

Let $n(\mathbf{r})$ denote the galaxy number density field. Since galaxies are discrete objects, this field is a sum of delta functions,

$$n(\mathbf{r}) = \sum_g \delta_D(\mathbf{r} - \mathbf{r}_g). \quad (1)$$

Averaged over an ensemble of universes, however, the mean number density $\bar{n}(\mathbf{r}) = \langle n(\mathbf{r}) \rangle$ is continuous and non-negative. (Note that in a statistically homogeneous universe, the mean number density should be uniform, i.e. independent of \mathbf{r} . However we treat $n(\mathbf{r})$ as the *observed* number density for some particular survey, which varies with position according to the survey's selection criteria.)

The probability of finding a galaxy within an infinitesimal volume δV_1 centered at \mathbf{r}_1 is

$$\delta P_1 = \bar{n}(\mathbf{r}_1) \delta V_1. \quad (2)$$

In the absence of clustering, the presence or absense of a second galaxy in a disjoint volume δV_2 at \mathbf{r}_2 is independent of the presence or absence of the first, so the probability of finding galaxies in both volumes is

$$\delta P_2 = \bar{n}(\mathbf{r}_1) \bar{n}(\mathbf{r}_2) \delta V_1 \delta V_2. \quad (3)$$

This generalizes to three or more galaxies, e.g.

$$\delta P_3 = \bar{n}(\mathbf{r}_1) \bar{n}(\mathbf{r}_2) \bar{n}(\mathbf{r}_3) \delta V_1 \delta V_2 \delta V_3, \quad (4)$$

$$\delta P_4 = \bar{n}(\mathbf{r}_1) \bar{n}(\mathbf{r}_2) \bar{n}(\mathbf{r}_3) \bar{n}(\mathbf{r}_4) \delta V_1 \delta V_2 \delta V_3 \delta V_4, \quad (5)$$

and so on. In fact, in the absence of clustering, the mean number density gives us everything we can hope to know about the distribution of galaxies. The probability of finding k galaxies within the finite volume V in this case is given by the Poisson distribution,

$$P[k \text{ galaxies in } V] = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{where } \lambda = \int_V \bar{n}(\mathbf{r}) d^3r. \quad (6)$$

Of course, galaxies *do* exhibit clustering. In the most general case, we can characterize this clustering by introducing a series of multi-point correlation functions that define deviations from the no-clustering case above. Thus, for disjoint infinitesimal volumes δV_i , the probabilities above become

$$\delta P_1 = \bar{n}(\mathbf{r}_1) \delta V_1, \quad (7)$$

$$\delta P_2 = \bar{n}(\mathbf{r}_1) \bar{n}(\mathbf{r}_2) \delta V_1 \delta V_2 [1 + \xi_{12}]. \quad (8)$$

$$\delta P_3 = \bar{n}(\mathbf{r}_1) \bar{n}(\mathbf{r}_2) \bar{n}(\mathbf{r}_3) \delta V_1 \delta V_2 \delta V_3 [1 + \xi_{12} + \xi_{23} + \xi_{13} + \zeta_{123}], \quad (9)$$

$$\begin{aligned} \delta P_4 = & \bar{n}(\mathbf{r}_1) \bar{n}(\mathbf{r}_2) \bar{n}(\mathbf{r}_3) \bar{n}(\mathbf{r}_4) \delta V_1 \delta V_2 \delta V_3 \delta V_4 \\ & \times [1 + \xi_{12} + \xi_{13} + \xi_{14} + \xi_{23} + \xi_{24} + \xi_{34} + \zeta_{123} + \zeta_{124} + \zeta_{134} + \zeta_{234} \\ & + \xi_{12}\xi_{34} + \xi_{13}\xi_{24} + \xi_{14}\xi_{23} + \eta_{1234}], \end{aligned} \quad (10)$$

where ξ_{12} is shorthand for $\xi(\mathbf{r}_1, \mathbf{r}_2)$, etc. The arrangement of terms is chosen so that at each level, the newly introduced n -point correlation function represents n -point clustering *above* what you would already expect given knowledge of $(n-1)$ -point clustering. That is, ξ_{12} represents the *excess* probability

of finding galaxies at *both* \mathbf{r}_1 and \mathbf{r}_2 , given what we already know about finding galaxies at \mathbf{r}_1 or \mathbf{r}_2 separately. Likewise, ζ_{123} represents the excess probability of finding galaxies at \mathbf{r}_1 , \mathbf{r}_2 , *and* \mathbf{r}_3 , given what we know about the probability of finding a pair of galaxies at \mathbf{r}_1 and \mathbf{r}_2 , or a pair at \mathbf{r}_2 and \mathbf{r}_3 , and so on.

In this way, the statistical properties of galaxy clustering can be characterized by a hierarchy of correlation functions, the first three of which are denoted $\xi(\mathbf{r}_1, \mathbf{r}_2)$, $\zeta(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$, and $\eta(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4)$. However, there is one subtlety that has not yet been addressed. We were careful to require above that the volumes δV_i are *disjoint*. To understand what changes when we remove this restriction, it is convenient to imagine partitioning space into a mesh of cells, each of equal infinitesimal volume δV , with units chosen such that $\delta V = 1$. We label each cell by a subscripted index, so that cell a is centered at \mathbf{r}_a , and the number density averaged over the cell is n_a , with ensemble average $\langle n_a \rangle = \bar{n}_a$. Since the cells are infinitesimal, there can be at most one galaxy per cell, thus n_a is either 0 or 1. Additionally, $\bar{n}_a \ll 1$, since the probability of finding a galaxy within an infinitesimal cell is itself infinitesimal.

We now ask what is the probability of finding one galaxy in cell a and one galaxy in cell b . If the cells are disjoint, i.e. if $a \neq b$, then the answer is the same as before (remember $\delta V = 1$),

$$P[\text{galaxy in } a \text{ and galaxy in } b \mid a \neq b] = \bar{n}_a \bar{n}_b (1 + \xi_{ab}). \quad (11)$$

But if $a = b$, then the question is the same as asking what is the probability of finding a galaxy in a , so

$$P[\text{galaxy in } a \text{ and galaxy in } b \mid a = b] = \bar{n}_a \delta_{ab}, \quad (12)$$

where δ_{ab} is the Kronecker delta symbol. Together, then we have

$$P[\text{galaxy in } a \text{ and galaxy in } b] = \bar{n}_a \bar{n}_b (1 + \xi_{ab}) + \bar{n}_a \delta_{ab}. \quad (13)$$

The second term is commonly referred to as a *shot noise* contribution, and arises due to the discreteness of galaxies. (Note that $\bar{n}_a \ll 1$, so when $a = b$ the first term is negligible compared to the second.)

Rather than speaking of probabilities, we can interpret these results in terms of correlations between the values of n_a and n_b . The product $n_a n_b$ will be 1 if there is a galaxy in each cell, and 0 otherwise. Thus, averaged over an ensemble of universes, $\langle n_a n_b \rangle$ gives the probability of finding galaxies in both cells a and b , i.e.

$$\langle n_a n_b \rangle = \bar{n}_a \bar{n}_b (1 + \xi_{ab}) + \bar{n}_a \delta_{ab}. \quad (14)$$

In the continuum limit, this expression reads

$$\langle n(\mathbf{r}_1) n(\mathbf{r}_2) \rangle = \bar{n}(\mathbf{r}_1) \bar{n}(\mathbf{r}_2) [1 + \xi(\mathbf{r}_1, \mathbf{r}_2)] + \bar{n}(\mathbf{r}_1) \delta_D(\mathbf{r}_2 - \mathbf{r}_1). \quad (15)$$

A similar result holds for the 3-point quantity $\langle n_a n_b n_c \rangle$. For this case, we can have all three indices distinct, we can have two of the indices equal and the third distinct, or we can have all three indices equal. Correspondingly, there are three types of terms,

$$\begin{aligned} \langle n_a n_b n_c \rangle = & \bar{n}_a \bar{n}_b \bar{n}_c (1 + \xi_{ab} + \xi_{ac} + \xi_{bc} + \zeta_{abc}) \\ & + \bar{n}_a \bar{n}_b (1 + \xi_{ab}) (\delta_{bc} + \delta_{ac}) + \bar{n}_a \bar{n}_c (1 + \xi_{ac}) \delta_{ab} \\ & + \bar{n}_a \delta_{abc}, \end{aligned} \quad (16)$$

where δ_{abc} is a generalization of the Kronecker delta symbol, equal to 1 if $a = b = c$, 0 otherwise. We have written the second line in an unsymmetric fashion so as to minimize the length of the expression; note however that, when the Kronecker deltas are taken into account, the expression is indeed symmetric under permutations of the indices a, b, c .

Finally, for the 4-point quantity $\langle n_a n_b n_c n_d \rangle$, we have five types of terms, namely those for which (i) all indices are distinct, (ii) there are two pairs of equal indices, (iii) two of the indices are equal, but the other two are distinct, (iv) three of the indices are equal, distinct from the fourth, and (v) all four indices are equal. Thus

$$\begin{aligned}
\langle n_a n_b n_c n_d \rangle = & \bar{n}_a \bar{n}_b \bar{n}_c \bar{n}_d (1 + \xi_{ab} + \xi_{ac} + \xi_{ad} + \xi_{bc} + \xi_{bd} + \xi_{cd} + \zeta_{abc} + \zeta_{abd} + \zeta_{acd} + \zeta_{bcd} \\
& + \xi_{ab}\xi_{cd} + \xi_{ac}\xi_{bd} + \xi_{ad}\xi_{bc} + \eta_{abcd}) \\
& + \bar{n}_a \bar{n}_b (1 + \xi_{ab})(\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}) + \bar{n}_a \bar{n}_c (1 + \xi_{ac})\delta_{ab}\delta_{cd} \\
& + [\bar{n}_a \bar{n}_b \bar{n}_c (1 + \xi_{ab} + \xi_{ac} + \xi_{bc} + \zeta_{abc})(\delta_{ad} + \delta_{bd} + \delta_{cd}) \\
& + \bar{n}_a \bar{n}_b \bar{n}_d (1 + \xi_{ab} + \xi_{ad} + \xi_{bd} + \zeta_{abd})(\delta_{ac} + \delta_{bc}) \\
& + \bar{n}_a \bar{n}_c \bar{n}_d (1 + \xi_{ac} + \xi_{ad} + \xi_{cd} + \zeta_{acd})\delta_{ab}] \\
& + \bar{n}_a \bar{n}_b (1 + \xi_{ab})(\delta_{acd} + \delta_{bcd}) + \bar{n}_c \bar{n}_d (1 + \xi_{cd})(\delta_{abd} + \delta_{abc}) \\
& + \bar{n}_a \delta_{abcd}.
\end{aligned} \tag{17}$$

More useful expressions:

$$\langle (n_a - \bar{n}_a)(n_b - \bar{n}_b) \rangle = \bar{n}_a \bar{n}_b \xi_{ab} + \bar{n}_a \delta_{ab} \tag{18}$$

$$\langle (n_a - \bar{n}_a)(n_b - \bar{n}_b)(n_c - \bar{n}_c) \rangle = \bar{n}_a \bar{n}_b \bar{n}_c \zeta_{abc} + [\bar{n}_a \bar{n}_b \xi_{ab}(\delta_{ac} + \delta_{bc}) + \bar{n}_a \bar{n}_c \xi_{ac} \delta_{ab}] + \bar{n}_a \delta_{abc} \tag{19}$$

$$\begin{aligned}
\langle (n_a - \bar{n}_a)(n_b - \bar{n}_b)(n_c - \bar{n}_c)(n_d - \bar{n}_d) \rangle = & \bar{n}_a \bar{n}_b \bar{n}_c \bar{n}_d (\xi_{ab}\xi_{cd} + \xi_{ac}\xi_{bd} + \xi_{ad}\xi_{bc} + \eta_{abcd}) \\
& + \bar{n}_a \bar{n}_b \bar{n}_c (\xi_{ab}\delta_{cd} + \xi_{bc}\delta_{ad} + \xi_{ac}\delta_{bd} + \zeta_{abc}(\delta_{ad} + \delta_{bd} + \delta_{cd})) \\
& + \bar{n}_a \bar{n}_b \bar{n}_d (\xi_{ad}\delta_{bc} + \xi_{bd}\delta_{ac} + \zeta_{abd}(\delta_{ac} + \delta_{bc})) \\
& + \bar{n}_a \bar{n}_c \bar{n}_d (\xi_{cd}\delta_{ab} + \zeta_{acd}\delta_{ab}) \\
& + \bar{n}_a \bar{n}_b ((1 + \xi_{ab})(\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}) + \xi_{ab}(\delta_{acd} + \delta_{bcd})) \\
& + \bar{n}_a \bar{n}_c ((1 + \xi_{ac})\delta_{ab}\delta_{cd} + \xi_{ac}\delta_{abd}) \\
& + \bar{n}_a \bar{n}_d \xi_{ad}\delta_{abc} \\
& + \bar{n}_a \delta_{abcd}
\end{aligned} \tag{20}$$

$$\begin{aligned}
& \langle (n_a - \bar{n}_a)(n_b - \bar{n}_b)(n_c - \bar{n}_c)(n_d - \bar{n}_d) \rangle - \langle (n_a - \bar{n}_a)(n_b - \bar{n}_b) \rangle \langle (n_c - \bar{n}_c)(n_d - \bar{n}_d) \rangle \\
&= \bar{n}_a \bar{n}_b \bar{n}_c \bar{n}_d (\xi_{ac} \xi_{bd} + \xi_{ad} \xi_{bc} + \eta_{abcd}) \\
&\quad + \bar{n}_a \bar{n}_b \bar{n}_c (\xi_{bc} \delta_{ad} + \xi_{ac} \delta_{bd} + \zeta_{abc} (\delta_{ad} + \delta_{bd} + \delta_{cd})) \\
&\quad + \bar{n}_a \bar{n}_b \bar{n}_d (\xi_{ad} \delta_{bc} + \xi_{bd} \delta_{ac} + \zeta_{abd} (\delta_{ac} + \delta_{bc})) \\
&\quad + \bar{n}_a \bar{n}_c \bar{n}_d \zeta_{acd} \delta_{ab} \\
&\quad + \bar{n}_a \bar{n}_b ((1 + \xi_{ab})(\delta_{ac} \delta_{bd} + \delta_{ad} \delta_{bc}) + \xi_{ab} (\delta_{acd} + \delta_{bcd})) \\
&\quad + \bar{n}_a \bar{n}_c (\xi_{ac} \delta_{ab} \delta_{cd} + \xi_{ac} \delta_{abd}) \\
&\quad + \bar{n}_a \bar{n}_d \xi_{ad} \delta_{abc} \\
&\quad + \bar{n}_a \delta_{abcd} \tag{21} \\
&= (\bar{n}_a \bar{n}_c \xi_{ac} + \bar{n}_a \delta_{ac})(\bar{n}_b \bar{n}_d \xi_{bd} + \bar{n}_b \delta_{bd}) + (\bar{n}_a \bar{n}_d \xi_{ad} + \bar{n}_a \delta_{ad})(\bar{n}_b \bar{n}_c \xi_{bc} + \bar{n}_b \delta_{bc}) \\
&\quad + \bar{n}_a \bar{n}_b \bar{n}_c \bar{n}_d \eta_{abcd} \\
&\quad + \bar{n}_a \bar{n}_b \bar{n}_c \zeta_{abc} (\delta_{ad} + \delta_{bd} + \delta_{cd}) + \bar{n}_a \bar{n}_b \bar{n}_d \zeta_{abd} (\delta_{ac} + \delta_{bc}) + \bar{n}_a \bar{n}_c \bar{n}_d \zeta_{acd} \delta_{ab} \\
&\quad + \bar{n}_a \bar{n}_b \xi_{ab} (\delta_{ac} \delta_{bd} + \delta_{ad} \delta_{bc} + \delta_{acd} + \delta_{bcd}) + \bar{n}_a \bar{n}_c \xi_{ac} (\delta_{ab} \delta_{cd} + \delta_{abd}) + \bar{n}_a \bar{n}_d \xi_{ad} \delta_{abc} \\
&\quad + \bar{n}_a \delta_{abcd} \tag{22}
\end{aligned}$$