

Scatterplots

SIÂN E. WILLIAMS AND DANIEL B. WRIGHT

Volume 4, pp. 1798–1799

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Scatterplots

Scatterplots are typically used to display the relationship, or association, between two variables. Examples include the relationship between *age* and *salary* and that between inches of *rainfall* in a month and the number of *car accidents*. Both variables need to be measured on some continuum or scale. If there is a natural response variable or a predicted variable, then it should be placed on the *y*-axis. For example, age would be placed on the *x*-axis and salary on the *y*-axis because it is likely that you would hypothesize that salary is, in part, dependant on age rather than the other way round.

Consider the following example. The estimated number of days in which students expect to take to complete an essay is compared with the actual number of days taken to complete the essay. The scatterplot in Figure 1 shows the relationship between the estimated and actual number of days.

Most statistical packages allow various options to increase the amount of information presented. In Figure 1, a diagonal line is drawn, which corresponds to positions where estimated number of days equals actual number of days. Overestimators fall below the diagonal, and underestimators fall above the diagonal. You can see from this scatterplot that most students underestimated the time it took them to complete the essay. **Box plots** are also included here to provide univariate information.

Other possible options include adding different regression lines to the graph, having the size of the

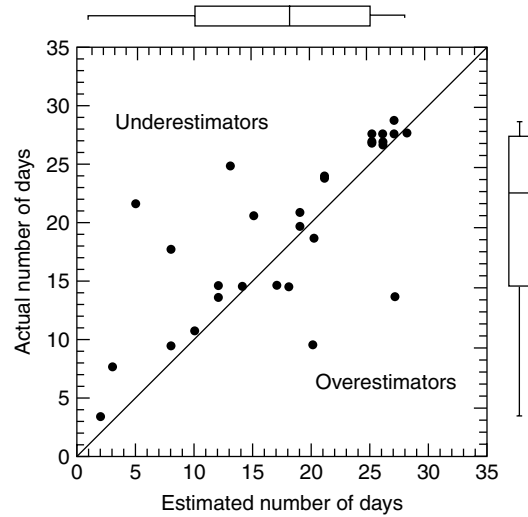


Figure 1 A scatterplot of estimated and actual essay completion times with overlapping case points

points represent their impact on the regression line, using ‘sunflowers’, and ‘jittering’. The use of the sunflowers option and jittering option allow multiple observations falling on the same location of the plot to be counted. Consider Figures 2(a), (b), and (c). Participants were presented with a cue event from their own autobiography and were asked whether that event prompted any other memory [2]. Because participants did this for several events, there were 1865 date estimates in total. If a standard scatterplot is produced comparing the year of the event with the year of the cueing event, the result is Figure 2(a). Because of the large number of events, and the fact that many

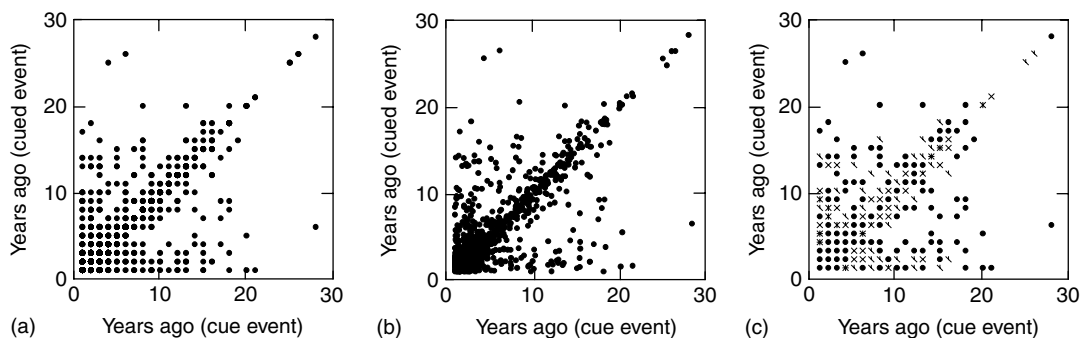


Figure 2 Scatterplots comparing the year of a remembered event with the year of the event that was used to cue the memory. Adapted from Wright, D.B. & Nunn, J.A. (2000). Similarities within event clusters in autobiographical memory, *Applied Cognitive Psychology* **14**, 479–489, with permission from John Wiley & Sons Ltd

2 Scatterplots

overlap, this graph does not allow the reader to determine how many events are represented by each point.

In Figure 2(b), each data point has had a random number (uniformly distributed between -0.45 and +0.45) added to both its horizontal and vertical component. The result is that coordinates with more data points have more dots around them. Jittering is particularly useful with large data sets, like this one. Figure 2(c) shows the sunflower option. Here, individual coordinates are represented with sunflowers. The number of petals represents the number of data points. This option is more useful with smaller data sets. More information on jittering and sunflowers can be found in, for example, [1].

It is important to realize that a scatterplot does not summarize the data; it shows each case in terms

of its value on variable x and its value on variable y . Therefore, the choice of regression line and the addition of other summary information can be vital for communicating the main features in your data.

References

- [1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.
- [2] Wright, D.B. & Nunn, J.A. (2000). Similarities within event clusters in autobiographical memory, *Applied Cognitive Psychology* **14**, 479–489.

SIÁN E. WILLIAMS AND DANIEL B. WRIGHT