 Lagunita is retiring and will shut down at 12 noon Pacific Time on March 31, 2020. A few courses may be open for self-enrollment for a limited time. We will continue to offer courses on other online learning platforms; visit <http://online.stanford.edu>.

Course > EDA: Examining Relationships > Case Q→Q: Scatterplots > Scatterplot: Introduction

 Bookmark this page

Scatterplot: Introduction

Learning Objective: Graphically display the relationship between two quantitative variables and describe: a) the overall pattern, and b) striking deviations from the pattern.

In the previous two cases we had a categorical explanatory variable, and therefore exploring the relationship between the two variables was done by comparing the distribution of the response variable for each category of the explanatory variable:

- In case C→Q we compared distributions of the quantitative response.
- In case C→C we compared distributions of the categorical response.

Case Q→Q is different in the sense that both variables (in particular the explanatory variable) are quantitative, and therefore, as you'll discover, this case will require a different kind of treatment and tools. Let's start with an example:

Example: Highway Signs


A Pennsylvania research firm conducted a study in which 30 drivers (of ages 18 to 82 years old) were sampled, and for each one, the maximum distance (in feet) at which he/she could read a newly designed sign was determined. The goal of this study was to explore the relationship between a driver's **age** and the **maximum distance** at which signs were legible, and then use the study's findings to improve safety for older drivers. (Reference: Utts and Heckard, *Mind on Statistics* (2002). Originally source: Data collected by Last Resource, Inc, Bellfonte, PA.)

Since the purpose of this study is to explore the effect of age on maximum legibility distance,

- the **explanatory** variable is **Age**, and

- the **response** variable is **Distance**.

Here is what the raw data look like:



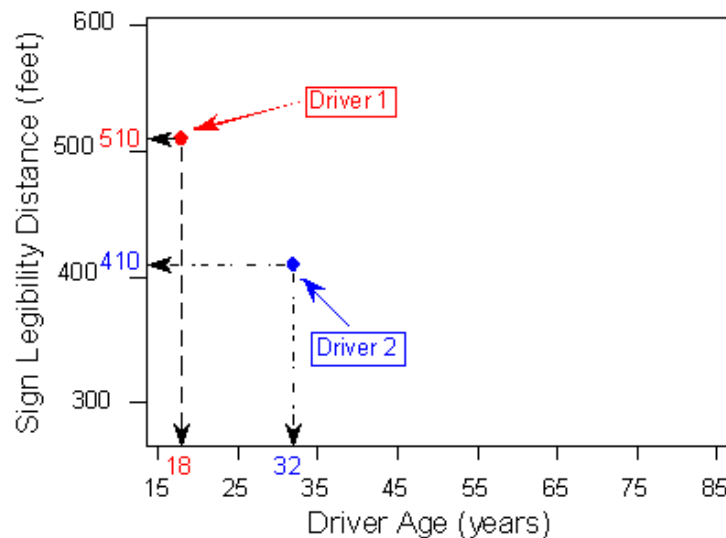
	Age	Distance
Driver 1	18	510
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

Note that the data structure is such that for each individual (in this case driver 1....driver 30) we have a pair of values (in this case representing the driver's age and distance). We can therefore think about these data as 30 pairs of values: (18, 510), (32, 410), (55, 420), ... , (82, 360).

The first step in exploring the relationship between driver age and sign legibility distance is to create an appropriate and informative graphical display. The appropriate graphical display for examining the relationship between two quantitative variables is the **scatterplot**. Here is how a scatterplot is constructed for our example:

To create a scatterplot, each pair of values is plotted, so that the value of the explanatory variable (X) is plotted on the horizontal axis, and the value of the response variable (Y) is plotted on the vertical axis. In other words, each individual (driver, in our example) appears on the scatterplot as a single point whose X-coordinate is the value of the explanatory variable for that individual, and whose Y-coordinate is the value of the response variable. Here is an illustration:

	Age (X)	Distance (Y)
Driver 1	18	510
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360



And here is the completed scatterplot:



Comment

It is important to mention again that when creating a scatterplot, the explanatory variable should always be plotted on the horizontal X-axis, and the response variable should be plotted on the vertical Y-axis. If in a specific example we do not have a clear distinction between explanatory and response

variables, each of the variables can be plotted on either axis.

Open Learning Initiative [↗](#)



[↗](#) Unless otherwise noted this work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License [↗](#).

© All Rights Reserved