Jittering

StatPREP Class Lesson

Orientation

Graphical displays are effective when they connect well with human visual and cognitive faculties. Humans are extremely good at some visual tasks and surprisingly poor at others. To illustrate, consider the task of identifying how many 3's there are in the following list:

5 4 6 6 9 5 5 5 0 5 0 5 7 6 5 4 6 6 9 3 2 6 2 7 6

2 4 0 0 9 8 4 2 3 9 4 7 3 1 2 1 6 8 7 7 2 7 1 8 0

6 0 5 1 0 1 6 3 1 8 3 3 3 0 2 0 8 5 8 8 2 5 7 1 5

7 1 4 9 4 0 0 0 5 2 2 9 0 6 8 2 7 4 0 6 2 4 5 0 9

The task is difficult because recognizing shapes is a tough cognitive task.

The task becomes much simpler if we take advantage of color, which our visual systems can process very quickly. Try finding the 3's again:

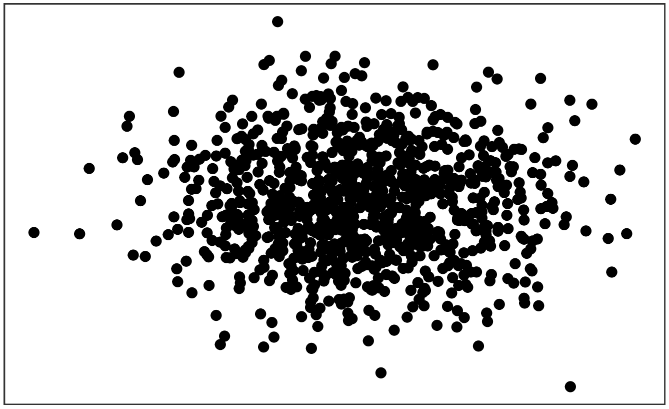
9 6 1 8 3 4 9 5 4 8 1 7 9 8 6 5 6 7 5 0 1 1 2 1 6

0 3 6 4 0 4 0 1 3 0 1 5 0 1 7 1 8 4 4 6 9 4 5 2 6

8 1 4 7 9 1 8 8 1 1 2 1 1 1 8 5 4 2 8 0 8 5 3 8 2

8 9 1 1 4 1 1 3 0 0 0 4 5 6 0 5 6 0 7 7 6 5 6 0 7

One visual task that people are good at might be called "seeing the forest despite the trees." When dots are scatter across a display, people can easily perceive differences in density from place to place. This despite the dots being all identical to one another.



This human capacity to see density is one of the factors that makes point plots an effective way to present data.

Regrettably, when one or both of the variables in a point plot are categorical, all of the dots at each level of the variable line up; multiple dots can be exactly in the same place in the graph. This is called *overplotting*. We perceive this multiplicity as a single dot and cannot see changes in density for different values of the variables.

Jittering and *transparency* are techniques to modify point plots so that we can see the relative density for different values of the variable.

Activity

Open up the jittering Little App‘. (See footnote[1]).

1. Select `NHANES` as the data frame, with `sex` as the response variable and `education` as the explanatory variable. Leave the sample size a n = 50.

   Count the points in the image.

   Are there n = 50 dots visible? . . .

2. Turn down the sample size to n = 5. There are now some places where there were dots for n = 50 but not for n = 5.

   What's going on at the places where there are not dots? . . .

   Look at the "statistics" tab. This contains a table of the points being displayed. Is the number of rows in the table consistent with n = 5? According to the table, are there any combinations of variables for which there is more than one row? Press "New Sample" if necessary so that there will be at least one repeated combination.

   Looking at the corresponding place in the graphic, is there any sign that the combination is repeated multiple times. . . .

3. Turn up the sample size to, say, n = 1000.

   With n = 1000, can you tell which values of the variables are most common? Why not? . .

4. Leaving n = 1000, move the horizontal slider that controls jittering width from zero to a value of about 0.1. Notice what's changed in the plot.

5. Move the vertical slider to move from zero vertical jittering to about 0.1.

   Is it possible to tell which combinations of the variables are most common? If so, explain how. . .

6. Select n to include the entire sampling frame: all the data. This will put about 7000 points in the graphic.

   Is it still possible to distinguish the relative frequency of the different combinations of the response and explanatory variables? . . .

[1] https://dtkaplan.shinyapps.io/LA_jitter/

Try changing the jittering sliders to make the differences in relative frequency more evident, while still keeping the square clouds of points separate.

Is your choice of jittering settings larger or smaller than it was when n = 1000 points were being displayed. . . .

7. Even with jittering, with large n there will be some overplotting. This is where transparency comes into the picture. Using transparency reveals such overplotting in terms of the perceived darkness of the display.

   Turn down the transparency slider from 1 to a small value that makes it easy to see where there is overplotting. (Hint: Depending on the jittering settings you selected in (4), you might need to make the value of the transparency slider very small.

   What are the values of the horizontal and vertical jittering controls and of the transparency control that make it easiest to distinguish the different density of data across the ten different variable combinations? . . .

8. Set the response variable to `income_poverty`, which is quantitative. (Leave the explanatory variable as `education`.) And turn the vertical jittering to zero.

   Can you tell which values of `income_poverty` are more common at the different education levels? Adjust the horizontal jittering and transparency to create a plot you think is effective at showing how the distributions of `income_poverty` change from one education level to another.

   Write down the horizontal jittering and transparency you think make the most effective graph. . . .

   Tell the story shown by the graph using everyday English terms. . . .

9. Check the “show violin plot” box underneath the jittering controls.

   Explain what is the relationship between the violin-like shapes and the density displayed by the jittered point plot. . . .

Version 0.3, 2019-05-28, Danny Kaplan, Word version