Online CDS 101 *Introduction to Computational and Data Sciences*

# Data visualization

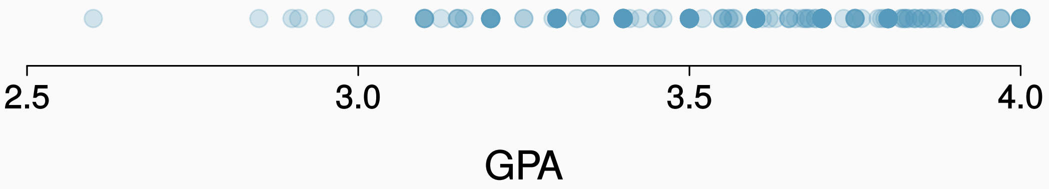
Examining numerical data

# Describing shapes of numerical distributions

- shape:
  - skewness: right-skewed, left-skewed, symmetric (skew is to the side of the longer tail)
  - modality: unimodal, bimodal, multimodal, uniform
- center: mean, median, mode (not always useful)
- spead: range, standard deviation, inter-quartile range
- unusual observations

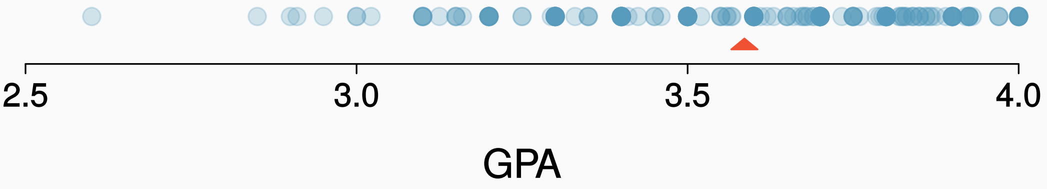
# Dot plots

Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.



How would you describe the distribution of GPAs in this data set? Make sure to say something about the center, shape, and spread of the distribution.

# Dot plots & mean



- The **mean**, also called the **average** (marked with a triangle in the above plot), is one way to measure the center of a **distribution** of data.

- The mean GPA is 3.59.

# Mean

- The **sample mean**, denoted as $\bar{\mathbf{x}}$, can be calculated as

  $$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

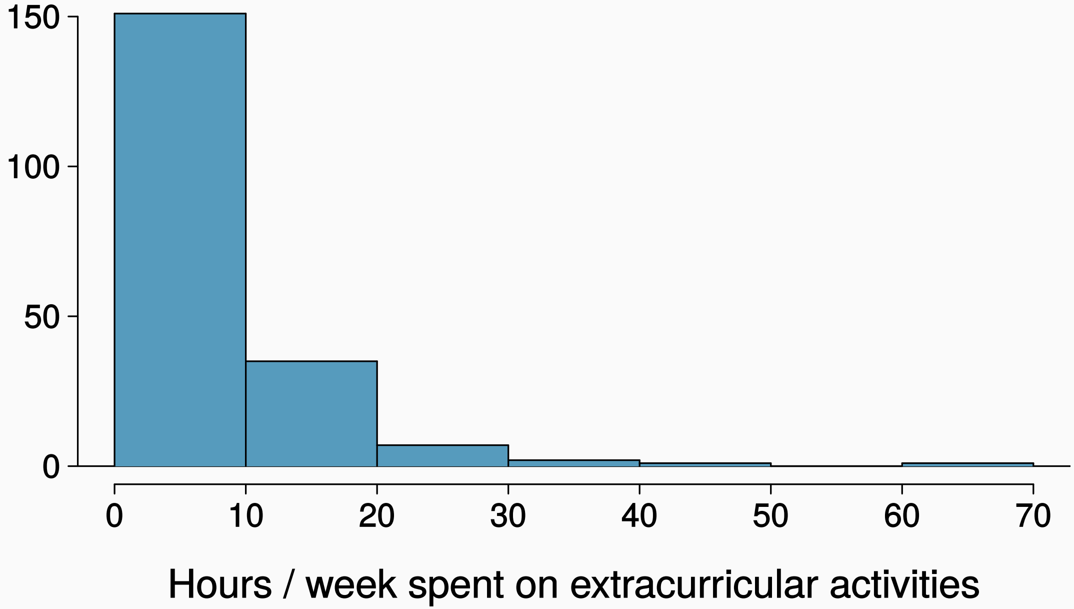
  where $x_1, x_2, \cdots, x_n$ represent the **n** observed values.

- The **population mean** is also computed the same way but is denoted as $\mu$. It is often not possible to calculate $\mu$ since population data are rarely available.

- The sample mean is a **sample statistic**, and serves as a **point estimate** of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

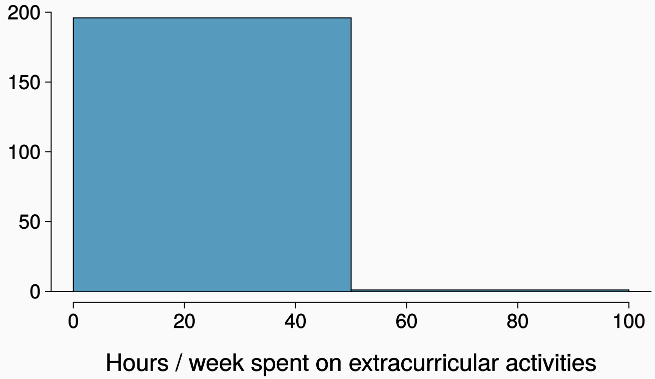
# Histograms

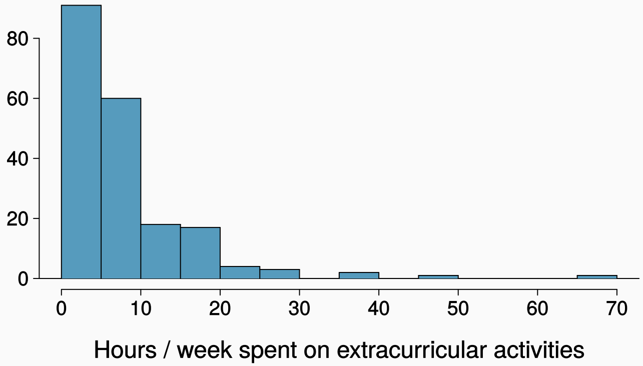
- Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the **shape** of the data distribution.
- The chosen **bin width** can alter the story the histogram is telling.

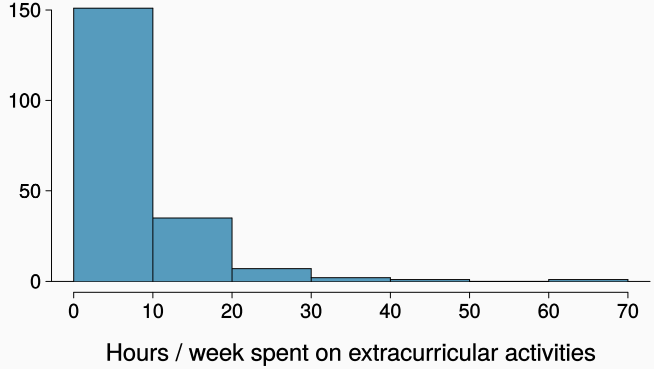


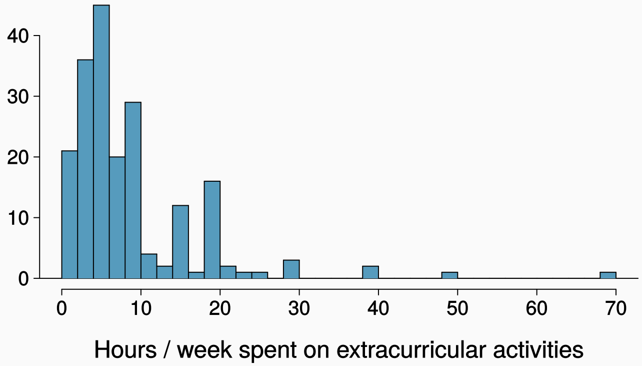
# Bin width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



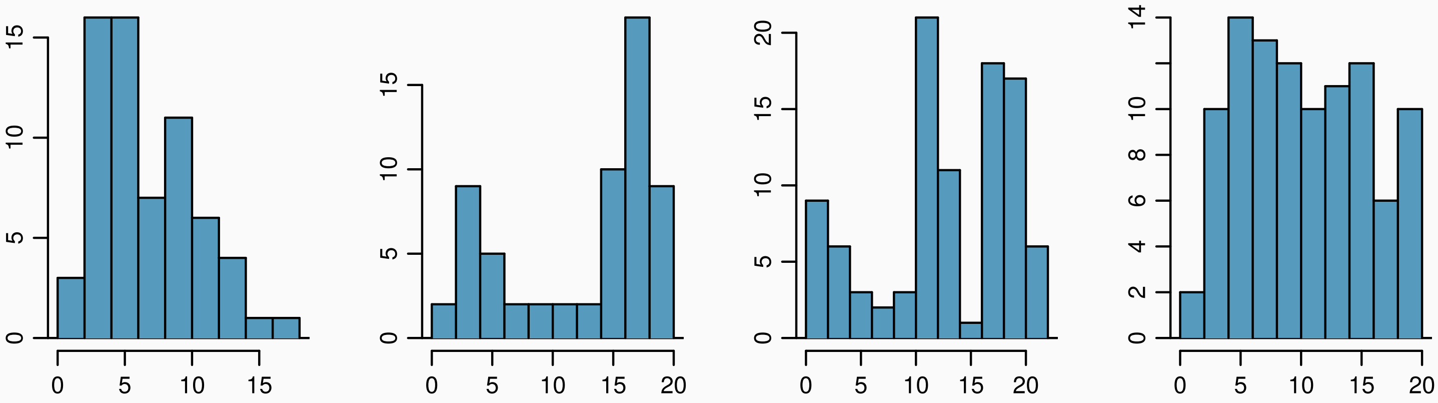






# Shape of a distribution: modality

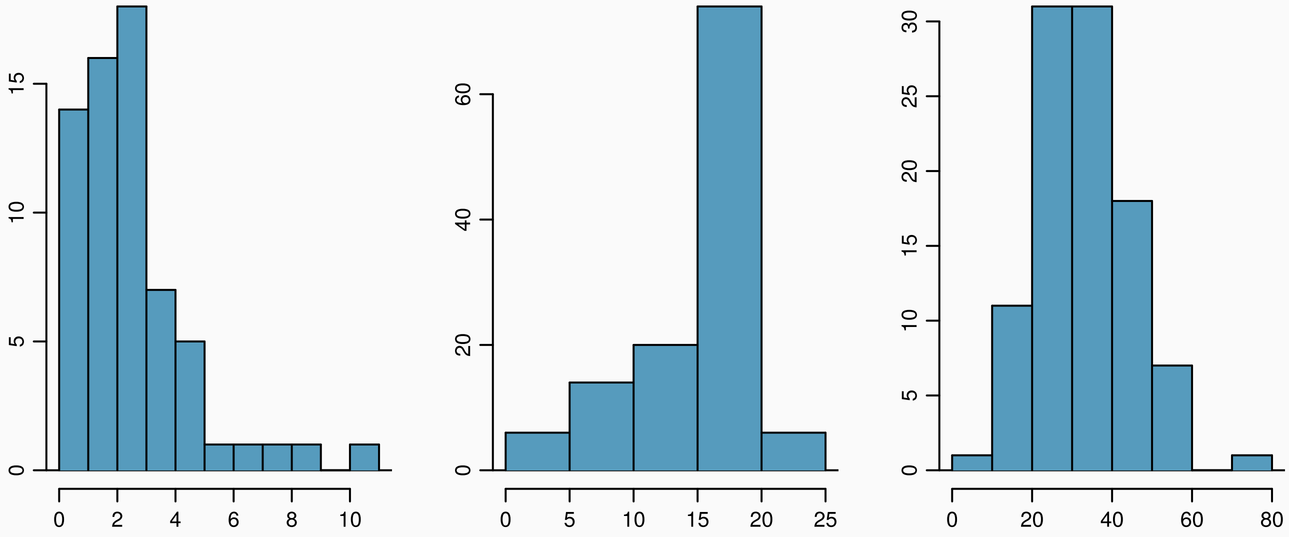
Does the histogram have a single prominent peak (unimodal), several prominent peaks (bimodal/multimodal), or no apparent peaks (uniform)?



In order to determine modality, step back and imagine a smooth curve over the histogram – imagine that the bars are wooden blocks and you drop a spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

# Shape of a distribution: skewness

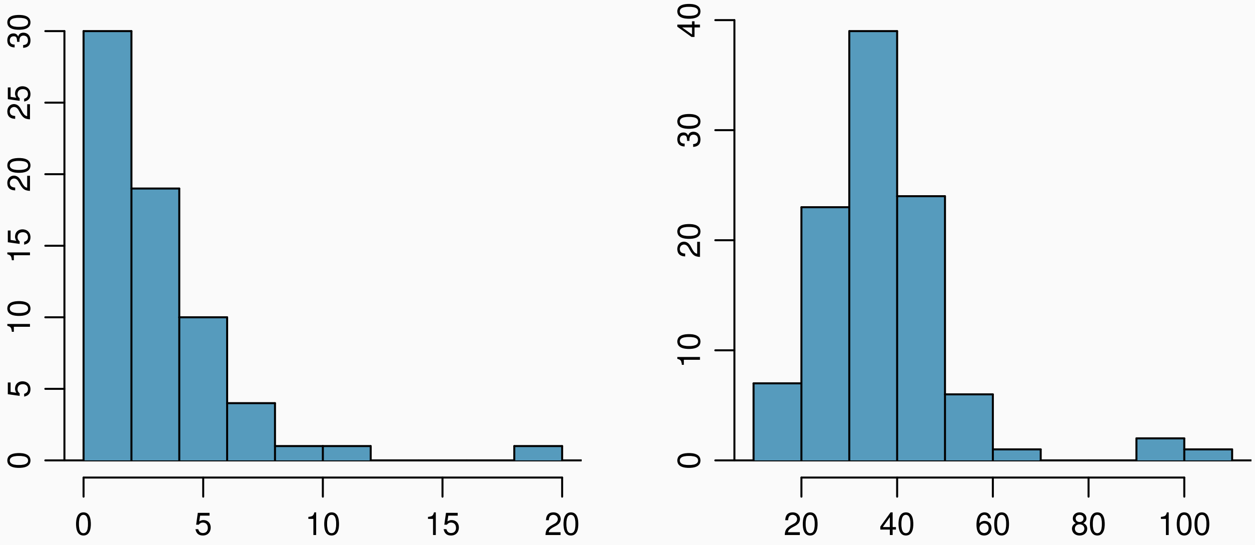
Is the histogram right skewed, left skewed, or symmetric?



Histograms are said to be skewed to the side of the long tail.

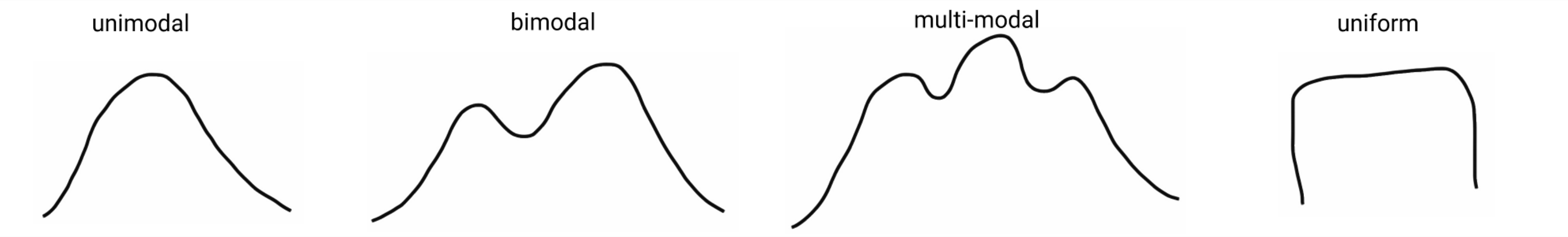
# Shape of a distribution: unusual observations

Are there any unusual observations or potential outliers?

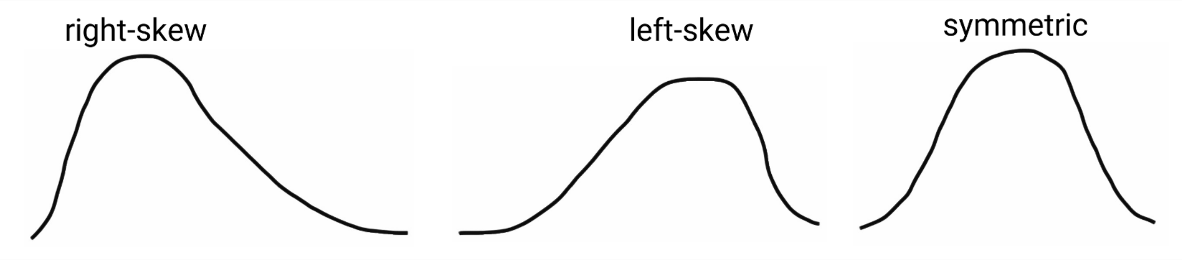


# Commonly observed shapes of distributions

Modality

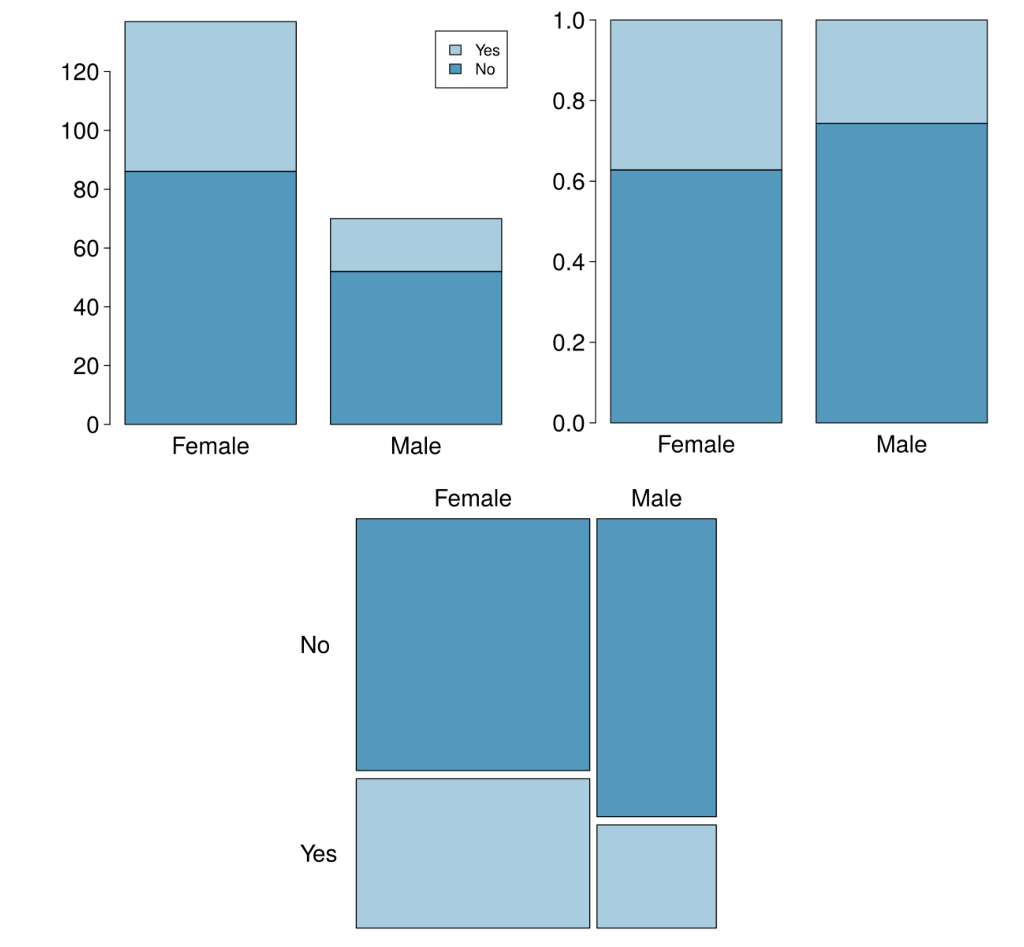


Skewness



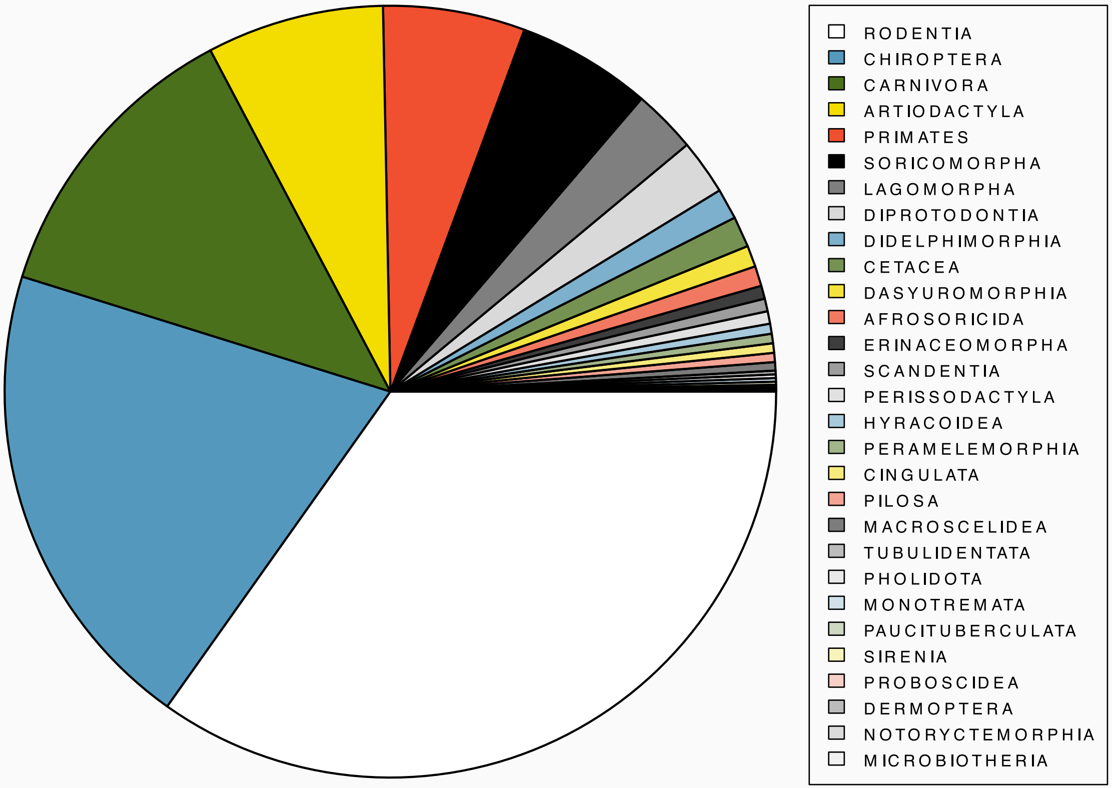
# What about categorical data?

Segmented bar and mosaic plots



# Pie charts

Can you tell which order encompasses the lowest percentage of mammal species?



Source: Data from www.departments.bucknell.edu/biology/resources/msw3/

# Credits

License

Acknowledgments

Content adapted from the Chapter 1 OpenIntro Statistics Slides developed by Mine Cetinkaya-Rundel and made available under the CC BY-SA 3.0 license.