STA 130 Tutorial 2

David Veitch

University of Toronto

daveveitch.github.io

Jan 18, 2019

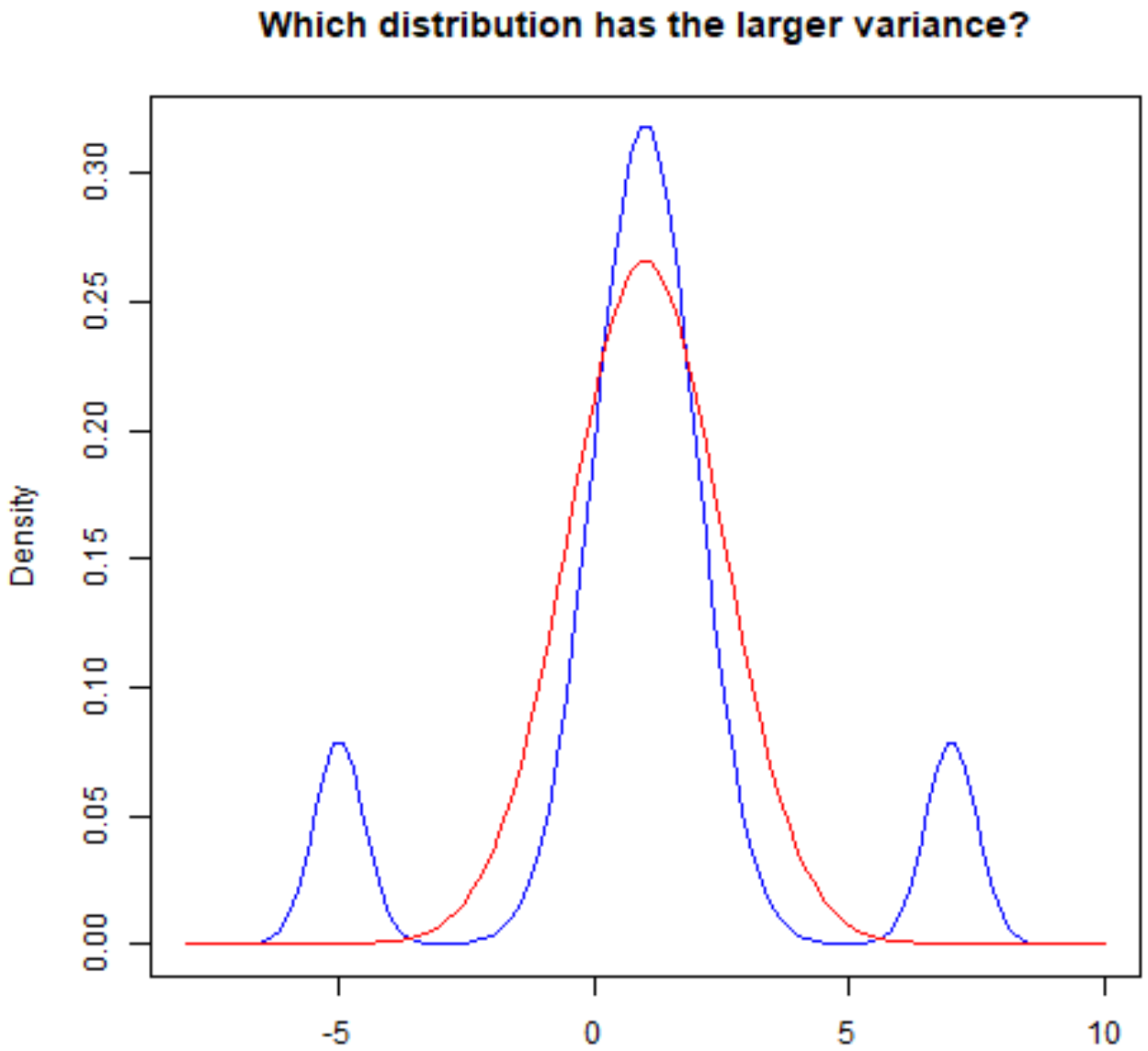
Agenda

1. Question 1 of Homework
   - Vocabulary List
   - Mean and Variance
2. Question 3 Discussion
3. Written Example
4. Assignment

Vocabulary List

- Dataframe
- Matrix
- Vector
- Average
- Standard Deviation
- Variance
- Missing Data
- Types of variables: character, numeric, factor

Mean and Variance



Mean and Variance

Answer: Although I have not analytically worked this out I suspect the blue has greater variance since the left and right modes will dominate.

Mean and Variance

Mean

$$\mu = \mathrm{E}[X] = \sum_{i=1}^{k} x_i f(x_i)$$

$$\mu = \mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f(x) dx$$

Variance

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \sum_{i=1}^{k} (x_i - \mu)^2\, f(x_i)$$

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x) dx$$

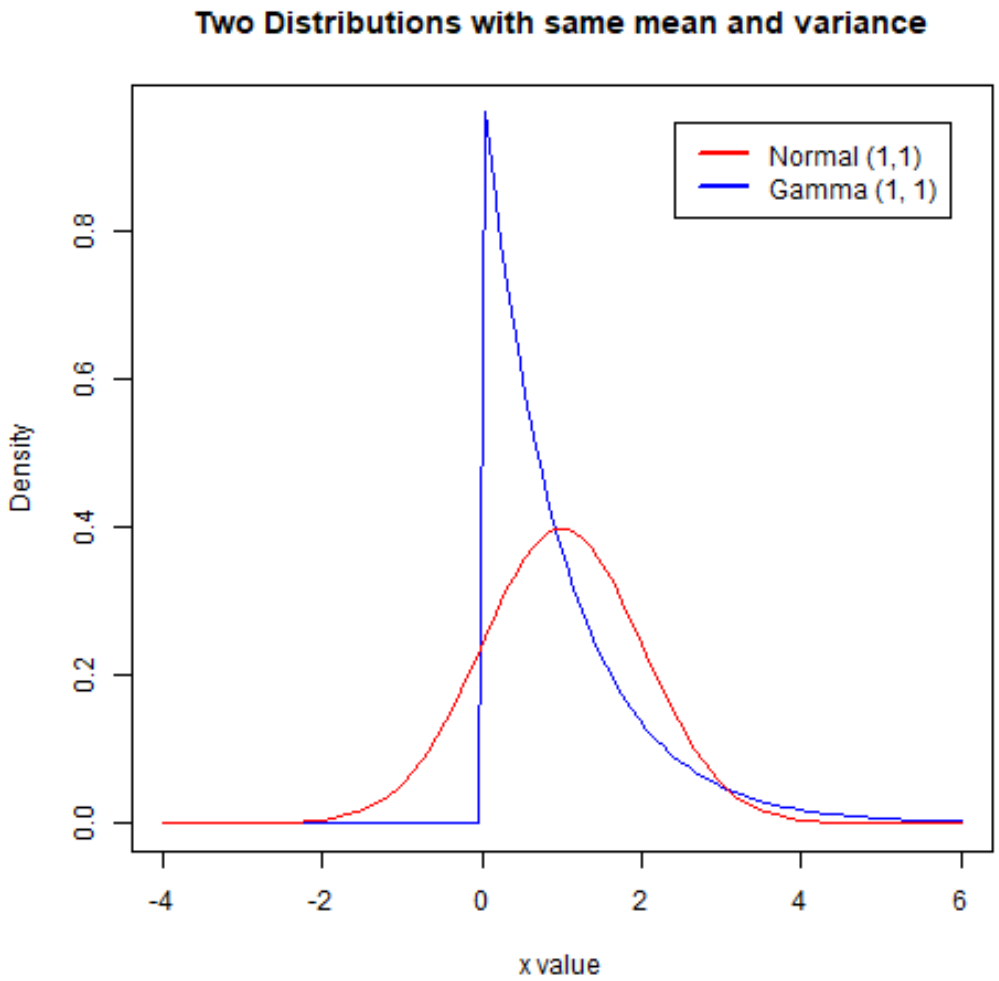
Note that $f(x)$ is the probability distribution function of x.

Mean and Variance

Does the mean and variance uniquely specify a distribution?

Mean and Variance

No!



Question 3 Discussion

- Describe what you did to create the variables.
- Explain why you did it this way.
- Compare graphs or summary statistics on the created variables.
- What were your main findings?

Key Things to Write

- Introduce your variables
- Define the problem
- How does data wrangling fit into problem?
- Summarize results

Example 1

For this question we used the Galton dataset, which provides data on children from in the 1880s. Because we were interested in investigating differences across families, we needed to create new variables which summed childrens characteristics by family ID. For example, we were interested in determining the number of children in each family. However, this value was repeated for every member of the families included in the Galton dataset. To make it easier to generate summaries, we made a new tidy data frame that included only one row per family. Using this new data set, we were more easily able to determine the number of children in each family group. The number of kids in the Galton data set families follow a positively skewed (or equivalently, right skewed) distribution. There were many more smaller families than largerquot; families. The number of kids per family ranged from 1 to around 15. The distribution appears to be bimodal, with quite a few families having 1-2 children, and 4-5 children. The family with around 15 kids appears to be an oulier because it is so much higher than the number of children in the other families.

Example 2

Because we wanted to do calculations based on family, it would be more convenient to put it in a tidy form. We calculated the number of children in each family.

Assignment

Explain what you discussed in your groups for question 3. You can use any figures you created.

- Were respondents familiar with reproducibility concerns in science?
- Were younger respondents more or less likely to report thinking that there is a reproducibility crisis in science? Why or why not?
- Is there variability in research reproducibility across scientific disciplines? If so, which disciplines are thought to be the most reproducible? The least?

Please submit before 4pm on Quercus!