LBL-36888
UC-605

Nearest Neighbor Analysis in One Dimension

M. Weingart and S. Selvin

Information and Computing Sciences Division
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

Division of Biostatistics
School of Public Health
University of California, Berkeley
Berkeley, California 94720

February 1995

This work was supported by the Director, Office of Epidemiologic Studies; Office of Health; Office of Environment, Safety and Health; U.S. Department of Energy under Contract No. DE–AC03–76SF00098.

Abstract

This work presents the derivation of the first and second moments of the nearest neighbor distances and their mean in one–dimesion. Five methods of edge effects correction are described and the means of the corrected nearest neighbor distances are compared to the uncorrected one using large scale computer simulations.

Contents

Acknowledgements

The authors are grateful to Mark van der Laan from the Division of Biostatistics, School of Public Health, University of California at Berkeley, for his assistance with the theoretical aspects of this work.

We would like to acknowledge the support from colleagues Deane W. Merrill and Mark Durst from the Information and Computing Science division at Lawrence Berkeley Laboratory, who engaged in weekly conversations that resulted in contributions to the simulations presented in this work.

1.1 Introduction

Since its initial presentation by Clark and Evans [1], nearest neighbor analysis for spatial randomness has gained considerable popularity in fields as diverse as geography, ecology, archaeology, cell biology, forestry, meteorology, and epidemiology.

Epidemiologists are often interested in determining whether disease cases are clustered, dispersed, or randomly distributed, since different patterns of disease incidence over time or space can provide clues to the etiology of the disease. An environmental hazard or a transmissable agent can produce a cluster of disease events, i.e. a set of events occuring unusually close to each other in time, space, or both time and space.

In spite of its wide applicability, few attempts have been made to adapt the nearest neighbor method to the analysis of points distributed along a line.

Clark and Evans [2, 3] presented an extension of the nearest neighbor method to cover distributions of points in dimensions other than two. This approach involves calculating the proportion of points which form reflexive pairs, i.e. pairs of points which are each other's nearest neighbors. For one dimension, if two–thirds of the points are paired, the pattern is assumed to be random. Smaller proportion indicates clustering while a larger proportion indicates a uniform distribution of the points. This technique was introduced into geographical literature by Dacey [4] who analysed the distribution of river towns.

Pinder and Witherick [5, 6] claimed that the formula for mean nearest neighbor in two dimensions can be easily modified to fit the one dimension situation. They carried out a small scale computer simulation to verify their claim. Unfortunately, the modified formula gives reasonable results only for large number of points.

Young [7] assumed that the contributions from the two extreme points to the sum of nearest neighbor distances are either the distance to the neighbors or the distance to the end of the line, whichever is the smaller. As it will be shown later, this approach corresponds to a specific alternative way to overcome the boundary problem, which arises in situations where at least one of the two extreme points is closer to the end–point of the line than to its neighbor.

Selkirk and Neave [8, 9, 10] were the first to recognize the boundary problem in the one dimension case and suggested three alternative ways which they considered as possible solutions. The first method was to consider a situation where n points are distributed around a circle (or any closed curve), a situation in which, according to Selkirk and Neave,

the boundary problem disappears. They claimed that while this situation is relatively less frequent in geographical examples, data which is in angular form can be analysed in this way, as well as data which is periodic in time. As it will be shown, a simple modification of their technique allows its use as an edge effects correction in situations where the points are distributed along a straight line.

Another approach suggested by Selkirk and Neave was to consider $n + 2$ points which are distributed along a straight line, including two at the boundaries of the line. In this case, data points necessarily occur at the end–points of the line, or part of the line. Selkirk and Neave claimed that this may arise in a number of ways; e.g. the extent of the line may be unknown or infinite and the investigator chooses that part of it terminating with two particular data points, or the line may be naturally defined by two data end–points.

The third solution presented by Selkirk and Neave was to consider the contribution of the two extreme points to the sum of nearest neighbor distances to be the distances to their neighbors. Actually, this approach does not involve any correction for the boundary problem and relates to what is defined in this paper as the uncorrected mean nearest neighbor distance.

This report outlines the theoretical derivation of the moments of the mean nearest neighbor distance in the one dimension case and the correction of its expected value in order to overcome the boundary problem. Section 2 presents the derivation of the moments of order statistics, for specific sample sizes and for the general case. These results are then used in Section 3 for the derivation of the moments of nearest neighbor distances, and in Section 4 for the derivation of the moments of the mean nearest neighbor distance. Section 5 presents the boundary problem and examines five alternative ways to compensate for it in the calculation of the expected value of the mean nearest neighbor distance. Section 6 presents the results from a large scale computer simulation which compares the various correction methods.

1.2 Moments of Order Statistics

1.2.1 A Random Sample of 2 Points from a Uniform Distribution

Let x_1 and x_2 be independent identically distributed random variables from a $U(0,1)$ distribution, which represent a random sample of two points on a line of length $L = 1$. Let $x_{(1)} \leq x_{(2)}$ denote the order statistics of the x's, which represent the position of the points on the line.

The joint density function of the two order statistics is:

$f_{(1),(2)}(x_{(1)}, x_{(2)}) = 2! \qquad 0 \leq x_{(1)} \leq x_{(2)} \leq 1$

The marginal density function of $x_{(i)}$, where $i = 1, 2$, can be obtained from the joint density function by integrating out the other variable. Following the procedure suggested by Tsuji [11], who derived the expected value of $x_{(i)}$, the marginal density function can then be used to calculate both the expected value and the variance of $x_{(i)}$.

Expected Value of $x_{(1)}$

$$\begin{aligned} E(x_{(1)}) &= \int_0^1 x_{(1)} f_{(1)}(x_{(1)}) dx_{(1)} \\ &= 2! \int_0^1 \int_{x_{(1)}}^1 x_{(1)} f(x_{(1)}) f(x_{(2)}) dx_{(2)} dx_{(1)} \\ &= 2! \int_0^1 x_{(1)} (1 - x_{(1)})^1 dx_{(1)} \end{aligned}$$

Let

$g(x) = -\frac{(1-x_{(1)})^2}{2} \qquad h(x) = x_{(1)}$

$g'(x) = (1 - x_{(1)})^1 \qquad h'(x) = 1$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned} &= 2! \left[0 + \int_0^1 \frac{(1 - x_{(1)})^2}{2} dx_{(1)} \right] \\ &= 2! \frac{1}{2} \int_0^1 (1 - x_{(1)})^2 dx_{(1)} \\ &= \frac{1}{3} \end{aligned}$$

Variance of $x_{(1)}$

$$\begin{aligned} E(x_{(1)}^2) &= \int_0^1 x_{(1)}^2 f_{(1)}(x_{(1)}) dx_{(1)} \\ &= 2! \int_0^1 \int_{x_{(1)}}^1 x_{(1)}^2 f(x_{(1)}) f(x_{(2)}) dx_{(2)} dx_{(1)} \\ &= 2! \int_0^1 x_{(1)}^2 (1 - x_{(1)})^1 dx_{(1)} \end{aligned}$$

Let

$g(x) = -\frac{(1-x_{(1)})^2}{2}$ $\qquad$ $h(x) = x_{(1)}^2$

$g'(x) = (1 - x_{(1)})^1$ $\qquad$ $h'(x) = 2x_{(1)}$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned} &= 2! \left[0 + 2 \int_0^1 x_{(1)} \frac{(1 - x_{(1)})^2}{2} dx_{(1)} \right] \\ &= 2!\, 2\, \frac{1}{2} \int_0^1 x_{(1)} (1 - x_{(1)})^2 dx_{(1)} \end{aligned}$$

Let

$g(x) = -\frac{(1-x_{(1)})^3}{3}$ $\qquad$ $h(x) = x_{(1)}$

$g'(x) = (1 - x_{(1)})^2$ $\qquad$ $h'(x) = 1$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned} &= 2!\, 2\, \frac{1}{2} \left[0 + \int_0^1 \frac{(1 - x_{(1)})^3}{3} dx_{(1)} \right] \\ &= 2!\, 2\, \frac{1}{2}\, \frac{1}{3} \int_0^1 (1 - x_{(1)})^3 dx_{(1)} \\ &= \frac{2}{3 \cdot 4} = \frac{2}{12} \end{aligned}$$

The variance of $x_{(1)}$ is then

$$Var(x_1) = E(x_{(1)}^2) - (E(x_{(1)}))^2 = \frac{2}{12} - (\frac{1}{3})^2 = \frac{1}{18}$$

Expected Value of $x_{(2)}$

$$\begin{aligned} E(x_{(2)}) &= \int_0^1 x_{(2)} f_{(2)}(x_{(2)}) dx_{(2)} \\ &= 2! \int_0^1 \int_{x_{(1)}}^1 x_{(2)} f(x_{(1)} f(x_{(2)}) dx_{(2)} dx_{(1)} \\ &= 2! \int_0^1 \int_{x_{(1)}}^1 x_{(2)} (1 - x_{(2)})^0 dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$g(x) = -\frac{(1-x_{(2)})^1}{1}$ $\qquad$ $h(x) = x_{(2)}$

$g'(x) = (1 - x_{(2)})^0$ $\qquad$ $h'(x) = 1$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned} &= 2! \int_0^1 [\, x_{(1)}(1 - x_{(1)}) + \int_{x_{(1)}}^1 (1 - x_{(2)}) dx_{(2)} \,] \, dx_{(1)} \\ &= E(x_{(1)}) + 2! \int_0^1 \int_{x_{(1)}}^1 (1 - x_{(2)}) dx_{(2)} dx_{(1)} \\ &= E(x_{(1)}) + \frac{1}{3} \\ &= \frac{1}{3} + \frac{1}{3} = \frac{2}{3} \end{aligned}$$

Variance of $x_{(2)}$

$$\begin{aligned} E(x_{(2)}^2) &= \int_0^1 x_{(2)}^2 f_{(2)}(x_{(2)}) dx_{(2)} \\ &= 2! \int_0^1 \int_{x_{(1)}}^1 x_{(2)}^2 f(x_{(1)}) f(x_{(2)}) dx_{(2)} dx_{(1)} \\ &= 2! \int_0^1 \int_{x_{(1)}}^1 x_{(2)}^2 (1-x_{(2)})^0 dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(2)})^1}{1} & h(x) &= x_{(2)}^2 \\ g'(x) &= (1-x_{(2)})^0 & h'(x) &= 2x_{(2)} \end{aligned}$$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned} &= 2! \int_0^1 [\, x_{(1)}^2 (1-x_{(1)}) + 2 \int_{x_{(1)}}^1 x_{(2)} (1-x_{(2)}) dx_{(2)} \,] \, dx_{(1)} \\ &= E(x_{(1)}^2) + 2!\, 2 \int_0^1 \int_{x_{(1)}}^1 x_{(2)} (1-x_{(2)})^1 dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(2)})^2}{2} & h(x) &= x_{(2)} \\ g'(x) &= (1-x_{(2)})^1 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned} &= E(x_{(1)}^2) + 2!\, 2 \int_0^1 [x_{(1)} \frac{(1-x_{(1)})^2}{2} + \int_{x_{(1)}}^1 \frac{(1-x_{(2)})^2}{2} dx_{(2)} \,] \, dx_{(1)} \\ &= E(x_{(1)}^2) + E(x_{(1)}^2) + 2!\, 2\, \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 (1-x_{(2)})^2 dx_{(2)} dx_{(1)} \\ &= 2E(x_{(1)}^2) + E(x_{(1)}^2) = \frac{2 \cdot 3}{2} E(x_{(1)}^2) \\ &= \frac{2 \cdot 3}{3 \cdot 4} = \frac{6}{12} \end{aligned}$$

The variance of $x_{(2)}$ is then

$$Var(x_2) = E(x_{(2)}^2) - (E(x_{(2)}))^2 = \frac{6}{12} - (\frac{2}{3})^2 = \frac{1}{18}$$

1.2.2 A Random Sample of 3 Points from a Uniform Distribution

Let x_1, x_2, x_3 be independent identically distributed random variables from a $U(0,1)$ distribution, which represent a random sample of three points on a line of length $L = 1$. Let $x_{(1)} \leq x_{(2)} \leq x_{(3)}$ denote the order statistics of the x's, which represent the position of the points on the line.

The joint density function of the three order statistics is:

$f_{(1),(2),(3)}(x_{(1)}, x_{(2)}, x_{(3)}) = 3! \qquad 0 \leq x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq 1$

The marginal density function of $x_{(i)}$, where $i = 1, 2, 3$, can be obtained from the joint density function by integrating out the other variables, and then be used to calculate the expected value and the variance of $x_{(i)}$.

Expected Value of $x_{(1)}$

$$\begin{aligned} E(x_{(1)}) &= \int_0^1 x_{(1)} f_{(1)}(x_{(1)}) dx_{(1)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(1)} f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 x_{(1)}(1 - x_{(2)}) dx_{(2)} dx_{(1)} \\ &= 3! \frac{1}{2} \int_0^1 x_{(1)}(1 - x_{(1)})^2 dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(1)})^3}{3} & h(x) &= x_{(1)} \\ g'(x) &= (1 - x_{(1)})^2 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned} &= 3! \frac{1}{2} [\, 0 + \int_0^1 \frac{(1 - x_{(1)})^3}{3} dx_{(1)} \,] \\ &= 3! \frac{1}{2} \frac{1}{3} \int_0^1 (1 - x_{(1)})^3 dx_{(1)} \\ &= \frac{1}{4} \end{aligned}$$

Variance of $x_{(1)}$

$$\begin{aligned} E(x_{(1)}^2) &= \int_0^1 x_{(1)}^2 f_{(1)}(x_{(1)}) dx_{(1)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(1)}^2 f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 x_{(1)}^2 (1 - x_{(2)}) dx_{(2)} dx_{(1)} \\ &= 3! \frac{1}{2} \int_0^1 x_{(1)}^2 (1 - x_{(1)})^2 dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(1)})^3}{3} & h(x) &= x_{(1)}^2 \\ g'(x) &= (1 - x_{(1)})^2 & h'(x) &= 2x_{(1)} \end{aligned}$$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned} &= 3! \frac{1}{2} [\, 0 + 2 \int_0^1 x_{(1)} \frac{(1 - x_{(1)})^3}{3} dx_{(1)} \,] \\ &= 3! \, 2 \, \frac{1}{2} \, \frac{1}{3} \int_0^1 x_{(1)} (1 - x_{(1)})^3 dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(1)})^4}{4} & h(x) &= x_{(1)} \\ g'(x) &= (1 - x_{(1)})^3 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned} &= 3! \, 2 \, \frac{1}{2} \, \frac{1}{3} [\, 0 + \int_0^1 \frac{(1 - x_{(1)})^4}{4} dx_{(1)} \,] \\ &= 3! \, 2 \, \frac{1}{2} \, \frac{1}{3} \, \frac{1}{4} \int_0^1 (1 - x_{(1)})^4 dx_{(1)} \\ &= \frac{2}{4 \cdot 5} = \frac{2}{20} \end{aligned}$$

The variance of $x_{(1)}$ is then

$$Var(x_1) = E(x_{(1)}^2) - (E(x_{(1)}))^2 = \frac{2}{20} - (\frac{1}{4})^2 = \frac{3}{80}$$

Expected Value of $x_{(2)}$

$$\begin{aligned} E(x_{(2)}) &= \int_0^1 x_{(2)} f_{(2)}(x_{(2)}) dx_{(2)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(2)} f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 x_{(2)} (1 - x_{(2)})^1 dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$g(x) = -\frac{(1-x_{(2)})^2}{2}$ $\qquad$ $h(x) = x_{(2)}$

$g'(x) = (1 - x_{(2)})^1$ $\qquad$ $h'(x) = 1$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned} &= 3! \int_0^1 [\, x_{(1)} \frac{(1 - x_{(1)})^2}{2} + \int_{x_{(1)}}^1 \frac{(1 - x_{(2)})^2}{2} dx_{(2)} \,] \, dx_{(1)} \\ &= E(x_{(1)}) + 3! \, \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 (1 - x_{(2)})^2 dx_{(2)} dx_{(1)} \\ &= E(x_{(1)}) + \frac{1}{4} \\ &= \frac{1}{4} + \frac{1}{4} = \frac{2}{4} \end{aligned}$$

Variance of $x_{(2)}$

$$\begin{aligned} E(x_{(2)}^2) &= \int_0^1 x_{(2)}^2 f_{(2)}(x_{(2)}) dx_{(2)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(2)}^2 f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 x_{(2)}^2 (1 - x_{(2)})^1 dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(2)})^2}{2} & h(x) &= x_{(2)}^2 \\ g'(x) &= (1 - x_{(2)})^1 & h'(x) &= 2x_{(2)} \end{aligned}$$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned} &= 3! \int_0^1 [\, x_{(1)}^2 \frac{(1 - x_{(1)})^2}{2} + 2 \int_{x_{(1)}}^1 x_{(2)} \frac{(1 - x_{(2)})^2}{2} dx_{(2)} \,] \, dx_{(1)} \\ &= E(x_{(1)}^2) + 3!\, 2 \int_0^1 \int_{x_{(1)}}^1 x_{(2)} \frac{(1 - x_{(2)})^2}{2} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(2)})^3}{3} & h(x) &= x_{(2)} \\ g'(x) &= (1 - x_{(2)})^2 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned} &= E(x_{(1)}^2) + 3!\, 2\, \frac{1}{2} \int_0^1 [\, x_{(1)} \frac{(1 - x_{(1)})^3}{3} + \int_{x_{(1)}}^1 \frac{(1 - x_{(2)})^3}{3} dx_{(2)} \,] \, dx_{(1)} \\ &= E(x_{(1)}^2) + E(x_{(1)}^2) + 3!\, 2\, \frac{1}{2}\, \frac{1}{3} \int_0^1 \int_{x_{(1)}}^1 (1 - x_{(2)})^3 dx_{(2)} dx_{(1)} \\ &= 2E(x_{(1)}^2) + E(x_{(1)}^2) = \frac{2 \cdot 3}{2} E(x_{(1)}^2) \\ &= \frac{2 \cdot 3}{4 \cdot 5} = \frac{6}{20} \end{aligned}$$

The variance of $x_{(2)}$ is then

$$Var(x_2) = E(x_{(2)}^2) - (E(x_{(2)}))^2 = \frac{6}{20} - (\frac{2}{4})^2 = \frac{4}{80}$$

Expected Value of $x_{(3)}$

$$\begin{aligned} E(x_{(3)}) &= \int_0^1 x_{(3)} f_{(3)}(x_{(3)}) dx_{(3)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(3)} f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(3)} (1 - x_{(3)})^0 dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$g(x) = -\frac{(1-x_{(3)})^1}{1}$ $\qquad$ $h(x) = x_{(3)}$

$g'(x) = (1 - x_{(3)})^0$ $\qquad$ $h'(x) = 1$

Integrating over $dx_{(3)}$ by parts gives

$$\begin{aligned} &= 3! \int_0^1 \int_{x_{(1)}}^1 [\, x_{(2)}(1 - x_{(2)}) + \int_{x_{(2)}}^1 (1 - x_{(3)}) dx_{(3)} \,] \, dx_{(2)} dx_{(1)} \\ &= E(x_{(2)}) + 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 (1 - x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\ &= E(x_{(2)}) + \frac{1}{4} \\ &= E(x_{(1)}) + \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4} \end{aligned}$$

Variance of $x_{(3)}$

$$\begin{aligned} E(x_{(3)}^2) &= \int_0^1 x_{(3)}^2 f_{(3)}(x_{(3)}) dx_{(3)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(3)}^2 f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 3! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(3)}^2 (1 - x_{(3)})^0 dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\frac{(1-x_{(3)})^1}{1} & h(x) &= x_{(3)}^2 \\ g'(x) &= (1 - x_{(3)})^0 & h'(x) &= 2x_{(3)} \end{aligned}$$

Integrating over $dx_{(3)}$ by parts gives

$$\begin{aligned} &= 3! \int_0^1 \int_{x_{(1)}}^1 [\, x_{(2)}^2 (1 - x_{(2)}) + 2 \int_{x_{(2)}}^1 x_{(3)} (1 - x_{(3)}) dx_{(3)} \,] \, dx_{(2)} dx_{(1)} \\ &= E(x_{(2)}^2) + 3!\, 2 \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(3)} (1 - x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\frac{(1-x_{(3)})^2}{2} & h(x) &= x_{(3)} \\ g'(x) &= (1 - x_{(3)})^1 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(3)}$ by parts gives

$$\begin{aligned} &= E(x_{(2)}^2) + 3!\, 2 \int_0^1 \int_{x_{(1)}}^1 [\, x_{(2)} \frac{(1 - x_{(2)})^2}{2} + \int_{x_{(2)}}^1 \frac{(1 - x_{(3)})^2}{2} dx_{(3)} \,] \, dx_{(2)} dx_{(1)} \\ &= E(x_{(2)}^2) + E(x_{(2)}^2) - E(x_{(1)}^2) + 3!\, 2\, \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 (1 - x_{(3)})^2 dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 2E(x_{(2)}^2) - E(x_{(1)}^2) + E(x_{(1)}^2) = \frac{3 \cdot 4}{2} E(x_{(1)}^2) \\ &= \frac{3 \cdot 4}{4 \cdot 5} = \frac{12}{20} \end{aligned}$$

The variance of $x_{(3)}$ is then

$$Var(x_3) = E(x_{(3)}^2) - (E(x_{(3)}))^2 = \frac{12}{20} - (\frac{3}{4})^2 = \frac{3}{80}$$

1.2.3 A Random Sample of 4 Points from a Uniform Distribution

Let x_1, x_2, x_3, x_4 be independent identically distributed random variables from a $\mathcal{U}(0,1)$ distribution, which represent a random sample of four points on a line of length $L = 1$. Let $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)}$ denote the order statistics of the x's, which represent the position of the points on the line.

The joint density function of the four order statistics is:

$f_{(1),(2),(3),(4)}(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}) = 4! \qquad 0 \leq x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)} \leq 1$

The marginal density function of $x_{(i)}$, where $i = 1, 2, 3, 4$, can be obtained from the joint density function by integrating out the other variables, and then be used to calculate the expected value and the variance of $x_{(i)}$.

Expected Value of $x_{(1)}$

$$\begin{aligned}
E(x_{(1)}) &= \int_0^1 x_{(1)} f_{(1)}(x_{(1)}) dx_{(1)} \\
&= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(1)} f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) f(x_{(4)}) dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\
&= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(1)} (1 - x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\
&= 4! \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 x_{(1)} (1 - x_{(2)})^2 dx_{(2)} dx_{(1)} \\
&= 4! \frac{1}{2} \frac{1}{3} \int_0^1 x_{(1)} (1 - x_{(1)})^3 dx_{(1)}
\end{aligned}$$

Let

$$\begin{aligned}
g(x) &= -\tfrac{(1-x_{(1)})^4}{4} & h(x) &= x_{(1)} \\
g'(x) &= (1 - x_{(1)})^3 & h'(x) &= 1
\end{aligned}$$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned}
&= 4! \frac{1}{2} \frac{1}{3} [\, 0 + \int_0^1 \frac{(1 - x_{(1)})^4}{4} dx_{(1)} \,] \\
&= 4! \frac{1}{2} \frac{1}{3} \frac{1}{4} \int_0^1 (1 - x_{(1)})^4 dx_{(1)} \\
&= \frac{1}{5}
\end{aligned}$$

Variance of $x_{(1)}$

$$\begin{aligned}
E(x_{(1)}^2) &= \int_0^1 x_{(1)}^2 f_{(1)}(x_{(1)})dx_{(1)} \\
&= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(1)}^2 f(x_{(1)})f(x_{(2)})f(x_{(3)})f(x_{(4)})dx_{(4)}dx_{(3)}dx_{(2)}dx_{(1)} \\
&= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(1)}^2(1-x_{(3)})dx_{(3)}dx_{(2)}dx_{(1)} \\
&= 4! \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 x_{(1)}^2(1-x_{(2)})^2 dx_{(2)}dx_{(1)} \\
&= 4! \frac{1}{2}\frac{1}{3} \int_0^1 x_{(1)}^2(1-x_{(1)})^3 dx_{(1)}
\end{aligned}$$

Let

$$\begin{aligned}
g(x) &= -\tfrac{(1-x_{(1)})^4}{4} & h(x) &= x_{(1)}^2 \\
g'(x) &= (1-x_{(1)})^3 & h'(x) &= 2x_{(1)}
\end{aligned}$$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned}
&= 4! \frac{1}{2}\frac{1}{3}[\,0 + 2\int_0^1 x_{(1)} \frac{(1-x_{(1)})^4}{4} dx_{(1)} \\
&= 4!\, 2 \frac{1}{2}\frac{1}{3}\frac{1}{4} \int_0^1 x_{(1)}(1-x_{(1)})^4 dx_{(1)}
\end{aligned}$$

Let

$$\begin{aligned}
g(x) &= -\tfrac{(1-x_{(1)})^5}{5} & h(x) &= x_{(1)} \\
g'(x) &= (1-x_{(1)})^4 & h'(x) &= 1
\end{aligned}$$

Integrating over $dx_{(1)}$ by parts gives

$$\begin{aligned}
&= 4!\, 2 \frac{1}{2}\frac{1}{3}\frac{1}{4}[\,0 + \int_0^1 \frac{(1-x_{(1)})^5}{5} dx_{(1)}\,] \\
&= 4!\, 2 \frac{1}{2}\frac{1}{3}\frac{1}{4}\frac{1}{5} \int_0^1 (1-x_{(1)})^5 dx_{(1)} \\
&= \frac{2}{5 \cdot 6} = \frac{2}{30}
\end{aligned}$$

The variance of $x_{(1)}$ is then

$$Var(x_1) = E(x_{(1)}^2) - (E(x_{(1)}))^2 = \frac{2}{30} - (\frac{1}{5})^2 = \frac{4}{150}$$

Expected Value of $x_{(2)}$

$$\begin{aligned}
E(x_{(2)}) &= \int_0^1 x_{(2)} f_{(2)}(x_{(2)}) dx_{(2)} \\
&= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(2)} f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) f(x_{(4)}) dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\
&= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(2)} (1 - x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\
&= 4! \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 x_{(2)} (1 - x_{(2)})^2 dx_{(2)} dx_{(1)}
\end{aligned}$$

Let

$$\begin{aligned}
g(x) &= -\tfrac{(1-x_{(2)})^3}{3} & h(x) &= x_{(2)} \\
g'(x) &= (1 - x_{(2)})^2 & h'(x) &= 1
\end{aligned}$$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned}
&= 4! \frac{1}{2} \int_0^1 \left[x_{(1)} \frac{(1 - x_{(1)})^3}{3} + \int_{x_{(1)}}^1 \frac{(1 - x_{(2)})^3}{3} dx_{(2)} \right] dx_{(1)} \\
&= E(x_{(1)}) + 4! \frac{1}{2} \frac{1}{3} \int_0^1 \int_{x_{(1)}}^1 (1 - x_{(2)})^3 dx_{(2)} dx_{(1)} \\
&= E(x_{(1)}) + \frac{1}{5} \\
&= \frac{1}{5} + \frac{1}{5} = \frac{2}{5}
\end{aligned}$$

Variance of $x_{(2)}$

$$\begin{aligned}
E(x_{(2)}^2) &= \int_0^1 x_{(2)}^2 f_{(2)}(x_{(2)}) dx_{(2)} \\
&= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(2)}^2 f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) f(x_{(4)}) dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\
&= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(2)}^2 (1 - x_{(3)}) dx_{(3)} dx_{(2)} dx_{(1)} \\
&= 4! \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 x_{(2)}^2 (1 - x_{(2)})^2 dx_{(2)} dx_{(1)}
\end{aligned}$$

Let

$$\begin{aligned}
g(x) &= -\tfrac{(1-x_{(2)})^3}{3} & h(x) &= x_{(2)}^2 \\
g'(x) &= (1 - x_{(2)})^2 & h'(x) &= 2x_{(2)}
\end{aligned}$$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned}
&= 4! \frac{1}{2} \int_0^1 [\, x_{(1)}^2 \frac{(1 - x_{(1)})^3}{3} + 2 \int_{x_{(1)}}^1 x_{(2)} \frac{(1 - x_{(2)})^3}{3} dx_{(2)} \,] dx_{(1)} \\
&= E(x_{(1)}^2) + 4! \, 2 \, \frac{1}{2} \frac{1}{3} \int_0^1 \int_{x_{(1)}}^1 x_{(2)} (1 - x_{(2)})^3 dx_{(2)} dx_{(1)}
\end{aligned}$$

$$\begin{aligned}
g(x) &= -\tfrac{(1-x_{(2)})^4}{4} & h(x) &= x_{(2)} \\
g'(x) &= (1 - x_{(2)})^3 & h'(x) &= 1
\end{aligned}$$

Integrating over $dx_{(2)}$ by parts gives

$$\begin{aligned}
&= E(x_{(1)}^2) + 4! \, 2 \, \frac{1}{2} \frac{1}{3} \int_0^1 [\, x_{(1)} \frac{(1 - x_{(1)})^4}{4} + \int_{x_{(1)}}^1 \frac{(1 - x_{(2)})^4}{4} dx_{(2)} \,] dx_{(1)} \\
&= E(x_{(1)}^2) + E(x_{(1)}^2) + 4! \, 2 \, \frac{1}{2} \frac{1}{3} \frac{1}{4} \int_0^1 \int_{x_{(1)}}^1 (1 - x_{(2)})^4 dx_{(2)} dx_{(1)} \\
&= 2E(x_{(1)}^2) + E(x_{(1)}^2) = \frac{2 \cdot 3}{2} E(x_{(1)}^2) \\
&= \frac{2 \cdot 3}{5 \cdot 6} = \frac{6}{30}
\end{aligned}$$

The variance of $x_{(2)}$ is then

$$Var(x_2) = E(x_{(2)}^2) - (E(x_{(2)}))^2 = \frac{6}{30} - (\frac{2}{5})^2 = \frac{6}{150}$$

Expected Value of $x_{(3)}$

$$\begin{aligned} E(x_{(3)}) &= \int_0^1 x_{(3)} f_{(3)}(x_{(3)}) dx_{(3)} \\ &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(3)} f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) f(x_{(4)}) dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(3)} (1 - x_{(3)})^1 dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(3)})^2}{2} & h(x) &= x_{(3)} \\ g'(x) &= (1 - x_{(3)})^1 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(3)}$ by parts gives

$$\begin{aligned} &= 4! \int_0^1 \int_{x_{(1)}}^1 [\, x_{(2)} \frac{(1 - x_{(2)})^2}{2} + \int_{x_{(2)}}^1 \frac{(1 - x_{(3)})^2}{2} dx_{(3)} \,] \, dx_{(2)} dx_{(1)} \\ &= E(x_{(2)}) + 4! \, \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 (1 - x_{(3)})^2 dx_{(3)} dx_{(2)} dx_{(1)} \\ &= E(x_{(2)}) + \frac{1}{5} \\ &= E(x_{(1)}) + \frac{1}{5} + \frac{1}{5} \\ &= \frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{3}{5} \end{aligned}$$

Variance of $x_{(3)}$

$$\begin{aligned} E(x_{(3)}^2) &= \int_0^1 x_{(3)}^2 f_{(3)}(x_{(3)}) dx_{(3)} \\ &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(3)}^2 f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) f(x_{(4)}) dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(3)}^2 (1 - x_{(3)})^1 dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(3)})^2}{2} & h(x) &= x_{(3)}^2 \\ g'(x) &= (1 - x_{(3)})^1 & h'(x) &= 2x_{(3)} \end{aligned}$$

Integrating over $dx_{(3)}$ by parts gives

$$\begin{aligned} &= 4! \int_0^1 \int_{x_{(1)}}^1 [\, x_{(2)}^2 \frac{(1 - x_{(2)})^2}{2} + \int_{x_{(2)}}^1 x_{(3)} (1 - x_{(3)})^2 dx_{(3)} \,] \, dx_{(2)} dx_{(1)} \\ &= E(x_{(2)}^2) + 4!\, 2\, \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 x_{(3)} (1 - x_{(3)})^2 dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(3)})^3}{3} & h(x) &= x_{(3)} \\ g'(x) &= (1 - x_{(3)})^2 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(3)}$ by parts gives

$$\begin{aligned} &= E(x_{(2)}^2) + 4!\, 2\, \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 [\, x_{(2)} \frac{(1 - x_{(2)})^3}{3} + \int_{x_{(2)}}^1 \frac{(1 - x_{(3)})^3}{3} dx_{(3)} \,] \, dx_{(2)} dx_{(1)} \\ &= E(x_{(2)}^2) + E(x_{(2)}^2) - E(x_{(1)}^2) + 4!\, 2\, \frac{1}{2} \frac{1}{3} \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 (1 - x_{(3)})^3 dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 2E(x_{(2)}^2) - E(x_{(1)}^2) + E(x_{(1)}^2) = \frac{3 \cdot 4}{2} E(x_{(1)}^2) \\ &= \frac{3 \cdot 4}{5 \cdot 6} = \frac{12}{30} \end{aligned}$$

The variance of $x_{(3)}$ is then

$$Var(x_3) = E(x_{(3)}^2) - (E(x_{(3)}))^2 = \frac{12}{30} - (\frac{3}{5})^2 = \frac{6}{150}$$

Expected Value of $x_{(4)}$

$$\begin{aligned} E(x_{(4)}) &= \int_0^1 x_{(4)} f_{(4)}(x_{(4)}) dx_{(4)} \\ &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(4)} f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) f(x_{(4)}) dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(4)} (1 - x_{(4)})^0 dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\frac{(1-x_{(4)})^1}{1} & h(x) &= x_{(4)} \\ g'(x) &= (1 - x_{(4)})^0 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(4)}$ by parts gives

$$\begin{aligned} &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 [\, x_{(3)}(1 - x_{(3)}) + \int_{x_{(3)}}^1 (1 - x_{(4)}) dx_{(4)} \,] \, dx_{(3)} dx_{(2)} dx_{(1)} \\ &= E(x_{(3)}) + 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 (1 - x_{(4)}) dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\ &= E(x_{(3)}) + \frac{1}{5} \\ &= E(x_{(2)}) + \frac{1}{5} + \frac{1}{5} \\ &= E(x_{(1)}) + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} \\ &= \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{4}{5} \end{aligned}$$

Variance of $x_{(4)}$

$$\begin{aligned} E(x_{(4)}^2) &= \int_0^1 x_{(4)}^2 f_{(4)}(x_{(4)}) dx_{(4)} \\ &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(4)}^2 f(x_{(1)}) f(x_{(2)}) f(x_{(3)}) f(x_{(4)}) dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(4)}^2 (1 - x_{(4)})^0 dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(4)})^1}{1} & h(x) &= x_{(4)}^2 \\ g'(x) &= (1 - x_{(4)})^0 & h'(x) &= 2x_{(4)} \end{aligned}$$

Integrating over $dx_{(4)}$ by parts gives

$$\begin{aligned} &= 4! \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 [\, x_{(3)}^2 (1 - x_{(3)}) + 2 \int_{x_{(3)}}^1 x_{(4)}) (1 - x_{(4)}) dx_{(4)} \,] \, dx_{(3)} dx_{(2)} dx_{(1)} \\ &= E(x_{(3)}^2) + 4!\, 2 \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 x_{(4)}) (1 - x_{(4)})^1 dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \end{aligned}$$

Let

$$\begin{aligned} g(x) &= -\tfrac{(1-x_{(4)})^2}{2} & h(x) &= x_{(4)} \\ g'(x) &= (1 - x_{(4)})^1 & h'(x) &= 1 \end{aligned}$$

Integrating over $dx_{(4)}$ by parts gives

$$\begin{aligned} &= E(x_{(3)}^2) + 4!\, 2 \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 [\, x_{(3)} \frac{(1 - x_{(3)})^2}{2} + \int_{x_{(3)}}^1 \frac{(1 - x_{(4)})^2}{2} dx_{(4)} \,] \, dx_{(3)} dx_{(2)} dx_{(1)} \\ &= E(x_{(3)}^2) + E(x_{(3)}^2) - E(x_{(2)}^2) + 4!\, 2\, \frac{1}{2} \int_0^1 \int_{x_{(1)}}^1 \int_{x_{(2)}}^1 \int_{x_{(3)}}^1 (1 - x_{(4)})^2 dx_{(4)} dx_{(3)} dx_{(2)} dx_{(1)} \\ &= 2E(x_{(3)}^2) - E(x_{(2)}^2) + E(x_{(1)}^2) = \frac{4 \cdot 5}{2} E(x_{(1)}^2) \\ &= \frac{4 \cdot 5}{5 \cdot 6} = \frac{20}{30} \end{aligned}$$

The variance of $x_{(4)}$ is then

$$Var(x_4) = E(x_{(4)}^2) - (E(x_{(4)}))^2 = \frac{20}{30} - (\frac{4}{5})^2 = \frac{4}{150}$$

1.2.4 The General Case

A Random Sample of n Points from a Uniform Distribution

Let $x_1, \ldots, x_n$ be independent identically distributed random variables from a $\mathcal{U}(0,1)$ distribution, which represent a random sample of n points on a line of length $L = 1$. Let $x_{(1)} \leq \ldots \leq x_{(n)}$ denote the order statistics of the x's, which represent the position of the points on the line. The joint density function of the n order statistics is:

$f_{(1),\ldots,(n)}(x_{(1)}, \ldots, x_{(n)}) = n! \qquad 0 \leq x_{(1)} \leq \ldots \leq x_{(n)} \leq 1$

The marginal density function of $x_{(i)}$, where $i = 1, \ldots, n$, can be obtained from the joint density function by integrating out the other variables.

$$\begin{aligned} f_{(i)}(x_{(i)}) &= n! f(x_{(i)}) \\ &\cdot \int_0^{x_{(i)}} \int_0^{x_{(i-1)}} \cdots \int_0^{x_{(3)}} \int_0^{x_{(2)}} f(x_{(1)}) \ldots f(x_{(i-1)}) dx_{(1)} \ldots dx_{(i-1)} \\ &\cdot \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-2)}}^1 \int_{x_{(n-1)}}^1 f(x_{(i+1)}) \ldots f(x_{(n)}) dx_{(n)} \ldots dx_{(i+1)} \end{aligned}$$

Then it can be used to calculate the expected value and the variance of $x_{(i)}$.

Expected Value of $x_{(i)}$

$$\begin{aligned} E(x_{(i)}) &= \int_0^1 x_{(i)} f_{(i)}(x_{(i)}) dx_{(i)} \\ &= n! \int_0^1 \int_{x_{(1)}}^1 \cdots \int_{x_{(i-1)}}^1 \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-1)}}^1 x_{(i)} dx_{(n)} \ldots dx_{(1)} \\ &= n! \int_0^1 \int_{x_{(1)}}^1 \cdots \int_{x_{(i-1)}}^1 \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-2)}}^1 x_{(i)} (1 - x_{(n-1)})^1 dx_{(n-1)} \ldots dx_{(1)} \\ &= \frac{n!}{2} \int_0^1 \int_{x_{(1)}}^1 \cdots \int_{x_{(i-1)}}^1 \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-3)}}^1 x_{(i)} (1 - x_{(n-2)})^2 dx_{(n-2)} \ldots dx_{(1)} \\ &= \frac{n!}{2 \cdot 3} \int_0^1 \int_{x_{(1)}}^1 \cdots \int_{x_{(i-1)}}^1 \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-4)}}^1 x_{(i)} (1 - x_{(n-3)})^3 dx_{(n-3)} \ldots dx_{(1)} \\ &= \\ &\vdots \\ &= \frac{n!}{2 \cdot 3 \cdot \ldots \cdot (n-i)} \int_0^1 \cdots \int_{x_{(i-1)}}^1 x_{(i)} (1 - x_{(i)})^{(n-i)} dx_{(i)} \ldots dx_{(1)} \end{aligned}$$

Integrating over $dx_{(i)}$ by parts gives

$$= \frac{n!}{2\cdot 3\cdot \ldots \cdot (n-i)} \int_0^1 \cdots \int_{x_{(i-2)}}^1 [x_{(i-1)} \frac{(1-x_{(i-1)})^{(n-i+1)}}{n-i+1} + \int_{x_{(i-1)}}^1 \frac{(1-x_{(i)})^{(n-i+1)}}{n-i+1} dx_{(i)}] \ldots dx_{(1)}$$

$$= E(x_{(i-1)}) + \frac{n!}{2\cdot 3\cdot \ldots \cdot (n-i)\cdot (n-i+1)} \int_0^1 \cdots \int_{x_{(i-1)}}^1 (1-x_{(i)})^{(n-i+1)} dx_{(i)} \ldots dx_{(1)}$$

$$= E(x_{(i-1)}) + \frac{n!}{2\cdot 3\cdot \ldots \cdot (n-i+1)\cdot (n-i+2)} \int_0^1 \cdots \int_{x_{(i-2)}}^1 (1-x_{(i-1)})^{(n-i+2)} dx_{(i-1)} \ldots dx_{(1)}$$

$$=$$

$$\vdots$$

$$= E(x_{(i-1)}) + \frac{n!}{2\cdot 3\cdot \ldots \cdot (n-i)\cdot (n-i+1)\cdot (n-i+2)\ldots (n+1)}$$

$$= E(x_{(i-1)}) + \frac{1}{n+1}$$

Applying this relation successively gives

$$\begin{aligned} E(x_{(i)}) &= E(x_{(i-1)}) + \frac{1}{n+1} \\ &= E(x_{(i-2)}) + \frac{2}{n+1} \\ &= E(x_{(i-3)}) + \frac{3}{n+1} \\ &= \\ &\vdots \\ &= E(x_{(i-(i-2))} + \frac{i-2}{n+1} \\ &= E(x_{(1)}) + \frac{i-1}{n+1} \\ &= \frac{1}{n+1} + \frac{i-1}{n+1} \\ &= \frac{i}{n+1} \end{aligned}$$

Variance of $x_{(i)}$

$$
\begin{aligned}
E(x_{(i)}^2) &= \int_0^1 x_{(i)}^2 f_{(i)}(x_{(i)}) dx_{(i)} \\
&= n! \int_0^1 \int_{x_{(1)}}^1 \cdots \int_{x_{(i-1)}}^1 \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-1)}}^1 x_{(i)}^2 dx_{(n)} \ldots dx_{(1)} \\
&= n! \int_0^1 \int_{x_{(1)}}^1 \cdots \int_{x_{(i-1)}}^1 \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-2)}}^1 x_{(i)}^2 (1 - x_{(n-1)})^1 dx_{(n-1)} \ldots dx_{(1)} \\
&= \frac{n!}{2} \int_0^1 \int_{x_{(1)}}^1 \cdots \int_{x_{(i-1)}}^1 \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-3)}}^1 x_{(i)}^2 (1 - x_{(n-2)})^2 dx_{(n-2)} \ldots dx_{(1)} \\
&= \frac{n!}{2 \cdot 3} \int_0^1 \int_{x_{(1)}}^1 \cdots \int_{x_{(i-1)}}^1 \int_{x_{(i)}}^1 \int_{x_{(i+1)}}^1 \cdots \int_{x_{(n-4)}}^1 x_{(i)}^2 (1 - x_{(n-3)})^3 dx_{(n-3)} \ldots dx_{(1)} \\
&= \\
&\vdots \\
&= \frac{n!}{2 \cdot 3 \cdot \ldots \cdot (n-i)} \int_0^1 \cdots \int_{x_{(i-1)}}^1 x_{(i)}^2 (1 - x_{(i)})^{(n-i)} dx_{(i)} \ldots dx_{(1)}
\end{aligned}
$$

Integrating over $dx_{(i)}$ by parts gives

$$
\begin{aligned}
&= \frac{n!}{2 \cdot 3 \cdot \ldots \cdot (n-i)} \int_0^1 \cdots \int_{x_{(i-2)}}^1 \\
&\quad [\, x_{(i-1)}^2 \frac{(1 - x_{(i-1)})^{(n-i+1)}}{n-i+1} + 2 \int_{x_{(i-1)}}^1 x_{(i)} \frac{(1 - x_{(i)})^{(n-i+1)}}{n-i+1} dx_{(i)} \,] \ldots dx_{(1)} \\
&= E(x_{(i-1)}^2) + \frac{2n!}{2 \cdot 3 \cdot \ldots \cdot (n-i) \cdot (n-i+1)} \int_0^1 \cdots \int_{x_{(i-1)}}^1 x_{(i)} (1 - x_{(i)})^{(n-i+1)} dx_{(i)} \ldots dx_{(1)}
\end{aligned}
$$

Integrating over $dx_{(i)}$ by parts gives

$$
\begin{aligned}
&= E(x_{(i-1)}^2) + \frac{2n!}{2 \cdot 3 \cdot \ldots \cdot (n-i) \cdot (n-i+1)} \int_0^1 \cdots \int_{x_{(i-2)}}^1 \\
&\quad [\, x_{(i-1)} \frac{(1 - x_{(i-1)})^{(n-i+2)}}{n-i+2} + \int_{x_{(i-1)}}^1 \frac{(1 - x_{(i)})^{(n-i+2)}}{n-i+2} dx_{(i)} \,] \ldots dx_{(1)}
\end{aligned}
$$

$$= E(x_{(i-1)}^2) + E(x_{(i-1)}^2) - E(x_{(i-2)}^2)$$

$$+ \frac{2n!}{2 \cdot 3 \cdot \ldots \cdot (n-i+1) \cdot (n-i+2)} \int_0^1 \cdots \int_{x_{(i-1)}}^1 (1 - x_{(i)})^{(n-i+2)} dx_{(i)} \ldots dx_{(1)}$$

$$= E(x_{(i-1)}^2) + E(x_{(i-1)}^2) - E(x_{(i-2)}^2)$$

$$+ \frac{2n!}{2 \cdot 3 \cdot \ldots \cdot (n-i+2) \cdot (n-i+3)} \int_0^1 \cdots \int_{x_{(i-2)}}^1 (1 - x_{(i-1)})^{(n-i+3)} dx_{(i-1)} \ldots dx_{(1)}$$

$$=$$

$$\vdots$$

$$= 2E(x_{(i-1)}^2) - E(x_{(i-2)}^2) + E(x_{(1)}^2)$$

Applying this relation successively gives

$$\begin{aligned}
E(x_{(i)}^2) &= 2E(x_{(i-1)}^2) - E(x_{(i-2)}^2) + E(x_{(1)}^2) \\
&= 3E(x_{(i-2)}^2) - 2E(x_{(i-3)}^2) + 3E(x_{(1)}^2) \\
&= 4E(x_{(i-3)}^2) - 3E(x_{(i-4)}^2) + 6E(x_{(1)}^2) \\
&= \\
&\vdots \\
&= (i-2)E(x_{(i-(i-3))}^2) - (i-3)E(x_{(i-(i-2))}^2) + \frac{(i-2)(i-3)}{2} E(x_{(1)}^2) \\
&= (i-1)E(x_{(2)}^2) - (i-2)E(x_{(1)}^2) + \frac{(i-1)(i-2)}{2} E(x_{(1)}^2) \\
&= 3(i-1)E(x_{(1)}^2) - (i-2)E(x_{(1)}^2) + \frac{(i-1)(i-2)}{2} E(x_{(1)}^2) \\
&= \frac{i(i+1)}{2} E(x_{(1)}^2) \\
&= \frac{i(i+1)}{(n+1)(n+2)}
\end{aligned}$$

The variance of $x_{(i)}$ is then

$$\begin{aligned}
Var(x_i) &= E(x_{(i)}^2) - (E(x_{(i)}))^2 \\
&= \frac{i(i+1)}{(n+1)(n+2)} - (\frac{i}{(n+1)})^2 \\
&= \frac{(\frac{i}{(n+1)})(1 - \frac{i}{(n+1)})}{n+2}
\end{aligned}$$

Covariance of $x_{(i)}$ and $x_{(j)}$

$$\begin{aligned} Cov(x_i, x_j) &= E(x_i x_j) - E(x_i)\,E(x_j) \\ &= E(x_i x_j) - \frac{ij}{(n+1)^2} \\ &= \int_0^1 \int_{x_{(i)}}^1 x_{(i)} x_{(j)} f_{(i),(j)}(x_{(i)}, x_{(j)}) dx_{(j)} dx_{(i)} - \frac{ij}{(n+1)^2} \end{aligned}$$

The calculation of the product moment $E(x_{(i)}x_{(j)})$ involves the joint density function $f_{(i),(j)}(x_{(i)}, x_{(j)})$. This density can be obtained by considering the joint density function of all n order statistics and then integrating out the other variables, $(x_{(1)}, \ldots, x_{(i-1)})$, $(x_{(i+1)}, \ldots, x_{(j-1)})$, and $(x_{(j+1)}, \ldots, x_{(n)})$

$$\begin{aligned} f_{(i),(j)}(x_{(i)}, x_{(j)}) &= n! f(x_{(i)}) f(x_{(j)}) \\ &= \cdot \int_0^{x_{(i)}} \cdots \int_0^{x_{(2)}} f(x_{(1)}) \cdots f(x_{(i-1)}) dx_{(1)} \cdots dx_{(i-1)} \\ &\quad \cdot \int_{x_{(i)}}^{x_{(j)}} \cdots \int_{x_{(i)}}^{x_{(i+2)}} f(x_{(i+2)}) \cdots f(x_{(j-1)}) dx_{(i+1)} \cdots dx_{(j-1)} \\ &\quad \cdot \int_{x_{(j)}}^{1} \cdots \int_{x_{(n-1)}}^{1} f(x_{(j+1)}) \cdots f(x_{(n)}) dx_{(n)} \cdots dx_{(j+1)} \\ &= n! \, \frac{x_{(i)}^{(i-1)}}{(i-1)!} \, \frac{(x_{(j)} - x_{(i)})^{(j-i-1)}}{(j-i-1)!} \, \frac{(1 - x_{(j)})^{(n-j)}}{(n-j)!} \end{aligned}$$

The product moment of $x_{(i)}$ and $x_{(j)}$ is then

$$\begin{aligned} E(x_i x_j) &= \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} \int_0^1 \int_0^{x_{(j)}} x_{(i)}^i (x_{(j)} - x_{(i)})^{(j-i-1)} x_{(j)} (1 - x_{(j)})^{(n-j)} dx_{(i)} dx_{(j)} \\ &= \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} \, \frac{i!\,(j-i-1)!}{j!} \int_0^1 x_{(j)}^j (1 - x_{(j)})^{(n-j)} dx_{(j)} \\ &= \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} \, \frac{i!\,(j-i-1)!}{j!} \, \frac{(j+1)!\,(n-j)!}{(n+2)!} \\ &= \frac{i\,(j+1)}{(n+1)\,(n+2)} \end{aligned}$$

and the covariance of $x_{(i)}$ and $x_{(j)}$ is

$$\begin{aligned} Cov(x_i, x_j) &= E(x_i x_j) - E(x_i)\, E(x_j) \\ &= \frac{i(j+1)}{(n+1)\,(n+2)} - \frac{ij}{(n+1)^2} \\ &= \frac{(\frac{i}{(n+1)})(1 - \frac{j}{(n+1)})}{n+2} \end{aligned}$$

1.2.5 The Homogeneous Poisson Process

In the previous sections the derivations of the moments of the order statistics were based on the model which postulates that n points are placed at random along an interval according to a uniform probability distribution. Those familiar with the areal nearest neighbor analysis will recall that a different model, the homogeneous Poisson process, is used to obtain the distribution of the nearest neighbor distances. According to the model, the number of points falling in a region with area A is assumed to have a Poisson distribution with expected number of points λA, where λ is the rate of the Poisson process.

A similar assumption can be made in the one dimension case. That is, the number of points falling on an interval of length L has a Poisson distribution with expected number of points λL, where λ is the rate of the Poisson process. At first glance this model does not appear to be related to the previous one, which postulates that the points are independent identically distributed random variables from a uniform distribution. However, a fundamental property of the homogeneous Poisson process is conditional uniformity. That is, given the number of points falling on an interval of specified length, and regardless of the rate λ, the conditional distribution of the ordered points is that of order statistics engendered by independent random variables, each uniformly distributed on the interval. This property allows the use of the theory of order statistics in the calculations of the moments of the nearest neighbor distances of random points from a one–dimensional homogenous Poisson process. The derivations of the expected value and the variance of the nearest neighbor distances and of their mean are described in the next two sections.

1.3 Moments of Nearest Neighbor Distance

As before, let $x_{(1)} \leq \ldots \leq x_{(n)}$ denote the order statistics of n independent identically distributed random points from a $\mathcal{U}(0,1)$ distribution or from an homogeneous Poisson process, conditioning on n.

Let $l_i = (x_{(i)} - x_{(i-1)}) + (x_{(i+1)} - x_{(i)}) = x_{(i+1)} - x_{(i-1)} \qquad 0 \leq l_i \leq 1, \qquad i = 2, \ldots, n-1$

Let $d_i = min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,]$ be the nearest neighbor distance of $x_{(i)}$

Then, according to Tsuji [11], given $x_{(i-1)}$ and $x_{(i+1)}$, there is no preferred position for $x_{(i)}$. That is, the conditional probability distribution of $x_{(i)}$ is constant within the region l_i. Therefore, given l_i, the conditional distribution of d_i is also uniform within l_i and defined as

$$f(d_i \mid l_i) = \begin{cases} \frac{1}{\frac{l_i}{2}} & \text{if } 0 \leq d_i \leq \frac{l_i}{2} \\ 0 & \text{otherwise} \end{cases}$$

Expected Value of d_i

Using the conditional distribution of d_i, its conditional expectation is

$$\begin{aligned} E(d_i \mid l_i) &= \int_0^{\frac{l_i}{2}} d_i \, f(d_i \mid l_i) \, d\, d_i \\ &= \frac{1}{\frac{l_i}{2}} \int_0^{\frac{l_i}{2}} d_i \, d\, d_i \\ &= \frac{2}{l_i} \frac{1}{2} \frac{(l_i)^2}{4} \\ &= \frac{l_i}{4} \end{aligned}$$

and the expected value of d_i is

$$\begin{aligned} E(d_i) = E[E(d_i \mid l_i)] &= E(\frac{l_i}{4}) \\ &= \frac{1}{4} E(l_i) \\ &= \frac{1}{4} E(x_{(i+1)} - x_{(i-1)}) \\ &= \frac{1}{4} (\frac{i+1}{n+1} - \frac{i-1}{n+1}) \\ &= \frac{1}{2(n+1)} \end{aligned}$$

Variance of d_i

$$\begin{aligned} Var(d_i) &= E(d_i^2) - (E(d_i))^2 \\ &= \int_0^{\frac{1}{2}} d_i^2 \, f(d_i) \, d\, d_i - (\frac{1}{2(n+1)})^2 \end{aligned}$$

The marginal density function of d_i can be derived by considering the joint density function of d_i and l_i and then integrating out l_i. The joint density function of d_i and l_i is no more than the product of the marginal density function of l_i and the conditional density function of d_i, given l_i, which is uniform within l_i.

Since l_i is a range of two order statistics, its marginal density function can be easily obtained from the basic theory of order statistics. It is well known that if $W_{j,k} = x_{(k)} - x_{(j)}$ is a range of two order statistics from n independent identically distributed $\mathcal{U}(0,1)$ random variables, it has a $\mathcal{B}eta(k-j,\ n-k+j+1)$ distribution, and hence depends only on $k-j$ and not on k and j individually. Especially, if $k = i+1$ and $j = i-1$, then $W_{(i-1),(i+1)} = l_i = x_{(i+1)} - x_{(i-1)}$ has a $\mathcal{B}eta(2,\ n-1)$ distribution.

It follows that the marginal density function of l_i is

$f(l_i) = n\,(n-1)\,l_i\,(1-l_i)^{(n-2)}$ for $0 \le l_i \le 1$

and the joint density function of d_i and l_i is

$$\begin{aligned} f(d_i, l_i) &= f(d_i \mid l_i)\, f(l_i) \\ &= \frac{2}{l_i}\, n\,(n-1)\, l_i\,(1-l_i)^{(n-2)} \\ &= 2\,n\,(n-1)\,(1-l_i)^{(n-2)} \end{aligned}$$

Integrating out l_i from the above joint density function gives the marginal density function of d_i

$$\begin{aligned} f(d_i) &= \int_{2di}^{1} f(d_i, l_i)\, d\, l_i \\ &= \int_{2di}^{1} 2\,n\,(n-1)\,(1-l_i)^{(n-2)} d\, l_i \\ &= 2\,n\,(n-1) \int_{2di}^{1} (1-l_i)^{(n-2)} d\, l_i \\ &= 2\,n\,(1-2\,di)^{(n-1)} \end{aligned}$$

The variance of d_i is then

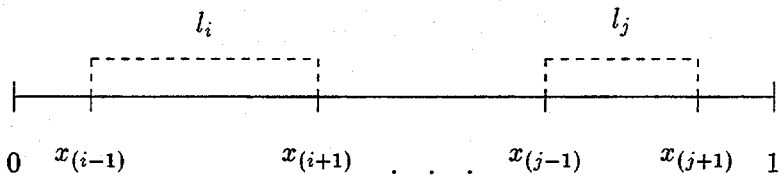
$$\begin{aligned}
Var(d_i) &= \int_0^{\frac{1}{2}} d_i^2 \, f(d_i) \, d \, d_i - (\frac{1}{2(n+1)})^2 \\
&= \int_0^{\frac{1}{2}} 2 \, n \, d_i^2 \, (1 - 2 \, di)^{(n-1)} d \, d_i - \frac{1}{4(n+1)^2} \\
&= 2 \, n \int_0^{\frac{1}{2}} d_i^2 \, (1 - 2 \, di)^{(n-1)} d \, d_i - \frac{1}{4(n+1)^2} \\
&= \frac{1}{2(n+1)(n+2)} - \frac{1}{4(n+1)^2} \\
&= \frac{n}{4(n+1)^2(n+2)}
\end{aligned}$$

Covariance of d_i and d_j

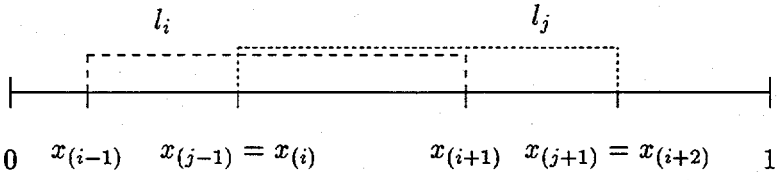
$$\begin{aligned} Cov(d_i, d_j) &= E(d_i d_j) - E(d_i)\, E(d_j) \\ &= E(d_i d_j) - (\frac{1}{2\,(n+1)})^2 \end{aligned}$$

There are two possible cases for which $E(d_i d_j)$ gets different values, and therefore $Cov(d_i, d_j)$ gets different values.

Case 1: $|(i) - (j)| \geq 2$, i.e. l_i and l_j do not overlap



Case 2: $|(i) - (j)| = 1$, i.e. l_i and l_j overlap.



In both cases, the calculation of the product moment $E(d_i d_j)$ involves the joint density function of four order statistics $x_{(i-1)}, x_{(i+1)}, x_{(j-1)}$ and $x_{(j+1)}$. As before, this density can be obtained by considering the joint density function of all n order statistics and then integrating out the other variables, $(x_{(1)}, \ldots, x_{(i-2)})$, $x_{(i)}$, $(x_{(i+2)}, \ldots, x_{(j-2)})$, $x_{(j)}$, and $(x_{(j+2)}, \ldots, x_{(n)})$

$$f_{(i-1),(i+1),(j-1),(j+1)}(x_{(i-1)}, x_{(i+1)}, x_{(j-1)}, x_{(j+1)})$$

$$= n! f(x_{(i-1)}) f(x_{(i+1)}) f(x_{(j-1)}) f(x_{(j+1)})$$

$$\cdot \int_0^{x_{(i-1)}} \cdots \int_0^{x_{(2)}} f(x_{(1)}) \cdots f(x_{(i-2)}) dx_{(1)} \cdots dx_{(i-2)}$$

$$\cdot \int_{x_{(i-1)}}^{x_{(i+1)}} f(x_{(i)}) dx_{(i)}$$

$$\cdot \int_{x_{(i+1)}}^{x_{(j-1)}} \cdots \int_{x_{(i+1)}}^{x_{(i+3)}} f(x_{(i+2)}) \cdots f(x_{(j-2)}) dx_{(i+2)} \cdots dx_{(j-2)}$$

$$\cdot \int_{x_{(j-1)}}^{x_{(j+1)}} f(x_{(j)}) dx_{(j)}$$

$$\cdot \int_{x_{(j+1)}}^{1} \cdots \int_{x_{(n-1)}}^{1} f(x_{(j+2)}) \cdots f(x_{(n)}) dx_{(n)} \cdots dx_{(j+2)}$$

$$= n!$$

$$\cdot \frac{x_{(i-1)}^{((i-1)-1)}}{((i-1)-1)!}$$

$$\cdot \frac{(x_{(i+1)} - x_{(i-1)})^{((i+1)-(i-1)-1)}}{((i+1)-(i-1)-1)!}$$

$$\cdot \frac{(x_{(j-1)} - x_{(i+1)})^{((j-1)-(i+1)-1)}}{((j-1)-(i+1)-1)!}$$

$$\cdot \frac{(x_{(j+1)} - x_{(j-1)})^{((j+1)-(j-1)-1)}}{((j+1)-(j-1)-1)!}$$

$$\cdot \frac{(1 - x_{(j+1)})^{(n-(j+1))}}{(n-(j+1))!}$$

$$= \frac{n!}{(i-2)!(j-i-3)!(n-j-1)!}$$

$$x_{(i-1)}^{(i-2)}(x_{(i+1)} - x_{(i-1)})(x_{(j-1)} - x_{(i+1)})^{(j-i-3)}(x_{(j+1)} - x_{(j-1)})(1 - x_{(j+1)})^{(n-j-1)}$$

Case 1: $|(i) - (j)| \geq 2$

Since l_i and l_j do not overlap, $(d_i \mid l_i = l_1)$ and $(d_j \mid l_j = l_2)$ are independent.
This implies that

$$\begin{aligned} E(d_i d_j | l_i = l_1, l_j = l_2) &= E(d_i \mid l_i = l_1) E(d_j \mid l_j = l_2) \\ &= \frac{l_1}{4}\frac{l_2}{4} \\ &= \frac{1}{16}(x_{(i+1)} - x_{(i-1)})(x_{(j+1)} - x_{(j-1)}) \end{aligned}$$

The conditional expectation can then be used to calculate the product moment.

$$E\,[\,E(d_i d_j \mid l_i = l_1,\ l_j = l_2)\,]$$

$$= E\,[\,\frac{1}{16}(x_{(i+1)} - x_{(i-1)})(x_{(j+1)} - x_{(j-1)})\,]$$

$$= \frac{1}{16}\int_0^1 \int_0^{x_{(j+1)}} \int_0^{x_{(j-1)}} \int_0^{x_{(i+1)}} (x_{(i+1)} - x_{(i-1)})\,(x_{(j+1)} - x_{(j-1)})$$

$$\cdot f_{(i-1),(i+1),(j-1),(j+1)}(x_{(i-1)}, x_{(i+1)}, x_{(j-1)}, x_{(j+1)}) dx_{(i-1)} dx_{(i+1)} dx_{(j-1)} dx_{(j+1)}$$

For simplicity, let

$x_{(i-1)}=a$	$(i-1)=(i)$
$x_{(i+1)}=b$	$(i+1)=(j)$
$x_{(j-1)}=c$	$(j-1)=(k)$
$x_{(j+1)}=d$	$(j+1)=(l)$

then solving the last equation gives

$$\frac{1}{16}\int_0^1 \int_0^d \int_0^c \int_0^b (b-a)\,(d-c)\, f_{(i),(j),(k),(l)}(a,b,c,d)\, da\, db\, dc\, dd$$

$$= \frac{1}{16} \frac{n!}{(i-1)!\,(j-i-1)!\,(k-j-1)!\,(l-k-1)!\,(n-l)!}$$

$$\cdot \int_0^1 \int_0^d \int_0^c \int_0^b a^{(i-1)}(b-a)^{(j-i)}(c-b)^{(k-j-1)}(d-c)^{(l-k)}(1-d)^{(n-l)} da\,db\,dc\,dd$$

$$= \frac{1}{16} \frac{n!}{(i-1)!\,(j-i-1)!\,(k-j-1)!\,(l-k-1)!\,(n-l)!} \frac{(i-1)!\,(j-i)!}{j!}$$

$$\cdot \int_0^1 \int_0^d \int_0^c b^j(c-b)^{(k-j-1)}(d-c)^{(l-k)}(1-d)^{(n-l)} db\,dc\,dd$$

$$= \frac{1}{16} \frac{n!}{(i-1)!\,(j-i-1)!\,(k-j-1)!\,(l-k-1)!\,(n-l)!} \frac{(i-1)!\,(j-i)!}{j!} \frac{j!\,(k-j-1)!}{k!}$$

$$\cdot \int_0^1 \int_0^d c^k(d-c)^{(l-k)}(1-d)^{(n-l)} dc\,dd$$

$$= \frac{1}{16} \frac{n!}{(i-1)!\,(j-i-1)!\,(k-j-1)!\,(l-k-1)!\,(n-l)!} \frac{(i-1)!\,(j-i)!}{j!} \frac{j!\,(k-j-1)!}{k!} \frac{k!\,(l-k)!}{(l+1)!}$$

$$\cdot \int_0^1 d^{(l+1)}(1-d)^{(n-l)} dd$$

$$= \frac{1}{16} \frac{n!\,(i-1)!\,(j-i)!\,j!\,(k-j-1)!\,k!\,(l-k)!\,(l+1)!\,(n-l)!}{(i-1)!\,(j-i-1)!\,(k-j-1)!\,(l-k-1)!\,(n-l)!\,j!\,k!\,(l+1)!\,(n+2)!}$$

$$= \frac{1}{16} \frac{(j-i)!\,(l-k)!}{(n+1)(n+2)}$$

Switching back to $(i-1)$, $(i+1)$, $(j-1)$, and $(j+1)$ gives

$$E\,[\,E(d_i d_j \mid l_i = l_1,\ l_j = l_2)\,] = \frac{1}{4\,(n+1)\,(n+2)}$$

Substituting into the covariance formula yields

$$\begin{aligned} Cov(d_i, d_j) &= E(d_i d_j) - E(d_i)\,E(d_j) \\ &= E\,[\,E(d_i d_j \mid l_i = l_1,\ l_j = l_2)\,] - \left(\frac{1}{2\,(n+1)}\right)^2 \\ &= \frac{1}{4\,(n+1)\,(n+2)} - \frac{1}{4\,(n+1)^2} \\ &= -\frac{1}{4\,(n+1)^2\,(n+2)} \end{aligned}$$

Case 2: $|(i)-(j)|=1$

In this case the four order statistics are actually $x_{(i-1)}$, $x_{(i)}$, $x_{(i+1)}$, and $x_{(i+2)}$, where $x_{(i)}$ was previously $x_{(j-1)}$ and $x_{(i+2)}$ was previously $x_{(j+1)}$. Since l_i and l_j overlap, $(d_i \mid l_i = l_1)$ and $(d_j \mid l_j = l_2)$ are not independent.

The joint density of the four order statistics is

$$f_{(i-1),(i),(i+1),(i+2)}(x_{(i-1)},x_{(i)},x_{(i+1)},x_{(i+2)}) = \frac{n!}{(i-2)!\,(n-i-2)!}\,x_{(i-1)}^{(i-2)}(1-x_{(i+2)})^{(n-i-2)}$$

Furthermore, $d_i \cdot d_j$ can get four different values as follows

$$d_i{\cdot}d_j = \begin{cases} (x_{(i)}-x_{(i-1)})(x_{(i+1)}-x_{(i)}) & \text{if } (2x_{(i)}-x_{(i+1)}) < x_{(i-1)} \text{ and } (2x_{(i+1)}-x_{(i+2)}) < x_{(i)} \\ (x_{(i)}-x_{(i-1)})(x_{(i+2)}-x_{(i+1)}) & \text{if } (2x_{(i)}-x_{(i+1)}) < x_{(i-1)} \text{ and } x_{(i)} < (2x_{(i+1)}-x_{(i+2)}) \\ (x_{(i+1)}-x_{(i)})(x_{(i+1)}-x_{(i)}) & \text{if } x_{(i-1)} < (2x_{(i)}-x_{(i+1)}) \text{ and } (2x_{(i+1)}-x_{(i+2)}) < x_{(i)} \\ (x_{(i+1)}-x_{(i)})(x_{(i+2)}-x_{(i+1)}) & \text{if } x_{(i-1)} < (2x_{(i)}-x_{(i+1)}) \text{ and } x_{(i)} < (2x_{(i+1)}-x_{(i+2)}) \end{cases}$$

Similarly to the calculation for Case 1, let

$x_{(i-1)}=a$

$x_{(i)} \;=b$

$x_{(i+1)}=c$

$x_{(i+2)}=d$

This implies that

$$d_i \cdot d_j = \begin{cases} (b-a)(c-b) & \text{if } (2b-c) < a \text{ and } (2c-d) < b \\ (b-a)(d-c) & \text{if } (2b-c) < a \text{ and } b < (2c-d) \\ (c-b)(c-b) & \text{if } a < (2b-c) \text{ and } (2c-d) < b \\ (c-b)(d-c) & \text{if } a < (2b-c) \text{ and } b < (2c-d) \end{cases}$$

The product moment of d_i and d_j is then

$$E(d_i d_j) = \frac{n!}{(i-2)!\,(n-i-2)!}\int_0^1\int_0^d\int_0^c\int_0^b d_i\,d_j\,a^{(i-2)}(1-d)^{(n-i-2)}\,da\,db\,dc\,dd$$

$$= \frac{n!}{(i-2)!\,(n-i-2)!}\, I((2c-d)<b) \int_0^1 \int_0^d \int_0^c \int_{max(0,(2b-c))}^b$$

$$a^{(i-2)}\,(b-a)\,(c-b)\,(1-d)^{(n-i-2)}\,da\,db\,dc\,dd$$

$$+ \frac{n!}{(i-2)!\,(n-i-2)!}\, I(b<(2c-d)) \int_0^1 \int_0^d \int_0^c \int_{max(0,(2b-c))}^b$$

$$a^{(i-2)}\,(b-a)\,(d-c)\,(1-d)^{(n-i-2)}\,da\,db\,dc\,dd$$

$$+ \frac{n!}{(i-2)!\,(n-i-2)!}\, I((2c-d)<b)\, I(0<(2b-c)) \int_0^1 \int_0^d \int_0^c \int_0^{(2b-c)}$$

$$a^{(i-2)}\,(c-b)\,(c-b)\,(1-d)^{(n-i-2)}\,da\,db\,dc\,dd$$

$$+ \frac{n!}{(i-2)!\,(n-i-2)!}\, I(b<(2c-d))\, I(0<(2b-c)) \int_0^1 \int_0^d \int_0^c \int_0^{(2b-c)}$$

$$a^{(i-2)}\,(c-b)\,(d-c)\,(1-d)^{(n-i-2)}\,da\,db\,dc\,dd$$

$$= \frac{n!}{(i-2)!\,(n-i-2)!}\, \frac{1}{(i)(i-1)} \int_0^1 \int_0^d \int_{max((2c-d),\frac{c}{2})}^c$$

$$\{\, b^i - [\,(2b-c)^{(i-1)}(2b-c-bi+ci)\,]\,\}\,(c-b)(1-d)^{(n-i-2)}\,db\,dc\,dd$$

$$+ \frac{n!}{(i-2)!\,(n-i-2)!}\, \frac{1}{(i)(i-1)}\, I((2c-d)<\frac{c}{2}) \int_0^1 \int_0^d \int_{max(0,(2c-d))}^{\frac{c}{2}}$$

$$(c-b)\,b^i\,(1-d)^{(n-i-2)}\,db\,dc\,dd$$

$$+ \frac{n!}{(i-2)!\,(n-i-2)!}\, \frac{1}{(i)(i-1)}\, I(\frac{c}{2}<(2c-d)) \int_0^1 \int_0^d \int_{\frac{c}{2}}^{(2c-d)}$$

$$\{b^i - [\,(2b-c)^{(i-1)}(2b-c-bi+ci)\,]\}(d-c)\,(1-d)^{(n-i-2)}\,db\,dc\,dd$$

$$+ \frac{n!}{(i-2)!\,(n-i-2)!}\,\frac{1}{(i)(i-1)}\int_0^1\int_0^d\int_0^{min((2c-d),\frac{c}{2})}$$

$$b^i\,(d-c)\,(1-d)^{(n-i-2)}\,db\,dc\,dd$$

$$+ \frac{n!}{(i-2)!\,(n-i-2)!}\,\frac{i}{(i)(i-1)}\int_0^1\int_0^d\int_{max((2c-d),\frac{c}{2})}^{c}$$

$$(c-b)^2\,(2b-c)^{(i-1)}\,(1-d)^{(n-i-2)}\,db\,dc\,dd$$

$$+ \frac{n!}{(i-2)!\,(n-i-2)!}\,\frac{i}{(i)(i-1)}\,I(\frac{c}{2}<(2c-d))\int_0^1\int_0^d\int_{\frac{c}{2}}^{(2c-d)}$$

$$(c-b)\,(2b-c)^{(i-1)}\,(d-c)\,(1-d)^{(n-i-2)}\,db\,dc\,dd$$

$$=$$

$$\vdots$$

$$= \frac{1}{3\,(n+1)\,(n+2)}$$

Substituting into the covariance formula yields

$$\begin{aligned} Cov(d_i,d_j) &= E(d_id_j) - E(d_i)\,E(d_j) \\ &= \frac{1}{3\,(n+1)\,(n+2)} - \frac{1}{4\,(n+1)^2} \\ &= -\frac{n-2}{12\,(n+1)^2\,(n+2)} \end{aligned}$$

Thus, for both cases, 1 and 2,

$$Cov(d_i,d_j) = \begin{cases} -\frac{1}{4\,(n+1)^2\,(n+2)} & \text{if } |(i)-(j)| \geq 2 \\ \frac{n-2}{12\,(n+1)^2\,(n+2)} & \text{if } |(i)-(j)| = 1 \end{cases}$$

1.4 Moments of Mean Nearest Neighbor Distance

Let $x_{(1)} \leq \ldots \leq x_{(n)}$ denote the order statistics of n independent identically distributed random points from a $\mathcal{U}(0,1)$ distribution or from an homogeneous Poisson process, conditioning on n. The mean nearest neighbor distance can be calculated as follows:

$$\bar{d} = \frac{1}{n}\{(x_{(2)} - x_{(1)}) + \sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})] + (x_{(n)} - x_{(n-1)})\}$$

Expected Value of $\bar{d}$

$$\begin{aligned}
E(\bar{d}) &= \frac{1}{n} E\{(x_{(2)} - x_{(1)}) + \sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})] + (x_{(n)} - x_{(n-1)})\} \\
&= \frac{1}{n}\{E(x_{(2)} - x_{(1)}) + E(\sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})]) + E(x_{(n)} - x_{(n-1)})\} \\
&= \frac{1}{n}\{\frac{(2-1)}{(n+1)} + \frac{n-2}{2(n+1)} + \frac{(n-(n-1))}{(n+1)}\} \\
&= \frac{1}{n}\{\frac{1}{(n+1)} + \frac{n-2}{2(n+1)} + \frac{1}{(n+1)}\} \\
&= \frac{n+2}{2n(n+1)}
\end{aligned}$$

Variance of $\bar{d}$

Let $d_1 = x_{(2)} - x_{(1)}$ and $d_n = x_{(n)} - x_{(n-1)}$

$$\begin{aligned}
Var(\bar{d}) &= \frac{1}{n^2} Var\{d_1 + \sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})] + d_n\} \\
&= \frac{1}{n^2}\{Var(d_1) + \sum_{i=2}^{n-1} Var(d_i) + Var(d_n) \\
&\quad + 2\sum_{i=2}^{n-1} Cov(d_1, d_i) + 2\,Cov(d_1, d_n) \\
&\quad + 2\sum_{i=2, i<j}^{n-1} Cov(d_i, d_j) \\
&\quad + 2\sum_{i=2}^{n-1} Cov(d_i, d_n)\}
\end{aligned}$$

$\underline{Var(d_1)}$

$$\begin{aligned} Var(d_1) &= Var(x_{(2)} - x_{(1)}) \\ &= Var(x_{(1)}) + Var(x_{(2)}) - 2\,Cov(x_{(1)}, x_{(2)}) \\ &= \frac{n}{(n+1)^2\,(n+2)} + \frac{2(n-1)}{(n+1)^2\,(n+2)} - \frac{2(n-1)}{(n+1)^2\,(n+2)} \\ &= \frac{n}{(n+1)^2\,(n+2)} \end{aligned}$$

$\underline{Var(d_n)}$

$$\begin{aligned} Var(d_n) &= Var(x_{(n)} - x_{(n-1)}) \\ &= Var(x_{(n-1)}) + Var(x_{(n)}) - 2\,Cov(x_{(n-1)}, x_{(n)}) \\ &= \frac{2(n-1)}{(n+1)^2\,(n+2)} + \frac{n}{(n+1)^2\,(n+2)} - \frac{2(n-1)}{(n+1)^2\,(n+2)} \\ &= \frac{n}{(n+1)^2\,(n+2)} \end{aligned}$$

$\underline{Cov(d_1, d_i)}$

The covariance of d_1 and d_i has two different values, for $i = 2$ and for $3 \leq i \leq n-1$.

For $\mathbf{i = 2}$

$$\begin{aligned} Cov(d_1, d_2) &= E(d_1 d_2) - E(d_1)E(d_2) \\ &= E(d_1 d_2) - E(x_{(2)} - x_{(1)})\; E(\; min[\,(x_{(2)} - x_{(1)}), (x_{(3)} - x_{(2)})\,]\,) \\ &= E(d_1 d_2) - \frac{1}{2(n+1)^2} \end{aligned}$$

The joint density function of $x_{(1)}, x_{(2)}$ and $x_{(3)}$ is

$$f_{(1),(2),(3)}(x_{(1)}, x_{(2)}, x_{(3)}) = n(n-1)(n-2)(1 - x_{(3)})^{(n-3)}$$

Here

$$d_1 \cdot d_2 = \begin{cases} (x_{(2)} - x_{(1)})(x_{(2)} - x_{(1)}) & \text{if } (2x_{(2)} - x_{(3)}) < x_{(1)} \\ (x_{(2)} - x_{(1)})(x_{(3)} - x_{(2)}) & \text{if } x_{(1)} < (2x_{(2)} - x_{(3)}) \end{cases}$$

As before, let

$x_{(1)} = a$

$x_{(2)} = b$

$x_{(3)} = c$

This implies that

$$d_1 \cdot d_2 = \begin{cases} (b-a)(b-a) & \text{if } (2b-c) < a \\ (b-a)(c-b) & \text{if } a < (2b-c) \end{cases}$$

The product moment of d_1 and d_2 is then

$$\begin{aligned}
E(d_1 d_2) &= n(n-1)(n-2) \int_0^1 \int_0^c \int_0^b d_1\, d_2\, (1-c)^{(n-3)}\, da\, db\, dc \\
&= n(n-1)(n-2) \int_0^1 \int_0^c \int_{max(0,(2b-c))}^b (b-a)^2\, (1-c)^{(n-3)}\, da\, db\, dc \\
&\quad + n(n-1)(n-2) I(0 < (2b-c)) \int_0^1 \int_0^c \int_0^{(2b-c)} (b-a)\,(c-b)\,(1-c)^{(n-3)}\, da\, db\, dc \\
&= n(n-1)(n-2) \int_0^1 \int_{\frac{c}{2}}^c \frac{b^3}{3} (1-c)^{(n-3)}\, db\, dc \\
&\quad - n(n-1)(n-2) \int_0^1 \int_{\frac{c}{2}}^c b^2\,(2b-c)\,(1-c)^{(n-3)}\, db\, dc \\
&\quad + n(n-1)(n-2) \int_0^1 \int_{\frac{c}{2}}^c b\,(2b-c)^2\,(1-c)^{(n-3)}\, db\, dc \\
&\quad - n(n-1)(n-2) \int_0^1 \int_{\frac{c}{2}}^c \frac{(2b-c)^3}{3} (1-c)^{(n-3)}\, db\, dc \\
&\quad + n(n-1)(n-2) \int_0^1 \int_0^{\frac{c}{2}} \frac{b^3}{3} (1-c)^{(n-3)}\, db\, dc \\
&\quad + n(n-1)(n-2) \int_0^1 \int_{\frac{c}{2}}^c b\,(c-b)\,(2b-c)\,(1-c)^{(n-3)}\, db\, dc \\
&\quad - n(n-1)(n-2) \int_0^1 \int_{\frac{c}{2}}^c (c-b) \frac{(2b-c)^2}{2} (1-c)^{(n-3)}\, db\, dc \\
&= \\
&\vdots \\
&= \frac{24\, n\, (n-1)\, (n-2)}{32\, (n-2)\, (n-1)\, n\, (n+1)\, (n+2)} \\
&= \frac{3}{4\, (n+1)\, (n+2)}
\end{aligned}$$

Substituting into the covariance formula yields

$$\begin{aligned} Cov(d_1, d_2) &= E(d_1 d_2) - E(d_1)E(d_2) \\ &= \frac{3}{4\,(n+1)\,(n+2)} - \frac{1}{2\,(n+1)^2} \\ &= \frac{n-1}{4\,(n+1)^2\,(n+2)} \end{aligned}$$

For $\mathbf{3 \leq i \leq n-1}$

$$\begin{aligned} Cov(d_1, d_i) &= E(d_1 d_i) - E(d_1)E(d_i) \\ &= E(d_1 d_i) - E(x_{(2)} - x_{(1)})\, E(\, min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,]\,) \\ &= E(d_1 d_i) - \frac{1}{2(n+1)^2} \end{aligned}$$

The joint density function of $x_{(1)}, x_{(2)}, x_{(i-1)}, x_{(i)}$ and $x_{(i+1)}$ is

$$\begin{aligned} & f_{(1),(2),(i-1),(i),(i+1)}(x_{(1)}, x_{(2)}, x_{(i-1)}, x_{(i)}, x_{(i+1)}) \\ &= \frac{n!}{(i-4)!\,(n-i-1)!}\,(x_{(i-1)} - x_{(2)})^{(i-4)}(1 - x_{(i+1)})^{(n-i-1)} \end{aligned}$$

Here

$$d_1 \cdot d_i = \begin{cases} (x_{(2)} - x_{(1)})(x_{(i)} - x_{(i-1)}) & \text{if } (2x_{(i)} - x_{(i+1)}) < x_{(i-1)} \\ (x_{(2)} - x_{(1)})(x_{(i+1)} - x_{(i)}) & \text{if } x_{(i-1)} < (2x_{(i)} - x_{(i+1)}) \end{cases}$$

As before, let

$x_{(1)} = a$
$x_{(2)} = b$
$x_{(i-1)} = c$
$x_{(i)} = d$
$x_{(i+1)} = e$

This implies that

$$d_1 \cdot d_i = \begin{cases} (b-a)(d-c) & \text{if } (2d-e) < c \\ (b-a)(e-d) & \text{if } c < (2d-e) \end{cases}$$

The product moment of d_1 and d_i is then

$$\begin{aligned} & E(d_1 d_i) \\ &= \frac{n!}{(i-4)!\,(n-i-1)!} \int_0^1 \int_0^e \int_0^d \int_0^c \int_0^b d_i\, d_j\, (c-b)^{(i-4)}(1-e)^{(n-i-1)}\, da\, db\, dc\, dd\, de \end{aligned}$$

$$= \frac{n!}{(i-4)!\,(n-i-1)!} I((2d-e)<c) \int_0^1 \int_0^e \int_0^d \int_0^c \int_0^b$$

$$(b-a)(c-b)^{(i-4)}(d-c)(1-e)^{(n-i-1)}\,dadbdcddde$$

$$+ \frac{n!}{(i-4)!\,(n-i-1)!} I(c<(2d-e)) \int_0^1 \int_0^e \int_0^d \int_0^c \int_0^b$$

$$(b-a)(c-b)^{(i-4)}(e-d)(1-e)^{(n-i-1)}\,dadbdcddde$$

$$= \frac{n!}{(i-4)!\,(n-i-1)!} I((2d-e)<c) \int_0^1 \int_0^e \int_0^d \int_0^c$$

$$\frac{b^2}{2}(c-b)^{(i-4)}(d-c)(1-e)^{(n-i-1)}\,dbdcddde$$

$$+ \frac{n!}{(i-4)!\,(n-i-1)!} I(c<(2d-e)) \int_0^1 \int_0^e \int_0^d \int_0^c$$

$$\frac{b^2}{2}(c-b)^{(i-4)}(e-d)(1-e)^{(n-i-1)}\,dbdcddde$$

$$=$$

$$\vdots$$

$$= \frac{(i+2)!\,(n-i-1)!\,n!}{2\,(i+2)!\,(n-i-1)!\,(n+2)!}$$

$$= \frac{1}{2\,(n+1)\,(n+2)}$$

Substituting into the covariance formula yields

$$\begin{aligned} Cov(d_1, d_i) &= E(d_1 d_i) - E(d_1)E(d_2) \\ &= \frac{1}{2\,(n+1)\,(n+2)} - \frac{1}{2\,(n+1)^2} \\ &= -\frac{1}{2\,(n+1)^2\,(n+2)} \end{aligned}$$

$\underline{Cov(d_1, d_n)}$

Following similar procedure gives the product moment $E(d_1 d_n)$.

Substituting into the covariance formula yields

$$\begin{aligned} Cov(d_1, d_n) &= E(d_1 d_n) - E(d_1)E(d_n) \\ &= \frac{1}{(n+1)(n+2)} - \frac{1}{(n+1)^2} \\ &= -\frac{1}{(n+1)^2 (n+2)} \end{aligned}$$

$\underline{Cov(d_i, d_n)}$

Similar to the covariance of d_1 and d_i, the covariance of d_i and d_n has two different values, for $i = n-1$ and for $2 \leq i \leq n-2$.

For $\mathbf{i = n - 1}$

$$\begin{aligned} Cov(d_{n-1}, d_n) &= E(d_{n-1}d_n) - E(d_{n-1})E(d_n) \\ &= E(d_{n-1}d_n) - E(\, min[\,(x_{(n-1)} - x_{(n-2)}), (x_{(n)} - x_{(n-1)})\,]\,)\, E(x_{(n)} - x_{(n-1)}) \\ &= \frac{3}{4(n+1)(n+2)} - \frac{1}{2(n+1)^2} \\ &= \frac{n-1}{4(n+1)^2(n+2)} \end{aligned}$$

For $\mathbf{2 \leq i \leq n - 2}$

$$\begin{aligned} Cov(d_i, d_n) &= E(d_i d_n) - E(d_i)E(d_n) \\ &= E(\, min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,]\,)\, E(x_{(n)} - x_{(n-1)}) \\ &= \frac{1}{2(n+1)(n+2)} - \frac{1}{2(n+1)^2} \\ &= -\frac{1}{2(n+1)^2(n+2)} \end{aligned}$$

Finally, substituting all the above results into the formula of the $Var(\bar{d})$ gives

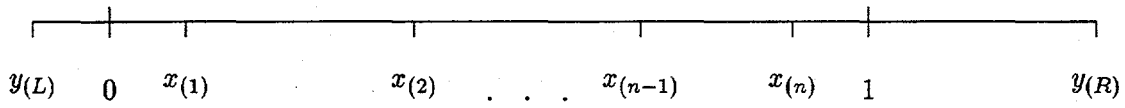
$$
\begin{aligned}
Var(\bar{d}) &= \frac{1}{n^2} Var\left\{ d_1 + \sum_{i=2}^{n-1} min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,] + d_n \right\} \\
&= \frac{1}{n^2} \left\{ \frac{n}{(n+1)^2\,(n+2)} + \frac{n\,(n-2)}{4\,(n+1)^2\,(n+2)} + \frac{n}{(n+1)^2\,(n+2)} \right. \\
&\quad + \frac{2\,(n-1)}{4\,(n+1)^2\,(n+2)} - \frac{2\,(n-3)}{2\,(n+1)^2\,(n+2)} - \frac{2}{(n+1)^2\,(n+2)} \\
&\quad + \frac{2\,(n-2)\,(n-3)}{12\,(n+1)^2\,(n+2)} - \frac{2\,\frac{(n-3)(n-4)}{2}}{4\,(n+1)^2\,(n+2)} \\
&\quad \left. + \frac{2\,(n-1)}{4\,(n+1)^2\,(n+2)} - \frac{2\,(n-3)}{2\,(n+1)^2\,(n+2)} \right\} \\
&= \frac{1}{n^2} \frac{2\,n^2 + 17\,n + 12}{12\,(n+1)^2\,(n+2)}
\end{aligned}
$$

1.5 Edge Effects Correction

In Section 1.4 the calculation of the mean nearest neighbor distance was based on the assumption that the contributions from the two extreme points are the distances to their neighbors. This can be readily seen in the formula for $\bar{d}$,

$$\bar{d} = \frac{1}{n}\{\underbrace{(x_{(2)} - x_{(1)})}_{d_1} + \sum_{i=2}^{n-1} min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,] + \underbrace{(x_{(n)} - x_{(n-1)})}_{d_n}\}$$

Unfortunately, these contributions may be biased due to the edge effects.
The edge effect, which is also known as the boundary problem, arises whenever at least one of the two extreme points is closer to the end–point of the line than to its neighbor. In this case, the distance from the extreme point in one direction is truncated and cannot be compared to the distance to its neighbor on the study interval. This implies that the nearest distance may be smaller than the one that was used in the calculation of the mean nearest neighbor distance. Consequently the contribution of this extreme point to the mean is larger than it should be. The figure below illustrates these two possible situations.



The smallest point, $x_{(1)}$, is closer to the interval border than to its neighbor $x_{(2)}$. The distance to the point on the other side of the border, $y_{(L)}$, is smaller than the distance to $x_{(2)}$. Thus the contribution of $x_{(1)}$ to the mean nearest neighbor distance is larger than it should be. On the other hand, although the largest point, $x_{(n)}$, is closer to the border than to its neighbor, $x_{(n-1)}$, there is no need for correction. As it is shown, the distance from this point to its neighbor is smaller than the distance to the point outside the interval.

Since there is no information on the location of the points outside the interval boundaries the correction for the edge effects relies either on the information about the location of the random points on the study interval, or on the assumption that the same processes responsible for the location of the points on the interval are operating beyond its boundaries.

There are five possible methods which can be used to correct for the edge effects: the 'Circle' method, the 'Boundary' method, the 'Mirror' method, the 'Expected Value' method, and the 'Random Points' method. The 'Boundary' and the 'Expected Value' methods are based only on the information from the given set of random points and their distribution along the interval. The other three correction methods are based on the assumption that the points pattern inside the interval is the same as the one outside its boundaries. These methods are described in the following sections.

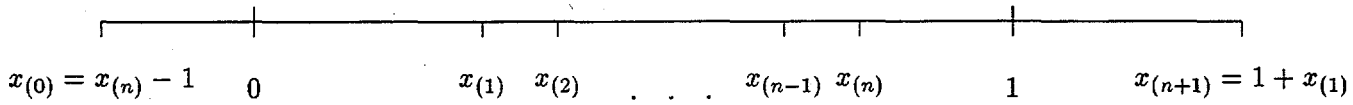
1.5.1 The 'Circle' Method

One way to compensate for the edge effects is based on the conversion of the straight line (or any other non closed curve) to a circle (or any other closed curve) by joining together its opposite edges. On the circle, the previously two exterme order statistics, $x_{(1)}$ and $x_{(n)}$, become candidates for being each other's nearest neighbor. The basic assumption is that the same processes responsible for the location of the points on the unit interval are operating beyond its boundaries. Therefore, these two points are assumed to be representatives of the other points outside the interval.

Another useful way to think about it is to consider $n+2$, instead of n, order statistics

$x_{(0)} < x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)} < x_{(n+1)}$

where $x_{(0)} = -(1 - x_{(n)})$ and $x_{(n+1)} = 1 + x_{(1)}$



The corrected expected value of $\bar{d}$ can be calculated following the same procedure which had been used in Sections 1.3 and 1.4

Let $l_i = (x_{(i)} - x_{(i-1)}) + (x_{(i+1)} - x_{(i)}) = x_{(i+1)} - x_{(i-1)}$ $\quad 0 \leq l_i \leq 1, \quad i = 1, \ldots, n$

Let $d_i = min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,]$ $\quad 0 \leq d_i \leq \frac{l_i}{2}, \quad i = 1, \ldots, n$

In particular,

$l_1 = x_{(2)} - x_{(0)} = 1 + x_{(2)} - x_{(n)}$

$l_n = x_{(n+1)} - x_{(n-1)} = 1 + x_{(1)} - x_{(n-1)}$

$d_1 = min[\,(x_{(1)} - x_{(0)}), (x_{(2)} - x_{(1)})\,] = min[\,(1 + x_{(1)} - x_{(n)}), (x_{(2)} - x_{(1)})\,]$

$d_n = min[\,(x_{(n)} - x_{(n-1)}), (x_{(n+1)} - x_{(n)})\,] = min[\,(x_{(n)} - x_{(n-1)}), (1 + x_{(1)} - x_{(n)})\,]$

Then,

$$\bar{d} = \frac{1}{n}\,\{\, d_1 + \sum_{i=2}^{n-1} min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,] + d_n \,\}$$

The expected values of d_1 and d_n can then be calculated using their conditional expectations, given l_1 and l_n.

$$\begin{aligned} E(d_1) = E[E(d_1 \mid l_1)] &= E(\frac{l_1}{4}) \\ &= \frac{1}{4} E(1 + x_{(2)} - x_{(n)}) \\ &= \frac{1}{4} \{ \frac{n+1}{n+1} + \frac{2}{n+1} - \frac{n}{n+1} \} \\ &= \frac{3}{4(n+1)} \end{aligned}$$

Similarly,

$$\begin{aligned} E(d_n) = E[E(d_n \mid l_n)] &= E(\frac{l_n}{4}) \\ &= \frac{1}{4} E(1 + x_{(1)} - x_{(n-1)}) \\ &= \frac{1}{4} \{ \frac{n+1}{n+1} + \frac{1}{n+1} - \frac{n-1}{n+1} \} \\ &= \frac{3}{4(n+1)} \end{aligned}$$

The corrected expected value of $\bar{d}$ is then

$$\begin{aligned} E(\bar{d}) &= \frac{1}{n} E \{ d_1 + \sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})] + d_n \} \\ &= \frac{1}{n} \{ E(d_1) + E(\sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})]) + E(d_n) \} \\ &= \frac{1}{n} \{ \frac{3}{4(n+1)} + \frac{n-2}{2(n+1)} + \frac{3}{4(n+1)} \} \\ &= \frac{1}{2\,n} \end{aligned}$$

Furthermore, it is possible to calculate the difference between the uncorrected and the corrected mean nearest neighbor distance.

$$\begin{aligned} DIFFERENCE &= \frac{n+2}{2n(n+1)} - \frac{1}{2n} \\ &= \frac{1}{2n(n+1)} \end{aligned}$$

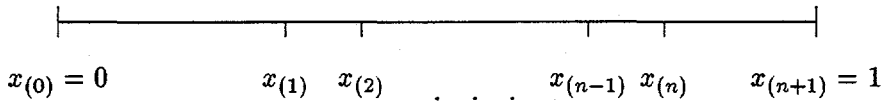
1.5.2 The 'Boundary' Method

An alternative way to overcome the boundary problem is to consider the distance from each extreme point to the nearest border as its possible nearest neighbor distance. Since there is no information on the point pattern outside the unit interval, this method relies only on the one available, that is, the location of the extreme points and there distance to the interval boundaries. The 'Boundary' method can be easily generalized to two dimensions, even when the shape of study area is irregular.

This method implies that there are $n+2$, instead of n, order statistics

$x_{(0)} \leq x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)} \leq x_{(n+1)}$

where $x_{(0)} = 0$ and $x_{(n+1)} = 1$



The corrected expected value of $\bar{d}$ can then be calculated as before

Let

$l_1 = x_{(2)} - x_{(0)} = x_{(2)}$

$l_n = x_{(n+1)} - x_{(n-1)} = 1 - x_{(n-1)}$

$d_1 = min[\,(x_{(1)} - x_{(0)}), (x_{(2)} - x_{(1)})\,] = min[\,(x_{(1)}), (x_{(2)} - x_{(1)})\,]$

$d_n = min[\,(x_{(n)} - x_{(n-1)}), (x_{(n+1)} - x_{(n)})\,] = min[\,(x_{(n)} - x_{(n-1)}), (1 - x_{(n)})\,]$

The expected values of d_1 and d_n are then

$$\begin{aligned} E(d_1) = E[E(d_1 \mid l_1)] &= E(\frac{l_1}{4}) \\ &= \frac{1}{4} E(x_{(2)}) \\ &= \frac{1}{4} \frac{2}{n+1} \\ &= \frac{1}{2(n+1)} \end{aligned}$$

Similarly,

$$\begin{aligned} E(d_n) = E[E(d_n \mid l_n)] &= E(\frac{l_n}{4}) \\ &= \frac{1}{4} E(1 - x_{(n-1)}) \\ &= \frac{1}{4} \{ \frac{n+1}{n+1} - \frac{n-1}{n+1} \} \\ &= \frac{1}{2(n+1)} \end{aligned}$$

The corrected expected value of $\bar{d}$ is then

$$\begin{aligned} E(\bar{d}) &= \frac{1}{n} E \{ d_1 + \sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})] + d_n \} \\ &= \frac{1}{n} \{ E(d_1) + E(\sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})]) + E(d_n) \} \\ &= \frac{1}{n} \{ \frac{1}{2(n+1)} + \frac{n-2}{2(n+1)} + \frac{1}{2(n+1)} \} \\ &= \frac{1}{2(n+1)} \end{aligned}$$

The difference between the uncorrected and the corrected mean nearest neighbor distance is then

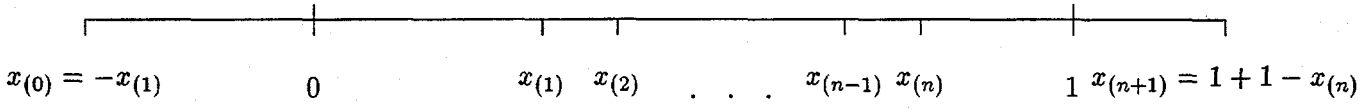
$$\begin{aligned} DIFFERENCE &= \frac{n+2}{2n(n+1)} - \frac{1}{2(n+1)} \\ &= \frac{1}{n(n+1)} \end{aligned}$$

1.5.3 The 'Mirror' Method

Another way to overcome the edge effects is to assume that on the other side of the border there is a point which is located at the same distance from it as the extreme point. Thus, a distance, which is exactly twice the distance to the border, becomes a possible nearest neighbor distance. As in the 'Circle' methods, the assumption is that the same processes responsible for the location of the points on the unit interval are operating beyond its boundaries. The 'Mirror' method, as the 'Boundary' method, can be easily generalized to a two–dimensions situation, even if the shape of study area is irregular, while the 'Circle' method can be generalized to two dimensions only if the study area has a rectangular shape. Another way to think about it is to consider again $n+2$ order statistics

$x_{(0)} < x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)} < x_{(n+1)}$

where $x_{(0)} = -x_{(1)}$ and $x_{(n+1)} = 1 + (1 - x_{(n)})$



Let

$l_1 = x_{(2)} - x_{(0)} = x_{(2)} + x_{(1)}$

$l_n = x_{(n+1)} - x_{(n-1)} = 2 - x_{(n)} - x_{(n-1)}$

$d_1 = min[\,(x_{(1)} - x_{(0)}), (x_{(2)} - x_{(1)})\,] = min[\,(2x_{(1)}), (x_{(2)} - x_{(1)})\,]$

$d_n = min[\,(x_{(n)} - x_{(n-1)}), (x_{(n+1)} - x_{(n)})\,] = min[\,(x_{(n)} - x_{(n-1)}), (2 - 2\,x_{(n)})\,]$

Note that given l_1, the conditional distribution of d_1 is uniform $(0,\ \frac{2}{3}\,x_{(2)})$

and given l_n, the conditional distribution of d_n is uniform $(0,\ \frac{2}{3}\,(1 - x_{(n-1)}))$

Then,

$$E(d_1 \mid l_1) \;=\; \int_0^{\frac{2x_{(2)}}{3}} d_1\, f(d_1 \mid l_1)\, d\, d_1 = \frac{1}{\frac{2\,x_{(2)}}{3}} \int_0^{\frac{2x_{(2)}}{3}} d_1\, d\, d_1 = \frac{x_{(2)}}{3}$$

and the expected value of d_1 is

$$E(d_1) = E[E(d_1 \mid l_1)] \;=\; E(\frac{x_{(2)}}{3}) = \frac{2}{3\,(n+1)}$$

Similarly,

$$E(d_n \mid l_n) = \int_0^{\frac{2(1-x_{(n-1)})}{3}} d_n \, f(d_n \mid l_n) \, d\, d_n = \frac{1}{\frac{2\,(1-x_{(n-1)})}{3}} \int_0^{\frac{2(1-x_{(n-1)})}{3}} d_n \, d\, d_n = \frac{1-x_{(n-1)}}{3}$$

and the expected value of d_n is

$$E(d_n) = E[E(d_n \mid l_n)] = E(\frac{1-x_{(n-1)}}{3}) = \frac{2}{3\,(n+1)}$$

The corrected expected value of $\bar{d}$ is

$$\begin{aligned} E(\bar{d}) &= \frac{1}{n} E\,\{\, d_1 + \sum_{i=2}^{n-1} min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,] + d_n \,\} \\ &= \frac{1}{n}\,\{\, E(d_1) + E(\sum_{i=2}^{n-1} min[\,(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})\,]) + E(d_n)\,\} \\ &= \frac{1}{n}\,\{\, \frac{2}{3(n+1)} + \frac{n-2}{2(n+1)} + \frac{2}{3(n+1)}\,\} \\ &= \frac{3n+2}{6n(n+1)} \end{aligned}$$

The difference between the uncorrected and the corrected mean nearest neighbor distance is then

$$\begin{aligned} DIFFERENCE &= \frac{n+2}{2n(n+1)} - \frac{3n+2}{6n(n+1)} \\ &= \frac{2}{3n(n+1)} \end{aligned}$$

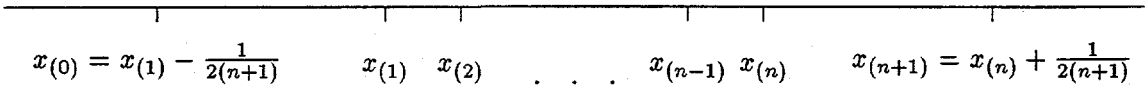
1.5.4 The 'Expected Value' Method

An alternative solution the boundary problem is to consider the expected value of the nearest neighbor distances for a unit interval as a possible distance of each of the extreme points. This method, like the 'Boundary' method, relies only on information about the distribution of the set of points on the interval. The major advantage of this method over the others is that the correction factor is a constant which depends only on the number of points and not on their location. Furthermore, it is based on information from all the points while the others are based only on the two extreme order statistics.

The 'Expected Value' method implies having $n+2$ order statistics

$x_{(0)} < x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)} < x_{(n+1)}$

where $x_{(0)} = x_{(1)} - \frac{1}{2(n+1)}$ and $x_{(n+1)} = x_{(n)} + \frac{1}{2(n+1)}$



Let

$l_1 = x_{(2)} - x_{(0)} = x_{(2)} - x_{(1)} + \frac{1}{2(n+1)}$

$l_n = x_{(n+1)} - x_{(n-1)} = x_{(n)} - x_{(n-1)} + \frac{1}{2(n+1)}$

$d_1 = min[\,(x_{(1)} - x_{(0)}), (x_{(2)} - x_{(1)})\,] = min[\,(\frac{1}{2(n+1)}), (x_{(2)} - x_{(1)})\,]$

$d_n = min[\,(x_{(n)} - x_{(n-1)}), (x_{(n+1)} - x_{(n)})\,] = min[\,(x_{(n)} - x_{(n-1)}), (\frac{1}{2(n+1)})\,]$

The expected values of d_1 and d_n are then

$$\begin{aligned} E(d_1) = E[E(d_1 \mid l_1)] &= E(\frac{l_1}{4}) \\ &= \frac{1}{4}\, E(x_{(2)} - x_{(1)} + \frac{1}{2(n+1)}) \\ &= \frac{1}{4}\,\{\frac{2}{n+1} - \frac{1}{n+1} + \frac{1}{2(n+1)}\} \\ &= \frac{3}{8(n+1)} \end{aligned}$$

Similarly,

$$\begin{aligned} E(d_n) = E[E(d_n \mid l_n)] &= E(\frac{l_n}{4}) \\ &= \frac{1}{4} E(x_{(n)} - x_{(n-1)} + \frac{1}{2(n+1)}) \\ &= \frac{1}{4} \{ \frac{n}{n+1} - \frac{n-1}{n+1} + \frac{1}{2(n+1)} \} \\ &= \frac{3}{8(n+1)} \end{aligned}$$

The corrected expected value of $\bar{d}$ is then

$$\begin{aligned} E(\bar{d}) &= \frac{1}{n} E \{ d_1 + \sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})] + d_n \} \\ &= \frac{1}{n} \{ E(d_1) + E(\sum_{i=2}^{n-1} min[(x_{(i)} - x_{(i-1)}), (x_{(i+1)} - x_{(i)})]) + E(d_n) \} \\ &= \frac{1}{n} \{ \frac{3}{8(n+1)} + \frac{n-2}{2(n+1)} + \frac{3}{8(n+1)} \} \\ &= \frac{2n-1}{4n(n+1)} \end{aligned}$$

The difference between the uncorrected and the corrected mean nearest neighbor distance is then

$$\begin{aligned} DIFFERENCE &= \frac{n+2}{2n(n+1)} - \frac{2n-1}{4n(n+1)} \\ &= \frac{5}{4n(n+1)} \end{aligned}$$

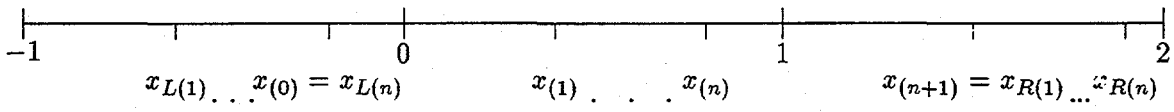
1.5.5 The 'Random Points' Method

An alternative strategy for overcoming the boundary problem is to distribute random points on both sides of the study interval with the same intensity as the one on the interval. This can be done by considering two more unit length intervals, which are located on its both sides, and n independent uniformly distributed random points on each of them. After ordering the points, the largest one on the left hand side interval becomes a candidate for being $x_{(1)}$'s nearest neighbor, and the smallest one on the right hand side intervals becomes a candidate for being $x_{(n)}$'s nearest neighbor.

Another useful way to think about it is to consider $n+2$ order statistics

$x_{(0)} < x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)} < x_{(n+1)}$

where $x_{(0)}$ is the largest order statistic on the left hand side interval and $x_{(n+1)}$ is the smallest order statistics on the right hand side interval.



This implies that the expected location of $x_{(0)}$ with respect to 0 is the same as the expected location of $x_{(n)}$ with respect to 1. Furthermore, the expected location of $x_{(n+1)}$ with respect to 1 is the same as the expected location of $x_{(1)}$ with respect to 0. Thus, $x_{(0)}$ can be considered as $-(1-x_{(n)})$ and $x_{(n+1)}$ can be considered as $1+x_{(1)}$. That is, the expected location of $x_{(0)}$ and $x_{(n+1)}$ is the same as their location in the 'Circle' method. In other words, over many trials the results from the 'Random Points' method will agree with those from the 'Circle' method. The advantage of the 'Circle' method is that it is not sensitive to random fluctuations outside the study interval and the correction factor depends only on the data set and not on additional random processes.

1.5.6 Summary of the Correction Methods

Table 1 below summarizes the differences between the expected values of the uncorrected and corrected means of nearest neighbor distances, which were derived in Sections 1.5.1 through 1.5.4.

Table 1: Differences Between Expected Values

Method of Correction	*Expected Value*	*Difference*
'Circle'	$\frac{1}{2n}$	$\frac{1}{2}\ \frac{1}{n(n+1)}$
'Mirror'	$\frac{3n+2}{6n(n+1)}$	$\frac{2}{3}\ \frac{1}{n(n+1)}$
'Boundary'	$\frac{1}{2(n+1)}$	$\frac{1}{n(n+1)}$
'Expected Value'	$\frac{2n-1}{4n(n+1)}$	$\frac{5}{4}\ \frac{1}{n(n+1)}$

The differences between the expected values are of magnitude $\frac{1}{n(n+1)}$. This implies that they decrease as the sample size increases. Since the correction procedures are applied to two nearest neighbor distances at most, their effect on the expected mean value gets smaller as the number of points gets larger. Furthermore, all differences are in the same direction and are greater than zero. This reflects the process of the edge effects correction. While the uncorrected nearest neighbor distance is the distance between the extreme point and its neighbor, the corrected distance is the minimum between that distance and a correction factor. Consequently, the expected value of the corrected mean can be equal to or smaller than the expected value of the uncorrected mean.

The 'Circle' method yields a corrected expected value which is the most similar to the uncorrected one. The 'Expected Value' method yields the least similar corrected expected value. A further investigation of these expected values and their standard deviations is needed in order to evaluate which correction method gives the best results.

1.6 Simulation Results

The simulations were conducted using a prescribed number of random points (n) on a unit length line between 0 and 1. The set of points was then used to calculate n nearest neighbor distances and the uncorrected mean nearest neighbor distance. The corrected nearest neighbor distances and their mean were calculated using four different methods of correction: the 'Circle', the 'Boundary', the 'Mirror', and the 'Expected Value'.
The process was simulated 10,000 times for each value of n. Then the overall uncorrected and corrected means and their standard deviations were calculated.

The 'Random Points' method was applied differently. For each set of n random points on a unit length line between 0 and 1, two new sets of n random points were generated. One set on a unit length line between -1 and 0, and the other on a unit length line between 1 and 2. The largest point from the left hand side set and the smallest point from the right hand side set became candidates for being nearest neighbors. The corrected mean nearest neighbor distance was then calculated. This proccess was repeated 100 times for each set of n random points and the overall mean was calculated. After repeating the process 100 times, a new set of random points was generated and the whole process was repeated another 100 times. This procedure was conducted 1000 times for each n and the overall corrected mean and its standard deviation were then calculated.

The results of the simulations are summarized in Table 2 on the next page. The table's rows correspond to the different methods of correction and its columns correspond to four sample sizes, $n = 5$, 10, 100, 200. For each sample size, the table presents the uncorrected and the five corrected means of nearest neighbor distances, together with their standard deviations.

The differences between the uncorrected and the corrected means and between the various corrected means get smaller as the sample size gets larger. This result corresponds with the calculations which are shown in Table 1 and indicate that those differences are of magnitude $\frac{1}{n(n+1)}$. The dependence of those differences on the sample size reflects the fundamental property of the mean. As it is well known, the mean is a summary statistic which is sensitive to the contributions of all random points. However, as explained in Section 1.5.6, the effect of a change in one or two nearest neighbor distances decreases as the sample size increases.

Table 2: Simulation Results for the Correction Methods

Method of Correction	*Sample Size*							
	$n = 5$		$n = 10$		$n = 100$		$n = 200$	
	Mean	*Std.Dev.*	*Mean*	*Std.Dev.*	*Mean*	*Std.Dev.*	*Mean*	*Std.Dev.*
'Circle'	0.099801	0.033005	0.050098	0.012229	0.004998	0.000407	0.002500	0.000144
'Boundary'	0.083148	0.028466	0.045527	0.011447	0.004947	0.000404	0.002487	0.000143
'Mirror'	0.094232	0.032919	0.048561	0.012296	0.004981	0.000408	0.002496	0.000144
'Expected Value'	0.076858	0.027920	0.043687	0.011629	0.004927	0.000404	0.002482	0.000143
'Random Points'	0.101118	0.036059	0.050247	0.012848	0.004998	0.000408	0.002499	0.000144
No Correction	0.116456	0.043828	0.054594	0.014630	0.005048	0.000414	0.002512	0.000145

For each sample size, the uncorrected mean nearest neighbor distance is larger than all corrected means. This result reflects the process of correction. While the uncorrected nearest neighbor distance is the distance between the extreme point and its neighbor, the corrected distance is the **minimum** between the above distance and a correction factor. This implies that the corrected mean can be equal to or lower than the uncorrected one.

Similarly, for each sample size, the standard deviation of the uncorrected mean nearest neighbor distance is larger than the standard deviations of the corrected means.

A comparison of the different correction methods reveals that two of them, the

'Circle' and the 'Random Points', yield similar values of the mean nearest neighbor, even for small sample sizes. This result corresponds to the similarity between these methods, as was explained in Section 1.5.5. That is, the expected location of the two candidates for being the extreme points' nearest neighbors are the same in both methods. Consequently, the means of the nearest neighbor distances are expected to be similar. However, the standard deviation of the mean nearest neighbor distance in the 'Circle' method is smaller than the one in the 'Random Points' method. This reflects the advantage of the 'Circle' method where the mean depends only on the data set and not on additional random processes outside the study interval.

Another correction method, the 'Mirror' method, also gives similar values of the mean nearest neighbor distance, especially as the sample size increases. An important result is that the standard deviation of the mean is smaller than the one from the 'Random Points' method, and similar to the one from the 'Circle' method. Consequently, the 'Mirror' method can be assumed a possible alternative to the 'Circle' method. This assumption is especially important for the two–dimensions case since the 'Mirror' method can be easily generalized to two dimensions, even if the shape of study area is irregular, while the 'Circle' method can be generalized to two dimensions only if the study area has a rectangular shape.

The 'Expected Value' and the 'Boundary' methods yield the smallest values of the mean nearest neighbor distance and its standard deviation. Those values of the standard deviations correspond to the common characteristic of the two methods. Both methods, unlike the others, rely only on information from the given set of random points. The 'Boundary' correction factor is based on the distance to the interval boundaries and the 'Expected Value' correction factor is based on the number of random points on the interval. The other correction methods rely on the assumption that the same processes responsible for the location of the points on the interval are operating beyond its boundaries. The calculation of the correction factors introduce additional variablity which affects the values of the mean's standard deviation.

Bibliography

[1] P. J. Clark and F. C. Evans. Distance to nearest neighbour as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453, 1954.

[2] P. J. Clark and F. C. Evans. On some aspectes of spatial patterns in biological populations. *Science*, 121:397–398, 1955.

[3] P. J. Clark. Grouping in spatial distributions. *Science*, 123:373–374, 1956.

[4] M. F. Dacey. The spacing of river towns. *Annals of the Association of Americal Geographers*, 50:59–61, 1960.

[5] D. A. Pinder and M. E. Witherick. Nearest–neighbour analysis of linear point pattern. *Tijdschrift voor Economische en Sociale Geografie*, 64:160–163, 1973.

[6] D. A. Pinder and M. E. Witherick. A modification of nearest–neighbour analysis for use in linear situations. *Geography*, 60:16–23, 1975.

[7] D. L. Young. The linear nearest neighbor statistic. *Biometrika*, 69:477–480, 1982.

[8] H. R. Neave and K. E. Selkirk. Nearest–neighbour analysis of the distribution of points on a circle. Technical Report 05–83, Nottingham Statistics Group Research, Nottingham University Department of Mathematics, 1983.

[9] H. R. Neave and K. E. Selkirk. Nearest–neighbour analysis of the distribution of points on a line. Technical Report 08–83, Nottingham Statistics Group Research, Nottingham University Nottingham University Department of Mathematics, 1983.

[10] K. E. Selkirk and H. R. Neave. Nearest–neighbour analysis of one–dimensional distributions of points. *Tijdschrift voor Economische en Sociale Geografie*, 75(5):356–362, 1984.

[11] L. J. S. Tsuji. A measure of spacing in one dimension. *Mathematical Biosciences*, 109:11–17, 1992.