IVY Plots and Gaussian Interval Plots

Jyotirmoy Sarkar
Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, Indiana, USA; e-mail: jsarkar@iupui.edu

Mamunur Rashid
Department of Mathematics, DePauw University, Indiana, USA; e-mail: mrashid@depauw.edu

Summary
While a dot plot depicts data on a quantitative variable without distortion, a boxplot shows only the five-number summary. For large data, to aid in counting, we propose an IVY plot as a companion to a dot plot. Also, for large data, if the variable is approximately normally distributed, as a companion to a box plot, we propose a Gaussian Interval plot that depicts the five-number summary, the mean, the SD, the sample size, and the counts of outliers. We hope these enhanced visualizations will add value to the commonly used methods.

Keywords: Teaching statistics; Five-number summary; Seven-number summary; Gaussian intervals

INTRODUCTION
To comprehend a single quantitative variable, one can use a stem-and-leaf plot, or a dot plot with little or no distortion of information. Sometimes to focus on the overall distribution, one uses a histogram which shows the number (or proportion) of values within contiguous subintervals, but the exact values are lost. Thereafter, eliminating a lot of details, but keeping a few essential features, a boxplot (Spear, 1952, 1969) displays only the five-number summary, and identifies potential outliers. Sometimes it is also called the box-and-whiskers plot, which was introduced by John Tukey (1977) as part of his toolkit for exploratory data analysis; and it is especially useful for comparing several subgroups.

Some important statistics, such as the mean and the standard deviation (SD) are not shown in a standard boxplot. These are sometimes superimposed on a boxplot in the form of an arrow whose tail locates the mean and length represents the SD (see, for example, Sarkar and Rashid (2019)), thereby depicting the seven-number summary. Sarkar and Rashid (2015, 2016) present alternative ways to visualize the mean and the SD using the empirical cumulative distribution function (also called the step plot).

This is the author's manuscript of the article published in final edited form as:

The purpose of this paper is twofold: For a large data set, (1) we propose an IVY plot as a companion to a dot plot; and (2) as a companion to a boxplot, we propose a Gaussian Interval (GI) plot which displays the seven-number-summary, the sample size, and potential outliers, including their counts. To compare several subgroups, one may draw several GI plots on the same scale. Whereas a boxplot is a nonparametric method applicable to any numerical variable, when the variable is approximately normally distributed, a GI plot is more powerful in identifying potential outliers.

We have used the freeware R to draw all graphs in this paper. Packages are being developed and will be available to all users in due time.

**DOT PLOT AND IVY PLOT**

Although an IVY plot is more appropriate for moderate to large data sets, we begin with a small data set simply to illustrate how to draw an IVY plot.

**EXAMPLE 1**

Suppose that 30 students in a *Regression Analysis* course earned homework scores (out of 50 points) as summarized in the following table.

<table>
<thead>
<tr>
<th>Table 1. Homework scores (sorted) of 30 students</th>
</tr>
</thead>
</table>

To understand the distribution of a raw data set, typically we draw a dot corresponding to each given value at its numerical location on a horizontal number line drawn to scale. If the same value occurs more than once, then the dots are stacked vertically. Such a diagram is called a dot plot or a dotchart (or a stripchart or a stripplot in the R programming language). For other kinds of dot plots, such as Cleveland dot plots and histodot plots, see Wilkinson (1999).

The sample mean $\bar{x}$ is a measure of the central location of a variable. Sometimes, borrowing from *Physics*, the mean is shown as a fulcrum under the dot plot denoting a pivot that balances a weightless number line on which each ball of unit mass exerts a downward pressure. Such a diagram explains why the mean is sensitive to a change in one or more numbers.

A widely used measure of spread is the standard deviation (SD) $s$, or the root ‘mean’ squared deviation from the mean, which indicates *on average* how far each value deviates from the mean. Sarkar and Rashid (2019) exhibited the sample SD by drawing an arrow below the number line, with the tail of the arrow exactly at the fulcrum and the arrowhead extending one SD to the right of the fulcrum. See figure 1.
Fig. 1. Under a dot plot the mean $\bar{x} = 41.8$ is shown as both a fulcrum and the tail of an arrow; the SD $s = 3$ is shown as the length of the arrow.

For a large data set, the dot plot suffers from two drawbacks: (1) the dots take up too much space; (2) the dots must be carefully counted to find out how many numbers are tied at each value. To overcome these drawbacks, we propose an IVY plot: It uses left-, right- and vertical notches for each observation, and up to five notches are joined to form a schematic diagram of one pinnate compound leaf of an ivy creeper. Thereafter, the notches start afresh, and the leaves (and perhaps a partial leaf with one to four leaflets) are vertically stacked. To aid in counting the leaves, a small extra vertical space may be added after every fifth leaf (or 25 leaflets). Furthermore, the fulcrum at the mean is augmented by a vertical segment whose height is proportional to the sample size $n$, which is also printed next to it. See figure 2.

The name “IVY” comes from the use of the three upper-case letters I, V, Y having 1, 2, 2 top endpoints, respectively, indicating the frequency of tied values. Higher frequencies are formed by stacking these letters vertically as shown in the legend of figure 2. Moreover, the notation for frequency 5 resembles the five leaflets of an ivy leaf.

| =Frequency 1,  =Frequency 2,  =Frequency 3,  =Frequency 4, and  =Frequency 5 |

Fig 2. (a) In an IVY plot, observations are represented by left-, right- and vertical notches to save space and to expedite counting. (b) An embellished IVY plot also shows the mean, the standard deviation, and the sample size. The mean $\bar{x} = 41.8$ is shown in three ways: as a fulcrum, as a vertical segment, and as the tail of an arrow. The SD $s = 3$ is shown as the length of the arrow. The height of the vertical segment at the mean is proportional to the sample size $n$. 

3
Remark 1: We depict in figure 2 an IVY plot for a small data set simply to develop the notion and to introduce the frequency symbols (explained in the legend). But we do not recommend using such an IVY plot because, for a small data set, an IVY plot distorts the proportionality of frequencies at the different values. On the other hand, for a large dataset as in Example 2, where the impact of distortion will be minimal, we strongly recommend using an IVY plot.

An IVY plot is more effective for a moderate to large data set because of the ease of counting. In contrast, tallying the dots from a dot plot requires careful counting and is prone to mistakes. Likewise, from a frequency histogram, while one may read off the frequencies of the bins from the vertical scale, one cannot retrieve the actual values within the bins. Moreover, the distribution depicted by a histogram depends very much on the choice of the bin width. Thus, an IVY plot merges the advantages of a dot plot and a histogram, while simultaneously avoids their drawbacks: It preserves the exact values and assists in counting the frequencies. Example 2 illustrates such advantages of an IVY plot for a large data set.

EXAMPLE 2

This data set consists of the overall GPA (grade point average) scores (correct to the nearest tenth) of 257 high school students in the Class of 2020 in Putnam county. The county has decided to issue graduation diplomas to students with GPA at least 2.0 and award a laptop computer to every student with a perfect GPA of 4.0.

How many students failed to earn a graduation diploma and how many students were awarded a laptop? The answers (12 and 21) can be read off with some effort from the dot plot (by counting the dots carefully); but the same information is obtained almost effortlessly from the IVY plot. Furthermore, we constructed two histograms with different bin widths (0.2 and 0.4) in figures 3(c) and 3(d), from which we cannot answer the above questions. While we can extract the frequencies of the bins by referring to the vertical axis, we cannot retrieve the exact values within the bins or their respective frequencies. The distributions depicted by figures 3(a) and 3(b) are identical; the distribution depicted by figure 3(c) is reasonably close to them; but the distribution depicted by figure 3(d), though it chooses the bin width according to the commonly adopted Sturge’s rule, is quite different—nay, even misleading. This is a shortcoming of a standard histogram.
**BOXPLOT AND GI PLOT**

The standard boxplot depicts only five percentiles: It shows a box stretching from the first quartile $Q_1$ to the third quartile $Q_3$, together with a vertical line at the median or the second quartile $Q_2$. The box plot also has two whiskers which are supposed to extend left from the first quartile $Q_1$, and right from the third quartile $Q_3$, for up to one-and-a-half times the interquartile range $Q_3 - Q_1$. All values farther away than the whiskers’ intended expanse are flagged as potential outliers (marked by the symbol *); and the whiskers are shortened to reach up to the most extreme values within their intended expanse. Thus, the minimum, as well as the maximum, is either flagged as an outlier or marked as an end point of a whisker. Points that are more than three times the interquartile range $Q_3 - Q_1$ away from the boundaries of the box may be flagged as extreme outliers (marked by a different symbol such as +). See figure 4(a), which has no lower outlier, one upper outlier, and no extreme outlier.

Whereas one can draw a boxplot after knowing only five percentiles, any rearrangement of other values (without changing these five percentiles) has no impact on the boxplot even though the
distribution may change drastically! This is a drawback of the standard boxplot, which can be mitigated by superimposing on the boxplot a fulcrum at the mean and a right arrow spanning the interval \((\bar{x}, \bar{x} + s)\). Such an enhanced boxplot depicts the seven-number summary (five percentiles, the mean, and the SD). See figure 4(b).

![Fig. 4. (a) A standard boxplot shows the five-number summary; and (b) an enhanced boxplot shows the seven-number summary. Both plots flag potential outlier(s)](image)

Moreover, a standard boxplot, though it flags the outlier values, does not report the frequencies of such values. Therefore, we are motivated to construct a Gaussian Interval (GI) plot which will show the frequencies of the outliers.

**Construction of a GI Plot**

In the special case when the variable is roughly normally distributed, one may construct the \(c\)-SD GI around the mean and flag any value outside this interval as a potential outlier. There are four steps to drawing a GI plot (see figure 5):

1) This step is optional. Eliminate the inter-quartile-box in favor of a thin line segment flanked by brackets ([ and ]) to denote the first and the third quartiles \(Q_1\) and \(Q_3\), respectively, and a vertical notch (\(|\) at the median \(Q_2\). To draw attention to the middle 50% of the data, one may wish to reintroduce the box by joining the top (and the bottom) ends of the brackets.

2) Draw a solid arrow spanning the right SD interval \((\bar{x}, \bar{x} + s)\).

3) Draw a dotted line intending to span the interval \((\bar{x} - cs, \bar{x} + cs)\). (A portion of the dotted line will be hidden behind the thin line and the solid arrow; this is intentional.) In practice, the dotted line may be shortened to the most extreme values within its intended expanse. This constitutes the \(c\)-SD Gaussian Interval, which need not remain symmetric about the mean \(\bar{x}\). Any data outside the expanse of the dotted line are flagged as outliers, with frequencies shown by using left-, right- and vertical notches as in an IVY plot.

4) Place a fulcrum indicating the mean; and then draw a thick vertical line joining the fulcrum to the horizontal axis, giving the appearance that a vertical rod is holding the GI plot in balance. Write the sample size alongside this vertical line. When multiple GI plots corresponding to several subgroups are drawn in the same figure, as in figure 7(b) the heights of the vertical rods are proportional to the sample sizes.

Thus, a \(c\)-SD GI plot is strictly more informative than the usual box plot.
Fig 5. A Gaussian Interval plot shows the seven-number summary, and the sample size $n$. It also flags potential outliers (showing their frequencies) that are farther than $c$-SD from the mean in either direction.

**CHOOSING $c$**

How to choose the multiplier $c$? If one wishes to flag a fraction $p$ of extreme values ($p/2$ on each side), under the assumption of normal distribution, one should choose the multiplier to be the $(1 - p/2)$-th percentile of the standard normal distribution (using code `qnorm(1-p/2)` in R). For example, if $p = 0.01$, then the multiplier is 2.5758; if $p = 0.02$, then the multiplier is 2.3263. Conversely, when the multiplier is 2, the associated $p = 0.0455$; and when the multiplier is 2.70, which corresponds to the $1.5 \times IQR$ criterion applied to a normally distributed variable, the associated $p = 0.007$. In figures 5–8, we have chosen $c = 1.96$, wishing to flag about 5 percent of normally distributed data. Readers may choose other values of $c$.

**DEPICTING LARGE DATA**

To convey the impact of our proposed modification, in figure 6, we show the standard box plot and the proposed GI plot of the large data set introduced in Example 2 showing the overall GPA of 257 students. How many outliers are flagged? The box plot shows three values flagged as outliers; it does not tell us how many items are tied at those values. The GI plot, using ($\bar{x} = 3.12$, $s = 0.59$, $c = 1.96$), flags five outlying values; but more informatively, exactly twelve outliers: three at GPA 1.2, five at 1.5, two at 1.7, one at 1.8, and one at 1.9.

![Diagram](image-url)
What can we say about the distribution of GPA for those students who will receive a graduation diploma (with GPA at least 2.0)? For this subset of students, the dot plot and the IVY plot need not be drawn again. One can simply erase the dots and the leaflets below 2.0 from those plots. However, once the twelve smallest values of GPA are eliminated, for the subset of students who will receive a graduation diploma, the quartiles increase, the mean increases and the SD decreases from the corresponding quantities for the entire student body. Hence, for the 245 students who will receive a graduation diploma, the boxplot and the GI plots must be redrawn. This we will do next. Moreover, we will separate the diploma recipients into two groups—urban and rural—based on the students’ residential addresses. In Figure 7, we draw the usual boxplots for the two groups, and then we draw the GI plots using \((n = 105, \bar{x} = 3.34, s = 0.4129, c = 1.96)\) for the urban group, and \((n = 140, \bar{x} = 3.08, s = 0.5037, c = 1.96)\) for the rural group.

For the 105 urban graduates’ GPA, both the boxplot and the GI plot indicate a slight left-skewness (since \(Q_2 - Q_1\) is larger than \(Q_3 - Q_2\) and the left whisker is longer than the right whisker); but for the 140 rural graduates’ GPA, both plots give conflicting indication of skewness suggesting lack of skewness. The boxplots do not show the mean GPA of the two residential groups; but the GI plots show that the mean for the urban group is larger than the mean for the rural group, and SD for the urban group is smaller than the SD for the rural group. Most strikingly, whereas the boxplots of urban and rural groups fail to identify any outlier, the GI plots (with \(c = 1.96\)) identify seven outliers: two urban students’ GPA at 2.3, and five rural students’ GPA at 2.0! Finally, the sample sizes are missing from the boxplots; but they are depicted in the GI plots as heights of vertical rods through the fulcrums.

Fig. 7. (a) A standard boxplot and (b) a 1.96-SD GI plot of 245 GPA scores of graduates separated by urban and rural residence.
DISCUSSION

To efficiently summarize large data on a quantitative variable, we propose an IVY plot as an enhancement to a dot plot, and a GI plot as an enhancement to a traditional boxplot. Compared to a dot plot, the IVY plot saves space, and facilitates counting the frequencies of each value. While a standard boxplot provides a five-number summary and detects outliers, the proposed GI plot depicts some additional information such as the sample size, the mean, the SD, and the exact frequencies of outliers. Such information is useful when several groups are compared in the same graph. Admittedly, the GI plot makes good sense only under the assumption that the variable is approximately normally distributed. However, when the assumption holds, the GI plot is more powerful in detecting potential outliers.

We hope that the proposed visualizations will empower students and users of statistics.

ACKNOWLEDGEMENT

The authors are indebted to the Editor and three anonymous referees for suggesting numerous improvements to an earlier draft.

REFERENCES


