# Proof of a conjecture of Narayana on dominance refinements of the Smirnov two-sample tests

***Citation for published version (APA):***
Di Bucchianico, A., & Loeb, D. E. (1994). Proof of a conjecture of Narayana on dominance refinements of the Smirnov two-sample tests. (Memorandum COSOR; Vol. 9418). Eindhoven: Technische Universiteit Eindhoven.

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 94-18

# Proof of a conjecture of Narayana on dominance refinements of the Smirnov two-sample test

A. Di Bucchianico
D. Loeb

Eindhoven, June 1994
The Netherlands

# Proof of a conjecture of Narayana on dominance refinements of the Smirnov two-sample test

A. Di Bucchianico[*]
Department of Mathematics and Computing Science
Eindhoven University of Technology
P. O. Box 513
5600 MB Eindhoven, The Netherlands
sandro@win.tue.nl

D. Loeb[*]
LaBRI (URA CNRS 1304)
Université de Bordeaux I
33405 Talence, France
loeb@labri.u-bordeaux.fr

## Abstract

**We prove the following conjecture of Narayana: there are no dominance refinements of the Smirnov two-sample test if and only if the two sample sizes are relatively prime.**

**Keywords** Smirnov two-sample test, dominance refinement, Gnedenko path, dominance.

**AMS classification** 62G10, 05A15

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from continuous distribution functions $F$ and $G$, respectively. In order to test nonparametrically whether $X_1$ is stochastically smaller than $Y_1$, one often uses the Smirnov statistic $D^+_{mn}$ defined by

$$D^+_{mn} = \sup_t \left(F_m(t) - G_n(t)\right), \qquad (1)$$

where $F_m$ and $G_n$ are the empirical distribution functions of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ respectively. A convenient way to study the distribution of $D^+_{mn}$ is the so-called Gnedenko path. The Gnedenko path $\omega$ of the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ is defined as follows: $\omega$ is a path from $(0,0)$ to $(m,n)$ with unit steps $\omega_i$ to the east or north. If

[*] Authors supported by NATO CRG 930554. Second author partially supported by EC grant CHRX-CT93-0400 and the PRC Maths-Info.

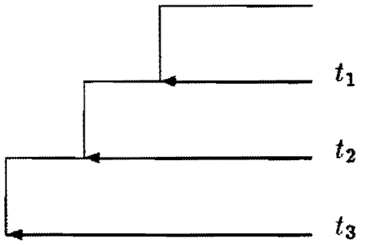

Figure 1: Representation of a Path

the $i$th value of the ordered combined sample comes from $X_1, \ldots, X_m$, then $\omega_i$ is a step east; otherwise, it is a step north. Since we assume that $F$ and $G$ are continuous, the probability of a tie (*i.e.*, $X_i = Y_j$) is zero. Hence, $\omega$ is almost surely well-defined. It is easy to see that under $H_0 : F = G$, all paths from $(0,0)$ to $(m,n)$ are equiprobable, *i.e.*, $\mathbf{P}(w) = 1/\binom{m+n}{n}$ for all paths $w$ (see *e.g.*, Hájek (1969, Theorem 5C)). Now,

$$mnD^{+}_{mn} \leq r$$

if and only if all vertices $(x, y)$ of the Gnedenko path satisfy

$$nx - my \geq r$$

In other words, $mnD^{+}_{mn} > r$ if and only if $\omega$ passes below the line $nx - my = r$. A convenient way to describe a path $\omega$ is to represent it by a $n$-tuple $\langle t_1, \ldots, t_n \rangle$, where $t_i$ is the minimal horizontal distance from $(m, n - i)$ to $\omega$ (see Figure 1). The path $\langle s_1, \ldots, s_n \rangle$ is said to **dominate** $\langle t_1, \ldots, t_n \rangle$ if $s_i \geq t_i$ for $i = 1, \ldots, n$. There is clearly a minimal path $\langle t_1, \ldots, t_n \rangle$ called the $r$-**profile** that lies above (possibly touching) the line $nx - my = r$ (cf. Figure 1).

Thus, we may cast the (upper-tailed) Smirnov two-sample test completely in terms of Gnedenko paths as follows: $mnD^{+}_{mn} \leq r$ if and only if the Gnedenko path dominates the $r$-profile. Thus, the Smirnov two-sample test is completely characterized by its $r$-profiles (*i.e.*, we regard the test as a set of critical regions, indexed by its natural levels). This formulation shows that we attain more levels if we can insert intermediate paths between consecutive $r$-profiles of the Smirnov two-sample test (see Narayana (1979, Chapter 2)). A set of paths totally ordered by dominance is said to be a **dominance refinement** of any set of paths included in it. The Smirnov test is of course a **trivial** dominance refinements of itself. A set of paths is **saturated** if it has no nontrivial dominance refinement. Of course, there exist other ways of refining the Smirnov test. Each partition of the set of paths with a common value $r$ of the statistic $D^{+}_{mn}$ (*i.e.*, all paths that touch but do not cross the line $nx - my = r$) yields a refinement of the Smirnov test. *E.g.*, we can divide the paths that touch but do not cross the line $nx - my = r$ according to the number of times that they touch the line $nx - my = r$.

Dominance refinements partition the set of paths with a common value of $D^+_{mn}$ into dominance regions, *i.e.*, collections of paths that dominate a given critical path. An advantage of dominance refinements is that they can be described very efficiently by simply listing the critical paths. Hence, the refined test can be performed graphically. Another reason for considering dominance refinements (or the notion of dominance itself) is the following relation with MPR tests (= most powerful rank tests). If $F$ and $G$ have densities $f$ and $g$ respectively, and the likelihood ratio $f/g$ is increasing (as is the case for the Lehmann alternatives $H_a : F = G^k$, $k > 0$), then $s$ dominates $t$ implies $\mathbf{P}(t|F = G^k) \geq \mathbf{P}(s|F = G^k)$ (see Savage (1956)). Thus, if $s$ is in the critical region of an MPR test, then all paths dominated by $s$ must also belong to this critical region. Thus, an MPR test at a fixed level is a dominance test in the terminology of Narayana (see Narayana (1979, Chapter 3, p. 35)). Conversely, dominance tests are good approximations for MPR tests (see Narayana (1979, Chapter 3, pp. 44-45)).

Narayana (1975) stated without proof that dominance refinements of the Smirnov two-sample test exist if and only if $\gcd(m,n) > 1$. This result was stated later as a conjecture in Narayana (1979, Exercise 9, p. 30). The purpose of this paper is to prove this conjecture.

Let us look at two examples in order to get a feeling for the Narayana conjecture.

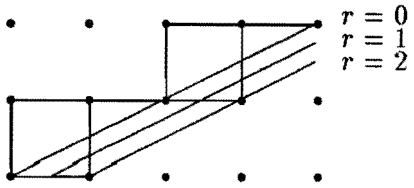

Figure 2: $m = 4$ and $n = 2$

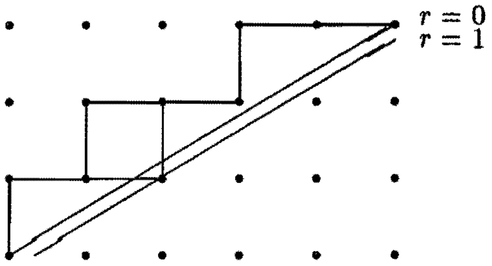

Figure 3: $m = 5$ and $n = 3$.

In Figure 2, the 0-profile and the 1-profile coincide and are equal to the path $\langle 2,4\rangle$; whereas, the 2-profile is the path $\langle 1,3\rangle$. Thus, we see that there are two intermediate paths between the 1-profile and the 2-profile: $\langle 1,4\rangle$ and $\langle 2,3\rangle$. Inserting either of these paths, we obtain a refinement of the Smirnov test. Note that the 2-profile differs from the 1-profile by the possibility to go through the points $(1,0)$ and $(3,1)$, which both lie on the line $2x - 4y = 2$.

In Figure 3, the 0-profile is the path $\langle 2,4,5\rangle$ and the 1-profile is the path $\langle 2,3,5\rangle$. Thus, there is no intermediate path between the profiles; this is also true for the other pairs of consecutive profiles. In other words, there is no refinement. Note that there is only one lattice point on each line of the form $3x - 5y = r$ and that no profiles coincide.

These examples indicate that the existence of dominance refinements depends on the number of lattice points on lines of the form $nx - my = r$. We first enumerate these points and then prove the Narayana conjecture.

**Lemma.** *Let $m$ and $n$ be positive integers. The number of integer solutions $(x, y)$ of*

$$nx - my = r$$

*with the additional constraints*

$$1 \leq x \leq m \text{ and } 0 \leq y \leq n-1$$

*is*

$$\begin{cases} d & \text{if } r = 0, \\ 1 + \left[\frac{d}{n}\left(n - 1 - \max\left(0, \frac{n-r}{m}\right)\right)\right] & \text{if } d \text{ divides } r \text{ and } -nm + m + n \leq r \leq -1, \\ 1 + \left[\frac{d}{m}\left(m - \max\left(1, \frac{r}{n}\right)\right)\right] & \text{if } d \text{ divides } r \text{ and } 1 \leq r \leq nm, \text{ and} \\ 0 & \text{otherwise} \end{cases}$$

*where $d = \gcd(m, n)$ and $[x]$ is the largest integer less than or equal to $x$.*

*Proof*: If $r = d$, then by Euclid's Lemma there exist integer solutions $(x, y)$ of $nx - my = r$. Obviously, this also holds if $r$ is a multiple of $d$. Conversely, if there exists an integer solution $(x, y)$ of $nx - my = r$, then $r$ is a multiple of $d$, since $d$ divides both $m$ and $n$. Thus, integer solutions of $nx - my = r$ exist if and only if $r$ is a multiple of $d$. This proves the last statement.

If $t$, $x$, and $y$ are integers and $nx - my = r$, then $x' := x + tm/d$ and $y' := y + tn/d$ satisfy $nx' - my' = r$. Conversely, if $nx - my = r$ and $nx' - my' = r$, then subtraction yields $n(x - x') = m(y' - y)$. Cancelling the common factor $d$ and using the uniqueness of prime factorizations, we see that there exists an integer $t$ such that $x - x' = tm/d$ and $y' - y = tn/d$. Since $(0, 0)$ does not satisfy the constraints, it immediately follows that there are $d$ integer solutions for $r = 0$.

If $r$ is a negative multiple of $d$, then we must have $r \geq -nm + m + n$, since this corresponds to the extremal solution $x = 1$ and $y = n - 1$. If $x = 1$, then $y = \max(1, (n-r)/m)$ is a (possibly non-integer) solution of $nx - my = r$. Since admissible $y$-values differ by a multiple of $n/d$, the second statement follows.

If $r$ is a positive multiple of $d$, then we must have $r \leq nm$, since this corresponds to the extremal solution $x = m$ and $y = 0$. If $y = 0$, then $x = \max(0, r/n)$ is a (possibly non-integer) solution of $nx - my = r$. Since admissible $x$-values differ by a multiple of $m/d$, the third statement follows. □

**Theorem.** *The Smirnov upper-tailed two-sample test with sample sizes $m$ and $n$ is saturated if and only if $\gcd(m, n) = 1$. In general, the number of dominance refinements (including the trival one) of the Smirnov test is given by the product*

$$\prod_{r=1}^{mn} \sum_{\ell=1}^{\alpha_r} \ell!\, S(\alpha_r, \ell),$$

*where $d = \gcd(m,n)$, $S(k,\ell)$ denotes the Stirling number of the second kind, $\alpha_r = 1 + \left[\frac{d}{m}\left(m - \max\left(1, \frac{r}{n}\right)\right)\right]$ if $d$ divides $r$ and 1 otherwise, and we note that the sum makes no contribution to the product unless $\alpha_r > 1$. The number of saturated dominance refinements of the Smirnov test is given by the product*

$$\prod_{r=1}^{mn} \alpha_r!.$$

*Proof*: We use the representation of the Smirnov test as a set of $r$-profiles. It is convenient to single out the special cases $n = 1$ and $m = n$. If $n = 1$, then all paths from $(0,0)$ to $(m,1)$ are profiles. Hence, there does not exist a dominance refinement in this case. If $m = n$, then the 0-profile is the path $\langle 1,2,\ldots,n\rangle$ and the 1-profile is the path $\langle 0,1,2,\ldots,n-1\rangle$. Thus, dominance refinements exist (*e.g.*, add the path $\langle 0,2,3,\ldots,n\rangle$ to the profiles).

By symmetry, we now assume without loss of generality that $m > n > 1$. Let $A_r$ be the set of all integer solutions $(x,y)$ of the equation $nx - my = r$ with additional constraints $1 \le x \le m$ and $1 \le y \le n$. The next step of the proof consists in showing that dominance refinements exist if and only if there exists an integer $r$ such that the line $nx - my = r$ contains at least two points in the set $A_r$.

Fix an integer $r$ such that $1 \le r \le mn$. Let $(a,b)$ be an arbitrary point of the $r$-profile such that $1 \le a \le m$ and $0 \le b \le n-1$. If $(a,b)$ is on the line $nx - my = r$, then the $r$-profile includes the points $(a-1,b)$, $(a,b)$, and $(a,b+1)$. Since the horizontal and vertical distances between the lines $nx - my = r$ and $nx - my = r-1$ are both strictly smaller than 1, the $(r-1)$-profile must include the points $(a-1,b)$, $(a-1,b+1)$, and $(a,b+1)$.

If $(a,b)$ is not on the line $nx - my = r$, then it follows from the defining minimality property that the $r$-profile must include the points $(a-1,b)$, $(a,b)$, and $(a,b+1)$. In order to show that the $(r-1)$-profile must include these three points too, we need to distinguish three cases.

- $(a,b)$ lies above $nx - my = r-1$ : since $(a,b)$ belongs to the $r$-profile, the vertical and horizontal distances from $(a,b)$ to the line $nx - my = r$, and hence the line $nx - my = r-1$, are strictly less than one. Thus, the $(r-1)$-profile must include the points $(a-1,b)$, $(a,b)$, and $(a,b+1)$.
- $(a,b)$ lies on $nx - my = r-1$ : it follows from minimality that the $(r-1)$-profile must include the points $(a-1,b)$, $(a,b)$, and $(a,b+1)$.
- $(a,b)$ lies below $nx - my = r-1$ : this case cannot occur, since $(a,b)$ lies on the line $nx - my = na - mb$.

Thus, we have shown that dominance refinements exist if and only if there exists a line $nx - my = r$ that contains at least two points in the set $A_r$. The existence part of the theorem now follows from the lemma.

If there are $k$ ($k \geq 2$) points on the line $nx - my = r$ ($1 \leq r \leq mn$), then refinements are possible by inserting chains of paths between the $r$-profile and the $(r-1)$-profile. If we represent profiles by the representation of Figure 1, then we see that the representations of the $r$-profile and the $(r-1)$-profile are the same, except at $k$ places where they differ by one. If we renumber those places to $1, \ldots, k$, then we see that a chain of paths between the $r$-profile and the $(r-1)$-profile is nothing but a chain of subsets of $\{1, \ldots, k\}$. If we look at the differences of consecutive elements of such chains, then we obtain ordered partitions of the set $\{1, \ldots, k\}$. The number of partitions of the set $\{1, \ldots, k\}$ into $\ell$ blocks is $S(k, \ell)$, the Stirling number of the second kind (see *e.g.*, Berge (1971)). Thus, if there are $k$ ($k \geq 2$) points on the line $nx - my = r$ ($1 \leq r \leq mn$), then the number of chains between the $r$-profile and the $(r-1)$-profile (including the trivial chain) equals $\sum_{\ell=1}^{k} \ell!\, S(k, \ell)$. The enumeration part of the theorem now follows from the lemma. □

**Remark.** We saw in the examples (and in the proof above) that the existence of dominance refinements depends on the number of lattice points on lines of the form $nx - my = r$. In the same way, the number of natural levels, *i.e.*, the number of distinct profiles, also depends on the number of lattice points on lines of the form $nx - my = r$. In fact, it follows from our lemma that the number of natural levels of an upper-tailed Smirnov test with sample sizes $m$ and $n$ equals $mn/\gcd(m, n)$. *E.g.*, if $m = n = 10$, then the test has 10 natural levels; whereas, if $m = 10$ and $n = 9$, then the test has 90 natural levels.

In our theorem, we only considered the upper-tailed Smirnov test based on $D^{+}_{mn} = \sup_t (F_m(t) - G_n(t))$. Of course, similar results exist for the Smirnov tests based on $D^{-}_{mn} = \sup_t (G_n(t) - F_m(t))$ or $D_{mn} = \sup_t |F_m(t) - G_n(t)|$.

**Acknowledgement.** We would like to thank John Einmahl for stimulating discussions on this paper.

## References

[1] Berge, C. (1971). *Principles of Combinatorics.* Academic Press, New York.

[2] Hájek, J. (1969). *A Course in Nonparametric Statistics.* Holden-Day, San Francisco.

[3] Narayana, T.V. (1975). Chaînes de Young et tests non-paramétriques. *Comp. Rend. Acad. Sci. Paris* 281, 1075-1076.

[4] Narayana, T.V. (1979). *Lattice Path Combinatorics with Statistical Applications.* University of Toronto Press, Toronto.

[5] Savage, I.R. (1956). Contributions to the theory of rank order statistics: the two-sample case. *Ann. Math. Statist.* 27, 590-615.