Statistical Hypothesis Test for the Difference Between Hirsch Indices of Two Pareto-Distributed Random Samples

Marek Gagolewski[1,2]

This is a revised version of the paper:
Gagolewski M., *Statistical Hypothesis Test for the Difference between Hirsch Indices of Two Pareto-Distributed Random Samples*, In: Kruse R. et al (Eds.), *Synergies of Soft Computing and Statistics for Intelligent Data Analysis* (AISC 190), Springer-Verlag, Heidelberg, 2013, pp. 359-367, doi:10.1007/978-3-642-33042-1_39

Abstract In this paper we discuss the construction of a new parametric statistical hypothesis test for the equality of probability distributions. The test bases on the difference between Hirsch's h-indices of two equal-length i.i.d. random samples. For the sake of illustration, we analyze its power in case of Pareto-distributed input data. It turns out that the test is very conservative and has wide acceptance regions, which puts in question the appropriateness of the h-index usage in scientific quality control and decision making.

1 Introduction

The process of data aggregation [cf. 7] consists in a proper synthesis of many numerical values into a single one, representative for the whole input in some sense. It plays a key role in many theoretical and practical domains, such as statistics, decision making, computer science, operational research, and management.

Particularly, in scientific quality control and research policy one often combines citation numbers in order to assess or just rank scientists, institutes, etc. Among the most notable and popular citation indices we have the Hirsch's h-index, which continues to be a subject of intensive and interesting debate since its introduction in 2005. Of course, the usage of the h-index is not solely limited to this particular domain of interest [cf. 6].

In this paper we deal with a highly important problem of comparing h-index values of two equal-length inputs and determining whether they differ significantly. We propose and analyze a statistical hypothesis test that may give us more insight into the very nature of the h-index.

Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland, gagolews@ibspan.waw.pl · Faculty of Mathematics and Information Science, Warsaw University of Technology, pl. Politechniki 1, 00-661 Warsaw, Poland

2 The h-index and its distribution

Let us first recall the definition of Hirsch's h-index [8].

Definition 1. Let $n \in \mathbb{N}$. The ***h*-index** is a function $\mathsf{H} : \mathbb{R}_{0+}^n \to \{0, 1, \dots, n\}$ such that

$$\mathsf{H}(\mathbf{x}) = \begin{cases} \max\{h = 1, \dots, n : x_{(n-h+1)} \geq h\} & \text{if } x_{(n)} \geq 1, \\ 0 & \text{otherwise,} \end{cases} \qquad (1)$$

where $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}_{0+}^n$ and $x_{(i)}$ denotes the ith order statistic, i.e. the ith smallest value in $\mathbf{x}$.

Interestingly, the h-index is a symmetric maxitive aggregation operator [cf. 6]. It is because (1) may be equivalently written as $\mathsf{H}(\mathbf{x}) = \bigvee_{i=1}^n \lfloor x_{(n-i+1)} \rfloor \wedge i$. Therefore, if $\mathbf{x} \in \mathbb{N}_0^n$ then H reduces itself to an ordered weighted maximum (OWMax) operator [2, 7], which in turn is equivalent to Sugeno integral of $\mathbf{x}$ w.r.t. some fuzzy (nonadditive) measure; see [4] for the proof. Please note that basic statistical properties of OWMax operators have already been examined in [5]: it turns out that they are asymptotically normally distributed and they are strongly consistent estimators of a distribution's parameter of location.

The exact distribution of H is given by the following theorem.

Theorem 1. *Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sequence of i.i.d. random variables with a continuous c.d.f. F defined on $\mathbb{R}_{0+}$. Then the c.d.f. of $\mathsf{H}(\mathbf{X})$ for $x \in [0, n)$ is given by $D_n(x) = \mathcal{I}\left(F\left(\lfloor x+1 \rfloor^{-0}\right); n - \lfloor x \rfloor, \lfloor x \rfloor + 1\right)$, where $\mathcal{I}(p; a, b)$ is the regularized incomplete beta function.*

Proof. For $i = 1, 2, \dots, n$ the c.d.f. of the ith order statistic, $X_{(i)}$, is given by $F_{(i)}(x) = \mathrm{P}(X_{(i)} \leq x) = \mathcal{I}(F(x); i, n-i+1)$ [cf. 1]. Note that supp $\mathsf{H}(\mathbf{X}) \subseteq \{0, 1, \dots, n\}$. Hence, $D_n(x) = 1$ for $x \geq n$. By (1) we have:

$$\begin{array}{rclcl} \mathrm{P}(\mathsf{H}(\mathbf{X}) < 1) & = & \mathrm{P}(X_{(n)} < 1) & = & \mathcal{I}(F(1^{-0}); n, 1), \\ \mathrm{P}(\mathsf{H}(\mathbf{X}) < 2) & = & \mathrm{P}(X_{(n-1)} < 2) & = & \mathcal{I}(F(2^{-0}); n-1, 2), \\ & & \dots & & \dots \\ \mathrm{P}(\mathsf{H}(\mathbf{X}) < n) & = & \mathrm{P}(X_{(1)} < n) & = & \mathcal{I}(F(n^{-0}); 1, n), \quad \text{QED.} \end{array} \qquad \square$$

As a consequence, for all $h = 0, \dots, n-1$ it holds $D_n(h) = \mathrm{P}(Z \leq h)$, where $Z \sim \mathrm{Bin}(n, 1 - F(h + 1^{-0}))$. We see that the values of the c.d.f. and the p.m.f. of the h-index in most cases may only be determined numerically. For convenience, they have been implemented in the `CITAN` package [3] for `R`.

3 Test for the difference between two h-indices

Given two equal-length vectors of observations, one may be interested whether their Hirsch's indices differ significantly. More formally, let $\Theta = (0, n)$ be a parameter space that induces an identifiable statistical model $(\mathbb{R}_{0+}, \{\mathrm{P}_\theta : \theta \in \Theta\})^n$ in which $\mathbb{E}_\theta \mathsf{H} = \theta$ for all $\theta \in \Theta$, and $\mathrm{P}_\theta(\mathsf{H} = i)$ is a continuous function of θ for all i. Moreover, let $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d P_{θ_x} and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ i.i.d P_{θ_y}, where $\theta_x, \theta_y \in \Theta$. We would like to construct a statistical test φ which verifies at given significance level α the null hypothesis $H_0 : \theta_x = \theta_y$ against the alternative $H_1 : \theta_x \neq \theta_y$.

The most natural test statistic is of course $T(\mathbf{X}, \mathbf{Y}) = \mathsf{H}(\mathbf{Y}) - \mathsf{H}(\mathbf{X})$. Obviously, under H_0 the distribution of T is symmetric around 0. Unfortunately, it may not be independent of the values of unknown parameters $\theta_x = \theta_y$. We therefore expect that by setting an acceptance region with bounds determined by functions of only α and n (an approach traditionally used in mathematical statistics) we will not obtain test of satisfactory power in result.

Denote by $\mathcal{B}$ the set of all 0–1 symmetric square matrices $B = (b_{ij})$, $i, j \in \{0, \ldots, n\}$, such that (i) $b_{ii} = 0$ for all i, and (ii) $b_{ij} = 1 \Longrightarrow b_{i,j+1} = 1$ for $i < j < n$. Each $B \in \mathcal{B}$ generates a statistical hypothesis test

$$\varphi_B(\mathbf{X}, \mathbf{Y}) = b_{\mathsf{H}(\mathbf{X}), \mathsf{H}(\mathbf{Y})}. \tag{2}$$

Such test bases on the test statistic T and has acceptance regions that depend on the value of the h-index in one of the samples. E.g. if we observed $\mathsf{H}(\mathbf{x}) = i$ and $\mathsf{H}(\mathbf{y}) = j$ then $b_{ij} = 1$ would indicate that H_0 should be rejected.

Please note that there is a bijection between $\mathcal{B}$ and the set of integer-valued sequences $\{(v_0, \ldots, v_n) : (\forall i)\ 0 \leq v_i \leq n - i\}$, as we may set $v_i = \sum_{j=i+1}^{n} (1 - b_{ij})$ for $i = 0, \ldots, n$ and $b_{ij} = b_{ji} = \mathbf{I}(j - i > v_i)$ for $0 \leq i \leq j \leq n$. Therefore, the acceptance region of T is given by $[-v_{\mathsf{H}(\mathbf{X}) \wedge \mathsf{H}(\mathbf{Y})}; v_{\mathsf{H}(\mathbf{X}) \wedge \mathsf{H}(\mathbf{Y})}]$. Additionally, we have $|\mathcal{B}| = (n + 1)!$

The power function of a test φ_B, reflecting the probability of rejecting H_0 for given $\theta_x, \theta_y \in \Theta$, is given by

$$\pi_B(\theta_x, \theta_y) = \sum_{i=0}^{n} \sum_{j=0}^{n} b_{ij}\, \mathrm{P}_{\theta_x}(\mathsf{H} = i)\, \mathrm{P}_{\theta_y}(\mathsf{H} = j). \tag{3}$$

Let $\mathcal{B}_\alpha = \{B \in \mathcal{B} : \sup_{\theta \in \Theta} \pi_B(\theta, \theta) \leq \alpha\}$ denote the set of all matrices which generate tests at significance level α. Our main task may be formulated formally as an optimization problem. We would like to find the matrix $B^* \in \mathcal{B}_\alpha$ which minimizes expected probability of committing Type II error, i.e.

$$B^* := \operatorname*{arg\,min}_{B \in \mathcal{B}_\alpha} \mathbb{E}\, \mathcal{L}(B) = \operatorname*{arg\,min}_{B \in \mathcal{B}_\alpha} \iint_{\Theta^2} (1 - \pi_B(\theta_x, \theta_y))\, w(\theta_x, \theta_y)\, d\theta_x\, d\theta_y, \tag{4}$$

where w is a prior distribution. If prior w is uniform (assumed by default when we have no knowledge of or preference for the underlying distribution parameters) then it may be shown that it holds:

$$\mathbb{E}\,\mathcal{L}(B) = \sum_{i=0}^{n}\sum_{j=0}^{n}(1-b_{ij}) \int_{\Theta} \mathrm{P}_\theta(\mathsf{H}=i)\,d\theta \int_{\Theta} \mathrm{P}_\theta(\mathsf{H}=j)\,d\theta. \tag{5}$$

Note that if the uniformly most powerful (UMP) test (in this class of tests) $\varphi_{B^{**}}$ exists then $\varphi_{B^{**}} = \varphi_{B^*}$ for any w such that $\operatorname{supp} w = \Theta^2$. Unfortunately, as the whole search space is $O(n!)$, in practice we may only seek for an *approximate* solution of (4), B^+, which may be computed in a sensible amount of time.

Let us introduce the following strict partial ordering relation over $\mathcal{B}$. We write $B \prec B'$ iff $B \neq B'$, $(\forall i,j)$ $b'_{ij} = 0 \Longrightarrow b_{ij} = 0$, and $b_{ij} = 1 \Longrightarrow b'_{ij} = 1$. Intuitively, if $B \prec B'$ then B' may be obtained from B by substituting some "1"s for "0"s. In such case eq. (3) implies that $(\forall \theta_x, \theta_y \in \Theta)$ $\pi_B(\theta_x,\theta_y) \leq \pi_{B'}(\theta_x,\theta_y)$.

For brevity, we will also write $B \prec^1 B'$ iff $B \prec B'$ and $\sum_i \sum_{j \geq i} b_{ij} = \sum_i \sum_{j\geq i} b'_{ij} - 1$. We propose the following algorithm for obtaining an approximation of B^*.

1. Calculate upper bound matrix $B^{(0)}$: For given $i < j$ we set $b^{(0)}_{ij} = 0$ iff $\max_\theta \sum_{k=j}^{n} \mathrm{P}_\theta(\mathsf{H}=i)\mathrm{P}_\theta(\mathsf{H}=k) > \alpha/2$, as surely is such case rejection of H_0 would lead to violation of given significance level.
2. If $B^{(0)} \in \mathcal{B}_\alpha$ then return $B^* := B^{(0)}$ as result (it is easily seen that $B^{(0)}$ is UMP).
3. Otherwise we generate a sequence $B^{(0)} \succ^1 B^{(1)} \succ^1 \cdots \succ^1 B^{(k)}$ such that $B^{(k-1)} \notin \mathcal{B}_\alpha$ and $B^{(k)} \in \mathcal{B}_\alpha$ by applying:

```
for $k := 1, 2, \ldots$ do
    if $(\exists B \prec^1 B^{(k-1)} : B \in \mathcal{B}_\alpha)$
        $B^{(k)} := \operatorname*{arg\,min}_{B \in \mathcal{B}_\alpha, B \prec^1 B^{(k-1)}} \mathbb{E}\,\mathcal{L}(B)$;
        proceed to Step #4;
    else
        $B^{(k)} := \operatorname*{arg\,min}_{B \in \mathcal{B}, B \prec^1 B^{(k-1)}} \int_\Theta \pi_B(\theta,\theta)\,\mathbf{I}(\pi_B(\theta,\theta) > \alpha)\,d\theta$;
```

4. Improve $B^{(k)}$: Find $B^+ \succeq B^{(k)}$ such that $B^+ \in \mathcal{B}_\alpha$ and $(\forall B \succ B^+)$ $B \notin \mathcal{B}_\alpha$ by applying:

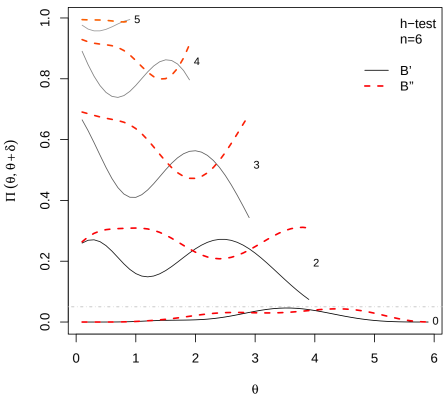
```
$B^+ := B^{(k)}$;
while $(\exists B \succ^1 B^+ : B \in \mathcal{B}_\alpha)$ do
    $B^+ := \operatorname*{arg\,min}_{B \in \mathcal{B}_\alpha, B \succ^1 B^+} \mathbb{E}\,\mathcal{L}(B)$;
return $B^+$ as result;
```

This procedure successively substitutes "1"s for "0"s in the initial upper bound matrix $B^{(0)}$ at positions which result in the greatest overall reduction of "oversized" power, down to the desired value α. This greedy approach — although quite fast to compute (we approximate the integrals by probing the power function at sufficiently many points in Θ) — does not of course guarantee convergence to optimal solution. However, the numerical results presented in the next section suggest that, at least in the considered cases, the solutions are close to optimal in terms of loss. The problem of finding accurate approximation of $\mathbb{E}\,\mathcal{L}(B^*)$ is left for further research.

4 Numerical results

We say that a random variable X follows a Pareto distribution with shape parameter $k > 0$, denoted $X \sim \mathrm{Par}(k)$, if its cumulative distribution function is given by $F(x) = 1 - 1/(1+x)^k$ for $x \geq 0$. Although F is continuous, it is quite often used by bibliometricians to model citation distribution (or different non integer-valued paper quality metrics). Note we have $\mathsf{H}(\mathbf{X}) = \mathsf{H}(\lfloor \mathbf{X} \rfloor)$. For any n, we apply a reparametrization of the shape parameter and set $\theta_n(k) := \mathbb{E}_k \mathsf{H}(X_1, \ldots, X_n)$ (it is a decreasing bijection). It may be shown that in result we obtain a statistical model that fulfills the assumptions stated in Sec. 3.

From now on, let us fix $\alpha = 0.05$. For $n \leq 5$ it holds $B^{(0)} \in \mathcal{B}_\alpha$, therefore $\varphi_{B^{(0)}}$ is uniformly most powerful in this class of tests ($\mathbb{E}\,\mathcal{L}(B^{(0)}) = 0.677$). However, e.g. for $n = 6$ we have $B^{(0)} \notin \mathcal{B}_\alpha$. In this case there are two maximal tests $\varphi_{B'}$ and $\varphi_{B''}$ (the latter is outputted by the above algorithm) in the sense that it holds $\neg(B' \prec B'')$, $\neg(B'' \prec B')$, $(B \succ B') \vee (B \succ B'') \Rightarrow B \notin \mathcal{B}_\alpha$, and $B \in \mathcal{B}_\alpha \Rightarrow (B \preceq B') \vee (B \preceq B'')$. As a consequence, the UMP test in this class does not exist (cf. Fig. 1). We have $\mathbb{E}\,\mathcal{L}(B') = 0.691$ and $\mathbb{E}\,\mathcal{L}(B'') = 0.647$. Obviously, if we assume no prior knowledge of θ then $\varphi_{B''}$ is the preferred choice for practical purposes.



i	0	1	2	3	4	5	6
v'_i	2	**3**	2	**3**	2	1	0

i	0	1	2	3	4	5	6
v''_i	2	2	**3**	2	2	1	0

i	0	1	2	3	4	5	6
$v^{(0)}_i$	2	2	2	2	2	1	0

Fig. 1 Power functions of two optimal h-tests $\varphi_{B'}, \varphi_{B''}$ for $n = 6$ and $\alpha = 0.05$; shift value δ is printed on the right of each curve.

Table 1 Computed acceptance region bounds; $n = 25$, $\alpha = 0.05$.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
v_i^+	1	2	2	3	3	4	5	4	5	6	5	5	6	5	6	5	5	5	5	4	4	3	3	2	1	0

Table 2 Computed acceptance region bounds; $n = 50$, $\alpha = 0.05$. Values improved in Step #4 of the algorithm are marked in bold.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
v_i^+	1	2	2	3	3	4	4	**4**	5	5	6	6	6	**6**	7	7	7	7	**7**	8	8	8	**7**	8	8	8
i		26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
v_i^+		9	8	8	8	8	8	7	8	7	8	7	**6**	7	7	6	6	5	5	5	4	3	3	2	1	0

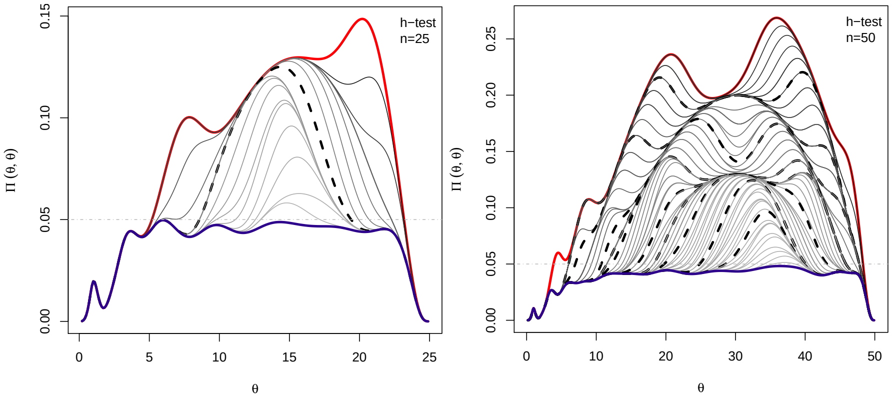


Fig. 2 Probabilities of committing Type I error at consecutive iterations of Step #3 of the algorithm; every 10th curve is dashed and marked in bold.

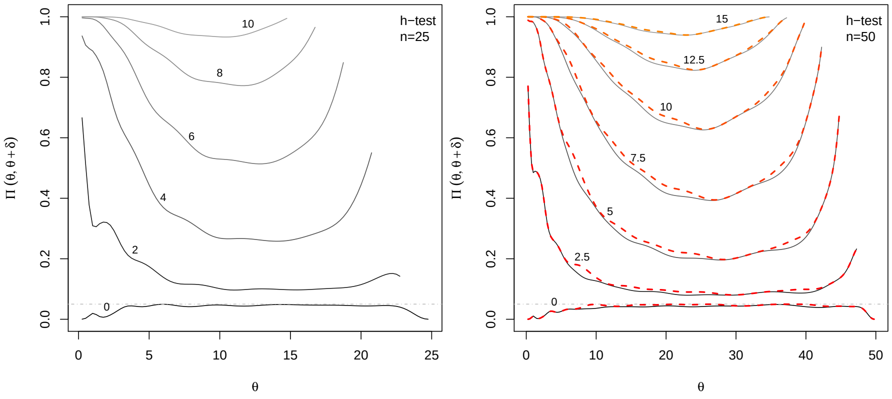


Fig. 3 Power functions of $\varphi_{B(20)} = \varphi_{B^+}$ ($n = 25$, left plot), and $\varphi_{B(110)} \prec \varphi_{B^+}$ ($n = 50$, right plot); shift value δ is printed above each curve.

We will study more deeply the two following cases. For $n = 25$ we get $k = 20$ and $\mathbb{E}\,\mathcal{L}(B^{(k)}) = \mathbb{E}\,\mathcal{L}(B^{+}) = 0.336$ (see Tab. 1 for the resulting acceptance region bounds), On the other hand, for $n = 50$ we have $k = 110$, $\mathbb{E}\,\mathcal{L}(B^{(k)}) = 0.251$, and $\mathbb{E}\,\mathcal{L}(B^{+}) = 0.247$ (cf. Tab. 2). Fig. 2 shows the plots of $\pi_{B^{(i)}}(\theta, \theta)$ for $i = 0, \ldots, k$ (cf. Step #3 of the algorithm). Additionally, in Fig. 3 we depict the plot of $\pi_{B^{(i)}}(\theta, \theta + \delta)$ and $\pi_{B^{(+)}}(\theta, \theta + \delta)$ for different values of δ. We see that the improvement of $B^{(110)}$ for $n = 50$ does not result in a drastic decrease of expected loss.

Let us compare the power of the computed h-tests with some other tests for equality of distribution parameters. The parametric F-test bases on a test statistic $T(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \log(1 + X_i) / \sum_{i=1}^{n} \log(1 + Y_i)$ which, under H_0, has Snedecor's F distribution with $(2n, 2n)$ degrees of freedom. We also consider 3 non-parametric tools: the Wilcoxon rank sum test, the discretized Wilcoxon test (computed on $\lfloor \mathbf{X} \rfloor$ and $\lfloor \mathbf{Y} \rfloor$), and the Kolmogorov-Smirnov test.

The plots of the examined tests' estimated power functions values, generated using $M = 25000$ Monte Carlo samples, are depicted in Fig. 4. The constructed h-tests are outperformed by the F-test and Wilcoxon's test, and often by the KS test. We also observe that their power is quite small for $\theta \simeq n$, which is due to the property $\mathsf{H}(\mathbf{X} \wedge n) = \mathsf{H}(\mathbf{X})$: here the h-index "ignores" some important information. What is more, we see that in the considered cases discretization of observations did not result in a significant reduction of power of the Wilcoxon test.

5 Conclusion

We should be very cautious while using the Hirsch index in decision making. For example, let us consider two authors A and B with 25 papers each, and whose h-indices are 12 and 16, respectively. Then — assuming that their citation counts follow Pareto distributions — at 0.05 significance level we cannot state that their output quality differs significantly, as $T = (16 - 12) \in [-v_{12}^{+}; v_{12}^{+}] = [-6; 6]$.

In future work we will consider the construction of h-tests in non-identifiable statistical models, and for samples of non necessarily equal lengths.

References

[1] David HA, Nagaraja HN (2003) Order statistics. Wiley
[2] Dubois D, Prade H, Testemale C (1988) Weighted fuzzy pattern matching. Fuzzy Sets and Systems 28:313–331
[3] Gagolewski M (2011) Bibliometric impact assessment with R and the CITAN package. Journal of Informetrics 5(4):678–692

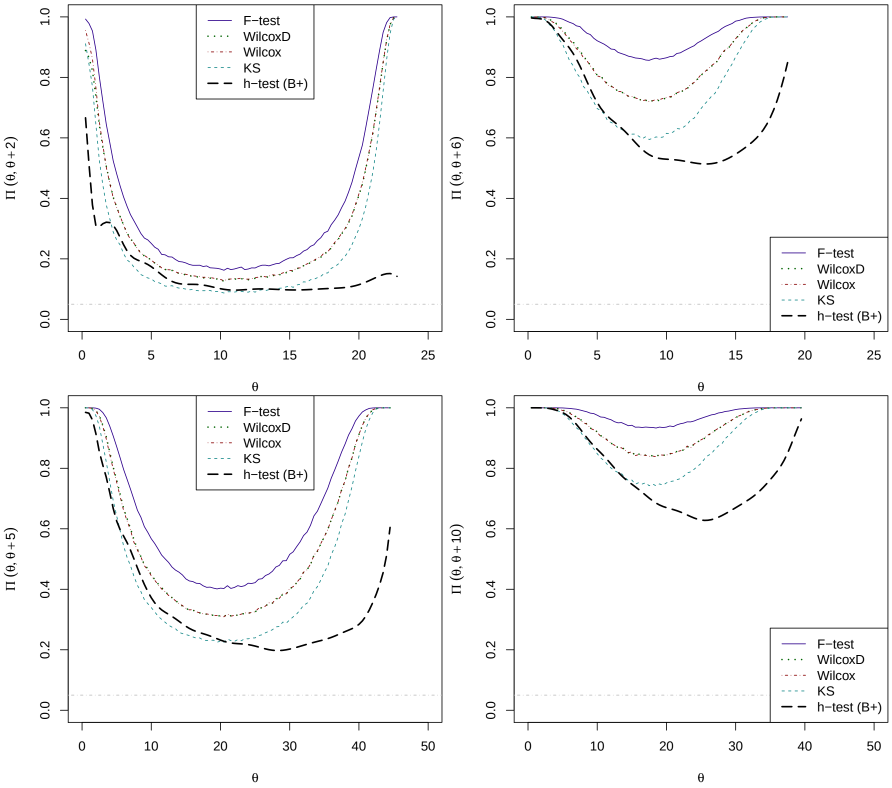


Fig. 4 Power functions of different two-sample tests; $n = 25$ (above) and $n = 50$ (below), $\alpha = 0.05$, $M = 25000$ MC iterations.

[4] Gagolewski M, Grzegorzewski P (2010) Arity-monotonic extended aggregation operators. In: Hüllermeier E, Kruse R, Hoffmann F (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems, CCIS 80, Springer-Verlag, pp 693–702

[5] Gagolewski M, Grzegorzewski P (2010) S-statistics and their basic properties. In: Borgelt C et al (ed) Combining Soft Computing and Statistical Methods in Data Analysis, Springer-Verlag, pp 281–288

[6] Gagolewski M, Grzegorzewski P (2011) Axiomatic characterizations of (quasi-) L-statistics and S-statistics and the Producer Assessment Problem. In: Galichet S, Montero J, Mauris G (eds) Proc. Eusflat/LFA 2011, pp 53–58

[7] Grabisch M, Pap E, Marichal JL, Mesiar R (2009) Aggregation functions. Cambridge

[8] Hirsch JE (2005) An index to quantify individual's scientific research output. Proceedings of the National Academy of Sciences 102(46):16569–16572

This is a revised version of the paper:
Gagolewski M., *Statistical Hypothesis Test for the Difference between Hirsch Indices of Two Pareto-Distributed Random Samples*, In: Kruse R. et al (Eds.), *Synergies of Soft Computing and Statistics for Intelligent Data Analysis* (AISC 190), Springer-Verlag, Heidelberg, 2013, pp. 359-367, doi:10.1007/978-3-642-33042-1_39