Population — collection of units

Variate — characteristic of a unit; something measured

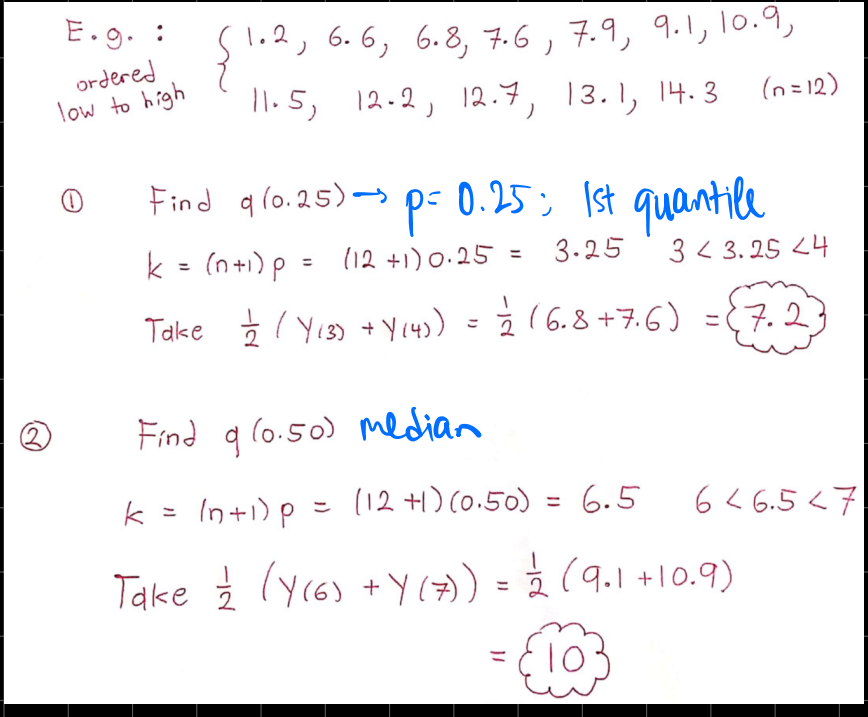
- Variates can be discrete/continuous, categorical, ordinal, etc

A *process* is something that causes variates to change

Ex. A group of STAT 231 students are asked about their favourite type of doughnut from the campus coffee shop (chocolate glaze, honey dip, apple fritter, or none of the above). Here, doughnut preference is an example of which type of variate?

- Categorical
- If in terms of the number of students that prefer each type, it's discrete

Quantiles



Interquartile Range (IQR): q(0.75) - q(0.25)

Histograms

Histograms represent frequencies in a dataset {a1, a2, …, a_n} using rectangles

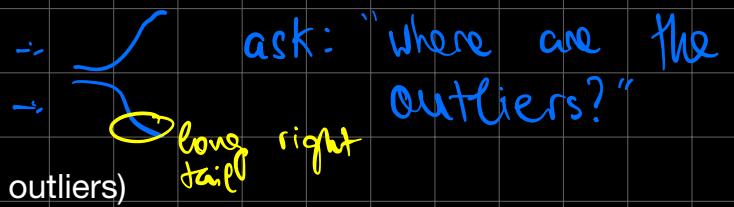
- *Standard frequency histogram*: intervals have equal length
    - Given an interval j, its height is either the frequency f_j or the relative frequency f_j / n
- *Relative frequency histogram*: intervals have different lengths

- The height of an interval j is $\frac{\frac{f_j}{n}}{a_j - a_{j-1}}$

Measures of shape

Measured relative to the *normal distribution*

- Skewness — whether distribution is shifted left/right
  - Left: negative
  - Right: positive
- Kurtosis — heaviness of tails (higher kurtosis = more outliers)
  - Kurtosis > 3: outliers occur more frequently than in normal distribution
  - Always positive

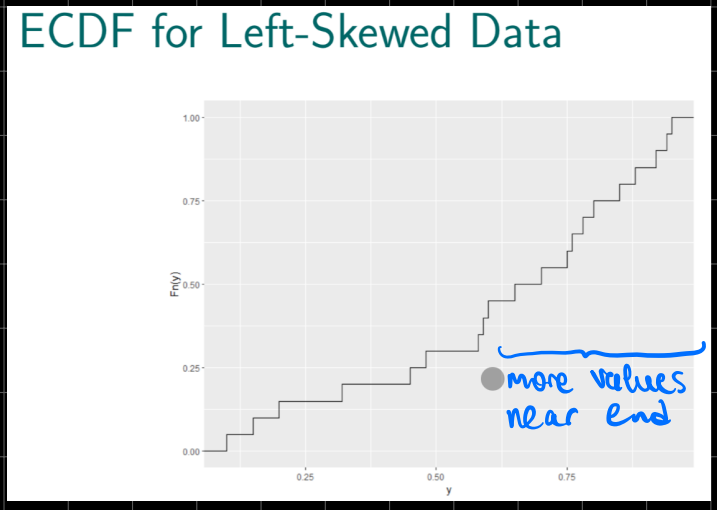


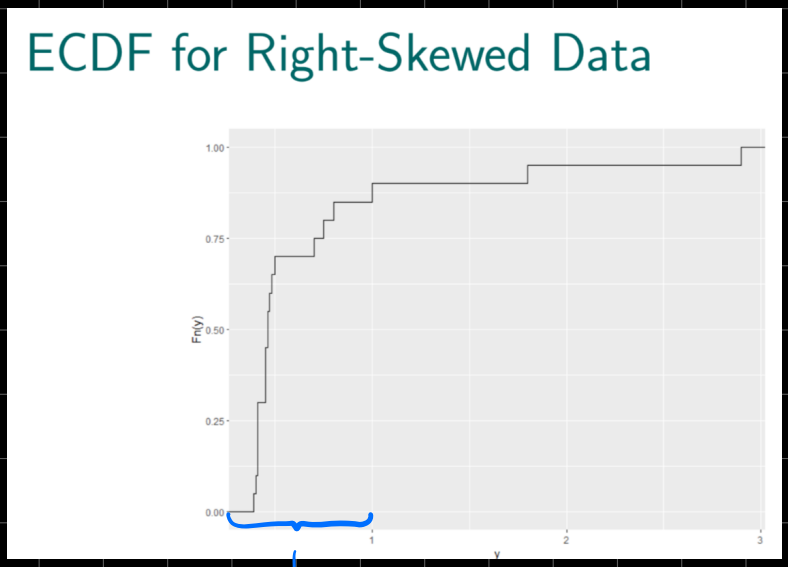
Empirical CDFs

Suppose we have an ordered dataset of n=10 observations, and suppose that they are from an unknown CDF.

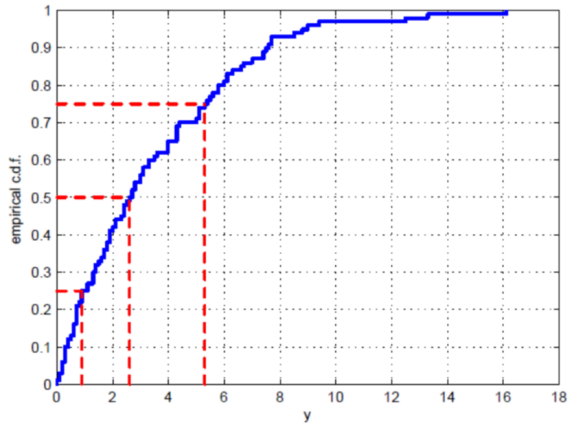
We can estimate the CDF, F(y), for any value of y by getting the proportion of values ≤ y:

$$\hat{F}(y) = \frac{\text{Number of values in the set} \{y_1, \ldots, y_n\} \leq y}{n}$$

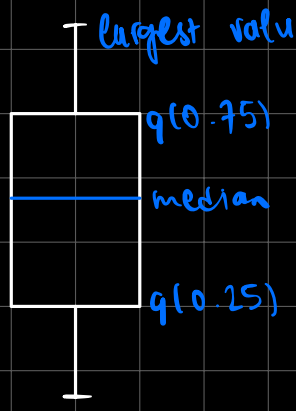


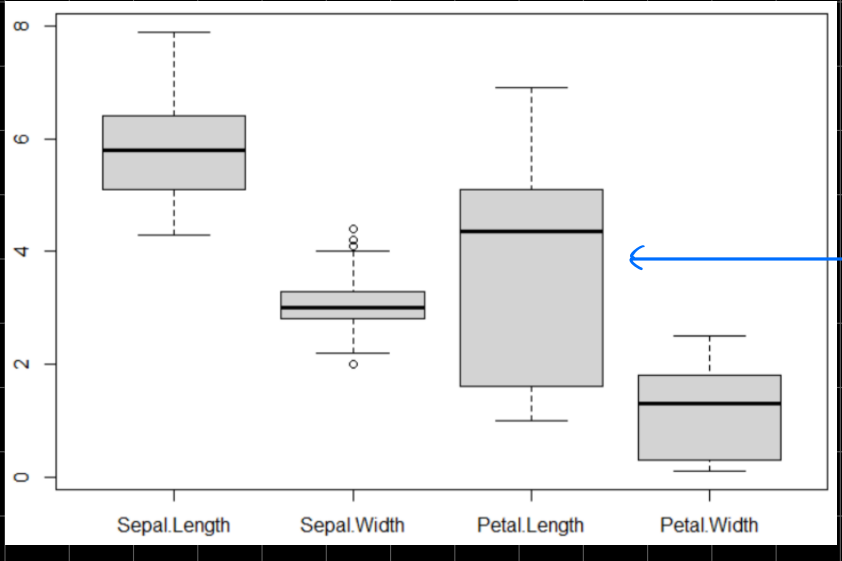


Reading ECDFs: Estimating Quantiles

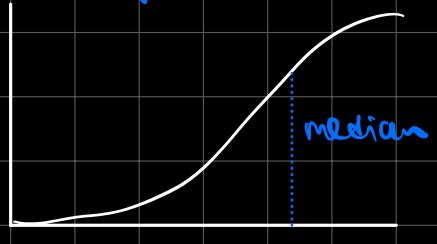


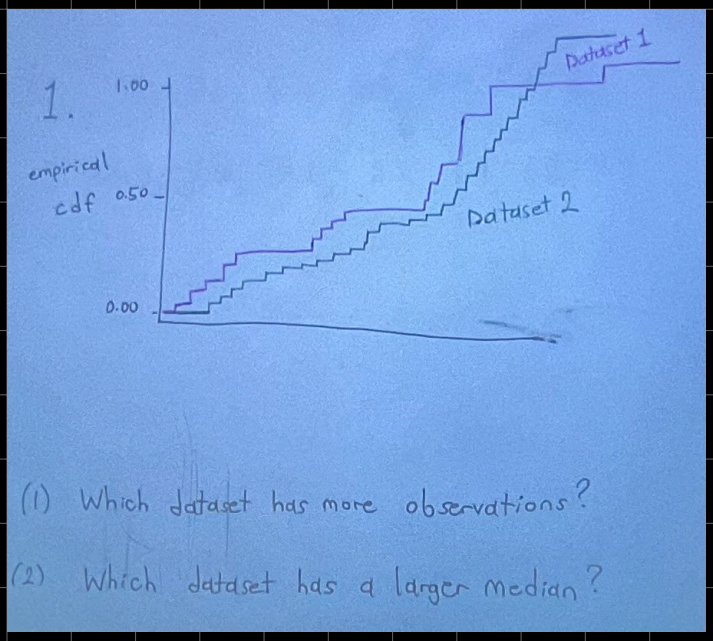
Boxplots





left skew: median is relatively high





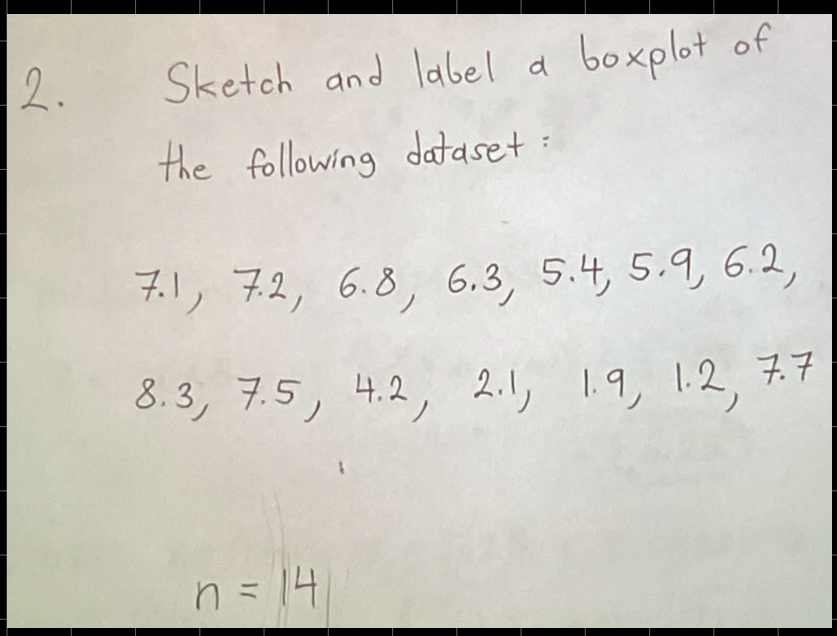
(1) Dataset 2 has more observations, since its steps are smaller

Dataset 1 has “lags”:

(2) Dataset 2 has a larger median

Describing the data:

- Dataset 1 looks more evenly distributed at the start, and more extreme values (high jumps) toward the end. This signifies a *right* skew.
- Dataset 2 is symmetric.



Sorted: 1.2, 1.9, 2.1, 4.2, 5.4, 5.9, 6.2, 6.3, 6.8, 7.1, 7.2, 7.5, 7.7, 8.3

Median: 6.25

more data closer to end
⇒ left skew

Calculating q(0.25): (0.25)(15) = 3.75

3 < 3.5 < 4

So q(0.25) = 0.5(2.1 + 4.2) = 3.15

Calculating q(0.75): (0.75)(15) = 11.25

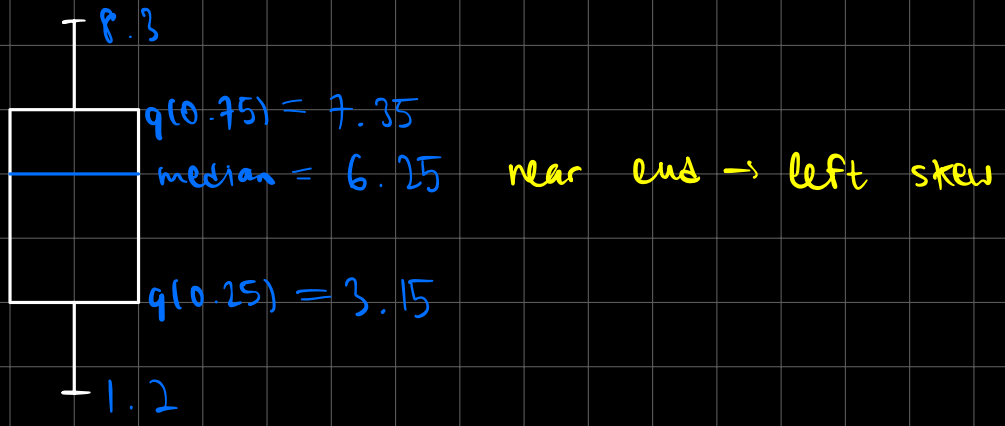
11 < 11.25 < 12

So q(0.75) = 0.5(7.2 + 7.5) = 7.35
11th term 12th

IQR = q(0.75) - q(0.25) = 4.2

So the whiskers of the box plot are:

- *Lower line:* q(0.25) - 1.5 • IQR = -3.15 —> max(min(data), -3.15)) = 1.2
- *Upper line:* q(0.75) + 1.5 • IQR = 13.65 —> min(max(data), 13.65) = 8.3



Relative Risk

Ex.

	CHD	No CHD	Total
Placebo	189	10845	11034
Daily Aspirin	104	10933	11037
Total	293	21778	22071

Table 1.5: Physicians' Health Study

The relative risk of CHD in the placebo group compared to the daily aspirin group is

$$\frac{189/11034}{104/11037} = 1.82$$

proportion of placebo group w/ CHD

proportion of aspirin group w/ CHD