of rank is a demanding task which requires very big sample sizes unless the differences in distribution really are substantial.

## 4 Non-transitivity

Non-transitivity may occur in pair-wise ranking in theoretical distributions as well as in empirical distributions. It means that some pair-wise ordering contradicts consequences of other pair-wise orderings. The problem is known in non-parametric statistics for a long time. The problem is discussed e.g. in Brown and Hettmansperger (2002).

Let us now consider our individual rank definition for an example where three units A, B and C have got the pair-wise orderings A>B, B>C and C>A, and all three units being superior to $n_1$ other units and inferior to $n_2$ other units. Following our definition the unit A will get the rank $n_1 + 2$ since it is considered superior to B and inferior to C. For this determination of the rank (theoretically or empirically) of the unit A, the internal ranking between B and C has no influence. And when making the confidence set of the theoretical rank of A all other units have a clear position in one of the categories inferior to, equal to and superior to unit A, determined by the pair-wise measure in relation to the unit A only. Individually the units B and C in the example would get the same rank as A, which seems illogical if one thinks of a global common ranking. Non-transitivity may appear in even more complicated arrangements than in the above example. We are, however, only interested in confidence intervals for each unit separately here, and we do not consider those intervals in some multiple inference setting. In our definitions and in our confidence interval construction we never work with a global ranking. For each single unit we make two separate multiple tests, one in each direction, in order to find how many other units can be declared different in that direction bases on difference parameters and empirical measures in relation to the unit for which we are making the confidence interval. The only difference parameters involved and the only empirical difference statistics involved are those between the unit for which we construct the confidence interval and one of the other units. When we make individual confidence intervals for the ranks of all m units, we make 2m separate multiple test in all. In the step-down Bonferroni type of procedure suggested, everything is based on p values in tests for individual pairs of units. The only requirement is that the test for each pair gives a correct p value for that particular pair itself.

## References


Altman, D. G. (1991). *Practical Statistics for Medical Research*. Chapman and Hall, London.

Brown, B. M., Hettmansperger, T. P., (2002). Kruskal-Wallis, multiple comparisons and Efron dice. *Austral. New Zealand J. Statist.* **44** (4), 427-438

Cox, D. R., Hinkley, D.V., (1974). *Theoretical Statistics*. Chapman and Hall, London.

Holm, S., (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65-70.

Lehmann, E. L., (1999). *Elements of Large-Sample Theory*. Springer-Verlag, New York

Maritz, J. S., (1981). *Distribution-Free Statistical Methods*. Chapman and Hall, London

Serfling , R. J., (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons. New York.