Data and Point Plots

StatPREP Class Lesson

Orientation

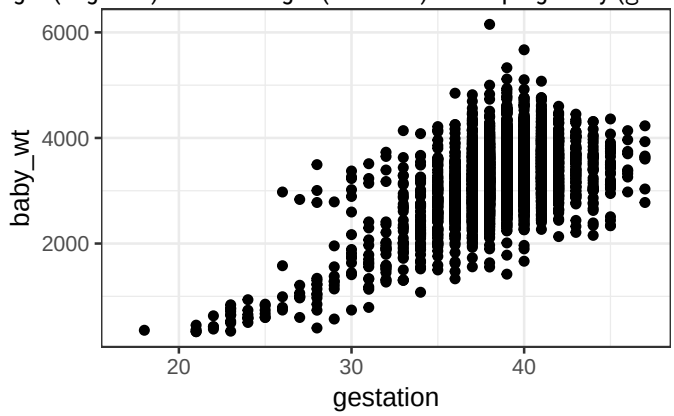
Data can be many things, but one of the most common formats is a data frame, a kind of spreadsheet of rows and columns. We'll work with the data frame `Births_2014`, which is based on data published by the US Centers for Disease Control. `Births_2014` has 100,000 rows. Each row reports a live birth in the US in 2014. There are dozens of variables, a few of which are shown below.

sex	baby_wt	gestation	delivery	age_mother	wic
M	4479	41	spontaneous	28	n
F	3203	39	vacuum	28	n
F	3590	39	spontaneous	31	y
F	3771	41	cesarean	34	n
F	3335	39	cesarean	38	NA
M	3750	41	spontaneous	19	y

It's hard to draw much of a conclusion by looking directly at a large data frame. But a graphical display of data can help.

A *point plot* is a basic statistical graphic that displays two variables from a data frame. Point plots[1]. One variable is represented on the vertical axis, another variable on the horizontal axis. Like the following point plot of the baby's weight (in grams) and the length (in weeks) of the pregnancy (`gestation`).

[1] Sometimes referred to as "scatterplots"



Activity

1. Find in the graph the dot corresponding to the first row in the data table above, the one for a male baby delivered spontaneously to a 28 year-old mother.

2. Describe the overall pattern shown in the graph as a whole. Use whatever form of description you think is appropriate.
3. Of course, weight differs from one baby to another. In other words, weight *varies*. Describe how much *variation* there is in babies' weight, according to the graph.
4. Describe how much *variation* there is in gestation length.
5. At which length of gestation are the heaviest babies born?

Activity

Open the Point Plot Little App.

1. Set the data source to `Births_2014` and choose `baby_wt` as the response variable and `gestation` as the explanatory variable. The resulting plot should look much like the graph seen in the introduction to this lesson. But plot seen in the Little App will have many fewer points. Change the sample size to $n = 5$.

2. Open the "Statistics" tab under the main graph. This tab displays the same data as in the plot, but in data-frame format.

   - For each of the $n = 5$ rows of the data frame, find the corresponding point in the graphic.

3. Change the *explanatory* variable to `sex`.

   - For each of the $n = 5$ rows of the data frame displayed in the Statistics tab, find the corresponding point in the graphic.
   - Change n to 500. In the `baby_wt` versus `sex` graph, all the points are lined up in two columns. Explain why.

4. Change the *response* variable to `delivery`, keeping the explanatory variable as `sex`.

   - For a few of the rows of the data frame shown in the Statistics tab, find the corresponding point in the graphic.
   - Make sure that n is something large, say $n = 500$. There aren't 500 points in the `delivery` versus `sex` graph. Explain why?

5. Check the "jitter categorical variables" box in the controls. The display changes and now there are many more points in the plot.

   - For a few of the rows in the data frame shown in the Statistics tab, find the corresponding point in the graphic. Are you able to uniquely identify in the graph the specific point corresponding to each row? Explain how you can do this or why it's not possible.

Version 0.1, 2019-04-15, Daniel Kaplan