STA610 Case Study 1

Emily Gentles, Weiyi Liu, Jack McCarthy, Qinzhe Wang

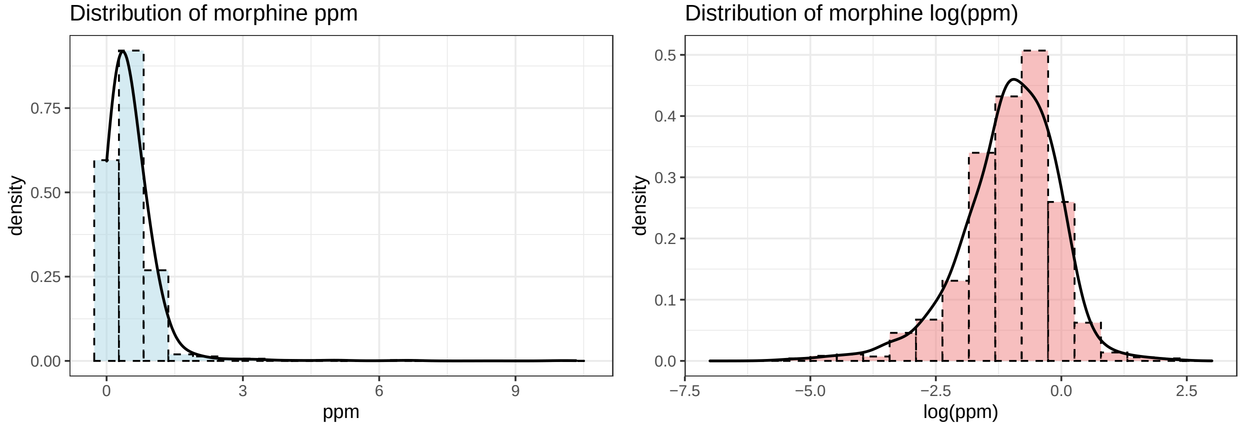
9/24/2021

Introduction

EDA

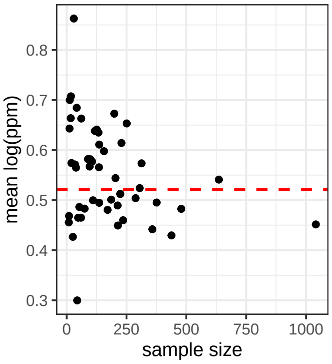
Response Distribution

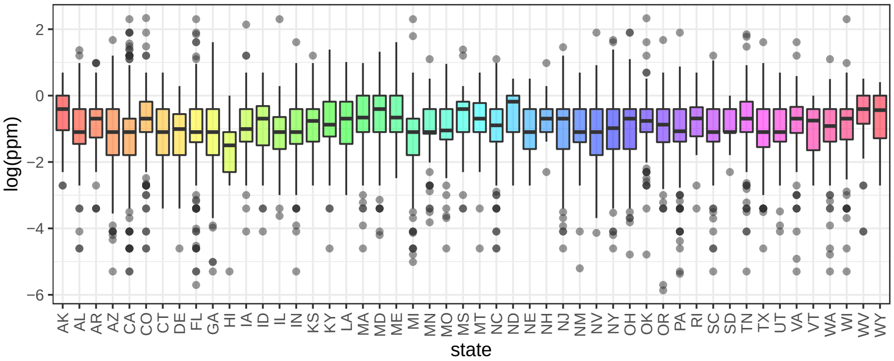
First, a look at the distributions of the response variable “ppm”. Observations with ppm between the 0.1 and 99.9 percentiles were considered so as to avoid the influence of extreme outliers on the analysis of the ppm distribution.



The distribution of ppm is clearly right-skewed, and it is strictly nonnegative in value, so a log transformation may be appropriate. The distribution of log(ppm) is given above, and appears closer to the desired normal.

state vs. log(ppm)

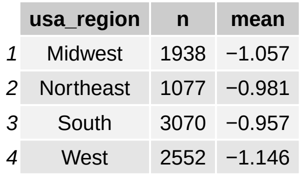




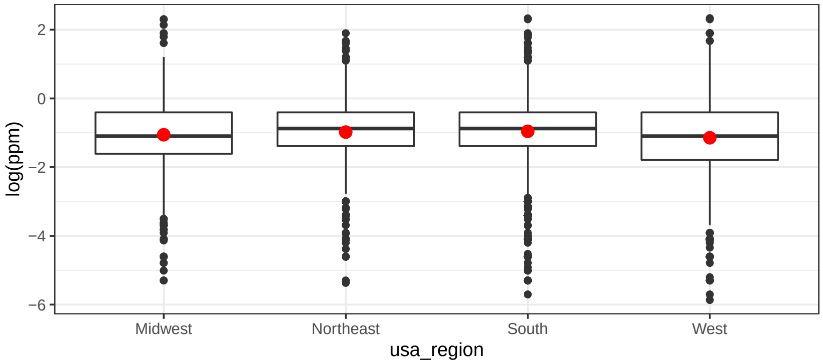
We observe that the within-state means for states with higher sample sizes in general adhere more closely to the grand mean. It is also evident that the log(ppm) distributions differ little as compared to the within-state variance. This is conducive to the borrowing of information between states.

region vs. log(ppm)

We also have access to the broader region in which a purchase is made. This could be useful if we wanted to develop a simpler model that still captured variation by purchase location.



	usa_region	n	mean
1	Midwest	1938	−1.057
2	Northeast	1077	−0.981
3	South	3070	−0.957
4	West	2552	−1.146


log(ppm)
Midwest
Northeast
South
West
usa_region