



Distribution of Minimum Distance Among N Random Points in d Dimensions

M. Fischler

April 29, 2002

Abstract

The Minimum Distance test in the DIEHARD suite for validating random number generators often indicates “false positives,” rejecting the quality of what is actually a good generator. A reason for this is presented: The test has enough sensitivity to detect the discrepancy between the approximate theoretical distribution of minimum distance used, and the actual distribution. We present next-order corrections to the theoretical expected distribution, for the 2-dimensional case used in DIEHARD and for higher dimensions. The corrected expectation will eliminate false positives arising from theoretical distribution discrepancies, at any practical test sensitivity level.

1 Motivation

1.1 The Minimum Distance Test

One of the tests in Marsaglia’s “DIEHARD” suite for pseudo-random number generators is a 2-dimensional minimum distance test: Plot a set of N points in the unit square, where each point has coordinates given by a consecutive pair of variates. Determine the minimum distance between any pair of these points. Repeat this procedure M times, to get M variates. Finally, compare (using a Kolmogorov-Smirnov or related test) the cumulative distribution function for these minimum distance variates with the theoretical expectation for the distribution of minimum distance.

Such a test has good resolution power for detecting clumping or banding in 2-dimensional space, since such clumping will almost surely skew the distribution of minimum distances toward the smaller distances. A generalization of the test to d -dimensional space gives a test for improper distribution in d -tuples of variates.

The observation, when applying the Minimum Distance test to generators in CLHEP which otherwise appear to be excellent (including some such as RanLux64, which can be proven to have excellent properties) is that these distributions fail the 2-dimensional Minimum Distance test for $N = 8000$ and $M = 100$.

At the time (1997) this was done, these failures were ascribed to one of three possible errors:

1. The original form of this test was coded in Fortran; the actual test applied was created using *f2c*. The original Fortran contained some “clever” index manipulation for efficiency, and it was thought that this manipulation might not have survived the transition to C, which uses a different array index ordering.

2. Some other aspect of coding, most likely the K-S test procedure itself, might have been incorrect.
3. We may not have understood what constitutes failure. In point of fact, Marsaglia mentions that quite good engines nonetheless have P-values like .9995, and that he would not reject without seeing numbers as bad as .99999. However, clearly this contradicts the definition we would like to use for P-value.

At any rate, these generators were, in the end, accepted as passing despite what should have been a nominal failure in this test.

1.2 Corrections to the Minimum Distance Distribution

In October 1998, we realized that the underlying theoretical expectation for the distribution of minimum distance was correct only to lowest order, and that this discrepancy would show up as apparent failures of distributions whenever the number of trials M was large. In particular, there was no hope of studying the apparent inconsistency by going to more statistics, since we would then *expect* the distribution to be “wrong” with respect to our (incorrect) theoretical distribution.

This paper provides a next-order correction to the approximate theoretical distribution used in the DIEHARD test, which will permit testing generators with significantly more resolution power and fewer false rejections.

We will compute the Cumulative Distribution Function for the square of the minimum distance s^2 between pairs of N points selected at random in a unit square with periodic boundary conditions:

$$P(s^2 < a^2) \tag{1}$$

for a given a .

In Marsaglia’s test, and according to Luscher’s explanation, the distribution is taken to be exponential with average a^2 value of $2/[\pi N(N-1)]$. The actual distribution differs from this, by a correction factor of $(1 + O(Na^2))$. The consequence of this error is that the number of trials which can validly be done with a fixed N to test randomness of a distribution is limited: For a sufficiently large number of trials, the sensitivity of the test becomes comparable to the error in the theoretical distribution, and even a good random generator will be rejected. Though the details are sensitive to the square of the coefficient of Na^2 in the actual distribution, this undesirable situation limits the sensitivity of the test. For $N = 8000$, which probes clustering on a scale of roughly 10^{-4} , one dare not exceed a few hundred trials without understanding and incorporating that correction coefficient.

On the other hand, if the error in the assumed distribution were of order N^2a^4 (or $1/N^2$) then you could confidently use up to a million trials; the sensitivity of the test becomes limited only by how much computation time you are willing to spend. This will be the case when using the corrected theoretical distribution.

To derive the distribution, imagine placing N balls of radius a in the square, and compute the probability that no center ever lies inside a ball already placed. At each step there is some area excluded from the area safe to place a ball.

The $O(Na^2)$ or $O(1/N)$ terms in the distribution come from two places. Firstly, the cumulative distribution involves a product of N decreasing terms each a bit smaller than 1; the exponential distribution is obtained if you add the logs of those terms and expand each log in the small distance away from 1. This discards terms of order N^3a^4 , which is $O(1/N)$. Secondly, that product assumes no overlap in excluded volume, and therefore

tends to underestimate the probability of large distances. The overlap terms again lead to an error of order $N^3 a^4$.

In the rest of this note, we will derive the distribution to first order “for practice,” then study the geometry of those overlaps, and use that to find the next order correction. Then we discuss the situation in higher dimensions. Lastly, a discussion of technique for finding the minimum distance variate is presented.

2 Lowest Order Calculation

Sometimes when two balls are placed, the amount of volume they cover is less than $2\pi a^2$ because though the center of each is outside the other, the two circles overlap. Ignoring this overlap, which leads to next-order corrections in Na^2 , we can immediately write the c.d.f.:

$$P(s^2 < a^2) = 1 - \prod_{k=1}^N (1 - (k-1)u) \quad (2)$$

where we will always designate $u \equiv \pi a^2$.

(To successfully place N balls of area u , you must survive about N chances to hit a filled area which is typically $N/2u$. This simple consideration argues that when $N^2u \gg 1$ the probability of succeeding is negligible. So we can always assume that N^2u is not much more than 1. The useful equivalent statement is that $Nu = O(1/N)$, which is small.)

Assuming N is even, we can group the first and N -th term, the second and $(N-1)$ -st term, and so forth. Multiplying two such terms gives a trinomial with a fixed coefficient of u :

$$P(s^2 < a^2) = 1 - \prod_{k=1}^{N/2} (1 - (N-1)u + (k-1)(N-k)u^2) \quad (3)$$

There are (at least) three ways of approximating this, all of which agree to zero-th order:

We can multiply out the polynomial, discarding terms whose N exponent is less than twice their u exponent. By considering terms with just a single $(N-1)u$ and the rest ones in the product, you get:

$$P(s^2 < a^2) \approx (N/2)(N-1)u \approx (N^2/2)u \quad (4)$$

This is correct until u becomes of order $(1/N^2)$; but it is not very useful because when u reaches that size, it exceeds 1—it is not a plausible form for a cumulative distribution function.

Instead, we can look at the log of each term. To first order in Nu we can ignore the expressions that include k , and are left with

$$\log(1 - P(s^2 < a^2)) \approx -(N/2)(N-1)u \quad (5)$$

$$P(s^2 < a^2) \approx 1 - e^{-(N/2)(N-1)u} \quad (6)$$

This is the exponential distribution that is used by Marsaglia, and it is a perfectly good distribution and matches the actual one with errors of order $1/N$.

Let’s consider a third way of doing the expansion, which will be extended when we do the higher order terms. Each term in the product can be written as $1 - (N-1)u$ times some correction factor which is one, to order N^2u^2 . As long as we are interested only

in $O(Nu)$ accuracy, we can ignore those corrections. (We could not ignore corrections of $O(Nu)$, because there are $N/2$ such contributions.)

$$P(s^2 < a^2) = 1 - (1 - (N - 1)u)^{N/2} \prod_{k=1}^{N/2} \left(1 + \frac{(k - 1)(N - k + 1)u^2}{1 - (N - 1)u} \right) \quad (7)$$

The product here is simply 1, with a correction of order Nu , so

$$P(s^2 < a^2) \approx 1 - (1 - (N - 1)u)^{N/2} \quad (8)$$

We stress that this is equivalent (within corrections of order $1/N$) to the exponential distribution derived above. But you need to pin the distribution off at $(N - 1)\pi a^2 = 1$. If it is uncomfortable to do this, then the exponential form is superior.

3 Overlap of Coverage

When going to the next order in Nu , two contributions must be considered. The first is that the terms involving k in equation 7 must be treated to first order in Nu . The second effect, discussed in this section, is that when two balls are placed the excluded volume is not always $2u$. Sometimes the centers are separated by a distance between a and $2a$, in which case the excluded volume is $2u - q$, where q is the overlap area. After placing k balls, the (expected) actual excluded area is $k(k - 1)/2 \langle q \rangle$ where $\langle q \rangle$ is the expected overlap of two randomly placed balls of area u .

(We shall treat the overlap as always being its expected value, even though it is slightly skewed (given success in the previous k placements) toward larger values. Correction of this, along with correction for triple-discounting any triple-overlap regions, and denominators correcting the expected value of overlap because the new ball does not have the entire volume to land in, would be $O(N^2 u^2)$ effects.)

To compute $\langle q \rangle$:

Consider circle O of radius a centered at point O , and circle P of the same radius centered at point P , such that $\overline{OP} = r$ with $a < r < 2a$. Label the points of intersection of the two circles X and Y , and the intersection of lines OP and XY as point Z . Also, label the intersection of circle P with line OP as point W . Then the overlap area in that diagram is composed of four small areas, each of which is congruent to the area bounded by Z, W, X . That area is given by the difference between the areas of circular wedge PWX and $\triangle PZX$.

Let $\angle ZPX$ be θ . Then:

$$\text{Area (wedge } PWX) = \frac{1}{2} a^2 \theta \quad (9)$$

$$\text{Area } (\triangle PZX) = \frac{1}{2} \overline{PZ} \overline{ZX} = \frac{1}{2} a^2 \cos \theta \sin \theta \quad (10)$$

$$\text{total overlap} = 2a^2(\theta - \cos \theta \sin \theta) \quad (11)$$

The expected overlap is the integral of this over all possible positions in the unit square of point P (since the boundary conditions are periodic, we can take O to be at the origin, which in turn we can take to be the center of the unit square). The radial distance of P from the origin is $r = 2a \cos \theta$. The overlap is zero unless $a < r < 2a$, which corresponds to $0 < \theta < \pi/3$. By axial symmetry, the area element corresponding to $d\theta$ is $2\pi r d\theta$. So we have:

$$\langle q \rangle = \int_{r=a}^{r=2a} 2\pi r d\theta 2a^2(\theta - \cos \theta \sin \theta) \quad (12)$$

or with $dr = -2a \sin \theta d\theta$,

$$\langle q \rangle = \int_{\theta=0}^{\pi/3} 16\pi a^4 (\theta \sin \theta \cos \theta - \cos^2 \theta \sin^2 \theta) d\theta \quad (13)$$

$$= 16\pi a^4 \int_{\theta=0}^{\pi/3} \left(\frac{1}{2} \theta \sin 2\theta - \frac{1}{4} \sin^2 2\theta \right) d\theta \quad (14)$$

$$= 8\pi a^4 \int_{\alpha=0}^{2\pi/3} \left(\frac{1}{4} \alpha \sin \alpha - \frac{1}{4} \sin^2 \alpha \right) d\alpha \quad (15)$$

$$= 2\pi a^4 \left[\sin \alpha - \alpha \cos \alpha + \frac{1}{2} \sin \alpha \cos \alpha - \frac{1}{2} \alpha \right]_0^{2\pi/3} \quad (16)$$

$$= 2\pi a^4 \left[\frac{\sqrt{3}}{2} + \frac{\pi}{3} - \frac{\sqrt{3}}{8} - \frac{\pi}{3} \right] = \frac{3\sqrt{3}}{4} \pi a^4 \approx 4.081 a^4 \quad (17)$$

That is, after k disjoint balls of radius a have been placed, the remaining open space for the next ball is (at average)

$$1 - ku + \frac{k(k-1)}{2} \frac{3\sqrt{3}}{4\pi} u^2 \quad (18)$$

$$= 1 - ku + Q \frac{k(k-1)}{2} u^2 \quad (19)$$

$$Q \equiv \frac{3\sqrt{3}}{4\pi} \approx 0.4135 \quad (20)$$

4 More Accurate Distribution of Minimum Distance

To take the distribution to the next order in Nu we need to include the effect of the overlap (but only to the lowest non-zero order) and we need to keep one more order of terms when expanding the product.

Including the overlap, equation 2 defining the probability becomes

$$P(s^2 < a^2) = 1 - \prod_{k=1}^N \left(1 - (k-1)u + \frac{(k-1)(k-2)}{2} Qu^2 \right) \quad (21)$$

Again we combine the first and last terms:

$$P(s^2 < a^2) = 1 - \prod_{k=1}^{N/2} \left\{ 1 - (N-1)u + \left[(k-1)(N-k) + \frac{Q}{2}(k-1)(k-2) + (N-k)(N-k+1) \right] u^2 + O(N^3 u^3) + O(N^4 u^4) \right\}$$

In those terms, we count a factor of k or of N as a power of N . Because of the sum up to $N/2$ they pick up another power of N . So the u^2 term contributes $O(N^3 u^2)$ which is $O(Nu)$; we must keep it. But the last two terms are of $O(N^4 u^3)$ and $O(N^5 u^4)$ which are respectively second and third order small in Nu so we can discard them.

Now (as in the expansion method that led to the geometric distribution to lowest order) we factor $1 - (N-1)u$ out of each pair of terms. The residue is of the form $1 + \frac{u^2 M(N,k)}{1 - (N-1)u}$.

That denominator can be treated as 1 since the u^2 term is already going to contribute $O(Nu)$ to the distribution. So the analogue of equation 7 is:

$$P(s^2 < a^2) = 1 - (1 - (N-1)u)^{N/2} \times \prod_{k=1}^{N/2} \left(1 + \left[(k-1)(N-k) + \frac{Q}{2} ((k-1)(k-2) + (N-k)(N-k+1)) \right] u^2 \right) \quad (22)$$

In the product here, a term with a factor of u^2 will be too small to matter unless it is multiplied by two powers of k and N . Such a term will contribute at $O(N^3 u^2) \approx O(Nu)$; terms involving just one power would contribute at $O(u)$ which for large N is negligible. Similarly, in formulas for $\sum k$ and $\sum k^2$, only the leading- N term need be kept. Thus the product, to this order, can be expressed as

$$1 + \sum_{k=1}^{N/2} \left(Nk - k^2 + \frac{Q}{2} (k^2 + N^2 - 2Nk + k^2) \right) u^2 \quad (23)$$

or

$$1 + \left(\frac{N^3}{8} - \frac{N^3}{24} + Q \left(\frac{N^3}{48} + \frac{N^3}{4} - \frac{N^3}{8} + \frac{N^3}{48} \right) \right) u^2 \quad (24)$$

Replacing u with πa^2 , we obtain the result:

$$P(s^2 < a^2) = 1 - (1 - (N-1)\pi a^2)^{N/2} \left(1 + \frac{\pi^2 N^3 a^4}{12} (1 + 2Q) \right) \quad (25)$$

and this is accurate neglecting terms smaller than $O(1/N)$.

This behaves reasonably for a cumulative distribution function, until pushed to large a ; when $(N-1)\pi a^2 > 1$ you have to pin the distribution at a c.d.f. of 1. Although the likelihood of a being anywhere near that big is infinitesimal, an exponential factor would be better. With some care, we can express this result in that manner. When Nu is small,

$$\prod_1^{N/2} (1 - (N-1)u) = 1 - \frac{N}{2}(N-1)u + \frac{1}{2} \frac{N}{2} \left(\frac{N}{2} - 1 \right) (N-1)^2 u^2 + \dots \quad (26)$$

while

$$e^{-\frac{N(N-1)}{2}u} = 1 - \frac{N}{2}(N-1)u + \frac{N^2(N-1)^2}{8} u^2 + \dots \quad (27)$$

We will need to keep terms of order $N^3 u^2$ and higher—and the $N^4 u^2$ terms match.

$$\prod_1^{N/2} (1 - (N-1)u) = e^{-\frac{N(N-1)}{2}u} \left(1 + \frac{N^3}{4} u^2 + \dots \right) \quad (28)$$

Finally, the cumulative distribution function, expressed as an exponential times a polynomial series and including corrections of order $1/N$, is:

$$P(s^2 < a^2) = 1 - e^{-\frac{N(N-1)}{2}\pi a^2} \left(1 + \frac{2+Q}{6} \pi^2 N^3 a^4 \right) \quad (29)$$

The sensible behavior of equation 29 when a grows large makes it very useful.

As to using the more exact distribution in the context of a minimum distance test for randomness: For each trial with minimum distance squared a^2 , instead of forming a variate from $1 - e^{-x a^2}$ you merely form that variate from $1 - e^{-x a^2} (1 + y a^4)$. The additional accuracy costs very little.

5 How Many Trials Can Be Done?

The Minimum Distance test of the quality of a random numbers generator probes whether there are any areas in the square which are more densely populated than they ought to be. N is taken to be large so as to probe this issue at a granularity of $1/N$, that is, (with $N = 8000$) if there were an area of length scale $1/50000$ with too high a density, the test would fail to perceive that.

The number of trials in the test determines the sensitivity to that fluctuation. For example, with one trial, you would not expect to detect a hint of an incorrect distribution unless the typical high-density area were ten times as dense as it should be. By this reasoning, one would like to use at least 2500 trials, to be sensitive to small-grain 20% density fluctuations. (Note that the test in DIEHARD uses only 100 trials, and would probably not detect a problem in a generator that had areas with 1.5 times the proper density.)

Now that we know the first order error in the exponential approximation to the true distribution, we can examine how many trials one could use before the deviation from the proper distribution might cause “false positives” declaring a good random engine as flawed. That is a danger if the assumed c.d.f. ever deviates from the actual c.d.f. by at least about $.1/\sqrt{T}$ where T is the number of trials done. That is, T trials probes the distribution with a sensitivity of $.1/\sqrt{T}$, and if we do not stray by more than a tenth on that scale, then the error in our approximate c.d.f. is completely unimportant.

The maximum deviation happens when $a^2 \approx 4/(\pi N^2)$ and the deviation is about $\frac{16(2+Q)}{6N}$. For $N = 8000$, this is .0007. This gives $T \approx 200$ so we are quite comfortable at 100 trials, but would be pushing things at a thousand, with a non-negligible risk of false rejections. And there would be a very high chance of false rejection were we to probe with as many as the desired 2500 trials.

On the other hand, with the corrected distribution in equation 29, the error will be down by approximately another factor of N . The coefficient of the next correction term is probably about three times that of the present one (based on the difference between the exponential and the product at that level—the other deviations are comparable in scale). But even taking that coefficient to be twenty instead of three, with the accurate c.d.f of equation 29, we should be completely safe taking up to twenty million trials. Practical time considerations limit our ability to probe beyond about a million trials of 8000 numbers; thus the present approximation is accurate enough that further work to extend it to third order is not warranted.

6 Higher Dimensions

The merit of the Minimum Distance test lies in the fact that it will detect any rank-deficiency with stripe-spacing bigger than the expected distance, and any clumping or fluctuations in the density on a scale larger than that distance. These are as useful in higher dimensions as in two. The only hurdle is that you may need more points in each trial to get to a given characteristic distance.

The issues we will discuss here are

- What is the expected distribution, to first order, for the d -dimensional case, with $d > 2$?
- How many trials are we likely to be able to do for d -dimensional points, and is the resulting accuracy good enough that the test is practical in that dimension?

- Will we need higher-order corrections in dimensions past two, or is the first order distribution adequate?
- If we need higher order corrections, what are the coefficients for each dimension?

6.1 Lowest Order Distribution in d Dimensions

If we stick to lowest order so that overlaps are irrelevant, the entire section 2 for computing the lowest order distribution function remains applicable, except that u , which is the volume of the sphere of radius a , is no longer πa^2 but instead is given by:

$$u = \frac{\pi^{d/2} a^d}{(d/2)!} \quad d \text{ even} \quad (30)$$

$$u = \frac{(2\pi)^{(d-1)/2} a^d}{d!!} \quad d \text{ odd} \quad (31)$$

The fact that a^d appears instead of a^2 merely implies that the lowest order distribution is an exponential when expressed as a distribution of s^d rather than s^2 .

Later when we can discuss how many points to use in each trial, it will be useful to know something about the characteristic distances probed by a given N value in d -dimensional space.

The characteristic minimum distance comes when $N(N-1)/2$ times that volume is roughly one. Using Stirling's approximation

$$d! \approx (d/e)^d \sqrt{2\pi d} \quad (32)$$

we find for the even cases (the odd cases will be quite similar) that

$$a \approx \sqrt[d]{2\sqrt{\pi d}} \sqrt{\frac{d}{2\pi e}} N^{-(2/d)} \quad (33)$$

This is a slowly-varying function of d (which stays between about 1.4 and 2 for $d < 16$) times $N^{-(2/d)}$.

6.2 Higher Order Corrections in d Dimensions

The reasoning that leads to a correction of the distribution given by equation (6) does not change when the expression for u in terms of a changes. The contributions still come from two sources: Overlaps of d -spheres, and expanding the product that leads to the exponential. Although geometry will dictate the size of the former correction, at the very least the latter is of the same form (and thus as important) in higher dimensions as it is in 2 dimensions.

In Section 3, one step was to compute the expected overlap $\langle q \rangle$ of two randomly placed balls of radius a , situated such that the center line and the line from either radius to any intersection point form an angle θ . This overlap volume is twice the difference between the volume of a wedge of a sphere of radius a , subtending a central angle θ and a right circular d -cone of side a and fulcrum half-angle θ (in 2 dimensions this is an ordinary triangle with base angle 2θ).

For example, in 2 dimensions, the cone is a triangle with a base of $2a \sin \theta$ and height of $a \cos \theta$, while the wedge has area $a^2 \theta$. Twice the difference is $2a^2(\theta - \cos \theta \sin \theta)$, as in

equation (11). We then multiply this by the area element at a radius of r , $2\pi r dr$, and integrate from $r = a$ to $r = 2a$ to obtain

$$\langle q \rangle_{d=2} = \frac{3\sqrt{3}}{4}\pi a^4 \quad (34)$$

We need to do this in higher dimensions; we will present the results for up to 5 dimensions. The technique pattern is the same for all cases: We find the base $d - 1$ volume of the cone, multiply by the height $a \cos \theta$ and divide by d to get the volume of the cone. We perform a trivial integral in d dimensional polar coordinates to get the volume of the wedge, and take twice the difference, which is the overlap for some fixed value of θ . We find the volume element in the r integration as the surface area of a d -sphere or radius r , and always have $dr = -2a \sin \theta d\theta$ to turn the overlap integral into an integral over θ from $\pi/3$ to 0. This gives $\langle q \rangle$. We finally find the correction coefficient Q by dividing u^2 where u is the volume of the d -sphere of radius a .

For $d = 3$:

$$\text{base} = \pi a^2 \sin^2 \theta \quad (35)$$

$$\text{cone} = \frac{\pi}{3} a^3 \cos \theta \sin^2 \theta \quad (36)$$

$$\text{wedge} = \int_0^a dr \int_0^\theta r d\vartheta \int_0^{2\pi} r \sin \theta d\phi \quad (37)$$

$$\text{wedge} = \frac{2\pi}{3} a^3 (1 - \cos \theta) \quad (38)$$

$$\text{overlap} = \frac{8\pi}{3} a^3 (2 + \cos \theta) \sin^4 \frac{\theta}{2} \quad (39)$$

$$\text{r-volume element} \rightarrow -32\pi a^3 \cos^2 \theta \sin \theta d\theta \quad (40)$$

$$\langle q \rangle = -\frac{256\pi^2}{3} a^6 \int_{\pi/3}^0 \cos^2 \theta (2 + \cos \theta) \sin^4 \frac{\theta}{2} \sin \theta d\theta \quad (41)$$

$$\langle q \rangle = \frac{17\pi^2}{18} a^6 \quad (42)$$

$$Q = \frac{17}{32} \approx .5312 \quad (43)$$

For $d = 4$:

$$\text{base} = \frac{4}{3} \pi a^3 \sin^3 \theta \quad (44)$$

$$\text{cone} = \frac{\pi}{3} a^4 \cos \theta \sin^3 \theta \quad (45)$$

$$\text{wedge} = \int_0^a dr \int_0^\theta r d\vartheta \int_0^\pi r \sin \theta d\phi \int_0^{2\pi} r \sin \theta \sin \phi d\psi \quad (46)$$

$$\text{wedge} = \pi a^4 \left(\frac{\theta}{2} - \frac{1}{4} \cos \theta \right) \quad (47)$$

$$\text{overlap} = \frac{\pi}{12} a^4 (12\theta - 8 \sin 2\theta + \sin 4\theta) \quad (48)$$

$$\text{r-volume element} \rightarrow -32\pi^2 a^4 \cos^3 \theta \sin \theta d\theta \quad (49)$$

$$\langle q \rangle = -\frac{8\pi^3}{3} a^8 \int_{\pi/3}^0 \cos^3 \theta \sin \theta (12\theta - 8 \sin 2\theta + \sin 4\theta) d\theta \quad (50)$$

$$\langle q \rangle = \frac{9\sqrt{3}}{32} \pi^3 a^8 \quad (51)$$

$$Q = \frac{9\sqrt{3}}{8\pi} \approx .6202 \quad (52)$$

For $d = 5$:

$$\text{base} = \frac{1}{2}\pi^2 a^4 \sin^4 \theta \quad (53)$$

$$\text{cone} = \frac{\pi^2}{10} a^5 \cos \theta \sin^4 \theta \quad (54)$$

$$\text{wedge} = \int_0^a dr \int_0^\theta r d\vartheta \int_0^\pi r \sin \theta d\phi \int_0^\pi r \sin \theta \sin \phi d\psi \int_0^{2\pi} r \sin \theta \sin \phi \sin \psi d\rho \quad (55)$$

$$\text{wedge} = \frac{2\pi^2}{5} a^5 \left(\frac{2}{3} - \frac{3}{4} \cos \theta + \frac{1}{12} \cos 3\theta \right) \quad (56)$$

$$\text{overlap} = \frac{4\pi^2}{15} a^5 (19 + 18 \cos \theta + 3 \cos 2\theta) \sin^6 \frac{\theta}{2} \quad (57)$$

$$\text{r-volume element} \rightarrow -\frac{128}{3} \pi^2 a^5 \cos^4 \theta \sin \theta d\theta \quad (58)$$

$$\langle q \rangle = -\frac{512\pi^4}{45} a^{10} \int_{\pi/3}^0 \cos^4 \theta (19 + 18 \cos \theta + 3 \cos 2\theta) \sin^6 \frac{\theta}{2} \sin \theta d\theta \quad (59)$$

$$\langle q \rangle = \frac{353}{3600} \pi^4 a^{10} \quad (60)$$

$$Q = \frac{353}{256} \approx 1.3789 \quad (61)$$

Although a slight trend toward increasing Q can be perceived, we see that this correction term is roughly the same magnitude for $d = 2$ through $d = 5$.

The rest of the reasoning in Section 4 holds intact, so that in d dimensions, the cumulative distribution function, expressed as an exponential times a polynomial series and including corrections of order $1/N$, is:

$$P(s^2 < a^2) = 1 - e^{-\frac{N(N-1)}{2}u(a)} \left(1 + \frac{2 + Q_d}{6} N^3 u(a)^2 \right) \quad (62)$$

where Q_d is the correction coefficient for d dimensions, and $u(a)$ is the volume of a d -sphere of radius a .

6.3 Are Higher Order Corrections Needed for Higher Dimensions?

At arbitrary d , the maximum deviation from the lowest order c.d.f. still comes at $u \approx 4/N^2$ and the deviation is about $\frac{16(2+Q_d)}{6N}$. So at first thought, the same argument that shows the first order correction to be important in d dimensions applies for higher d , with almost the same quantitative results (since $(2 + Q_d)$ does not change much).

However, a subtle factor does matter: Since a fixed value of N probes, for higher d , a coarser granularity in d -space (the granularity behaving like $N^{-2/d}$) a tester might well choose to use a slightly higher N for 3 dimensions than for 2, compensating for the extra time needed per trial by doing fewer trials. A prudent tester might trade off such that both the granularity and sensitivity degrade slightly, rather than taking the entire hit on granularity. If that is done, N increases, and thus the deviation of the lowest order approximation decreases as $1/N$. Also, if the number of trials decrease, then the acceptable deviation level rises.

This argument shows that the necessity of higher order corrections diminishes in higher dimensions. In particular, we can justify disregarding the corrections in, say, 6 or more dimensions. One might in fact be tempted to disregard the corrections for all $d > 2$ but since the cost of applying the correction is tiny (one the calculation of the coefficient Q_d has been done), we suggest correcting the distribution in all cases where Q_d is available.

7 Technique for Computing the Minimum Distance in d Dimensions

To evaluate the probative power of a minimum distance test in d dimensions we should know something about the time taken to do one trial of N points. This, in turn, requires some assumption about how the minimum distance would be found. The naive method would involve calculating the distance of each point to each other point; this requires $O(N^2)$ distance computations for a trial of N points.

For large N , a superior technique is to store all the d -tuples of variates, sort by first coordinate, then for each d -tuple, compute the distance to all the later d -tuples until the difference in first coordinate exceeds the minimum distance found thus far. Schematically:

```
double mindist = 1;
for (int i = 0; i < N; ++i) {
    dTuple p = v[i];
    for (int j = 0; j < N; ++j) {
        if ( v[j][0]-p[0] > mindist ) break;
        if ( distance ( p, v[j] ) < mindist ) mindist = distance ( p, v[j] );
    }
}
```

This technique takes $O(N \log N)$ for the sort, followed by the distance computation, which turns out (non-trivially) to be $O(N)$. The overall time is $O(N \log N)$, which is good.

A better technique for finding the minimum distance given N points in a trial might be to divide the unit hypercube into many smaller hypercubic regions. If the side of the box is greater than the minimum distance between the points, then the two minimally distant points must lie in the same box or in touching boxes. So if we choose boxes at least as large as the greatest possible minimum distance, then we need only check distances against the few points lying in the same box, plus, for points closer to some box side(s) (or corners) than the current minimum, against the points lying in the boxes neighboring along those sides. We can be sure to catch the actual minimum in that way.

Since we want the boxes to be small if possible, we would like to know a relatively small upper bound on the minimum distance. We can construct that by the following reasoning: If for a given set of points $\{P\}$ the minimum distance is a , then for any given point $p \in \{P\}$, we can assign to p all the volume inside a d -sphere of radius $a/2$ around p . With that assignment, no point in the unit cube can be in the d -spheres assigned to more than one point p . Now, assume we were to know that it is impossible to pack N d -spheres of radius ρ into a unit cube. Then if the minimum distance for any set $\{P\}$ of N points were greater than or equal to 2ρ , the assignment discussed above provides a packing of N d -spheres of radius at least ρ into the unit cube—contradicting the assumption. Thus 2ρ is an upper bound for the possible values of minimum distance.

So we could ask how tightly can we pack N d -spheres. Optimal sphere packing is a tough problem; but we easily can construct a ρ (though not a smallest possible such ρ) such

that it is plainly impossible to pack N d -spheres of radius ρ into a unit cube. Consider each such sphere to fully contain a small cube of diagonal 2ρ , of side $2\rho/\sqrt{d}$. If the cubes can't fit into the unit cube, then the spheres certainly can't. The number of small cubes of side s/\sqrt{d} that can be fit into the unit d -cube is $(\sqrt{d}/s)^d$, ignoring small boundary effects. This implies that the minimum distance can be at most $\sqrt{d}/\sqrt[d]{N}$. That is, if we choose the division to be into boxes of shorter than $\sqrt{d}/\sqrt[d]{N}$ in each dimension, then we are guaranteed that the two closest points lie in the same box or in touching boxes.

The technique derived from this reasoning would be to create a list for each of those boxes; by the time all N points are added most lists will have several points (because of our conservative choice for size) but not many. When a point is added, check the distances between it and the contents of its list and the lists for the appropriate neighbor boxes. This is a hash-table technique; it uses extra memory but its time-complexity grows linearly with N .

That said, the simpler first good method, which relies on sorting, may be good enough for the granularity and number of trials one would want. The loss of a factor of $\log N$ in the sort step may be compensated by the simplicity of this method, for any practical value of N .