CEE 260/MIE 273: Probability and Statistics in Civil Engineering

M1c: Case studies and experiments

Prof. Oke

UMass Amherst

College of Engineering

September 8, 2025

Outline

Your expectations (Fall 2024)

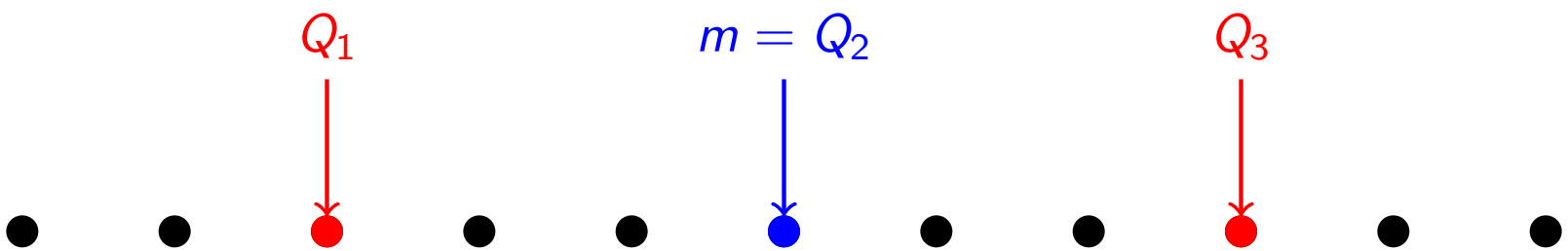
Quantiles

Quantiles are cutoff points that partition an ordered sample or dataset into equal-sized groups.

- A **median** splits a sample into two: m
- Two **terciles** split a sample into 3 groups: T_1, T_2
- Three **quartiles** split a sample into 4 groups: Q_1, Q_2, Q_3
- Four **quintiles** split a sample into 5 groups: QU_1, QU_2, QU_3, QU_4
- ...
- Ninety-nine **[per]centiles** split a sample into 100 groups: $P_1, \ldots, P_{99}$

Quantiles (cont.)

Certain quantiles are equivalent to others:



- The median is the second quartile Q_2
- The 25th percentile is equivalent to the first quartile Q_1

Sextiles ($S_1, S_2, \ldots$) partition a distribution into 6 equal groups.

1. How many sextiles are there?
2. The second sextile S_2 can be expressed as which tercile?[a]

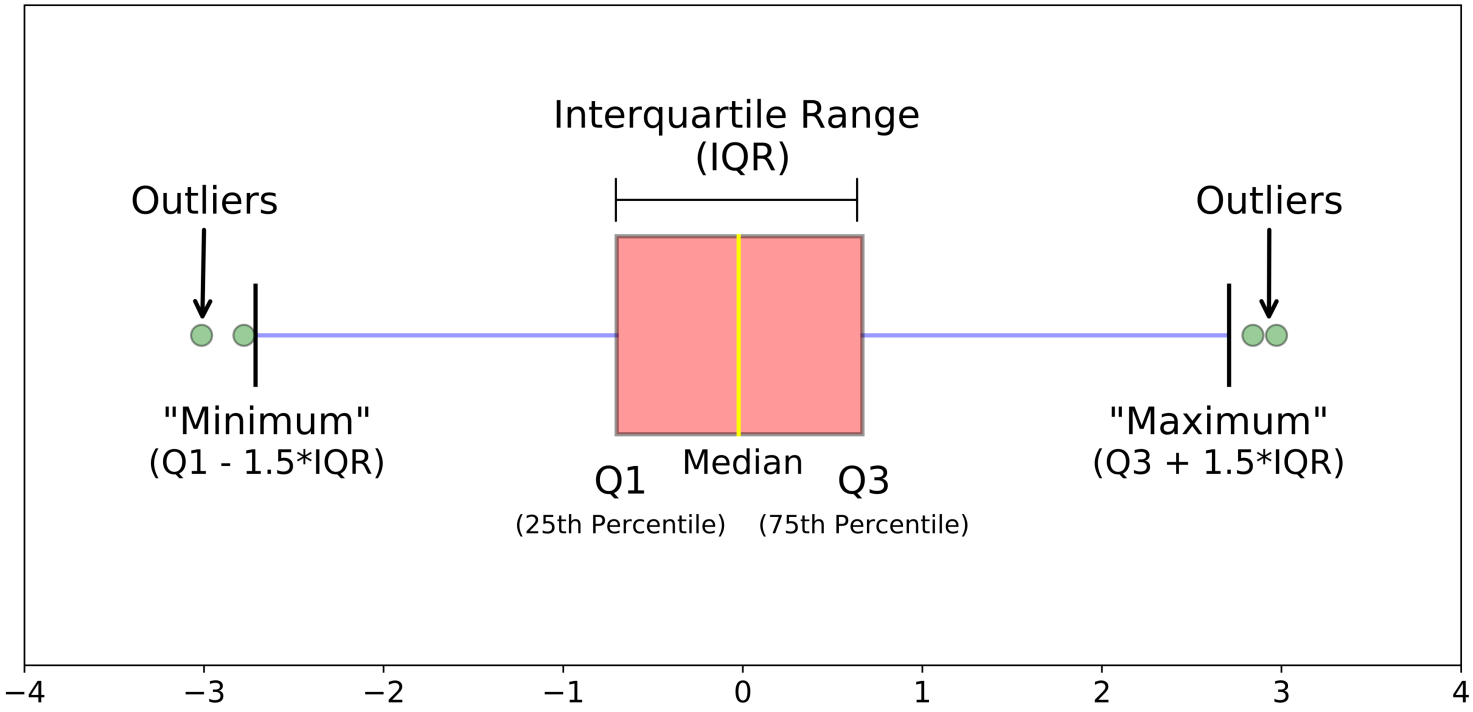
Answers: Q1: There are 5 sextiles; Q2: $S_2 = T_1$ (first tercile)

[a] Recall that a tercile splits a sample into 3 equal groups.

Using boxplots

A boxplot[1] displays the distribution of data

- Useful for identifying outliers
- Efficient for comparing multiple datasets
- The lines indicating the "maximum/minimum" points (excluding outliers) are called *whiskers*



[1] Figure source: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

MATLAB

We will begin our introduction to MATLAB via basic statistical analyses.

For an introduction, visit: https://www.mathworks.com/help/matlab/getting-started-with-matlab.html.

Summarizing data

Example 1: Walking cadence

In the article "Can We Really Walk Straight?" (*Amer. J. of Physical Anthropology*, 1992: 19–27) reported on an experiment in which each of 20 healthy men was asked to walk as straight as possible to a target 60m away at normal speed.

Consider the following observations on cadence (number of strides per second):

.95 .85 .92 .95 .93 .86 1.00 .92 .85 .81 .78 .93 .93 1.05 .93 1.06 .96 .81 .96

Summarize the data; interpret and discuss.

Example 2: Iris dataset

We will explore the properties and applications of boxplots using the following datasets:

- Popularized in Ronald A. Fisher's classic 1936 paper, "The Use of Multiple Measurements in Taxonomic Problems" https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x
- Data collected by Edgar Anderson on various measurements of 3 species of Iris flowers
- Also found on the UCI Machine Learning Repository. Data available by default on MATLAB

Example 2: Iris species

Three species of the *Iris* flower:

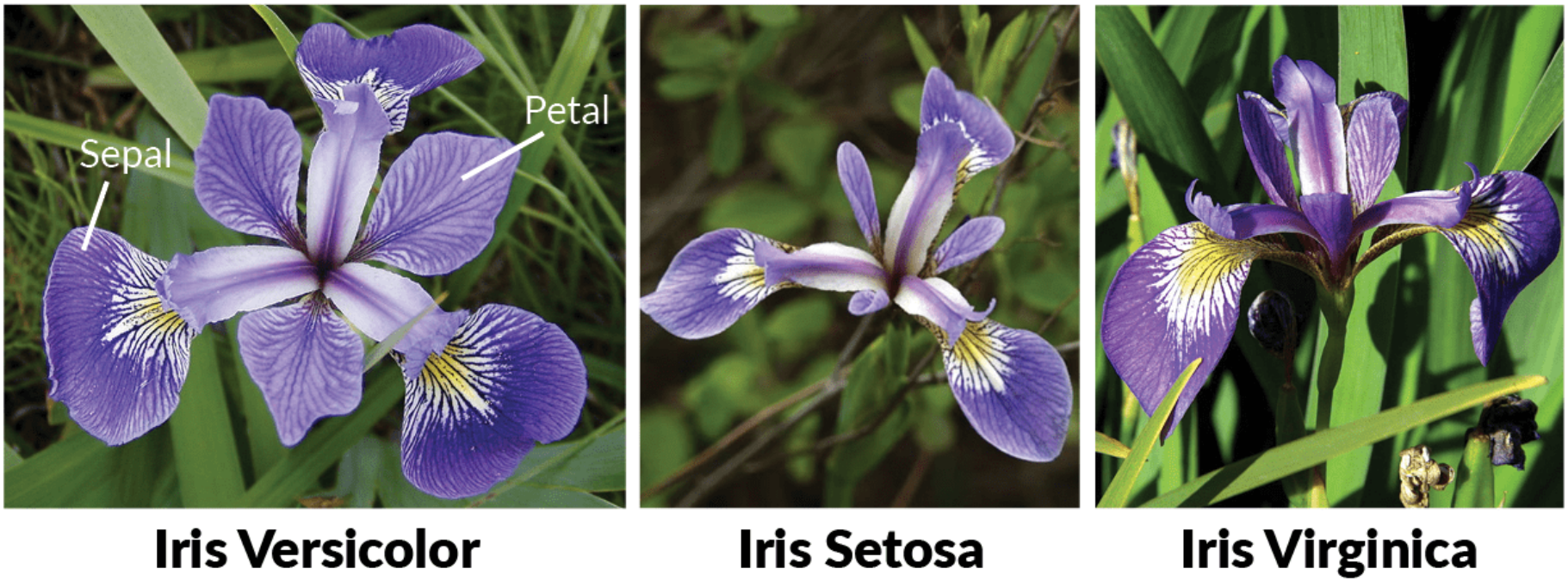


Figure: Iris species (Source:
https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Machine+Learning+R/iris-machinelearning.png)

Example 2: Iris flower measurements

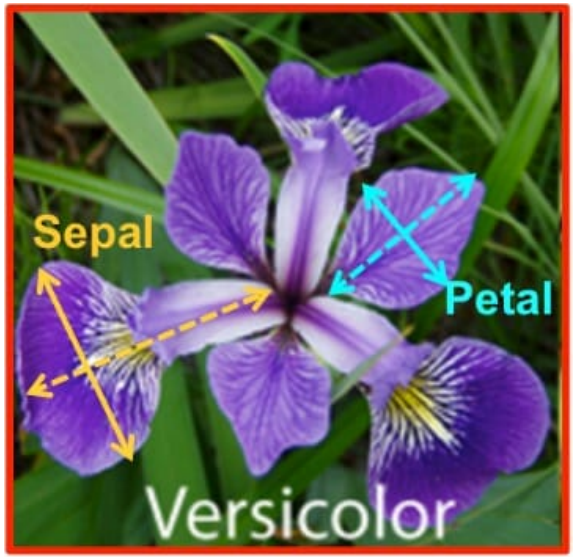


Figure: *Iris versicolor* sepal and petal measurements (Source: https://yculz33w9skgdkhey8rajqm6-wpengine.netdna-ssl.com/wp-content/uploads/2018/11/versicolor.jpg)

Are there significant differences in the petal/sepal width/length in each species?

Example 3: "Daphne and Santa Cruz."

- The data set consists of measurements of beak sizes in mm of one species of Darwin's ground finch (Geospiza fortis) taken at Daphne Island and at Santa Cruz Island in the Galapagos by Peter and Rosemary Grant.
- Data was extracted from http://wps.prenhall.com/esm_freeman_evol_3/0,8018,8412374-,00.html.
- The original data is summarized in the article: "The classical case of character release: Darwin's finches (Geospiza) on Isla Daphne Major, Galapagos" by P. T. Boag and P. R. Grant that appeared in Biological Journal of the Linnean Society 22:243-287 (1984).