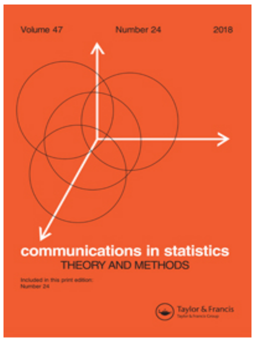
Confidence intervals for distributional positions

Citation for the original published paper (version of record):
Holm, S., Larsson, R. (2024). Confidence intervals for distributional positions. Communications in Statistics - Theory and Methods, 53(7): 2645-2660.
http://dx.doi.org/10.1080/03610926.2022.2148468



Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Confidence intervals for distributional positions

Sture Holm & Rolf Larsson

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2022.2148468

Confidence intervals for distributional positions

Sture Holm[a] and Rolf Larsson[b]

[a]Department of Mathematical Sciences, Chalmers University, Gothenburg, Sweden; [b]Department of Mathematics, Uppsala University, Uppsala, Sweden

ABSTRACT
A common problem with ranking lists (e.g., regarding success/failure proportions of treatments at hospitals) is that smaller units tend to end up either in the top or in the bottom just by pure chance. To alleviate this problem, we propose a method that, for a given unit, gives a confidence interval for the position of this unit within the distribution of the other units. The confidence interval is based on asymptotic normality. The method is illustrated by an empirical example. The small sample confidence level is investigated in a simulation study.

ARTICLE HISTORY
Received 21 September 2020
Accepted 14 November 2022

KEYWORDS
Ranking; confidence interval; asymptotic normality

1. Introduction

It is rather common that ranking data are presented in media and elsewhere in the society. Mostly no real statistical analysis of the data is made at all. Here, we will suggest a general method to obtain a suitable type of confidence statements and prove an asymptotic result on the confidence degree.

We formulate the statistical problem in the following model. There are $n+1$ units and we call one of them the investigated unit and assign it index 0. The other n units are called reference units with indexes $1, 2, \dots, n$. Each unit has its own value of a parameter of interest. It is natural to consider the parameters to be random and independent.

Denote the parameters in the reference group by Z_i, $i = 1, 2, \dots n$, and suppose that they have some unknown cumulative distribution function $F(z)$ and density function $f(z)$. The unit parameters are not directly observable, but only estimated with some estimates Y_i, $i = 1, 2, \dots, n$. These are supposed to be based on samples of sizes N_i and have variance estimates V_i. For the investigated unit we use the same notation with index 0.

Typically, Z_i is a probability (for example the risk for complications of a surgery at hospital i), Y_i is a relative frequency (the corresponding observed proportion) and $V_i = Y_i(1 - Y_i)/N_i$. In other cases Z_i is an expectation parameter, Y_i is an empirical mean and V_i is an empirical variance.

A rank of Y_0 in the set Y_i, $i = 0, 1, 2, \dots n$ gives some information on the position of the investigated unit in the distribution of the reference units. It gives an estimate of the percentile in the distribution of the reference unit parameters as well as an estimate

CONTACT Rolf Larsson rolf.larsson@math.uu.se Department of Mathematics, Uppsala University, P.O. Box 480, SE-751 06 Uppsala, Sweden.

of the distributional position of the investigated unit. The latter is here defined as $p = F(Z_0)$.

The aim of this paper is to suggest a method to get a confidence interval for the distributional position p of the investigated unit and to prove that the intended confidence degree is asymptotically correct. The confidence interval (region) at level $1-\alpha$ (say) is constructed as the set of $p = p_0$ such that the null hypothesis H_0: $p = p_0$ is not rejected vs. the alternative hypothesis H_1: $p \neq p_0$ at level α.

A confidence interval for the distributional position of the investigated unit means that with a small risk of error, we can exclude positions which are too small or too large. There is always a correspondence between confidence intervals and tests for different parameters (here distributional positions). Test aimed to exclude a parameter value will be based on the difference between the observation unit estimate and an estimate of the reference unit distribution divided by a standard error. It will be seen that there are two sources of variation involved in the problem, the variances for the unit estimates and an intrinsic variance due to the finiteness of the number of comparison units.

It should be mentioned that the above problem has been addressed in Klein, Wright, and Wieczorek (2020), but their approach is different. They construct joint confidence intervals for ranks of units on the basis of simultaneous confidence intervals for individual specific parameters. In contrast to this, our approach is more of a non parametric nature, in that we consider the unit parameters to be random.

In section 2 we will motivate our procedure. Section 3 states its asymptotic statistical properties and gives heuristic motivations. In section 4, the asymptotic result is proved. Section 5 gives an example of the use of the method, with data on mortality during treatment after a heart attack for 70 Swedish hospitals in the years 2007 to 2009. In section 6, we describe and show results of a conducted simulation study, and conclude in section 7.

2. A simplified motivation for the method

In a typical real life situation our problem is quite irregular. The sample sizes for the units are usually not the same and the characteristic unit parameters Z_i, $i = 1, 2, \ldots, n$, are random and may have any unknown distribution. Here, in order to get a basic idea of a suitable procedure, we will consider a very simplified situation. Thus, we suppose for the moment that all sample sizes are so big that the characteristic unit parameters can be considered known.

For a given distributional position $p \in (0, 1)$, the number of units with random parameters below p in a group of n independent reference units follow a binomial distribution with parameters n and p. We cannot be sure which unit estimate is in fact closest to the percentile. Thus, even if there is an extremely good precision in the estimates of the unit parameters, there is a random variation in the estimation of the population percentiles due to the finiteness of the number of units. We can call this the intrinsic variation.

This variation means that we are not sure of which order statistic is the one closest to the percentile. This can naturally be taken into consideration by weighting the order statistics suitably.

There are n ordered statistics in the reference group. In the uniform (0, 1) distribution the order statistics have expectations

$$\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}. \tag{1}$$

Any other continuous distribution can be considered to be a transformation of this uniform distribution with the inverse cumulative distribution function F^{-1}. As a first rough attempt, one could estimate the $m/(n+1)$ percentile $F^{-1}\left(\frac{m}{n+1}\right)$ in the general distribution by the m:th order statistic. A better estimate, however, can be obtained by using a weighted mean of "nearby" order statistics.

Consider any percentile corresponding to p in the general distribution of the n reference units. If one of the observations is in a small neighborhood of p, the number of smaller observations among the other $n-1$ reference units is binomial distributed with parameters $n-1$ and p. This shows that not only the order statistic m would be a reasonable estimate of $F^{-1}\left(\frac{m}{n+1}\right)$, but that also the order statistics with similar numbers would be quite reasonable. And if they are weighted suitably we could get an estimate with smaller variance and small bias.

Now, we consider estimates with weights determined by a binomial distribution with parameters $n-1$ and some q. This distribution has n possible outcomes $0, 1, 2, \ldots, n-1$. We attach those probabilities as weights to the order statistics with numbers $1, 2, 3, \ldots, n$. If the distribution of the reference units is uniform with parameters (0, 1) the expectation of the order statistics is as in (1) and the expectation of the weighted mean is

$$\sum_{k=0}^{n-1} \binom{n-1}{k} q^k (1-q)^{n-1-k} \frac{k+1}{n+1} = \frac{1+(n-1)q}{n+1}.$$

Thus we get an unbiased estimate of the percentile

$$p = \frac{1+(n-1)q}{n+1}$$

by using these weights for the order statistics. This holds not only for the uniform distribution with parameters (0, 1) (where it is in fact known), but also for any uniform distribution. For a general continuous distribution this is not exactly true, but it would hold approximately for big numbers of units n if the density f is differentiable.

Thus now, in order to get a basis for a test of the null hypothesis that the investigated unit has distributional position p, we calculate a "binomial parameter" q by

$$q = \frac{p(n+1)-1}{n-1}. \tag{2}$$

Then, the weight for order statistic number k is the probability for outcome in the point $k-1$ in the binomial distribution with parameters $n-1$ and q, that is

$$b_k = \binom{n-1}{k-1} q^{k-1} (1-q)^{(n-1)-(k-1)} = \binom{n-1}{k-1} q^{k-1} (1-q)^{n-k}, \tag{3}$$

and so, the statistic that we will study in the sequel is based on

$$R_p = Y_0 - \sum_{k=1}^{n} b_k Y^{(k)}, \tag{4}$$

where $Y^{(k)}$ is the kth order statistic in $(Y_1, \dots, Y_n)$. Note that, by construction of the b_k and the fact that $E(Y_0) = p$, we have $E(R_p) = 0$.

Observe that there is a p dependence hidden in $q = q(p)$. As a function of p, $q(p)$ is a suitable smoothed estimate of the cumulative distribution function.

Already in an idealized situation with extremely big unit sample sizes, there is an intrinsic variance. In a realistic case we must also consider the variance arising from the errors in the internal estimation of parameters in units based on finite sample sizes N_i, $i = 1, 2, \dots, n$. The total variance in the test statistic will consist of two components, this variance and a variance due to the finiteness of sample size for the different reference units.

In the next section we will give a heuristic motivation of formulas for variance estimation which will later be proved in an asymptotic theorem in section 4.

3. Heuristics on asymptotic variance

In the next section we will formulate and prove asymptotic properties for the suggested test statistic $R_p/\hat{\sigma}_R$, where $\hat{\sigma}_R^2$ estimates the variance of R_p, which is defined later in this section. In this section we will now only give a sketch of the results in order to explain basic ideas.

We now consider $\hat{\sigma}_R^2$. There are two components of variation in the estimate, the randomness in the order statistics due to finite sample size in each unit and an intrinsic variance due to finiteness of the number of units. The second one remains true even if the variance for each individual unit is very small due to big sample size in the estimates for individual units. From now on we assume that the number of observations in each unit is big enough such that the unit's expected value is exactly obtained.

Consider a random sample $Y_1, \dots, Y_n$ from a distribution with density function $f(\cdot)$. Assume that the set of order statistics is

$Y^{(1)} < Y^{(2)} < \dots < Y^{(n)}$ with corresponding p values $p_i = F(Y^{(i)})$. Let $f_i = f(Y^{(i)})$. We have the following theorem, first given by Smirnov (1935) generalized by Mosteller (1946), also given in Wilks (1962).

Theorem 1. *For each integer n, define a sequence of integers* $0 < n_{1n} < n_{2n} < \dots < n_{kn} < n$. *Assume that*

1. *As* $n \to \infty$, $\lim_{n\to\infty} \frac{n_{in}}{n} \to \lambda_i$ *for* $i = 1, 2, \dots, k$, *where* $\lambda_1 < \lambda_2 < \dots < \lambda_k$.
2. *The probability density function f(x) is continuous, and does not vanish in the neighborhood of* u_i, *where*

$$\int_{-\infty}^{u_i} f(x)dx = \lambda_i,$$

for $i = 1, 2, \dots, k$.

If $x(1), \dots, x(n)$ *are drawn from f(x) satisfying condition 2, and if* $n_{1n}, \dots, n_{kn}$ *satisfy condition 1, then the* $x(n_{in})$, $i = 1, 2, \dots, k$, *are asymptotically multivariate normal distributed, with expectations* u_i, *variances*

$$c_{ii} = \frac{\lambda_i(1-\lambda_i)}{nf(u_i)^2} \tag{5}$$

for $i = 1, \ldots, k$, *and covariances*

$$c_{ij} = \frac{\lambda_i(1-\lambda_j)}{nf(u_i)f(u_j)} \tag{6}$$

for all $1 \le i < j \le k$.

For the case of extremely well estimated unit values, we use the c_{ij} together with the binomial weights in a calculation of the intrinsic variance of the estimate R_p. Theorem 1 is applied for the p values $p_i = \lambda_i$, $i = 1, 2, \ldots, n$. We use the probability points $p_i = i/(n+1)$, since these are the expectations of the order statistics for the uniform distribution on the interval (0, 1). With these values thought of as members of the infinite sequence, we will employ the asymptotic result by Mosteller (1946) in a finite setting.

The densities $f_i = f(Y^{(i)})$ may be estimated by the reciprocal of the mean length between the observations Y_j (say) at the positional point p_i. In a point p close to p_i, the numbers of observation points below and above p are approximate Poisson distributed with intensities $\lambda_- = if_i$ and $\lambda_+ = (n-i)f_i$, respectively. The approximate mean length between observation points close to p_i is then written as

$$l_i = \frac{1}{nf_i}.$$

An empirical estimate, $\hat{l}_i$, will be obtained by again using the binomial weights with a small modification at the ends, according to

$$\hat{l}_i = \begin{cases} Y^{(2)} - Y^{(1)}, & i = 1, \\ \left(Y^{(i+1)} - Y^{(i-1)}\right)/2, & i = 2, 3, \ldots, n-1, \\ Y^{(n)} - Y^{(n-1)}, & i = n. \end{cases} \tag{7}$$

From (5)–(7), we now have the empirically estimated variances for order statistics

$$\hat{c}_{ii} = np_i(1-p_i)\hat{l}_i^2 \tag{8}$$

and covariances

$$\hat{c}_{ij} = np_i(1-p_j)\hat{l}_i\hat{l}_j \tag{9}$$

for $i < k$.

Next, we use this asymptotic result on our estimate R_p in (4). We get the estimated variance by the sum

$$\hat{\sigma}_R^2 = \mathrm{Var}(Y_0) + \sum_i np_i(1-p_i)\hat{l}_i^2 b_i^2 + 2\sum_{i<j} np_i(1-p_j)\hat{l}_i\hat{l}_j b_i b_j, \tag{10}$$

where $\mathrm{Var}(Y_0) = Y_0(1-Y_0)/N_0$, Y_0 being the observed proportion and N_0 the number of observations of the reference unit.

Here, observe that both the estimate itself and the estimate of its variance depend on the value p to be tested through the p-dependence in b_k. For each value p we now have a test statistic, R_p, and an estimate of its variance, $\hat{\sigma}_R^2$. The ratio of the estimate and its estimated standard error, i.e., $R_p/\hat{\sigma}_R$, is the derived test statistic, which has an asymptotic normal distribution with parameters 0 (mean) and 1 (variance). This is proved in the next section.

In other words, when testing H_0: $p = p_0$ vs H_1: $p \neq p_0$, we reject at level α if $|R_{p_0}/\hat{\sigma}_R|$ is greater than the standard normal $\alpha/2$ percentile, $\lambda_{\alpha/2}$ say. The $1-\alpha$ confidence region consists of the $p = p_0$ which are such that H_0 is not rejected at level α, i.e., such that $|R_p/\hat{\sigma}_R| \leq \lambda_{\alpha/2}$.

We have not proved that the derived test statistic is monotone, so in principle we have to check this test statistic not only at the boundaries (where its absolute value equals $\lambda_{\alpha/2}$) but also in all values of p outside the boundaries. In practice the derived test statistic will appear to be monotone in most cases. If, however, the dependence is not monotone, we have to find the most distant values of p where the hypothesis is "barely rejected".

4. Asymptotic theorem

Assume at first that sample sizes N_i, $i = 1, 2, \ldots, n$, are fixed. We want to prove that our test statistic, $R_p/\hat{\sigma}_R$ where R_p is as in (4) and $\hat{\sigma}_R^2$ is as in (10), tends in distribution to standard normal as the number of units, n, tends to infinity.

Consider (4). We note that R_p is a linear combination of Y_0 and the order statistics $Y^{(1)}, \ldots, Y^{(n)}$. Thus, by Mosteller's theorem (section 3), R_p is asymptotically normal. We have also seen that $E(R_p) = 0$ (section 2). Moreover, by (4),

$$\mathrm{Var}(R_p) = \mathrm{Var}(Y_0) + \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j \mathrm{Cov}(Y^{(i)}, Y^{(j)}). \tag{11}$$

Now, via Mosteller's theorem, for $i \leq j$ we get that as $n \to \infty$,

$$n\mathrm{Cov}(Y^{(i)}, Y^{(j)}) \to \frac{p_i(1-p_j)}{f_i f_j}, \tag{12}$$

where p_i and f_i are the values of the distribution and density functions of $Y^{(i)}$ respectively, hence fixed quantities. Furthermore, conditionally on Z_0, $\mathrm{Var}(Y_0) = Z_0(1 - Z_0)/N_0$, so it is clear that n times the unconditional variance of Y_0 is asymptotically finite as (n, N_0) tend to infinity in such a way that $n/N_0 \to c < \infty$. In the following, we call this limit v_0.

Equations (11) and (12) imply

$$n\mathrm{Var}(R_p) \to v_0 + \lim_{n\to\infty}\left\{\sum_{i=1}^{n} b_i^2 \frac{p_i(1-p_i)}{f_i f_j} + 2\sum_{i<j} b_i b_j \frac{p_i(1-p_j)}{f_i f_j}\right\}. \tag{13}$$

To see that the limit of the double sum in (13) is finite, we note that the double sum is equal to the variance of $\sum_{1\leq k\leq n} b_k Y^{(k)}$. The proof that the limit of this variance is finite is given in Appendix 1.

Because of the Slutsky theorem, the proof will be completed if we can show that n times the estimated variance, given in (10), tends to the same limit as in (13). To this end, we note that

$$n\hat{\sigma}_R^2 = n\text{Var}(Y_0) + \sum_{i=1}^{n} b_i^2 p_i(1-p_i)(n\hat{l}_i)^2 + 2\sum_{i<j} b_i b_j p_i(1-p_j)(n\hat{l}_i)(n\hat{l}_j). \tag{14}$$

But then the result follows from the fact that per (7), the $n\hat{l}_i$ converge to $1/f_i$ as $n \to \infty$.

5. An empirical example

In this section, in order to show how the method works, we present an example. Details of the algorithms are given in Appendix 2. The data describes mortality during treatment after a heart attack for 70 Swedish hospitals in the years 2007 to 2009. The results are ordered according to adverse event rates. See Table 1.

We will construct a 95% confidence interval for unit number 19 ('Akademiska sjukhuset', Uppsala) which has complication risk $Y_0 = 0.124$ in $N_0 = 2691$ operated patients. The other 69 hospitals will serve as reference units. The test statistic includes a p-dependent part, which is an estimate of the p:th quantile (called "moving average" in Figure 1). Figure 1 shows the dependence together with the order statistics in the reference group. The order statistics are placed in the points $1/70, 2/70, \ldots, 69/70$.

The measurement standard error for the investigated unit is 0.0063. The standard error in the estimate of the percentiles in the reference distribution depends on p, is fairly constant and less than 0.01 for $p \leq 0.8$, but then seems to increase noticeably for $p > 0.8$. See Figure 2.

Table 1. Mortality during treatment after heart attack for 70 Swedish hospitals in the years 2007 to 2009.

Region no.	1	2	3	4	5	6	7	8	9	10
Risk	.0967	.0992	.1055	.1082	.1109	.1110	.1110	.1114	.1132	.1133
Size	362	2309	1678	804	2245	1559	2081	1086	1872	1465
Region no.	11	12	13	14	15	16	17	18	19	20
Risk	.1141	.1141	.1145	.1168	.1173	.1196	.1198	.1198	.1241	.1243
Size	3505	841	2489	2175	392	736	2279	409	2691	732
Region no.	21	22	23	24	25	26	27	28	29	30
Risk	.1263	.1285	.1287	.1309	.1315	.1316	.1340	.1346	.1354	.1359
Size	388	599	785	596	540	904	1977	6002	1167	655
Region no.	31	32	33	34	35	36	37	38	39	40
Risk	.1373	.1374	.1375	.1377	.1378	.1380	.1386	.1418	.1439	.1439
Size	2833	3587	669	2259	1655	1471	1717	2369	834	660
Region no.	41	42	43	44	45	46	47	48	49	50
Risk	.1447	.1451	.1456	.1464	.1466	.1492	.1496	.1497	.1522	.1533
Size	1092	448	309	3006	805	516	1210	1630	552	1957
Region no.	51	52	53	54	55	56	57	58	59	60
Risk	.1555	.1563	.1628	.1645	.1652	.1666	.1679	.1720	.1744	.1783
Size	1383	646	596	310	690	2473	137	436	1640	415
Region no.	61	62	63	64	65	66	67	68	69	70
Risk	.1813	.1839	.1940	.1973	.2085	.2093	.2173	.2275	.2362	.2880
Size	888	892	928	446	470	430	635	422	436	125

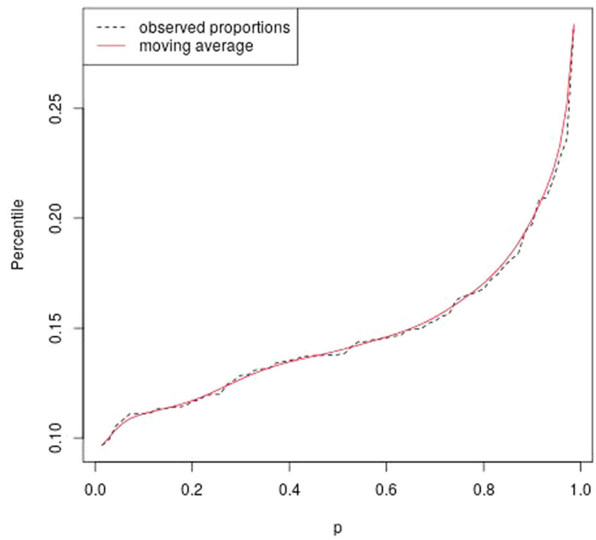


Figure 1. Quantile estimates.

Now, for each p we construct a t type test as the difference between the estimate for the investigated unit and the estimated p percentile in the reference distribution, divided by the standard error for the estimate difference, i.e., $R_p/\hat{\sigma}_R$. It is depicted in Figure 3.

The approximate 95% two-sided confidence interval for p is $(0.10, 0.46)$. Approximately, it can be obtained by reading off the ±1.96 points in Figure 3 (see the dotted and dashed lines), and for better precision it is obtained by examining the test statistic calculation for different values of p in the neighborhood of the bounds of the confidence interval.

It may be thought that we have then neglected the random measurement errors, but this is not the case. Our estimated variance is made for the cumulative distribution function including this variance. It is only the minor influence on the estimate itself which is neglected here. If the curvature of the cumulative distribution function is big, the correction can be considerable. The distribution for the true unit parameters has smaller variance than the distribution for the parameters obtained by measurement including measurement errors. The correction aims at adjusting for this and tends generally to make the estimated distribution having smaller variance. This means in practice that the confidence interval will be longer with the correction, which is natural.

6. Simulation study

In this section, we present simulations of empirical coverage probabilities for finite sample sizes. The simulations have been run in Matlab 2019a. Program codes are available upon request.

The sample sizes for different units (N_i) were assumed to be large. The estimate for the unit of interest, Y_0, was calculated as the pth percentile from a beta distribution with parameters a and b. For the reference units, $Y_1, \ldots, Y_n$ were generated from the

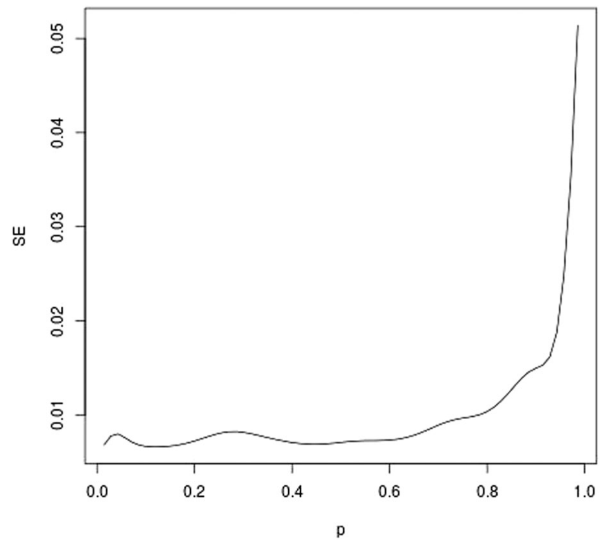


Figure 2. Standard errors.

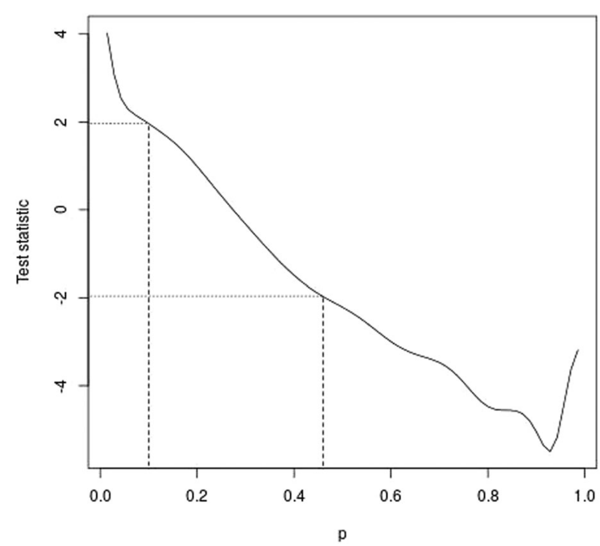


Figure 3. Test statistics.

same beta distribution, using 10^7 replicates. Then, the empirical coverage was computed as the percentage of times where $-\lambda_{\alpha/2} \le R_p/\hat{\sigma}_R \le \lambda_{\alpha/2}$, where λ_α is defined through $P(Z > \lambda_\alpha) = \alpha$ for Z standard normal.

We used $1 - \alpha = 0.95$, $n \in \{20, 40, ..., 300, 340, ..., 500\}$, $p \in \{0.2, 0.5\}$ and $(a, b) \in \{(10, 10), (15, 5), (5, 15), (50, 50), (75, 25), (25, 75)\}$, excluding $(a, b) \in \{(5, 15), (25, 75)\}$ for $p = 0.5$ because of symmetry.

In the $(a, b) = (10, 10)$ case, we added

$n \in \{750, 1000, 1250, 1500, 2000, 2500, 3000, 4000, 5000\}$. For these n, to calculate the b_k coefficients we used the improved Stirling approximation,

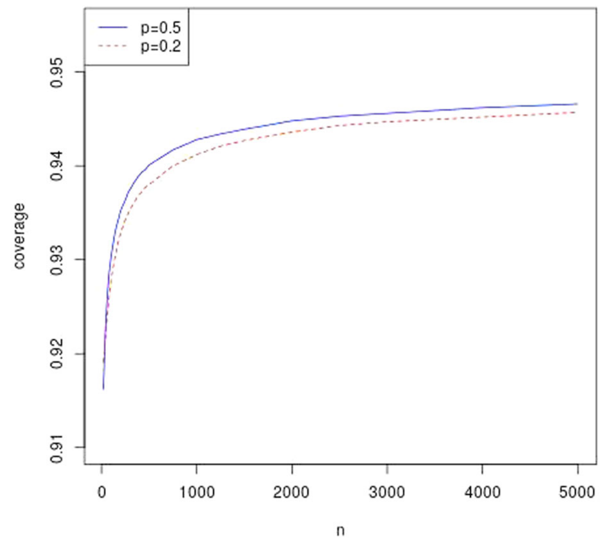


Figure 4. Coverage, (a, b) = (10, 10).

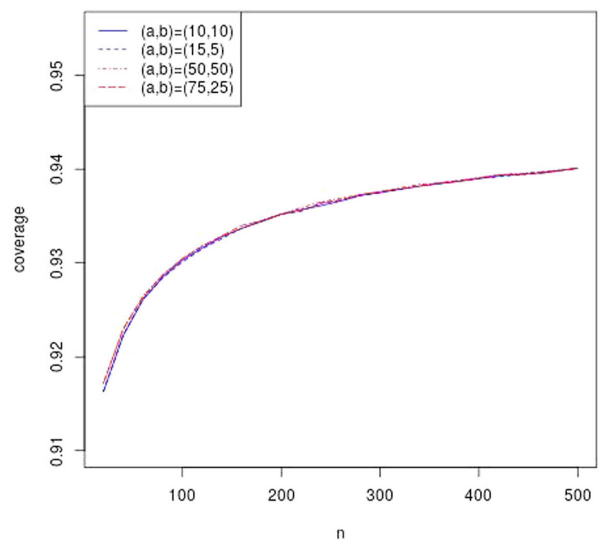


Figure 5. Coverage, $p = 0.5$.

$$n! \approx \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \left(1 + \frac{1}{12n}\right),$$

cf Wrench (1968). This is a way for the computer to be able to compute $n!$ for large values of n efficiently.

The simulation results are depicted in Figures 4–6 (see also Table 2). These figures show that the real confidence degree (empirical coverage) approaches the nominal confidence degree from below. That is quite natural since the estimate of difference is normalized by dividing by the estimated standard deviation, which gives an extra random variation. It is the same type of effect as if, in the simple situation of estimating the

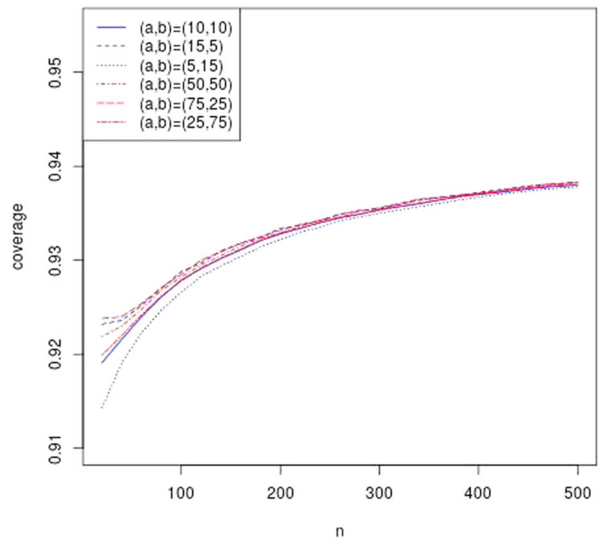


Figure 6. Coverage, $p = 0.2$.

Table 2. Empirical coverage probabilities for the case $(a, b) = (10, 10)$.

	Coverage probability	
Sample size	$p = 0.2$	$p = 0.5$
20	0.9191	0.9162
40	0.9216	0.9223
60	0.9240	0.9261
80	0.9261	0.9286
100	0.9279	0.9303
120	0.9291	0.9316
140	0.9302	0.9328
160	0.9311	0.9337
180	0.9321	0.9344
200	0.9328	0.9352
220	0.9334	0.9357
240	0.9340	0.9361
260	0.9345	0.9366
280	0.9349	0.9372
300	0.9354	0.9375
340	0.9360	0.9382
380	0.9368	0.9388
420	0.9372	0.9393
460	0.9377	0.9396
500	0.9380	0.9401
750	0.9400	0.9417
1000	0.9412	0.9428
1250	0.9421	0.9434
1500	0.9427	0.9439
2000	0.9436	0.9448
2500	0.9443	0.9453
3000	0.9447	0.9456
4000	0.9452	0.9462
5000	0.9457	0.9466

theoretical mean with series of normally distributed observations, the asymptotic normal distribution percentage value is used instead of the t distribution function value. The convergence here seems to be quite slow. The reason is that the effective number of

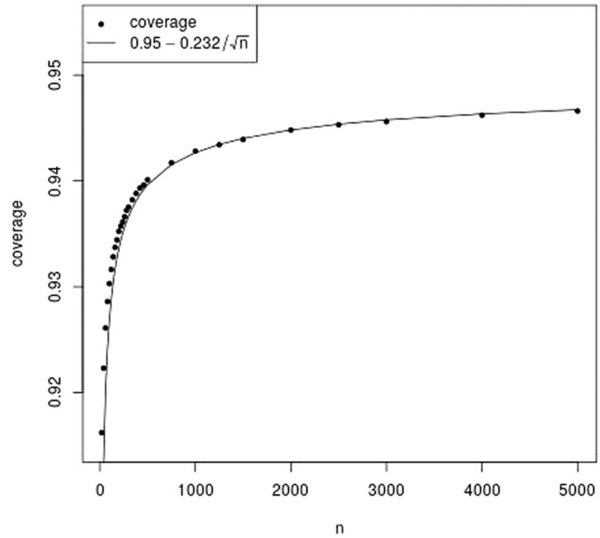


Figure 7. Coverage, (a, b) = (10, 10) $p = 0.5$.

Table 3. Mean confidence interval length for the case $(a, b) = (10, 10)$.

	Mean confidence interval length	
Sample size	$p = 0.2$	$p = 0.5$
20	0.304	0.378
40	0.227	0.274
60	0.188	0.229
80	0.163	0.202
100	0.146	0.182
120	0.134	0.168
140	0.124	0.156
160	0.117	0.147
180	0.110	0.139
200	0.105	0.132
220	0.100	0.126
240	0.096	0.121
260	0.093	0.117
280	0.089	0.112
300	0.087	0.109
340	0.082	0.102
380	0.077	0.097
420	0.074	0.093
460	0.071	0.089
500	0.068	0.085

observations used for estimating a percentile in the reference distribution is of the order of square root of sample size. This effect is greater for $p = 0.2$ than for $p = 0.5$ since the binomial variance is smaller.

Another observation from the figures is that for $p = 0.2$, the coverage is higher for higher $a + b$ and for higher a.

A further illustration of the convergence issues is given in Figure 7. Here, as in Figure 4, we have plotted the simulated coverage for the case $(a, b) = (10, 10)$, $p = 0.5$, but now along with the approximation

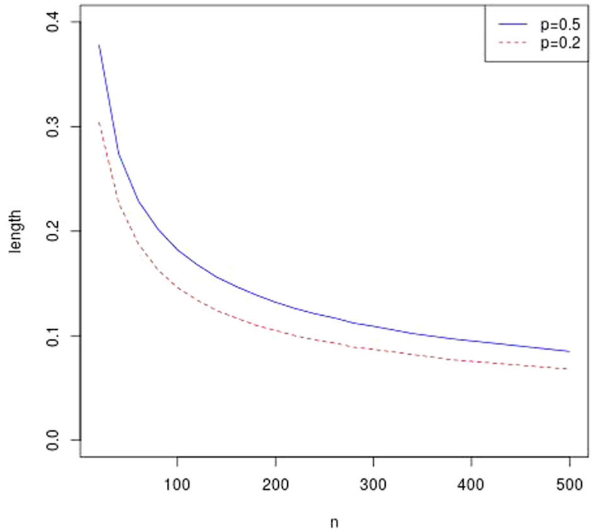


Figure 8. Mean c.i. length, (a, b) = (10, 10).

$$coverage \approx 0.95 - \frac{0.232}{\sqrt{n}}.$$

The constant 0.232 is obtained from a regression without intercept of 0.95 minus the simulated coverage on $1/\sqrt{n}$, for the nine n values of 750 and higher. For such large n, this approximation is seen to fit very well to the simulated coverage values.

Finally, for the case $(a, b) = (10, 10)$ and $p \in \{0.2, 0.5\}$, we have simulated the average lengths of the confidence intervals, see Table 3 and Figure 8. Because these simulations were quite time consuming, we stopped at $n = 500$, and only 10^5 replicates were used. (By comparing to preliminary simulations, we have found that the results are quite stable up to three decimal points.) The main findings are that the confidence intervals are shorter for $p = 0.2$ than for $p = 0.5$, and that as the sample size n increases, the lengths decay about proportionally to $1/\sqrt{n}$.

7. Conclusion

In this paper, we have suggested a statistical method to analyze ranking lists. For a given unit, the method gives a confidence interval for the distributional position of the unit within the distribution of the other (reference) units. This gives a way to statistically judge if the ranking of a certain unit as particularly good or bad is really significant, or if it was ranked in an extreme way mainly by chance, which is something that could happen to relatively small sized units.

The method is quite general, and works under a minimum of assumptions, as long as the number of inverstigated units as well as the numbers of observations for the units are sufficiently large.

One problem still not solved by our method is how to deal with simultaneous confidence intervals for the positions of several units. This could be an issue for further work on the subject.

Acknowledgments

We would like to thank Svante Janson, Dept. of Mathematics, Uppsala University, for helping out with the proof in the appendix that the variance is asymptotically finite. We are also grateful to the anonymous referee, for many constructive comments and remarks that greatly helped to improve the paper.

References

Janson, S., T. Luczak, and A. Rucinski. 2000. *Random Graphs*. New York: John Wiley & Sons.
Klein, M., T. Wright, and J. Wieczorek. 2020. A joint confidence region for an overall ranking of populations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 69 (3):589–606. doi:10.1111/rssc.12402.
Mosteller, F. C. 1946. On some useful" inefficient" statistics. *The Annals of Mathematical Statistics* 17 (4):377–408. doi:10.1214/aoms/1177730881.
Smirnov, N. 1935. Ûber die Verteiling des allgemeinen Gliedes in der Variationsreihe. *Metron* 12: 59–81.
Wilks, S. 1962. *Mathematical Statistics*. New York: John Wiley & Sons.
Wrench, J. W. 1968. Concerning two series for the gamma function. *Mathematics of Computation* 22 (103):617–26. doi:10.1090/S0025-5718-1968-0237078-4.

Appendix 1

In this Appendix, we prove that the limit of the variance of

$$W_n = \sum_{k=1}^{n} b_k Y^{(k)},$$

as $n \to \infty$, is finite.

Let X be a discrete random variable with probability function b_{k+1} as in (3), i.e., binomial with parameters $m = n-1$ and q. Furthermore, we may assume that X is independent of all Y_i. Moreover, let $Z = Y^{(X+1)}$. Then, it follows that $W_n = E(Z|Y_1, \ldots, Y_n)$.

Suppose that a is so small that

$$F(a) = P(Y_i \le a) < qe^{-2}/2 \tag{15}$$

and that $mq \ge 8$. Now,

$$P(Z \le a, X > mq/2) \le P\left(Y^{([mq/2])} \le a\right), \tag{16}$$

where $[v]$ is the integer part of v. Since $F(Y)$ is uniform on the unit interval, the r.h.s. of (16), which says that at least $[mq/2]$ of the Y_is are less than a, equals the probability that a binomial $(n, F(a))$ variate is greater than or equal to $mq/2$.

From corollary 2.4 of Janson et al. [1], this can be majorized by

$$\begin{aligned}\exp\left[-\frac{mq}{2}\left\{\log\left(\frac{q}{2F(a)}\right)-1\right\}\right] &\le \exp\left[-\frac{mq}{4}\log\left(\frac{q}{2F(a)}\right)\right]\\ \le \exp\left[-\left\{\frac{mq}{8}+\frac{mq}{8}\log\left(\frac{q}{2F(a)}\right)\right\}\right] &\le \frac{2F(a)}{q}\exp(-c_1 m),\end{aligned}$$

for a constant c_1. Moreover, it follows similarly that

$$\begin{aligned}P(Z \le a, X \le mq/2) &\le P(Y^{(1)} \le a)P(X \le mq/2)\\ &\le nF(a)C_1 \exp(-c_2 m) \le F(a)\exp(-c_3 m),\end{aligned}$$

($P(Y^{(1)} \leq a) \leq mF(a)$ follows because the l.h.s. equals the probability that some $Y_i \leq a$, which in turn is smaller than the r.h.s.) for some constants C_1, c_2, c_3, and so, putting the results together we find

$$P(Z \leq a) \leq \exp(-c_4 m)F(a), \tag{17}$$

for some constant c_4, for all large enough m and a with $F(a) < qe^{-2}/2$.

Now, take $b > 0$ such that $P(Y_i \leq -b) \leq q/2$. Together with (17), we get ($I\{A\}$ is the indicator of the event A)

$$E\left(Z^2 I\{Z \leq -b\}\right) = \sum_{k=0}^{\infty} E\left(Z^2 I\{-2^{k+1}b < Z \leq -2^k b\}\right)$$
$$\leq \sum_{k=0}^{\infty} 2^{2k+2} b^2 P\left(Z \leq -2^k b\right) \leq \sum_{k=0}^{\infty} 4b^2 2^{2k} C_2 e^{-c_5 m} P\left(Y_i \leq -2^k b\right)$$
$$= C_2 e^{-c_5 m} E\left(\sum_{2^k \leq -Y_i/b} 2^{2k}\right) \leq C_2 e^{-c_6 m} E(Y_i^2) = C_3 < \infty,$$

for constants C_2, C_3, c_5, c_6. (The last inequality follows since the sum is on the same order of magnitude as its last term.)

Similarly, by symmetry, for b sufficiently large,

$$E\left(Z^2 I\{Z \geq b\}\right) \leq C_4 < \infty$$

for some constant C_4, and we have

$$E(Z^2) \leq C_3 + C_4 + b^2 < \infty.$$

Finally, by the Jensen inequality,

$$W^2 = \{E(Z|Y_1, ..., Y_n)\}^2 \leq E(Z^2|Y_1, ..., Y_n),$$

and by taking expectations w.r.t. the Y_is, $E(W^2) \leq E(Z^2)$, and we are done.

Appendix 2

This Appendix gives some further details on the algorithms of Sections 5 and 6. Program codes in Matlab2019a may be obtained at request.

To obtain the confidence interval for p and Figures 1–3, the algorithm is as follows:

1. Use the observed proportion Y_0 and the number of observations N_0 for the reference unit to calculate $\text{Var}(Y_0) = Y_0(1 - Y_0)/N_0$.
2. Given the sorted input data $Y^{(i)}$, calculate the weights $\hat{l}_i$ from (7).
3. For $i = 1, ..., n$, do
   (a) Let $p_i = i/(n + 1)$.
   (b) Calculate the b_k coefficients from (3), where for q we use (2) with $p = p_i$.
   (c) Use the b_k to calculate the correction term $\sum_{k=1}^n b_k Y^{(k)}$ in (4).
   (d) Use $\text{Var}(Y_0)$, p_i, the $\hat{l}_i$ and the b_i to calculate the estimated variance $\hat{\sigma}_R^2$ in (10).
   (e) Plug in Y_0 and the correction term to get R_p from (4), then use $\hat{\sigma}_R$ to obtain the observed test statistic $t(i) = R_p/\hat{\sigma}_R$.
4. Compute the upper limit of the $1 - \alpha$ confidence interval as the proportion of $t(i)$ greater than $-\lambda_{\alpha/2}$.
5. Compute the lower limit of the $1 - \alpha$ confidence interval as the proportion of $t(i)$ greater than $\lambda_{\alpha/2}$.
6. For Figure 1, plot the correction terms $\sum_{k=1}^n b_k Y^{(k)}$ and the sorted input data $Y^{(i)}$ vs the p_i.

7. For Figure 2, plot the estimated standard errors $\hat{\sigma}_R$ vs the p_i.
8. For Figure 3, plot the test statistics $t(i)$ vs the p_i.

Items 4 and 5 may seem counterintuitive, but this is because the larger the p_i, the smaller the $t(i)$. Cf Figure 3.

In the coverage probability simulations, the algorithm may be simplified (and quicker to execute), because for this purpose, we only need to check if p is inside the confidence interval. Moreover, to simplify we skip the $\mathrm{Var}(Y_0)$ term (assuming that the corresponding sample size, N_0, is very large). In this case, for each replicate we use the following algorithm:

1. Generate Y_0 and the $Y^{(i)}$ from the specified distribution, including the given p and n. It holds that $p = F(Y_0)$, where F is the distributon function used in the data generation.
2. Given the $Y^{(i)}$, calculate the weights $\hat{l}_i$ from (7).
3. Calculate the b_k coefficients from (3), where for q we use (2) with the given p.
4. Use the b_k to calculate the correction term $\sum_{k=1}^{n} b_k Y^{(k)}$ in (4).
5. Use p, the $\hat{l}_i$ and the b_i to calculate the estimated variance $\hat{\sigma}_R^2$ in (10), neglecting $\mathrm{Var}(Y_0)$.
6. Plug in Y_0 and the correction term to get R_p from (4), then use $\hat{\sigma}_R$ to obtain the observed test statistic $t = R_p/\hat{\sigma}_R$.
7. The replicate generates a p inside the $1-\alpha$ confidence interval if $t^2 < \chi_\alpha^2(1)$, where $\chi_\alpha^2(1)$ is given by $P\{Q > \chi_\alpha^2(1)\} = \alpha$ for a variate Q that is chi square with one degree of freedom.