Conditional Expected Value and Distribution

Vehicle Performance Data

Bayes' Rule Practice Problem

February 14, 2018; Revised April 4, 2018

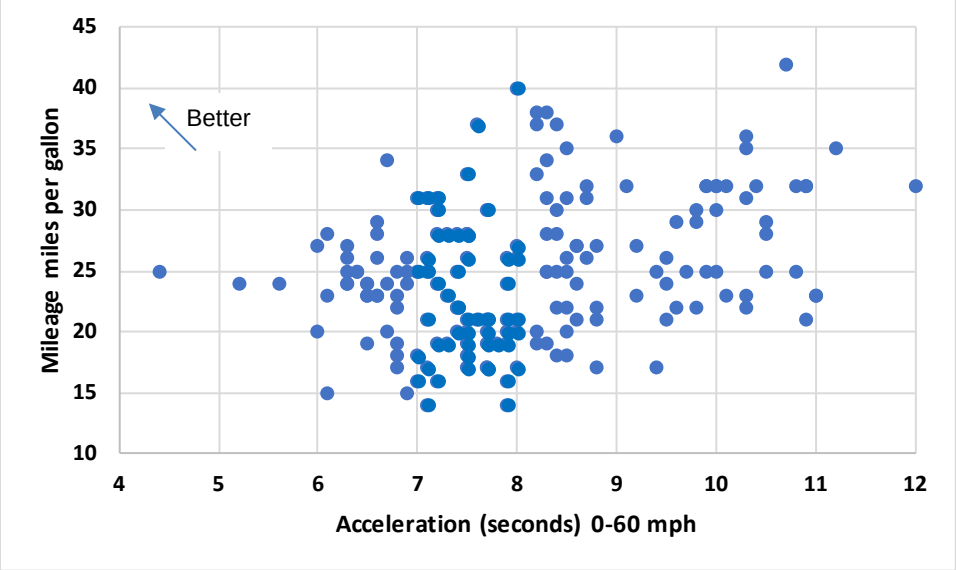
BACKGROUND

Consumer Reports (CR) is an independent, non-profit organization that tests and rates a variety of consumer products. Automobiles and light trucks are an item category important to consumers. The accompanying worksheet, **`mileage.csv`**, contains a table of 173 records of two parameters:

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

The original data are from https://www.consumerreports.org/cro/news/2013/06/fuel-economy-vs-performance/index.htm. Let's assume that the CS report is a random sampling of a much larger population of vehicles. I omitted manufacturer names and models, excluded all other parameters, excluded electric and hybrid vehicles, and disguised the data slightly with light dithering.

Here is an XY chart of the data:



Problem

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

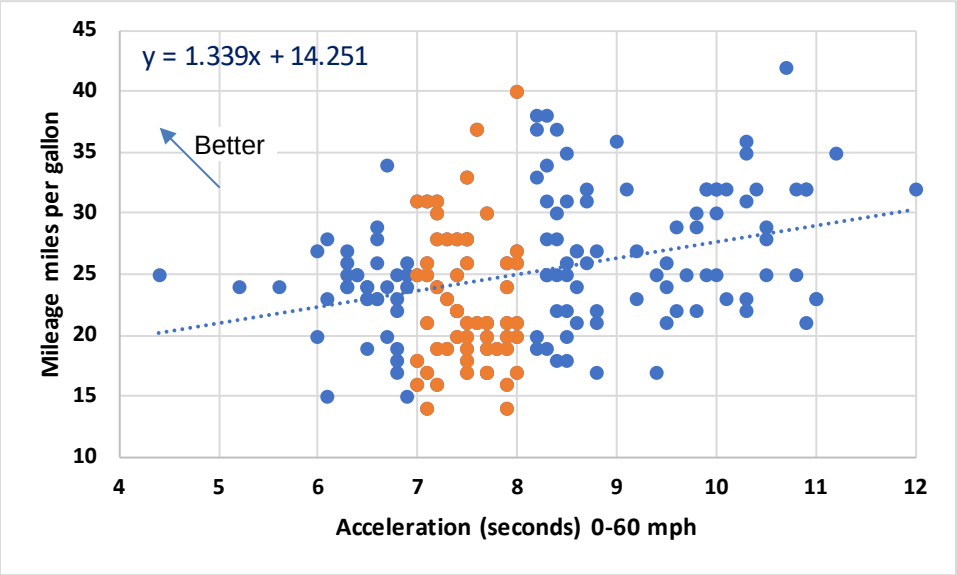
1. Calculate the mean Mileage (an *EV* approximation) for a vehicle given that it accelerates to 60 mph in approximately 7.5 seconds.
2. Plot the Mileage distribution given that Acceleration is about 7.5 seconds.

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

Stop! Please attempt to solve the problem before continuing.

Solution

Here is an XY chart of the 173 data points:



Visually, Mileage and Acceleration appear to have a positive correlation. The Pearson correlation coefficient (Excel's CORREL function) is 0.33. The Excel's linear trendline (dotted line) also confirms it. One would expect heavier vehicles generally to have lower Mileage and slower Acceleration.

Acceleration measurements are to 1 decimal place, and Mileage values are integers. There are nine vehicles reported to Accelerate in 7.5 seconds. The average Mileage of those is **23.33 mpg**. That's one answer to our problem. However, nine is a small sample size, and I suggest expanded the range to include more points. I selected the range 7.0 to 8.0 seconds, inclusive. This subset (orange points in the chart) contains 56 of the 173 points in `mileage.csv`. Orange appears lighter-gray if this document is printed in grayscale.

Acceleration	No. Points
7	4
7.1	6
7.2	6
7.3	3
7.4	5
7.5	9
7.6	2
7.7	7
7.8	1
7.9	7
8	6
	56

Mileage	No. Points	Wt. Avg.
14	2	0.50
16	3	0.86
17	5	1.52
18	2	0.64
19	6	2.04
20	5	1.79
21	7	2.63
22	2	0.79
23	1	0.41
24	2	0.86
25	3	1.34
26	4	1.86
27	1	0.48
28	5	2.50
30	2	1.07
31	3	1.66
33	1	0.59
37	1	0.66
40	1	0.71
	56	22.89

The right table, above, calculates mean Mileage in the 56 cells where 7.0 ≤ Mileage ≤ 8 mpg. This **22.89 mpg** *sample mean* is an estimate of EV Mileage given that Acceleration is about 7.5 seconds.

Check for Reasonableness

Excel's linear-regression trendline through all 173 points has the equation:

$$\text{Mileage} = 14.251 + 1.339(\text{Acceleration})$$

So, at Acceleration = 7.5 s, the regression equation estimates:

$$\text{Mileage} = 14.251 + 1.339(\text{Acceleration}) \text{ mpg}$$

With Acceleration = 7.5 s,

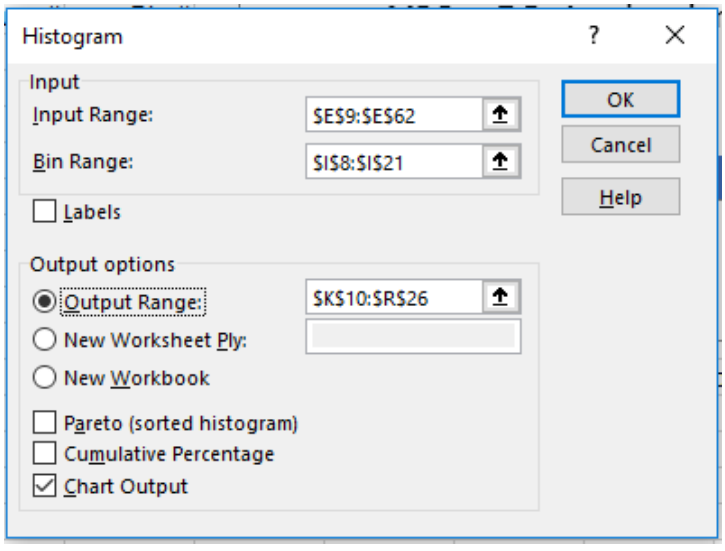
$$\text{EV Mileage} \mid 7.5 \text{ s Acceleration}$$
$$= 14.251 + 1.339(7.5) = 24.47 \text{ mpg}$$

Histogram Plot

Excel's Data Analysis toolkit provides an adequate histogram charting capability. Define data bins so that the boundaries are pleasantly rounded. Beware: Excel is not always consistent in sorting data into bins whose boundaries are the same as some data. I experienced that problem in this example. To ensure correct bin counts, define the bin boundaries to be between possible values.

Access Excel's Histogram tool using this menu sequence: Data / Data Analysis / Histogram.

Here is the Histogram set-up screen and process for preparing a frequency histogram for Acceleration in the 7 to 8-second range:



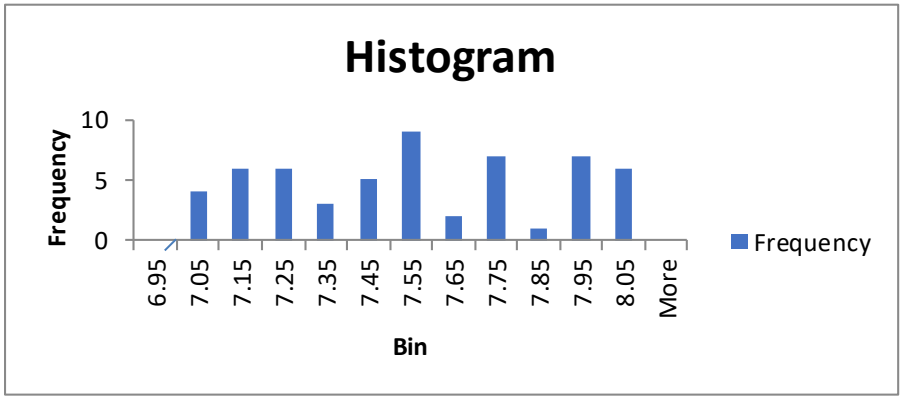
Input Range	The data first need sorting by Acceleration. Then select the Acceleration values in the 7-8 second range, inclusive.
Bin Range	Select the predetermined bin bounds. Otherwise, Excel will assign bin cutoffs (usually poorly). I used bins {6.95, 7.05, …, 8.05}.

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.

Be sure to check the Chart Output checkbox.

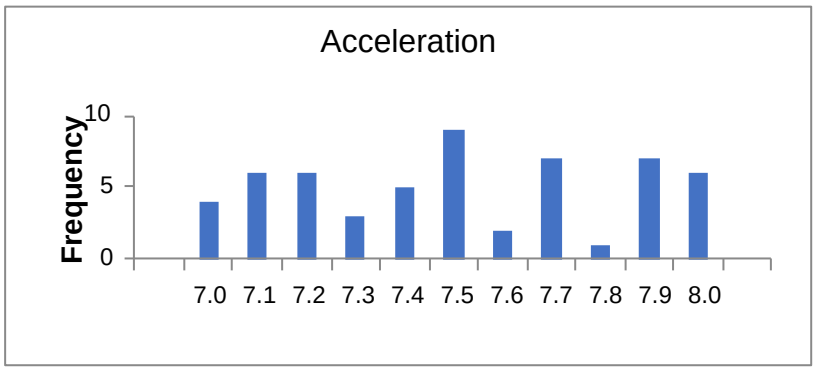
Question 2: Below is the produced histogram data and chart. A line-fit through the bar tops (not shown) is nearly horizontal.

Bin	*Frequency*
6.95	0
7.05	4
7.15	6
7.25	6
7.35	3
7.45	5
7.55	9
7.65	2
7.75	7
7.85	1
7.95	7
8.05	6
More	0
	56

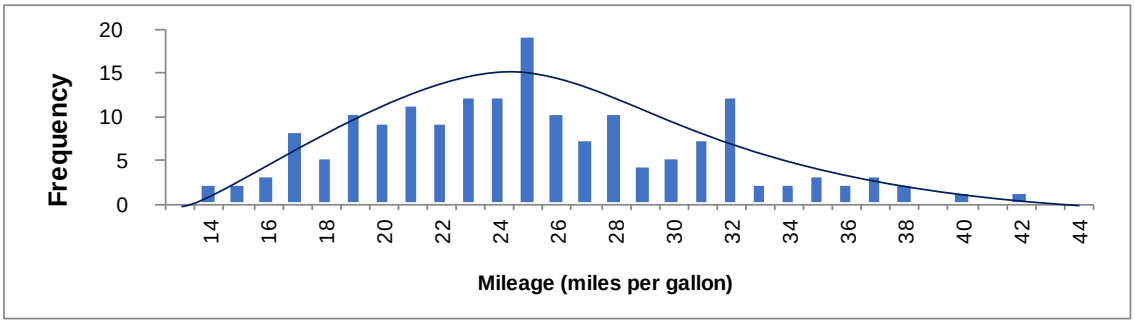


Microsoft's programmers could do a better job of histogram *x*-axis labeling. Usually, bin value labels along the *x*-axis on the chart above refer to the *next* tick mark, between bars, at the right. Here, Mileage values are integers.

After some cleanup using PowerPoint, my modified chart looks like this:



And here, prepared similarly, is a histogram showing the Mileage frequency distribution:



The swoop is my hand-drawn approximation of what the PDF might look like if the 173 samples are from a continuous distribution.

Bayesian Analysis

I call working with the points dataset in this manner "Bayes by Cloud."

What does this have to do with Bayes' rule?

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

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

Most often, we use a Bayes' rule formula with discrete probabilities. We obtain conditional probabilities by *partitioning* the joint probability table (or probability tree or Venn diagram) and normalizing the probabilities that remain.

In the formula, new, usually imperfect information *A conditions* P(*B*). In the vehicle example, both the Acceleration and Mileage are continuous variables, though they may look discrete from rounding. Even if the measurements are highly accurate, their relationship is fuzzy.

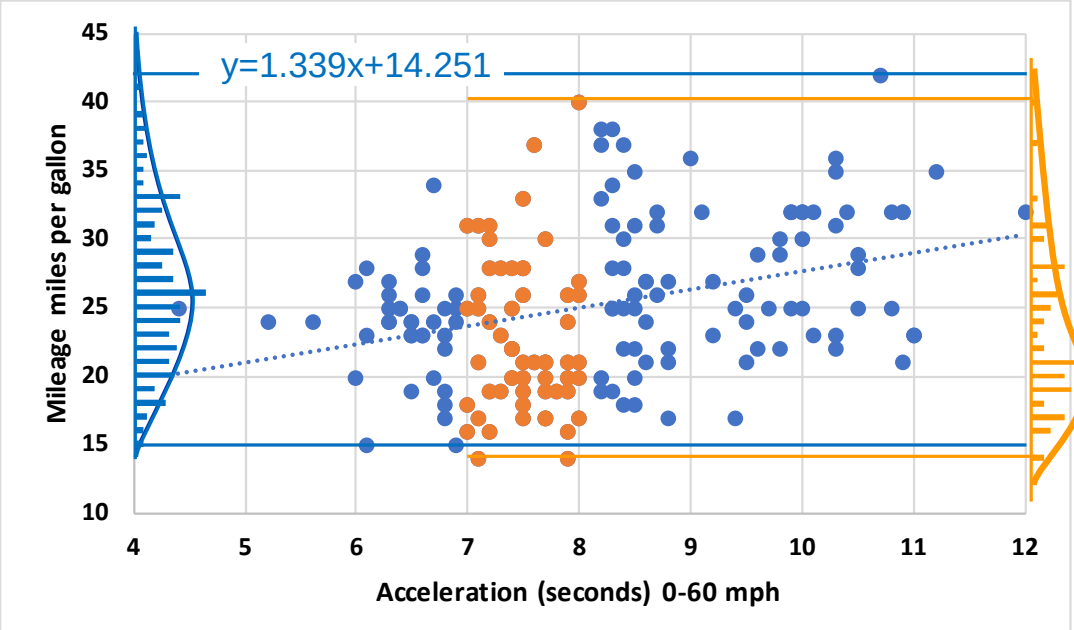
In the data cloud context of our problem, the conditioning is by partitioning a large dataset (points in *n* dimensions), excluding the now-irrelevant points, and normalizing what remains. The remaining subset is the basis for conditioned EV estimates and conditional frequency distributions (FDs). Conditioning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

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

The histogram at the left (turned sideways) shows the Mileage frequency distribution for all points.

The probability of randomly-choosing vehicle type having Acceleration in the 7 to 8 mpg range is approximately 56/173 = 0.324. This is the fraction of points that are orange.

Knowing that Acceleration is about 7.5 s, we have the conditional distribution drawn at the right.



Projecting the orange points (lighter-gray) to the right *y*-axis histogram:

- Though not by much, conditioning by Acceleration being 7-8 s reduces both the mean and standard deviation. In almost all cases, additional information (Acceleration) reduces uncertainty in the conditioned distribution (Mileage).
- EV Mileage given Acceleration ≅ 7.5 seconds is **22.89 mpg** (retaining unwarranted precision to show the calculations). This is the average *y*-axis values of the 56 orange-highlighted points. The right table on page 2 showed this mean calculation.
- The histogram (at the right) is the Mileage frequency distribution for the orange-highlighted points. This is the conditional frequency distribution (FD) of Mileage given Acceleration is about 7.5 seconds.
- The swoop curves overlaying the histogram bars are my attempts at approximating the shapes of inferred PDFs.

Was it necessary to expand the range of Acceleration points?

If you are interested, here is some more fun with probabilities. Out of curiosity, I checked the calculation EV Mileage calculation for different Acceleration ranges about 7.5 s.

Range	7.5	7.4-7.6	7.3-7.7	7.2-7.8	7.1-7.9	7-8
n points	9	16	26	33	46	56
sample mean	23.33	24.06	23.08	23.24	22.63	22.89
std. dev. (*s*)	5.57	5.60	5.23	5.31	5.30	5.68
SEM	1.86	1.40	1.03	0.92	0.78	0.76
T value for 68% CI	1.067	1.034	1.020	1.016	1.011	1.009
SEM ' = 68% CI	1.98	1.45	1.05	0.94	0.79	0.77

The EV Mileage given Acceleration values are in the third row. Coincidentally, rather than having abundant statistical support, the sample mean ($\bar{x}$) changed little with increased range. The sample standard deviation (*s*) is also relatively unchanged across its row.

The *standard error of the mean* (*SEM*) is a popular tool for testing convergence in Monte Carlo simulation. *SEM* measures how much confidence we have in the sample mean ($\bar{x}$) as an estimate for the true but unknown PDF mean (μ).

$$SEM = \bar{x} / \sqrt{n}$$

Notice how the values in the *SEM* row (highlighted) decrease as *n* increases. The *EV* calculation uncertainty fell 60% by including more points. So, for the last column, the estimate of Mileage given Acceleration = 7.5 s = 22.89 ± 0.76 with a 68% confidence. (The 68% *confidence interval* (CI) is from knowing that 68% of a normal distribution is within ± one standard deviation.)

You may recall that the formula for sample variance (s^2) divides by *n*-1 instead of *n*. The is a longstanding adjustment to correct for underestimating sample variance with *n* in the denominator. Well, there is a correction to the correction. The *Student T distribution* adjusts the *SEM* for small sample sizes, even as small as *n* = 2. I used T values at 68% so that *SEM*s and the T-based alternative *SEM* 's are comparable. At *n* =16 and higher, there isn't much difference.

■