International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 20(2), 2020, pp 99-116

Shutter Plot: A Visual Display of Summary Statistics over a Scatter Plot

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

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

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

[*]Correspondence should be addressed to Mamunur Rashid
(Email: mrashid@depauw.edu)

[Received June 23, 2020; Accepted October 20, 2020]

Abstract

While a dot plot of one variable is naturally extended to a scatter plot of two variables, how should a box plot of one variable be extended to two variables? We propose a shutter plot that depicts the means and the standard deviations of both variables, the two regression lines and the coefficients of correlation and determination over a scatter plot. By showing all relevant summary statistics simultaneously, a shutter plot captures all aspects of a linear relationship, including flagging potential outliers, and helps the readers make good decisions.

Keywords: Scatter plot, Regression lines, Correlation coefficient, Coefficient of Determination.

AMS Classification: 62J05, 62-09.

1. Introduction

For one quantitative variable, the complete raw data are displayed without distortion in a dot plot. Oftentimes, selected summary statistics of this variable are also depicted in a condensed form: For example, the quartiles (and potential outliers) are shown in a boxplot, which sometimes is also superimposed with the mean as a fulcrum and the standard deviation (SD) as the length of a segment. We review these techniques in Section 2.

On the contrary, for two quantitative variables, the unabridged raw data are displayed in a scatter plot to suggest the relationship between the variables; and sometimes one or both regression lines are superimposed on the scatter plot; but no other summary statistics are shown in the graph! We remedy this omission by constructing a shutter plot, which superimposes on a scatter plot multiple rectangles which together depict the means and the SDs of both variables, the two regression lines and the coefficients of correlation (r) and determination (r^2). (The terminology will be explained in Section 3.) Furthermore, we identify potential x-outliers, y-outliers and residual-outliers.

2. Visualizing Univariate Data and Their Summary Statistics

When data are gathered on a quantitative variable measured on n observational units, we obtain n numbers $\{x_1, x_2, \ldots, x_n\}$. Oftentimes, the order in which the data points are collected is irrelevant; and the data may be sorted in ascending or descending order. Below is an example.

Example 1. The heights *x* of 24 football players are given (in nearest centimetre) in Table 1.

Table 1: The heights (cm) of 24 football players.

159	182	163	163	155	168	168	165	169	173	164	160	164	170
169	165	179	173	174	188	172	185	178	168				

To visualize the entire dataset without distortion one good idea is to plot the data points as dots along the number line, with multiple data points having the same value shown as vertically stacked dots. Such a diagram, shown in Figure 1(a), is called the dot plot or dotchart (also known as stripchart or stripplot in R programming language). For other kinds of dot plots, such as Cleveland dot plots and histodot plots, see Wilkinson (1999).

Furthermore, the mean of the data may be shown as a fulcrum under the number line, since Physics tells us that if data points are balls of unit mass along the weightless number line then the center of gravity of the balls is at the mean. See Devore (2015). Likewise, the SD may be shown as the length of an arrow extending from the mean to a value one SD above the mean.

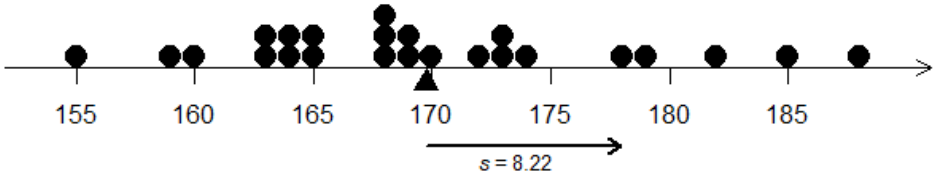


Figure 1: The unabridged data are shown as a dot plot. Additionally, the mean is shown as a fulcrum at $\bar{x} = 169.75$; and the SD $s = 8.22$ is shown as the length of a right-arrow parallel to (and at a small distance from) the number line. Also, the tail of the arrow begins at the mean.

The dots in a dot plot are often replaced by other symbols such as an 'X'. Some authors prefer to replace the dots with vertical dashes to minimize overlaps; but they still must stack the exactly tied values or add a small jitter to distinguish them. For aesthetic reasons, sometimes instead of bottom-justifying the dots (or any other symbol), they are (vertically) center-justified. Sarkar and Rashid (2016) recommends visualizing the raw data as the empirical cumulative distribution function (also called the step plot) essentially obtained by staggering the vertical dashes so that the top of a left dash and the bottom of the very next right dash are at the same height, with the leftmost dash starting at height 0 and the rightmost dash reaching height 1; and then visualize the mean as a vertical line that equalizes areas to its two sides. Also, Sarkar and Rashid (2019a) depicts the SD starting from a step plot.

We recommend displaying the entire data when the sample size is small (say, $n < 20$). For larger values of n, oftentimes for ease of comprehension, instead of (or in addition to) displaying the entire data, it is preferable to display some summary statistics. A boxplot depicts the five-number summary or the summary-5 statistics (minimum, the three quartiles Q_1, Q_2, Q_3, which are also known as the 25^{th}, 50^{th}, 75^{th} percentiles, and maximum). Figure 2(a) shows the standard boxplot produced by many statistics software. Recall that the whiskers extend on each side of the box for up to 1.5 times the inter-quartile range $Q_3 - Q_1$; and any values beyond the whiskers are flagged as potential outliers (shown by the symbol X).

To efficiently depict the summary statistics of bivariate data, we make the following three modifications to the standard boxplot of Figure 2(a) and construct Figure 2(b):

1) While the length of the box denotes the inter-quartile range, the width of the box is totally non-informative. Hence, we eliminate the inter-quartile-box in favor of a solid line segment flanked by open and close parentheses to denote the first and the third quartiles respectively. This simplified diagram we call the five-number line.
2) We superimpose a thick right-arrow on top of the five-number line extending from the mean at the tail end to the mean plus one SD at the arrowhead. This augmented diagram we call the seven-number line, since it depicts the summary-7 statistics (that is, the summary-5 statistics, the mean and the SD).
3) We print the sample size to the right of the seven-number line. The sample size is often omitted during summary, causing readers to give an undue credence to a small dataset or to gain a false aura of comparability when in fact the sample sizes of two datasets are vastly different. We strongly urge all users to declare the sample size of each summarized dataset, especially when multiple seven-number lines are displayed one the same graph.

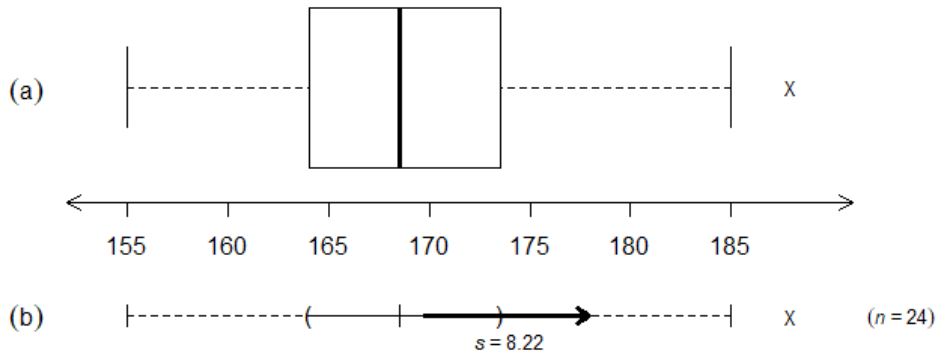


Figure 2: (a) A standard boxplot of the data in Example 1 shows the summary-5 statistics; but the width of the box is non-informative. **(b)** The seven-number line depicts the summary-7 statistics consisting of the summary-5 statistics, the mean (at the tail of the thick arrow) and the standard deviation (as the length of the arrow). Both diagrams flag potential outlier(s) by the symbol X.

The primary objective of this paper is to develop a graphical display of summary statistics for two quantitative variables whose interrelation is being studied. While each variable can be studied separately using the univariate methods mentioned

above, we focus on depicting the summary statistics computed using the two variables simultaneously—statistics such as the two regression lines and the coefficients of correlation (r) and determination (r^2). Furthermore, we identify potential x-outliers, y-outliers and residual-outliers.

3. Visualizing Bivariate Data and Their Summary Statistics

When data are gathered on two quantitative variables measured on each of n observational units, we end up with n pairs of numbers $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Again, the order in which the data points are collected may be irrelevant; and the data may be sorted in ascending or descending order with respect to either variable x or variable y. The pairs cannot be sorted simultaneously as there is no natural bivariate ordering. Below is an example.

Example 2. (Example 1 continued) The heights x (in nearest centimeter) and weight y (in nearest kilogram) of 24 football players are given in Table 2.

Table 2: The heights (cm) and weights (kg) of 24 football players.

Height	159	182	163	163	155	168	168	165	169	173	164	160	164
Weight	57	75	65	68	63	67	70	75	71	70	65	65	69
Height	170	169	165	179	173	174	188	172	185	178	168		
Weight	69	65	67	74	70	72	67	70	78	72	69		

To visualize the entire dataset without distortion, we customarily look at the scatter plot, as in Figure 3(a) which shows height in the horizontal direction, weight in the vertical direction and each datum as a dot in the coordinate plane. By default, multiple data points having the same pair of values are superimposed, so that the viewer cannot tell how many data points each dot represents. To overcome this shortcoming, most software can report the frequency of each dot when an appropriate input is given through a command or a menu. For our

purpose, such refinement is unnecessary. Figure 3(a) depicts the bivariate data of Example 2; but it does not depict any summary statistics.

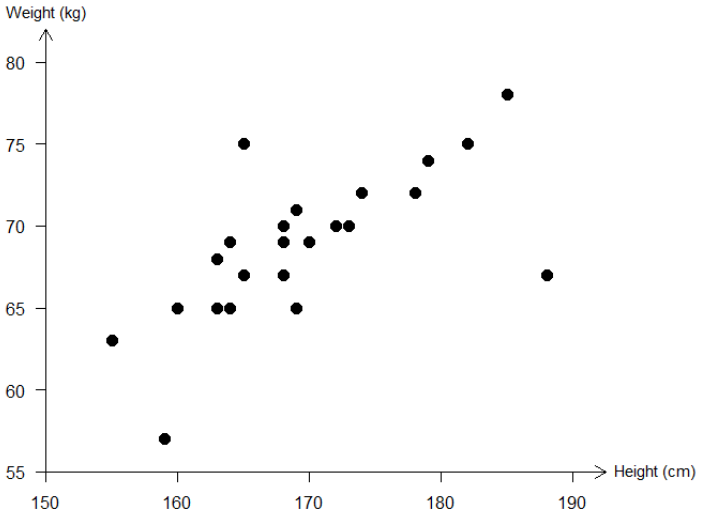
Sometimes the scatter plot is superimposed with the least-squares regression line (of y on x) given by

$$\hat{y} = a + bx = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x}) \tag{1}$$

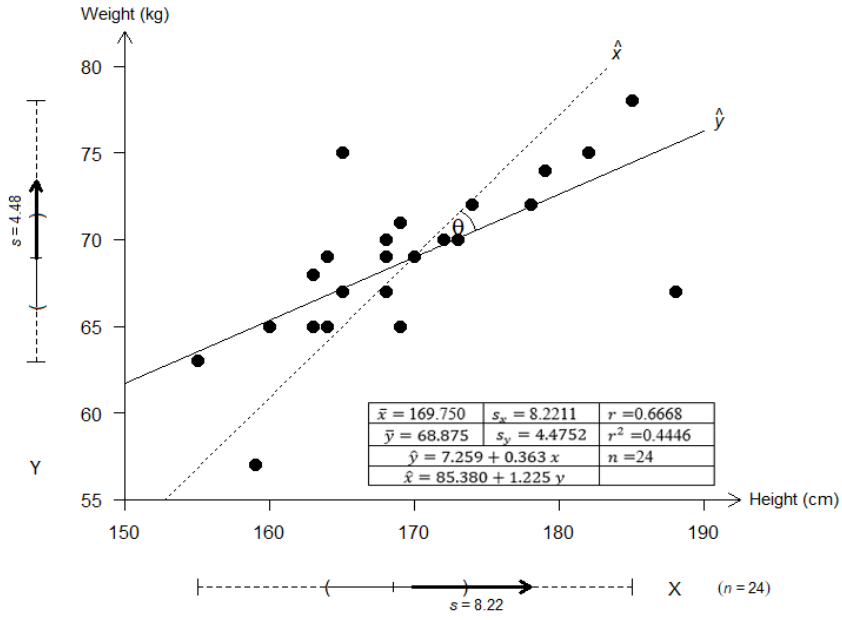
which, of course, passes through the mean vector $(\bar{x}, \bar{y})$; but this vector may or may not be identified on the diagram. On rare occasions, when the other least-squares regression line (of x on y) given by

$$\hat{x} = c + dy = \bar{x} + r\frac{s_x}{s_y}(y - \bar{y}) \tag{2}$$

is also depicted, we can figure out the mean vector as the point of intersection of the two regression lines. As in the case of a dot plot, one can imagine that if on a weightless, rigid plane, metal balls of unit weight are placed at the scatter points, then a fulcum located exactly under the mean vector $I = (\bar{x}, \bar{y})$ will keep the system in balance. The other bivariate summary statistics are usually not depicted; only their numerical values are printed either in an inset or in the caption or in the accompanying text. See, for example, Figure 3(b) for such an embellished scatter diagram, where we have also shown the univariate seven-number summary lines, showing the summary-7 statistics, for variable height at the bottom margin and for variable weight at the left margin of the scatter plot. The outliers are marked with X and Y respectively.



(a)



(b)

Figure 3: Height and weight of football players are positively correlated as seen in (a) a bare bone scatter plot; and (b) an embellished scatter plot showing both least-squares regression lines and listing all bivariate summary statistics in an inset.

We have augmented the latter diagram with univariate summary-7 statistics at the bottom margin and the left margin respectively, using which two outliers are flagged—one high outlier with respect to height (marked as X) and a different low outlier with respect to weight (marked as Y). One can locate the corresponding points in the scatter diagram. The regression outlier(s) are not flagged.

How can we visualize most of these univariate and bivariate summary statistics printed in the inset of Figure 3(b)? The answer to this question is the main contribution of this paper.

In Figure 4, we depict all relevant bivariate summary statistics needed in the study of correlation and regression. We call it a shutter plot because it superimposes on the scatter plot additional rectangles: Just as a virtual rectangular box guides the photographer focus on stress-worthy items during a photo session, a shutter plot focuses the reader's attention to the two sets of univariate summary-7 statistics, the two regression lines and coefficients of correlation and determination. For the time being, let us assume $r > 0$; we will visit the other case afterwards. The following steps are needed:

1) Starting from Figure 3(a), let us draw the univariate seven-number summary lines—not in the margins, but parallel to the respective axes—so that they intersect at $I = (\bar{x}, \bar{y})$, the bivariate mean. In the north-east quarter with respect to the relocated seven-number lines, we draw a rectangle that measures s_x in the horizontal direction and s_y in the vertical direction; and call it the bivariate-SD-rectangle, and its diagonal through the bivariate mean I we call the bivariate-SD-line (dotted). Once the bivariate-SD-rectangle is drawn, there is no need for the univariate SD-arrows; and as such we eliminate them.
2) Next, we split the bivariate-SD-rectangle by a horizontal line at a distance $r \cdot s_y$ from the bottom edge; and call the lower part the $\hat{y}$-rectangle, and its diagonal through the bivariate mean I is indeed the least squares regression line of y on x or simply the $\hat{y}$-line (solid line given by (1)). Thus, r is the ratio of heights of $\hat{y}$-rectangle and bivariate-SD-rectangle.
3) If through the point where the upper edge of the $\hat{y}$-rectangle intersects the bivariate-SD-line we draw a vertical line, then we split the bivariate-SD-rectangle a second time by a vertical line at a distance $r \cdot s_x$ from the left edge; and call the left part the $\hat{x}$-rectangle, and its diagonal through the

bivariate mean is indeed the least squares regression line of x on y (dashed line given by (2)) or simply the $\hat{x}$-line. Thus, r is also the ratio of widths of $\hat{x}$-rectangle and bivariate-SD-rectangle.

4) The overlap between the $\hat{x}$-rectangle and the $\hat{y}$-rectangle (which are both bottom aligned and left aligned) is, of course, another rectangle, which we shade and call the determination-rectangle, since its area as a fraction of the area of the SD-rectangle is indeed the coefficient of determination r^2.
5) Although the sample size n can be recovered by counting the dots in the scatter plot (together with their frequencies, if given), to aid the viewer we highly recommend printing the sample size (to the right of the SD-rectangle).
6) The x-outliers and the y-outliers are already flagged using univariate boxplot method. Alternatively, we may superimpose dotted vertical lines $\bar{x} \mp 2s_x$ and declare as x-outliers points (marked as X) falling outside the vertical strip between these lines. (Here, the multiplier 2 is chosen arbitrarily: One may replace it with 2.75 or 3 or any other multiplier.) Similarly, we may superimpose dotted horizontal lines $\bar{y} \mp 2s_y$ and declare as y-outliers points (marked as Y) falling outside the horizontal strip between these lines. To keep the picture uncluttered, such boundary lines need not be shown.
7) Furthermore, we may draw dotted curves (as functions of x) given by

$$\bar{y} + r\frac{s_y}{s_x}(x - \bar{x}) \mp t^* \sqrt{1 + \frac{1}{n} + \frac{\left(\frac{x-\bar{x}}{s_x}\right)^2}{n-1}} \cdot \sqrt{1 - r^2} \cdot s_y \qquad (3)$$

where t^* is a critical value from a t-distribution with $(n-2)$ degrees of freedom such that the tail area is 2.5% (or 1%, or any other small percentage). Note that these curves are above and below the $\hat{y}$-line at a *vertical* distance increasing with respect to the distance of x from $\bar{x}$. These curves are called prediction bands. See, for example, Wackerly, Mendenhall and Schaffer (2008). Any scatter point falling outside these prediction bands is a potential regression outlier (marked as R). Again, to keep the shutter plot uncluttered, such boundary curves need not be shown.

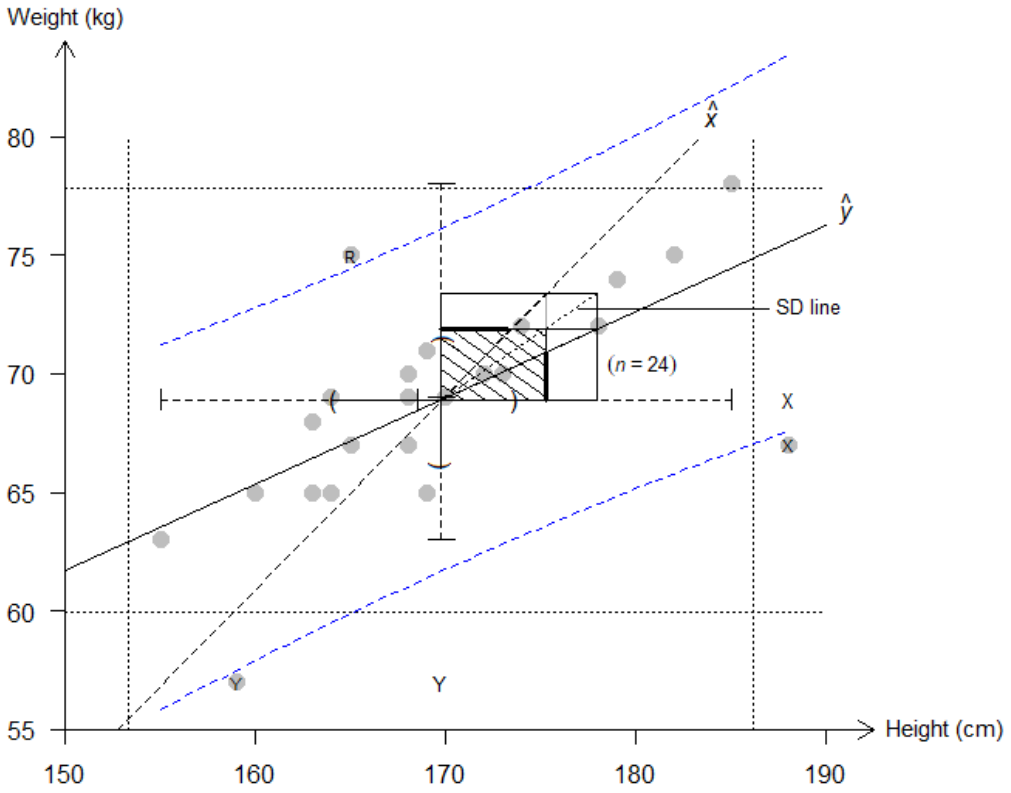


Figure 4: Height and weight of football players are positively correlated with $r = 0.6668$, which is both the ratio of heights of the $\hat{y}$-rectangle and the bivariate-SD-rectangle and the ratio of widths of the $\hat{x}$-rectangle and the bivariate-SD-rectangle. Also shown are the two regression lines, which are diagonals through $I = (\bar{x}, \bar{y})$ of the $\hat{y}$-rectangle and the $\hat{x}$-rectangle respectively. The coefficient of determination r^2 is shown as the ratio of the area of the determination rectangle (shaded) to the area of the bivariate-SD-rectangle. Potential x-outliers, y-outliers and regression outliers are marked with X, Y and R symbols.

The purpose of Figure 4 is to depict the summary statistics without actually printing them. For the benefit of the readers who wish to verify these quantities, as shown in the legend of Figure 3(b), we list them here again: $n = 24;\ \bar{x} = 169.750;\ s_x = 8.2211, \bar{y} = 68.875;\ s_y = 4.4752; r = 0.6668;\ r^2 = 0.4446;\ \hat{y} = 7.259 + 0.363x;\ \hat{x} = 85.380 + 1.225y.$

So far, we have assumed that $r > 0$. If, on the other hand, $r < 0$, then we draw the above-mentioned rectangles and diagonals in the south-east (or the north-west) quarter with respect to the relocated seven-number lines. The rest of the displays are exactly like the previous case. See Example 3 and Figure 5.

Example 3. Fifty-two members of IU Health Plan registered for a weight loss program for a six-week period. Their average daily calorie intakes and the net weight loss by the end of the program were tabulated. The entire bivariate data and all univariate and bivariate summary statistics are depicted in Figure 5. As such, the actual data table is no longer necessary.

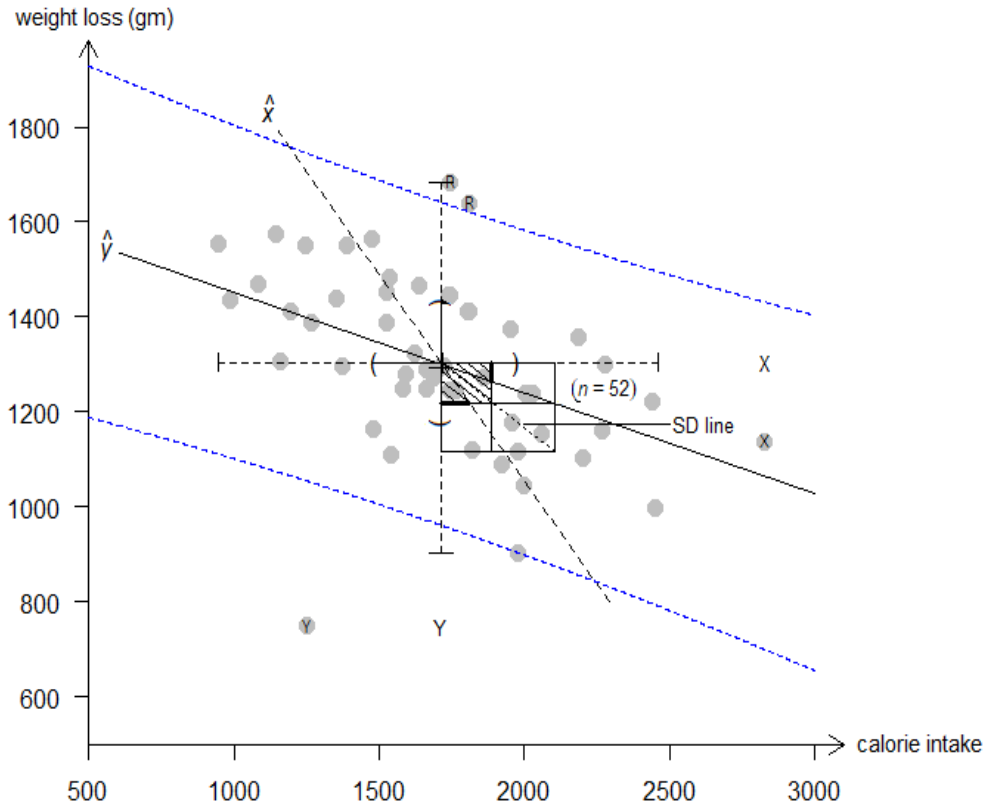


Figure 5: Average daily calorie intake and weight loss are negatively correlated with $r = -0.4470$, which is both the ratio of depths of the $\hat{y}$-rectangle and the bivariate-SD-rectangle and the ratio of widths of the $\hat{x}$-rectangle and the bivariate-

SD-rectangle. Also shown are the two regression lines, which are diagonals through $I = (\bar{x}, \bar{y})$ of the $\hat{y}$-rectangle and the $\hat{x}$-rectangle respectively. The coefficient of determination r^2 is shown as the ratio of the area of the determination rectangle (shaded) to the area of the bivariate-SD-rectangle. Potential x-outliers, y-outliers and regression outliers are marked with X, Y and R symbols.

4. Conclusion

Our goal has been to visually display all univariate and bivariate summary statistics used in studying interrelation between two quantitative variables. To do so efficiently, we have modified the traditional boxplot to a seven-number line. Thereafter, we constructed the shutter plot, which superimposes on the scatter plot a second picture containing several rectangles and their diagonals. We can read off all bivariate summary statistics from the shutter plot.

The following properties of the shutter plot are noteworthy:

1) The diagonal of the bivariate-SD-rectangle (dotted line) is also the diagonal of the determination-rectangle; and the ratio of the latter to the former also represents r.
2) The coefficient of determination r^2 can be viewed alternatively as the width of the horizontal thick line segment (intercepted by the vertical line through I and the $\hat{x}$-line) along the top border of the $\hat{y}$-rectangle relative to the width of the bivariate-SD-rectangle, and also as the height of the vertical thick line segment (intercepted by the horizontal line through I and the $\hat{y}$-line) along the right border of the $\hat{x}$-rectangle relative to the height of the bivariate-SD-rectangle. See Figures 4 and 5.
3) If we write $\tau = s_y/s_x$ as the slope of the bivariate-SD-line, and we write θ as the angle between the two regression lines $\hat{y}$ and $\hat{x}$ (see Figure 4), then

$$\tan\theta = \frac{r^{-1} - r}{\tau^{-1} + \tau} \tag{4}$$

The proof uses the trigonometric identity for the tangent of the sum of two angles and some algebraic manipulations; and we leave it to the interested reader.

In the extreme cases, when $r = \pm 1$, we have $\theta = 0$ and the two regressions lines are identical; and in the uncorrelated case, when $r = 0$, we have $\theta = \pi/2$ and the two regressions lines are respectively horizontal and vertical. Furthermore, when $s_y = s_x$, as is the case when each variable is replaced by its standardized version, we have $\tau = 1$; consequently, $\tan\theta = (1 - r^2)/(2r)$.

4) From the traditional display of both regression lines on a scatter diagram, as in Figure 3(b), without even knowing the equations of the regression lines, we can decipher r^2, r and $\tau = s_y/s_x$ as follows: From $I(\bar{x}, \bar{y})$, the intersection of the two regression lines, draw a horizontal segment IH (of an arbitrary length); then through H draw a vertical line cutting the $\hat{y}$-line and the $\hat{x}$-line at E and F respectively. Then $r^2 = HE/HF$. Next, take the geometric mean of HE and HF, say HG. See, for example, Sarkar and Rashid (2019b) for how to draw a geometric mean of two line-segments. Then $r = HE/HG$ and $\tau = HG/HI$. See Figure 6. However, it is not possible to decipher the individual standard deviations s_x and s_y, since one can arbitrarily magnify either or both variables without changing r.

We strongly advocate keeping the scatter plot as an integral part of the shutter plot because many different bivariate data sets may give rise to the exact same bivariate summary statistics. See Anscombe (1973), Chatterjee and Firat (2007), and Matejka and Fitzmaurice (2017).

Finally, the notions developed here can be extended to visualize the correlation between two random variables—either discrete or continuous—and the least squares regression lines of each variable on the other.

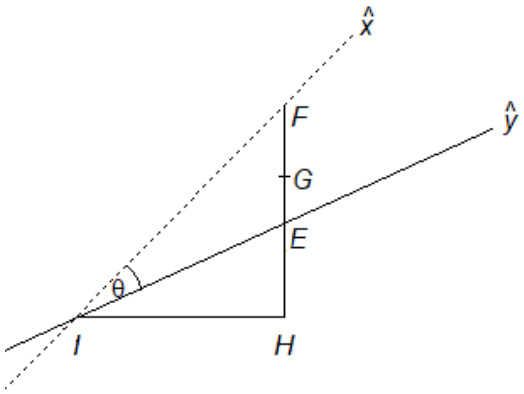


Figure 6: Given only the two regression lines intersecting at $I(\bar{x}, \bar{y})$, how can one find r^2, r and $\tau = s_y/s_x$ without using the equations of the regression lines or even the scales of the two axes? Draw a horizontal segment IH of arbitrary length; then draw its perpendicular intersecting the regression lines $\hat{y}$ and $\hat{x}$ at E and F respectively. Then $r^2 = HE/HF$. Next, draw HG, the geometric mean of HE and HF. Then $r = HE/HG = HG/HF$ and $\tau = HG/HI$.

Acknowledgments: We thank our students and colleagues for partaking in a pop quiz where we asked them the question posed in Figure 6. We also thank the referee for correcting errors, filling-in omissions and suggesting to include the R codes.

Reference

[1] Anscombe, F.J. (1973). Graphs in Statistical analysis. The American Statistician, 27(1): 17-21.

[2] Chatterjee, S. and Firat, A. (2007). Generating data with identical statistics but dissimilar graphics: A follow up to the Anscombe dataset. The American Statistician, 61(3): 248–254.

[3] Devore, J. (2015). Probability and Statistics for Engineering and Sciences (9^{th}edn). Boston, MA: Brooks/Cole, Cengage Learning.

[4] Matejka, J. and Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems: 1290–1294.

[5] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[6] 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.

[7] Sarkar, J. and Rashid, M. (2019a). Have you seen the standard deviation?", Nepalese Journal of Statistics, 3, 1-10, 2019.

[8] Sarkar, J. and Rashid, M. (2019b), "Euclidean plane geometry suffices to visualize the standard deviation," The Mathematics Student, 88(3-4), 2019.

[9] Wackerly, D.D., Mendenhall, W. and Scheaffer, R.L. (2008).Mathematical Statistics with Applications.7th ed. Brookes/Cole, Cengage Learning, Belmont, CA.

[10] Wilkinson, L. (1999). Dot Plot. The American Statistician, 53(3), 276-281.

Appendix

R Codes for producing Shutter plot (Figure 4)

```
dev.new(width=12,height=10)

ht=c(159,182,163,163,155,168,168,165,169,173,164,160,164,170,169,1
     65,179,173,174,188,172,185,178,168)

wt=c( 57, 75, 65, 68, 63, 67, 70, 75, 71, 70, 65, 65, 69,  69, 65,
     67, 74, 70, 72, 67, 70, 78, 72, 69)

# Scatter plot

par(mai=c(1.6,1.6,1,1.2))

x1<-min(ht);x2<-max(ht);y1<-min(wt);y2<-min(wt)

plot(ht,wt,xlim=c(150,190),ylim=c(55,80),xaxs="i", yaxs="i",

frame.plot=FALSE,xlab="",ylab="",las=1,pch=19,col="grey",cex=1.5,x
     pd=TRUE)

arrows(150,55,192.5,55,code = 2, xpd = TRUE, length=.12,xpd=TRUE)

arrows(150,55,150,84,code = 2, xpd = TRUE, length=.12,xpd=TRUE)

text(196.2,55, expression("Height (cm)"),xpd=TRUE,cex=.9,xpd=TRUE)

text(150,85,expression("Weight (kg)"),xpd=TRUE,cex=.9)

rect(mean(ht),mean(wt),mean(ht)+sd(ht),mean(wt)+sd(wt))

segments(mean(ht),mean(wt),mean(ht)+sd(ht),mean(wt)+sd(wt),lty=3)

segments(mean(ht),mean(wt)+cor(ht,wt)*sd(wt),mean(ht)+sd(ht),mean(
     wt)+cor(ht,wt)*sd(wt))

segments(mean(ht),mean(wt),mean(ht)+sd(ht),mean(wt)+cor(ht,wt)*sd(
     wt))

reg1<-lm(wt~ht)

abline(reg1)

text(191,77,expression(hat(italic(y))),cex=1,xpd=T)

segments(mean(ht)+cor(ht,wt)*sd(ht),mean(wt),mean(ht)+cor(ht,wt)*s
     d(ht),mean(wt)+sd(wt),col="darkgrey")

segments(mean(ht),mean(wt),mean(ht)+cor(ht,wt)*sd(ht),mean(wt)+sd(
     wt),lty=2)

segments(mean(ht),mean(wt),mean(ht)+cor(ht,wt)*sd(ht),mean(wt)+sd(
     wt),lty=2)
```

```
segments(152.755,55,183.38,80,lty=2)

text(184,81,expression(hat(italic(x))),cex=1,xpd=T)

segments(mean(ht),mean(wt)+cor(ht,wt)*sd(wt),       mean(ht)+3.48,
     mean(wt)+cor(ht,wt)*sd(wt),lwd=3)

segments(mean(ht)+cor(ht,wt)*sd(ht),mean(wt),mean(ht)+cor(ht,wt)*s
     d(ht),mean(wt)+2,lwd=3,xpd=TRUE)

rect(mean(ht),mean(wt),mean(ht)+cor(ht,wt)*sd(ht),mean(wt)+cor(ht,
     wt)*sd(wt),density=12,angle=145)

segments(177,72.7, 187,72.7)

text(189.1,72.9, expression("SD line"),xpd=TRUE,cex=.8)

text(184,72.1,expression((italic(n)~'='~24)),xpd=TRUE,cex=.8)

abline(v=mean(ht)-2*sd(ht),lty=3)

abline(v=mean(ht)+2*sd(ht),lty=3)

text(188,67,expression("X"),xpd=TRUE,cex=.7)

abline(h=mean(wt)-2*sd(wt),lty=3)

abline(h=mean(wt)+2*sd(wt),lty=3)

text(159,57,expression("Y"),xpd=TRUE,cex=.7)

summary(ht)

newx<- seq(155, 188, by=.5)

pred_interval<-     predict(reg1,     newdata=data.frame(ht=newx),
     interval="prediction",  level = 0.95)

lines(newx, pred_interval[,2], col="blue", lty=2,xpd=T)

lines(newx, pred_interval[,3], col="blue", lty=2,xpd=T)

length(newx);length(pred_interval[,2])

text(165,75,expression("R"),xpd=TRUE,cex=.7)

# BOTTOM Quartile plot

par(mai=c(2.05,2.08,1,1.28)) # bottom, left, top, right

par(new=TRUE)

boxplot(ht,horizontal=TRUE,las=1,frame.plot=FALSE,boxwex=0,

axes=FALSE,staplewex = 9,col="black",outpch = "X", outcex =.8)

boxplot.stats(ht)
```

```
segments(boxplot.stats(ht)$stats[1],.98,boxplot.stats(ht)$stats[1]
     ,1.02,xpd=TRUE)

text(boxplot.stats(ht)$stats[2],1,expression("("),xpd=TRUE)

segments(boxplot.stats(ht)$stats[3],.98,boxplot.stats(ht)$stats[3]
     ,1.02,xpd=TRUE)

text(boxplot.stats(ht)$stats[4],1,expression(")"),xpd=TRUE)

segments(boxplot.stats(ht)$stats[5],.98,boxplot.stats(ht)$stats[5]
     ,1.02,xpd=TRUE)

# LEFT Quartile plot

par(mai=c(1.78,1.70,1.18,1.36)) # bottom, left, top, right

par(new=TRUE)

boxplot(wt,las=1,frame.plot=FALSE,boxwex=0,staplewex           =9,
     axes=FALSE,

col="black",outpch = "Y", outcex = .8)

boxplot.stats(wt)

segments(.98,boxplot.stats(wt)$stats[1],1.02,boxplot.stats(wt)$sta
     ts[1],xpd=TRUE)

segments(.995,boxplot.stats(wt)$stats[2],1.005,boxplot.stats(wt)$s
     tats[2],xpd=TRUE)

segments(.98,boxplot.stats(wt)$stats[3],1.02,boxplot.stats(wt)$sta
     ts[3],xpd=TRUE)

segments(.995,boxplot.stats(wt)$stats[4],1.005,boxplot.stats(wt)$s
     tats[4],xpd=TRUE)

segments(.98,boxplot.stats(wt)$stats[5],1.02,boxplot.stats(wt)$sta
     ts[5],xpd=TRUE)
```