THE ASTROPHYSICAL JOURNAL, 412:64–71, 1993 July 20

BIAS AND VARIANCE OF ANGULAR CORRELATION FUNCTIONS

STEPHEN D. LANDY AND ALEXANDER S. SZALAY[1]
Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218

Received 1992 October 9; accepted 1993 January 29

ABSTRACT

We present a general method for calculating the bias and variance of estimators for $w(\theta)$ based on galaxy-galaxy (DD), random-random (RR), and galaxy-random (DR) pair counts and describe a procedure for quickly estimating these quantities given an arbitrary two-point correlation function and sampling geometry. These results, based conditionally upon the number counts, are accurate for both high and low number counts. We show explicit analytical results for the variances in the estimators DD/RR, DD/DR, which turn out to be considerably larger than the common wisdom Poisson estimate and report a small bias in DD/DR in addition to that due to the integral constraint. Further, we introduce and recommend an improved estimator $(DD - 2DR + RR)/RR$, whose variance is nearly Poisson.

Subject headings: galaxies: clustering — methods: numerical

1. INTRODUCTION

A common statistic characterizing the clustering of galaxies is the galaxy correlation function. It has been estimated using angular positions of galaxies in magnitude-limited samples, as well as in surveys with redshift information. Since catalogs with redshifts are still relatively small, angular galaxy correlations are more readily available and with additional information and assumptions can be used to estimate the spatial function (Rubin 1954; Limber 1954; Fall & Tremaine 1977; Fall 1979).

The angular correlation function $w(\theta)$ is the projection of the spatial function on the sky and is defined in terms of the joint probability δP of finding two galaxies separated by an angular distance θ with respect to that expected for a random distribution (see Peebles 1980, § 45),

$$\delta P = N^2[1 + w(\theta)]\delta\Omega_1\,\delta\Omega_2\ , \qquad (1)$$

where $\delta\Omega_1$ and $\delta\Omega_2$ are elements of solid angle and N is the mean surface density of objects. If $w(\theta)$ is zero, the distribution is homogeneous.

In practice, pair counts of galaxies are measured as a function of angular separation from photographic plates or CCD images. The usual estimator for $w(\theta)$ is then given by the ratio of the number of pairs of galaxies counted in the sample DD to that expected for a random distribution RR with the same mean density and sampling geometry, suitably normalized:

$$1 + \hat{w}(\theta) = \frac{DD}{RR}\ . \qquad (2)$$

Other common estimators involve ratios of galaxy pair counts to cross-correlated pair counts of data and randomly distributed points.

The major problems in estimating the angular correlation function involve the lack of knowledge of the underlying uniform density convolved with true and artificial large-scale gradients, small sample sizes, and interactions of the sample data with the boundaries of the sampling space. Different estimators have been proposed to overcome these various problems and measure the correlation function with minimum bias and variance (Peebles & Hauser 1974; Peebles 1975; Sharp 1979; Shanks et al. 1980; Hewett 1982; A. J. Hamilton 1992, private communication). Generally, it has been assumed that the variance in these estimators is the Poisson error of the bin counts (Peebles 1980, § 48).

Until recently most work has been done with strongly clustered, bright galaxies and relatively large data sets. However, as research has expanded into faint galaxy correlations with small field sizes (Koo & Szalay 1984; Pritchet & Hartwick 1987, Efstathiou et al. 1991; Neuschaefer, Windhorst, & Dressler 1991) new difficulties have arisen from working with smaller data sets where the quantity of interest is both the amplitude of the correlations as well as the relative changes in the amplitude. Here small effects become important, and investigating the uncertainty and biases in $w(\theta)$ estimators in this regime is the focus of this paper.

In the following section we develop a general procedure for analyzing the bias and variance of *any* estimator for $w(\theta)$ which is constructed out of pair counts and/or cross-correlations between data and random points. Section 3 contains the explicit calculation of the quantities introduced in the preceeding section with the added benefit that they are conditional upon the number counts for any given data set. This method gives accurate results in both high and low number count regimes for an arbitrary two-point correlation function and sampling geometry. In § 4 we apply these results to common estimators of $w(\theta)$ and show that the variances are larger than common knowledge would predict, in agreement with other studies (Ripley 1988), and that the estimator DD/DR has a small additional bias not previously reported. We also introduce the estimator $(DD - 2DR + RR)/RR$ which minimizes the variance to the Poisson level. Section 5 contains Monte Carlo verification of our results, in a regime relevant to faint galaxy counts, along with a simple prescription for estimating the bias and variance given an arbitrary geometry, correlation function, and estimator.

2. CONSTRUCTING ESTIMATORS FOR $w(\theta)$

2.1. *Definitions and Notation*

The angular two-point correlation function is the relative probability of finding a pair of galaxies separated by certain angular distance with respect to that for a uniform distribu-

[1] Also Department of Physics, Eötvös University, Puskin U. 5–7, Budapest, Hungary.

tion. Estimators for $w(\theta)$ are generally constructed out of ratios between three fundamental quantities. These are the number of pairs of galaxies DD, the number of pairs given a cross-correlation between the galaxies and a random distribution DR, and the number of pairs for a random distribution RR, all suitably normalized and with regard to the geometry in which the data were taken. We will take their explicit dependence on θ to be understood.

These pair counts are random variables, whose randomness has two different aspects. First, they depend on the total number of galaxies on the plate, n, which is a random variable itself. This fixes the total number of pairs in DD as $n(n-1)/2$. However, even for a fixed n there will be variations in the pair counts in the different bins due to the many different ways the galaxies can be distributed over the survey area. We will calculate the conditional average and variance of the estimator $\hat{w}(\theta)$ for a fixed n. This most resembles a real observation with a known number of galaxies, but with an unknown expectation value for n. Later, we average over the distribution $P(n)$, the probability of having exactly n galaxies in the survey area.

The conditional averages with a fixed number of data and random points n and n_r will be given by $\langle DD\rangle$, $\langle DR\rangle$, and $\langle RR\rangle$. The symbol $\langle\ldots\rangle \equiv \langle\ldots \mid n, n_r\rangle$ will denote the conditional average of the enclosed quantity throughout.

The expectation value for RR is proportional to the fraction of the total pairs in the bin with separations of $\theta \pm d\theta/2$, for a uniform distribution. This fraction is a geometric quantity which we seek from RR, given the complicated but known shape and excisions of a real survey. It can be determined with an arbitrary accuracy, by taking either a very large number of random points and/or repeating the Monte Carlo calculation many times over.

In this notation the two most common estimators for $w(\theta)$, like data-data pairs over random-random pairs and data-data pairs over data-random pairs, are expressed in the following simple way:

$$1 + \hat{w}_1(\theta) = \frac{DD}{RR}\,\frac{n_r(n_r-1)}{n(n-1)}\,,$$

$$1 + \hat{w}_2(\theta) = \frac{DD}{DR}\,\frac{2n_r}{n-1}\,, \tag{3}$$

where the second term is the normalization. Below we will consider both these and other estimators and calculate their expectations and variances.

2.2. *Fluctuations of the Pair Counts*

To calculate the bias and variances of estimators of this general form, it is useful to express the three pair counts in terms of fluctuations about their means. Let

$$DD = \langle DD\rangle(1+\alpha)\,,$$

$$DR = \langle DR\rangle(1+\beta)\,,$$

$$RR = \langle RR\rangle(1+\gamma)\,, \tag{4}$$

where α, β, and γ are the fluctuations about the mean for a given realization of the data or random set Then by definition $\langle\alpha\rangle = \langle\beta\rangle = \langle\gamma\rangle = 0$. The variance of γ can be made arbitrarily small by choosing a large enough n_r. Hereafter we will consider this to be zero, but will provide a recipe on how to estimate its magnitude in a practical case. Using a similar procedure, all fluctuations in DR will be due to variations in the data points only. In essence, the random part behaves as a continuum and will be taken as such throughout our calculations.

Using this method the mean and variance of any estimator constructed out of combinations of these pair counts can be expressed in terms of the mean pair counts and these fluctuations. For example, the mean and variance of the estimator DD/DR to second order in the fluctuations are

$$1 + \hat{w}_2(\theta) = \left\langle\frac{DD}{DR}\right\rangle = \frac{2n_r}{n-1}\,\frac{\langle DD\rangle}{\langle DR\rangle}\left\langle\frac{1+\alpha}{1+\beta}\right\rangle$$

$$\simeq \frac{2n_r}{n-1}\,\frac{\langle DD\rangle}{\langle DR\rangle}\,(1 - \langle\alpha\beta\rangle + \langle\beta^2\rangle)\,,$$

$$\mathrm{var}\,[\hat{w}_2(\theta)] \simeq \left(\frac{2n_r}{n-1}\,\frac{\langle DD\rangle}{\langle DR\rangle}\right)^2(\langle\alpha^2\rangle + \langle\beta^2\rangle - 2\langle\alpha\beta\rangle)\,. \tag{5}$$

The second moments of these fluctuations are simply the normalized variances of the pair counts

$$\langle\alpha^2\rangle = \frac{\langle DD\cdot DD\rangle - \langle DD\rangle^2}{\langle DD\rangle^2}\,,$$

$$\langle\beta^2\rangle = \frac{\langle DR\cdot DR\rangle - \langle DR\rangle^2}{\langle DR\rangle^2}\,. \tag{6}$$

The presence of $\langle\alpha\beta\rangle$ in these quantities it due to the correlated nature of DD and DR and is given by

$$\langle\alpha\beta\rangle = \frac{\langle DD\cdot DR\rangle - \langle DD\rangle\langle DR\rangle}{\langle DD\rangle\langle DR\rangle}\,. \tag{7}$$

Therefore, in order to calculate the bias and variance of an estimator based on these pair counts we must calculate the first and second moments of the pair counts as well as their cross-correlations. Explicit calculation of these quantities is the focus of § 3.

3. CALCULATION OF PAIR COUNTS AND FLUCTUATIONS

3.1. *Uncorrelated Pair Counts* $\langle DD\rangle$

First consider n points to be distributed in a uniform random fashion over the survey area Ω. Their distribution, while multinomial, is very well approximated with an infinitesimal Poisson process. Divide the sampling space Ω into K cells. In the limit of large K, the number of galaxies found in any cell v_i is either zero or one, a greater number being an infinitesimal of higher order. The probability of finding an object in any cell is $\langle v\rangle = n/K$. The expectation for a second cell becomes $\langle v'\rangle = (n-1)/(K-1)$, etc.

The expected number of pairs in the *uncorrelated* case for an angular separation θ given K cells can be expressed as

$$\langle DD\rangle = \left\langle\sum_{i<j}^{K} v_i v_j \Theta_{ij}^{\theta}\right\rangle = \sum_{i<j}^{K}\langle v_i v_j\rangle\Theta_{ij}^{\theta}\,, \tag{8}$$

where Θ_{ij}^{θ} is one, if cells i and j are separated by a distance $\theta \pm d\theta/2$, and zero otherwise. In our analysis we will only be concerned with the case $i \neq j$, that is, we are working with pair counts, not cell counts. For a given sample with n random points, the probability that both will contain a point is

$$\langle v_i v_j\rangle = \frac{n(n-1)}{K(K-1)}\,. \tag{9}$$

Since this quantity is a constant, it can be taken outside the summation, and we are left with $\sum_{i<j}^{K} \Theta_{ij}^{\theta}$. It is useful to rewrite this as

$$\sum_{i<j}^{K} \Theta_{ij}^{\theta} = \frac{K(K-1)}{2} G_p(\theta) , \qquad (10)$$

where $G_p(\theta)$ is a dimensionless geometric form factor which is equal to the fraction of unique cell pairs separated by a distance $\theta \pm d\theta/2$. Therefore, the expected number of pairs for a sample with n galaxies is

$$\langle DD \rangle = \frac{n(n-1)}{K(K-1)} G_p(\theta) \frac{K(K-1)}{2} = \frac{n(n-1)}{2} G_p(\theta) , \qquad (11)$$

independent of K. The quantity $G_p(\theta)$ is identical to the probability of finding any two randomly placed points separated by a distance $\theta \pm d\theta/2$ in the same geometry. For a large number of realizations we can estimate the ensemble average of the number of pairs as

$$\begin{aligned} E_N(DD) &= \sum_{n=2}^{\infty} \langle DD \rangle P(n) \\ &= G_p(\theta) \sum_{n=2}^{\infty} \frac{n(n-1)}{2} P(n) = \frac{N^2}{2} G_p(\theta) , \qquad (12) \end{aligned}$$

where $P(n)$ is the probability of obtaining n galaxies from a Poisson distribution with mean N.

To determine the second moment of the pair counts we will follow the same method as above and take the ensemble average over n at the end:

$$\begin{aligned} \langle DD \cdot DD \rangle &= \left\langle \sum_{i<j}^{K} v_i v_j \Theta_{ij}^{\theta} \sum_{k<l}^{K} v_k v_l \Theta_{kl}^{\theta} \right\rangle \\ &= \sum_{i<j}^{K} \sum_{k<l}^{K} \langle v_i v_j v_k v_l \rangle \Theta_{ij}^{\theta} \Theta_{kl}^{\theta} . \qquad (13) \end{aligned}$$

The total number of nonzero terms in this sum is $[G_p(\theta)K(K-1)/2]^2$. Since there are either zero or one galaxies in each cell v_i, the expectation value of all the moments of v_i are equal to $v = \langle v_i \rangle = \langle v_i^2 \rangle$. Therefore, it is useful to divide the terms of this equation into three separate sums which depend on the relative values of the individual indices, considering degenerate configurations explicitly, similar to the approach of Peebles (1980, § 48).

Since $i < j$ and $k < l$, in the sum there are three terms corresponding to various degeneracies in the indices. The first term consists of the cases in which no indices overlap. Given K cells there are

$$\binom{K}{2}\binom{K-2}{2} = K(K-1)(K-2)(K-3)/4$$

such combinations. Let $G_q(\theta)$ be the fraction of these which satisfy the constraint that the members of the pairs i, j and k, l are separated by a distance $\theta \pm d\theta/2$, that is,

$$\sum_{i,j,k,l}^{*} \Theta_{ij}^{\theta} \Theta_{kl}^{\theta} = \frac{K(K-1)(K-2)(K-3)}{4} G_q(\theta) , \qquad (14)$$

where $\sum^*$ indicates that all summing indices are different. In a similar manner, the second term consists of those cases in which one of the two indices k, l overlaps with either i or j. The total number of these cases is

$$2\binom{K}{2}(K-2) = K(K-1)(K-2) .$$

Let us pick this degenerate index as i. Then the fraction $G_t(\theta)$ of all triplets of cells, which given one point as the center, the other two are within a distance $\theta \pm d\theta/2$ of the first, becomes

$$\sum_{i,j,k}^{*} \Theta_{ij}^{\theta} \Theta_{ik}^{\theta} = K(K-1)(K-2) G_t(\theta) . \qquad (15)$$

The last term consists of the case in which two pairs overlap. As above, this is denoted by the fraction $G_p(\theta)$ of $K(K-1)/2$ pairs of cells which are separated by a distance $\theta \pm d\theta/2$. As a check on pair conservation,

$$\begin{aligned} \left[\frac{K(K-1)}{2}\right]^2 &= \frac{K(K-1)(K-2)(K-3)}{4} \\ &\quad + K(K-1)(K-2) + \frac{K(K-1)}{2} . \qquad (16) \end{aligned}$$

Multiplying these cases by their respective probabilities cancels all K dependence as before, in equation (11). Summing, we get

$$\begin{aligned} \langle DD \cdot DD \rangle &= \frac{n(n-1)(n-2)(n-3)}{4} \times G_q(\theta) \\ &\quad + n(n-1)(n-2)G_t(\theta) + \frac{n(n-1)}{2} G_p(\theta) . \qquad (17) \end{aligned}$$

This equation can be simplified further by taking into consideration the constraint that the total number of nonzero terms in this sum equals the number of pairs squared:

$$\begin{aligned} \left[G_p(\theta) \frac{K(K-1)}{2}\right]^2 &= G_q(\theta) \frac{K(K-1)(K-2)(K-3)}{4} \\ &\quad + G_t(\theta)K(K-1)(K-2) + G_p(\theta) \frac{K(K-1)}{2} . \qquad (18) \end{aligned}$$

Dividing both sides by K^4 and taking the limit for large K gives

$$G_q(\theta) = G_p^2(\theta) . \qquad (19)$$

Therefore, the concise expression for the expected variance of the number of pairs is

$$\begin{aligned} \mathrm{var}\,(DD) &= \frac{n(n-1)(n-2)(n-3)}{4} G_p^2(\theta) - \left[\frac{n(n-1)}{2} G_p(\theta)\right]^2 \\ &\quad + n(n-1)(n-2)G_t(\theta) + \frac{n(n-1)}{2} G_p(\theta) . \qquad (20) \end{aligned}$$

Here only the last term is DD, the Poisson noise. When n is distributed as a Poisson random variable with mean N, the variance over the Poisson ensemble becomes

$$\mathrm{var}_N\,(DD) = N^3 G_t(\theta) + \frac{N^2}{2} G_p(\theta) . \qquad (21)$$

It has been generally believed that the variance in pair counts is Poisson in the number of pairs var$(DD) = \langle DD \rangle$ (see Peebles 1980, § 48). The last term in equation (21) equals the expected number of pairs; however, the additional term $N^3 G_t(\theta)$ can be relatively large depending on the bin width. In two dimensions, the coefficient $G_t(\theta)$ depends on the second power of the bin width while $G_p(\theta)$ depends on the first. Therefore, the variance in pair counts only approaches that of the expected number of pairs in the limit of an infinitesimal bin width. Note also that for n relatively small, this equation would greatly overestimate the exact variance as given by equation (20) and illustrates the utility of working with conditional pair counts for small samples. The value of the conditional variance, on the other hand, can be greater or less than the expected number of pairs depending upon the relationship between n, $G_p(\theta)$, and $G_t(\theta)$. Calculating $\langle \alpha^2 \rangle$ we obtain

$$\langle \alpha^2 \rangle = \frac{2}{n(n-1)} \left\{ 2(n-2)\left[\frac{G_t(\theta)}{G_p^2(\theta)} - 1\right] + \frac{1}{G_p(\theta)} - 1 \right\} . \quad (22)$$

3.2. *Correlated Pair Counts* $\langle DD \rangle$

Several complications arise when generalizing the above method to the case of nonzero correlations. The first concerns the sample size in relation to the correlation length. We will explicitly add the correlations now to the derivation. Let us define w_Ω as the mean of the two-point correlation function over the sampling geometry,

$$w_\Omega = \int_\Omega G_p(\theta) w(\theta) d\Omega . \quad (23)$$

The normalization of $G_p(\theta)$ is equal to one by definition.

We can calculate the conditional expectation for the number of pairs at a separation $\theta \pm d\theta/2$ given that our correlated sample contains n galaxies. Now the expectation value for the infinitesimal pair count becomes

$$\langle v_i v_j \rangle = \frac{n(n-1)}{K(K-1)} [1 + w(\theta)] . \quad (24)$$

The inclusion of correlations will also change the normalization. Let C_Ω be this unknown normalization constant for the conditional expectation, such that after integrating over θ we retrieve the correct total number of pairs. Thus

$$\langle DD \rangle = C_\Omega \frac{n(n-1)}{2} G_p(\theta)[1 + w(\theta)] . \quad (25)$$

Integrating over θ we obtain

$$\frac{2}{n(n-1)} \int_\Omega \langle DD \rangle d\Omega = C_\Omega \int_\Omega G_p(\theta)[1 + w(\theta)] d\Omega$$
$$= C_\Omega(1 + w_\Omega) = 1 . \quad (26)$$

Therefore

$$C_\Omega = \frac{1}{1 + w_\Omega} , \quad (27)$$

and the expected number of pairs given n galaxies is

$$\langle DD \rangle = \frac{n(n-1)}{2} G_p(\theta) \frac{1 + w(\theta)}{1 + w_\Omega} . \quad (28)$$

The constant is the integral constraint correction used when estimating $w(\theta)$ with an unknown mean surface density (Peebles 1974). The next complication arises in calculating the expected number of pairs squared. Looking again at equation (17), the three terms must be modified in order to take into account the relative excess probabilities of finding pairs, triplets, and quadruplets given the correlation function. Here we will assume that the correlations are weak enough that the higher order (three- and four-point) correlations are negligible. Also we neglect quantities which are second order in $w(\theta)$. For recent results in the strong clustering regime, but neglecting the discreteness, see Mo, Jing, & Börner (1992).

The first term in the case of zero correlations is simply the expectation of quadruplets with unique indices. In the case of nonzero correlations it becomes a weighted average over possible distinct pairs, the members of which are constrained to lie within a distance $\theta \pm d\theta/2$ of each other while the relative position of the two pairs is averaged over the sampling space. As an approximation, we will work to first order in the two-point function. The excess probabilities are given in terms of two-point correlation functions between each pair of points, thus

$$\langle v_i v_j v_k v_l \rangle = \frac{n(n-1)(n-2)(n-3)}{K(K-1)(K-2)(K-3)}$$
$$\times (1 + w_{ij} + w_{ik} + w_{il} + w_{jk} + w_{jl} + w_{kl}) . \quad (29)$$

The pairs ij and kl are constrained to lie within $\theta \pm d\theta/2$ of each other, and so $w_{ij} = w_{kl} = w(\theta)$, while the other four pair separations are averaged over the sampling space, somewhat constrained by the two fixed pairs ij and kl. Here we make the approximation that the average value for w_{ik}, etc., over the sampling space is w_Ω. This becomes a poor approximation as the separation approaches the sample size, however, the error is definitely second order, and the exact result could be calculated numerically if a higher degree of accuracy was needed. Normalizing as above we get

$$\frac{n(n-1)(n-2)(n-3)}{4} G_p^2(\theta) \frac{1 - 2w(\theta)}{1 + 2w_\Omega} . \quad (30)$$

Similarly, the second term is given by the average over the three pairs given three points, with two points constrained to lie within $\theta \pm d\theta/2$ of a third, and by consequence within twice that distance of each other. This last average we will approximate by $w(\theta)$ as well, thus

$$n(n-1)(n-2)G_t(\theta) \frac{1 + 3w(\theta)}{1 + 3w_\Omega} . \quad (31)$$

The last term is simply the expected number of pairs given the two-point function and a separation $\theta \pm d\theta/2$,

$$\langle DD \rangle = \frac{n(n-1)}{2} G_p(\theta) \frac{1 + w(\theta)}{1 + w_\Omega} . \quad (32)$$

Therefore, the expected number of pairs squared is

$$\langle DD \cdot DD \rangle = \frac{n(n-1)(n-2)(n-3)}{4} G_p^2(\theta) \frac{1 + 2w(\theta)}{1 + 2w_\Omega}$$
$$+ n(n-1)(n-2)G_t(\theta) \frac{1 + 3w(\theta)}{1 + 3w_\Omega}$$
$$+ \frac{n(n-1)}{2} G_p(\theta) \frac{1 + w(\theta)}{1 + w_\Omega} . \quad (33)$$

The conditional variance is calculated by simply subtracting the square of the expected number of pairs. Therefore, $\langle\alpha^2\rangle$ is now given by

$$\langle\alpha^2\rangle = \frac{2}{n(n-1)}\left\{2(n-2)\left[\frac{G_t(\theta)}{G_p^2(\theta)}\,\frac{1+w(\theta)}{1+w_\Omega}-1\right] + \frac{1}{G_p(\theta)}\,\frac{1+w_\Omega}{1+w(\theta)}-1\right\}. \qquad (34)$$

In this expression both $w(\theta)$ and w_Ω appear only next to unity, thus they are negligible as far as the variance is concerned since we are assuming the weak correlation regime. Consequently we get a considerably simplified expression in agreement with the case of zero correlations, equation (22).

3.3. *Data-Random Pair Counts* (*DR*)

We will proceed in a similar manner in calculating the mean and variance of the data-random cross-correlation pair counts. For each hypothetical data set we would first cross-correlate it with a large number of random sets and take the average results to approach the continuum limit. Next we take an ensemble average over different realizations of the data, conditional upon n to calculate the mean and variance. The calculation here follows closely that of §§ 3.1 and 3.2. Since random and data points are uncorrelated, the mean number of data-random pairs is simply

$$\langle DR\rangle = nn_r\,G_p(\theta)\,. \qquad (35)$$

The equation for the second moment is given by

$$\langle DR\cdot DR\rangle = \left\langle \sum_{i\neq j}^{K} v_i\rho_j\Theta_{ij}^{\theta}\sum_{k\neq l}^{K} v_k\rho_l\Theta_{kl}^{\theta}\right\rangle , \qquad (36)$$

where ρ_l is the continuum expectation that cell l contains a random point. This calculation proceeds almost identically as that explicitly shown for the expected variance of pairs in § 3.2, except that when the indices of two random points overlap in this calculation, the joint probability is given by $(n_r/K)^2$ rather than n_r/K since we are working with an average over random sets for each data set. The solution is given by

$$\langle DR\cdot DR\rangle = nn_r^2[G_p^2(\theta)(n-1)+G_t(\theta)] \qquad (37)$$

and

$$\langle\beta^2\rangle = \frac{1}{n}\left[\frac{G_t(\theta)}{G_p^2(\theta)}-1\right]. \qquad (38)$$

3.4. *Cross-Correlations between DD and DR*

As the last step, we calculate

$$\langle DD\cdot DR\rangle = \left\langle \sum_{i<j}^{K} v_i v_j\Theta_{ij}^{\theta}\sum_{k\neq l}^{K} v_k\rho_l\Theta_{kl}^{\theta}\right\rangle . \qquad (39)$$

We have to distinguish the case when all of i, j, k are different and the case when k coincides with either i or j. Summing up the different cases we obtain

$$\langle DD\cdot DR\rangle = \frac{n(n-1)}{2}\,n_r[(n-2)G_p^2(\theta)+2G_t(\theta)]\,; \qquad (40)$$

thus

$$\langle\alpha\beta\rangle = \frac{2}{n}\left[\frac{G_t(\theta)}{G_p^2(\theta)}-1\right]. \qquad (41)$$

4. APPLICATION TO VARIOUS ESTIMATORS

At this point let us introduce a new quantity depending on the survey geometry and number counts

$$t = \frac{1}{n}\left[\frac{G_t(\theta)}{G_p^2(\theta)}-1\right]. \qquad (42)$$

This is the combination that appears everwhere in the variances of the fluctuations. For θ small, $G_t(\theta)\simeq G_p^2(\theta)$ but always $t>0$, as can be seen in equation (44) below. It is also convenient to introduce

$$p = \frac{2}{n(n-1)}\left[\frac{1}{G_p(\theta)}-2\frac{G_t(\theta)}{G_p^2(\theta)}+1\right]\simeq\frac{2}{n(n-1)G_p(\theta)}\,, \qquad (43)$$

the inverse of the pair counts, the usual Poisson error. Using these expressions $\langle\alpha^2\rangle$, $\langle\beta^2\rangle$, and $\langle\alpha\beta\rangle$ can be rewritten as

$$\langle\alpha^2\rangle = 4t+p\,,$$
$$\langle\beta^2\rangle = t\,,$$
$$\langle\alpha\beta\rangle = 2t\,. \qquad (44)$$

Here it is easy to see that the first quantity in $\langle\alpha^2\rangle$ which is dominated by the triplet term t enters into the variance as the first power of $1/n$ while the second term p, which is approximately equal to the inverse pair counts, depends on $(1/n)^2$. This relationship becomes important below. From $\langle\alpha^2\rangle$ one can also see that as a general rule the variance in excess of inverse pair counts becomes significant when

$$n > \frac{2\{[1/G_p(\theta)]-1\}}{[G_t(\theta)/G_p^2(\theta)]-1}-3\,. \qquad (45)$$

4.1. *Estimators for* $w(\theta)$

At this point we are ready to consider various estimators. For the sake of simplicity, let us introduce the pair counts normalized for both the auto- and cross-correlations as

$$d = \frac{DD}{G_p(\theta)n(n-1)/2}\,;\quad \langle d\rangle = \frac{1+w(\theta)}{1+w_\Omega}\,;$$
$$x = \frac{DR}{G_p(\theta)nn_r}\,;\quad \langle x\rangle = 1\,. \qquad (46)$$

Out of these we construct four different estimators shown in simplified notation as

$$DD/RR\colon\quad 1+\hat{w}_1(\theta) = d$$
$$DD/DR\colon\quad 1+\hat{w}_2(\theta) = \frac{d}{x}$$
$$DD/DR^2\colon\quad 1+\hat{w}_3(\theta) = \frac{d}{x^2} \qquad (47)$$
$$(DD-2DR+RR)/RR\colon\quad 1+\hat{w}_4(\theta) = d-2x+2\,,$$

DD/DR^2 having been suggested by A. J. Hamilton (1992, private communication). In terms of fluctuations about the mean pair counts and substituting for $\langle\alpha^2\rangle$, $\langle\beta^2\rangle$, and $\langle\alpha\beta\rangle$, these estimators and their variances become

$$\begin{aligned} 1 + \langle \hat{w}_1(\theta) \rangle &= \langle d \rangle , & \text{var}\,[\hat{w}_1(\theta)] &\simeq \langle d \rangle^2 (4t + p) , \\ 1 + \langle \hat{w}_2(\theta) \rangle &\simeq \langle d \rangle (1 - t) , & \text{var}\,[\hat{w}_2(\theta)] &\simeq \langle d \rangle^2 (t + p) , \\ 1 + \langle \hat{w}_3(\theta) \rangle &\simeq \langle d \rangle (1 - t) , & \text{var}\,[\hat{w}_3(\theta)] &\simeq \langle d \rangle^2 (p) , \\ 1 + \langle \hat{w}_4(\theta) \rangle &= \langle d \rangle , & \text{var}\,[\hat{w}_4(\theta)] &\simeq \langle d \rangle^2 (p) . \end{aligned} \tag{48}$$

4.2. *Strategies*

While the variance of the first two estimators are first order in $(1/n)$, the third and fourth are second order. In fact, the variance of these last two estimators is very nearly inverse in the pair counts, closely approximating a Poisson variance. However, the third estimator has a small bias in addition to that due to the integral constraint. This additional bias in the second and third estimators is a consequence of the fact that they are constructed out of the ratio of two random numbers. Therefore, the fourth estimator is preferred when estimating $w(\theta)$.

Although this estimator appears similar to that introduced by Sharp (1979) and discussed by Hewett (1982), it is different. That estimator, which is given by $(DD - DR)/RR$, was primarily aimed at edge effects and has a variance identical to that for DD/DR discussed above. Beyond the advantage of reduced edge effects, the main emphasis here is on the smallest possible variance. In the variance of the correlation function there is a large contribution from $\langle DD \cdot DD \rangle$. The largest term, quadratic in the number of galaxies, cancels with the square of the pair counts. However, as we have shown, the dominant remaining source of noise is proportional to n^3 rather than the usual n^2 from the pair counts themselves. Through the inclusion of the appropriately scaled DR component, we have eliminated the n^3 part of the variance as well, leaving only the Poisson term behind.

It is also important to note that these results assume an ideal world wherein galaxies are distributed in an infinitesimal Poisson process and perfectly observed. Other errors certainly arise due to variable seeing conditions across the sky. If these materialize as a systematic modulation of the galaxy density on scales comparable to the survey area, one can also show that our estimator is optimal. One can also have several areas, observed in the same geometry, with both Poisson and systematic variation in the number counts from field to field. If one first estimated a mean density from the different areas, and then used that for normalization, a serious bias and variance would be introduced. Our method, based upon the conditional average, is clearly preferred.

The question arises as to whether it is advantageous to estimate $w(\theta)$ from one large field with n galaxies or, rather, to break the field up into m subfields with a mean of n/m galaxies per field and average the results. If the nosie went as $1/n$, it would not matter whether the field is split into smaller bits. Using the optimal $1/n^2$ estimator, breaking the field up would increase the variance by at least a factor of m. Additional error is also introduced by the $1/m$ loss in total pairs.

5. MONTE CARLO VERIFICATION AND ANALYSIS

Given the present interest in estimating the angular correlation functions of faint galaxies, we will apply our results in a regime with a density, number count, and bin width approximating that of Efstathiou et al. (1991) only as a typical application of our result. In the following section a simple procedure for estimating $G_p(\theta)$ and $G_t(\theta)$ is described, and in § 5.2 we examine results for estimators DD/RR, DD/DR, and $(DD - 2DR + RR)/RR$.

5.1. *Estimating $G_p(\theta)$ and $G_t(\theta)$*

Although for a circular geometry $G_p(\theta)$ can be derived analytically, $G_t(\theta)$ is much more difficult and may not be expressible in closed form. In any event, most data are not given in a perfect geometric form as some areas are usually masked out due to meteor trails, bad pixel lines, bright stars, etc. However, these quantities can be estimated to an arbitrarily high accuracy in the following simple way. $G_p(\theta)$ is the probability of finding any two randomly distributed points separated by a distance $\theta \pm d\theta/2$ while $G_t(\theta)$ is the probability *given one point* of finding two others at a distance $\theta \pm d\theta/2$ of the first. Therefore, these quantities are given by

$$\begin{aligned} G_p(\theta) &= \frac{\langle n_p(\theta) \rangle}{n(n-1)/2} , \\ G_t(\theta) &= \frac{\langle n_t(\theta) \rangle}{n(n-1)(n-2)/2} , \end{aligned} \tag{49}$$

where n points are randomly distributed over the sampling geometry, $\langle n_p(\theta) \rangle$ is the average number of unique pairs over an ensemble of random data sets, and $\langle n_t(\theta) \rangle$ is the average number of unique triplets given one point as the center. The number of points and realizations can be chosen depending upon accuracy, computer time, etc.

Since the two quantities $G_p(\theta)$ and $G_t(\theta)$ must be estimated, there arises the question of their errors. However, since the variance of an average scales as inverse in the number of runs, 100 runs using the same number of points as in the data set under consideration should reduce the contribution to the variance from these quantities by this same factor compared to the data. The question also arises as to whether it is better to estimate $G_p(\theta)$ from m runs with n points per run or, rather, from one run with mn points. Since the variance of $G_p(\theta)$ scales as $1/n$, either method gives the same $1/n$ variance. Nevertheless, the first method reduces the computational work by a factor of m and so is to be preferred.

5.2. *DD/RR, DD/DF, and $(DD - 2DR + RR)/RR$*

Our Monte Carlo results are based on 1570 galaxies in a circular geometry of radius 237″. We estimate $w(\theta)$ in 15 logarithmic bins of equal size with a lower limit of 3″.5. The Monte Carlo realizations were generated as follows:

1. We chose 1570 random points within a circle of radius 237″ using RAN3 (see Press et al. 1986).
2. Distances between unique pairs of points were calculated and summed for each bin DD. For DR, the original set of random points for each run was cross-correlated with 100 additional sets of 1570 random points. The average of this cross-correlation gave DR for that data set. These were then used to calcuate $1 + \hat{w}(\theta)$ for that run.
3. One hundred runs were made, and the mean and variance of $\hat{w}(\theta)$ were calculated.
4. $G_p(\theta)$ and $G_t(\theta)$ were calculated using equation (49) for 100 runs with 1570 random points per run in the same geometry.

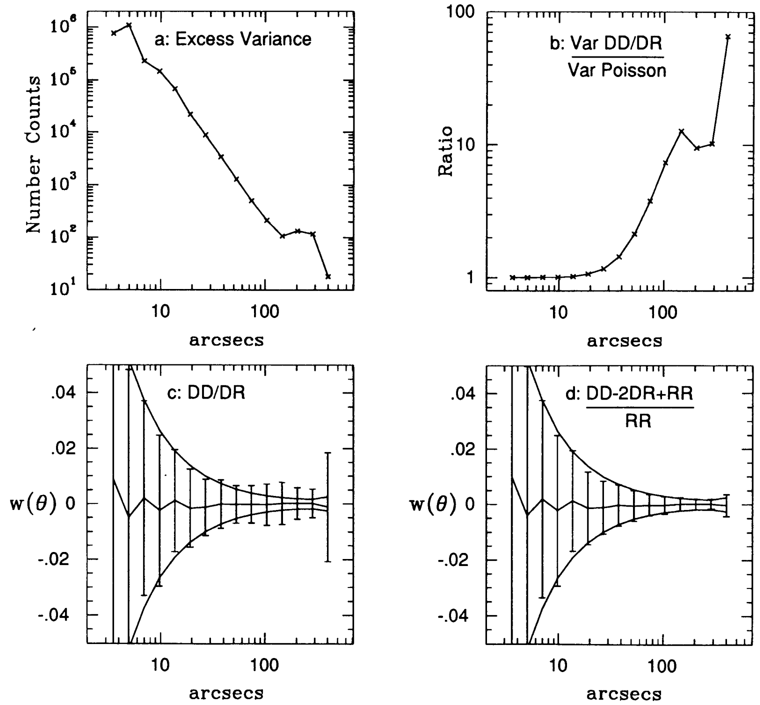

FIG. 1.—(*a*) Eq. (45) is plotted as a function of the point separation. Total number counts above the curve indicate the dominance of the triplet term over the Poisson. Data sets with number counts falling above this line would have a variance significantly greater than Poisson, using an estimator such as *DD*/*DR*. (*b*) The variance ratio of the *DD*/*DR* estimator to the inverse pair counts for 1570 points based upon $G_p(\theta)$ and $G_t(\theta)$ from the Monte Carlo runs. Note that this *underestimation* of the true variance scales approximately linearly with the number counts and thus becomes worse for higher number counts in the same geometry. For the estimator *DD*/*RR* all values would increase by a factor of 4. (*c*) The mean of the estimator *DD*/*DR* with error bars from the Monte Carlo runs. The additional bias due to the random nature of the denominator is insignificant at this number count. The envelope shows the standard errors assuming a Poisson variance. In agreement with (*b*), the true standard deviations have been underestimated up to a factor of 8. (*d*) The estimator (*DD* − 2*DR* + *RR*)/*RR* with error bars from the Monte Carlo runs. The envelope is the standard error assuming a Poisson variance. Notice the excellent agreement.

The line in Figure 1*a* shows *number* counts as a function of separation for this geometry and bin width where the excess variance due to the third-order term becomes comparable to the Poisson variance (eg. [45]). When the total number count is above the curve the variance is dominated by the triplet term rather than the Poisson. Data sets with number counts falling above this line would have a variance significantly greater than Poisson, using an estimator such as *DD*/*DR*.

Figure 1*b* presents the ratio of analytical variance of the *DD*/*DR* estimator to the inverse pair counts for 1570 points based upon $G_p(\theta)$ and $G_t(\theta)$ from the Monte Carlo runs. It is important to note here that since the variance of *DD*/*DR* scales as $(1/n)$ and the pair counts as $(1/n)^2$, the relative error in assuming inverse pair counts for the variance would *increase* approximately linearly as a function of an increasing number of data points. For the estimator *DD*/*RR*, this ratio would increase by approximately a factor of 4 (not shown). The ratio of the variance of the estimator (*DD* − 2*DR* + *RR*)/*RR* to the inverse pair counts is approximately one over the entire range.

Figure 1*c* shows the mean of the estimator *DD*/*DR* along with error bars from the Monte Carlo runs. The additional bias in this estimator is insignificant for this number count. The envelope shows the standard errors assuming a Poisson variance. In agreement with Figure 1*b*, the standard errors are seen to be underestimated by a factor as large as 8.

Figure 1*d* shows the estimator (*DD* − 2*DR* + *RR*)/*RR* along with error bars from the Monte Carlo runs. The envelope is again the standard error assuming a Poisson variance. Note the excellent agreement and significant reduction in variance.

6. CONCLUSION

We have developed a general method which can be used to calculate easily the bias and variance of any estimator derived from galaxy-galaxy (*DD*), random-random (*RR*), and galaxy-random (*DR*) pair counts and present a simple procedure for estimating them to high accuracy for any sampling geometry, point density, and correlation function. As these results are developed in terms of conditional expectations in the number of data points, they are valid for both high and low number counts.

Using this method the variances of the estimators *DD*/*RR* and *DD*/*DR* are derived and shown to be significantly greater than the common wisdom Poisson estimate, that is, inverse pair count. An improved estimator (*DD* − 2*DR* + *RR*)/*RR* for $w(\theta)$ is introduced whose variance is effectively Poisson. This estimator which *improves* the estimation of the angular correlation function is straightforward and costs the researcher no additional computational work. Researchers who wish to assume errors which are Poisson in the pair counts should take advantage of this new estimator. This is especially important when researchers are calculating maximum likelihood fits to correlation functions, pushing their results to the limits of their data such as with sparse data sets, or when they have few fields and thus cannot independently estimate the error.

These calculations can easily be generalized to provide errors for the spatial galaxy correlation function and to construct optimum estimators for higher order correlations. Also, they may be used to estimate the cross-correlation between different angular bins, in order to fit a correlation function with a proper maximum likelihood method. We hope that using conditional estimators has clarified the issues relating to the sources of the statistical noise in correlation functions, and that this work will be extended to various more specialized applications.

We would like to thank István Szapudi for useful conversations and Robert Welsh for computer resources, and acknowledge support from NSF AST-9020380, an NSF-Hungary exchange grant in cosmology, and a grant from OTKA in Hungary.

REFERENCES

Efstathiou, G., Bernstein, G., Katz, N., Tyson, J. A., & Guhathakurta, P. 1991, ApJ, 380, L47
Fall, S. M. 1979, Rev. Mod. Phys., 51, 21
Fall, S. M., & Tremaine, S. 1977, ApJ, 216, 682
Hewett, H. C. 1982, MNRAS, 201, 867
Koo, D. C., & Szalay, A. S. 1984, ApJ, 282, 390
Limber, D. N. 1954, ApJ, 119, 655
Mo, H. J., Jing, Y. P., & Börner, G. 1992, ApJ, 392, 452
Neuschaefer, L. W., Windhorst, R. A., & Dressler, A. 1991, ApJ, 382, 32
Peebles, P. J. E. 1974, A&A, 32, 197
———. 1975, ApJ, 185, 413
———. 1980, The Large Scale Structure of the Universe (Princeton: Princeton Univ. Press)
Peebles, P. J. E., & Hauser, M. G. 1974, ApJS, 28, 19
Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. 1986, Numerical Recipes "The Art of Scientific Computing" (Cambridge: Cambridge Univ. Press)
Pritchet, C. J., & Hartwick, F. D. A. 1987, ApJ, 320, 464
Ripley, B. D. 1988, Statistical Inference for Spatial Processes (Cambridge: Cambridge Univ. Press)
Rubin, V. C. 1954, Proc. Natl. Acad. Sci., 40, 541
Shanks, T., Fong, R., Ellis, R. S., & MacGillivray, H. T. 1980, MNRAS, 192, 209
Sharp, N. A. 1979, A&A, 74, 308