

Distance to neighbours in a set of points on the sphere

prepared by: F. Mignard
affiliation : OCA
reference: GAIA-C4-TN-OCA-FM-035-1
issue: 1
revision: 1
date: 06/02/2008
status: Issued

Abstract

This note presents few results on the statistical distribution of the distances between a point and its closest, second or kth closest neighbour in a random and uniform distribution on a sphere and offers sample applications to small field astronomy. The correctness of the theoretical distributions is supported by comparisons to Monte-Carlo simulations.

Document History

Issue	Revision	Date	Author	Comment
1	0	06-02-2008	FM	Final version
1	0	24-01-2008	FM	First draft

Contents

1 Introduction **4**

2 The nearest neighbour **4**

3 Distribution of the *k*th neighbour **8**

4 Number of background stars around a bright target **14**

5 Number of optical binary stars **16**

1 Introduction

In small field astronomy it is commonplace to consider that the stars have a uniform spatial distribution and that basic statistical properties regarding the background light or the number of chance alignments between two sources can be assessed with probabilistic method. In most case this amounts to looking for the probable number of stars in the immediate vicinity of a target direction. As soon as one wishes to go beyond a very close neighbour (very close means here, much closer than the average separation between two stars, a quantity of the order of $1/\rho^{1/2}$ where ρ is the local surface density) great care must be exercised to count properly the more distant neighbours to draw valuable conclusions from statistical inference. This is the purpose of this note to derive the probability distribution for the closest, second and kth neighbours within a simple framework of uniform and random distribution of points on a sphere or on a plane normal to the line of sight and to illustrate their properties on few astronomically oriented applications. While nothing is totally new here, it is brought into a consistent and unique mathematical framework useful for further discussions.

2 The nearest neighbour

In the following I try to use (consistently I hope) the word *point* to refer to a mathematical point in space or on a surface and the word *source* when dealing with an entry selected in the random drawing of points, usually on a surface. This distinction is useful to avoid language ambiguity, in particular when the context is not sufficient to make the distinction. A source is then an element of a finite data set of points.

Let $\mathscr{A}$ be the area where N sources are uniformly scattered. The expected number of sources in any area S within $\mathscr{A}$ is asymptotically proportional to S as $N \to \infty$. The area $\mathscr{A}$ could be a finite subset of a plane or a unit sphere, in which case one has $\mathscr{A} = 4\pi$ sr or $42,300$ square degrees. The number of sources per unit of surface (surface density) is then $\rho = N/\mathscr{A}$.

For a uniform density, the actual number of sources to be found in a particular area S in a random drawing is given by a Poisson law (see the Appendix below for the conditions of applicability), that is to say the probability that there are exactly k points, $k = 0,1,2,\cdots$ in an area S is given by,

$$\mathscr{P}(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \tag{1}$$

where $\lambda = \rho S$ is the mean number of sources expected in the area S. Two different proofs of this important and widely used result are given in the Appendix at the end of this note.

Let A be a point located at the center of a circular area $S = \pi r^2$. A could be a pure reference abstract mathematical point on the surface never drawn in the random population or could be connected to a source located there. The only important issue is that the small disk associated is deterministically determined by the location of the point and not by either a random process or a *a priori* selection independent of the source distribution.

Now if the distance D to the nearest neighbour of the reference point is less than r this means that there is at least one source in the area πr^2. Given the Poisson distribution, the probability that there is no source in an area S is

$$\mathscr{P}(k=0) = e^{-\lambda}, \tag{2}$$

and the probability that there is at least one source is $1-\exp(-\lambda)$. Therefore the probability that the distance D to the closest neighbour is less than r is given by,

$$\mathscr{P}(D<r) = 1 - e^{-\pi\rho r^2}, \tag{3}$$

which leads to the PDF for the distance to the nearest neighbour as,

$$f(r) = 2\pi\rho\, r\, e^{-\pi\rho r^2}. \tag{4}$$

For the following it is convenient to introduce a characteristic distance

$$\sigma = 1/(2\pi\rho)^{1/2}, \tag{5}$$

typically the average distance between a source and its nearest neighbour in the data set (nearest source). $\pi\sigma^2$ is also the area in which one expects to have about one source around a reference point (or around another source located at the center of the disk). Then the PDF of the random variable **D** becomes,

$$\boxed{f(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}}} \tag{6}$$

which happens to be a Rayleigh distribution.

Remark: Although a spherical distribution is not excluded in the above derivation, I have made the approximation that the density is large enough to consider that the distances involved are such that $r \ll$ the sphere radius. When the density is low, the area must be expressed with the angular distance on a great circle and the area of a disk of radius θ becomes $2\pi(1-\cos\theta)$. But at this stage the validity of the Poisson approximation is no longer guaranteed and one must

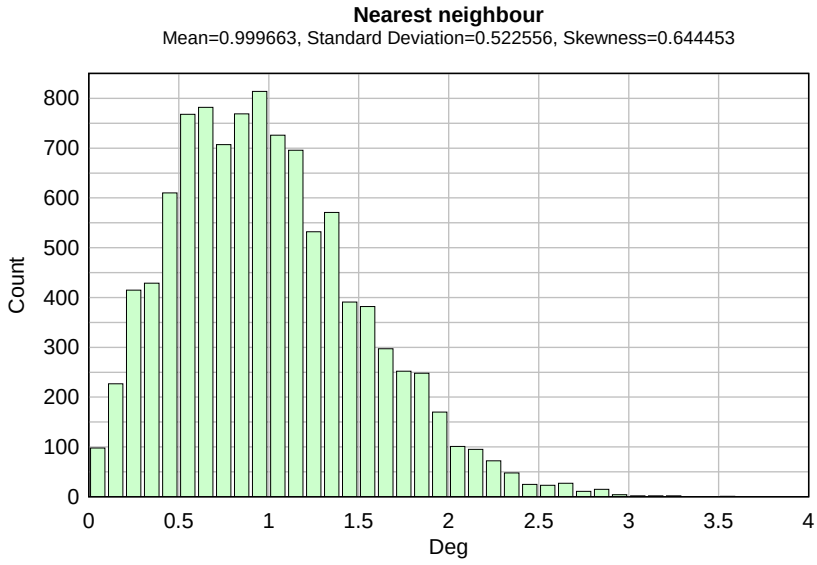


Figure 1: *Empirical distribution of the distances to the closest neighbours from a uniform distribution of* $10,300$ *points on a unit sphere. The distribution is a Rayleigh distribution with mean equal to 1.0*

return one step back and use the Binomial distribution as explained in the Appendix. If one chooses 10 points at random on a sphere, the distribution of the number of points within one hemisphere (selected in advance or at random after the points) is the same as the results of the number of heads in a coin tossing experiment and not amenable to a Poisson approximation.

In the usual situation it is not too difficult to find the main moments (Expectation, Variance and Skewness coefficient) by computing the integrals,

$$\mathbb{E}(\mathbf{D}) = \int_0^\infty r f(r) dr \tag{7}$$

$$\mathbb{V}(\mathbf{D}) = \int (r - \mathbb{E}(\mathbf{D}))^2 f(r) dr \tag{8}$$

$$\mathbb{S}(\mathbf{D}) = \int (r - \mathbb{E}(\mathbf{D}))^3 f(r) dr / (\mathbb{V}(\mathbf{D}))^{3/2} \tag{9}$$

This leads to,

$$\mathbb{E}(\mathbf{D}) = \left(\frac{\pi}{2}\right)^{1/2} \sigma, \tag{10}$$

$$\mathbb{V}(\mathbf{D}) = \frac{4-\pi}{2}\sigma^2, \tag{11}$$

$$\mathbb{S}(\mathbf{D}) = \frac{2\sqrt{\pi}(\pi-3)}{(4-\pi)^{3/2}}. \tag{12}$$

A simulation has been set up to generate a uniform distribution of n points on the unit-sphere. For each realisation of this distribution one computes the distance of each source to its nearest, second nearest, k-th nearest neighbour and determines the empirical moments of the distribution. The program is based on (1) and the associated fortran library available on-line in Stripack. It allows to create the Voronoï tessellation of a spherical distribution of sources and afterwards the Delaunay triangulation. It is easy then to add few modules for locating and sorting the neighbours of a source and later to investigate the statistical properties of these distances.

In the experiment described here I took $n = 10,313$, giving $\rho = 0.25$ point per square degree or $\sigma = 0.7979 \simeq 0.8$ degree. We expect that each source is at the center of a cell of area ~ 4 square degrees, and that the typical distance between nearest neighbour is ~ 2 degree. The Corresponding Voronoï diagram is shown in Fig. 2 and the associated Delaunay triangulation in Fig. 4. The distribution of the cell areas is shown in Fig. 3 with the values of the first moments on top of the plot.

From Eqs. 10-12 one should have a mean distance to the nearest neighbour of 1 degree, a standard deviation of 0.524 degree and a skewness of 0.618. The observed distribution of the nearest neighbours is shown in Fig. 1 and the empirical moments are in perfect agreement with the theoretical values. The overall shape matches also the Rayleigh distribution of same mean. It is important to notice that the initial guess regarding the typical separation was given by $1/\rho^{1/2}$ = 2 degrees whereas the average is twice as small. In addition from the Rayleigh distribution one can see that occurrences of distances to the closest neighbour as small as 0.25 times the mean are not exceptional. Therefore it is reasonable when dealing with numerous sources to quantify the likeliness of a perturber by taking a safety circle as small as $0.1/\rho^{1/2}$.

As an illustration, if we consider that at any particular time, the 300,000 minor planets observable by Gaia are uniformly distributed in a spherical domain within ± 15 deg. of ecliptic latitude, this gives a density of $\rho = 30$ objects per deg^2. There are 5 percent of the planets which have another planet within 30 arcsec, whereas $1/\rho^{1/2} \sim 0.2$ deg.

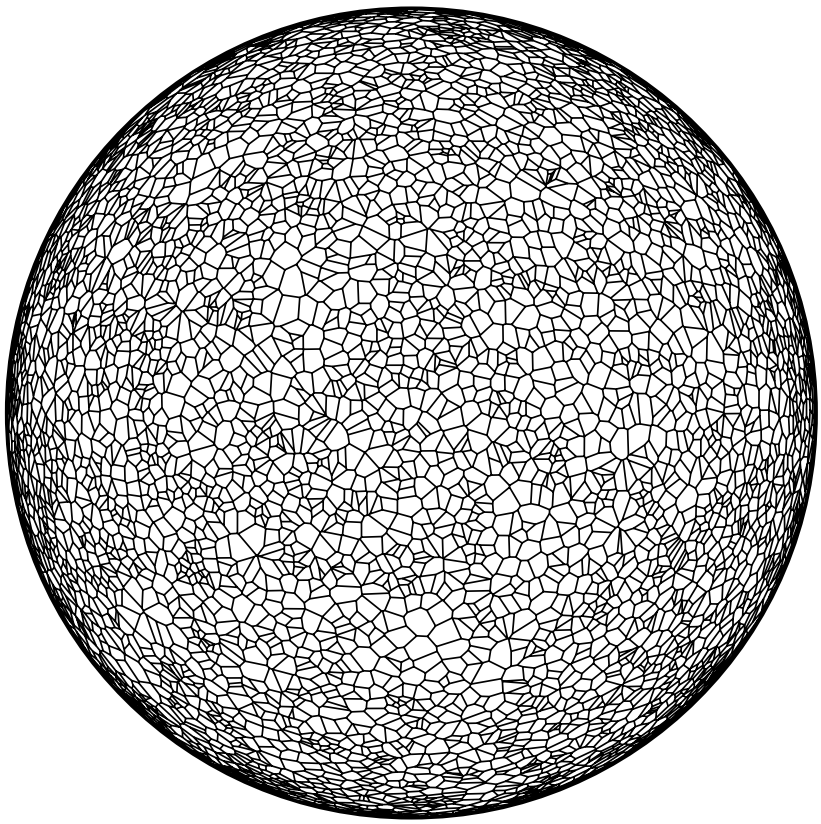


Figure 2: *Voronoï tessellation for a set of* $10,300$ *sources uniformly and randomly distributed on a unit sphere. Each cell contains one source of the data set in its interior and a cell is defined by the set of points of the sphere which are closer to the source in the cell than to any other source of the data set.*

3 Distribution of the kth neighbour

One can extend the method used to derive the probability distribution of the distance to the nearest neighbour to the second, third or further neighbours. Each source is surrounded by other sources and the distances between the reference source and its neighbours can be ranked in increasing order. The smallest distance corresponds to the nearest neighbour, the second smallest to the second neighbour, and more generally one defines the kth neighbour as the point whose distance to the reference source is the kth in the sorted set of distances. The probability that the distance to the kth neighbour is less than r is the same as the probability that there are at least k sources around the reference sources at a distance $d < r$. This is also given by

$$1 - P(n \leq k-1)$$

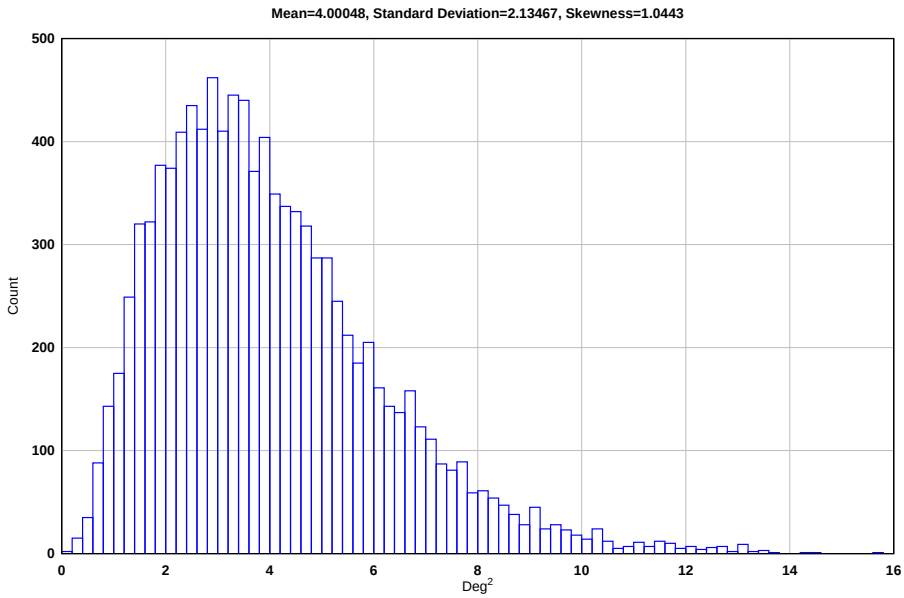


Figure 3: *Histogram giving of the size (in deg^2) of the cells in the Voronoï tessellation of a set of* $10,300$ *sources uniformly and randomly distributed on a unit sphere.*

or equivalently,

$$1-[P(n=0)+P(n=1)+\cdots+P(n=k-1)]$$

and explicitly,

$$P(D_k<r)=1-e^{-\pi\rho r^2}\left(1+\pi\rho r^2+\frac{(\pi\rho r^2)^2}{2!}+\cdots+\frac{(\pi\rho r^2)^{k-1}}{(k-1)!}\right) \tag{13}$$

and this yields the PDF for the distribution of the kth neighbour,

$$f_k(D_k=r)=2\pi\rho r\frac{(\pi\rho r^2)^{k-1}}{(k-1)!}e^{-\pi\rho r^2} \tag{14}$$

and by using σ as before,

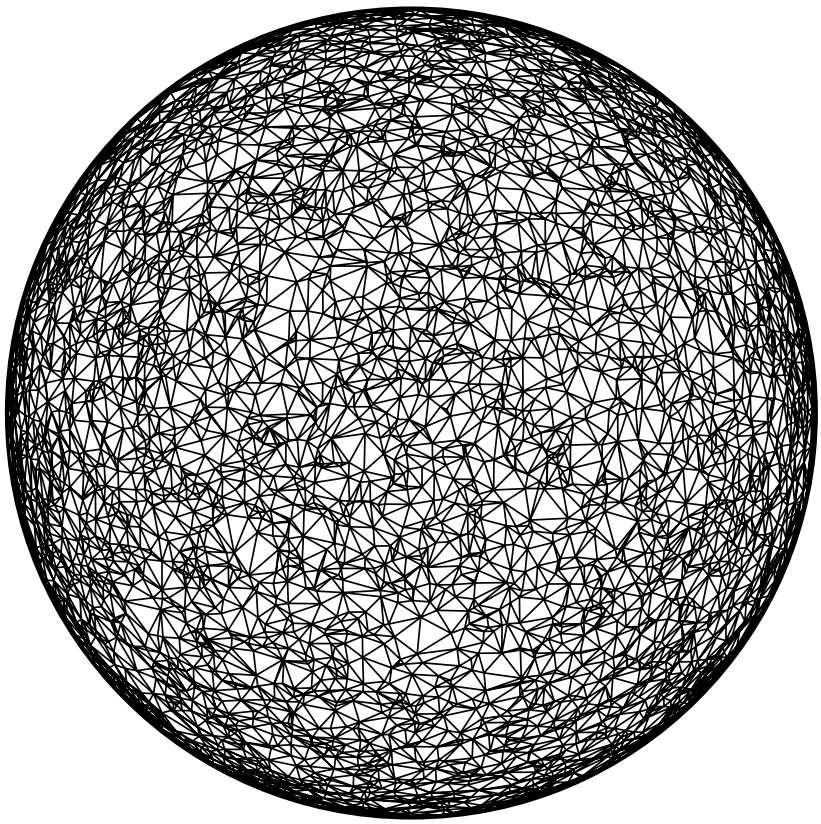


Figure 4: *Delaunay triangulation for a set of* $10,300$ *sources uniformly and randomly distributed on a unit sphere. The sources are at the vertices of the triangles and the number of edges originating at a vertex is equal to the number of sides of the Voronoï cell of the vertex source.*

$$\boxed{f_k(D_k = r) = \frac{1}{2^{k-1}\,(k-1)!\,\sigma}\left(\frac{r}{\sigma}\right)^{2k-1} e^{-\frac{r^2}{2\sigma^2}}} \tag{15}$$

Using $u = r^2/\sigma^2$ this reduces to the distribution for u given by,

$$f_k(u) = \frac{1}{2^k\,(k-1)!}\,u^{k-1}\,e^{-\frac{u}{2}} \tag{16}$$

which is the same as a χ^2_{2k} with $2k$ degrees of freedom.

Using Eqs. 7-9 with 15 on obtains the following expressions for the mean and variance:

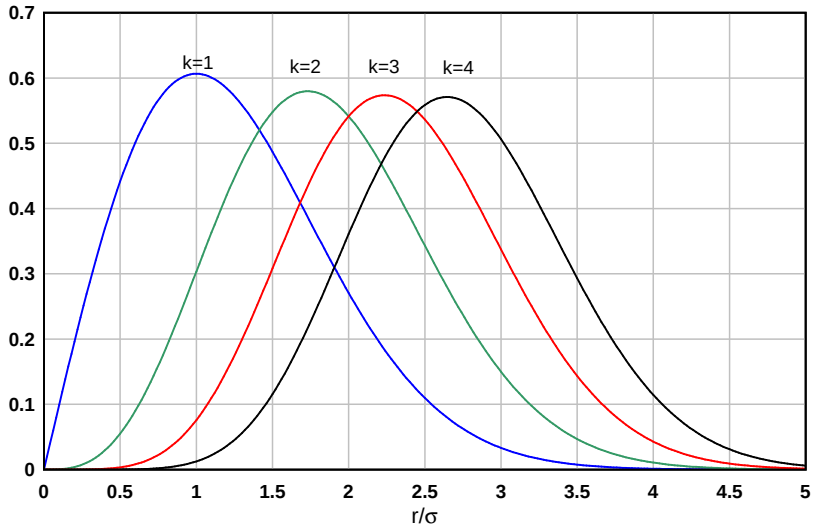


Figure 5: *Probability distribution of the kth neighbour as a function of the dimensionless quantity r/σ. The tendency toward a Gaussian shape is quickly visible even for small k.*

$$\mathbb{E}(\mathbf{D}_k) = \frac{\sqrt{2}\sigma}{(k-1)!}\Gamma(k+\frac{1}{2}) \tag{17}$$

$$\mathbb{V}(\mathbf{D}_k) = \left[2k - \frac{2\left(\Gamma(k+\frac{1}{2})\right)^2}{(k-1)!^2}\right]\sigma^2 \tag{18}$$

and a very complicated expression not shown here for the skewness. This can be specialised to the cases of interest as:

Second closest neighbour:

$$\mathbb{E}(\mathbf{D}_2) = \frac{3}{2}\left(\frac{\pi}{2}\right)^{1/2}\sigma \tag{19}$$

$$\mathbb{V}(\mathbf{D}_2) = \frac{32-9\pi}{8}\sigma^2 \tag{20}$$

$$\mathbb{S}(\mathbf{D}_2) = \frac{6\sqrt{\pi}(9\pi-28)}{(32-9\pi)^{3/2}} \tag{21}$$

Third closest neighbour:

$$\mathbb{E}(\mathbf{D}_3) = \frac{15}{8}\left(\frac{\pi}{2}\right)^{1/2}\sigma \tag{22}$$
$$\mathbb{V}(\mathbf{D}_3) = \frac{768-225\pi}{128}\sigma^2 \tag{23}$$
$$\mathbb{S}(\mathbf{D}_3) = \frac{10\sqrt{\pi}(225\pi-704)}{(768-225\pi)^{3/2}} \tag{24}$$

Fourth closest neighbour:

$$\mathbb{E}(\mathbf{D}_4) = \frac{35}{16}\left(\frac{\pi}{2}\right)^{1/2}\sigma \tag{25}$$
$$\mathbb{V}(\mathbf{D}_4) = \frac{4096-1225\pi}{512}\sigma^2 \tag{26}$$
$$\mathbb{S}(\mathbf{D}_4) = \frac{350\sqrt{\pi}(245\pi-768)}{(4096-1225\pi)^{3/2}} \tag{27}$$

and the general form for the kth closest neighbour,

$$\mathbb{E}(\mathbf{D}_k) = \frac{(2k-1)!!}{2^{k-1}(k-1)!}\left(\frac{\pi}{2}\right)^{1/2}\sigma \tag{28}$$
$$\mathbb{V}(\mathbf{D}_k) = \left[2k-\left(\frac{(2k-1)!!}{(k-1)!}\right)^2\frac{\pi}{2^{2k-1}}\right]\sigma^2 \tag{29}$$

Using

$$(2k-1)!! = \frac{1}{2^k}\frac{(2k)!}{k!} \tag{30}$$

and the Stirling formula

$$n! \approx \sqrt{2\pi n}\frac{n^n}{e^n}\left(1+\frac{1}{12n}+\cdots\right) \tag{31}$$

(where the term in $1/12n$ is needed for the variance) one has after some algebra for $k \gg 1$,

$$\mathbb{E}(\mathbf{D}_k) \approx \sqrt{2k}\,\sigma \tag{32}$$
$$\mathbb{V}(\mathbf{D}_k) \approx \sigma^2/2. \tag{33}$$

This result could also have been deduced from the well known moments of the χ^2 distribution. From the properties of the χ^2 distribution, the probability distribution of the kth neighbour is asymptotically normal with the above mean and variance. One finds that the average distance of the kth closest neighbour increases as the square root of k. This also could have been hinted, since within a disk of radius D_k and with $k \gg 1$ one should find k sources (the first k neighbours to the central point, whether there is a source or not). Therefore setting $\rho\,\pi D_k^2 = k$ this leads immediately with (5) to the same as (32). One should also notice that the standard deviation gets constant with large k.

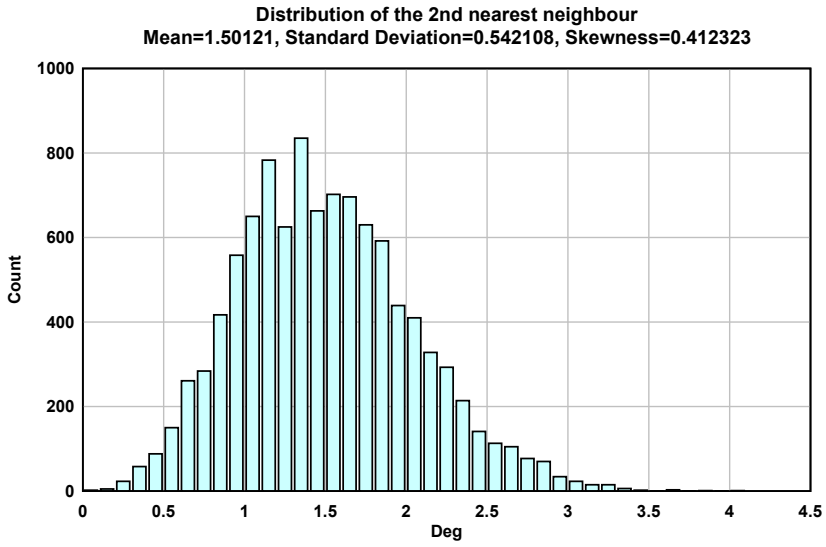


Figure 6: *Empirical distribution of the distances to the 2nd closest neighbours from a uniform distribution of* $10,300$ *points on a unit sphere. The values of the first moments agree fully with the expected values from Eqs. 19- 21.*

Table 1: *Comparison of the theoretical moments for the distribution of the kth closest neighbour to the empirical values found in a simulation with 10,313 points uniformly distributed on a sphere. The near perfect agreement between the expected and observed values supports the overall correctness of the probability approach.*

	$\mathbb{E}(\mathbf{D}_k)$		$(\mathbb{V}(\mathbf{D}_k))^{1/2}$		$\mathbb{S}(\mathbf{D}_k)$	
k	Comp.	Obs.	Comp.	Obs.	Comp.	Obs.
1	1.000	0.9996	0.523	0.522	0.631	0.644
2	1.500	1.5012	0.544	0.542	0.406	0.412
3	1.875	1.8731	0.551	0.547	0.318	0.331
4	2.188	2.1848	0.555	0.551	0.269	0.287

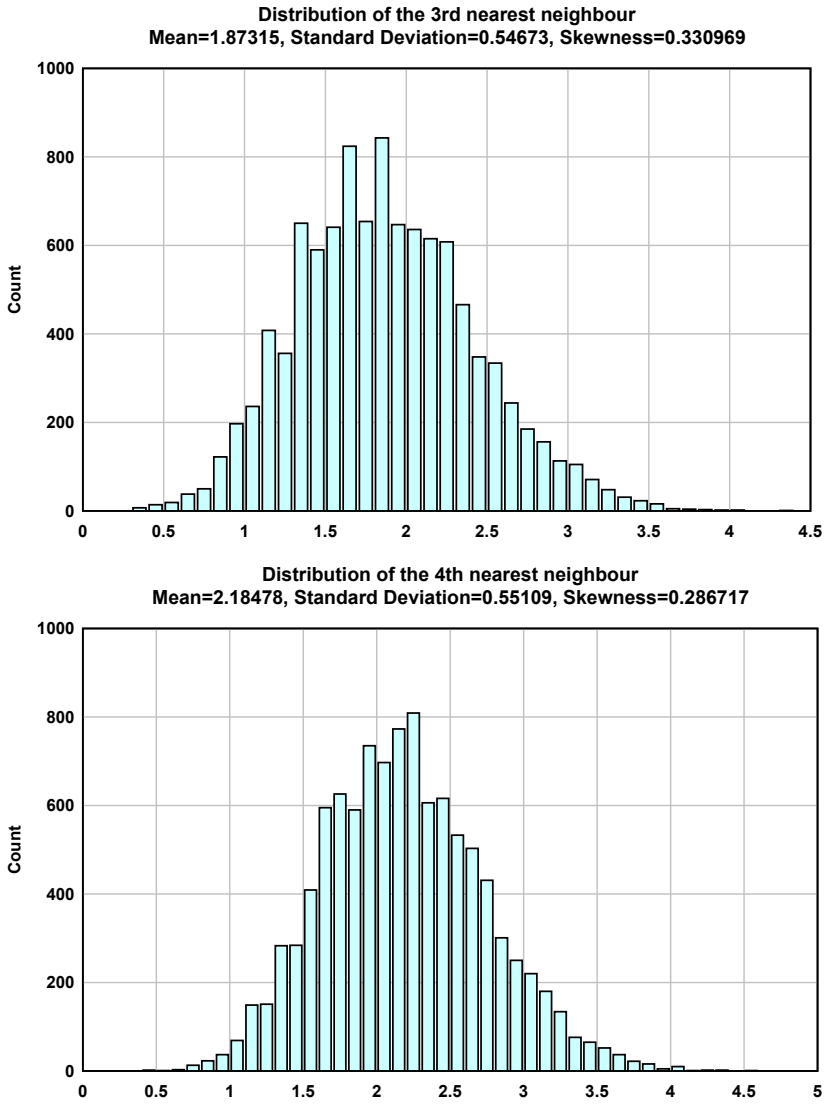


Figure 7: *Empirical distribution of the distances to the 3rd (top) and 4th (bottom) closest neighbours from a uniform distribution of* $10,300$ *points on a unit sphere.*

4 Number of background stars around a bright target

While the celestial sphere is a far cry from a uniform distribution, this assumption remains acceptable locally as soon as faint enough stars are included and provided one works on an area wide enough for the idea of density to become meaningful. The modeling with a sphere is mathematically very useful (besides its aesthetic value) to avoid the boundary effects, but the results remain applicable on any small area within a larger region on the sphere on which the surface density is defined. This is clearly the case with Gaia when one considers the occurrence of perturbing faint stars around a pointed target in a small field. In a presentation at the 4th

meeting of the DPAC CU5, A. Brown (2) has used a very simple estimate for the probability of having at least one background source within an area S around a target. With the notation adopted here, Brown's expression is the same as,

$$P(n \geq 1) = \frac{1 - e^{-\frac{r^2}{2\sigma^2}} - \frac{r^2}{2\sigma^2} e^{-\frac{r^2}{2\sigma^2}}}{1 - e^{-\frac{r^2}{2\sigma^2}}} \tag{34}$$

This is essentially the expression of a conditional probability of having at least two sources within the area S with the condition that there is at least one (the target), or mathematically $P(n \geq 2 | n \geq 1)$. The idea is clear: I have a target which belongs to the distribution and this must be accounted for when evaluating the number of cases where a second source is in the vicinity of the target. However this is not equivalent to the probability of having at least one star around a target, since the latter is just a reference source not selected at random. One considers a target, or equivalently a direction on the sky (with or without a source, this is not the problem) and one asks ourselves about the probability of having a source within a certain distance of that target direction. In term of frequency, Eq. 34 is the answer to the following question: given a set of sources on a sphere, one covers the sphere with non-overlapping tiles of area S. Some tiles are empty areas (empty fields for the astronomer), other have one, two or more data points. Among those having at least one source, what is the number of tiles with at least a second source? This is precisely the probability expressed in Eq. 34. The probability function is somewhat re-scaled to allow for the fact the empty tiles do not belong to the probability space anymore.

The actual problem amounts to something different: around each data point (the targets), I consider a tile of area S (say a disk centered at the target source) and ask for the number of cases where another source is in the tile. To be perfectly correct, one can decouple the distribution of the targets (the bright stars) and that of the perturbers (the faint stars compared to the target). For the problem to be meaningful, the target must be much less numerous so that the tile may contain a perturber, but not another potential target. Therefore the probability of having a background star is the same as having a nearest neighbour to the target direction within a disk of area S centered at the target source. This is given by Eq. 3 as,

$$\boxed{P(n \geq 1) = 1 - e^{-\frac{r^2}{2\sigma^2}}} \tag{35}$$

In the low probability regime (meaning uncrowded fields with few perturbers, if any, within the area of disturbance) the latter probability is about twice as large as (34). The two equations give the same result as soon as one can expect several background stars in the area under consideration; this is expected since the influence of the infrequent empty fields becomes negligible.

Combining the probability framework for the nearest or farther neighbours with the stellar dis-

tribution per range of magnitude should be the starting point to investigate the statistics of the stellar environment of a target source.

5 Number of optical binary stars

The use of probabilistic arguments to support the physical association between two angularly close stars dates back to J. Michell in 1767 and 1784. The chance grouping of stars along the line of sight appeared to him very improbable thus indicating that these stars were in fact physical pairs in space. The argument is simple, even though Michell used a rather convoluted mathematical approach to reach his conclusion.

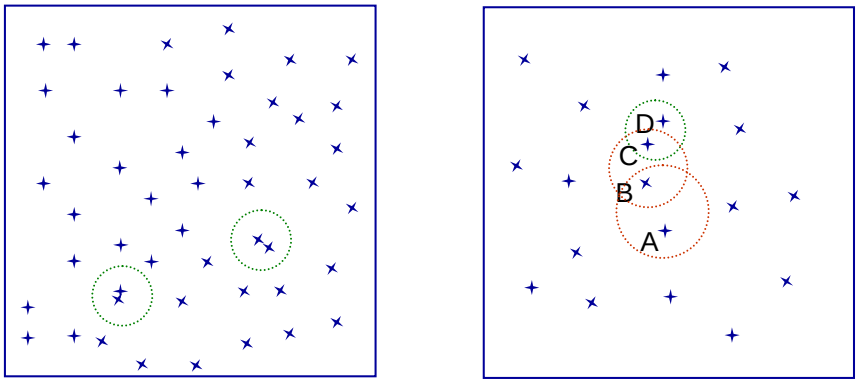


Figure 8: *Illustration of the concept of reciprocal pairs compared to the chained pairs. Left: two reciprocal pairs, where the two components (called A and B, but with no label shown) are mutual closest neighbours; During the systematic exploration of the sources this pair is counted twice, with A first for the reference source, then later with B. Right: B is the closest neighbour of A, and AB is detected as a pair when the reference source is A. However when B becomes the reference, its closest neighbour is C and BC is the new pair. Finally CD is a reciprocal pair, and this always happens at the end of a chain of successive closest neighbours which must terminate by a pair of mutual neighbours. At the opposite there is a first element in a chain that has no father but just a son. The whole set of sources on a sphere can be partitioned into disconnected chains of different lengths. The chains whose length is 2 are precisely the isolated reciprocal pairs.*

Given N sources uniformly distributed on the celestial sphere, how many pairs are found with angular separation less than r? A quick answer is given by considering that the probability to have a star in a disk of radius r centered at one point is $\rho\,\pi\,r^2$ if $r \ll \sigma$, that is to say smaller than the typical distance between two sources. Therefore the expected number of stars found with one close companion is $N\rho\,\pi\,r^2$ and since there are two stars in a pair, the number of pairs is $N\rho\,\pi\,r^2/2 = Nr^2/4\sigma^2$ (with Eq. 5). This is essentially the quantitative argument set forth by

Struve in 1852 to estimate the number of optical pairs resulting from chance alignment.

There are several hidden assumptions in this computation:

- A pair is defined by the fact that, given a reference source, one object is found at a distance less than r from that source. But there may be several sources meeting this criterion, and not all of them have been counted by the above simple counting method. What has been done is in fact searching the pairs by looking only at the closest neighbour. If the second nearest, or more, neighbours are at a distance less than r from the reference, they should be included as additional pairs. In this case the problem is much more complex.
- When I state that the pairs are counted twice (so the division by two) I consider that if B is found to be the closest neighbour of A, then A is the closest neighbour of B. Let's call these pairs the *reciprocal pairs* and they are the rule when $r \ll \sigma$, but not in general when $r \simeq \sigma$. A reciprocal pair is then a set of two sources close enough to each other so that the pair appears as an isolated system among the general distribution. As soon as one looks over areas around a source wide enough to contain at least one or two additional sources on the average, the pairs are no longer reciprocal and a more accurate definition must be used. The natural solution is to consider the nearest neighbour and to see how many are within a given distance from a given point.

Let's try to derive a more general expression with a wider range of validity. I consider only the definition of a pair based on the concept of closest neighbour. The probability that a source is within the distance r from the reference is given by Eq. 3. Now there is a certain fraction f of reciprocal pairs AB, where then A and B are mutually their closest neighbours. Hence there is the faction $1-f$ of pairs AB for which if B is the closest neighbour of A, there is a point C closer to B than A is. The pairs of the first group are counted twice when one goes over the whole population while that of the second only once. Therefore the expected number of pairs is,

$$N\left(1-e^{-\frac{r^2}{2\sigma^2}}\right)(f/2+(1-f)) \tag{36}$$

It remains to find the fraction f. Let B be the closest neighbour of A at a distance r; the probability that A is also the closest neighbour of B is equal to the probability that the closest neighbour of B different from A is at a distance $> r$. Therefore from Eq. 3 $f = \exp(-r^2/2\sigma^2)$ and the number of pairs

$$\boxed{N\left(1-e^{-\frac{r^2}{2\sigma^2}}\right)\left(1-\frac{1}{2}e^{-\frac{r^2}{2\sigma^2}}\right)} \tag{37}$$

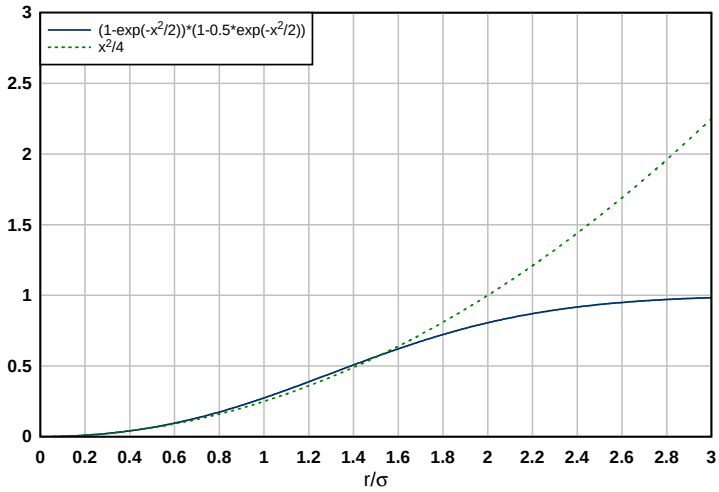


Figure 9: *Comparison between the standard estimation of number of optical pairs by the Struve approximation (dash line) and the counts based on the nearest neighbours of Eq. 37 (solid line).*

For $r \ll \sigma$, that is to say when only isolated reciprocal pairs are counted one recovers Struve relationship $Nr^2/4\sigma^2$ for the number of optical pairs. Equation (37) is more general and has the advantage also to draw the attention on the need for a better definition of an optical pair, when the separations are no longer very small compared to the average distance between sources. For large distances, Struve's formula becomes quickly meaningless with more pairs than sources, while the general formula approaches N when every sources has its nearest neighbour within the target area. Then every source belongs to a pair. This is shown in Fig. 9, where one can also notice that the range of validity of the Struve's formula is wider than one could have expected from its elementary derivation.

Acknowledgment

A. Bijaoui is gratefully acknowledged for his relevant comments on a preparatory version of this note.

Appendix: Poisson distribution

It has been taken for granted that the number of points falling into a small area of a sphere covered with a uniform and random distribution of points should follow a Poisson distribution. This is a very important result which rests upon few assumptions spelt out below. The derivations are classical and can be found in most probability textbooks, but not specifically in the context of spatial distribution.

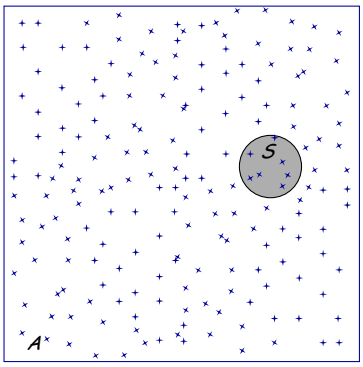


Figure 10: *The number of points found in a small area S within a subset of a plane of area $\mathscr{A}$ filled with N points randomly and uniformly distributed follows a Poisson distribution of parameter $NS/\mathscr{A}$.*

Derivation from Bernoulli process

As illustrated in Fig. 10, for a random and uniform distribution of sources over an area $\mathscr{A}$, the probability that a source falls onto the area S is just $p = S/\mathscr{A}$. This is the simple translation of the equidistribution definition which states that if N points are equidistributed on an area $\mathscr{A}$ the number of points n_S in any area $S \subset \mathscr{A}$ is such that $\lim n_S = NS/\mathscr{A}$ as $N \to \infty$. In the plane, uniform and random distributions are not the only ones to meet this criterion; this is also the case of regular lattice distributions with points at the intersections of two regular networks of straight lines. On the sphere, if we exclude the vertices of the five Platonic solids, there is nothing equivalent. However even in this case there is not a total equivalence between random uniform distribution and equidistribution of points. In the following I restrict to the random distribution. There are non-random algorithms (3) that can generate spherical distributions of points somewhat more *regular*, although without symmetry, than the random distribution considered hereafter, and allowing to bound the error when computing an integral on the sphere with a Monte Carlo method. Some good physical approximations would be the equilibrium distribution of N electric charges free to move on the sphere. A significant literature exists on this problem under the name of the Tammes problem.

The random experiment regarding the outcome of where one source falls is a Bernoulli trial with a probability of success p. Now the probability distribution $P(r|N)$ of obtaining exactly r

successes out of N Bernoulli trials is given by a binomial distribution,

$$P(r|N) = \binom{N}{r} p^r (1-p)^{N-r} \tag{38}$$

Using $\lambda = NS/A$ for the expected number of sources ending in S after N trials, one has

$$P(r|N) = \binom{N}{r} \left(\frac{\lambda}{N}\right)^r \left(1 - \frac{\lambda}{N}\right)^{N-r} \tag{39}$$

When $S/A \ll 1$ and $N \gg 1$, such that $NS/A = \lambda$ remains bounded, the binomial distribution has significant values only around $r \approx \lambda \ll N$ and takes the limiting form,

$$P(r|\lambda) = \frac{\lambda^r}{r!} e^{-\lambda} \tag{40}$$

that is to say a Poisson distribution. The properties used are (i) the stationarity (p is constant), (ii) the independence of the trials, (iii) no two sources fall at the same point.

Remark: If there are N points in total in the area $\mathscr{A}$ or on the 4π area of a sphere and r in the small area S, there are $N-r$ in the area $\mathscr{A} - S$. If r is a random variable, the same is true for $N-r$. However while the probability distribution of r is very well approximated by a Poisson distribution, this becomes totally wrong for $N-r$, although the probability to have r sources in S is naturally the same as having $N-r$ in the complementary surface, or equivalently $P(r) = P'(N-r)$. But in the latter case the average number of sources in $\mathscr{A} - S$ does not remain bounded as $N \to \infty$ with $S \to 0$. However both distributions are rigourously binomial since $\binom{N}{r} = \binom{N}{N-r}$. This simple example helps grasp the importance of checking the validity of the underlying assumptions before relying to the Poisson approximation.

Derivation from waiting time process

In the context of spatial distribution of random points, it is useful to give another proof of the above result with a very different approach based on memoryless stochastic process. The proof is based on mathematical induction.

First consider the probability distribution of the waiting time $\mathscr{T}$ before a source falls in the area S from an arbitrary origin. Let $P(t)$ be the probability that $\mathscr{T} > t$, that is to say no successful outcome occurs before t. One assumes that the probability that a success occurs over an small interval dt is proportional to the length of the interval as $dp = \lambda\, dt$ and that the probability that

two successes arise over the same interval is negligibly small compared to dt ($\lim_{dt\to 0} P(r = 2|dt)/dt = 0$)). Therefore the probability that a success occurs between t and $t+dt$ is given by,

$$P(t) - P(t+dt) = P(t)\lambda\, dt \tag{41}$$

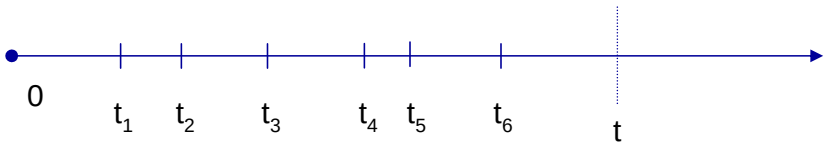


Figure 11: *Typical arrival of six events over a time t for a memoryless exponential stochastic process.*

where the right hand side expresses that this probability is also equal to the probability that the system does not change of state until t and does change during the subsequent interval dt. The solution of this difference equation satisfying $P(0) = 1$ is,

$$P(t) = e^{-\lambda t} \tag{42}$$

meaning that the waiting time follows an exponential law. The probability that no event arises over a duration t is equal to $e^{-\lambda t}$ and the probability that one event occurs precisely between t and $t+dt$ is $\lambda\, e^{-\lambda t} dt$.

Consider now the number of successes over a given duration t, that is to say how many times a success has occurred over this interval. If there are r such events, the outcomes occur at time $t_1, t_2, \cdots, t_r$ with $t_i \leq t$ for $i = 1, \cdots, r$. A typical occurrence for six events is shown in Fig. 11.

Compute first the probability that exactly *one* event occurs between 0 and t. It may happens at any time $\tau \leq t$ and then no new event occurs after τ, that is to say in the interval $[\tau, t]$, or equivalently because this is a memoryless process, no event occurs over the duration $t - \tau$. Therefore the probability can be written as,

$$P(r = 1) = \int_0^t e^{-\lambda\,\tau}\, \lambda\, e^{-\lambda(t-\tau)} d\tau = \lambda\, t\, e^{-\lambda t} \tag{43}$$

the same as a Poisson distribution for $r = 1$. Assuming that this result is general and that the probability distribution of the number of events over a duration t is given by (2), one has the

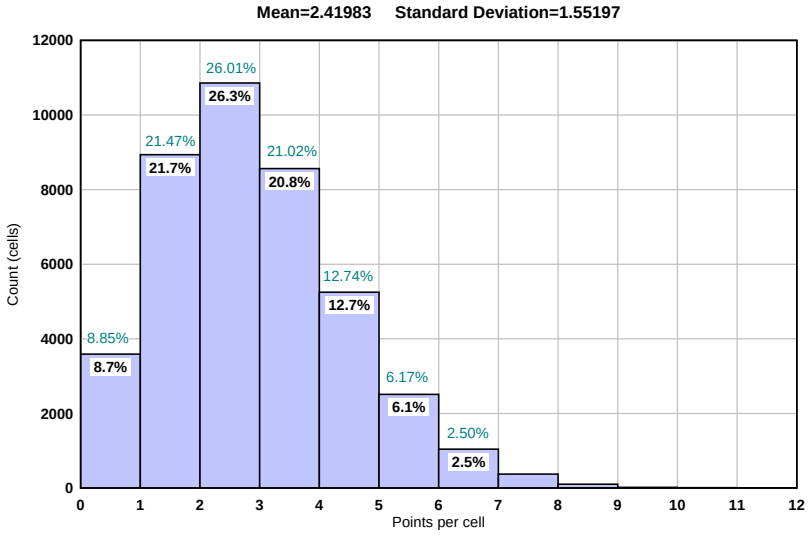


Figure 12: *Distribution of the number of points per cell of 1 deg^2 on a sphere for 100,000 uniformly and randomly distributed points. The Poisson expected percentages are in green above the columns and the observed values in black inside the columns. There are 41,252 cells each populated by 2.42 points in average.*

recurrence relationship for the probability of $r+1$ outcomes,

$$P(r+1) = \int_0^t \frac{(\lambda\,\tau)^r}{r!}\, e^{-\lambda\,\tau}\,\lambda\, e^{-\lambda(t-\tau)} d\tau = \frac{(\lambda\,t)^{(r+1)}}{(r+1)!}\, e^{-\lambda\,t} \tag{44}$$

where the expression under the integral expresses the probability that exactly r events occur over the duration τ, one in the interval $d\tau$ and none after that until the time t. The final result is clearly that under the assumption of a stationary memoryless stochastic arrival of the events, the number of outcomes over t follows a Poisson distribution.

For the spatial distribution of sources over an area S, it suffices to consider that sources fall randomly on the area A at regular intervals, such that there are N such intervals over the time t. This means that $\lambda\,t$ is the average number of points found in the area S and is equal to λ in the Bernoulli based derivation.

A simulation shows how nicely this distribution is matched by a random drawing of 100,000 points on a sphere. I have then divided the sphere into non-overlapping cell of 1 deg^2 and counted how many points fell in each cell. Since there are 41,252 cells, on should have $\lambda = 2.424$ points per cell. The observed distribution is plotted in Fig. 12 with the observed percentage written in the column and the theoretical Poisson value just above. The chance fluctuation is about $0.5/\sqrt{(100000)} \simeq 0.15\%$. Allowing for this stochastic margin, the agreement is just perfect.

References

[1] Renka, R., 1997, Delaunay Triangulation and Voronoï Diagram on the Surface of a Sphere, ACM Transactions on Mathematical Software, **23**, 3.

[2] Brown, A., 2007, DU18 Update, `http://www.rssd.esa.int/SA-general/Projects/GAIA/wiki/images/e/e1/CU5M04_Brown_DU18.pdf`

[3] Lubotzky, A., Phillips, R., Sarnak, P., 1986, Hecke Operators and Distributing Points on the Sphere I, Com. Pure & App. Math., **39**, 149-186.