TI 2010-083/1
Tinbergen Institute Discussion Paper

Identifying All Distinct Sample P-P Plots, with an Application to the Exact Finite Sample Distribution of the L_1-FCvM Test Statistic

Jeroen Hinloopen[1]
Rien Wagenvoort[2]

[1] *University of Amsterdam, and Tinbergen Institute;*
[2] *European Investment Bank, Luxemburg.*

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Identifying all distinct sample p-p plots, with an application to the exact finite sample distribution of the L_1-FCvM test statistic[*]

Jeroen Hinloopen[†]and Rien Wagenvoort[‡]

August 26, 2010

Abstract

P-p plots contain all the information that is needed for scale-invariant comparisons. Indeed, Empirical Distribution Function (EDF) tests translate sample p-p plots into a single number. In this paper we characterize the set of all distinct p-p plots for two balanced sample of size n absent ties. Distributions of EDF test statistics are embedded in this set. It is thus used to derive the exact finite sample distribution of the L_1-version of the Fisz-Cramér-von Mises test. Comparing this distribution with the (known) limiting distribution shows that the latter can always be used for hypothesis testing: although for finite samples the critical percentiles of the limiting distribution differ from the exact values, this will not lead to differences in the rejection of the underlying hypothesis.

Key words: Sample p-p plot, EDF test, finite sample distribution, limiting distribution.

JEL Classification: C12, C14, C46

[*] Thanks are due to Adriaan Soetevent, Jan Kiviet, Robert Waldmann, Sanne Zwart, and to seminar participants at the University of Amsterdam for useful comments and constructive suggestions. All computing codes (GAUSS) are available upon request. The usual disclaimer applies.

[†] Corresponding author: University of Amsterdam, Amsterdam School of Economics, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands; J.Hinloopen@uva.nl; www.fee.uva.nl/io/jhinloopen.

[‡] Correspondence: European Investment Bank, Economic and Financial Studies, 100 Boulevard Konrad Adenauer, Luxemburg, Luxemburg; wagenvoo@eib.org.

1 Introduction

Any two continuous distribution functions are conveniently compared graphically using the percentile-precentile (p-p) plot: the scatter plot of two distributions' percentiles (Wilk and Gnanadesikan, 1968). P-p plots have the desirable property that they contain all the information that is needed for scale-invariant comparisons (Holmgren, 1995). Little is known however of the statistical properties of sample p-p plots, which obtain in case two samples are compared. This is all the more surprising as any Empirical Distribution Function (EDF) test can be represented in a sample p-p plot. In this paper, therefore, we characterize the set of all distinct p-p plots for two balanced samples of size n absent ties.

P-p plots yield a straight 45-degree line when two identical distributions are compared. EDF tests use this property as they quantify in one way or another the distance between the p-p plot and the diagonal. For example, the Kolmogorov-Smirnov test considers the largest positive distance, the Kuiper test computes the sum of the maximum positive and negative distance, and the L_1 (L_2) version of the Fisz-Cramér-von Mises ($FCvM$) test sums up over all absolute (squared) distances (Stephens, 1974).[1] Applying an EDF test to small samples might be troublesome however because only the limiting distributions are known for any of the concomitant test statistics. How accurate these limiting distributions are for small samples is yet to be determined.[2,3]

[1] The area below the p-p plot corresponds to the Mann-Whitney-Wilcoxon U statistic (Bamber, 1975), which is used to test if one distribution first-order stochastically dominates another distribution.

[2] Sample sizes vary across disciplines. For instance, economics research that involves controlled laboratory experiments typically relies on a very limited number of independent observations.

[3] Monte Carlo and Bootstrapping exercises can retrieve approximations of finite sample distributions (see e.g. Henze (1996), Famoye (1999), and Olea and Pawlowsky-Glahn, 2009).

Using the characterization of all distinct sample p-p plots we retrieve the exact finite sample distribution of $FCvM_1$. That is, we order all distinct sample p-p plots according to the corresponding value of $FCvM_1$ and link these values to the relative frequency of occurrence of the underlying sample p-p plot. This also serves as an example as to how the exact finite sample distribution of other EDF test statistics could be retrieved.

We conclude by comparing the finite sample distribution of $FCvM_1$ with its (known) limiting distribution. It turns out that the latter can always be used for hypothesis testing: although for finite samples the critical percentiles of the limiting distribution differ from the exact values, this will not lead to differences in the rejection of the underlying hypothesis.

2 All distinct sample p-p plots

Consider the set of cumulative density functions Ξ_1. For $F_1, F_2 \in \Xi_1$ the p-p plot depicts for every domain value z from their joint support the percentiles of one distribution relative to the other:

$$z \longmapsto \begin{bmatrix} F_1(z) \\ F_2(z) \end{bmatrix}. \tag{1}$$

This is a plot in the 2-dimensional simplex that depicts the correspondence of F_1 and F_2 in probability space (Figure 1, panel a).[4]

Two discrete samples yield the *sample* p-p plot (Figure 1, panel b). Let $X_1 = \{x_{1,1}, ..., x_{1,n_1}\}$ be n_1 realizations of the random variable X_1 with discrete sample CDF F_{1,n_1}, and let $X_2 = \{x_{2,1}, ..., x_{2,n_2}\}$ be n_2 realizations of random variable X_2 with discrete sample CDF F_{2,n_2}. The set of ordered values of the joint support of X_1 and X_2 is $\{z_1, ...z_m\}$, whereby $m \leq n_1 + n_2$.

[4] Written as a function rather than a plot it reads as: $p \longmapsto F_1(F_2^{-1}(p))$, $0 \leq p \leq 1$,whereby $F_2^{-1}(p) = \inf\{x : F_2(x) \geq p\}$.

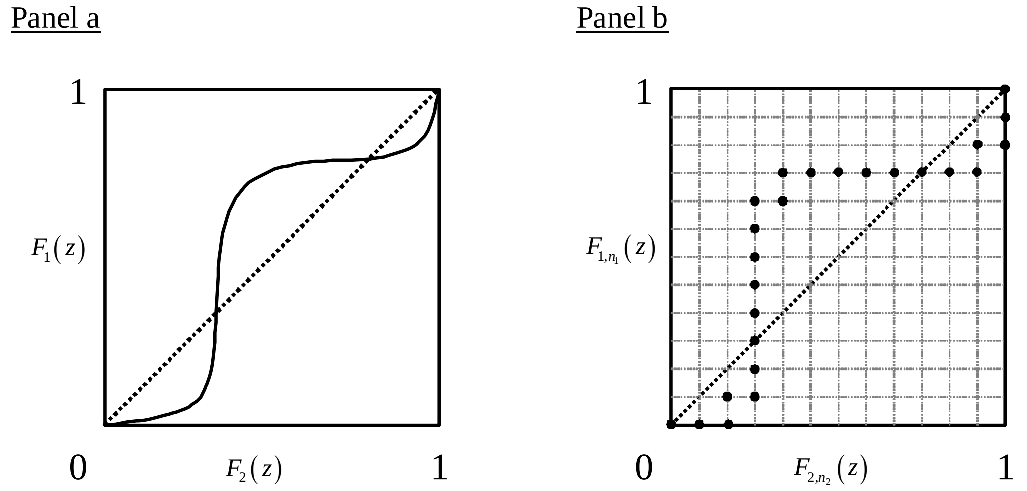

Figure 1: Continuous p-p plot (panel a) and corresponding discrete (sample) p-p plot (panel b).

In addition let z_0 denote $-\infty$ and define $\mathbf{z} \equiv \{z_0, ..., z_m\}$. The vertical coordinates of the sample p-p plot equal $P\left[X_1 \leq z\right] \; \forall \; z \in \mathbf{z}$, while the horizontal coordinates are given by $P\left[X_2 \leq z\right] \; \forall \; z \in \mathbf{z}$ (Bamber, 1975). The sample p-p plot thus reads as:

$$z(X_1, X_2) \longmapsto \begin{bmatrix} F_{1,n_1}(z) \\ F_{2,n_2}(z) \end{bmatrix}. \tag{2}$$

Let $\Xi_2 = \{F | \, \forall x, \, h \in \mathbb{R}$: $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$, $\lim_{h \longrightarrow 0} F(x+h) = F(x)$, and $a < b \Longrightarrow F(a) < F(b)$, for $F(a), F(b) \in (0,1)\}$. Note that $\Xi_2 \subset \Xi_1$, that functions belonging to Ξ_2 are continuous and strictly increasing on their support, and that mass points are absent. For the remainder we restrict the analysis to balanced samples absent ties:[5]

Assumption A1: $F_1, \, F_2 \in \Xi_2$.

[5] Within-sample ties eliminate points from the grid of the sample p-p plot while between-sample ties induce the continuous p-p plot to deviate from the grid lines. Both types make the number of distinct sample p-p plots not tractable analytically.

Assumption A2: $n_1 = n_2 = n$.

To identify the number of distinct sample p-p plots for sample size n we describe how the set of sample p-p plots develops when the sample expands. Increasing the sample size from $n = 0$ tot $n = 1$ creates two sample p-p plots: one going from (0,0) to (1,1) via (0,1) and another via (1,0), see panel a of Figure 2. Adding another observation creates six sample p-p plots in total. Three of these go through point a, and three run through point b, see panel b of Figure 2. Panel c applies when going from $n = 2$ to $n = 3$. As of points c and f there are four possible continuations of the p-p plot, while there are three possible continuations from points d and e onwards, leading to 20 different sample p-p plots in total, and so on.

2.1 Border point, border number and history number

To capture the recursive pattern in Figure 2 we first introduce two labels: the *p-p grid* refers to the grid of the sample p-p plot, and the *grid step* is a continuation of the sample p-p plot of length $1/n$ in either the horizontal or vertical direction. Next, we identify a group of special points on the p-p grid:

Definition 1 *Border point $x(n)$: any point on the p-p grid with coordinates $(1 - 1/n, c)$ or $(c, 1 - 1/n)$ for some $c \in \{0, ..., 1 - 2/n\}$.*

Point e in panel c is a border point, while it is not in panel d. Let $X(n)$ be the set of border points for sample size n. Border points have an important property (proofs of lemmata are in the Appendix, Section 5.2):

Lemma 1 *Given sample size n, any sample p-p plot passes through some $x(n) \in X(n)$.*

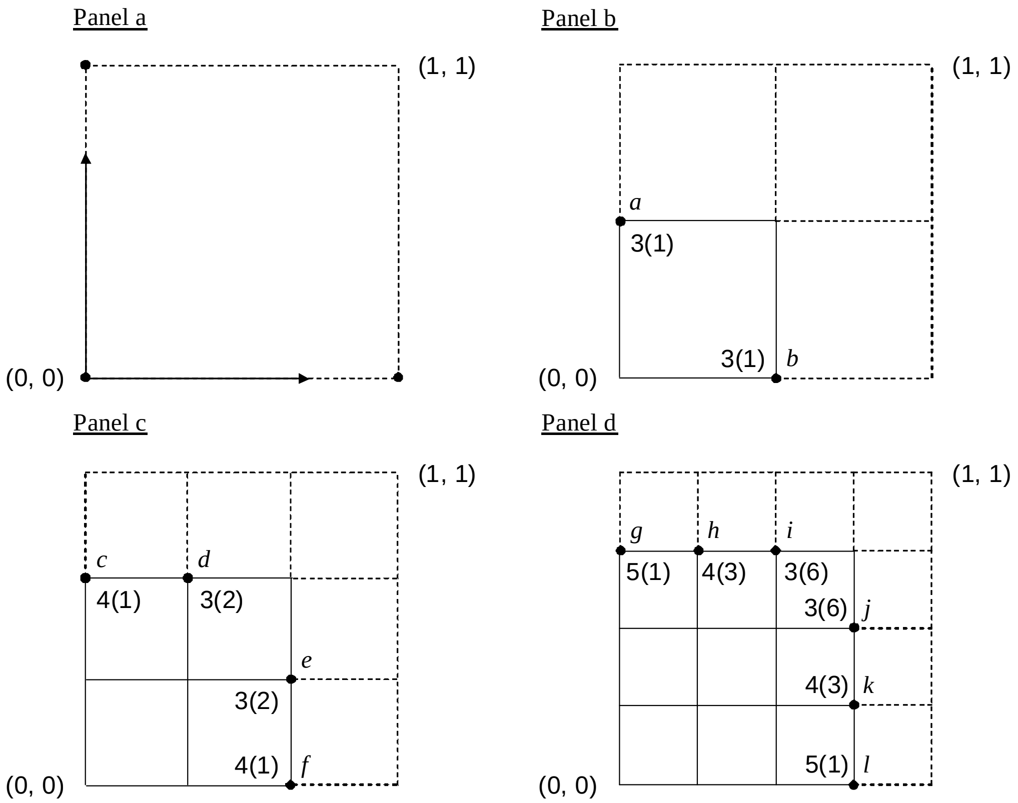

Figure 2: Sample p-p plots, border points, border numbers and history numbers; panels a through d respectively refer to n going from 0 to 1, from 1 to 2, from 2 to 3, and from 3 to 4.

Lemma 1 implies that all p-p plots passing through any $x(n) \in X(n)$ necessarily have passed through some $x(n-1) \in X(n-1)$. Accordingly, to reveal the pattern in the development of the set of distinct p-p plots, it suffices to keep track of what happens at the border points.

For each border point we introduce two numbers.

Definition 2 *Border number $bn(x(n))$: the number of border points $x(n+1)$ that can be reached from $x(n)$ onwards when the sample size increases to $n+1$.*

For example, as of point d in panel c three border points can be reached in panel d: h, i and j, yielding $bn(d) = 3$. Observe that as of point d onwards there are also three possible routes to arrive at (1,1). This correspondence holds in general (proofs of properties are in the Appendix, Section 5.1):

Property P1 *Border number $bn(x(n))$ coincides with the number of distinct continuations of the sample p-p plot from $x(n)$ onwards to (1,1).*

This property allows border numbers to be calculated:

Lemma 2 $bn(x(n)) = 1 + n(1-c)$.

Border numbers are also uniquely related to the shape of the p-p plot running through the underlying border point:

Property P2 *A higher border number implies a larger distance between the diagonal and the p-p plot.*

This second property will be instrumental for deriving the finite sample distribution of EDF statistics.

The second number associated with border points is related to the path that lead to it. Let $BN(x(n)) \subset X(n+1)$ be the set of border points that can be reached from $x(n)$ onwards when the sample size increases to $n+1$.

For instance, in Figure 2 panel c we have that $BN(d) = \{h, i, j\}$. Further, let $HN(x(n)) \subset X(n-1)$ be the set of border points that can reach $x(n)$ when the sample size increases from $n-1$ to n. For example, $HN(d) = \{a, b\}$. We then introduce:

Definition 3 *History number $hn(x(n))$: the number of distinct continuations of the sample p-p plot from all $x(n-1) \in HN(x(n))$ towards $x(n)$ without passing through any other border number $x\prime(n) \in X(n)$.*

For example, point h in panel d can be reached in three different ways from two border points in panel c: twice by passing d without going through c and once by passing c without going through g. This yields $hn(h) = 3$. In Figure 2 the border numbers are depicted with their concomitant history number in brackets. History numbers are uniquely related to the probability that a p-p plot passes through the underlying border point:

Property P3 *History number $hn(x(n))$ is proportional to the probability that $x(n)$ is part of some sample p-p plot.*

Property P3 is the second building block of the finite sample distribution of EDF statistics.

Intuitively, knowing how many distinct routes lead to a particular border point and how many distinct continuations of the p-p plot this border point allows, suffices to determine the number of distinct sample p-p plots. The next proposition formalizes this intuition (proofs of propositions are in the Appendix, Section 5.3):

Proposition 1 *For any $n > 1$ the number of distinct sample p-p plots equals:*

$$\Omega(n) = \sum_{x(n) \in X(n)} bn(x(n))hn(x(n)).$$

2.2 A logical tree

To calculate the number of distinct sample p-p plots thus requires all border points to be identified, together with their border numbers and history numbers, for any sample size n. For that we introduce a logical tree Γ as in Figure 3. It groups together all border points with the same border number by branch. Each branch reflects the history of the underlying p-p plot. There is a close relation between logical tree Γ and the p-p plots in Figure 2. Going from $n = 1$ to $n = 2$ creates two border points, a and b, that both have border number 3. These border points are grouped at the first node of the tree. In case $n = 3$ there are four border points, two of which have border number 3 (d and e), and two that have border number 4 (c and f). Logical tree Γ splits up accordingly: one branch that continues with border number 3 and another that continues with border number 4. And so on.

Let $N(\gamma_k(n))$ be the set of all border points $x(n)$ at node $\gamma_k(n)$ of logical tree Γ, where $k = 1, ...K(n)$ refers to a branch, $K(n)$ being the number of branches for sample size n. For instance, $N(\gamma_1(4)) = \{i, j\}$ at the upper most node. All points in $N(\gamma_k(n))$ have identical border numbers and history numbers. Property P1 implies that there is one, and only one, border number associated with any border point. These border numbers are included at the nodes in logical tree Γ and referred to as $bn(\gamma_k(n))$. For history numbers the situation is more involved, as they reflect the *total* number of distinct routes that the underlying p-p plot can have taken to reach a particular border point. For instance, border point i has history number 6. This is the sum of two distinct possibilities: it could have been reached through d or e, yielding $hn(i|\, d, e) = 4$, or via c or f, yielding $hn(i|\, c, f) = 2$. In logical tree Γ this distinction is made explicit: each border point enters with its 'net' history number. Accordingly, border points can enter the tree at more than

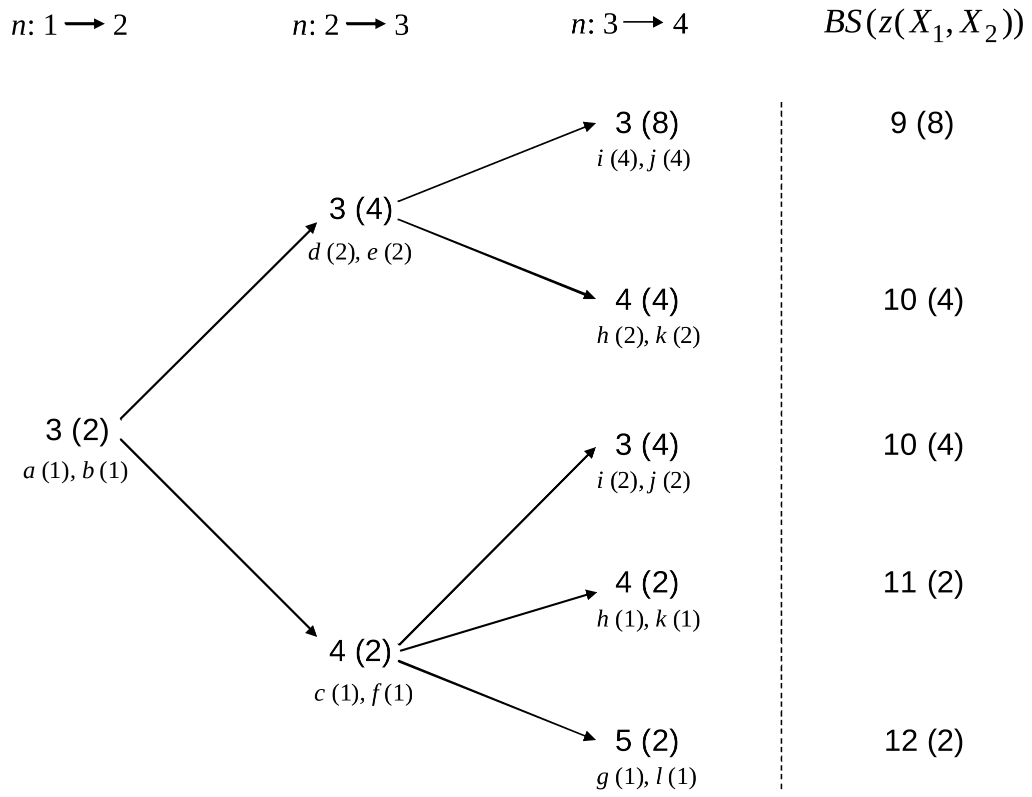

Figure 3: Logical tree Γ for the evolution of distinct p-p plots when the sample size increases.

one node. The numbers in brackets next to the border numbers $bn(\gamma_k(n))$ in logical tree Γ are therefore no history numbers, but the sum of the (net) history numbers of all border points at any node $\gamma(n)$. We label these history sums:

Definition 4 *History sum $HS(\gamma_k(n))$: the sum of all history numbers of the border points at node $\gamma_k(n)$, that is,*

$$HS(\gamma_k(n)) = \sum_{x(n)\in N(\gamma_k(n))} hn(x(n)).$$

Property P3 implies that history sums are proportional to the probability of reaching node $\gamma_k(n)$.

Considering then the development of border numbers in logical tree Γ suggests that border numbers for sample size n always give rise to the same sequence of border numbers for sample size $n+1$. For example, border number 3 is split into $\{3,4\}$ while border number 4 evolves into $\{3,4,5\}$, and so on. This is due to a recursive pattern indeed:

Lemma 3 *At any node $\gamma_k(n)$ in logical tree Γ, border number $bn(\gamma_k(n))$ splits into $bn(\gamma_k(n)){-}1$ border numbers with respective values $\{3,4,...,bn(\gamma_k(n))$ $+1\}$ when the sample size increases from n to $n+1$.*

Figure 3 suggests also an obvious recursive pattern in the evolution of history sums:

Lemma 4 *At any node $\gamma_k(n)$ in logical tree Γ history sum $HS(\gamma_k(n))$ splits into $bn(\gamma_k(n)){-}1$ history sums with respective values $\{2HS(\gamma_k(n)), HS(\gamma_k(n)),$ $..., HS(\gamma_k(n))\}$ for the border numbers $\{3,4,...,bn(\gamma_k(n)){+}1\}$ when the sample size increases from n to $n+1$.*

Lemmata 3 and 4 jointly describe how logical tree Γ evolves when the sample size increases. To describe this development in compact matrix notation we need to introduce one additional number. Let $BP(z(X_1, X_2))$ be the set of border points $x(i) \subset X(i)$, $i = 2, ..., n$, that sample p-p plot $z(X_1, X_2)$ crosses. For example, $BP(z(X_1, X_2)) = \{a, d, h, i\}$ implies that $z(X_1, X_2)$ crosses points a, d, h and i in Figure 2. The border sum is then the sum of all border numbers at these border points:

Definition 5 *Border sum $BS(z(X_1, X_2)$: the sum of all border numbers at the border points through which $z(X_1, X_2)$ passes, that is,*

$$BS(z(X_1, X_2)) = \sum_{x(i) \in BP(z(X_1, X_2))} bn(x(i))$$

Border sums are included to the right of the vertical dashed line in Figure 3, together with the history sums at the concomitant node $\gamma_k(n)$.

The compact matrix notation now follows. Let $BM(n) \equiv (\Delta_n | \, 2\Upsilon_n)$ for $n > 1$, where Δ_n is the upper triangular unit matrix of size n and Υ_n the unit vector of size n, and $BM(1) = [1 \; 2]$, a 1×2 matrix. Matrix $BM(n)$ contains the development factors of the history sums. For example, for $n = 3$ we have for $BM(3)$:

$$\begin{array}{c} \\ 5 \\ 4 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 6 & 5 & 4 & 3 \end{array} \\ \left[\begin{array}{cccc} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 2 \end{array}\right] \end{array},$$

where rows and columns respectively refer to the border numbers for $n = 3$ and $n = 4$. Further let $HM(1) \equiv 2$ and let $HM(n)$ be a $\Theta(n) \times n$ matrix with elements $hm_{jk}(n)$ defined as:

$$hm_{jk}(n) = \begin{cases} a_{(j-k+1)k}(n), & j = k, ..., \Theta(n-1) + k - 1, \; k = 1, ..., n, \\ 0 & otherwise, \end{cases}$$

with $\Theta(n) \equiv 1 + n(n-1)/2$, where $A(n+1) \equiv HM(n)BM(n)$, whereby $A(1) \equiv 2$, and a_{jk} is an entry from the auxiliary matrix $A(n+1)$. Note that

$A(n+1)$ is needed to construct $HM(n+1)$. Indeed, matrix $HM(n)$ contains the history sums themselves. For example, $HM(3)$ boils down to:

$$\begin{array}{c c} & \begin{array}{ccc} 5 & 4 & 3 \end{array} \\ \begin{array}{c} 12 \\ 11 \\ 10 \\ 9 \end{array} & \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 4 & 4 \\ 0 & 0 & 8 \end{bmatrix} \end{array},$$

where the columns refer to the distinct border numbers for $n = 3$, and where the rows refer to all distinct values of $BS(z(X_1, X_2))$ for $n = 3$. Observe that the number of distinct sample p-p plots equals the sum of all entries in $A(n)$: $\Omega(n) = \sum_i \sum_j a_{ij}(n)$. For instance:

$$A(4) = HM(3)BM(3) = \begin{bmatrix} 2 & 2 & 2 & 4 \\ 0 & 2 & 2 & 4 \\ 0 & 4 & 8 & 16 \\ 0 & 0 & 8 & 16 \end{bmatrix},$$

and $\sum_i \sum_j a_{ij}(4) = 70 = \Omega(4)$.

3 An application: the finite sample distribution of FCvM$_1$

The analysis of the previous section can be used to derive the finite sample distribution of $FCvM_1$. Note that a p-p plot coincides with the diagonal if, and only if, the two underlying distributions are identical. EDF tests are therefore based on the distance between the p-p plot and the diagonal (see Figure 4, panel a). For instance, the Kolmogorov-Smirnov test considers the largest absolute value of the maximum positive distance (D^+) and the maximum negative distance (D^-), the Kuiper test considers the sum of D^+ and D^-, the L_1 (L_2) version of $FCvM$ test sums up over all absolute (squared) distances d, and the Anderson-Darling test augments $FCvM_2$ by weighing every squared distance with the product of the distance between 0 and the centre of d, and the distance between 1 and the centre of d.

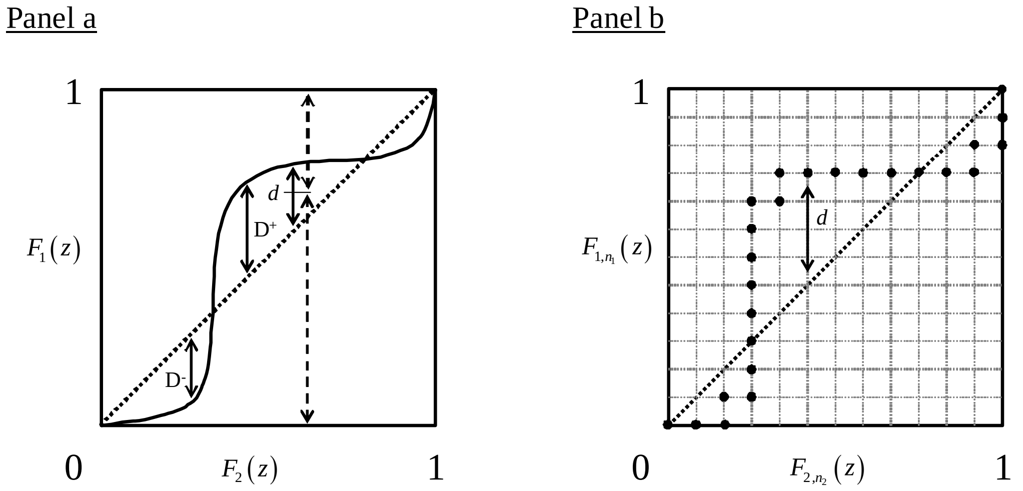

Figure 4: Continuous p-p plot (panel a) and corresponding discrete (sample) p-p plot (panel b).

3.1 Computation

Given X_1 and X_2, $FCvM_1$ equals (Schmidt and Trede, 1996):

$$FCvM_1(X_1, X_2) = S(n_1, n_2) \sum_{i=1}^{n_1+n_2} \left| F_{1,n_1}(z_i) - F_{2,n_2}(z_i) \right|, \tag{3}$$

where the sample size correction factor $S(n_1, n_2)$ follows Rosenblatt (1952) and Fisz (1960):

$$S(n_1, n_2) = \sqrt{\frac{n_1 n_2}{(n_1 + n_2)^3}}. \tag{4}$$

This correction factor speeds up convergence of the finite sample distribution towards the limiting distribution.

Observe two properties of $FCvM_1$:

Property P4 *Under A1 – A2, the number of distinct values of* $FCvM_1(X_1,X_2)$ *is* $\Theta(n) \equiv 1 + n(n-1)/2$.

Property P5 *Under A1 – A2, the vector containing all possible, distinct* *values of $FCvM_1(X_1,X_2)$ is* $\mathbf{FCvM}_1(n) \equiv \{FCvM_1^1, ..., FCvM_1^{\Theta(n)}\}$*, where* $FCvM_1^j = [n^2 - 2(j-1)] \Big/ 2\sqrt{2n^3}$, $j = 1, ..., \Theta(n)$.

3.2 Finite sample distribution

To derive the distribution of $FCvM_1(X_1, X_2)$ we order all distinct sample p-p plots $z(X_1, X_2)$ according to the corresponding value of $FCvM_1^j$ and link all values of $\mathbf{FCvM}_1(n)$ to the relative frequency of occurrence of the underlying p-p plot. First note that the value of $FCvM_1(X_1, X_2)$ is uniquely related to the border sum:

Lemma 5 *$BS(z(X_1, X_2)) \propto FCvM_1(X_1, X_2)$.*

Hence, to keep track of $\mathbf{FCvM}_1(n)$, it suffices to trace the development of $BS(z(X_1, X_2))$ for an expanding sample size.

Figure 3 also displays the 'history numbers' of border sums: the history sum of all border points at the concomitant node $\gamma_k(n)$. Recall that Property P3 implies that these history sums yield the frequency of obtaining any p-p plot that passes through the same border numbers as $z(X_1, X_2)$, be it in a possibly different order.[6] Indeed, let $HM_j(n)$ refer to row j of $HM(n)$. The following then holds:

Proposition 2 *Under* H_0 *and A1 – A2,* $FCvM_1^j(n)$ *has probability* $p^j(n)$*, where*

$$p^j(n) = \frac{HM_j(n)\Upsilon_n}{\Omega(n)},$$

$j = 1, ..., \Theta(n)$.

[6] At the same time, different p-p plots can have the same border sum while their history sums differ because the underlying p-p plots passes through a different set of border numbers. In Figure 3 this situation can arise for border points i, j, h and k.

As an illustration, let $n = 3$. There are four distinct values of $BS(z(X_1, X_2))$ that corresponds to the four different values of $FCvM_1(X_1, X_2)$ as given in Property P5: $(1, 7/9, 5/9, 3/9)\sqrt{3/8}$. Their respective relative frequencies then equal the probabilities that the corresponding values of $BS(z(X_1, X_2))$ emerge, which follow from the concomitant history sums: $(2/20, 2/20, 8/20/8/20)$.

3.3 Hypothesis testing

The exact critical percentiles, $FCvM_1^{\alpha}(n)$, are given in Table 1 for $n = 3, ..., 20$ and $\alpha = 0.90$, 0.95, 0.975, and 0.99.[7] High values of $FCvM_1^{\alpha}(n)$ imply a low probability that the underlying samples are drawn from the same distribution.

Schmidt and Trede (1995) note that the limiting distribution of $FCvM_1(X_1, X_2)$ corresponds to the limiting distribution of the L_1-norm of a Brownian bridge. Johnson and Killeen (1983) derive the analytical expression for the latter and tabulate its critical values. These values are in the last row of Table 1 and are denoted by $FCvM_1^{\alpha}(\infty)$.[8]

The question then is whether relying on the percentiles of the limiting distribution in case of small samples will lead to differences in the rejection of H_0. To answer this question we use Property P5 to examine whether $FCvM_1(X_1, X_2)$ can obtain a value in between the true critical percentiles and those of the limiting distribution. This turns out not to be possible:[9]

Proposition 3 $P[FCvM_1(X_1, X_2) < FCvM_1^{\alpha}(n)] = P[FCvM_1(X_1, X_2) < FCvM_1^{\alpha}(\infty)]$.

[7] The entries do not display a monotonously declining pattern because the critical values of $FCvM_1(X_1, X_2)$ are falling in sample size absent the sample size correction factor, whereas this factor itself is increasing in sample size.

[8] Johnson and Killeen (1983) do not report $FCvM_1^{0.975}(\infty)$.

[9] The computing code (GAUSS) is available upon request, as is all computing code used for this paper.

percentile n	90	95	97.5	99
3	0.6124			
4	0.5303	0.6187		
5	0.5376	0.6008	0.6641	0.7273
6	0.5292	0.5774	0.6736	0.7217
7	0.5154	0.5918	0.6682	0.7445
8	0.5000	0.5938	0.6563	0.7500
9	0.5107	0.5893	0.6678	0.7464
10	0.5143	0.5814	0.6485	0.7379
11	0.5136	0.5911	0.6493	0.7462
12	0.5103	0.5784	0.6634	0.7485
13	0.5054	0.5808	0.6562	0.7467
14	0.4995	0.5804	0.6614	0.7424
15	0.5051	0.5903	0.6634	0.7486
16	0.5082	0.5856	0.6629	0.7513
17	0.4994	0.5801	0.6608	0.7516
18	0.5000	0.5833	0.6574	0.7500
19	0.4995	0.5849	0.6617	0.7471
20	0.5060	0.5850	0.6562	0.7510
∞	0.4993	0.5821	*	0.7518

Table 1: Exact critical values of the FCvM1(X1,X2) under A1 - A2 at percentile 90, 95, 97.5, and 99.

That is, although for finite samples the critical percentiles of the limiting distribution differ from their true values, this will not lead to differences in the rejection of H_0.

4 Conclusions

For two balanced samples absent ties we characterize how the set of distinct sample p-p plots expands when the sample size increases. We then order all sample p-p plots according to the corresponding value of $FCvM_1$ and link these values to the relative frequency of occurrence of the underlying sample p-p plot. In this way we obtain the finite sample distribution of $FCvM_1$.

The (known) critical percentiles of the limiting distribution of $FCvM_1$ can thus be compared with the exact finite sample critical percentiles. This shows that using the former will not lead to differences in the rejection of the hypothesis that the distributions from which the two samples are drawn, are identical.

It is left for future research to examine whether our analysis of p-p plots can be used to derive the finite sample distribution of other EDF test statistics.

References

[1] Anderson, T. W. and D. A. Darling (1952), "Asymptotic theory of certain goodness of fit criteria based on stochastic processes", *Annals of Mathematical Statistics* **23**: 193 – 212.

[2] Bamber, D. (1975), "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph", *Journal of Mathematical Psychology* **12**: 387 – 415.

[3] Cramér, H. (1928), "On the composition of elementary errors II: statistical applications", *Skandinavisk Aktuarietidskrift* **11**: 141 – 180.

[4] Famoye, F. (1999), "EDF tests for the generalized Poisson distribution", *Journal of Statistical Computation and Simulation* **63**: 159 – 168.

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

[6] Henze, N. (1996), "Empirical-distribution-function goodness-of-fit tests for discrete models", *Canadian Journal of Statistics* **24**: 81 – 93.

[7] Holmgren, E. B. (1995), "The p-p plot as a method of comparing treatment effects", *Journal of the American Statistical Society* **90**: 360 – 365.

[8] Johnson, B. McK. and T. Killeen (1983), "An explicit formula for the C.D.F. of the L_1 norm of the Brownian bridge", *The Annals of Probability* **11**: 807 – 808.

[9] Kolmogorov, A. N. (1933), "Sulla determinizazione empirica delle leggi di probabilita", *Giornale dell 'Istituto Italiano Attuari* **4**: 1 – 11.

[10] Kuiper, N. H. (1960), "Tests concerning random points on a circle", *Koninklijke Nederlandse Akademie van Wetenschappen*, The Netherlands.

[11] Olea, R. A. and Pawlowsky-Glahn, V. (2009), "Kolmogorov-Smirnov test for spatially correlated data", *Stochastic Environmental Research and Risk Assessment* **23**: 749 – 757.

[12] Rosenblatt, M. (1952), "Limit theorems associated with variants of the von Mises statistic", *Annals of Mathematical Statistics* **23**: 617 – 623.

[13] Schmidt, F. and M. Trede (1995), “A distribution free test for the two sample problem for general alternatives”, *Computational Statistics & Data Analysis* **20**: 409 – 419.

[14] Schmidt, F. and M. Trede (1996), “An L_1-variant of the Cramér-von Mises test”, *Statistics and Probability Letters* **26**: 91 – 96.

[15] Smirnov, N. V. (1939), “On the deviation of the empirical distribution function”, *Rec. Math.* [*Mathematicheskii Sbornik*] *N.S.* **6**: 3 – 26.

[16] Stephens, M. A. (1974), “EDF statistics for goodness of fit and some comparisons”, *Journal of the American Statistical Society* **69**: 730 – 737.

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

[18] Wilk, M. B. and R. Gnanadesikan (1968), “Probability plotting methods for the analysis of data”, *Biometrika* **55**: 1 – 17.

5 Appendix

5.1 Proofs of properties

5.1.1 Proof of Property P1

Proof. Note that all border points in $BN(x(n))$ have one, and only one, coordinate equal to 1. Hence, from any point $x(n+1) \in BN(x(n))$ onwards there is one, and only one route towards (1,1): along the border of the p-p grid for which the coordinate of $x(n+1)$ equals 1 at sample size n. $\blacksquare$

5.1.2 Proof of Property P2

Proof. A higher border $bn(x(n))$ number is exclusively due to an increase in n, or a decrease in c (Lemma 2). The distance between $bn(x(n))$ and the diagonal of the p-p plot equals $|bn(x(n)) - c| = 1 + n - (n-1)c$. The property then follows as $\forall c \in [0, 1 - 2/n)$, $\partial\, |bn(x(n)) - c| / \partial n > 0$, and $\partial\, |bn(x(n)) - c| / \partial c < 0$. ■

5.1.3 Proof of Property P3

Proof. Lemma 1, Property P1 and Definition 3 jointly imply that the number of distinct sample p-p plots that pass through $x(n)$, without passing through any other $x\prime(n) \in X(n)$, equals $\Phi(n) = bn(x(n))hn(x(n))$. The property then follows as $\partial\Phi(n)/\partial hn(x(n)) = bn(x(n)) > 0$. ■

5.1.4 Proof of Property P4

Proof. Under A1-A2 the smallest value of $FCvM_1$ obtains when each next grid step is vertical (horizontal) after a horizontal (vertical) step. In that case, n distances d in Panel b of Figure 1 are equal to $1/n$ and n distances d are zero, yielding $FCvM_1 = 1/\sqrt{8n}$. $FCvM_1$ obtains its largest value when n consecutive grid steps are either vertical or horizontal, yielding $FCvM_1 = 1/\sqrt{8n}\left(2\sum_{i=1}^{n-1} i + n\right) = \sqrt{n/8}$. The smallest difference between two values of $FCvM_1$ is twice distance d of length $1/n$, multiplied by $S(n)$: $2/n\sqrt{8n} = 1 \Big/ \sqrt{2n^3}$. Hence, the number of distinct values of $FCvM_1$ is: $1 + \left(\sqrt{n/8} - 1/\sqrt{8n}\right) \Big/ \left(1/\sqrt{2n^3}\right) = 1 + n(n-1)/2$. ■

5.1.5 Proof of Property P5

Proof. First note that $FCvM_1^j$ is decreasing in j. Hence, $FCvM_1^1 = \sqrt{n/8}$ is the largest value of $FCvM_1$, while $FCvM_1^{\Theta(n)} = 1/\sqrt{8n}$ is the smallest value of $FCvM_n^1$. Because $FCvM_1^{j+1} - FCvM_1^j = 1/\sqrt{2n^3}$, which is the smallest difference between two values of $FCvM_1$ (see Property P4), the lemma follows. ■

5.2 Proofs of lemmata

5.2.1 Proof of Lemma 1

Proof. Because all p-p plots start at (0,0) and end at (1,1), both coordinates of the p-p plot run through the sequence $\{0, 1/n, , ..., 1 - 1/n, 1\}$, whereby each next grid step adjusts one coordinate only. Hence, if one coordinate is the first to equal $1 - 1/n$, the other must equal $c \in \{0, ..., 1 - 2/n\}$. ■

5.2.2 Proof of Lemma 2

Proof. Consider some border point $x(n)$ above the diagonal with coordinates $(c, 1 - 1/n)$, $c \in \{0, ..., 1 - 2/n\}$. All p-p plots emanating from $x(n)$ take one, and only one vertical grid step at one of the horizontal positions $i \in \{c, c + 1/n, ..., 1\}$, whereby the number of distinct values i is $1 + n(1 - c)$. An identical reasoning applies for any $x(n)$ with coordinates $(1 - 1/n, c)$, $c \in \{0, ..., 1 - 2/n\}$. ■

5.2.3 Proof of Lemma 3

Proof. From any border point $x(n)$ all border points with respective border numbers $\{3, 4, ..., bn(x(n))\}$ can be reached at least once. In addition, there is one border point with border number $bn(x(n)) + 1$ that can be reached as well. ■

5.2.4 Proof of Lemma 4

Proof. First, from any border point $x(n)$ all border points with respective border numbers $\{4, 5, ..., bn(x(n)) + 1\}$ can be reached once, and only once. Accordingly, the history numbers do not change when the sample size increases. Second, border points with border number 3 can be reached from two different border points that have the same border number and history number. And because these border points are grouped together in logical tree Γ, the history number of border points with border number 3 is twice the history number of the border point they emanate from. ■

5.2.5 Proof of Lemma 5

Proof. As $FCvM_1(X_1, X_2)$ sums up the absolute distances between all border points in $BP(z(X_1, X_2))$ and the diagonal, it follows from Property P2 that the value of $FCvM_1(X_1, X_2)$ is positively related to $BS(z(X_1, X_2))$. ■

5.3 Proofs of Propositions

5.3.1 Proof of Proposition 1

Proof. Recall from the proof of Property P3 that $\Phi(n) = bn(x(n))hn(x(n))$ is the number of distinct sample p-p plots that pass through $x(n)$, without passing through any other $x\prime(n) \in X(n)$. Summing up over all border points $x(n) \in X(n)$ then yields the proposition. ■

5.3.2 Proof of Proposition 2

Proof. By construction, $HM_j(n)$ refers to all $BS(z(X_1, X_2))$ with an identical value, which, according to Lemma 5, yield the same value for $FCvM_1(X_1, X_2)$. Hence, $HM_j(n)\Upsilon_n$ is the frequency of observing this particular value of $FCvM_1(X_1, X_2)$. ■