A Note on Choices of Bivariate Histogram Bin Shape

by

D.W. Scott¹

Technical Report 85-3, April 1985

¹Mathematical Sciences Department, Rice University, Houston, Texas 77251.
A Note on Choice of Bivariate Histogram Bin Shape

Abstract. Regular tiling of the plane may be accomplished with square, triangular, or hexagonal tiles. The statistical properties of bivariate histograms with square bin shapes are evaluated. Hexagonal bins are shown to be best but only marginally better than square bins, which are 96% efficient.

1. Introduction.

The square tiling pattern shown in Figure 1 is the standard basis of bivariate histogram construction. Improved histograms may be realized in two ways. First the square bins may be generalized to rectangular bins. It is possible to show that significant improvement in global error measures such as integrated mean squared error

\[ IMSE = \iint E[(\hat{f}(x,y) - f(x,y))^2 \, dx \, dy] \]

may be realized. Alternately or in conjunction, the pattern of bins may be rotated away from the co-ordinated direction. In this note I shall ignore the two methods of improvement mentioned above and focus on the shape of the histogram bias. For real applications the pattern of bins should be generalized to include stretching and possibly rotation.

2. Global Error for Several Bin Shapes

2.1 Square Bins

We illustrate the method of computing the integrated mean squared error with square bins. Consider a typical bin as shown in Figure 1b with origin at the center. Now

\[ \hat{f}(x,y) = \frac{v(x,y)}{nA^2} \quad (2.1) \]

where \( v(x,y) \) is the sample count for the bin containing \((x,y)\). Clearly \( E\hat{v}(x,y) = np(x,y) \) where

\[ p(x,y) = \int \int f(x,y) \, dx \, dy \quad (2.2) \]

for \((x,y)\) in this bin. Assuming
\[ f(x, y) = f(0, 0) + xf_x(0, 0) + yf_y(0, 0) + O(h^2) \]  \hfill (2.3)

it follows that

\[ p(x, y) = h^2f(0, 0) + O(h^4) \]  \hfill (2.4)

and using (2.3),

\[ Bins(x, y) = E\dot{f}(x, y) - f(x, y) = f_x(0, 0) - yf_y(0, 0) + o(h^2) \]  \hfill (2.6)

Hence

\[ \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} Bins(x, y)^2 \, dx \, dy = \frac{1}{12} h^4 f_x^2(0, 0)^2 + \frac{1}{12} h^4 f_y^2(0, 0)^2. \]  \hfill (2.7)

Equation (2.7) holds in every bin with (0,0) replaced by its center. Summing over all bins and using standard approximation, we find

\[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Bins(x, y)^2 \, dx \, dy = \frac{1}{12} h^2 f_x^2 + \frac{1}{12} h^2 f_y^2. \]  \hfill (2.8)

where

\[ I^2 = \int_{-\infty}^{\infty} f_x(x, y)^2 \, dx \, dy. \]  \hfill (2.9)

Now

\[ (2.10) \ Var \ \dot{f}(x, y) = \frac{f(0, 0)}{nh^2}. \]

Next integrate (2.10) for the bin and sum these expressions for all bins. Then

\[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Var \ \dot{f}(x, y) \, dx \, dy = \frac{1}{nh^2}. \]  \hfill (2.11)

Adding (2.8) and (2.11), we show

\[ IMSE = \frac{1}{nh^2} + \frac{h^2}{12} I^2_x + \frac{h^2}{12} I^2_y \]  \hfill (2.12)

2.2 Triangular Bin Choices

2.2.1 Diagonal Cuts
Consider the pattern shown in Figure 2a, which may be built from the basic 2-bin tile shown in Figure 2b, I proceed as in Section (2.1), computing bias and variance estimates for the basic shape in Figure 2b and extending to the plane. The result is

\[
IMSE = \frac{2}{nh^2} + \frac{1}{18} h^2 (I_x^2 + I_y^2 + I_{xy}),
\]  

(2.13)

where we have unfortunately encountered an additional term

\[
I_{xy} = \int \int f_x(x, y) f_y(x, y) \, dx \, dy.
\]  

(2.14)

Comparison with (2.12) is facilitated if we redefine \( h \) by multiplying by \( \sqrt{2} \). Then (2.13) becomes

\[
IMSE = \frac{1}{nh^2} + \frac{1}{9h^2} (I_x^2 + I_y^2 + I_{xy}).
\]  

(2.15)