Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 77-87

Depicting Bivariate Relationship with a Gaussian Ellipse

Jyotirmoy Sarkar[1] **and Mamunur Rashid**[2]

[1]*Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, USA*

[2]*Department of Mathematics, DePauw University, Greencastle, Indiana, USA*

Received: 15 June 2020; Revised: 01 August 2020; Accepted: 08 August 2020

Abstract

For data on two continuous variables, how should one depict the summary statistics (means, SDs, correlation coefficient, coefficient of determination, regression lines) so that their values can be read off easily from the depiction and potential outliers can be flagged also? We propose the Gaussian covariance ellipse as an answer that will benefit all users of statistics.

Key words: Five-number summary; Variance; Least squares lines; Bivariate normal distribution; x-, y- and regression outliers.

AMS Subject Classification: 62H20

1. Introduction

Appropriate graphical representation of data is necessary for easy comprehension of underlying information. For each type of variable and for each objective, one must choose the correct graph to depict the data. In this paper, we focus only on *quantitative variables* which take values on a continuous scale; that is, even though measurement limitations may force us (and ease of comprehension may prompt us) to report the value correct to an integer or up to a few decimal places, we recognize that finer values are surely possible. Some quantitative variables are measured only on a *difference scale*, where the difference between two values has a meaningful interpretation, but not their ratio; and other quantitative variables are measured on an *interval* or *ratio scale*, where the ratio of two values has a proper physical interpretation.

The objective of this paper is to depict the summary statistics of two quantitative variables that are related via the linear regression model or that exhibit a bivariate normal distribution. Section 2 identifies some commonly used bivariate statistics and poses the problem of depicting them efficiently. Section 3 depicts bivariate linear association for standardized data using a *correlation ellipse*; and Section 4 depicts bivariate summary statistics for raw data using a *covariance ellipse*. Section 5 highlights the sufficient statistics from which other bivariate summary statistics can be reconstructed; and Section 6 further reduces the sufficient statistics. Section 7 concludes the paper, interprets the covariance ellipse and poses some directions of future research.

Corresponding Author: Jyotirmoy Sarkar
Email: jsarkar@iupui.edu

2. Depicting Bivariate Summary Statistics: Statement of the Problem

Methods are well-known for depicting summary statistics of a quantitative variable. We refer the reader to Maverick (1932), Embse and Engebretsen (1996), and Sarkar and Rashid (2016, 2019). Also, above (or below) a dot plot or a histogram one can easily superimpose an arrow, whose tail shows the mean and length the SD. See, for example, Devore (2015) and Rashid and Sarkar (2018). Likewise, to depict the interrelations between two quantitative variables, the commonly used scatter plot can be augmented by the five-number-summary, the mean and the SD of each variable in the margins; that is, when the scatter plot is projected along each coordinate axis, the corresponding dot plots can be summarized using univariate methods. See an example given in Figure 1 with details found in Sarkar and Rashid (2020).

Throughout the paper we illustrate some visualization techniques using the following example involving the midterm exam score (x) and the final exam score (y) of 23 students in an *Introduction to Statistics* course.

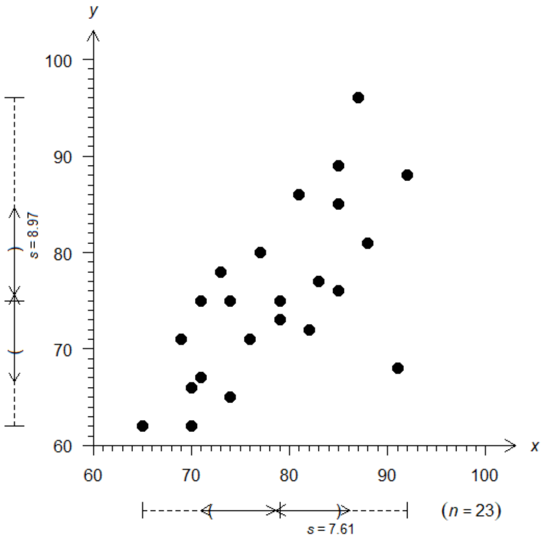


Figure 1: The midterm exam (x) and the final exam (y) scores of 23 students, with the five-number-summary, mean and SD of each score shown in the margins

To the astute reader we pose a quiz: "Projection of a scatter plot in which direction will cause the corresponding dot plot to exhibit the largest (or the smallest) SD? Alternatively, which linear combination of x and y has the largest (or the smallest) SD?" We urge the readers to jot down their answers before reading further. Our answer is given later in this paper.

Frequently used bivariate summary statistics include the correlation coefficient r, the least squares regression line $\hat{y}$ as a function of x, the *inverse* least squares regression line $\hat{x}$ as a function of y, and the coefficient of determination r^2. How should these bivariate statistics be depicted so that their numerical values can be easily read off from the scales of the axes?

As a solution to this question, we propose to draw the Gaussian covariance ellipse that fits inside the c-SD rectangle given by the c-SD boundaries $y = \bar{y} \pm cs_y$ and $x = \bar{x} \pm cs_x$. The diagonals of this rectangle intersect at the mean vector $(\bar{x}, \bar{y})$. We shall exhibit the Gaussian covariance ellipse in Section 4. But first, in Section 3, let us look at the Gaussian correlation ellipse, which only considers the standardized variables. It strips out the central location vector

$(\bar{x}, \bar{y})$, suppresses the scale factors s_x and s_y, and focuses on the correlation coefficient r, the coefficient of determination r^2 and the two regression lines (in standardized units).

3. Gaussian Correlation Ellipse

Let us first focus on the correlation coefficient r. To do so, we replace the variables by their standardized versions: Replace x by $\tilde{x} = (x - \bar{x})/s_x$ and y by $\tilde{y} = (y - \bar{y})/s_y$. Consequently, the mean vector for $(\tilde{x}, \tilde{y})$ is (0, 0), and $s_{\tilde{y}} = s_{\tilde{x}} = 1$. In particular, the c-SD rectangle for $(\tilde{x}, \tilde{y})$ is a square (so long as the scales of the two axes in the diagram are chosen to be the same). We inscribe in this square the c-SD Gaussian correlation ellipse whose two axes pass through $(0, 0)$ and have slopes 1 and -1, and is internally tangential to the c-SD square at exactly four points: bottommost point $B = (-rc, -c)$, topmost point $T = (rc, c)$, leftmost point $L = (-c, -rc)$ and rightmost point $R = (c, rc)$. Then LR is the regression line $\hat{\tilde{y}} = r\tilde{x}$ line, BT is the inverse regression line $\hat{\tilde{x}} = r\tilde{y}$ line. Furthermore, these two regression lines LR and BT intersect at the center (0, 0) of the c-SD correlation ellipse, which is also the point of intersection of the two diagonals of the c-SD square and the *center of gravity* of the scatter plot of standardized variables $(\tilde{x}, \tilde{y})$.

For the example data, shown in Figure 1, after standardizing the scores, Figure 2 depicts the *standard* correlation ellipse, where we have chosen $c = 1$, so that 39.35% of the data are expected to fall inside. We will say more about the choice of c towards the end of Section 4.

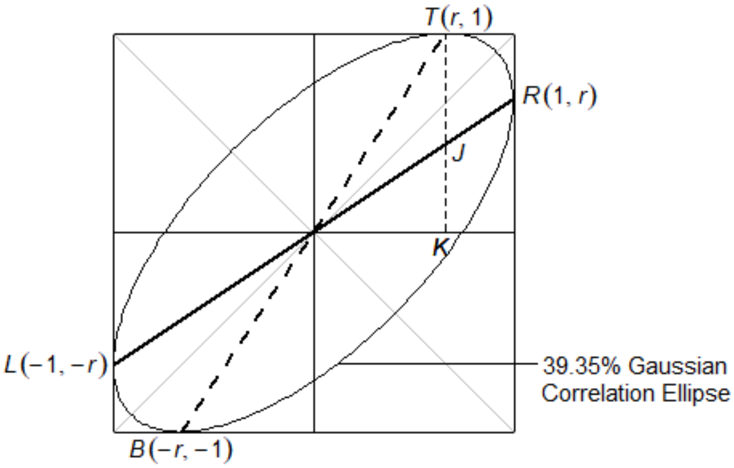


Figure 2: The 1-SD square and the 1-SD correlation ellipse for midterm and final exam scores of 23 students in an *Introduction to Statistics* course

Based on the standardized variables, the standard (that is, with $c = 1$) correlation ellipse is centered at the origin, has one axis with half-length a (to be determined in the next paragraph) in the direction $(1, 1)$ starting from the center, and has the other (orthogonal) axis with half-length b (also to be determined shortly) in the direction $(1, -1)$ starting from the origin. After normalizing the direction vectors (that is, dividing each vector by $\sqrt{2}$), the standard correlation ellipse can be described by the equation

$$\frac{(\tilde{x}+\tilde{y})^2}{2a^2} + \frac{(\tilde{x}-\tilde{y})^2}{2b^2} = 1. \tag{1}$$

Next, focusing on the upper $\tilde{y}$-value as a function of $\tilde{x}$ and implicitly differentiating with respect to $\tilde{x}$, we have

$$\tfrac{(\tilde{x}+\tilde{y})}{a^2}(1+\tilde{y}') + \tfrac{(\tilde{x}-\tilde{y})}{b^2}(1-\tilde{y}') = 0. \tag{2}$$

The standard correlation ellipse is internally tangential to the 1-SD square at the point $(r,1)$. Hence, we have $\tilde{y}'(r) = 0$ and $\tilde{y}(r) = 1$, and equations (1) and (2) yield

$$\tfrac{(r+1)^2}{2a^2} + \tfrac{(r-1)^2}{2b^2} = 1; \text{ and } \quad \tfrac{(r+1)}{a^2} + \tfrac{(r-1)}{b^2} = 0. \tag{3}$$

Solving the two equations in (3) simultaneously, we determine $a = \sqrt{1+r}$ and $b = \sqrt{1-r}$.

Having determined a and b, the standard correlation ellipse can be described by any one of the following equivalent equations {of which we prefer the last; that is, expression (4)}:

$$\frac{(\tilde{x}+\tilde{y})^2}{2(1+r)} + \frac{(\tilde{x}-\tilde{y})^2}{2(1-r)} = 1$$

$$\begin{pmatrix}\frac{\tilde{x}+\tilde{y}}{\sqrt{2}} & \frac{\tilde{x}-\tilde{y}}{\sqrt{2}}\end{pmatrix}\begin{bmatrix}1+r & 0\\ 0 & 1-r\end{bmatrix}^{-1}\begin{pmatrix}(\tilde{x}+\tilde{y})/\sqrt{2}\\ (\tilde{x}-\tilde{y})/\sqrt{2}\end{pmatrix} = 1$$

$$(\tilde{x} \quad \tilde{y})\begin{bmatrix}\sqrt{1/2} & \sqrt{1/2}\\ \sqrt{1/2} & -\sqrt{1/2}\end{bmatrix}\begin{bmatrix}1+r & 0\\ 0 & 1-r\end{bmatrix}^{-1}\begin{bmatrix}\sqrt{1/2} & \sqrt{1/2}\\ \sqrt{1/2} & -\sqrt{1/2}\end{bmatrix}\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix} = 1$$

$$(\tilde{x} \quad \tilde{y})\begin{bmatrix}1 & r\\ r & 1\end{bmatrix}^{-1}\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix} = 1. \tag{4}$$

The reader can verify that the correlation matrix $\begin{bmatrix}1 & r\\ r & 1\end{bmatrix}$ has eigen-values $(1+r)$ and $(1-r)$, and eigen-vectors $(\sqrt{1/2},\ \sqrt{1/2})$ and $(\sqrt{1/2},\ -\sqrt{1/2})$, respectively. The standard correlation ellipse is also given (among several equivalent expressions) by

$$(1-r^2)\tilde{x}^2 + (\tilde{y}-r\tilde{x})^2 = 1-r^2$$

or,

$$(\tilde{x}-r\tilde{y})^2 + (1-r^2)\tilde{y}^2 = 1-r^2.$$

The standard correlation ellipse has the following properties: It passes through, not just the already mentioned four points L, R, B, T, but also through other recognizable points on the $\tilde{x}$-axis, the $\tilde{y}$-axis, the major- and the minor axis such as

$$\pm(\sqrt{1-r^2},0), \pm(0,\sqrt{1-r^2}), \pm\left(\sqrt{\tfrac{1+r}{2}}, \text{sign}(r)\sqrt{\tfrac{1+r}{2}}\right), \pm\left(\sqrt{\tfrac{1-r}{2}}, -\text{sign}(r)\sqrt{\tfrac{1-r}{2}}\right).$$

Any vertical line segment terminated by the correlation ellipse is bisected by the $\hat{\tilde{y}} = r\tilde{x}$ line LR; and similarly, any horizontal line segment terminated by the correlation ellipse is bisected by the $\hat{\tilde{x}} = r\tilde{y}$ line BT. Hence, the standard correlation ellipse also passes through the following four points: $\pm(r, 2r^2 - 1)$ and $\pm(2r^2 - 1,\ r)$.

When the 1-SD square and the 1-SD Gaussian correlation ellipse of $(\tilde{x}, \tilde{y})$ are both horizontally and vertically dilated (magnified or expanded) by the same factor c we obtain the c-SD square and the c-SD Gaussian correlation ellipse of $(\tilde{x}, \tilde{y})$. To reiterate, the major axis of the c-SD correlation ellipse falls precisely on that diagonal of the c-SD square whose slope has the same sign as r. The ratio of the lengths of the two axes is $ca/[cb] = \sqrt{1+r}/\sqrt{1-r}$. Hence, the c-SD Gaussian correlation ellipse (for all c) has eccentricity

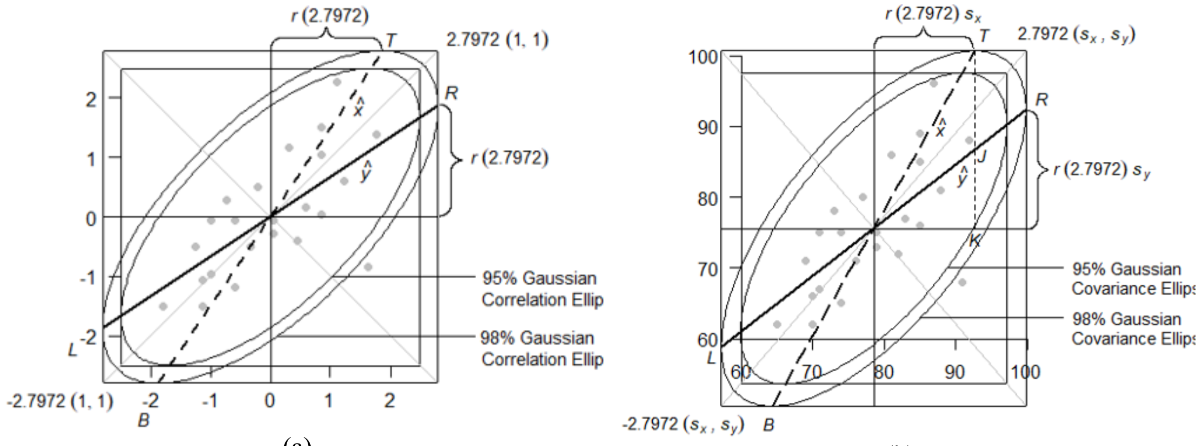
$$e = \frac{\sqrt{a^2-b^2}}{a} = \sqrt{\frac{2|r|}{1+|r|}}. \tag{5}$$

Specifically, when $r = 0$, eccentricity is 0 and the correlation ellipse is a circle; and when $|r| = 1$, eccentricity is 1 and the ellipse with a major axis of half-length $\sqrt{2}$ and a minor axis of half-length 0 collapses into a line segment of length $2\sqrt{2}$.

4. Gaussian Covariance Ellipse

In the more general case, when $s_x \neq s_y$, let us consider the shifted variables $u = x - \bar{x}$ and $v = y - \bar{y}$. Note that the mean vector for (u, v) is (0, 0), and $s_u = s_x \neq s_y = s_v$.

Starting from the c-SD square and the c-SD Gaussian correlation ellipse of $(\tilde{x}, \tilde{y})$, shown in Figure 3(a), if both are horizontally dilated by a factor $s_u = s_x$ and vertically dilated by a factor $s_v = s_y$, and then the image is translated by $(\bar{x}, \bar{y})$, the *transformed regions* are shown in Figure 3(b). We have chosen two different values, $c = 2.448$ and $c = 2.7972$, for reasons given at the end of this section. What exactly are the shapes of these transformed regions?



(a)

(b)

Figure 3: A scatter plot of midterm and final exam scores, together with (a) the c-SD square and the c-SD correlation ellipse for standardized data, and (b) the corresponding transformed regions after dilations and translation for raw data, choosing $c = 2.448$ and $c = 2.7972$. What shapes are these transformed regions?

It is trivial to see that each c-SD square in Figure 3(a) turns into a c-SD *rectangle* in Figure 3(b). But it is not easy to recognize that each c-SD correlation ellipse turns into another *ellipse*, which we shall call the c-SD covariance ellipse. Why is the dilations-translation of an ellipse another ellipse? How do the major- and the minor axes of the correlation ellipse morph into the corresponding axes of the covariance ellipse?

The answers to these questions are straightforward in the special case when $s_u = s_v = s$, say: The axes of the covariance ellipse coincide with the SD lines and their half-lengths are s-multiples of those of the correlation ellipse. To answer the questions in the more general case when $s_u \neq s_v$, we use matrix algebra. Note that after dilations, the standard correlation ellipse, given in (4), changes into

$$\begin{pmatrix} \frac{u}{s_u} & \frac{v}{s_v} \end{pmatrix} \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}^{-1} \begin{pmatrix} u/s_u \\ v/s_v \end{pmatrix} = 1$$

or equivalently,

$$(u \quad v) \begin{bmatrix} s_u^2 & r s_u s_v \\ r s_u s_v & s_v^2 \end{bmatrix}^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = 1, \tag{6}$$

which is an ellipse (called the *shifted* standard covariance ellipse, shifted because the mean is (0, 0) and standard because $c = 1$).

Let the eigen-values of the covariance matrix $S = \begin{bmatrix} s_u^2 & r s_u s_v \\ r s_u s_v & s_v^2 \end{bmatrix}$ be α and β; let the associated (orthonormal) eigen-vectors be $(e_{11}, \quad e_{12})$ and $(e_{21}, \quad e_{22})$ respectively; that is,

$$\begin{bmatrix} s_u^2 & r s_u s_v \\ r s_u s_v & s_v^2 \end{bmatrix} = \begin{bmatrix} e_{11} & e_{21} \\ e_{12} & e_{22} \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix};$$

and

$$\begin{bmatrix} e_{11} & e_{21} \\ e_{12} & e_{22} \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then, reversing the steps shown for the standard correlation ellipse, the shifted standard covariance ellipse, given in (6), becomes

$$\frac{(e_{11}u + e_{12}v)^2}{\alpha} + \frac{(e_{21}u + e_{22}v)^2}{\beta} = 1. \tag{7}$$

Returning to the original variables $(x = u + \bar{x},\ y = v + \bar{y})$ with mean vector $(\bar{x}, \quad \bar{y})$, the shifted standard correlation ellipse (7) becomes the general covariance ellipse (with arbitrary c) given by

$$\frac{(e_{11}(x-\bar{x}) + e_{12}(y-\bar{y}))^2}{\alpha} + \frac{(e_{21}(x-\bar{x}) + e_{22}(y-\bar{y}))^2}{\beta} = c^2. \tag{8}$$

Thus, matching (8) with (1), we note that the c-SD covariance ellipse has major and minor axes given by

$$\begin{aligned} &\text{Major axis: } y - \bar{y} = (e_{12}/e_{11})\ (x - \bar{x}) \text{ with half-length } c\sqrt{\alpha} \\ &\text{Minor axis: } y - \bar{y} = (e_{22}/e_{21})\ (x - \bar{x}) \text{ with half-length } c\sqrt{\beta}. \end{aligned} \tag{9}$$

To complete the discussion on the major and the minor axes, shown in (9), it remains to obtain the eigen-decomposition of the covariance matrix S. The eigen-values are the solutions τ of the quadrative equation $det\begin{bmatrix} s_x^2-\tau & rs_xs_y \\ rs_xs_y & s_y^2-\tau \end{bmatrix}=0$; or equivalently,

$$\tau^2-\left(s_x^2+s_y^2\right)\tau+s_x^2s_y^2(1-r^2)=0$$

or,

$$\tau=\frac{1}{2}\left\{\left(s_x^2+s_y^2\right)\pm\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}\right\}.$$

The larger eigen-value is $\alpha=\frac{1}{2}\left\{\left(s_x^2+s_y^2\right)+\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}\right\}$, and the eigen-vector associated with α satisfies $(s_x^2-\alpha)e_{11}+rs_xs_ye_{12}=0$, whence

$$\frac{e_{12}}{e_{11}}=\frac{\alpha-s_x^2}{rs_xs_y}=\frac{\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}-\left(s_x^2-s_y^2\right)}{2rs_xs_y}$$

which has the same sign as that of r.

Likewise, the smaller eigen-value is $\beta=\frac{1}{2}\left\{\left(s_x^2+s_y^2\right)-\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}\right\}$, and the eigen-vector associated with β satisfies $(s_x^2-\beta)e_{21}+rs_xs_ye_{22}=0$, whence

$$\frac{e_{22}}{e_{21}}=\frac{\beta-s_x^2}{rs_xs_y}=-\frac{\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}+\left(s_x^2-s_y^2\right)}{2rs_xs_y}$$

which has the opposite sign as that of r.

Moreover, instead of documenting the two eigen-vectors, it may suffice to record only the slope $m=e_{12}/e_{11}$ of the major axis (since the minor axis is orthogonal to the major axis, the slope of the minor axis is $-1/m=e_{22}/e_{21}$). Let us compare m with the ratio s_y/s_x of the two SDs in all possible cases:

1. When $r=0$, the major- and minor axes of the covariance ellipse may be chosen to coincide with the two coordinate axes. In this case, $m=0$ if $s_x>s_y$ and $m=\infty$ if $s_x<s_y$.
2. When $s_x=s_y$ and $r\neq0$, we have $m=sign(r)$; that is, the major axis falls on the SD line if $r>0$, and the minor axis falls on the SD line if $r<0$.
3. When $s_x<s_y$ and $r>0$, we have $m>s_y/s_x$; but when $s_x<s_y$ and $r<0$, we have $m<-s_y/s_x$; that is, the major axis is steeper than the SD line.
4. When $s_x>s_y$ and $r>0$, we have $m<s_y/s_x$; but when $s_x>s_y$ and $r<0$, we have $m>-s_y/s_x$; that is, the major axis is less steep than the SD line.

How should one choose the multiplier c to construct the c-SD rectangle? If one desires a fraction p of points to fall outside the covariance ellipse, one can choose c as the $(1-p)^{\text{th}}$ percentile of a chi-square distribution with two degrees of freedom, obtained from R using code: sqrt(qchisq(1–p,2). For example, 60.65% of points fall outside the 1–SD covariance ellipse; 5% of points will fall outside the 2.448–SD covariance ellipse; 1.11% outside the 3–SD covariance ellipse; 1% outside the 3.035–SD covariance ellipse. In Figure 3, we have used $c = 2.448$ and $c = 2.7972$ to flag the farthest (in 2-d sense) 5% and 2% of the scatter points. We recommend using $c = 2.8$, since this value is easy to remember and since with this choice, for a bivariate normal distribution, roughly 2% of the points fall outside the covariance ellipse.

Whereas the x-outliers and the y-outliers are already detected using the single variable c-SD line segments, the c-SD covariance ellipse is a handy tool to detect the regression outliers or bivariate outliers.

We mention a few properties of the c-SD Gaussian covariance ellipse: As mentioned before, the c-SD Gaussian covariance ellipse is internally tangent to the c-SD rectangle at four points: bottommost point $B = (\bar{x} - rcs_x, \bar{y} - cs_y)$, topmost point $T = (\bar{x} + rcs_x, \bar{y} + cs_y)$, leftmost point $L = (\bar{x} - cs_x, \bar{y} - rcs_y)$ and rightmost point $R = (\bar{x} + cs_x, \bar{y} + rcs_y)$. Moreover, LR is the $\hat{y}$ line, BT is the $\hat{x}$ line. These two regression lines LR and BT intersect at the center of the ellipse, which is also the point of intersection of the two diagonals of the c-SD rectangle and is also the mean vector $(\bar{x}, \bar{y})$. As it was for the correlation ellipse, any vertical line segment terminated by the covariance ellipse is bisected by the $\hat{y}$-line LR; and any horizontal line segment terminated by the covariance ellipse is bisected by the $\hat{x}$-line BT. Admittedly, the directions and lengths of the major- and the minor-axis of the c-SD Gaussian covariance ellipse, given in (9) and the discussion afterwards, are relatively more difficult to fathom. Nonetheless, astute students of statistics will do wisely to learn them.

5. Sufficiency

So far, we established that the c-SD rectangle and the c-SD covariance ellipse summarize all bivariate statistics mentioned in Section 2. Now we go a step further to claim that it suffices to draw only one c-SD Gaussian covariance ellipse (for any value of c) since all summary statistics can be recovered from it. How so?

Here is how: Refer to Figure 3(b) again. Given the c-SD covariance ellipse, the c-SD rectangle can be reconstructed by sandwiching the ellipse between lines parallel to the two coordinate axes. Hence, we can locate the four points of tangency B, T, L, R between the c-SD covariance ellipse and the c-SD rectangle. Then, using the points of tangency, we obtain the regression lines LR (for $\hat{y}$) and BT (for $\hat{x}$). The center of the ellipse is found either as the point of intersection between LR and BT, or the point of intersection of the two diagonals of the c-SD rectangle. The correlation coefficient r is the ratio of the horizontal distance between B and T to the horizontal distance between L and R of the c-SD rectangle (with sign positive if T is to the right of B, and negative otherwise); or equivalently, it is the ratio of the vertical distance between L and R to the vertical height BT of the c-SD rectangle (with sign positive if L is below R, and negative otherwise). The major and the minor axes are found (at least visually) as the largest and the smallest diameters (line segments passing through the center and terminated by the ellipse). If the two axes have half-lengths a and b respectively, we can also calculate $r = \frac{a^2-b^2}{a^2+b^2}$, since one can verify that $\frac{a}{b} = \sqrt{\frac{1+r}{1-r}}$. Finally, if the vertical line from T to the horizontal line $y = \bar{y}$ intersects LR (the $\hat{y}$ line) at J and ends at K, then the coefficient of determination is $r^2 = JK/TK$.

6. Further Reduction

To help the user decipher all bivariate summary statistics, we recommend superposing the entire c-SD covariance ellipse on the scatter plot. However, for mathematical completeness, we must mention that it suffices to superpose the four points of tangency B, T, L, R between the covariance ellipse and the c-SD rectangle. In fact, any three of these points will also suffice. For instance, given B, T, L, you can discover R as follows: Join BT; find its midpoint M; join LM and produce it to R such that $LM = MR$. Using these four points, we can obtain the c-SD rectangle, as explained below and shown in Figure 4.

Draw horizontal lines through B and T and vertical lines through L and R. Their points of intersection form the c-SD rectangle $EFGH$, whose vertices are labeled clockwise starting from the north-west corner. Then we have

$$s_x = \frac{1}{2c}EF,\ s_y = \frac{1}{2c}FG, \qquad \text{and} \quad r = 2\frac{ET}{EF} - 1 = 1 - 2\frac{TF}{EF}. \tag{10}$$

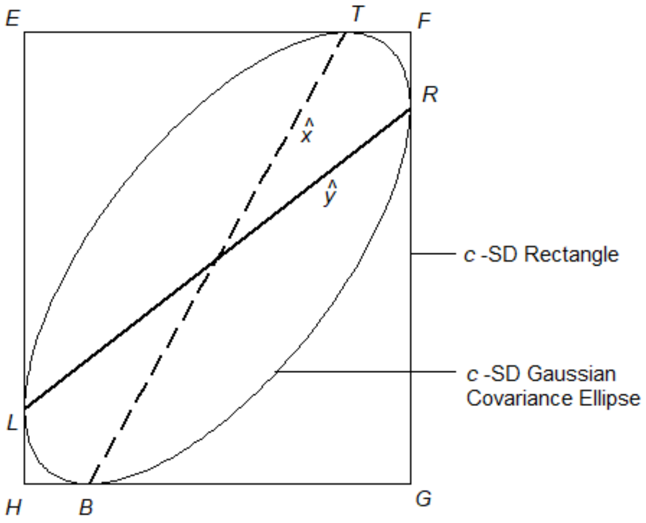


Figure 4: Any three points, out of the four points of tangency T, B, L, R, suffice to reconstruct the c-SD rectangle and the two regression lines

As already mentioned before, LR is the $\hat{y}$ line, BT is the $\hat{x}$ line. Furthermore, if we draw vertical lines through B and T, and horizontal lines through L and R, then their intersections form the inner rectangle $E'F'G'H'$ whose area as a proportion of the area of the outer rectangle $EFGH$ represents the coefficient of determination r^2. See Figure 5. In particular, as r approaches 0, the inner rectangle $E'F'G'H'$ reduces in size until it coincides with the center; and as r approaches 1, the inner rectangle $E'F'G'H'$ increases in size until it coincides with the outer rectangle $EFGH$.

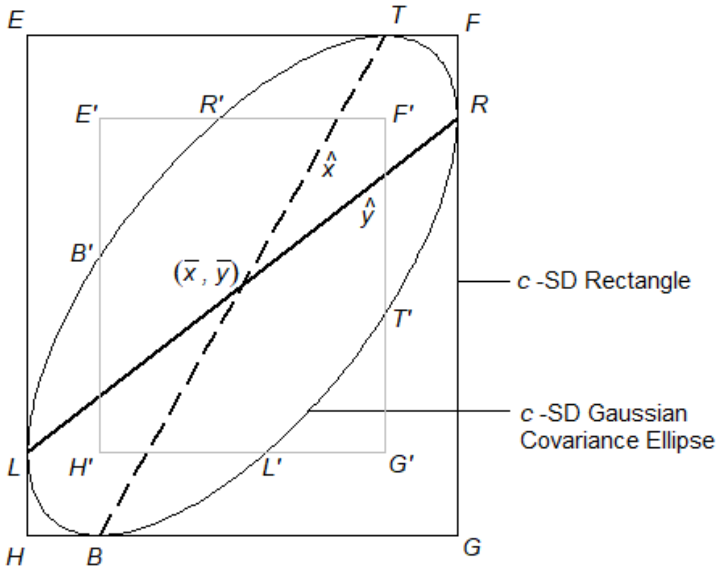


Figure 5: Any three of the four points of tangency B, T, L, R, suffice to calculate the coefficient of determination and the correlation coefficient.

Thereafter, we impose a coordinate system such that the point of intersection between BT and LR represents $(\bar{x}, \bar{y})$. Then using the SDs and the correlation given in (10) and scaling both variables in (6) by the same factor c, the c-SD covariance ellipse is given by

$$(x-\bar{x}, y-\bar{y})\begin{bmatrix} s_x^2 & rs_xs_y \\ rs_xs_y & s_y^2 \end{bmatrix}^{-1}\begin{pmatrix} x-\bar{x} \\ y-\bar{y} \end{pmatrix} = c^2. \tag{11}$$

We can draw (at least a free-hand sketch of) the c-SD covariance ellipse internally tangential to the c-SD rectangle at the four points B, T, L, R and passing through B', T', L', R' obtained by moving vertically points B, T towards the regression line $\hat{y}$ (or LR) and continuing equally far on the opposite side of LR, and similarly moving points L, R horizontally towards the regression line $\hat{x}$ (or BT) and continuing equally far on the opposite side of BT. Additional points on the ellipse are found by repeating the process.

7. Conclusion

While a box plot and a mean-SD arrow (or alternatively, a c-SD line) offer graphical summaries of one continuous variable, our proposed c-SD covariance ellipse does the same for two continuous variables. Using the c-SD covariance ellipse, we can recover the means, the SDs, the correlation coefficient r, the regression line $\hat{y}$ (as a linear function of x), the regression line $\hat{x}$ (as a linear function of y), and the coefficient of determination r^2. Thereafter, the equation of the ellipse can be recovered from (11). Moreover, scatter points outside the c-SD covariance ellipse (with a desired choice of c) are flagged as potential outliers. We hope that the c-SD covariance ellipse (or simply any three points of tangency between the c-SD covariance ellipse and the c-SD rectangle) will help users develop better intuitions about the important concepts of correlation, regression and bivariate outliers.

It is worth mentioning that the c-SD covariance ellipse, given in (11), is the shortest area region (of any shape) that captures inside it a specific fraction of the bivariate normal distribution approximately equal to the cumulative distribution function of a chi-square variable with two degrees of freedom {given by R codes: pchisq(c^2, 2) = pexp(c^2/2)}. Equivalently, the contour plots of a bivariate normal distribution form a family of ellipses for various values of c. What we have demonstrated in this paper is that given any one of these contour ellipses (even when not knowing the value of c) we can discover the mean vector, the SD line, the SD-ratio s_y/s_x, the correlation coefficient r, the two regression lines $\hat{y}, \hat{x}$, and the coefficient of determination r^2.

Here is the answer to the quiz we posed in the Section 2 regarding the largest (or the smallest) SD of a suitable projection of a scatter plot: The largest SD is attained when the scatter plot is projected on to the major axis of the c-SD covariance ellipse, for any c. Similarly, the smallest SD is attained when the scatter plot is projected on to the minor axis of the c-SD covariance ellipse.

We invite the interested reader to depict simultaneously the summary statistics involving three or more quantitative variables—specifically focusing on multiple correlation coefficient partial correlation coefficients, and principal components.

Acknowledgements

We thank our fellow faculty members and graduate students who participated in multiple pop quizzes on guessing the summary statistics based on the covariance ellipse. Special thanks are due to an anonymous reviewer for suggesting a more focused exposure of the methodology.

References

Devore, J. (2015). *Probability and Statistics for Engineering and Sciences*. Ninth Edition, Boston, MA: Brooks/Cole, Cengage Learning.

Embse, C. and Engebretsen, A. (1996). Visual representations of mean and standard deviation. *The Mathematics Teacher*, **89**(**8**), 688–692.

Maverick, L. A. (1932). Graphic presentation of standard deviation. *Journal of the American Statistical Association*, **27**(**179**), 287–297.

Rashid, M. and Sarkar, J. (2018). Cyber mentoring in an online *Introductory Statistics* course. *Educational Research Quarterly*, **41**(**3**), 25–38.

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.

Sarkar, J. and Rashid, M. (2019). Portraying standard deviation via revolution. *Journal of Probability and Statistical Science*, **17**(**1**), 109–119.

Sarkar, J. and Rashid, M. (2020). Shutter plot: A visual display of summary statistics over a scatter plot. *International Journal of Statistical Sciences*. To appear.