Conditional Expected Value and Distribution

Car Performance Data

Bayes' Rule Practice Problem

February 14, 2018

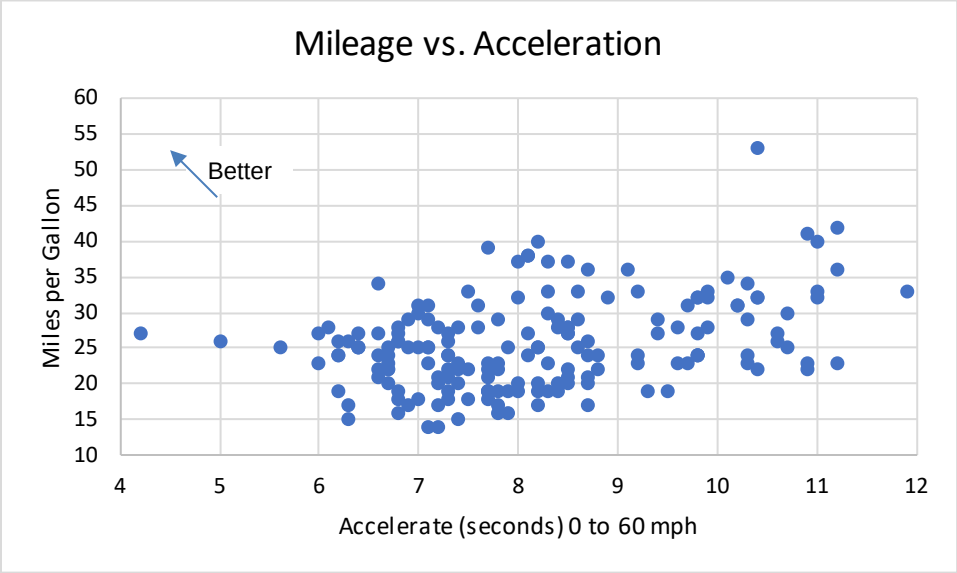
BACKGROUND

Consumer Reports (CR) is an independent, non-profit organization that tests and rates a variety of consumer products. Automobiles are an item category important to consumers. The accompanying worksheet, **`mileage.csv`**, contains a table of 177 records of two parameters:

- *Mileage*, measured in average miles per gallon (mpg) of gasoline
- *Acceleration*, measured as the time (in seconds) to accelerate from 0 to 60 miles per hour (mph) (apologies to readers who remember physics, where acceleration = distance unit / second^2)

The original data are from https://www.consumerreports.org/cro/news/2013/06/fuel-economy-vs-performance/index.htm. I omitted car makes and models, excluded all other parameters, excluded electric vehicles, and disguised the data slightly with light dithering.

Here is an XY chart of the data:



Problem

Assume each car model point is equally-probable. Suppose you learn for a car of interest that Acceleration is 7.5 seconds.

1. Calculate the mean Mileage (an *EV* approximation) for a vehicle given that it accelerates to 60 mph in approximately 7.5 seconds.
2. Plot the Mileage distribution given that Acceleration is 7.5 seconds.

Suggestion: First estimate both solutions by inspecting the XY chart.

Stop! Solve the problem before continuing.

Solution

Acceleration measurements are to 1 decimal place. There are only three vehicles that Accelerate in 7.5 seconds. Because three is an insufficient sample size, expand the range to include more points. I used the range 7.0 to 8.0 seconds, inclusive. This subset contains 54 of the 177 points in **`mileage.csv`**.

Value	No. points
7.0	4
7.1	6
7.2	5
7.3	8
7.4	5
7.5	3
7.6	2
7.7	8
7.8	6
7.9	3
8.0	4
	54

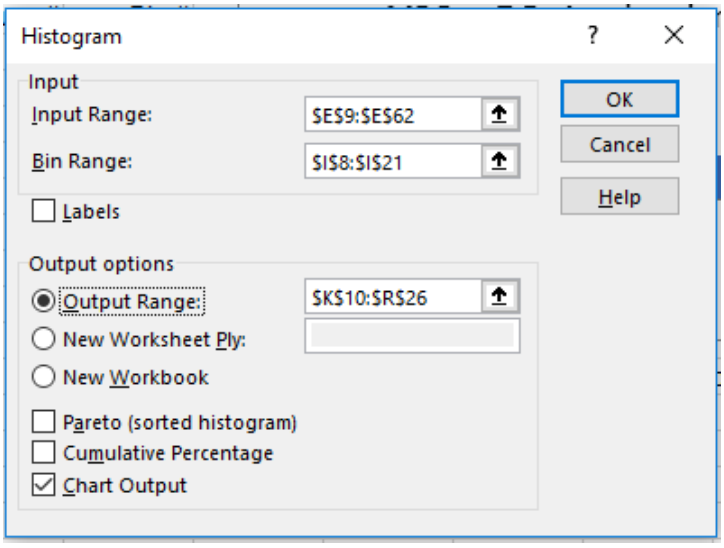
Bins
13
15
17
19
21
23
25
27
29
31
33
35
37
39

Question 1: The mean Mileage value (using Microsoft® Excel®'s **`Average`** function) in the 54 cells where 7.0 ≤ Mileage ≤ 8 mpg is **23.1 mpg**. This *sample mean* is an estimate of EV Mileage given Acceleration is 7.5 seconds.

Excel's Data Analysis toolkit provides an adequate histogram charting capability. It works better to define your own data bins so that the bin boundaries are pleasantly rounded. I used the bin boundaries, above, which are 2 mpg wide. Bin centers are even-numbers.

Access the Histogram tool from Excel's menu sequence: Data / Data Analysis / Histogram.

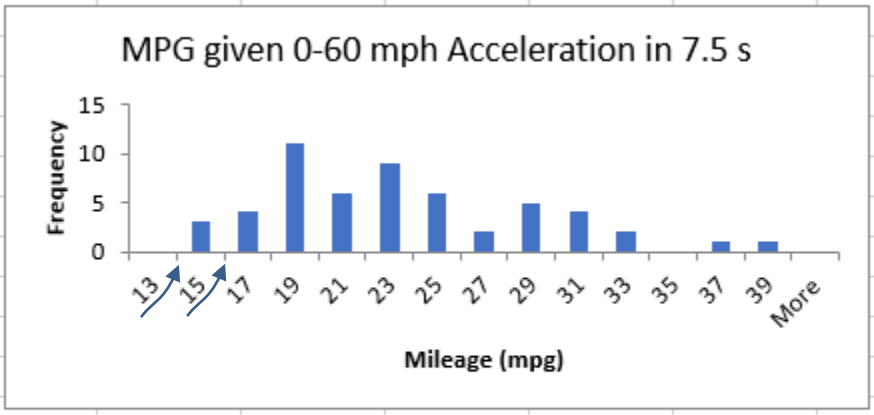
Here is the set-up screen:



Input Range	Select the 54 Mileage data values where Acceleration is 7 to 8 seconds, inclusive. The data first need sorting by Acceleration.
Bin Range	Select the predetermined bin bounds. Otherwise, Excel will assign bin cutoffs (usually poorly).
Output Range	Select a cell in the spreadsheet with substantial space down and right (at least number of bins + 2 rows × 8 columns). Alternately, you can direct Excel to put the output on a new worksheet (New Worksheet Ply) or in a New Workbook.

Question 2: Here is the produced histogram data and chart (after some cleanup):

Bin	*Frequency*
13	0
15	3
17	4
19	11
21	6
23	9
25	6
27	2
29	5
31	4
33	2
35	0
37	1
39	1
More	0



Bin counts are above lower-bound and through upper-bound.

Excel programmers could do a better job of *x*-axis labeling. The bin value labels along the *x*-axis refer to the *next* tick mark at the right. When some data values equal bin bounds, this requires additional care. Here, Mileage values are integers. As examples:

- The first bar represents three values: two 14s and one 15. The bar range is 13 < Mileage ≤ 15.
- The second bar represents four values: two 16s and two 17s. The bar range is 15 < Mileage ≤ 17.

Bayesian Analysis

Working with a dataset in this manner I call "Bayes by Cloud."

What does this have to do with Bayes' rule?

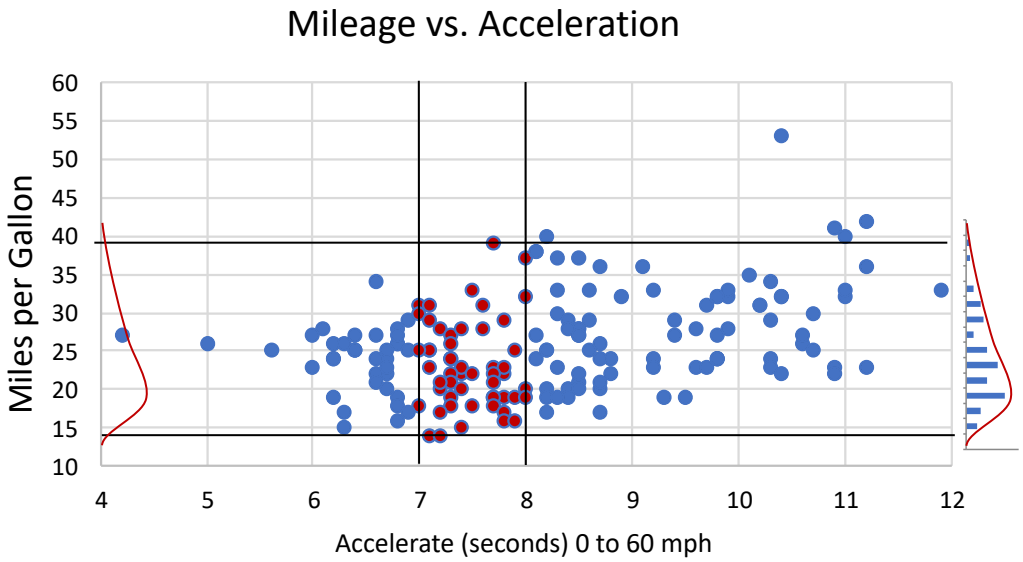
Recall the simple form of Bayes' rule to calculate a conditional probability:

$$P(B \mid A) = \frac{P(A \bullet B)}{P(A)}$$

P(B) is *conditioned* by new, usually imperfect information A. Most often, Bayes' rule is used with discrete probabilities. The conditioning eliminates some potential outcomes, and we normalize the probabilities of what remains.

In the data cloud context of our problem, the conditioning is by partitioning a large dataset (points in *n* dimensions) to exclude irrelevant records. The remaining subset is the basis for conditioned EV estimates and conditional frequency distributions (FDs). This is the basis for *Bayesian analysis*, a principal *analytics* operation on large datasets ("Big Data").

This next figure shows the car data subset defined by *conditioning* on Acceleration being approximately 7.5 seconds. The range 7.5 ± 0.5 seconds allows sufficient points to remain.



The probability of randomly-choosing car type having Acceleration in the 7 to 8 mpg range is approximately 54/177 = .305. This is the fraction of red-filled points.

Projecting the red-filled points to the *y*-axis:

- EV Mileage given Acceleration ≅ 7.5 seconds is **23.1 mpg**. This is the average *y*-axis values of the 54 red-highlighted points.
- The frequency histogram (at the right, turned sideways) summarizes the *y*-axis values of the red-highlighted points. This is the conditional frequency distribution (FD) of Mileage given Acceleration ≅ 7.5 seconds. The overlying red, smooth curve (and copied at left) is my attempt at approximating the shape of the conditional PDF.

■