Economic and Financial Report 2006/03

Comparing Distributions: The Harmonic Mass Index: Extension to m Samples

Rien Wagenvoort*

JEL Classification codes: C12 (hypothesis testing)
C14 (semi-parametric and non-parametric methods)

Notes

*** Rien Wagenvoort** is Senior Economist, Economic and Financial Studies division of the European Investment Bank
The author would like to thank Luis Gonzales Pacheco, Jeroen Hinloopen and Charles van Marrewijk for constructive comments.

Comparing Distributions: The Harmonic Mass Index: Extension to m Samples

Abstract

We extend the paper of Hinloopen and van Marrewijk (2005), who introduce the harmonic mass index to test whether two samples come from the same distribution, in the following directions. Firstly, we derive the Harmonic Weighted Mass (HWM) index for any number of samples. Secondly, this paper shows how to compute the HWM index without making any assumptions on the number of "ties" (i.e. identical observations) within or between samples. Thirdly, we investigate ties with a Monte Carlo analysis, and find that the critical percentiles as reported in Hinloopen and van Marrewijk (2005), for two samples that are free of ties, are fairly accurate approximations of the HWM percentiles for two samples with ties when the sample size exceeds 50 observations. Furthermore, our results show that these percentiles are fairly accurate as well for cases where there are more than two samples.

1 Introduction

Recently, Hinloopen and van Marrewijk (2005) have introduced a new non-parametric test to infer whether two samples come from the same distribution, the so-called Harmonic Mass (HM) index test. Like the Anderson-Darling (AD) test and the Fisz-Cramér-von Mises (FCM) test, the HM index compares the Empirical Distribution Function (EDF) of the two samples over their entire domain (see, Anderson and Darling (1952), Fisz (1960), and von Mises (1931)). Other EDF tests only consider differences in the sample distributions at particular entries on their domain. For instance, the Kolmogorov-Smirnov test is based on the maximum deviation between the two cumulative distribution functions (see Kolmogorov (1933) and Smirnov (1939)). The Kuiper test is based on the maximum deviation above and below the cumulative distribution function (see, Kuiper (1960)).

The key advantageous properties of the HM index test are the following. Firstly, no parametric specification of the population distribution is required. Secondly, the HM index test is distribution free. Thirdly, in contrast with the critical values of AD and FCM that are based on asymptotic results, critical values of the HM index can be exactly derived for balanced samples of any size.

The HM index quantifies a so-called Percentile-Percentile plot (PP-plot) with a single number. A PP-plot is a graphical tool to analyse the differences between two samples. It shows the cumulative probability of the elements of one (let's say first) sample against the corresponding cumulative probability for the other sample (i.e. at the elements of the first sample). The HM index for two samples is defined as the surface between the PP-plot-line and the diagonal, scaled by the factor two. Hinloopen and van Marrewijk (2005) compute the HM index under the strong assumption that there are no ties, i.e. identical observations, within and between samples.

We extend the work of Hinloopen and van Marrewijk (2005) in the following directions. Firstly, we derive the Harmonic Weighted Mass (HWM) index for any number of samples.[1] Secondly, this paper shows how to compute the HWM index without making any assumptions on the number of ties within or between samples, and derives the probability density of the HWM index values for samples with "between ties" but no "within ties". Thirdly, we investigate ties with a Monte Carlo analysis.

Under the null-hypothesis that distributions are equal, the analytically derived critical percentiles (at a significance level of 10% and lower) are lower for the samples with between ties (but no within ties) than for samples free of ties. The results of our Monte Carlo analysis show that the simulated percentiles of the HWM index for samples with within ties (but no between ties) are higher than for samples free of ties. We find that the critical percentiles as reported in Hinloopen and van Marrewijk (2005), for two samples that are free of ties, are fairly accurate approximations of the HWM percentiles for two samples with ties when the sample size exceeds 50 observations.

This paper is organised as follows. The HWM index is defined and computed in Section 2. Its critical percentiles are analytically derived in Section 3. Section 4 shows the results of our Monte Carlo study whereas Section 5 concludes. The proof of Proposition 1 is put in the Annex.

[1] Scholz and Stephens (1987) extend the Anderson-Darling one-sample test to K samples.

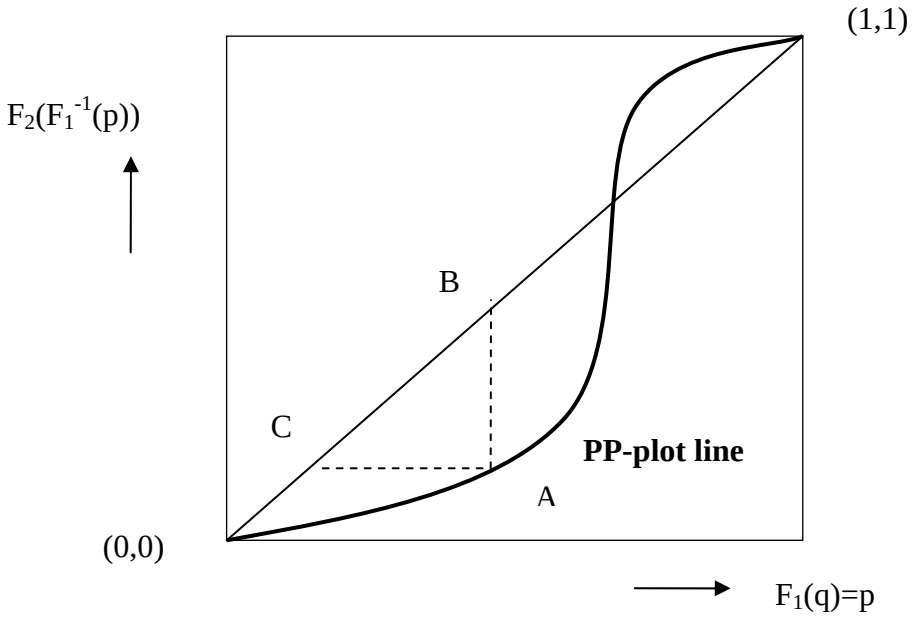
2 Definition and computation of the Harmonic Weighted Mass index

2.1 Notation and definition

Let $F_i(q)$ be the (population) cumulative distribution function of a 1-dimensional random variable q associated with sample i, and m be the number of samples. The objective of this paper is to test whether the m samples are drawn from the same distribution (i.e. $F_1 = F_2 = ... = F_m$).

We begin with the definition of the HWM index for two samples (*Case 1:* $m = 2$) as in Hinloopen and van Marrewijk (2005). Then, we extend the HWM index to more than two samples (*Case 2:* $m > 2$).

Figure 1: An example of a PP-plot



Case 1: $m = 2$

The harmonic mass index of Hinloopen and van Marrewijk (2005) quantifies a Percentile-Percentile plot with a single number. Figure 1 above shows an example of such a PP-plot. The axes of this plot are defined as follows. The cumulative probability associated with the first random variable ($F_1 \leq p$) is shown on the horizontal axis whereas the corresponding cumulative probability for the second random variable ($F_2(q)$ for $q = F_1^{-1}(p)$) is shown on the vertical axis. When $F_1(p) = F_2(p)\ \forall p$ then the PP-plot line coincides with the diagonal, which is a straight line from (0,0) to (1,1). The deviation of the PP-plot line from the diagonal at probability mass p shows to which extent at that point the distribution F_1 differs from distribution F_2. The harmonic mass index for two samples is defined as the surface between the PP-plot-line and the diagonal, scaled by the factor two. The scale factor ensures that the index takes values on the domain [0,1]. The index is equal to one when there is no overlap in the domain of F_1 and F_2 whereas the index is zero when F_1 and F_2 are identical.

Let $F_2(F_1^{-1}(p))$ denote $F_2(q)$ for q equal to the inverse of F_1 at probability p. The Harmonic Weighted Mass index is then defined as:

$$HWM(F_1, F_2) = 2\int_0^1 \sqrt{(p - F_2(F_1^{-1}(p)))^2}\, dp. \qquad (1)$$

In equation (1), F_1 is the distribution function of the base sample on the horizontal axis of Figure 1. The base sample is always on the diagonal whereas the other sample is below, above, or coincides with the diagonal. The choice of the base sample is irrelevant when there are two samples. To see this, let $p_1 = F_1$ and $p_2 = F_2$ at point A on the PP-plot line in Figure 1; p_2 is smaller than p_1. Note that the distance from A to B, i.e. $\sqrt{(p_1 - F_2(F_1^{-1}(p_1)))^2}$, is equal to the distance from C to A, i.e. $\sqrt{(p_2 - F_1(F_2^{-1}(p_2)))^2}$. Hence, the base sample choice is irrelevant. This regularity for $m = 2$ implies that one can choose a different base sample at each point p.

Let $M = \{1,...,m\} \subset N^+$ and $L(p)$ be a sample indicator function, $L(p) : [0,1] \to M$. $L(p)$ indicates the sample with minimum cumulative probability:

$$L(p) = \min\{F_i(F_j^{-1}(p))\} \qquad \forall i, j \in M \qquad (2)$$

Equation (2) first selects sample j that has the lowest value q for probability p. Then, $L(p)$ returns sample i that has the lowest cumulative probability given q. Note that i can be equal to j. In case $L(p)$ is not unique, it assumes the lowest indicator value. It means that sample 1 is indicated when the PP-plot line is above, or coincides with, the diagonal whereas sample 2 is indicated when the PP-plot line is below the diagonal. Furthermore, let $M^-(p)$ refer, for any $p \in [0,1]$, to all indicator values in M not identified by $L(p)$.

Then, equation (1) can be re-written as follows:

$$HWM(F_1, F_2) = 2\int_0^1 \sqrt{\sum_{j \in M^{-1}(p), i = L(p)} (p - F_j(F_i^{-1}(p)))^2}\, dp, \qquad (3)$$

Equation (3) gives the same index value as equation (1). Basically, equation (3) is based on a PP-plot line that is mirrored to one side of the diagonal.

Case 2: $m > 2$

Let $\{p, \sum_{j \in M^{-1}(p), i = L(p)} F_j(F_i^{-1}(p))\}$ be the coordinates of a "multi-dimensional" PP-plot line in m-dimensional space. The Harmonic Weighted Mass (HWM) index for $m > 2$ is defined as the surface between this multi-dimensional PP-plot line and the line which cuts all two-dimensional spaces in exact halves, scaled by the normalisation factor of $2/\sqrt{m-1}$:

$$HWM(F_1, F_2, ..., F_m) = \frac{2}{\sqrt{m-1}} \int_0^1 \sqrt{\sum_{j \in M^{-1}(p), i = L(p)} (p - F_j(F_i^{-1}(p)))^2}\, dp\,, \qquad (4)$$

The HWM index has the following properties:

P1 (equality): $HWM = 0 \leftrightarrow F_1 = F_2 = ... = F_m$

P2 (range): $HWM \in [0,1]$.

Having defined the HWM index at population level, we now turn to the computation of its empirical counterpart.

2.2 Computation of the empirical HWM index

For actual samples, the HWM index is computed by replacing the population distribution functions in (1) by their corresponding empirical distribution functions. Hinloopen and van Marrewijk (2005), based on Mushkudiani (2000), show that the empirical index is consistent, i.e. converges to the population index, and goes to zero when the samples are drawn from the same distribution.

We first show how to compute the HWM index when there are two samples ($m = 2$).

Case 1: $m = 2$

Let X_1 and X_2 be a vector that contain n_1 and n_2 drawings of two (one-dimensional) random variables respectively.

Let $Z = \{X_1, X_2\}$ be the ordered vector of n_1 observations of X_1 and n_2 observations of X_2.

Let $T_1(Z) = \{T_1(Z_1), .., T_1(Z_{n_1+n_2})\}$ be the vector of horizontal coordinates of the PP-plot.

Let $T_2(Z) = \{T_2(Z_1), .., T_2(Z_{n_1+n_2})\}$ be the vector of vertical coordinates of the PP-plot.

Note that going from the origin of the PP-plot with coordinates (0,0) to coordinates (1,1) means going from (0,0) to $(T_1(Z_1), T_2(Z_1))$, from $(T_1(Z_1), T_2(Z_1))$ to $(T_1(Z_2), T_2(Z_2))$, etc.

$T_1(Z_i)$ is equal to the number of observations in X_1 that are smaller or equal to Z_i divided by n_1; $T_2(Z_i)$ is equal to the number of observations in X_2 that are smaller or equal to Z_i divided by n_2.

The PP-plot line cuts the diagonal "from below" when $T_2(Z_i) < T_1(Z_i)$ and $T_2(Z_{i+1}) \geq T_1(Z_{i+1})$. The PP-plot line cuts the diagonal vertically from below when at the same time $T_1(Z_i) = T_1(Z_{i+1})$. The PP-plot line cuts the diagonal "from above" when $T_1(Z_i) < T_2(Z_i)$ and $T_1(Z_{i+1}) \geq T_2(Z_{i+1})$. The PP-plot line cuts the diagonal horizontally from above when at the same time $T_2(Z_i) = T_2(Z_{i+1})$.

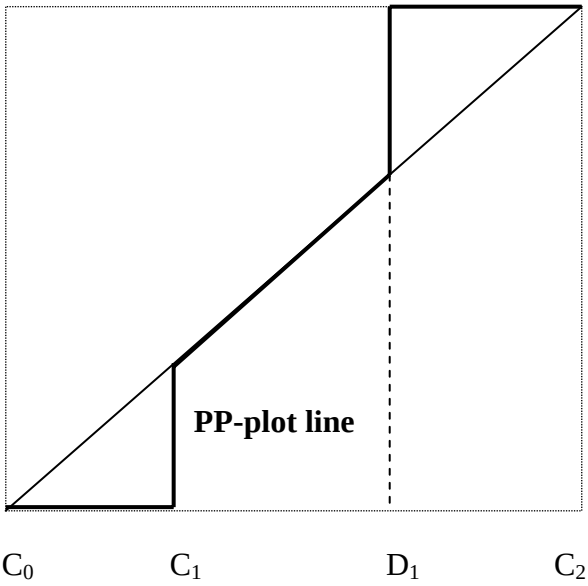
Note that cutting does not necessarily imply that the PP-plot line crosses the diagonal in the sense that at i it is below (above) the diagonal whereas at $i+1$ it is above (below) the diagonal.

Define C_j as the point where the PP-plot line cuts the diagonal either from below or above. We set C_0 at the coordinate associated with the first PP-plot line departure of the diagonal and C_J at the coordinate associated with the last PP-plot line retrieval of the diagonal. For example, when $T_1(Z_1) \neq T_2(Z_1)$ and $T_1(Z_{n_1+n_2-1}) \neq T_2(Z_{n_1+n_2-1})$ then $C_0 = 0$ and $C_J = 1$.

Define $D_{h(j)}$ as the point where the PP-plot line departs from the diagonal after coinciding with the diagonal between C_{j-1} and $D_{h(j)}$, not counting the first departure as we called this C_0. We assume that there are H of such new PP-plot line departures.

Let C^* be the ordered set of the first PP- plot line departure C_0, the cutting points C_j, and the new PP-plot line departures $D_{h(j)}$ and contain $J+1$ elements. For example, in Figure 2 below there are four elements in the set $C^*: C_0 = 0$ (first departure), C_1 (cutting from below), D_1 (new departure) and $C_2 = 1$ (last retrieval).

Figure 2: PP-plot line departures and retrievals of the diagonal



Define $j(i)$ as the next cutting or departure point when standing at point Z_i, and $I_i = 1$ when between i and $i+1$ there is a (additional) cut or departure of the diagonal, $I_i = 0$ otherwise. Furthermore, let $T_1(Z_0) = T_2(Z_0) = C_0^* = C_0$.

Proposition 1

For $m = 2$**:**

$$HWM = \sum_{j=1}^{J} (C_j^* - C_{j-1}^*)^2 - \sum_{h=1}^{H} (D_{h(j)} - C_{j-1}^*)^2 - \sum_{i=0}^{n_1+n_2-1} E_i \qquad (5)$$

where

$C_j^* = C_j = T_{1i} + \dfrac{(T_{1,i+1} - T_{1,i})(T_{1i} - T_{2i})}{(T_{2,i+1} - T_{2,i}) - (T_{1,i+1} - T_{1,i})}$ if at point j the diagonal is cut from below,

$C_j^* = C_j = T_{2i} + \dfrac{(T_{2,i+1} - T_{2,i})(T_{2i} - T_{1i})}{(T_{1,i+1} - T_{1,i}) - (T_{2,i+1} - T_{2,i})}$ if at point j the diagonal is cut from above,

$C_j^* = D_{h(j)}$ if point j is a new PP-plot line departure.

$$\begin{aligned} E_i = {} & 2(1 - I_i)(T_2(Z_{i+1}) - T_2(Z_i))(C_{j(i)} - T_1(Z_{i+1})) \\ & + (1 - I_i)((T_2(Z_{i+1}) - T_2(Z_i))(T_1(Z_{i+1}) - T_1(Z_i)) \\ & + I_i(C_{j(i)} - T_1(Z_i))(C_{j(i)} - T_2(Z_i)) \\ & + I_i(T_1(Z_{i+1}) - C_{j(i)})(T_2(Z_{i+1}) - C_{j(i)}) \\ & + 2I_i(C_{j(i+1)} - T_2(Z_{i+1}))(T_1(Z_{i+1}) - C_{j(i)}), \end{aligned}$$

when at the next cutting point $j(i)$ the diagonal is cut from below,

$$\begin{aligned} E_i = {} & 2(1 - I_i)(T_1(Z_{i+1}) - T_1(Z_i))(C_{j(i)} - T_2(Z_{i+1})) \\ & + (1 - I_i)((T_1(Z_{i+1}) - T_1(Z_i))(T_2(Z_{i+1}) - T_2(Z_i)) \\ & + I_i(C_{j(i)} - T_2(Z_i))(C_{j(i)} - T_1(Z_i)) \\ & + I_i(T_2(Z_{i+1}) - C_{j(i)})(T_1(Z_{i+1}) - C_{j(i)}) \\ & + 2I_i(C_{j(i+1)} - T_1(Z_{i+1}))(T_2(Z_{i+1}) - C_{j(i)}), \end{aligned}$$

when at the next cutting point $j(i)$ the diagonal is cut from above,

$E_i = 0$ when between i and $i+1$ the PP-plot line coincides with the diagonal (i.e. $T_1(Z_i) = T_2(Z_i)$ and $T_1(Z_{i+1}) = T_2(Z_{i+1})$).

Proof:
See Annex A.

Case 2: $m > 2$

A $m > 2$ comparison can be transformed into a classical PP-plot for $m = 2$.

Let $Z = \{X_1, X_2, ..., X_m\}$ be the ordered vector of n_1 observations of X_1, n_2 observations of X_2, n_3 observations of X_3, etc. Furthermore, let $P_j(Z_i)$ be equal to the number of observations in X_j that are smaller or equal to Z_i divided by n_j, $j = 1,...,m$.

The horizontal coordinate $T_1(Z_i)$ is equal to $P_1(Z_i)$ if $P_1(Z_i) \le P_2(Z_i)$, and $P_1(Z_i) \le P_3(Z_i)$, and, …, $P_1(Z_i) \le P_m(Z_i)$. $T_1(Z_i)$ is equal to $P_2(Z_i)$ if $P_2(Z_i) < P_1(Z_i)$ and $P_2(Z_i) \le P_3(Z_i)$, and, …, $P_2(Z_i) \le P_m(Z_i)$. Etc.

The vertical coordinates $T_2(Z_i)$ are computed as

$$T_2(Z_i) = T_1(Z_i) + \frac{\sqrt{(T_1(Z_i) - P_1(Z_i))^2 + ... + (T_1(Z_i) - P_m(Z_i))^2}}{\sqrt{m-1}},\ i = 1,...,n_1 + n_2 + ... + n_m.$$

Subsequently, the formulas of Proposition 1 can be applied.

3 Hypothesis testing with the HWM index

Hinloopen and van Marrewijk (2005) derive analytically the densities of the harmonic mass index under the null-hypothesis ($H_0 : F_1 = F_2$) and the following assumptions:

A1: $n_1 = n_2 = n$, $m = 2$

A2: There are no ties possible between and within samples.

Under conditions A1 and A2, the PP-plot line only consists of pieces of vertical and horizontal lines.

It is possible however to derive analytically the densities of the HWM index for another particular case, namely when ties between samples occur with probability 1/3. In this case, the PP-plot line goes possibly with an angle of 0, 45 and 90 degrees. When, in addition, within ties are allowed, then the PP-plot line can have any angle. Unfortunately, in that case, it is not possible to derive HWM densities analytically.

Consider the following modification of assumption A2:

A2*: There are no ties within samples.
Ties between samples occur with probability 1/3.

In this section we show the analytical derivation of HWM densities under both assumptions A1-A2 and assumptions A1-A2* because our proof is slightly different from the one of

Hinloopen and van Marrewijk (2005). In all other cases, one needs to resort to Monte Carlo simulation experiments, which is done in the next section.

Lemma 1

Under assumptions A1-A2, the number of possible distinct values of the HWM index is equal to $\Theta(n) = 1 + n(n-1)/2$.

*Under assumptions A1-A2**, the number of possible distinct values of the HWM index is equal to $\Theta(n) = 1 + n^2$.

Proof:

Let a "step" on a PP-plot line be an increase of $1/n$ on either the horizontal axis and/or the vertical axis.

Under assumptions A1-A2, the smallest value of the HWM index is obtained when after each step in the horizontal (vertical) direction the following step is made in the vertical (horizontal) direction. In that case, the value of the HWM index is equal to the (scaled) sum of n triangles where the surface of each triangle is $1/2n^2$. Hence, the smallest value of the HWM index is $2*(n/2n^2) = 1/n$. The largest value of the HWM index is obtained when n steps are made in the same direction. In that case, the HWM index is 1. The smallest possible difference in two HWM index values is $2/n^2$. Hence, the number of distinct values is $1 + \frac{1-1/n}{2/n^2} = 1 + n(n-1)/2$.

*Under assumptions A1-A2**, the HWM index is zero when the PP-plot line coincides with the diagonal, i.e. each step is a between tie. The smallest possible difference in two HWM index values is $1/n^2$. Hence, the number of distinct HWM values is $1 + \frac{1-0}{1/n^2} = 1 + n^2$.

Q.E.D.

Lemma 2

Under assumptions A1-A2, the vector $HWM_j(n) = 1 - 2(j-1)/n^2$, $j = 1,..,1 + n(n-1)/2$ contains all possible distinct HWM index values.

*Under assumptions A1-A2**, the vector $HWM_j(n) = 1 - (j-1)/n^2$, $j = 1,..,1 + n^2$ contains all possible distinct HWM index values.

Proof:

Under assumptions A1-A2, the largest HWM index value is obtained for $j = 1$ ($HWM_{j=1}(n) = 1$) whereas the smallest value is obtained for $j = 1 + n(n-1)/2$ ($HWM(n)_{j=1+n(n-1)/2} = 1/n$). $HWM(n)_j$ decreases with each step of j with $2/n^2$, which is equal to the smallest possible difference in two HWM index values.

*Under assumptions A1-A2**, the largest value is obtained for $j = 1$ ($HWM_{j=1}(n) = 1$) whereas the smallest HWM index value is obtained for $j = 1 + n^2$ ($HWM_{j=1+n^2}(n) = 0$). $HWM(n)_j$ decreases with each step of j with $1/n^2$, which is equal to the smallest possible difference in two HWM index values.

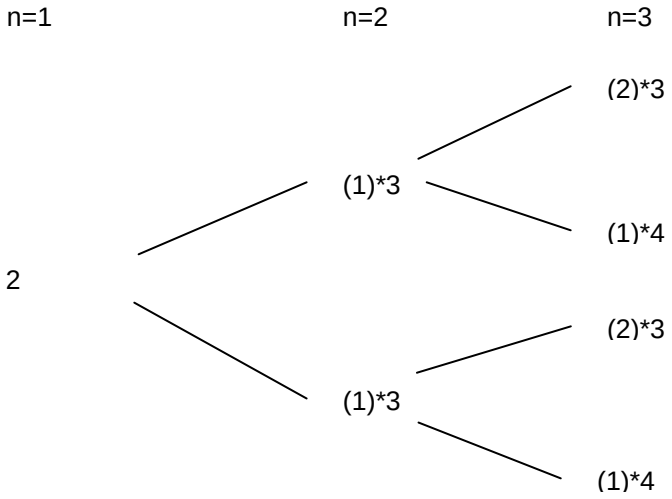
Q.E.D.

Definitions

• Let XX represent a possible route for $n = 2$ where two steps are made on the horizontal axis of the PP-plot. XXYX represents a possible route for $n = 3$ where first two steps are made on the horizontal axis of the PP-plot, then one step on the vertical axis, and then one step on the horizontal axis. Etc.

• Let a "split number" at a conjunction indicate the number of possible distinct routes until the next conjunction. Let the "split history" associated with a conjunction indicate the number of routes that were continued in the direction of the respective conjunction.

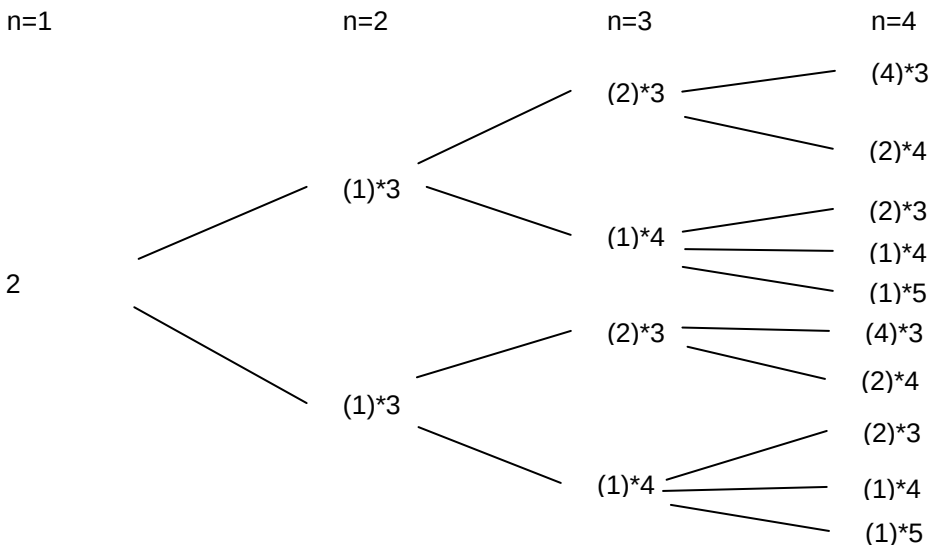
For example, under assumptions A1-A2, until $n = 2$, there are three possible routes, i.e. XX, XYX and XYY, which start in the X-direction and three possible routes, i.e. YY, YXY, and YXX, which start in the Y-direction. Among the three routes with the first step in the X-direction, one (i.e. XX) can be continued in four directions[2] whereas two (i.e. XYX and XYY) can be continued in three directions when the number of observations increases to $n = 3$. This is illustrated in the following tree:



In the tree above, at $n = 2$, 3 indicates the "split number" (i.e. at $n = 1$, after one step in the X or Y-direction, there are three possible continuations) whereas the number within brackets is the "split history" (i.e. at $n = 1$ one can either go in the X or Y-direction). At $n = 3$, there are two split numbers, i.e. 3 and 4, and the split history is (2) and (1) respectively.

[2] That is, XXX, XXYX, XXYYX, and XXYYY.

For $n = 4$ the tree can be continued as follows:



Notice that each split number is continued in the same "split proportions" and with the same split numbers wherever it is located. For example, both at $n = 2$ and $n = 3$ split number 3 is continued with split numbers 3 and 4 in the proportions 2:1. Split number 4 is continued with split proportions 2:1:1. Indeed, any split number k $(k \geq 3)$ has split proportions $2:1:1:..:1$ where there are $k-2$ ones in the sequence. Split proportions, however, are different under assumption A2* as will be demonstrated below.

• Let the matrix $B(n)$ contain the "split proportions" of split numbers 3 and higher. The rows and columns of $B(n)$ are defined according to decreasing split number. The last row of $B(n)$ contains the split proportions of the lowest split number. The first row of $B(n)$ contains the split proportions of the highest split number.

Under assumptions A1-A2, $B(n) \equiv (\Delta_n | 2\mathrm{Y}_n)$ for $n > 1$ where Y_n is a n-dimensional column unit vector and Δ_n is the n-dimensional upper triangular unit matrix. $B(1) = [1 \quad 2]$. $B(n)$ has n rows and $n+1$ columns.

For example, $B(4)$ is equal to the non-shaded matrix in the following table:

	Ending split number				
Starting split number	7	6	5	4	3
6	1	1	1	1	2
5	0	1	1	1	2
4	0	0	1	1	2
3	0	0	0	1	2

Under assumptions A1-A2*, $B(n)$ only consists of uneven split numbers. From split number 7 and higher, split proportions follow the following sequence:

	Ending split number						
Starting split number	15	13	11	9	7	5	3
13	1	2	2	2	2	3	1
11	0	1	2	2	2	3	1
9	0	0	1	2	2	3	1
7	0	0	0	1	2	3	1

Hence, the split proportions are (when standing at a starting split number): one time split number 3, three times split number 5, two times the other split numbers except for the highest split number which occurs one time. Furthermore, $B(1) = \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & 1 \end{bmatrix}$. $B(n)$ has $n+1$ rows and $n+2$ columns.

For example, $B(3)$ is equal to the non-shaded matrix in the following table:

	Ending split number				
Starting split number	11	9	7	5	3
9	1	2	2	3	1
7	0	1	2	3	1
5	0	0	1	3	1
3	0	0	0	2	1

- $A(n+1) \equiv M(n)B(n)$,

Under assumptions A1-A2,

$m_{jk}(n) = a_{(j-k+1)k}(n)$ for $j = k,..,\Theta(n-1)+k-1$, $k = 1,..,n$

$m_{jk}(n) = 0$ otherwise

$M(1) \equiv 2$, $A(1) = 2$.

*Under assumptions A1-A2**,

$m_{jk}(n) = a_{(j-k+1)k}(n)$ for $j = k,..,r+k-1$, $k = 1,..,n+1$, where r is equal to the number of rows plus the number of columns of $A(n)$ minus 1,

$m_{jk}(n) = 0$ otherwise.

$M(1) \equiv \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $A(1) = 3$.

- Let the "split sum" be the sum of the split numbers on a continuous branch (i.e. without counting the split numbers of sister branches in other directions), from start ($n = 1$) to end ($n = n+1$), of the trees developed above.

For example, *under assumptions A1-A2*, for $n=3$ (i.e. sum until split numbers at $n=4$ included) the highest branch of the tree consists of 2,3,3,3 with split sum equal to 11 (see tree above). The second highest branch of this tree consists of 2,3,4,3 with split sum equal to 12. Etc.

Under assumptions A1-A2, each continuous branch with the same split sum has the same HWM index value. Furthermore, the HWM index increases with the split sum. Hence, the highest HWM index value (=1) coincides with the highest split sum (=14, for $n=3$).

The rows of matrix M are defined according to descending split sum (i.e. the first row corresponds to the highest HWM index value whereas row $1+n(n-1)/2$ corresponds to the smallest HWM index value). The columns of matrix M are defined according to descending *ending* (i.e. at $n+1$) split number (i.e. the first column corresponds to the highest ending split number whereas column n corresponds to the lowest ending split number). M (n) contains the split history shown at $n+1$ in the tree and has $\Theta(n)$ rows and n columns. For example, $M(3)$ is equal to the non-shaded matrix in the following table:

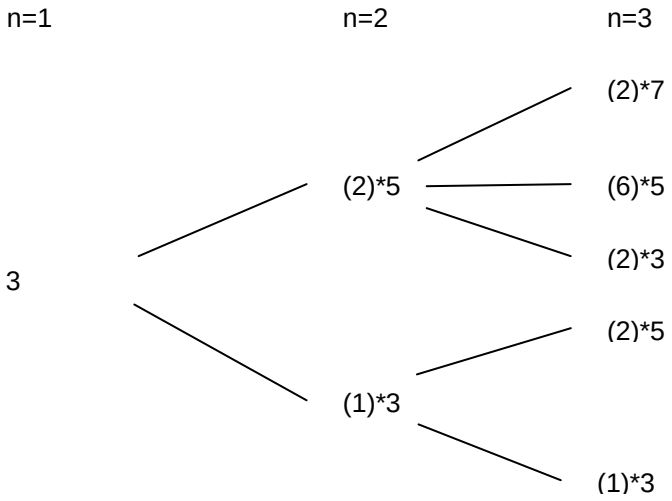
	Ending split number		
Split sum	5	4	3
14	2	0	0
13	0	2	0
12	0	4	4
11	0	0	8

• *Under assumptions A1-A2**, split numbers can be refined into "split refinement numbers" according to the following sequence:

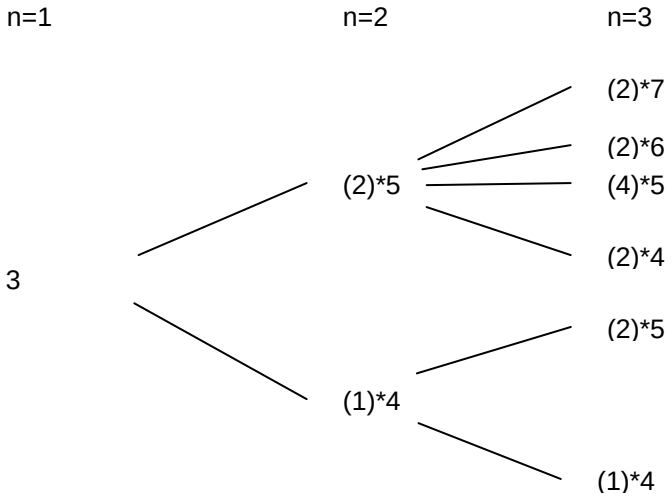
	Ending split refinement number							
Starting split refinement number	4	5	6	7	8	9	10	11
3	1	2						
4	1	2						
5	1	2	1	1				
6	1	2	1	1				
7	1	2	1	1	1	1		
8	1	2	1	1	1	1		
9	1	2	1	1	1	1	1	1

Hence, the split refinement proportions of even split refinement numbers are equal to the split refinement proportions of the respective closest but lower uneven split number. Split refinement proportions of uneven split numbers are equal to one for all split refinement numbers from 4 until the respective uneven split number plus 2 except split refinement number 5 which occurs twice.

For example, under *assumptions A1-A2**, until $n=3$, the tree of split numbers is as follows



whereas the tree of split refinement numbers is:



• Let the matrix $BR(n)$ contain the split refinement proportions of split refinement numbers 4 and higher. The rows and columns of $BR(n)$ are defined according to decreasing split refinement number. $BR(n)$ has $2n$ rows and $2n+2$ columns.

For example, $BR(2)$ is equal to the non-shaded matrix in the following table:

	Ending split refinement number					
Starting split refinement number	9	8	7	6	5	4
7	1	1	1	1	2	1
6	0	0	1	1	2	1
5	0	0	1	1	2	1
4	0	0	0	0	2	1

Furthermore, $BR(1) = \begin{bmatrix} 1 & 1 & 2 & 1 \\ 0 & 0 & 2 & 1 \end{bmatrix}$.

• Let the "split refinement sum" be the sum of the split refinement numbers from start ($n=1$) to end ($n=n+1$) on a continuous branch. Under A1-A2*, the HWM index increases with the split refinement sum. Let MR (n) contain the split refinement history shown at $n+1$ in the corresponding tree. The $\Theta(n)$ rows of matrix MR are defined according to decreasing split refinement sum whereas the columns of matrix MR are defined according to decreasing *ending* (i.e. at $n+1$) split refinement number and has $\Theta(n)$ rows and $2n$ columns. For example, $MR(2)$ is equal to the non-shaded matrix in the following table:

	Split refinement number			
Split refinement sum	7	6	5	4
15	2	0	0	0
14	0	2	0	0
13	0	0	4	0
12	0	0	2	2
11	0	0	0	1

• $AR(n+1) \equiv MR(n)BR(n)$,

*Under assumptions A1-A2**,

$$mr_{jk}(n) = ar_{(j-k+1)k}(n) \text{ for } j = k,..,\Theta(n-1)+k-1,\ k = 1,..,2n$$
$$mr_{jk}(n) = 0 \qquad \text{otherwise}$$

$$MR(1) = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix},\ AR(1) = 3.$$

Lemma 3

The total number of possible distinct PP-plot lines, $\Omega(n)$, is equal to the sum of all elements in $A(n)$, i.e. $\Omega(n) = \sum_i \sum_j A_{ij}(n)$.[3]

Proof:
Each element of A contains the product of the number of routes at a previous conjunction times the number of possible continuations. Summing over all conjunctions gives the total number of possible distinct PP-plot lines.

Q.E.D.

The number of possible distinct PP-plot lines increases with a factor 4 (approximately 6) under assumptions A1-A2 (A1-A2*). For example, $\Omega(n) = 9.05e+58$ under A1-A2 whereas

[3] In Hinloopen and van Marrewijk (2005), $\Omega(n)$ indicates half the number of possible distinct sample plots. Furthermore, note that under assumptions A1-A2*, $\Omega(n) = \sum_i \sum_j A_{ij}(n) = \sum_i \sum_j AR_{ij}(n)$.

$\Omega(n) = 2.05e + 75$ under A1-A2*, for $n = 100$. Hence, by allowing ties between samples, the HWM index values are substantially refined.

Proposition 2

Under $H_0 : F_1 = F_2$ and *assumptions A1-A2*, $HWM_j(n)$ has probability $p_j(n)$ where $p_j(n) = \frac{M_j(n)\mathrm{Y}_n}{\Omega_n}$, $j = 1,..,1 + n(n-1)/2$. $M_j(n)$ is row j of matrix M.

Under $H_0 : F_1 = F_2$ and *assumptions A1-A2**, $HWM_j(n)$ has probability $p_j(n)$ where $p_j(n) = \frac{MR_j(n)\mathrm{Y}_n}{\Omega_n}$, $j = 1,..,1 + n^2$. $MR_j(n)$ is row j of matrix MR.

Y_n is a n-dimensional column unit vector.

Proof:
$M_j(n)\mathrm{Y}_n$ and $MR_j(n)\mathrm{Y}_n$ are equal to the sum of all possible PP-plots with the same HWM index value under *assumptions A1-A2* and *assumptions A1-A2** respectively. The relative frequency of this HWM index value is equal to $M_j(n)\mathrm{Y}_n / \Omega(n)$ and $MR_j(n)\mathrm{Y}_n / \Omega(n)$ respectively.

Q.E.D.

Hypothesis testing with the HWM index works as follows. Samples with a high empirical HWM index value have a low probability that they come from the same distribution. The significance level of the HWM index test thus corresponds to the corresponding percentage of highest HWM values. The associated HWM percentiles, derived under assumptions A1-A2, are shown in Hinloopen and van Marrewijk (2005).

Table 1 below summarizes the results for the HWM percentiles, derived under both assumptions A1-A2 and A1-A2*, for $n = 5$, 50,100,150,200, and 250. It turns out that the HWM percentiles at the usual confidence levels (i.e. at 10%, 5%, 2.5% and 1%) are lower under A1-A2* than under A1-A2. Therefore, the HWM index test accepts H_0 more often than justified when A1-A2 percentiles are used instead of A1-A2* percentiles and ties can be present between samples. As a consequence, under H_0, the size distortion of the test under A1-A2 is always smaller than the chosen significance level.

However, under the alternative hypothesis that samples do not come from the same distribution, the power of the HWM test based on A1-A2 could be low for samples with less than 50 observations when there are ties between samples. For example, the difference between the HWM percentiles (A1:A2 – A1:A2*) is equal to 0.12 at $n = 5$ for all reported significance levels. That said, for samples with 50 observations or more, the differences in the HMW percentiles (first column to the right of Table 1) are relatively small and decreasing with the sample size (n). This suggests that the A1-A2 HWM percentiles are

fairly accurate approximations of the true percentiles when the sample size exceeds 50, even when as much as one-third of the samples consist of between ties.

But, can the same conclusion also be drawn for samples with ties within samples?

Table 1: Critical percentiles of the HWM index under $H_0 : F_1 = F_2$ [1]

N	HWM percentile under A1:A2	HWM percentile under A1:A2*	Difference (A1:A2 – A1:A2*)
		10% significance level	
5	0.680	0.560	0.120
50	0.200	0.168	0.032
100	0.141	0.119	0.023
150	0.115	0.097	0.018
200	0.100	0.084	0.016
250	0.089	0.075	0.014
		5% significance level	
5	0.760	0.640	0.120
50	0.234	0.196	0.037
100	0.165	0.139	0.026
150	0.134	0.113	0.021
200	0.117	0.098	0.019
250	0.104	0.088	0.017
		2.5% significance level	
5	0.840	0.720	0.120
50	0.264	0.222	0.042
100	0.186	0.157	0.030
150	0.152	0.128	0.024
200	0.132	0.111	0.021
250	0.118	0.099	0.019
		1% significance level	
5	0.920	0.800	0.120
50	0.300	0.254	0.046
100	0.212	0.179	0.033
150	0.174	0.146	0.027
200	0.150	0.127	0.024
250	0.134	0.113	0.021

[1] Assumption A1: $n_1 = n_2$, $m = 2$.

Assumption A2: There are no ties within and between samples.

Assumption A2*: There are no ties within samples. Ties between samples occur with probability 1/3.

4 Monte Carlo simulation experiments

To investigate within ties we resort to a Monte Carlo simulation experiment because analytical solutions for the HWM percentiles are not available. This is not necessarily a drawback since the HWM index is distribution free and, as a result, its critical values are independent of the distributional choices.

Let the population of X_1 and X_2 consists of integers, which are uniformly distributed over the range $[0, n*c]$. Here, n is the sample size of both X_1 and X_2, and the factor c determines the population size. We consider all values of c in the set [20,10,5,2] and, therefore, draw randomly 5,10,20 and 50 percent, respectively, of the population size. Note however that samples X_1 and X_2 do not necessarily consist of distinct numbers, and, therefore, do not necessarily include $n/n*c$ percent of the distinct elements of the population. Evidently, the probability of finding ties within and between samples is the highest when the sample size is half the population size, i.e. $c = 2$.

Table 2: Average percentage of ties in the Monte Carlo experiments (as a share of n)

	Only within ties	Between and within ties
	Sample size is 5% of population size	
5	2.1	4.5
50	2.5	4.8
100	2.5	4.8
150	2.5	4.8
200	2.5	4.8
250	2.5	4.8
	Sample size is 10% of population size	
5	4.3	8.5
50	5.0	9.3
100	5.0	9.3
150	5.1	9.3
200	5.0	9.3
250	5.1	9.4
	Sample size is 20% of population size	
5	8.5	16.1
50	10.1	17.5
100	10.2	17.5
150	10.2	17.6
200	10.3	17.5
250	10.3	17.6
	Sample size is 50% of population size	
5	22.4	34.2
50	26.1	36.5
100	26.3	36.6
150	26.4	36.7
200	26.5	36.7
250	26.5	36.7

Table 3: Analytical and simulated critical percentiles of the HWM index under $H_0 : F_1 = F_2$ at the 5% confidence level for two samples[1]

Sample size (n)	*Analytical* HWM percentile under A1:A2 (no ties)	*Analytical* HWM percentile under A1:A2* (only between ties)	*Simulated* HWM percentile under A1 (only within ties)	*Simulated* HWM percentile under A1 (between and within ties)	Difference in HWM percentiles (no ties – between and within ties)
Sample size is 5% of population size					
5	0.760	0.640	0.760	0.760	0.000
50	0.234	0.196	0.239	0.228	0.006
100	0.165	0.139	0.168	0.163	0.002
150	0.134	0.113	0.137	0.134	0.001
200	0.117	0.098	0.119	0.114	0.003
250	0.104	0.088	0.106	0.104	0.000
Sample size is 10% of population size					
5	0.760	0.640	0.760	0.720	0.040
50	0.234	0.196	0.246	0.230	0.004
100	0.165	0.139	0.175	0.164	0.001
150	0.134	0.113	0.140	0.136	-0.001
200	0.117	0.098	0.123	0.115	0.001
250	0.104	0.088	0.109	0.102	0.002
Sample size is 20% of population size					
5	0.760	0.640	0.840	0.720	0.040
50	0.234	0.196	0.257	0.233	0.000
100	0.165	0.139	0.180	0.161	0.003
150	0.134	0.113	0.146	0.132	0.002
200	0.117	0.098	0.130	0.114	0.002
250	0.104	0.088	0.115	0.104	0.000
Sample size is 50% of population size					
5	0.760	0.640	0.920	0.720	0.040
50	0.234	0.196	0.294	0.231	0.003
100	0.165	0.139	0.209	0.162	0.003
150	0.134	0.113	0.169	0.135	0.000
200	0.117	0.098	0.147	0.114	0.002
250	0.104	0.088	0.133	0.102	0.002

[1] Assumption A1: $n_1 = n_2, m = 2$.

Assumption A2: There are no ties within and between samples.
Assumption A2*: There are no ties within samples. Ties between samples occur with probability 1/3.

Based on the population described above, we do two experiments, in one experiment we allow only within ties whereas in the other experiment we allow both between and within ties. The cumulative distribution function of HWM is simulated for $n = 5$, 50,100,150,200 and 250, where for each sample size we make 10,000 runs.

Table 2 shows the average percentage of within and between ties as a share of the sample size over the 10,000 runs. The average share of ties varies from about 2.5% for a sample size equal to 5% of the population size (the highest panel in the table above of the experiment

with only within ties) to about 37% for a sample size equal to 50% of the population size (the lowest panel of the experiment with between and within ties). Hence, our experiments capture a wide range of tie proportions that is likely to cover most of the actual datasets in terms of tie representation.

Table 3 contains the simulated HWM percentiles and is organised as follows. The sample size is shown in the first column to the left. The analytically derived HWM percentile at the 5% confidence level, under A1-A2 (no ties) and A1-A2* (only between ties), are found in the second and third column respectively. The fourth and fifth column shows the corresponding percentile of the simulated HWM distribution for the experiment with only within ties and for the experiment with between and within ties respectively. The difference between the analytically derived percentile for assumptions A1-A2 (no ties), and the simulated percentile for samples with between and within ties is computed in the sixth column.

We recall from the previous section that between ties lead to lower HWM percentiles at the 5% significance level in comparison to samples that are free from ties. Table 3 demonstrates that within ties lead to higher HWM percentiles. It is striking to observe that the analytically derived percentiles under assumptions A1-A2 (the second column to the left) are almost identical to the simulated HWM percentiles (the fifth column to the left) when there are between and within ties. The difference between the respective percentiles (column six) is smaller than 0.05 for all sample sizes, and smaller than 0.01 for samples of 50 observations or more.

We ran a third simulation experiment to investigate the critical HWM percentiles for more than two samples. Here the population of X_1 and X_2 consists of numbers which are uniformly distributed over the range $[0,1]$. Therefore, in this case, the probability of drawing a tie is zero.

The highest horizontal panel of Table 4 shows the simulated critical percentiles for two up to five samples whereas the lowest horizontal panel contains the corresponding differences with the analytically derived percentiles under assumptions A1 and A2, i.e. $m = 2$. Again, the absolute differences in percentiles are relatively small and decreasing with the sample size but too important to be explained by random error only when $m > 2$. In all cases except for samples with 5 observations, the critical percentiles of the HWM index for more than two samples are slightly higher than the critical percentiles for two samples. The A1-A2 HWM percentiles are again fairly accurate approximations of the true percentiles when the sample size exceeds 50, even when there are more than two samples.

As a final remark, our simulation experiments only consider samples of equal size (assumption A1). Effective solutions to the comparison of samples of different sizes are given in Hinloopen and van Marrewijk (2005).

Table 4: Simulated critical percentiles of the HWM index under $H_0 : F_1 = F_2$ at the 5% confidence level for multiple ($m \geq 2$) samples

Sample size (n)	$m = 2$	$m = 3$	$m = 4$	$m = 5$
		Simulated HWM percentile		
5	0.760	0.737	0.732	0.732
50	0.236	0.239	0.245	0.247
100	0.164	0.171	0.175	0.176
150	0.134	0.138	0.144	0.143
200	0.117	0.118	0.123	0.124
250	0.106	0.107	0.110	0.110
	Difference with analytical percentile under assumptions A1 ($m = 2$) and A2[1]			
5	0.000	-0.023	-0.028	-0.028
50	0.002	0.006	0.012	0.014
100	-0.001	0.006	0.010	0.011
150	-0.001	0.003	0.009	0.008
200	0.000	0.002	0.007	0.008
250	0.002	0.003	0.006	0.005

[1] Assumption A1: $n_1 = n_2, m = 2$.

Assumption A2: There are no ties within and between samples.

5 Conclusions

We have shown how to compute the Harmonic Weighted Mass index for any number of samples. Under the null hypothesis, its percentiles are analytically derived for two samples with between ties but no within ties. The results of our simulation experiments reveal that the percentiles as reported in Hinloopen and van Marrewijk (2005), for two samples that are free of ties, are fairly accurate approximations of the HWM percentiles for two samples that contain between and within ties when the sample size exceeds 50. Furthermore, our results show that these percentiles are fairly accurate as well for cases where there are more than two samples.

It goes without saying that the HWM index test can be used in numerous applications: goodness of fit analysis, treatment effect analysis, event-study analysis, regime-switching analysis, frontier analysis, performance analysis, inequality analysis etc.

References

[1] Anderson, T.W., and Darling, D.A. (1952), "Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes", *Annals of Mathematical Statistics*, 23, 193-212.

[2] Fisz, M. (1960), "On a result by M. Rosenblatt concerning the von Mises-Smirnov test", *Annals of Mathematical Statistics*, 31, 427-429.

[3] Hinloopen, J. and C. van Marrewijk (2005), "Comparing Distributions: The Harmonic Mass Index", *Tinbergen Institute discussion paper,* TI 2005-122/1, The Netherlands.

[4] Kolmogorov, A.N. (1933), "Sulla determinazione empirica delle leggi di probabilita", Giorn. Ist. Ital. Attuari, 4, pp. 1-11.

[5] Kuiper, N.H. (1960), "Tests concerning random points on a circle", *Proc. Koninkl. Neder. Akad. Van Wetenschappen*.

[6] Mushkudiani, N. (2000), *Statistical applications of generalized quantiles: nonparametric tolerance regions and P-P plots*, Ph.D. dissertation, Eindhoven University of Technology.

[7] Scholz, F.W. and Stephens, M.A. (1987), "K-sample Anderson-Darling tests", *Journal of the American Statistical Association*, 82, 399, 918-924.

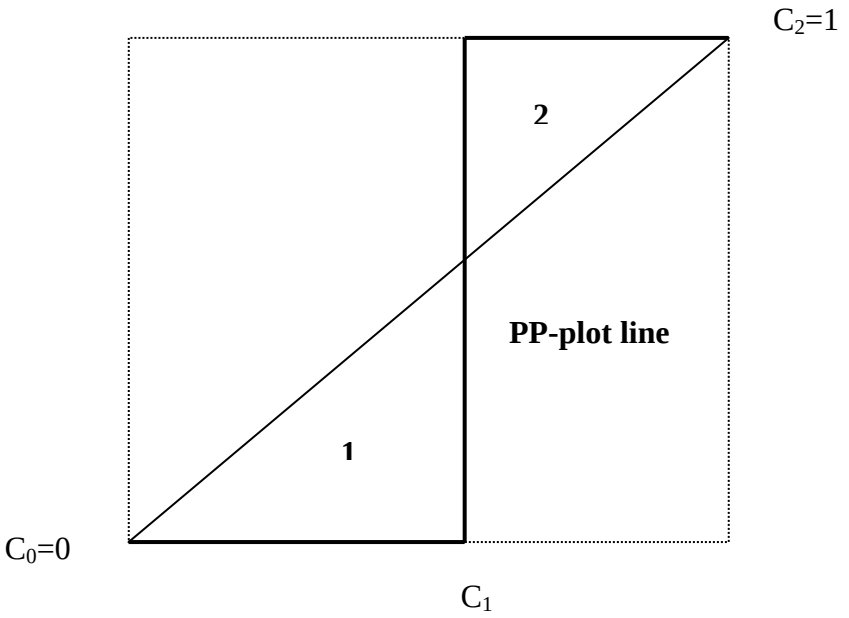
[8] Smirnov, N.V. (1939), "On the deviation of the empirical distribution function", *Rec. Math. (Mat. Sbornik) (NS)*, 6, pp. 3-26.

[9] Von Mises, R. (1931), *Wahrscheinlichkeitsrechnung*, Deuticke, Vienna.

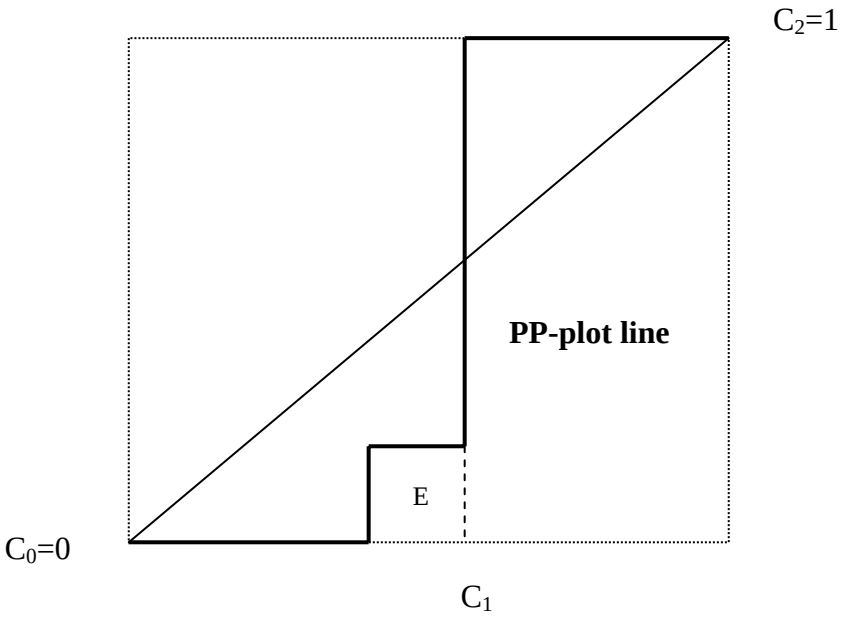
Annex

Proof of Proposition 1

When the PP-plot line is under the diagonal, the horizontal line from C_{j-1} to C_j (i.e. from $T_1 = C_{j-1}$ to $T_1 = C_j$ and $T_2 = C_{j-1}$), the vertical line from C_{j-1} to C_j (i.e. from $T_2 = C_{j-1}$ to $T_2 = C_j$ and $T_1 = C_j$), and part of the diagonal form a right triangle. Triangle 1 in the figure below is an example of such a triangle. Triangle 2 in this figure is an example of a triangle following from a PP-plot line that is above the diagonal.



E_i is defined as a "hole" in a triangle, which occurs, for example, when T_2 changes before C_j is reached while the PP-plot line is below the diagonal. This is illustrated in the figure below:



Step 1: Computing the surface of the triangles

The surface of the triangle from cutting C_{j-1} to C_j is equal to $(C_j - C_{j-1})^2/2$. The contribution of this triangle (when there are no holes) to the HWM index is $(C_j - C_{j-1})^2$.

However, the surface of triangles defined by the elements of C^* must not be counted when the PP-plot line coincides with the diagonal between C^*_{j-1} and C^*_j, that is, when $C^*_j = D_h(j)$ is a new PP-plot line departure. This explains why $\sum_{h=1}^{H}(D_h(j) - C^*_{j-1})^2$ is subtracted from $\sum_{j=1}^{J}(C^*_j - C^*_{j-1})^2$.

$C_j = T_1(Z_i)$ when there is a vertical cutting j between Z_i and Z_{i+1} whereas $C_j = T_2(Z_i)$ when there is a horizontal cutting j between Z_i and Z_{i+1}. Notice that cuttings are either vertical or horizontal when there are no ties in the data because, in that case, the PP-plot only consists of vertical and horizontal lines.

When there are ties, the PP-plot may consist of lines that are neither vertical nor horizontal. Such a situation occurs from i to $i+1$ (going from Z_i to Z_{i+1}) if both coordinate $T_1(Z_i)$ is different from coordinate $T_1(Z_{i+1})$ and coordinate $T_2(Z_i)$ is different from coordinate $T_2(Z_{i+1})$.

The figure below illustrates this possibility for a "cutting from below". The PP-plot goes with a straight line from coordinates (T_{1i}, T_{2i}) to $(T_{1,i+1}, T_{2,i+1})$. The diagonal goes from a straight line from coordinates (D_i, D_i) to (D_{i+1}, D_{i+1}). Note that, if cutting occurs from i to $i+1$, $C_j = D_i + g$ where g is the vertical distance between the crossing point (where the diagonal cuts the PP-plot) and the horizontal line at D_i.

In the figure below:

$a = D_i - T_{2i} = T_{1i} - T_{2i}$,

$b = D_{i+1} - T_{2i} = T_{1,i+1} - T_{2i}$,

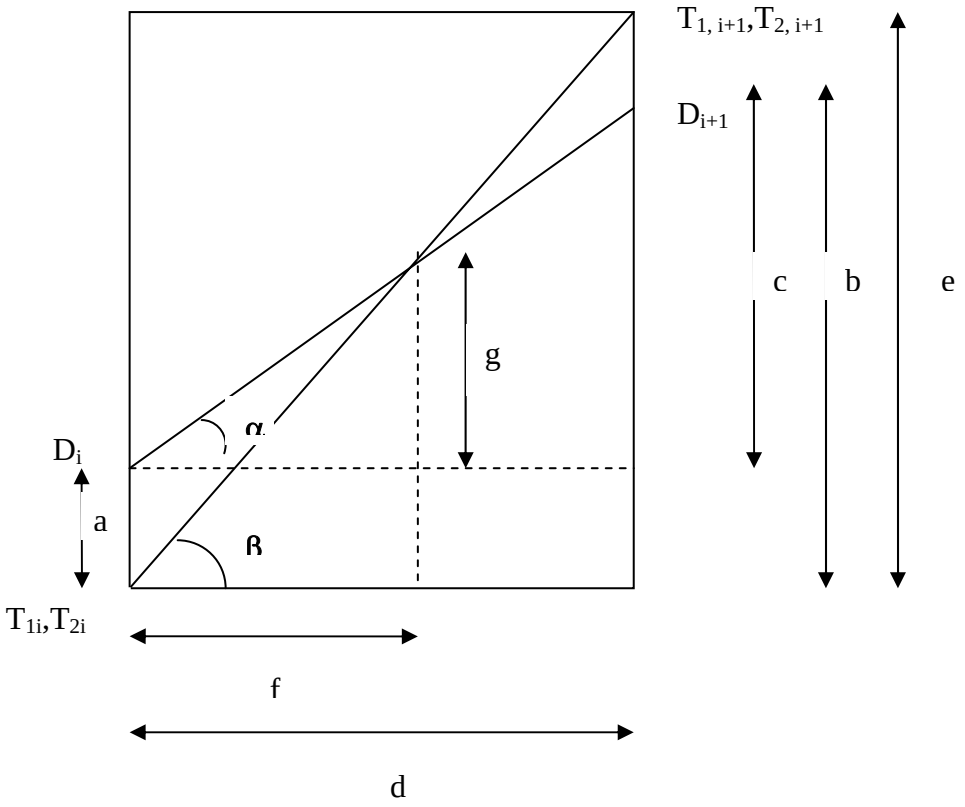
$c = b - a = T_{1,i+1} - T_{1,i}$,

$d = T_{1,i+1} - T_{1,i} = c$,

$e = T_{2,i+1} - T_{2,i}$,

$\tan(\alpha) = \frac{c}{d} = 1,\ \tan(\alpha) = \frac{g}{f} = 1$,

$\tan(\beta) = \frac{e}{d},\ \tan(\beta) = \frac{g+a}{f}$.



Using the fact that $\frac{c}{d} = \frac{g}{f}$ and $\frac{e}{d} = \frac{g+a}{f}$ one can derive that

$$g = \frac{ca}{e-c} = \frac{(T_{1,i+1} - T_{1,i})(T_{1i} - T_{2i})}{(T_{2,i+1} - T_{2,i}) - (T_{1,i+1} - T_{1,i})}$$

Furthermore, $D_i = T_{1i}$, hence

$$C_j = T_{1i} + \frac{(T_{1,i+1} - T_{1,i})(T_{1i} - T_{2i})}{(T_{2,i+1} - T_{2,i}) - (T_{1,i+1} - T_{1,i})}$$, when at j the diagonal is cut from below.

Applying the same calculation method for a cutting from above we find that

$$C_j = T_{2i} + \frac{(T_{2,i+1} - T_{2,i})(T_{2i} - T_{1i})}{(T_{1,i+1} - T_{1,i}) - (T_{2,i+1} - T_{2,i})}$$, when at j the diagonal is cut from above.

Step 2: Computing the holes in the triangles

We only discuss possible "holes" in triangles (as computed under step 1) that are below the diagonal. The same procedure can be applied for holes in triangles that are above the diagonal. A "hole" is defined as the (additional) area of a triangle that is taken away by a movement in the T_2 coordinate. Note that a movement in the T_2 coordinate does not necessarily lead to a hole.

There are three possible types of "holes" created in the surface of the triangles computed under step 1 when going from i to $i+1$:

(i) There is no cutting of the diagonal, there is a change in T_2 coordinate but there is no change in the T_1 coordinate. In this case, the hole is a rectangle with surface $(T_2(Z_{i+1}) - T_2(Z_i))(C_{j(i)} - T_1(Z_i))$ where $C_{j(i)}$ is the next cutting point. Taking into account the scaling factor of 2, the (negative) contribution of this rectangle to the HWM index is then:

$$2(1 - I_i)(T_2(Z_{i+1}) - T_2(Z_i))(C_{j(i)} - T_1(Z_i)), \qquad \text{(A1)}$$

where $I_i = 0$.

(ii) There is no cutting of the diagonal, and both coordinate T_1 and T_2 change. In this case, the hole in the triangle consists of the surface of a triangle and, possibly, of a rectangle. The contribution to the HWM index is:

$$\begin{aligned} &2(1 - I_i)(T_2(Z_{i+1}) - T_2(Z_i))(C_{j(i)} - T_1(Z_{i+1})) + \\ &(1 - I_i)(T_2(Z_{i+1}) - T_2(Z_i))(T_1(Z_{i+1}) - T_1(Z_i)) \end{aligned} \qquad \text{(A2)}$$

where $I_i = 0$.

The rectangle disappears when $C_{j(i)} = T_1(Z_{i+1})$. Note that when there is no change in the T_1 coordinate (i.e. $T_1(Z_{i+1}) = T_1(Z_i)$), formula (A2) boils down to formula (A1). Hence, only formula (A2) needs to be retained.

(iii) There is a cutting of the diagonal from i to $i+1$ and both coordinate T_1 and T_2 change. In this case, there can be holes in two triangles. There is a possible hole in the triangle below the diagonal; this hole is itself a triangle. There is a possible hole in the triangle above the diagonal; this hole is a triangle and a rectangle. The contribution of the surface of these holes to the HWM index is:

$$\begin{aligned} &I_i(C_{j(i)} - T_1(Z_i))(C_{j(i)} - T_2(Z_i)) + I_i(T_1(Z_{i+1}) - C_{j(i)})(T_2(Z_{i+1}) - C_{j(i)}) \\ &+ \; 2I_i(C_{j(i+1)} - T_2(Z_{i+1}))(T_1(Z_{i+1}) - C_{j(i)}) \end{aligned} \qquad \text{(A3)}$$

where $I_i = 1$.

Combining formulas (A2) and (A3) gives:

$$
\begin{aligned}
E_i = {} & 2(1-I_i)(T_2(Z_{i+1})-T_2(Z_i))(C_{j(i)}-T_1(Z_{i+1})) \\
& + (1-I_i)((T_2(Z_{i+1})-T_2(Z_i))(T_1(Z_{i+1})-T_1(Z_i)) \\
& + I_i(C_{j(i)}-T_1(Z_i))(C_{j(i)}-T_2(Z_i)) \\
& + I_i(T_1(Z_{i+1})-C_{j(i)})(T_2(Z_{i+1})-C_{j(i)}) \\
& + 2I_i(C_{j(i+1)}-T_2(Z_{i+1}))(T_1(Z_{i+1})-C_{j(i)}).
\end{aligned}
\qquad \text{(A4)}
$$

Notice that $E_i = 0$ when $T_1(Z_{i+1}) = T_1(Z_i)$ and $T_2(Z_{i+1}) = T_2(Z_i)$, i.e. Z_i is a "within" tie that is located before another within tie Z_{i+1}. However, E_i is not necessarily equal to zero when Z_i is a within tie but Z_{i+1} is not a within tie.

Finally, we set $E_i = 0$ when both Z_i and Z_{i+1} are on the diagonal (i.e. $T_1(Z_i) = T_2(Z_i)$ and $T_1(Z_{i+1}) = T_2(Z_{i+1})$).

Q.E.D.