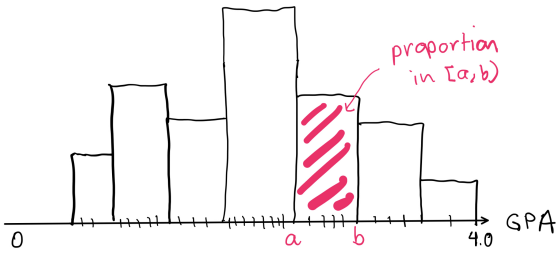
18.650: Fundamentals of Statistics Fall 2023

Lecture 2 — September 8, 2023

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Basic data visualization: the histogram

Suppose $x_1, \dots, x_n$ are the GPAs of the students in this class (we have $n = 130$). We can visualize the distribution of GPAs with a histogram.



- The area of the rectangle above $[a, b)$ is the proportion of GPAs between a and b:

$$\text{area} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(a \leq x_i < b), \tag{1}$$

  where $\mathbb{1}(\cdot)$ is called an *indicator* function. It evaluates to 1 if the statement inside the parentheses is true and it evaluates to 0 if the statement is false.

- Since area $= (b - a) \times$ height, we get the height of the column by dividing the area by $b - a$.

- **Caution:** *only if* the bins are equally spaced can we visually judge the proportions of GPAs in each bin by looking at the heights. If the bins are *not* equally spaced then the heights don't tell us everything (a column could be unusually tall if it corresponds to a very small bin size.)

1.1 Shapes

We can also smooth out the histogram with a "kernel density estimator" (KDE), which we'll learn about in December. A smoothed out histogram tells us about the *shape* of the distribution; see Figure 1.

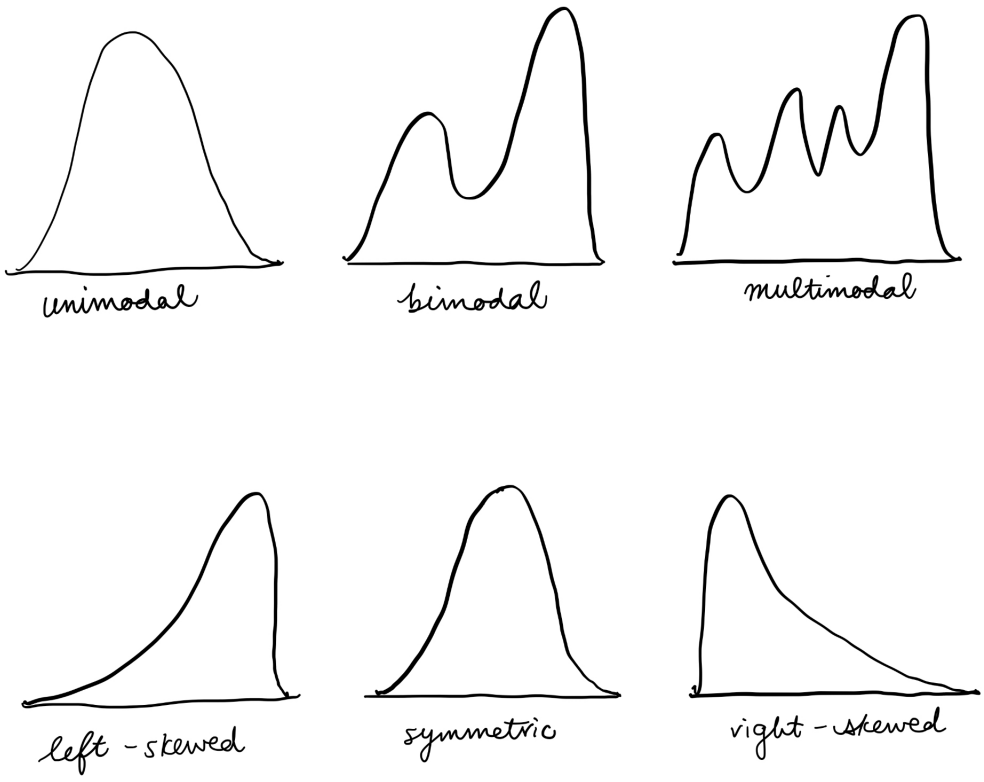


Figure 1: Shapes a distribution can take. Is there skew? Are there one or several modes?

2 Summary statistics

We can summarize the data with a few summary statistics:

- **mean**
- **standard deviation**
- **median**
  - splits the data in half:
    half of the data points $x_1, \ldots, x_n$ are to the left and half to the right
  - Formally: $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq \text{median}) = 1/2$.
- **quantiles**
  - a generalization of the median
  - **the q_α quantile** is a number such that α of the data is *above* it.
  - Formally: $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq q_\alpha) = 1 - \alpha$.

- Important special cases:
  1st quartile $Q_1 = q_{0.75}$ (3/4 of the data is to the right),
  3rd quartile $Q_3 = q_{0.25}$ (1/4 of the data is to the right).
- Note that $Q_1 < \text{median} < Q_3$. The median is the second quartile.

- **interquantile range IQR**$= Q_3 - Q_1$.

Figure 2 shows some of these statistics on a smoothed histogram.

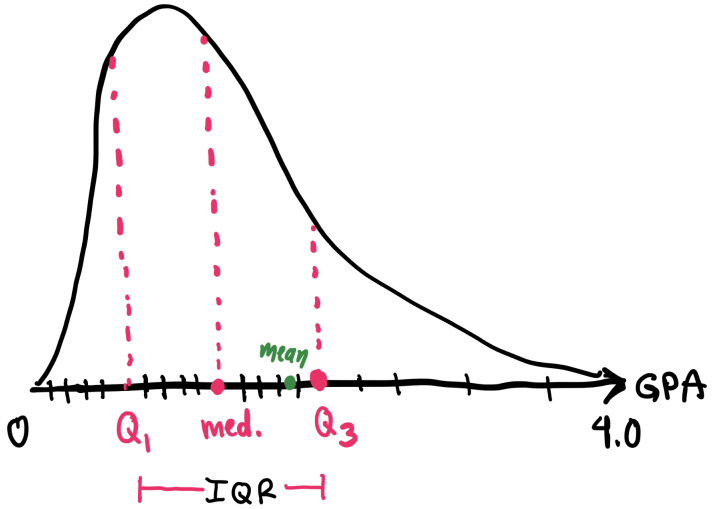


Figure 2: A smoothed histogram showing locations of the summary statistics. Note that the mean is to the right of the median because the distribution is right-skewed.

2.1 Robustness

Outliers are abnormally large or small values compared to the rest of the data. Formally:

$$x_i \text{ is an } \textbf{outlier} \text{ if } \quad x_i > Q_3 + 1.5\text{IQR} \quad \text{or} \quad x_i < Q_1 - 1.5\text{IQR}.$$

The mean and standard deviation are strongly affected by outliers. For example, very large outliers pull the mean to the right of the median, as in Figure 2. On the other hand, the median and IQR are robust to outliers (if the largest data point is doubled, say, this will not change the location of the median and IQR). The following table summarizes four important summary statistics.

location	spread	
mean	std. dev.	
median	*IQR*	← *robust*

3 Visualizing summary statistics

Another way to concisely depict summary statistics is with a *box plot*, also sometimes called a "box and whiskers" plot; see Figure 3. The left and right endpoints of the box are Q_1, Q_3 respectively, and a line is drawn in between to denote the location of the median. The box is our visual representation for 75% of the data, and the location of the median between Q_1 and Q_3 conveys whether the distribution is skewed in one direction.

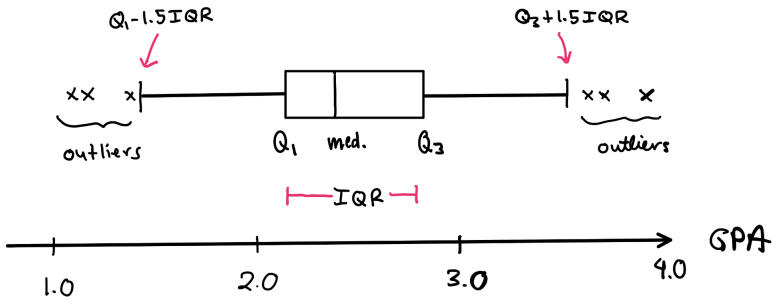


Figure 3: A box plot

The two line segments extending to the left and right from the box are the "whiskers". The length of the whiskers is 1.5IQR, so that by definition, the outliers are to the left of the left whisker and to the right of the right whisker. The locations of the outliers are indicated explicitly on the boxplot.

3.1 Data with multiple variables: scatterplot and comparative boxplot

So far we've only considered one-dimensional data. But we could also have e.g. pairs (X_i, Y_i). For example, X_i denotes the number of days a month a student smokes marijuana, and Y_i denotes the student's GPA. A common way to depict such data is with a scatterplot, as in Figure 4. We simply plot the location of each data point in the X-Y plane.

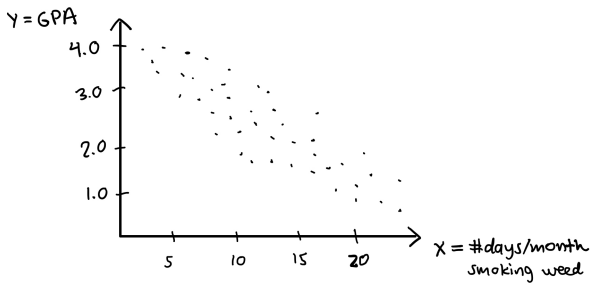


Figure 4: A scatterplot

If the X-values tend to be clustered or if the X-values are not numbers at all (e.g. $X_i \in \{\text{freshman, sophomore, junior, senior}\}$), then we can depict the data using several boxplots for the Y distributions, one for each cluster/category of X values. This lets us visualize the difference in the Y distributions across different X values. This is called a comparative boxplot; see Figure 5.

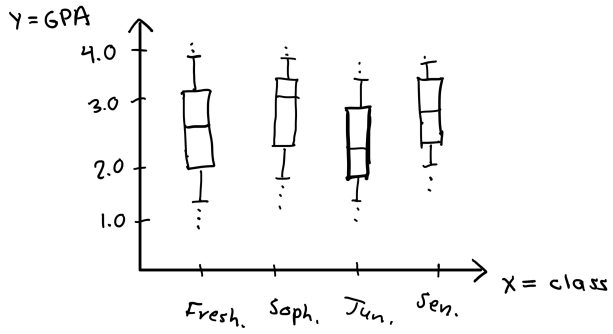


Figure 5: Comparative boxplots