EPSY 5261 : Introductory Statistical Methods

Day 5

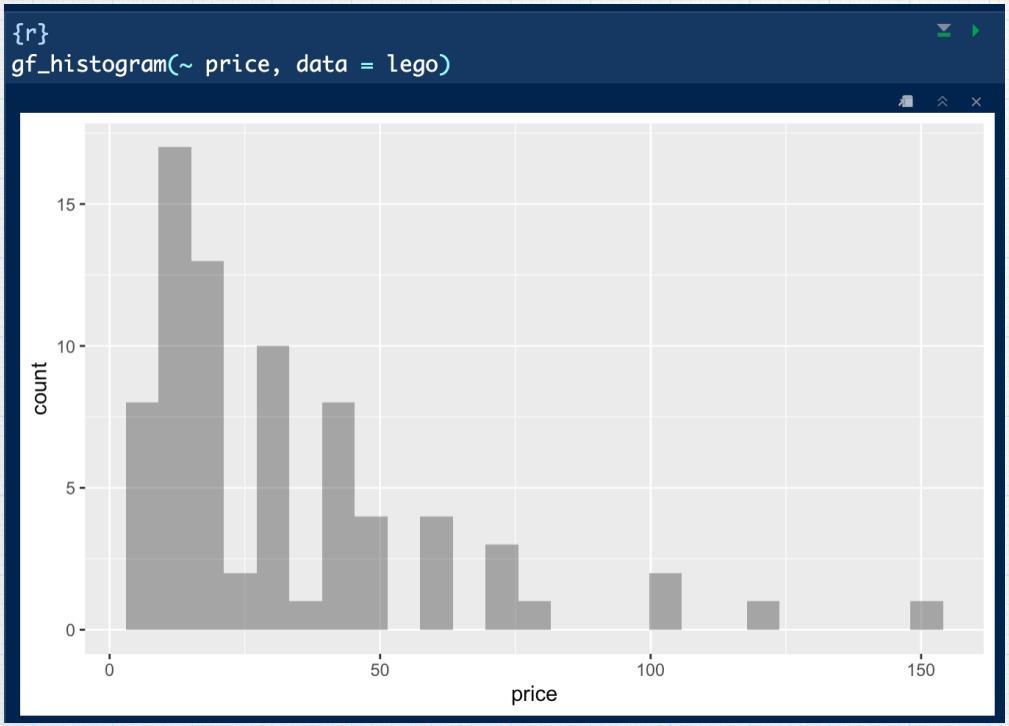
Visualization and Numerical Summaries (Part II)

Learning Goals

- At the end of this lesson, you should be able to…
  - Name and describe the three key features of a distribution
  - Identify and explain when to use the mean or median to describe the center of a distribution
  - Identify and explain when to use the standard deviation or IQR to describe the variability of a distribution
  - Describe distributions key features in the context of the data

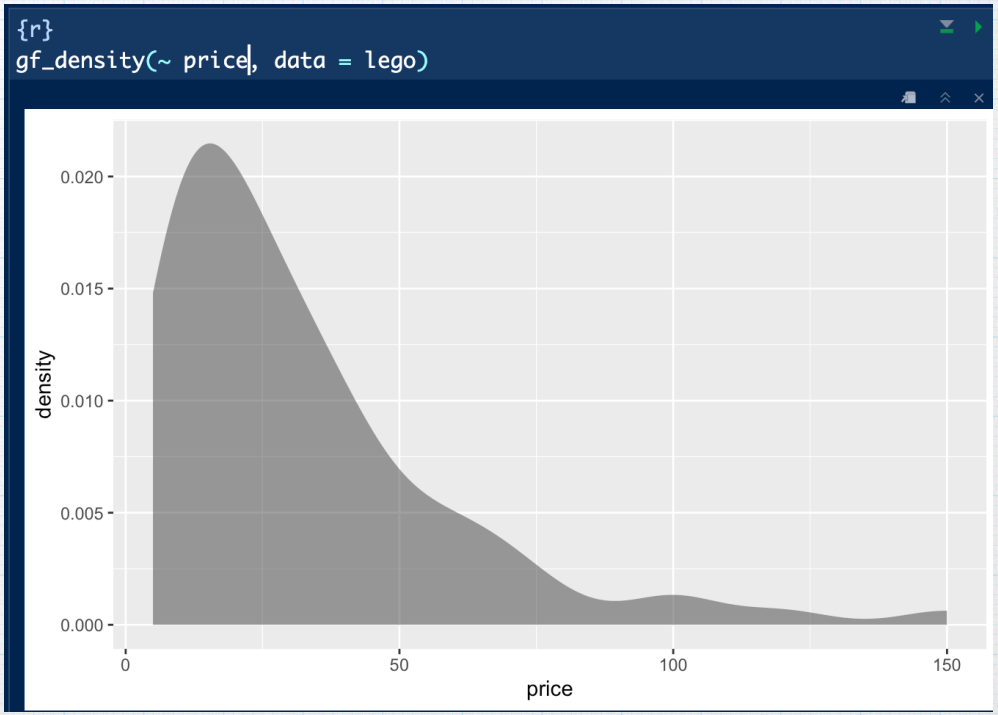
RECALL: Graphs for Quantitative Variables

```r
gf_histogram(~ price, data = lego)
```



Histogram of Price

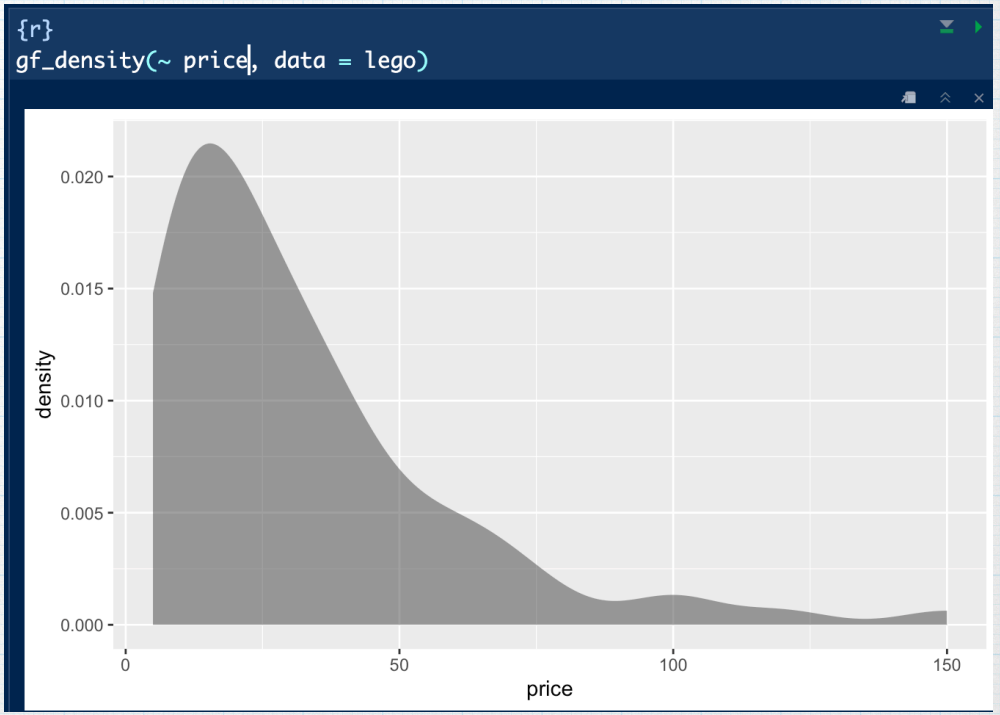
```r
gf_density(~ price, data = lego)
```



Density Plot of Price

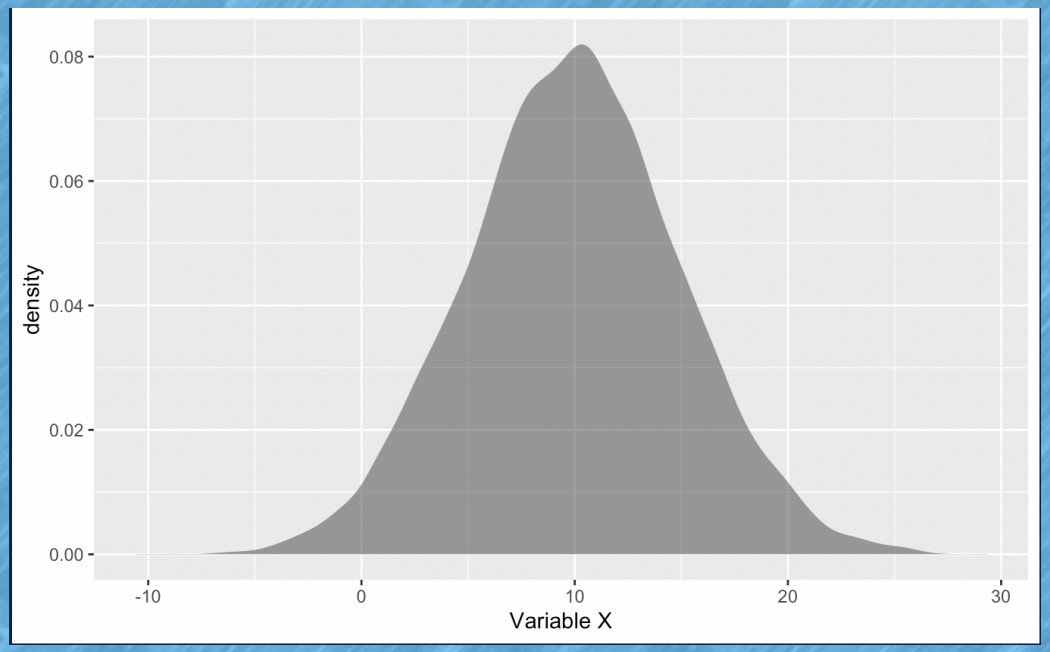
RECALL: Graphs for Quantitative Variables

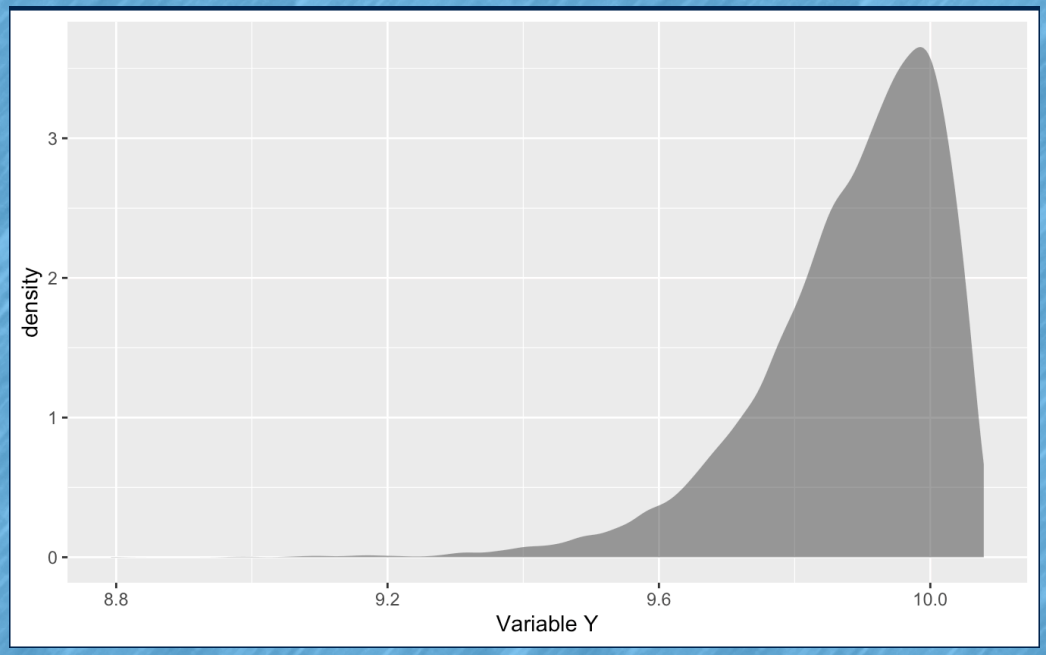
- These graphs allow us to see the **distribution** of the data
- We want to know what the variable "looks like"



Density Plot of Price

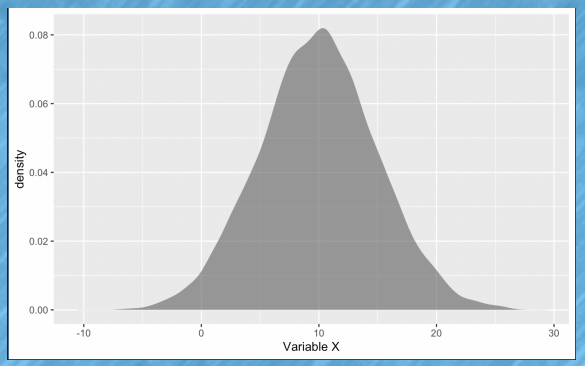
What is the big difference between these two graphs



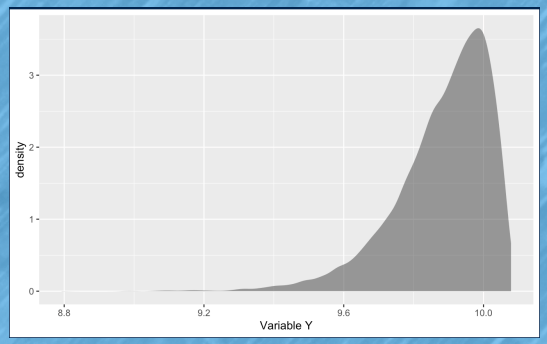


SHAPE

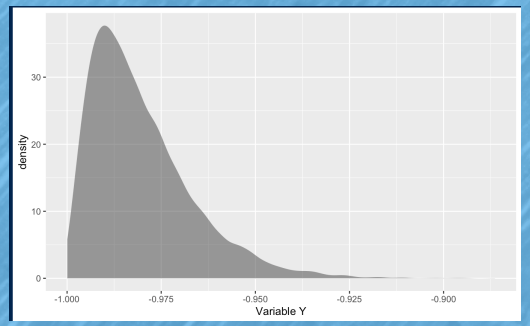
- Symmetric



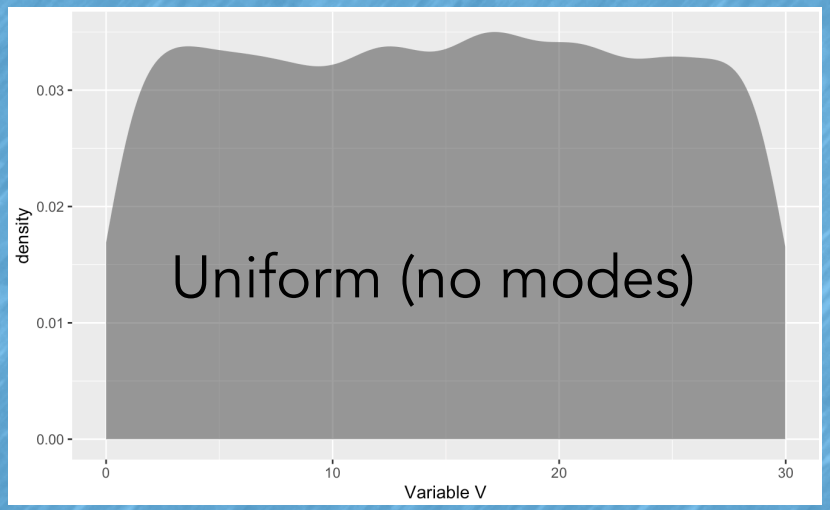
- Skewed Left (the tail is on the left)

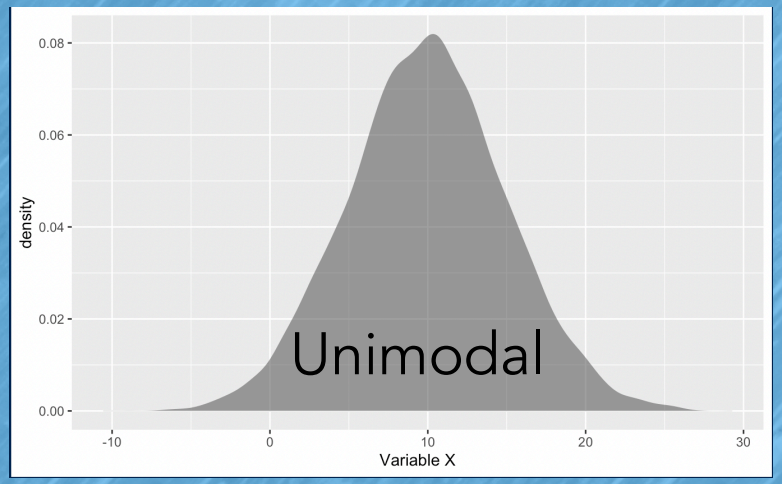


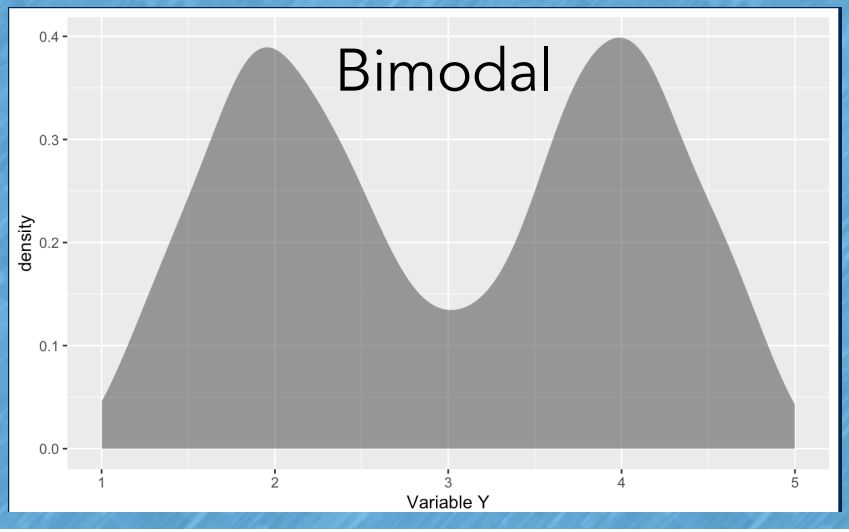
- Skewed Right (the tail is on the right)

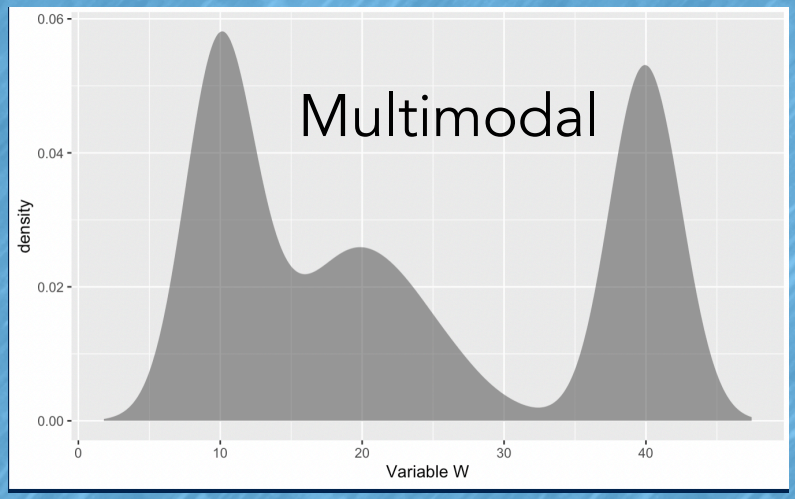


Modality

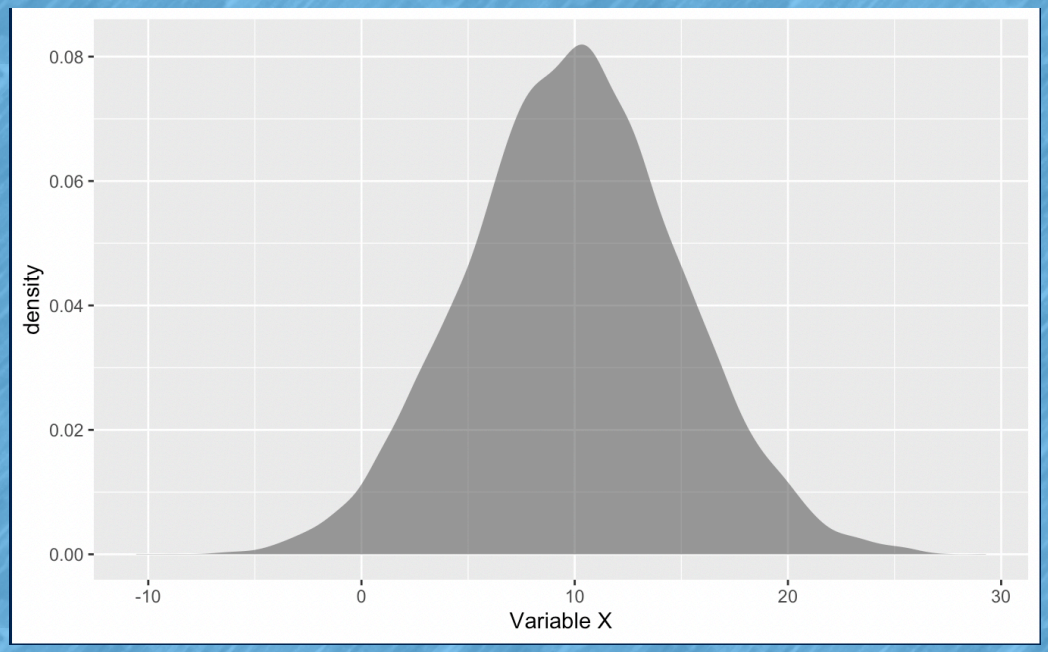

Uniform (no modes)
density
0.03
0.02
0.01
0.00
0
10
20
30
Variable V

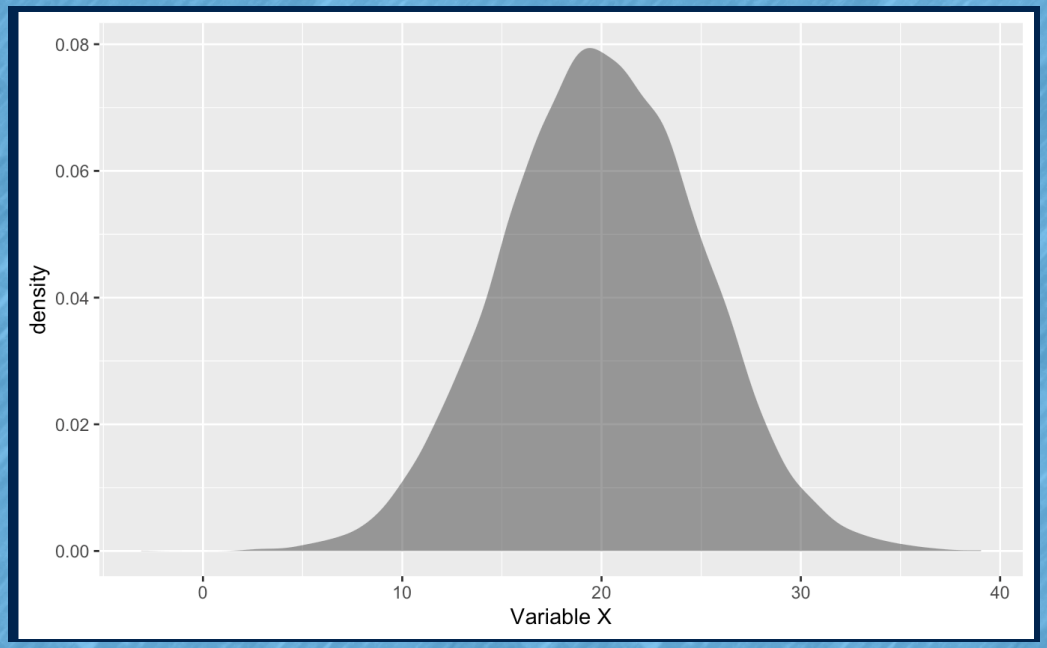

Unimodal
density
0.08
0.06
0.04
0.02
0.00
-10
0
10
20
30
Variable X


Bimodal
density
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
Variable Y


Multimodal
density
0.06
0.04
0.02
0.00
0
10
20
30
40
Variable W

What is the big difference between these two graphs?



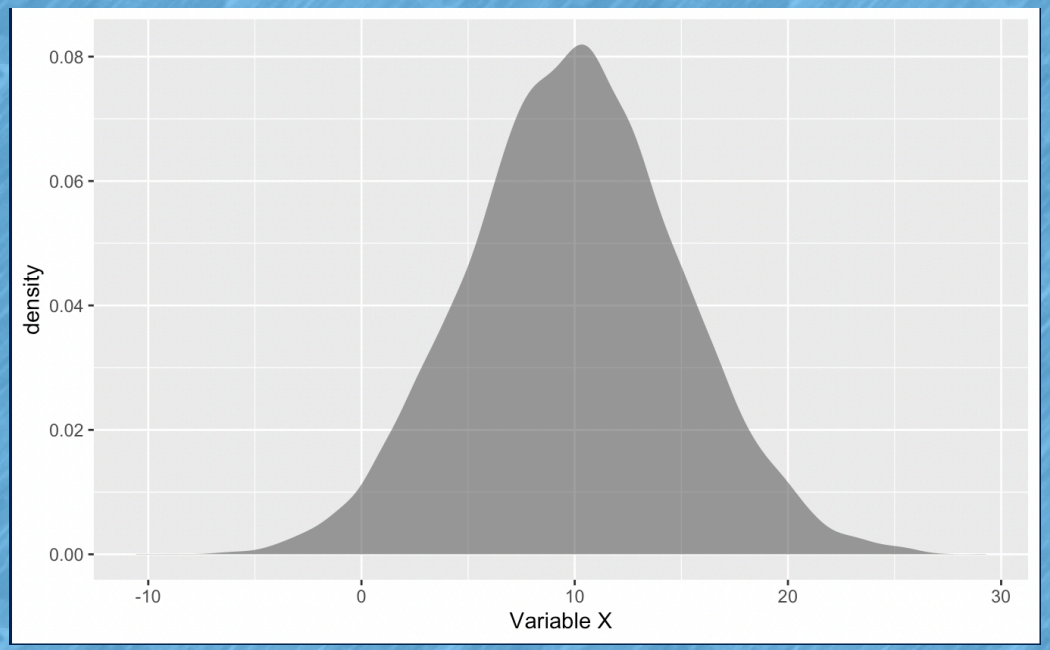


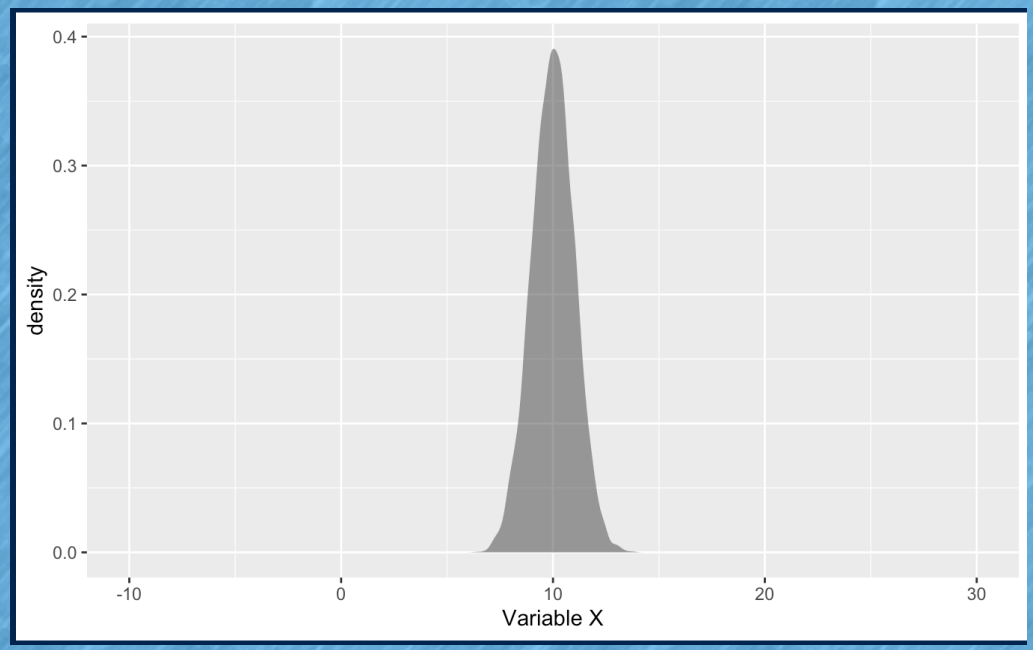
CENTER

Center

- A point where "most" of the data is located
- Often measured by the:
  - **Mean:** average of all data points
  - **Median:** middle point of the data (if put in numerical order)

What is the big difference between these two graphs?





VARIABILITY

Variability

- A measure of how distinct the points in your dataset are
- Often measured by the:
  - **Range:** the distance between the smallest and largest data value
  - **Standard deviation:** a measure of the average distance each point it from the mean
  - **IQR:** an interval outlining the the middle 50% of the data

Describing Distributions Activity

Summary

- The three key features of a distribution
  - Shape: symmetric/skewed? Unimodal/bimodal?
  - Center: at what single point is most of the data?
  - Variability: what range best describes where the majority of the data falls?
- When to use the mean or median to describe the center of a distribution
  - If we have a skewed distribution we will use the median to describe the center
  - If we have a symmetric distribution we could use either mean or median, but the mean is the most often used
- When to use the standard deviation or IQR to describe the variability of a distribution
  - If we have a skewed distribution we will use the IQR to describe the variability
  - If we have a symmetric distribution we could use either IQR or standard deviation, but typically we use the standard deviation