Journal of Probability and Statistical Science
24(2), 1-15Aug. 2026

Visualizing the Regression Lines and the Correlation Coefficient for Bivariate Discrete Random Variables

Jyotirmoy Sarkar
Department of Mathematical Sciences
Indiana University Indianapolis, USA

Mamunur Rashid
Department of Mathematical Sciences
DePauw University Indiana, USA

ABSTRACT

Sarkar and Rashid (2024) visualized the joint probabilities of two discrete random variables using a 2D disc plot from which they graphically obtained the conditional distributions of Y given $X = a$, and X given $Y = b$; and the marginal distributions of X, Y, $X + Y$, $X - Y$, and $cX + dY$. Sarkar and Rashid (2025) visualized the mean, mean deviation, and standard deviation of any discrete random variable using a disc plot. Here, starting with the 2D disc plot, we visualize the correlation coefficient between X and Y, and the linear regression of one variable from the other. Such visualizations help users better comprehend these concepts. Thus, this paper develops a unified geometric framework to visualize conditional means, the overall mean vector, least-squares regression lines, and the correlation coefficient, starting from the disc plot.

Keywords: Teaching statistics; joint distribution; disc plot; correlation; regression.

1. Introduction

Joint distributions of two or more random variables (RVs) frequently arise in probability and statistics. For a single discrete RV (DRV), we depict its probability mass function (PMF) using a 2D stick plot showing the value along the x-axis and the associated probability $p_X(x)$ as the height of a stick standing at x parallel to the y-axis. See Sarkar and Rashid (2019). Further, they depict the mean and the standard deviation (SD) as the length of a single-headed arrow, starting at the mean. Recall that the mean of any DRV is $\bar{x} = \sum x \cdot p\,(x)$, the mean deviation (MD) is $\bar{d} = \sum |x - \bar{x}| \cdot p\,(x)$, the mean square is $E[X^2] = \sum x^2 \cdot p(x)$, the mean square deviation or the variance is $\sigma^2 = \sum (x - \bar{x})^2 \cdot p\,(x) = E[X^2] - \bar{x}^2$, and the SD is $\sigma = \sqrt{\sigma^2}$.

To visualize two DRVs simultaneously, one may consider a bivariate stick diagram — a 3D picture showing the values of X and Y (on the xy-plane) and the associated probability $p(x, y)$ as the height of the stick orthogonal to the xy-plane (or parallel to the z-axis). See Wackerly et al. (2011). However, recognizing the difficulty many students face in understanding and further processing 3D pictures, Sarkar and Rashid (2024) proposed displaying the joint PMF of two DRVs using a disc plot, in which each disc is centered at the vector (x, y) and has area proportional to the joint probability $p(x, y)$ at (x, y). Equivalently, the disc's radius is proportional to $\sqrt{p(x, y)}$. Further, they obtain the conditional distributions of Y given $X = a$, and X given $Y = b$; the marginal distributions of X, Y, $X + Y$, $X - Y$, and $cX + dY$. For each DRV, Sarkar and Rashid (2025) visualize the mean, mean deviation (MD), and standard deviation (SD) starting with a disc plot, even when the numerical values of x, y, $p(x, y)$ are undisclosed. See Sarkar and Rashid (2016) for visualizing the same quantities for a data set. Here, starting with the bivariate 2D disc plot, we visualize the regression line for predicting Y from X, the correlation coefficient between X and Y, and the regression line for predicting X from Y.

- Received August 30, 2026, in final form August 2026.
- Mamunur Rashid (corresponding author) is affiliated with Department of Mathematical Sciences, DePauw University, Greencastle, Indiana, USA
mrashid@depauw.edu

The purpose is not computational efficiency. Rather, the constructions reveal how regression and correlation emerge from the geometry of conditional means and variability. The resulting pictures make visible relationships that are often hidden in algebraic formulas.

All pictures are drawn using the freeware R, see R Core Team (2022).

2. Weighted Average

Figure 1 shows the weighted average of two bivariate points with weights given by their associated probabilities p and q (represented by the green and black discs centered at the points and with radii $\sqrt{p}$ and $\sqrt{q}$, so that the disc areas are πp and πq, respectively). The weighted average is the orange disc with center at M (explained below) and associated weight $p+q$ (represented by the area $\pi(p+q)$ of the orange circle, the sum of the areas of the green and the black circles). Note that the radius of the orange circle is $\sqrt{p+q}$.

To find M, follow these steps sequentially (See Sarkar and Rashid (2025) for details):

0. Join OP and draw lines through O and P orthogonal to OP.
1. Duplicate the green circle at center O and radius PQ, so that the duplicated green circle is concentric with the black circle having a common center O.
2. Let A and F be the east and the south ends of the new green circle and B and C be the north, and the east ends of the black circle; clearly, OA=OF=PQ=$\sqrt{p}$ and OB=OC=OF=$\sqrt{q}$. Through C, draw a line parallel to AB that intersects extended OB at D.
3. Through D, draw a line parallel to OP, cutting extended PQ at E.
4. Let EF cut OP at M, the center of the circle representing the weighted average. Note that OM:MP = p:q.
5. Draw the orange circle centered at M with radius $\sqrt{p+q}$; hence, area $\pi(p+q)$.

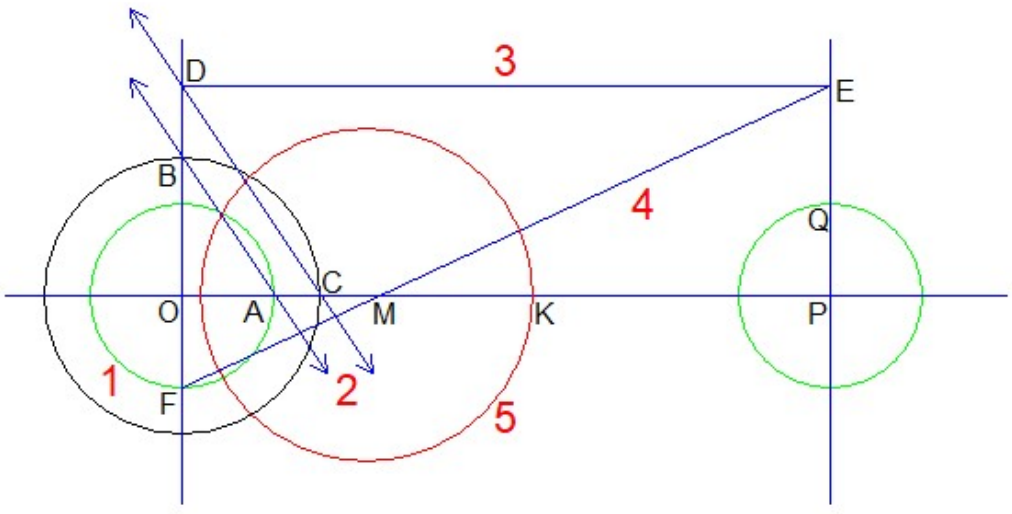


Figure 1: The weighted average of two points O and P with weights p and q, proportional to the areas of green and black circles of radii OB=$\sqrt{p}$ and PQ=$\sqrt{q}$, respectively, is a point M on OP with an associated weight of $p+q$ represented by the area $\pi(p+q)$ of the orange circle.

Weighted Average of Many Points

To construct the weighted average of more than two bivariate points, first take the weighted average of any two points. Treat this weighted average and the remaining (n-2) original points as a set of (n-1) points. Then repeat the process until only one point remains, which is the weighted average of all original points. It is a *theorem* that the final weighted average does not depend on the order in which the points are combined. See Sarkar and Rashid (2024).

Remark 1: This weighted average technique works even when the points are not on a line. We will apply this method in the next sections for points that are either along a line or in a plane. In fact, the construction works in three or higher dimensions also, where spheres or hyperspheres replace the discs.

Weighted Average Principle

The weighted average of points P_i with weights w_i is the unique point

$$M = \frac{\sum w_i P_i}{\sum w_i}.$$

The geometric construction of Figure 1 produces this point using only ruler-and-parallel-line constructions.

3. Regression Lines and Correlation

We illustrate how to calculate the LS-LR lines and the correlation coefficient in two examples.

EXAMPLE 1

Do research mathematicians keep on submitting research papers after retirement? Mathematicians who retired within the last three years reported the number of years after retirement (X) and the number of research papers they submitted within the last year (Y). Table 1 presents the joint PMF and some additional calculations explained below.

Table 1. The joint PMF of years since retirement (X) and the number of research papers submitted during the last year (Y), together with additional calculations

		i				
$y \setminus x$		1	2	3	**Total**	
j	0	.03	.05	.16	**.24**	
	1	.07	.20	.02	**.29**	
	2	.10	.20	.01	**.31**	
	3	.10	.05	.01	**.16**	
Total		**.30**	**.50**	**.20**	**1.00**	
$\bar{y}_i$		.57/.30 =1.90	.75/.50 =1.5	.07/.20 =0.35	$\bar{x} = 1.90$ $\bar{y} = 1.39$	$s_x = 0.7000$ $s_y = 1.0188$
b_i		-.5667	1.1	-.9455	$b = -0.7373$	$r = -0.5066$
w_i		.243	.005	.242	$w_+ = .490$	

The column sums represent the marginal distribution of X and the row sums represent the marginal distribution of Y. Recall that $\bar{x} = \sum x_i p_{i+}$. Given each x_i, the conditional means are defined as

$$\bar{y}_i = \frac{\sum_{j=0}^{3} y_j p_{ij}}{\sum_{j=0}^{3} p_{ij}} = \frac{\sum_{j=0}^{3} y_j p_{ij}}{p_{i+}}.$$

One can check that $\bar{y} = \sum \bar{y}_i p_{i+} = \sum_{i=1}^{3} \sum_{j=0}^{3} y_j p_{ij}$. The conditional slopes are defined as $b_i = \frac{\bar{y}_i - \bar{y}}{x_i - \bar{x}}$.

Also, we define weights $w_i = (x_i - \bar{x})^2 \, p_{i+}$, which will be used in a later section. We normalize the weights by dividing each weight by their sum w_+.

The joint probabilities are depicted in Figure 2 using a disc plot, with the area of each (red) disc proportional to the probability p_{ij} (that is, with radii $\sqrt{p_{ij}}$). Furthermore, Figure 2 shows the conditional weighted averages $\bar{y}_i = E[Y|X = x_i]$ (purple discs with radii $r_i = \sqrt{p_{i+}} = \sqrt{\sum_j p_{ij}}$), using the method illustrated in Figure 1.

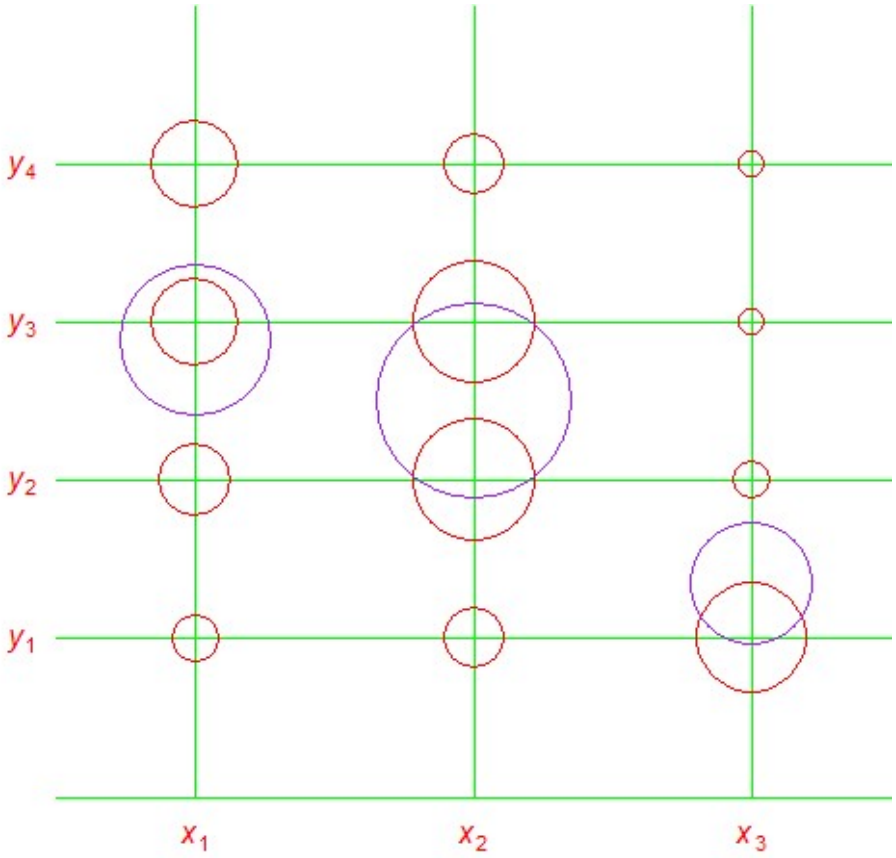


Figure 2: Visualizing the joint *distribution* of (X, Y) using a 2D disc plot and constructing the conditional means $\bar{y}_i$, given x_i

We shall use these purple discs to construct in order: (1) the overall mean vector, (2) the LS-LR line for predicting Y from X, (3) the correlation coefficient between X and Y, and (4) the LS-LR line for predicting X from Y.

The Overall Mean Vector

Using the method described in Figure 1, in Figure 3, we construct the weighted average of the three purple discs (they are not necessarily on a line, but the method still works) to obtain a single black disc with a center at $(\bar{x}, \bar{y})$ and radius 1 or area π (proportional to the total weight 1 of all purple discs).

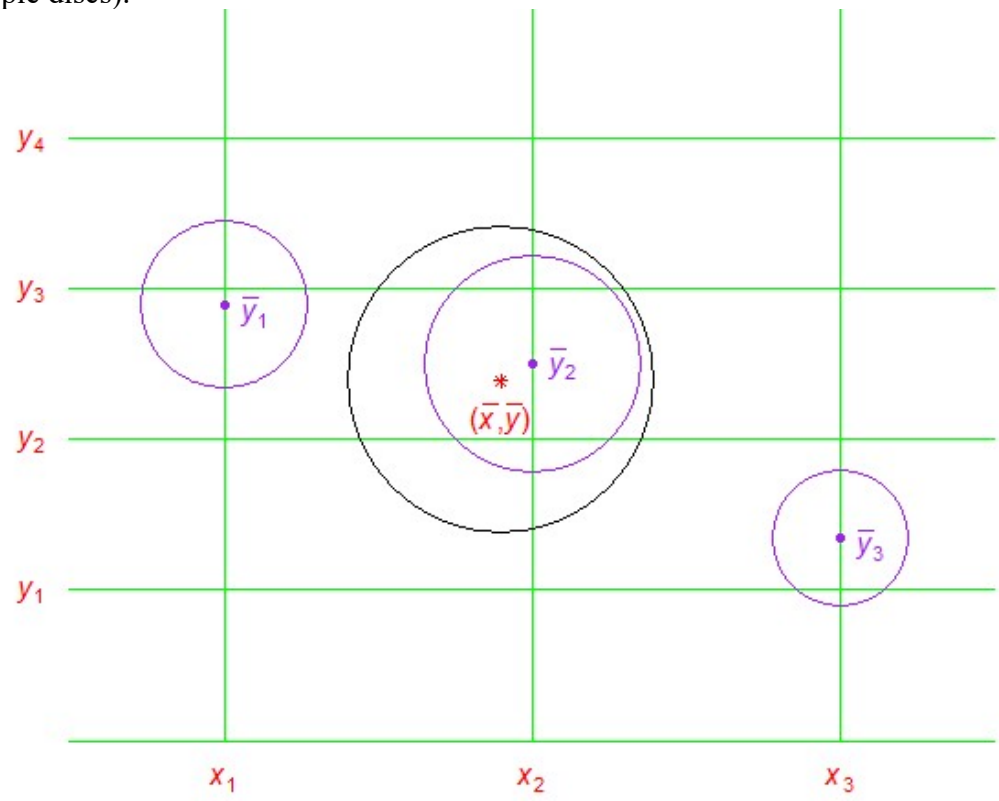


Figure 3: The overall mean $(\bar{x}, \bar{y})$ is the weighted average of the conditional means $\bar{y}_i$ obtained by the method of Figure 1.

Thus, the overall mean vector is obtained by averaging the conditional means using the marginal probabilities as weights.

Revising Weights by Linearizing Products

In the next section, we will choose the weight of each conditional mean $\bar{y}_i$, given x_i, to be proportional to the product of two linear distances. Specifically, we will draw a disc with radius proportional to $|x_i - \bar{x}|\, r_i$. Therefore, we need to know how to linearize a product $d \cdot r$ with respect to an arbitrarily fixed constant D to obtain the quotient dr/D. The technique is explained in Figure 4: Let the vertical line through D intersect the (extended) top edge of the $d \times r$ rectangle at P. Join OP and let it cut the (extended) vertical line through d at Q. Then the segment dQ has length dr/D.

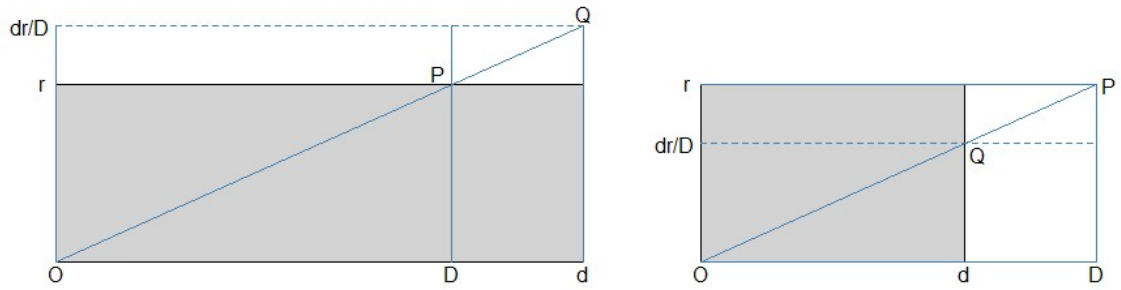


Figure 4: How to linearize the area of a rectangle with sides d and r with respect to an arbitrary constant *D*. Equivalently, how to draw a line segment of length dr/D.

The LS-LR Line $\hat{y}(x)$

Let us describe how, in Figure 5, we construct the LS-LR line to predict Y from *X*. The justification, adapted from Wackerly *et al.* (2014), is postponed to the Appendix.

1. Draw a reference vertical line $x = R$ to the right of all purple discs centered at $(x_i, \bar{y}_i)$'s.
2. Let the line joining each $(x_i, \bar{y}_i)$ to $(\bar{x}, \bar{y})$ meet the vertical line $x = R$ at (R, v_i).
3. Draw discs with centers at (R, v_i) and normalized weights $w_i \propto p_{i+}(x_i - \bar{x})^2$, so that $\sum_i w_i = 1$. The radii are proportional to $|x_i - \bar{x}|\, r_i$, by the method explained in Figure 4.
4. Take the weighted average $(R, \bar{v})$ of all m newly created discs centered on the vertical line $x = R$, using the method explained in Figure 1. There is no need to draw the circle corresponding to $(R, \bar{v})$; its center is sufficient.
5. Join $(R, \bar{v})$ to $(\bar{x}, \bar{y})$. It is the LS-LR line $\hat{y}(x)$.

The central claim of this paper is that the regression slope is itself a weighted average of the conditional slopes b_i. Figure 5 gives a purely geometric realization of this averaging process. The claim is rigorously proved in the Appendix.

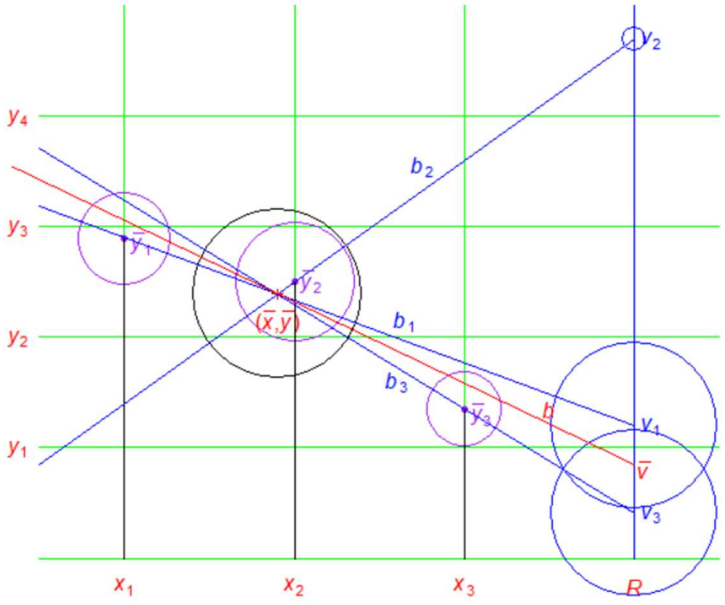


Figure 5: The LS-LR line b joins $(\bar{x}, \bar{y})$ to $(R, \bar{v})$, the weighted average of the discs at points $(R, \bar{v}_i)$'s with areas proportional to $p_{i+}(x_i - \bar{x})^2$

The Correlation Coefficient

Starting with the disc plot of the discrete bivariate PMF, first construct the disc plots of the marginal PMFs of RVs X and Y. Using the technique in Sarkar and Rashid (2025), visualize the mean and SD of the discrete RVs X and Y.

Having found the SDs of X and Y, in Figure 6, we superimpose the mean-SD rectangle $(\bar{x} \pm s_x\ ,\ \bar{y} \pm s_y)$ with center $M = (\bar{x}\ ,\ \bar{y})$, and the LS-LR line $\hat{y}(x)$ for predicting Y from X. The construction of $\hat{y}(x)$ is already explained in Figure 5. Recall that (see also the Appendix) $\hat{y}(x) = a + bx = \bar{y} + b(x - \bar{x}) = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$.

Because the slope of the regression line equals $r\, s_y/s_x$, the regression line intersects the boundary of the mean-SD rectangle at a point whose vertical displacement is proportional to r. As such, r is a signed ratio bounded between -1 and 1.

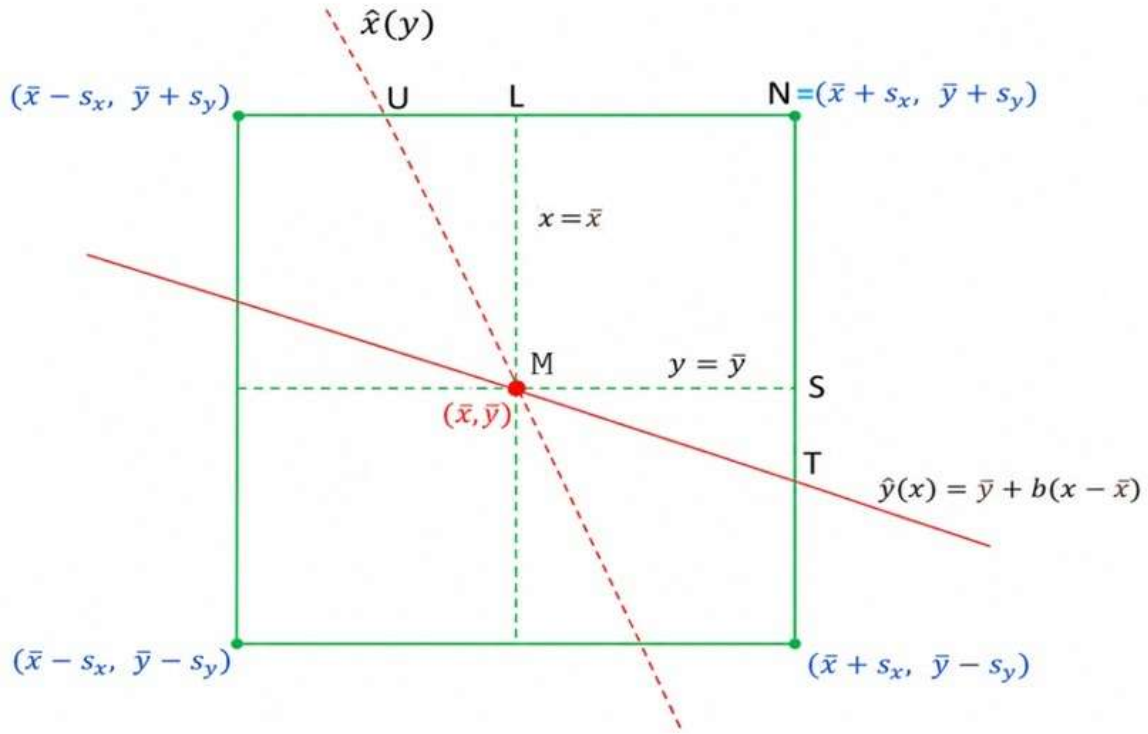


Figure 6: Superimpose the mean-SD rectangle and $\hat{y}(x)$. The correlation coefficient is r =ST/SN. If you pick U on the top edge such that LU/LN= r, then UM is $\hat{x}(y)$.

Let $y = \bar{y}$ and $\hat{y}(x)$ intersect the right edge $x = \bar{x} + s_x$ of the rectangle at S and T, respectively. Then the correlation coefficient is the ratio ST/SN, where ST is positive if T is above S and negative if T is below S, and N is the top-right corner of the rectangle.

Remark 2: The correlation coefficient can be visualized equally well from any c-magnification of the mean-SD rectangle $(\bar{x} \pm cs_x\ ,\ \bar{y} \pm cs_y)$ and the LS-LR line $\hat{y}(x)$. By default, $c = 1$.

The LS-LR Line $\hat{x}(y)$

Next, in Figure 6, on the top edge of the mean-SD rectangle, let L be the midpoint, and U be the point such that LU/LN= r =ST/SN. Then UM is the LS-LR Line $\hat{x}(y)$.

We end the paper by giving a second example (with a positive correlation) and its geometric visualization. We encourage the reader to supply all justifications as was given in Example 1.

EXAMPLE 2

How many adults with a driver's license live in each house in a housing subdivision? Typically, some drivers may own more than one car, and more than one driver may drive the same car. Let X denote the number of cars parked in the garage or the driveway of each house, and Y denote the number of licensed adults living in the house. Their joint PMF is given in Table 2.

Table 2: The joint PMF of the number of cars (X) and the number of licensed drivers (Y)

$y \setminus x$	1	2	3	4	**Total**	
1	.08	.02	.00	.00	**.10**	
2	.02	.18	.10	.02	**.32**	
3	.00	.08	.20	.05	**.33**	
4	.00	.02	.15	.08	**.25**	
Total	**.10**	**.30**	**.45**	**.15**	**1.00**	
$\bar{y}_i$	1.20	2.33	3.11	3.40	$\bar{x} = 2.65$ $\bar{y} = 2.73$	$s_x = .8530$ $s_y = .9472$
b_i	153/165	40/65	38/35	67/135	$b = .7230$	$r = .6511$
w_i	.27225	.12675	.055125	.273375	$w_+ = .7275$	

In Figure 7, starting with the joint *distribution* of (X, Y) shown as a 2D disc plot, we construct the conditional means $\bar{y}_i$, given x_i, and the overall mean $(\bar{x}, \bar{y})$**.** Figure 8 shows the LS-LR line with slope b, which joins $(\bar{x}, \bar{y})$ to $(R, \bar{v})$, the weighted average of the discs at points $(R, \bar{v}_i)$'s with areas proportional to $p_{i+}(x_i - \bar{x})^2$. In Figure 9, we superimpose the mean-SD rectangle and the LS-LR line $\hat{y}(x)$. The correlation coefficient is r =ST/SN. If on the top edge, we pick the midpoint L and the point U such that LU/LN= r, then UM is the LS-LR line $\hat{x}(y)$.

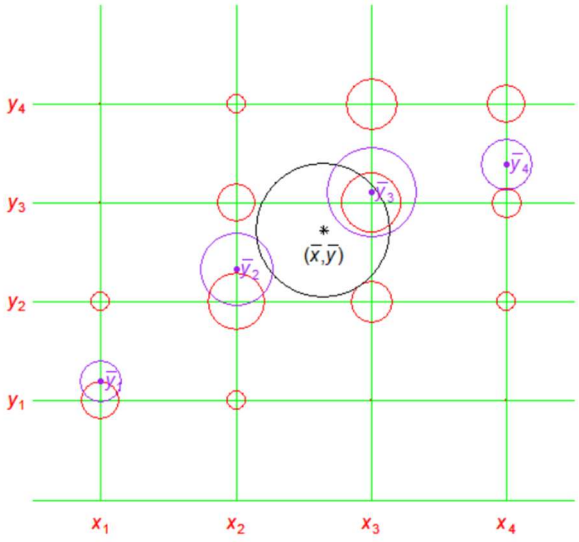


Figure 7: Starting with the joint *distribution* of (X, Y) shown as a 2D disc plot, we construct the conditional means $\bar{y}_i$, given x_i, and the overall mean $(\bar{x}, \bar{y})$**.**

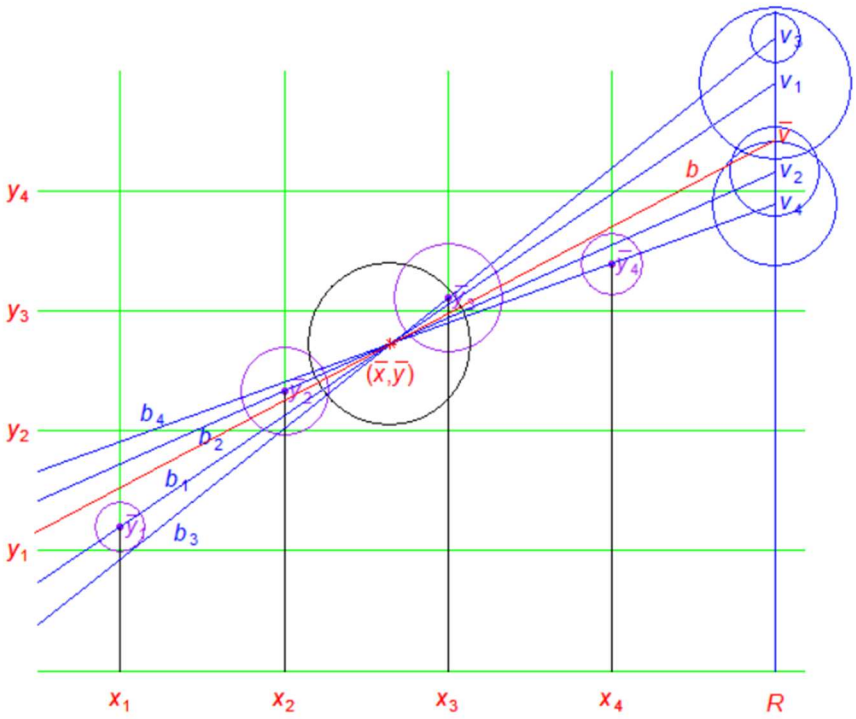


Figure 8: The LS-LR line b joins $(\bar{x},\ \bar{y})$ to $(R, \bar{v})$, the weighted average of the discs at points $(R, \bar{v}_i)$'s with areas proportional to $p_{i+}(x_i - \bar{x})^2$.

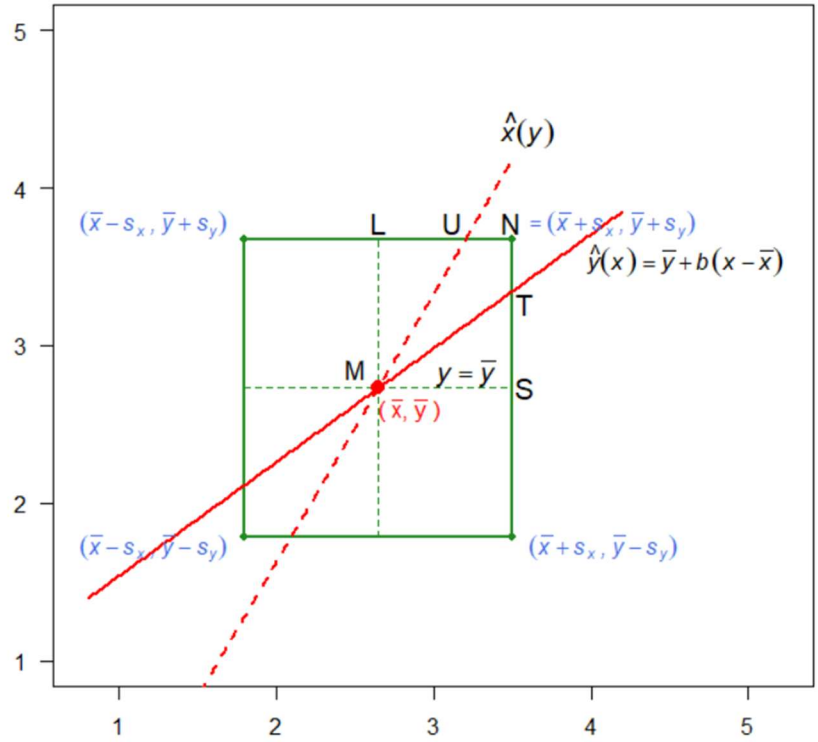


Figure 9: Superimpose the mean-SD rectangle and $\hat{y}(x)$. The correlation coefficient is r =ST/SN. If on the top edge, L is the midpoint and U is the point such that LU/LN= r, then UM is $\hat{x}(y)$.

4. Conclusion

To efficiently visualize the joint distribution of two DRVs, Sarkar and Rashid (2024) promoted 2D disc plots. They also yield the marginal distributions of X and Y, conditional distributions of $Y|X = a$ and $X|Y = b$, and distributions of the sum $X + Y$, difference $X - Y$, and any other linear combination. Sarkar and Rashid (2025) visualized the mean, MD, and SD of any DRV using a disc plot. Here, we have shown that the 2D disc plots also yield the two LS-LR lines and the correlation coefficient, even when the values of $x, y, p(x, y)$ are not disclosed!

Thus, this paper develops a unified geometric framework that begins with the joint distribution of two discrete random variables and graphically derives conditional means, the overall mean vector, least-squares regression lines, and the correlation coefficient.

We hope the visualizations will empower students and users of probability and statistics. They will also be able to visualize the corresponding elements from a scatterplot of a random sample of bivariate data.

References

1. R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Sarkar, J. and Rashid, M. (2015). A Geometric View of the Mean of a Set of Numbers. *Teaching Statistics—An International J. for Teachers*, **38**(3), 77-82.
3. Sarkar, J. and Rashid, M. (2016). Visualizing Mean, Median, Mean Deviation, and Standard Deviation of a Set of Numbers. *The American Statistician*, **70**(3), 304-312.
4. Sarkar, J. and Rashid, M. (2019). Euclidean plane geometry suffices to visualize the standard deviation. *The Mathematics Student*, **88**(3-4), 5-16.
5. Sarkar, J. and Rashid, M. (2024). Visualizing Distributions Derived from the Joint Distribution of Two Discrete Random Variables. *Teaching Statistics—An International J. for Teachers*, **46**(3), 141-151.
6. Sarkar, J. and Rashid, M. (2025). Visualizing the Mean, the Mean Deviation, and the Standard Deviation from the Disc Plot of a Discrete Random Variable. *Journal of Probability and Statistical Science*, **24**(1), 1-15.
7. Wackerly, D., Mendenhall, W., and Scheaffer, R. (2014). *Mathematical Statistics with Applications* (8th Ed), Brooks/Cole Cengage Learning, CA, USA.

APPENDIX

Lemma 1: Suppose that (X, Y) are bivariate DRVs with a support set $S = \{(x_i, y_j): i = 1,2, \dots, m; j = 1,2, \dots, n\}$ and joint PMF $P\{X = x_i, Y = y_{ij}\} = p_{ij}$. Define the conditional means $E[Y|x = x_i] = \bar{y}_i = \sum_{j=1}^{n} p_{ij} y_{ij} / p_{i+}$, the marginal PMF of X as $\mathrm{P\{X = x_i\}} = p_{i+} = \sum_{j=1}^{n} p_{ij}$, the overall mean of X as $E[X] = \bar{x} = \sum_{j=1}^{n} p_{i+} x_i$, and the overall mean of Y as $E[Y] = \bar{y} = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} y_j = \sum_{j=1}^{n} p_{i+} \bar{y}_i$. Then the slope b of the LS-LR Line $\hat{y}(x)$ is the weighted average of the slopes of the m lines joining the conditional means $(x_i, \bar{y}_i)$ to the overall mean $(\bar{x}, \bar{y})$, with weights proportional to $(x_i - \bar{x})^2 \, p_{i+}$.

Proof: The proof is adapted from Wackerly et al. (2014).

The LS-LR line is $\hat{y}(x) = a + bx$, where the intercept a and the slope b are chosen to minimize $Q(a, b) = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \left(y_j - a - bx_i\right)^2$. The solutions must satisfy:

(1) the first-order conditions $Q_{10}(a, b) = 0$, $Q_{01}(a, b) = 0$; and

(2) the second-order condition that the matrix $M = \begin{bmatrix} Q_{20} & Q_{11} \\ Q_{11} & Q_{02} \end{bmatrix}$ of second-order partial derivatives is positive definite.

The first-order conditions yield

$$0 = Q_{10}(a, b) = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \left(y_j - a - bx_i\right)(-2),$$

$$0 = Q_{01}(a, b) = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \left(y_j - a - bx_i\right)(-2x_i).$$

After some algebraic simplifications, the first-order conditions become

$$\begin{cases} \bar{y} = a + b\bar{x}, \\ \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} y_j x_i = a\bar{x} + b \sum_{i=1}^{m} p_{i+} x_i^2 \end{cases}$$

whence, by subtracting $\bar{x}$ times the first equation from the second, we have

$$b = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} y_j x_i - \bar{x}\,\bar{y}}{\sum_{i=1}^{m} p_{i+} x_i^2 - \bar{x}^2} = \frac{\sum_{i=1}^{m} p_{i+} (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sum_{i=1}^{m} p_{i+} (x_i - \bar{x})^2} = r \frac{s_y}{s_x},$$

where the correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^{m} p_{i+} (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} p_{i+} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{m} p_{i+} (\tilde{y}_i - \bar{y})^2}} = \frac{s_{xy}}{s_x \, s_y}.$$

Moreover, from the second-last equation, we have

$$b = \sum_{i=1}^{m} \frac{p_{i+} (x_i - \bar{x})^2}{\sum_{i=1}^{m} p_{i+} (x_i - \bar{x})^2} \times \frac{(\bar{y}_i - \bar{y})}{(x_i - \bar{x})} = \sum_{i=1}^{m} w_i \, b_i,$$

which is a weighted average of b_i with weights $w_i \propto p_{i+} (x_i - \bar{x})^2$ such that $\sum_i w_i = 1$.

It remains to verify the second-order condition. Towards that end, note that

$$M = \begin{bmatrix} Q_{20} & Q_{11} \\ Q_{11} & Q_{02} \end{bmatrix} = 2 \begin{bmatrix} \sum\sum p_{ij} & \sum\sum p_{ij} x_i \\ \sum\sum p_{ij} x_i & \sum\sum p_{ij} x_i^2 \end{bmatrix} = 2 \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & E[X^2] \end{bmatrix}$$

has diagonal entries positive and $\det(M) = E[X^2] - \bar{x}^2 = \mathrm{var}(X) > 0$.