Class 06

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-09-11

Packages Used Today

NONE!

Numerical Variables

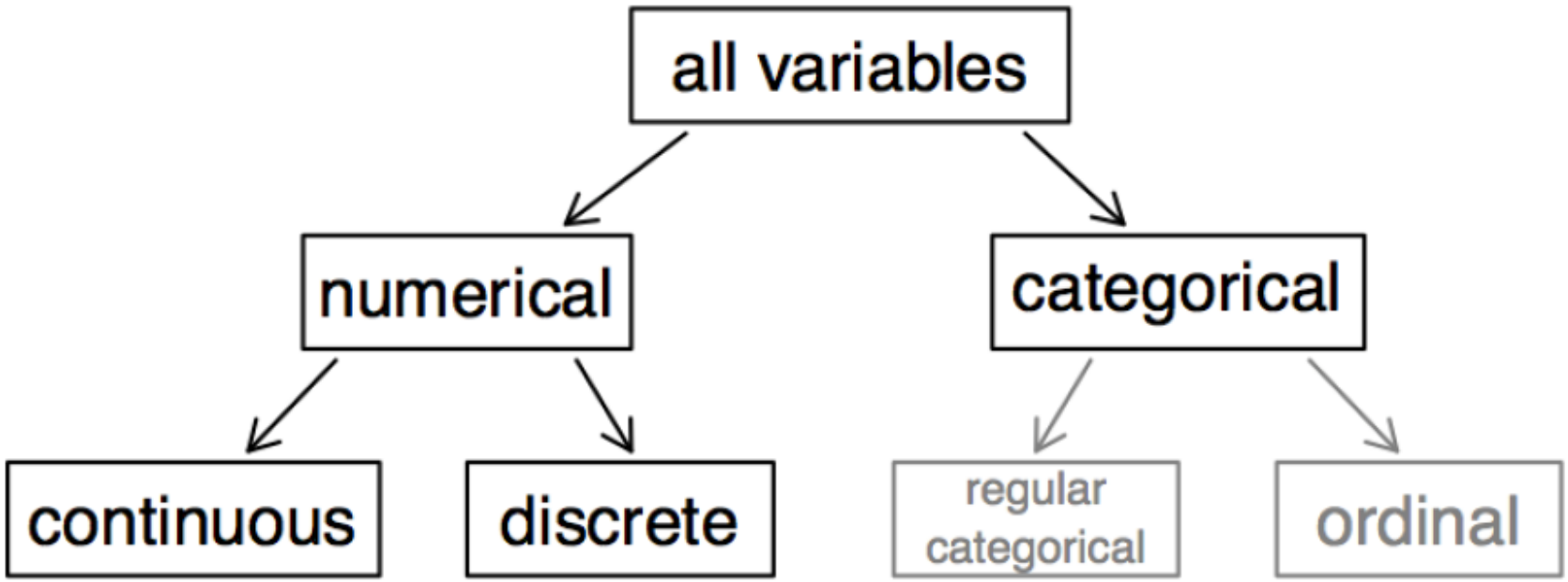


Figure 1: Numerical variables can be continuous or discrete.

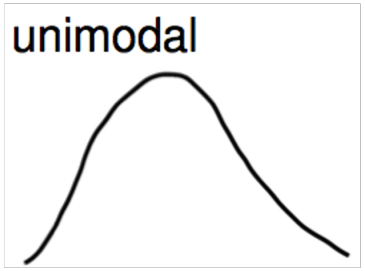
Describing numerical distributions

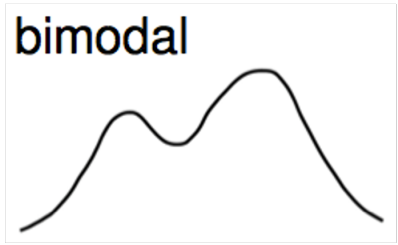
The "shape" of numerical data is called its **distribution**.

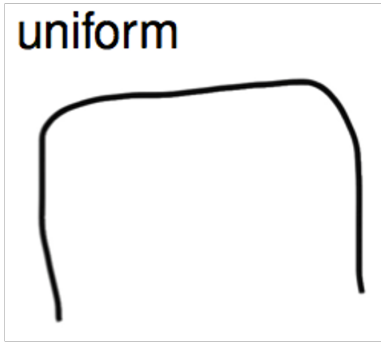
- **Location:** the "center" of the data
  - The value(s) around which most observations are clustered
- **Scale:** the "spread" of the data
  - How variable the observations are around that "center"

Describing distribution shapes

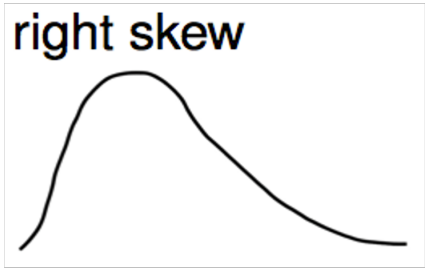
Modality

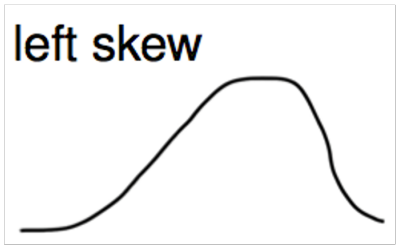






Skewness





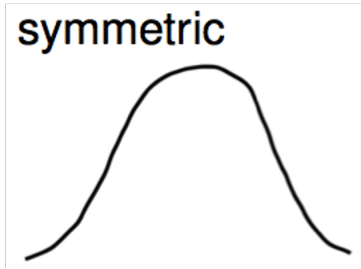


Figure 2: Commonly observed patterns in numerical distributions

Describing a distribution's **location**

The **location** of a numerical variable's distribution can be thought of as the "center" of the data, around which the bulk of the observations cluster.

- **Mean:** the sum of a values divided by the number of observations (i.e. "average")
- **Median:** the value in the exact middle of the data
- **Mode:** the most common value in the data (for discrete variables)

Describing a distribution's **scale**

How far is each data value from the mean?

- **Variance:** s^2, the sum of the squared differences between each observation's value and the sample mean $\bar{x}$ divided by $n - 1$
- **Standard deviation:** s, the square root of the variance
- **Range:** minimum to maximum
- **Interquartile Range (IQR):** 25th percentile to 75th percentile, the middle 50% of the data

Robust statistics

The **median** and **interquartile range** are considered to be **robust statistics** for the numerical summary of data because they are less sensitive to **skew** and **outliers** than the **mean**, **variance**, and **standard deviation**.

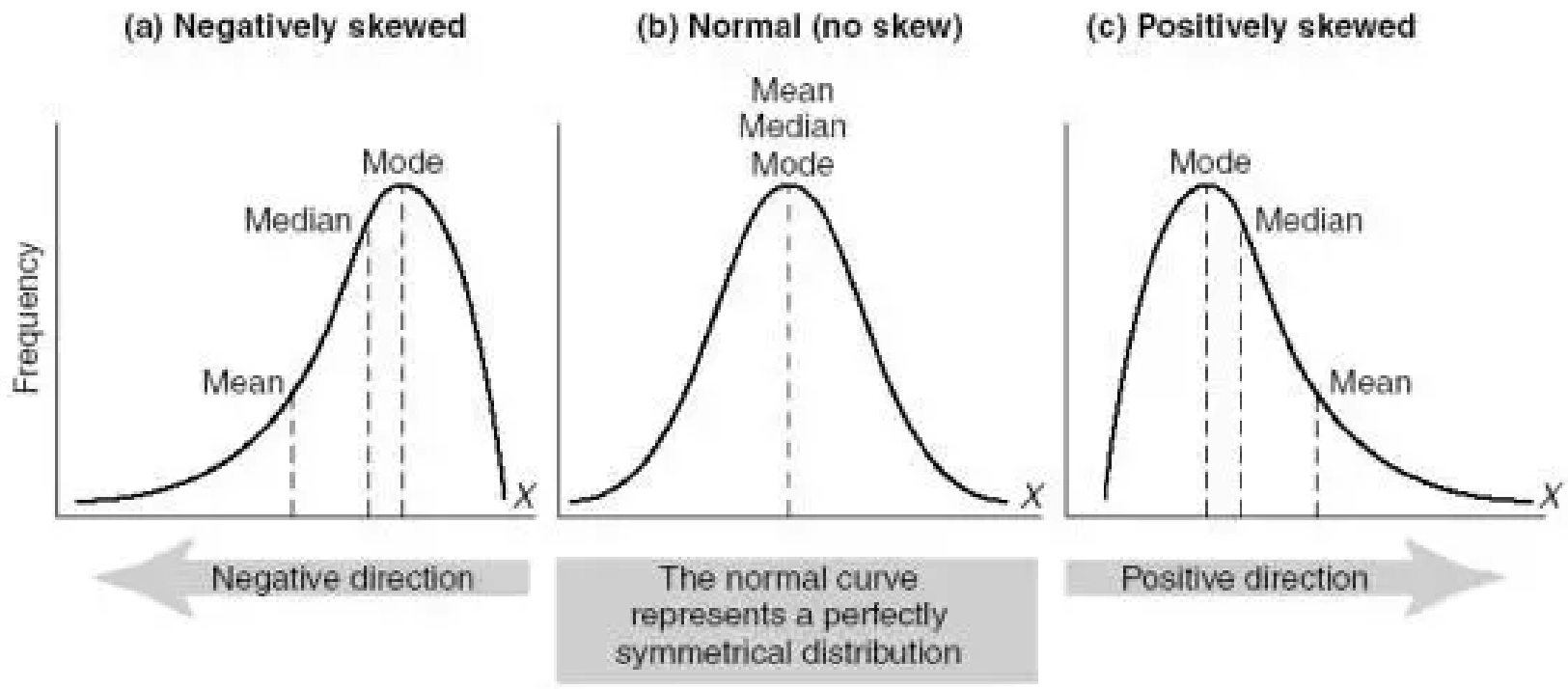


FIGURE 15.6 Examples of normal and skewed distributions

Figure 3: The presence of outliers and/or skew in a numerical variable's distribution affects how well summary statistics describe a distribution's

5-Number Summary of Numerical Data

1. Minimum value
2. 1st quartile (Q1, 25th percentile)
3. Median (Q2, 50th percentile)
4. 3rd quartile (Q3, 75th percentile)
5. Maximum value

Choosing Summary Statistics for Numerical Data

- The **mean** and **standard deviation** are really only appropriate for a certain type of unimodal, symmetric distribution called the **normal distribution** and often misused
- Most real world data will be best described by the **median** and **interquartile region** as part of a 5-number summary

The Normal Distribution

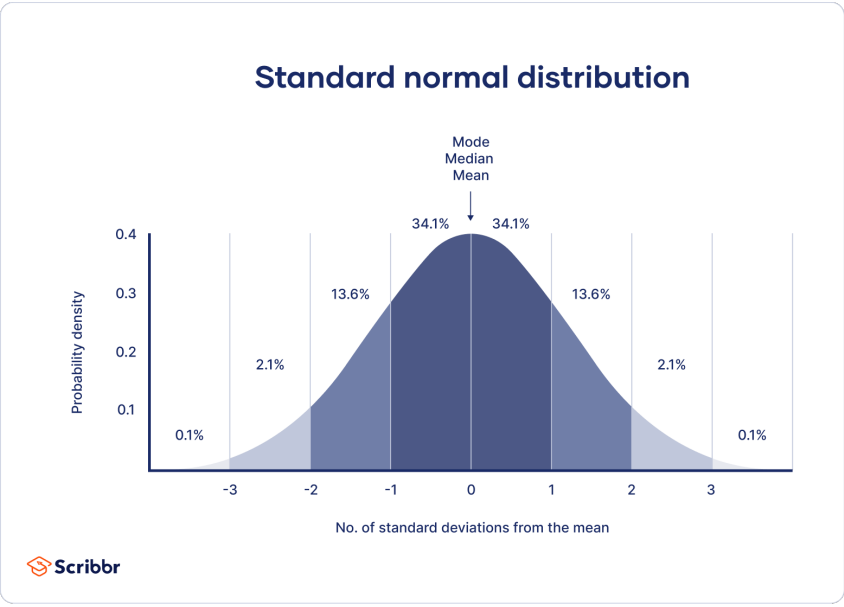
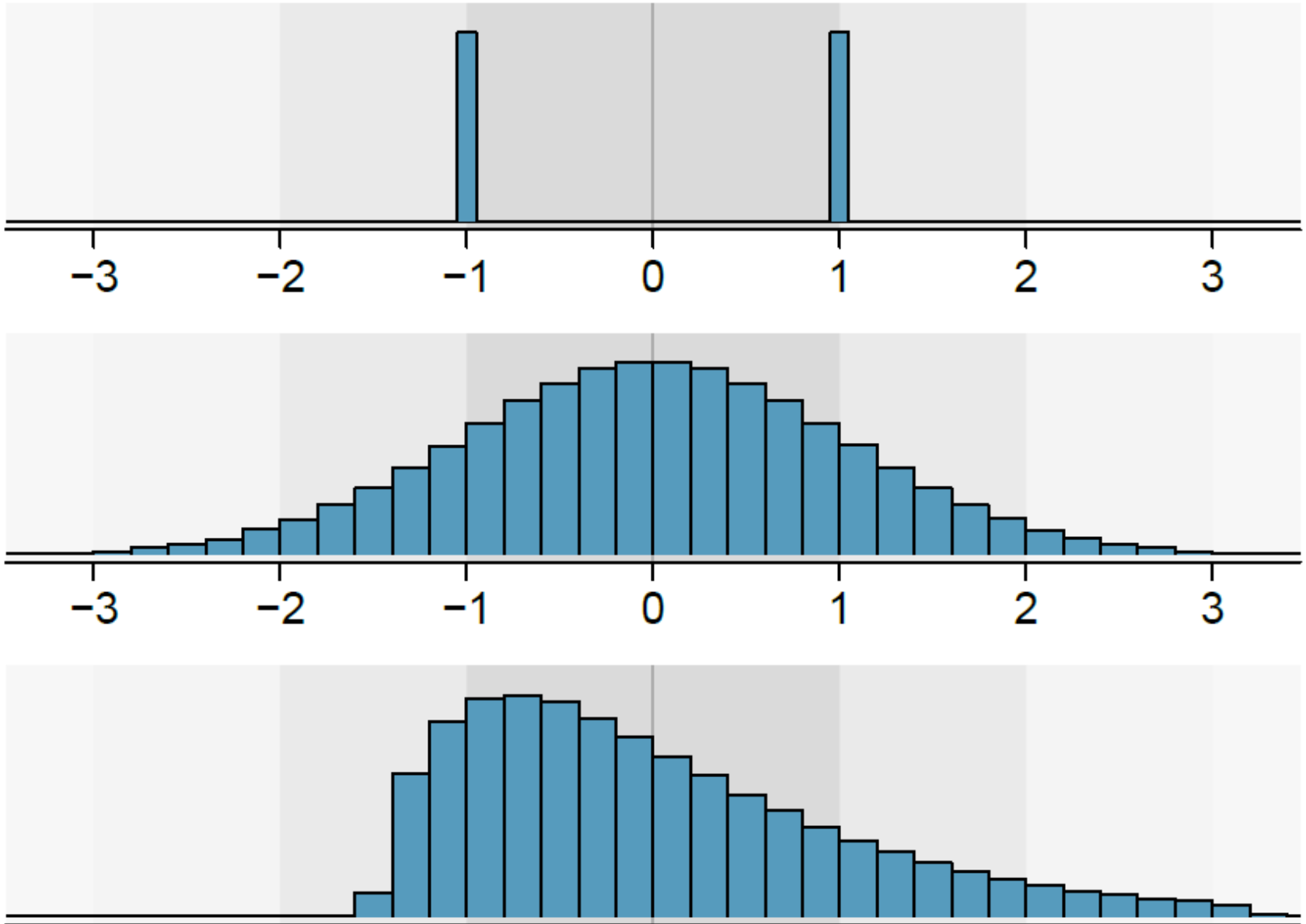


Figure 4: The percentage of the observations which fall within +/- 1, 2, and 3 standard deviations from the mean when data is normally distributed.

- **Normal distributions** are unimodal and symmetric
- The **mean** and the **median** of normally distributed data will be approximately equal
- **Normally distributed** variables are desirable in statistics but rare in practice

Using the mean +/- standard deviation to describe non-normal distributions



Visualizing Numerical Data

- Dot plot
- Histogram
- Density Curve
- Boxplot
- Violin plot
- QQ plot

How to Read a Dot Plot

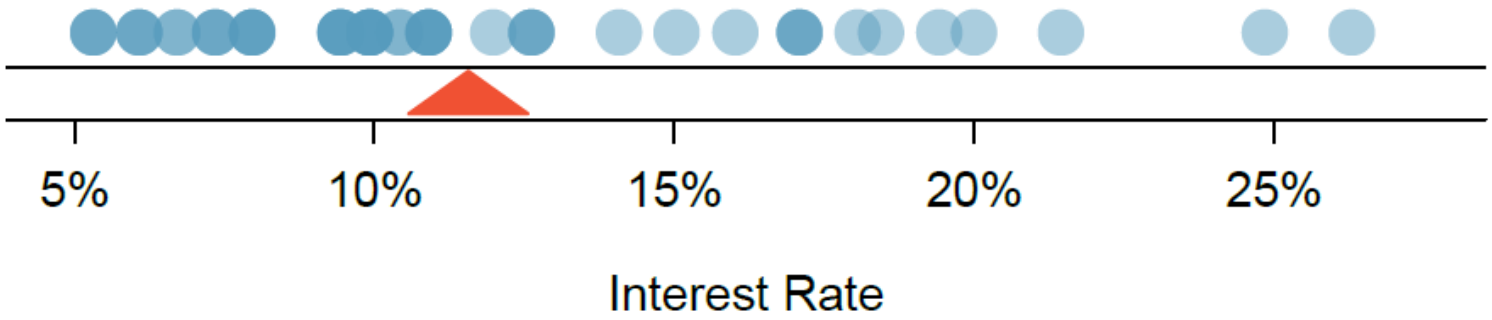


Figure 6: There is a single axis (x) along with a dot marking each data point. The points are usually slightly transparent, so you can see when points are overlapping.

How to read a stacked dot plot

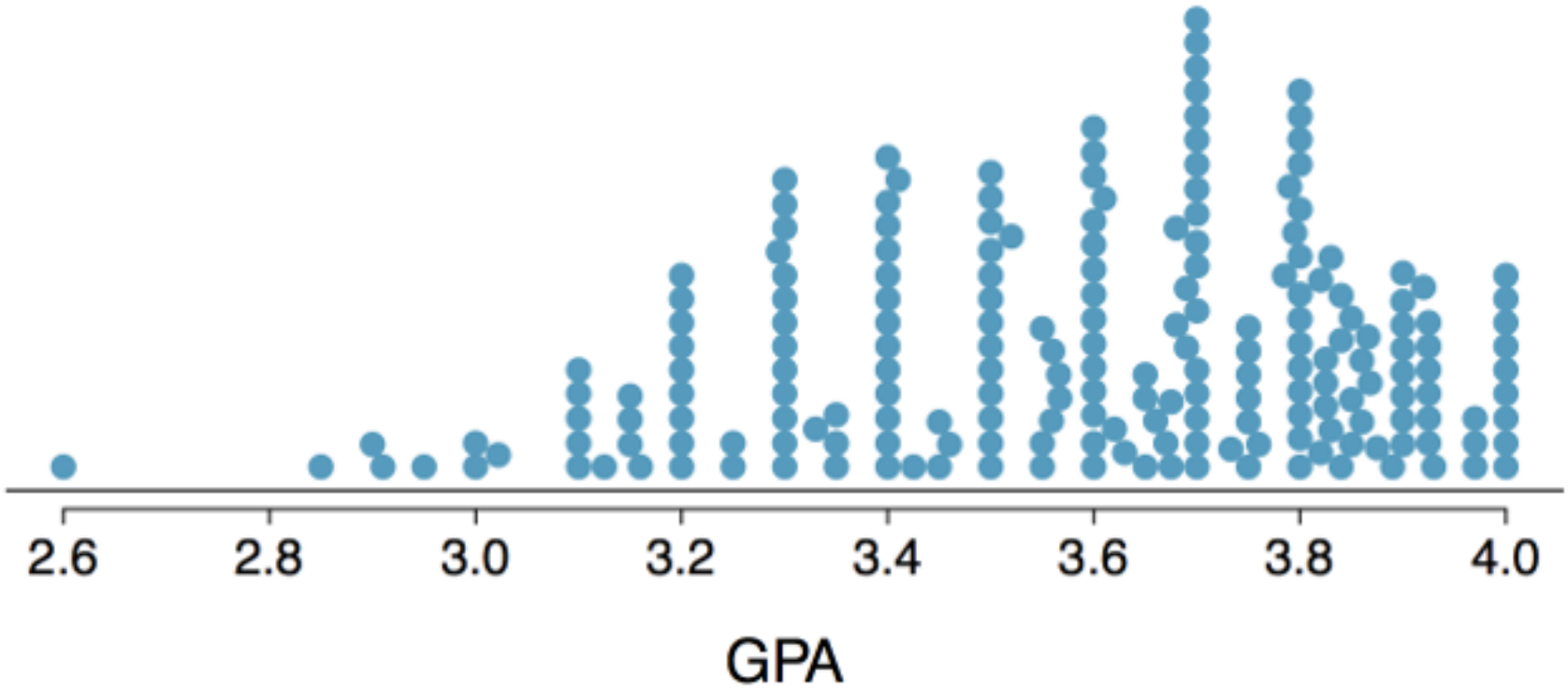
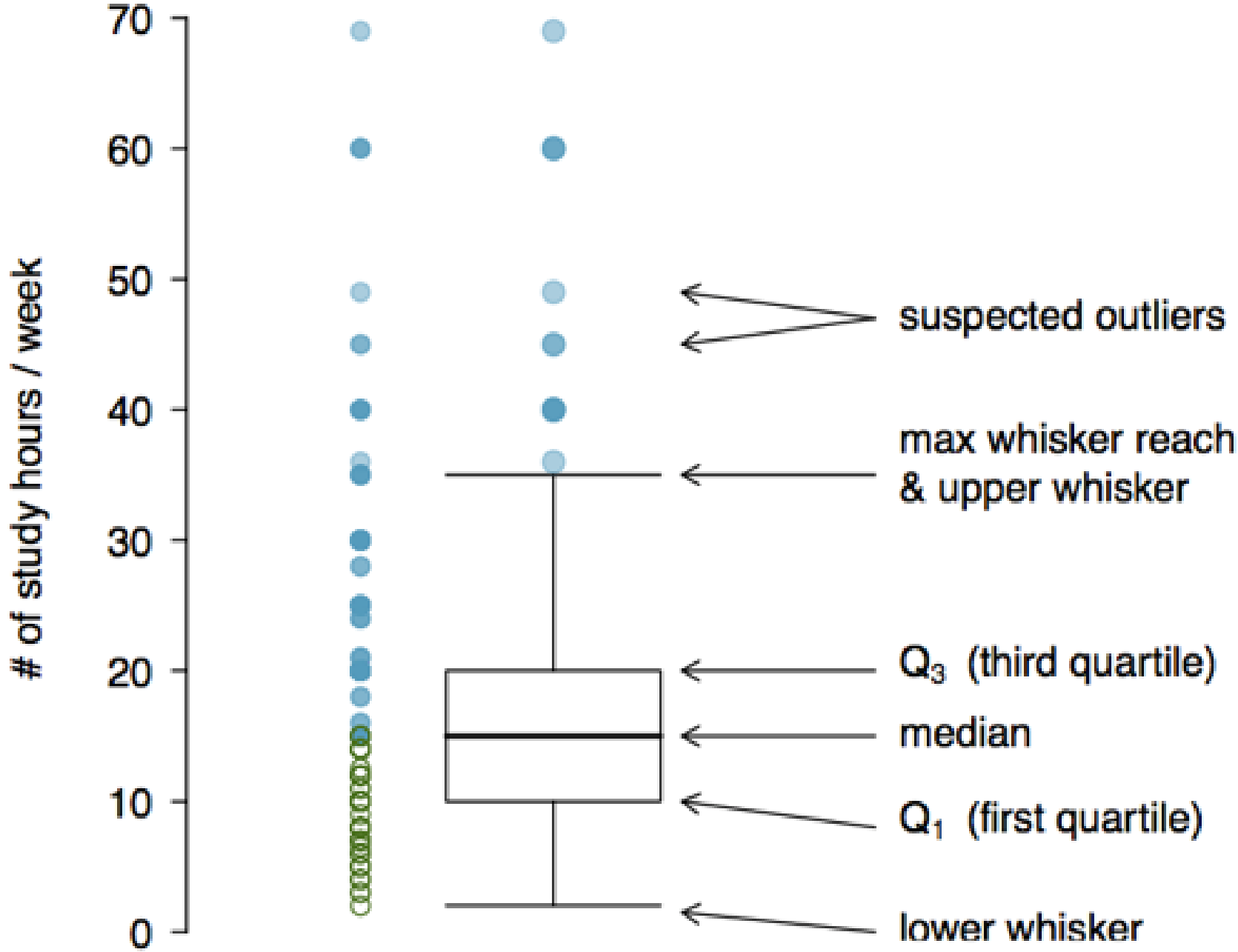


Figure 7: In a stacked dot plot, multiple observations at a single value are stacked on top of each other.

How to Read a Histogram



Histograms for different distributions

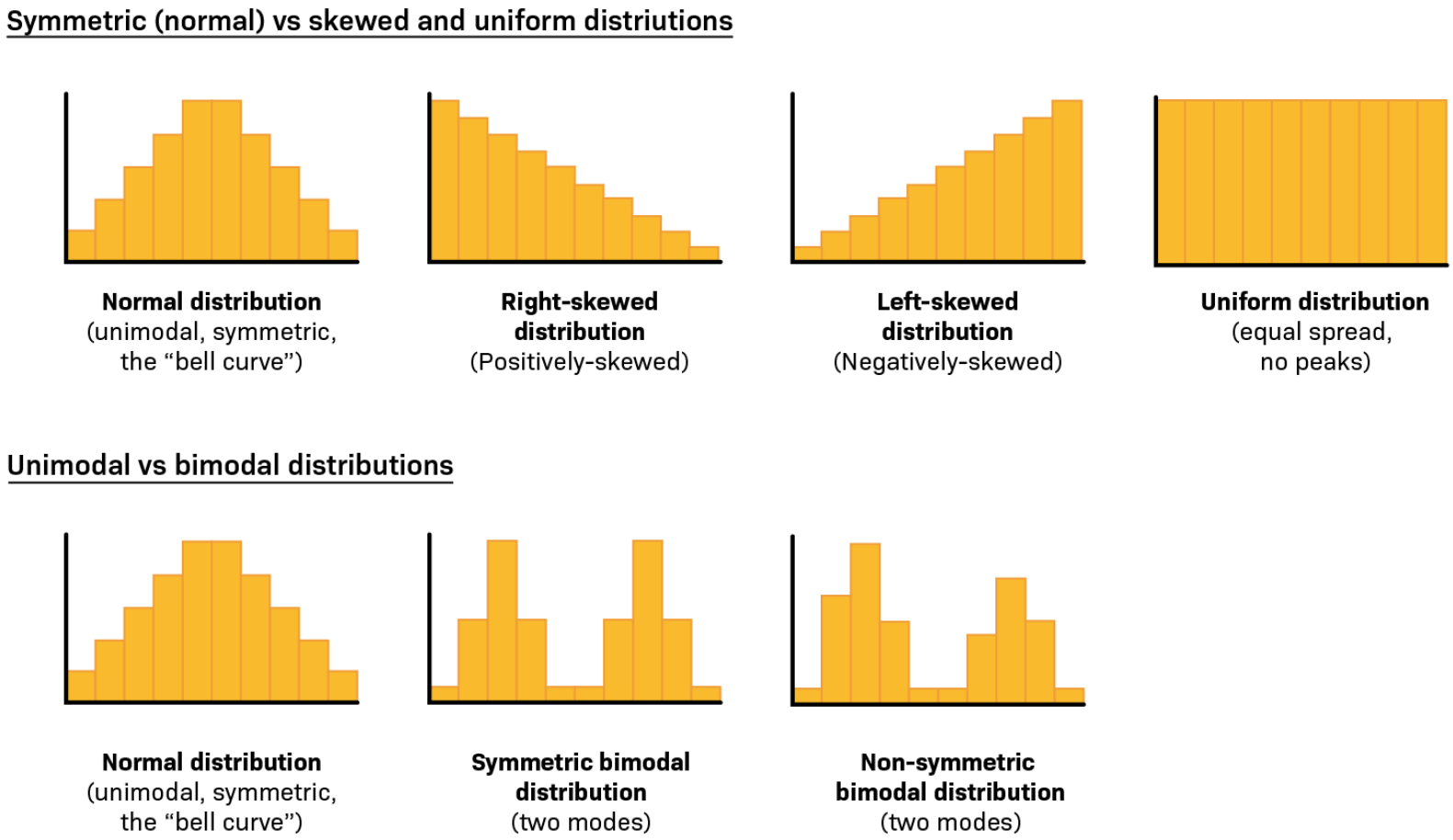


Figure 9: Examples of the different distribution shapes as histograms

Histograms and skew

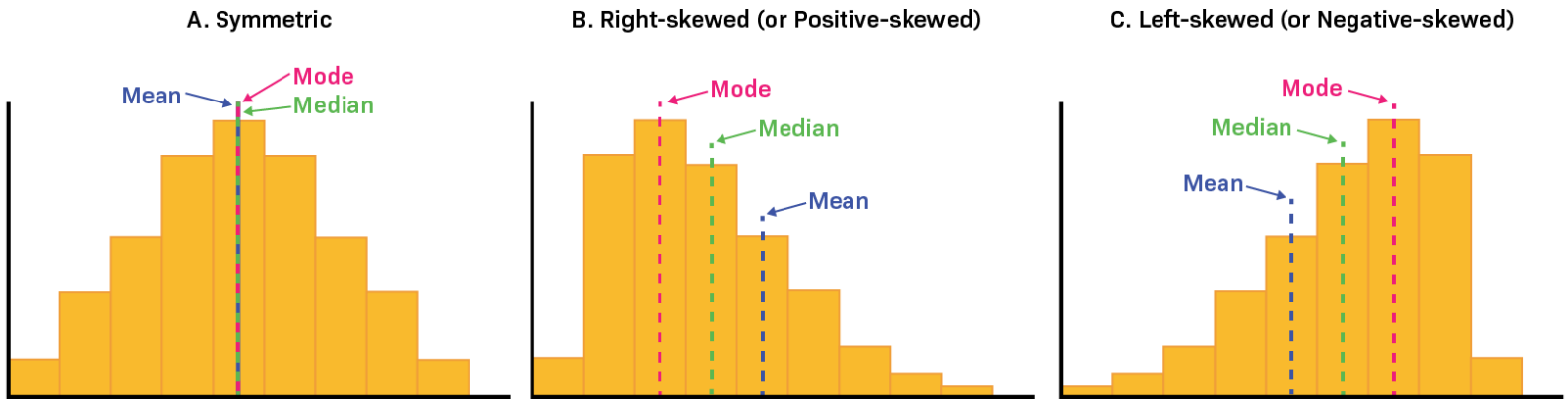


Figure 10: When histograms are skewed, the mean and the median may occur in 2 different bins.

Histograms and outliers

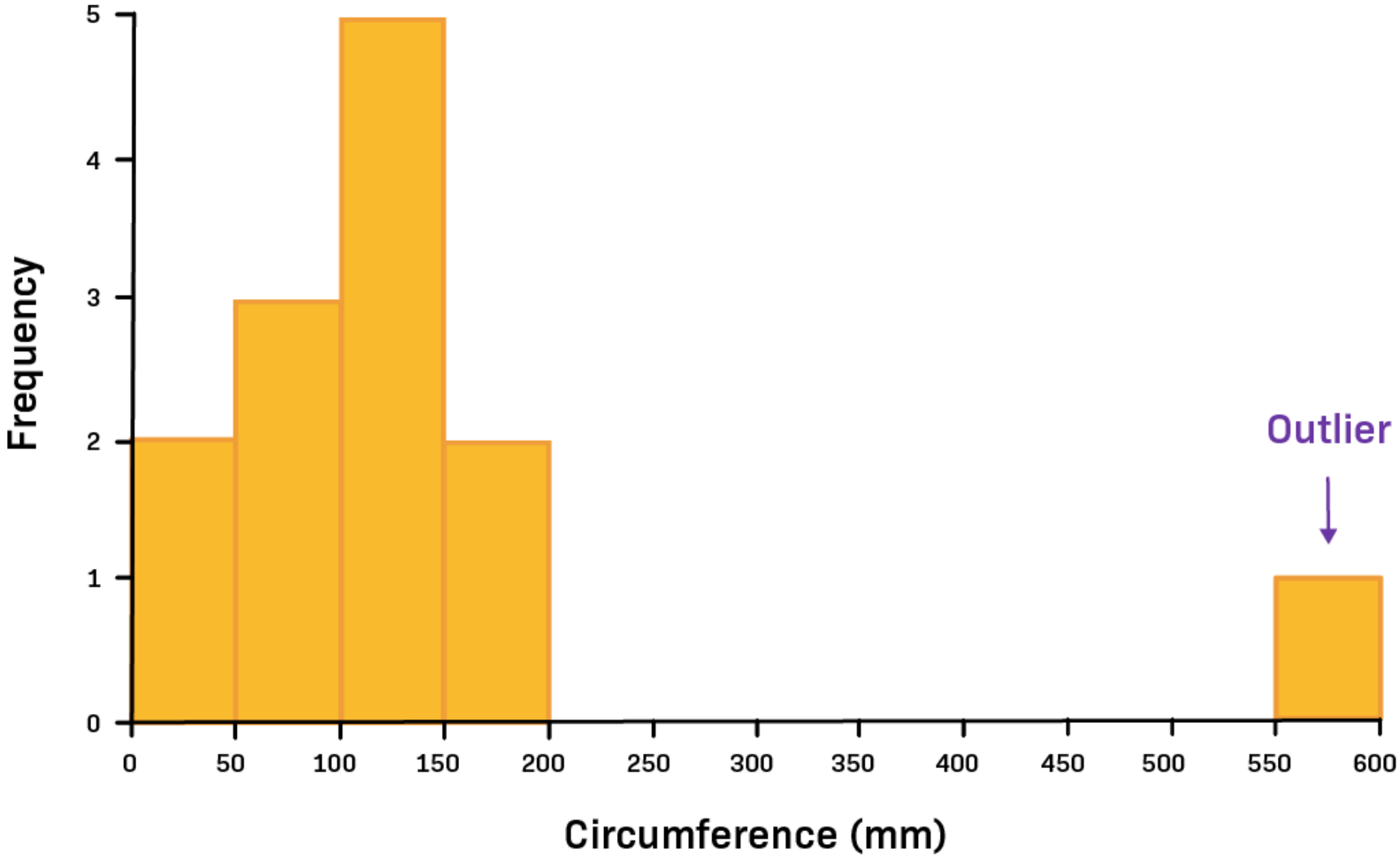


Figure 11: Outliers are easy to spot on a histogram

Histograms and modality

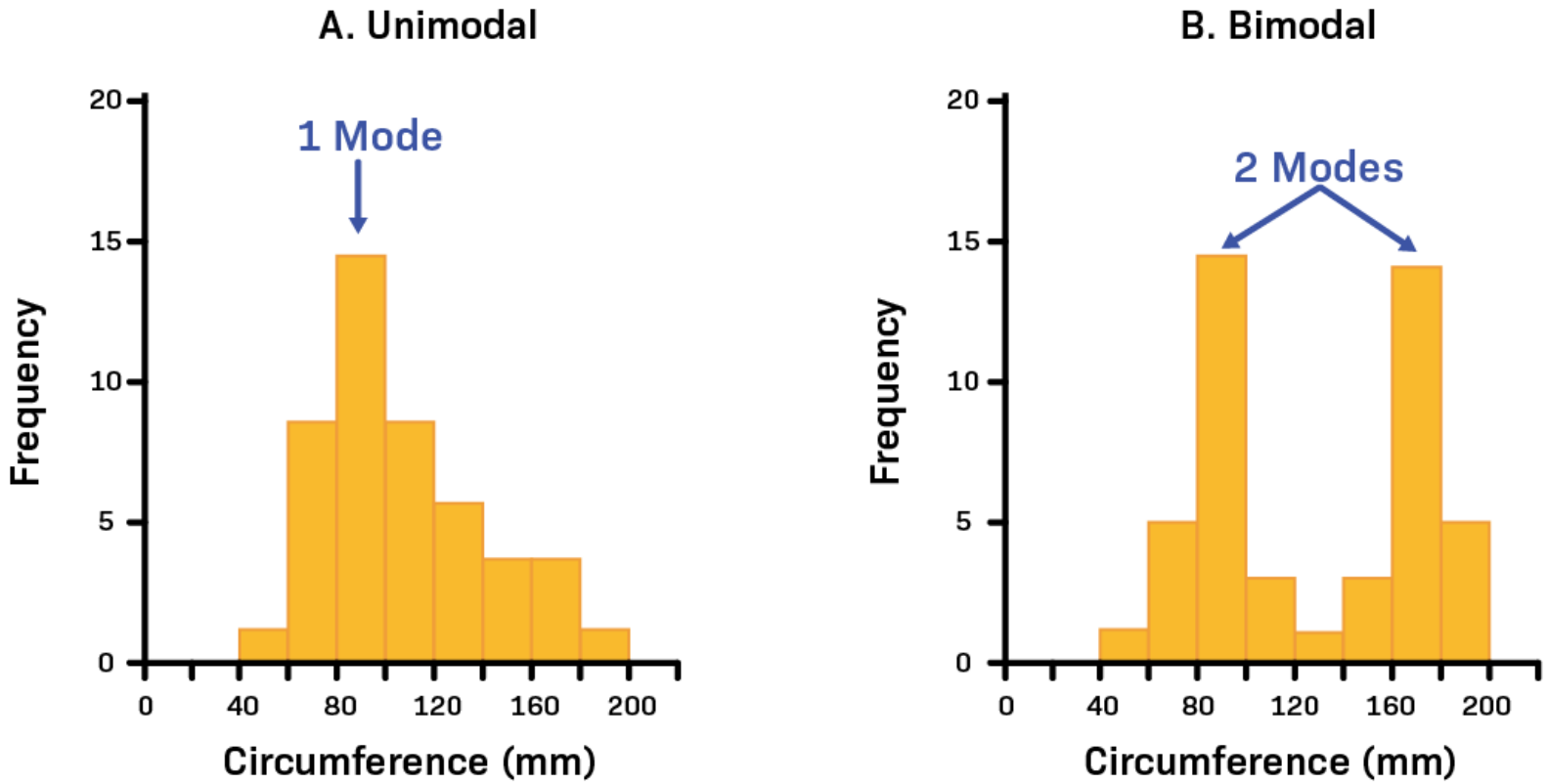


Figure 12: Modality is easy to spot on a histogram.

Choosing a bin width for your histogram

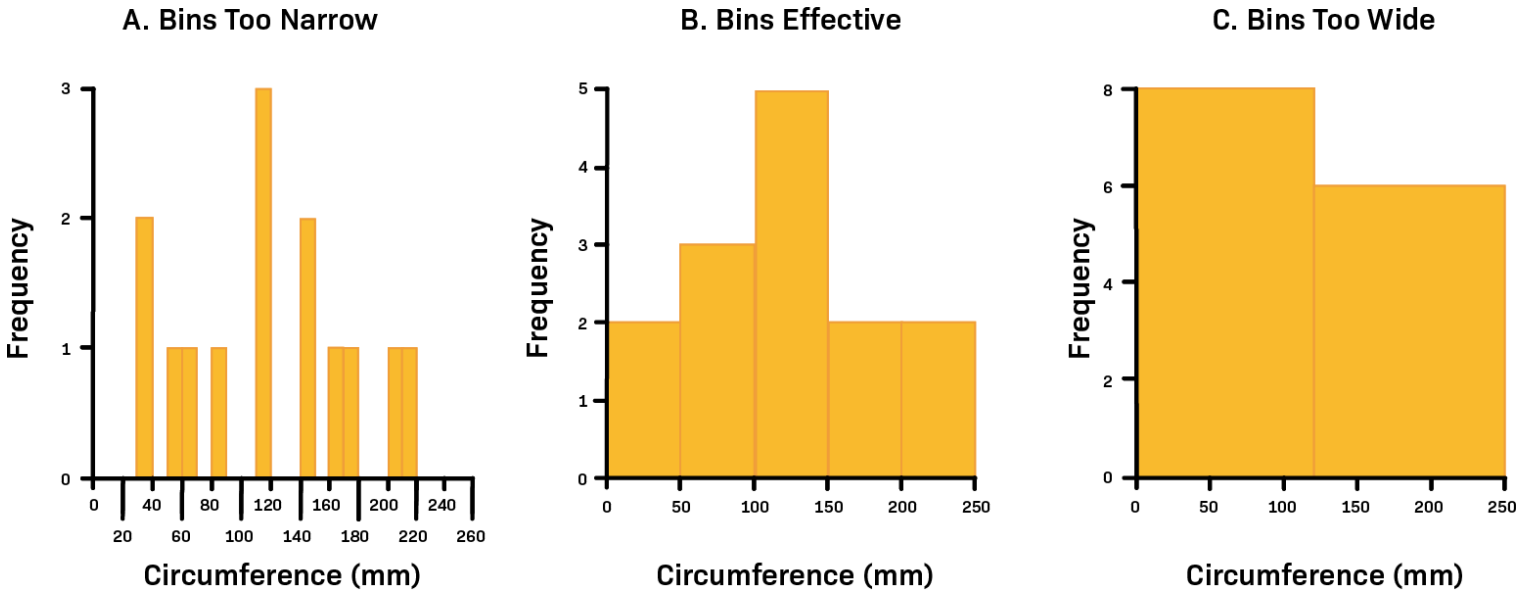
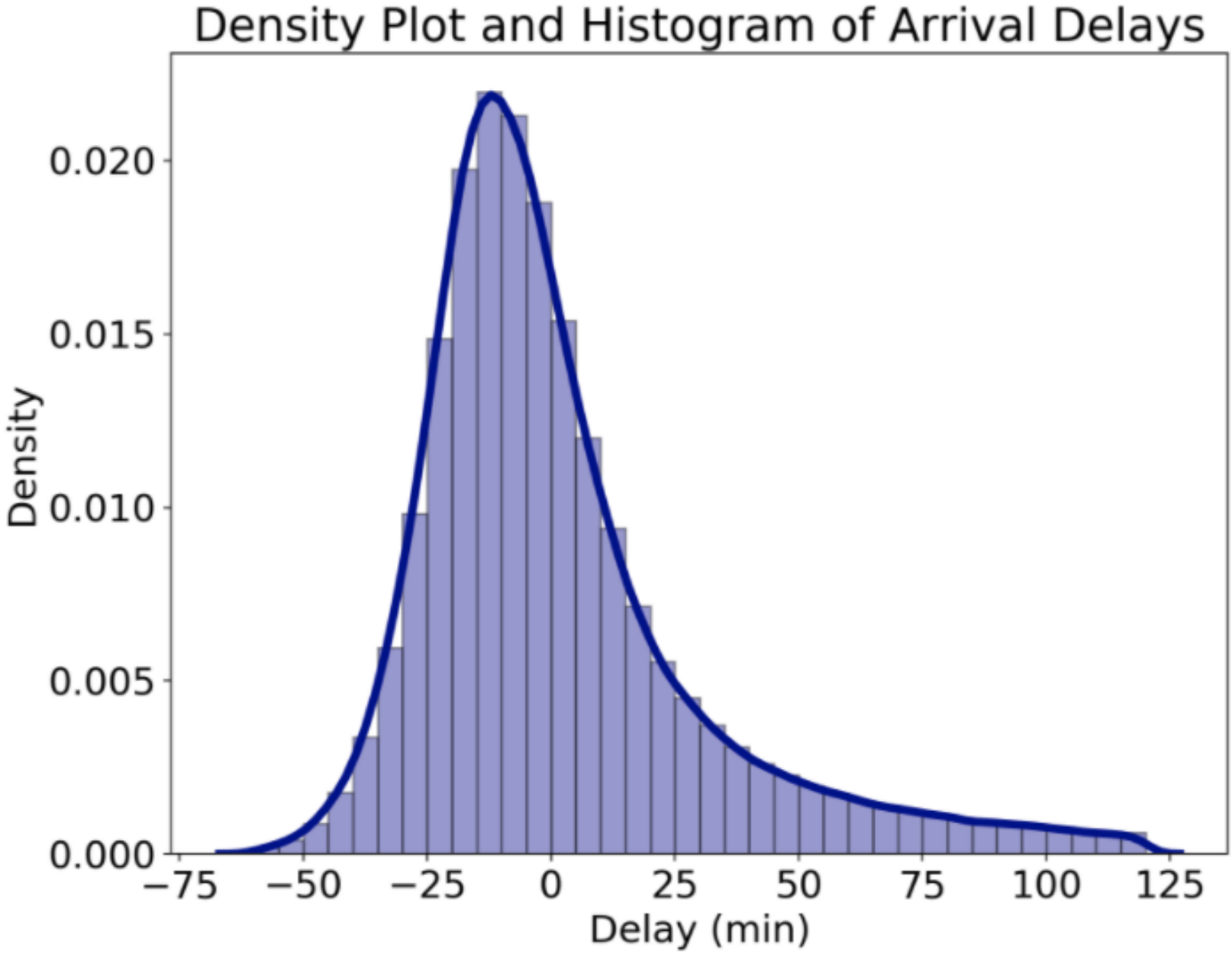
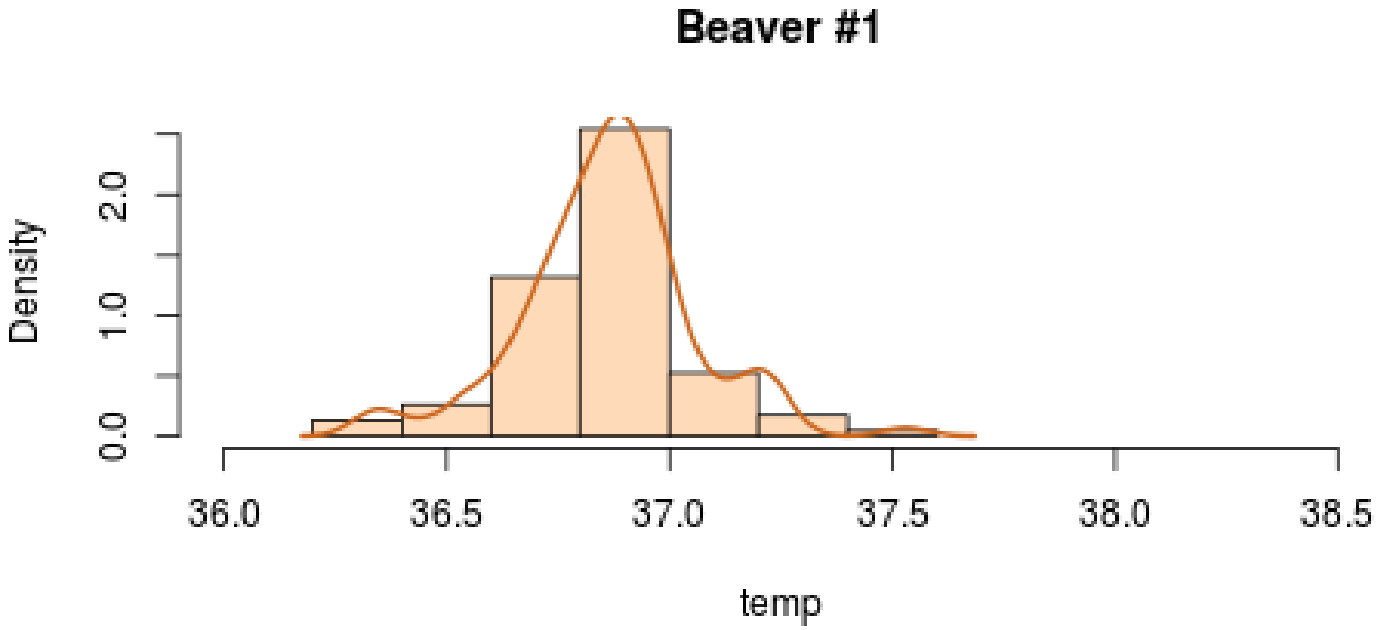


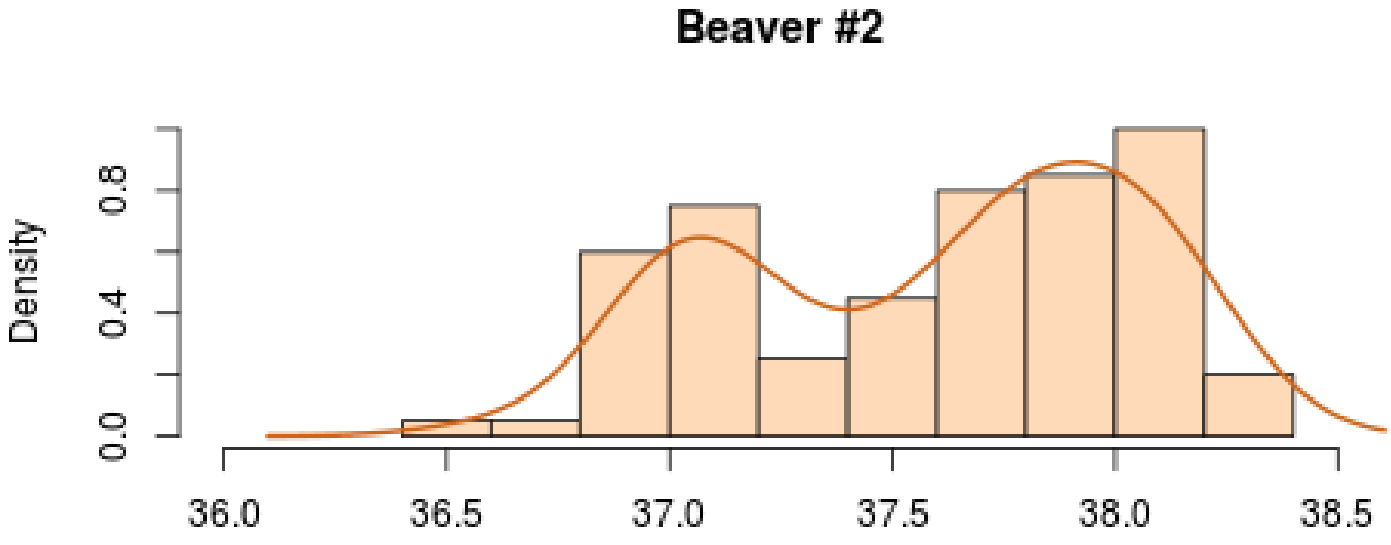
Figure 13: Bins that are too narrow may produce gaps. Bins that are too wide can hide the "shape" of the distribution.

Histograms –> Density Plots

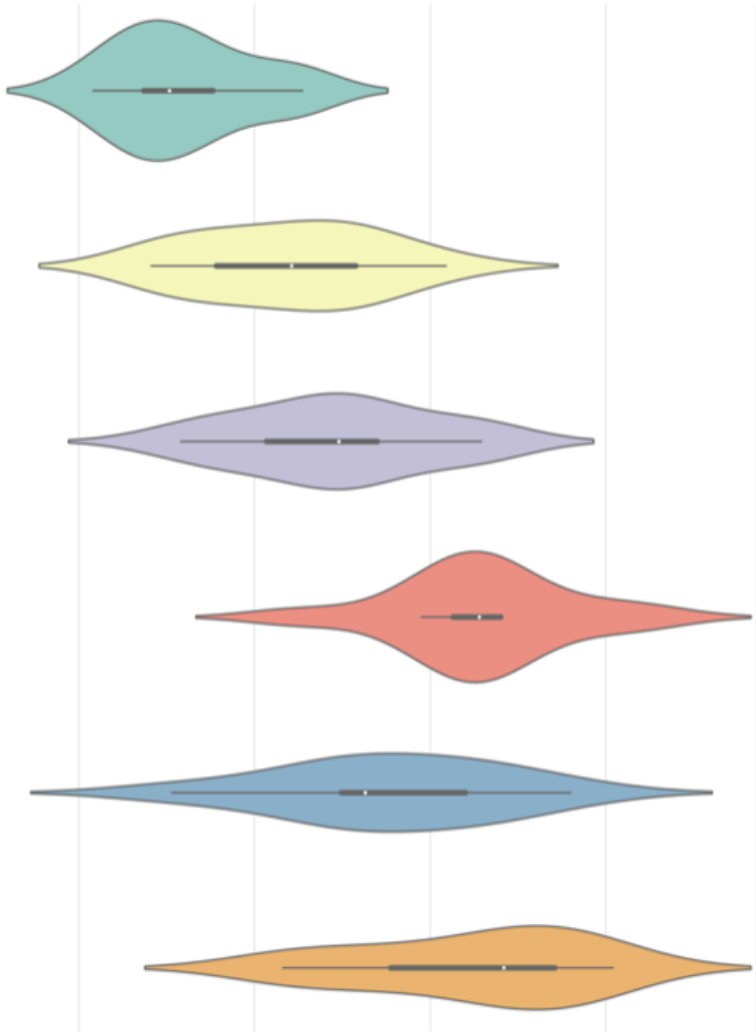


Histograms –> Density Plots





Density Plots –> Violin Plots



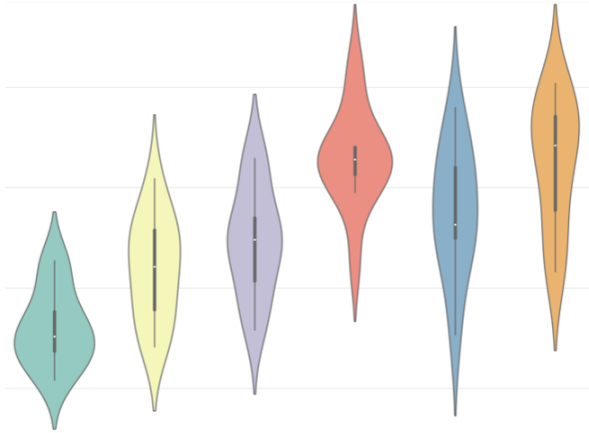


Figure 17: A violin plot of a variable is a mirrored image of its density curve. It is often plotted vertically, whereas density curves are usually plotted horizontally.

All together now

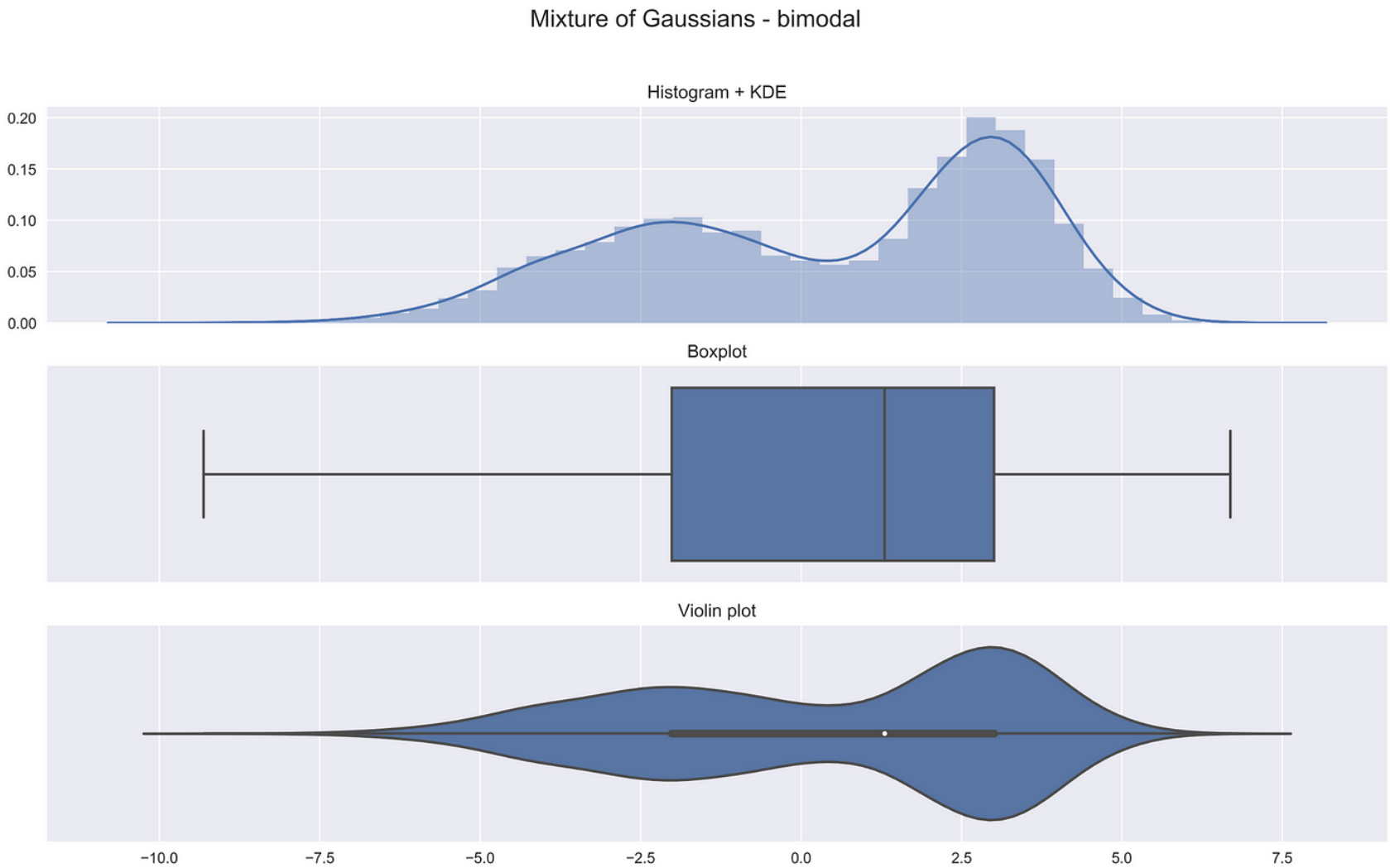
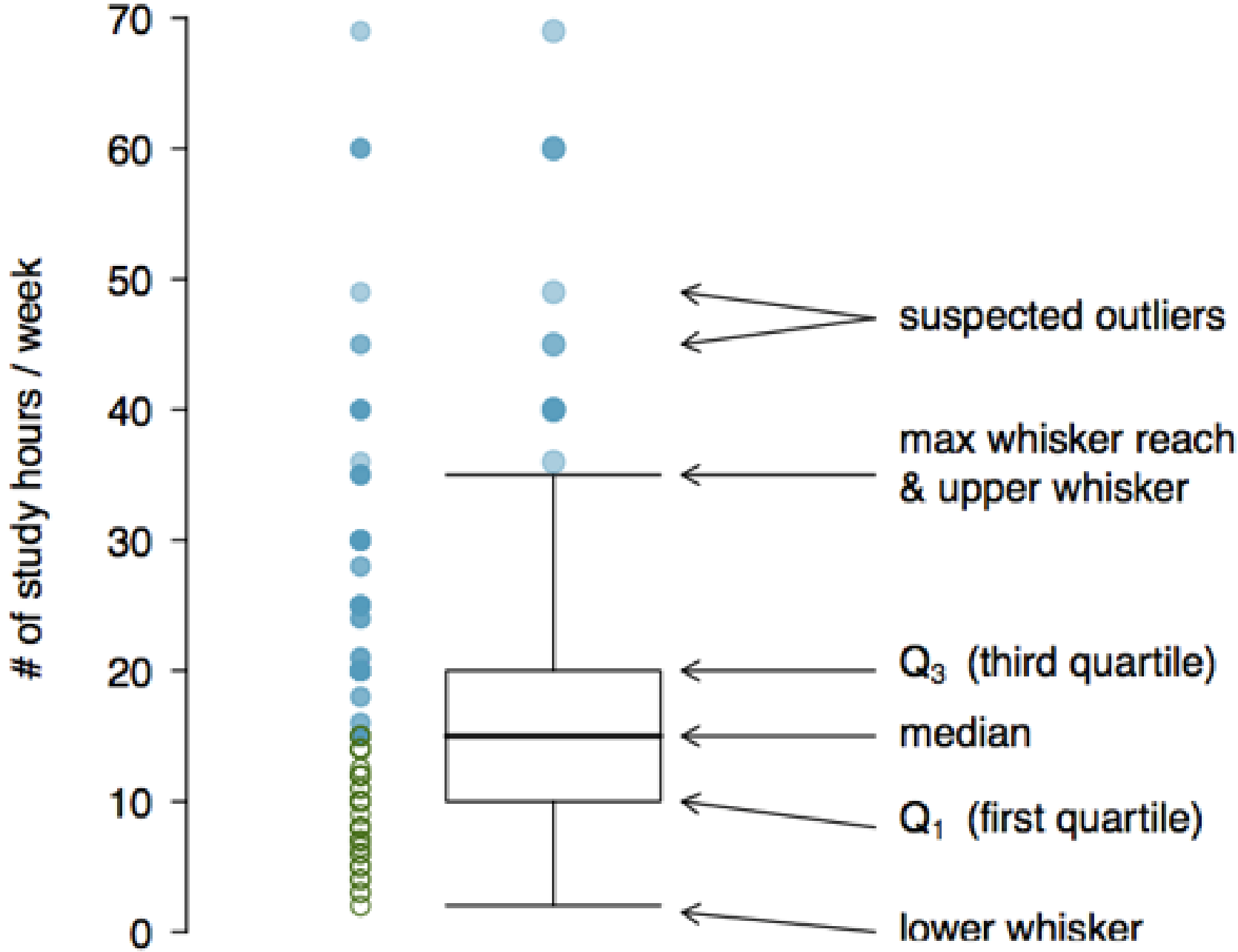


Figure 18: A histogram with a density curve overlaid, a violin plot, and a boxplot for the same distribution

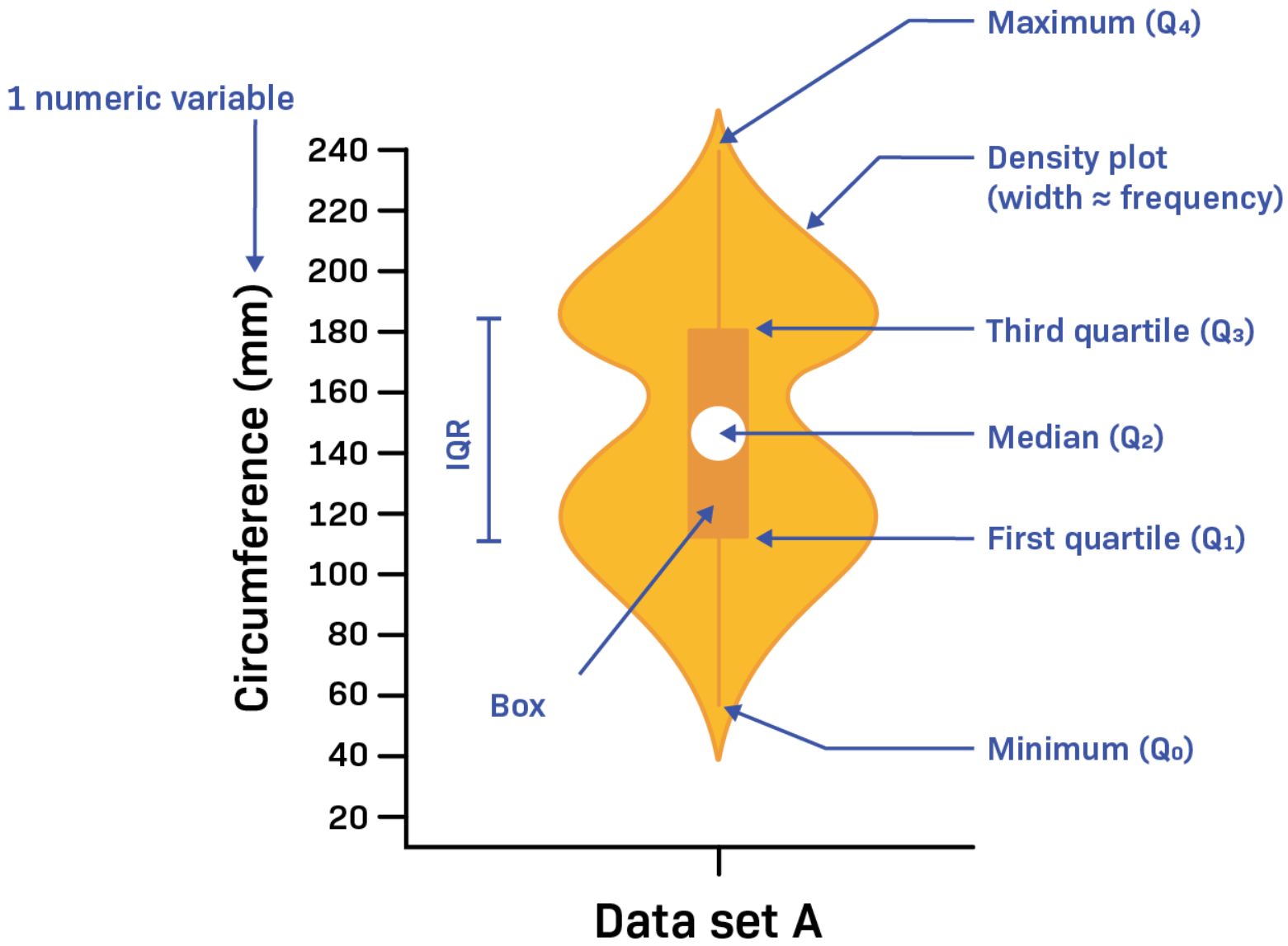
Anatomy of a Boxplot



Boxplot whiskers and outliers

- The ***whiskers*** of a boxplot (the lines extending out from the box) are 1.5 times the ***interquartile region*** long
  - Min whisker: Q1 - 1.5 x IQR
  - Max whisker: Q3 + 1.5 x IQR
- If a point is outside this range, it is considered to be a potential ***outlier***

Combining strategies: density + numerical summary



Combining strategies: violin + boxplot

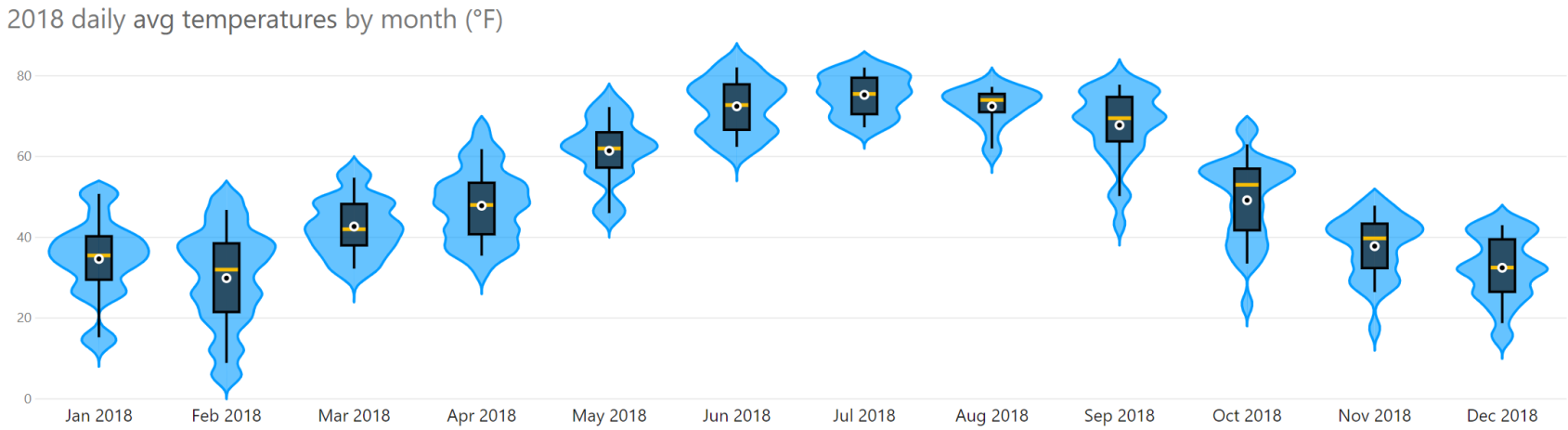


Figure 21: Some visualizations add a point to the boxplot indicating the location of the mean. If the mean is meaningfully different than the median, you have outliers and/or a skewed distribution.

Combining strategies: raincloud plots

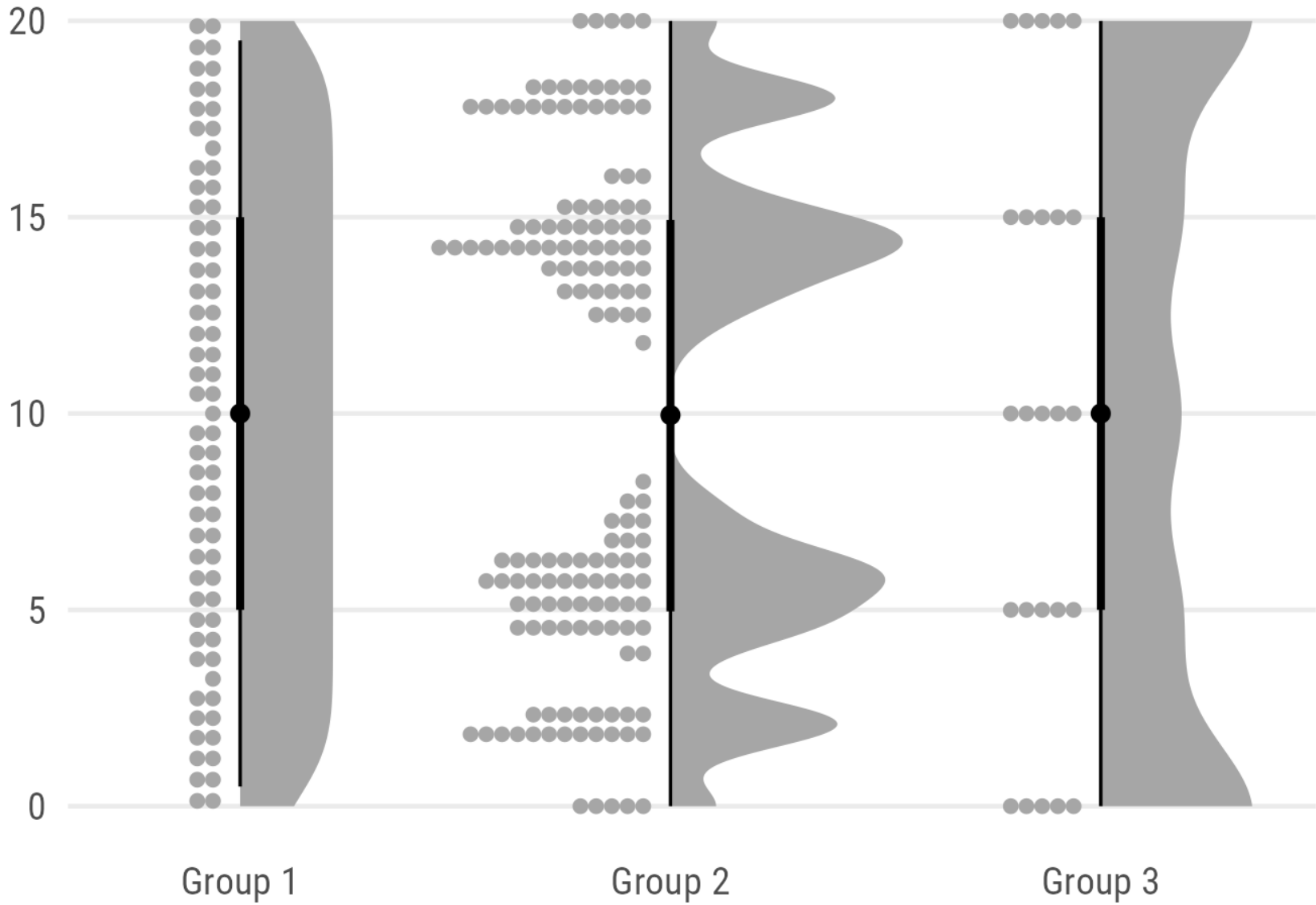


Figure 22: Raincloud plots combine density curves, boxplots, and stacked

Distribution Checklist

- What is the **modality** of the distribution?
  - How many "peaks" are there?
- Is the distribution **skewed** or **symmetric**?
  - Is there a longer "tail" on the left or right side?
- Are there any **outliers**?
  - How extreme are the most extreme values?
- What are the appropriate **summary statistics** for a distribution with this shape?
  - Would the mean+standard deviation or the median+IQR more accurately describe this data?

Example: The Penguins!

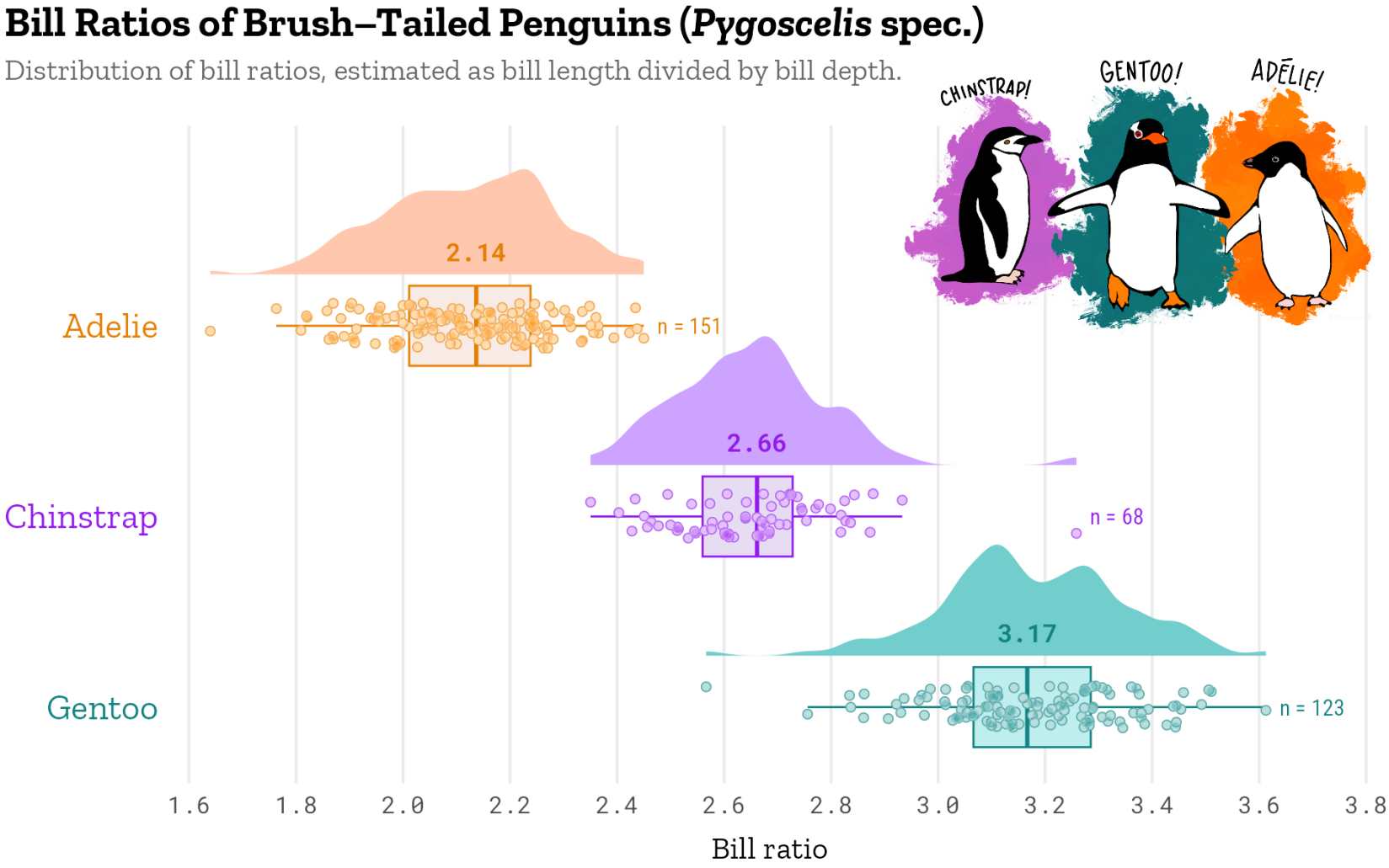


Figure 23: What is the modality of each distribution? Are they skewed?

Example: `datasets::iris` data set

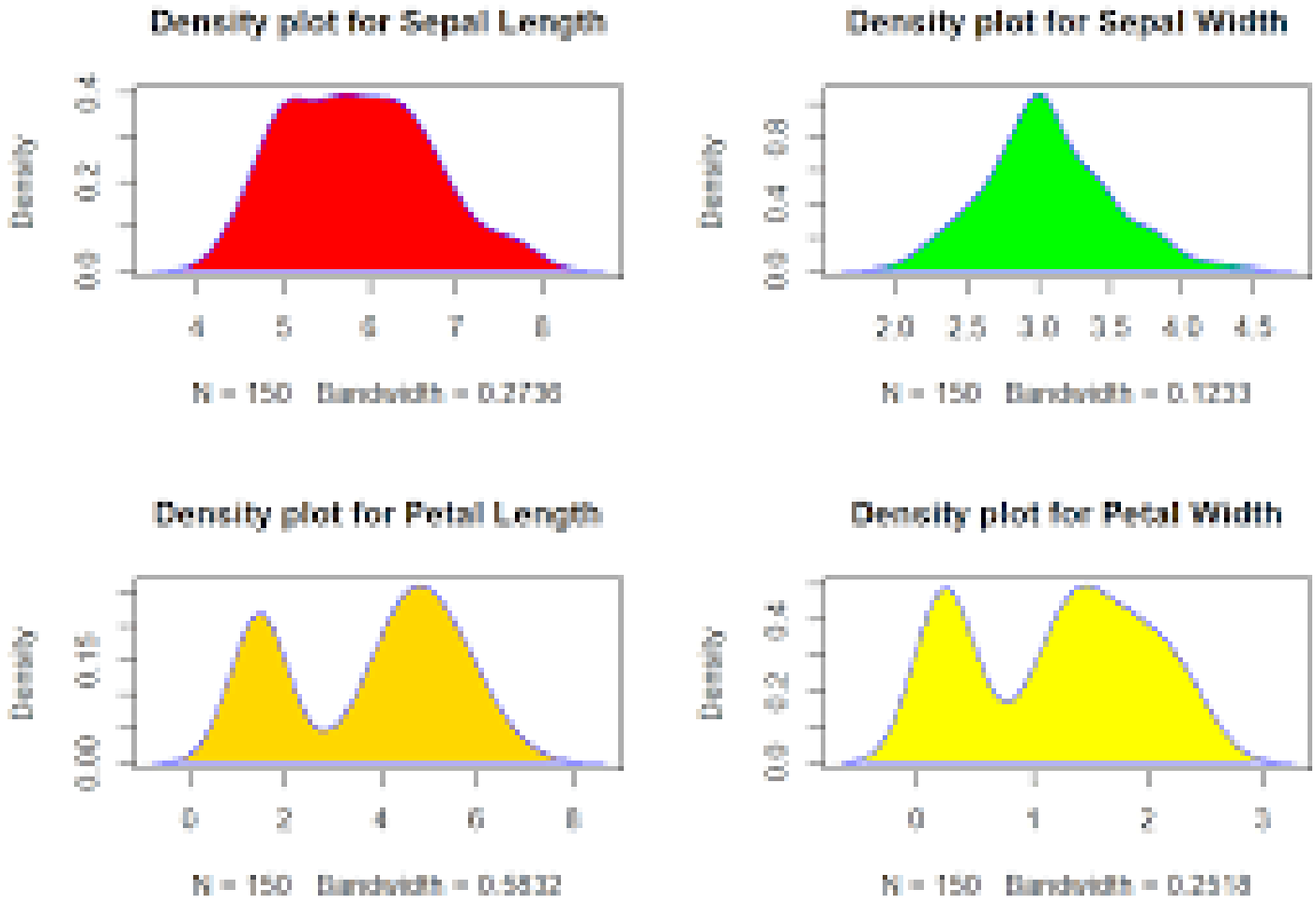


Figure 24: Describe the shape of these different distributions. Do any of them look normally distributed?

Example: `datasets::iris` data set

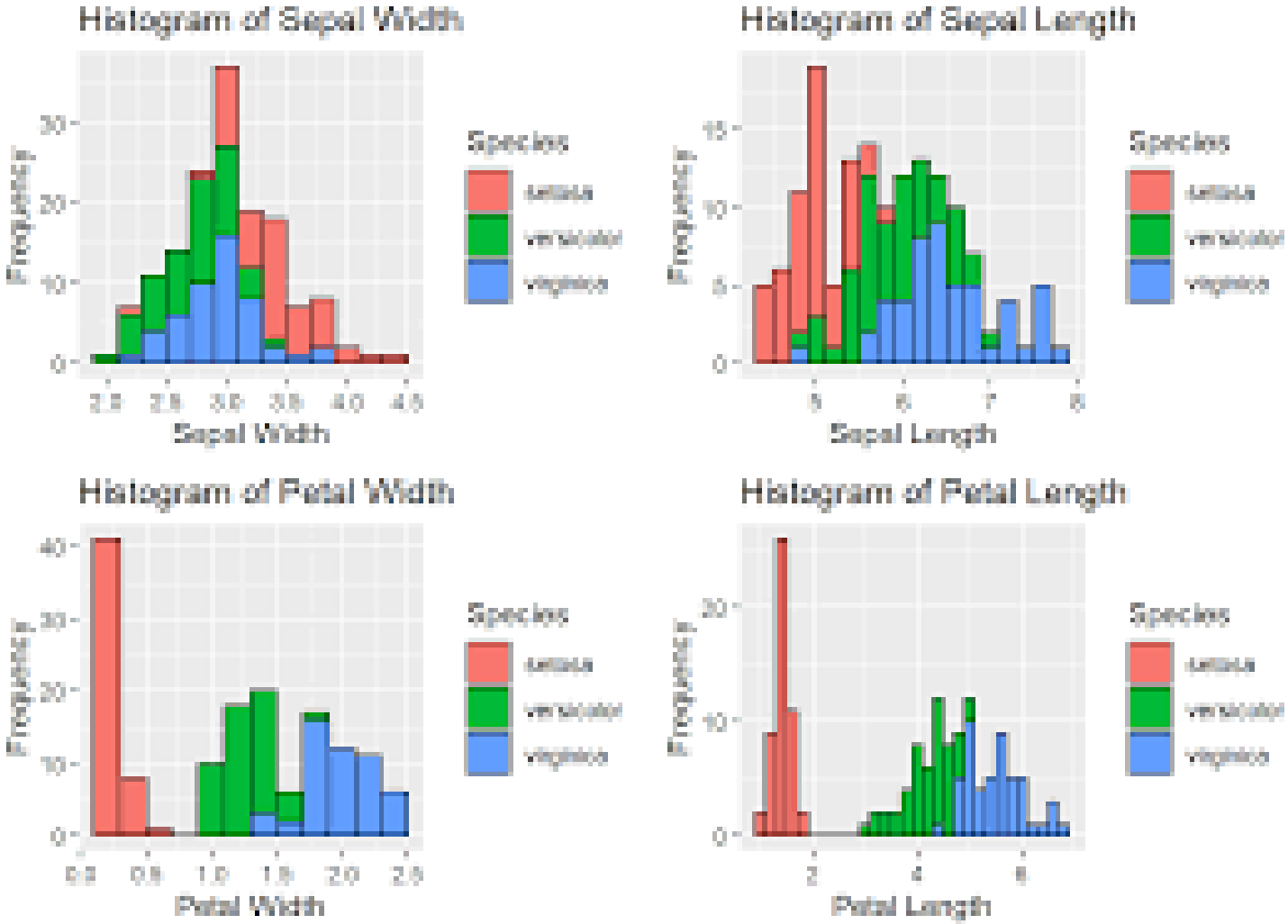


Figure 25: When a distribution has multiple modes or is unusually distributed, it may be better to visualize the data separated by a categorical variable.

Example: `datasets::iris` data set

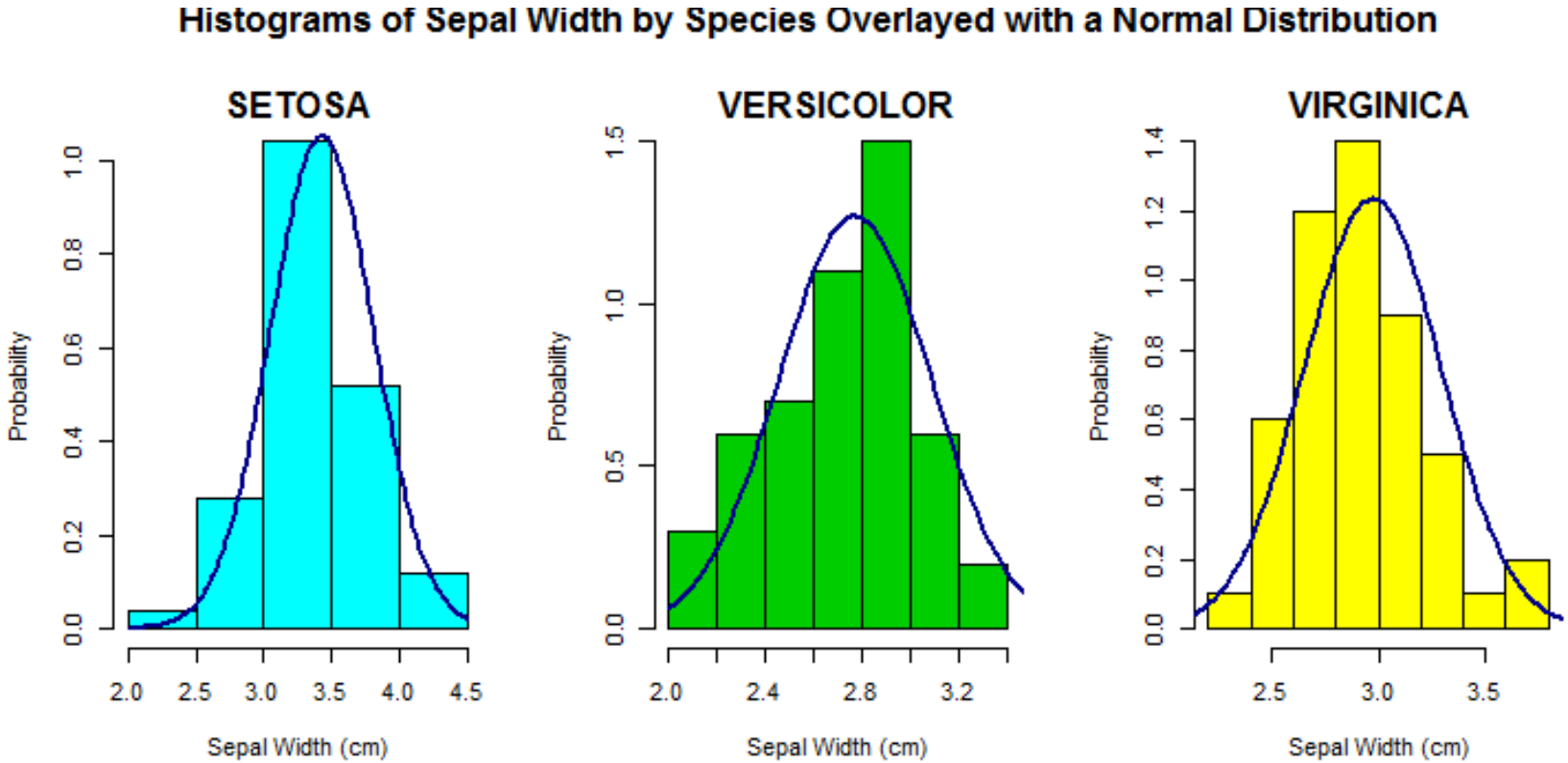


Figure 26: What type of special distribution is this? What summary statistics best describe this type of distribution?

Next time: Categorical Data

- Analyze contingency (e.g. 2x2) tables
- Summarizing categorical variables with proportions
- Comparison of numerical data between categorical groups

Next time: Visualizing Data

- Recognize common visualization techniques / plots
  - Numerical: Dot plots, histograms, density plots, QQ plots, box plots, violin plots
  - Categorical: bar plots, mosaic plots, tree map
- Build basic visualizations in `R` using `ggplot2`
- Data visualization do's and dont's