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

M1b: Summarizing Data

Prof. Oke

UMass Amherst

College of Engineering

September 4, 2025

Outline

Recap from Lecture 1a

- Two categories of uncertainty: aleatory and epistemic
- Visualize distributions via histograms and scattegrams
- Measures of location: mean, median, mode
- Measures of dispersion: range, variance, standard deviation, coefficient of variation

Objectives of today's lecture

- Understand quantiles and boxplots
- Learn how to use Python to generate descriptive statistics and graphical summaries of datasets (using Colab)

Class expectations (Fall 2020)

Class expectations (Fall 2021)

Class expectations (Fall 2024)

Sample mean

A **sample** is a finite set of n observations:

$$(x_1, x_2, \ldots, x_n) \tag{1}$$

Sample mean

This is the average of a sample:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2}$$

Standard deviation and variance

The sample variance and standard deviation are **measures of dispersion**.

Sample variance

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3}$$

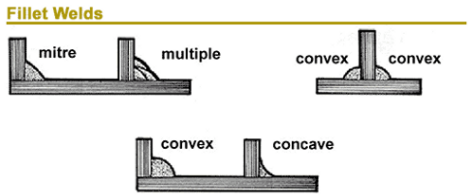
Sample standard deviation

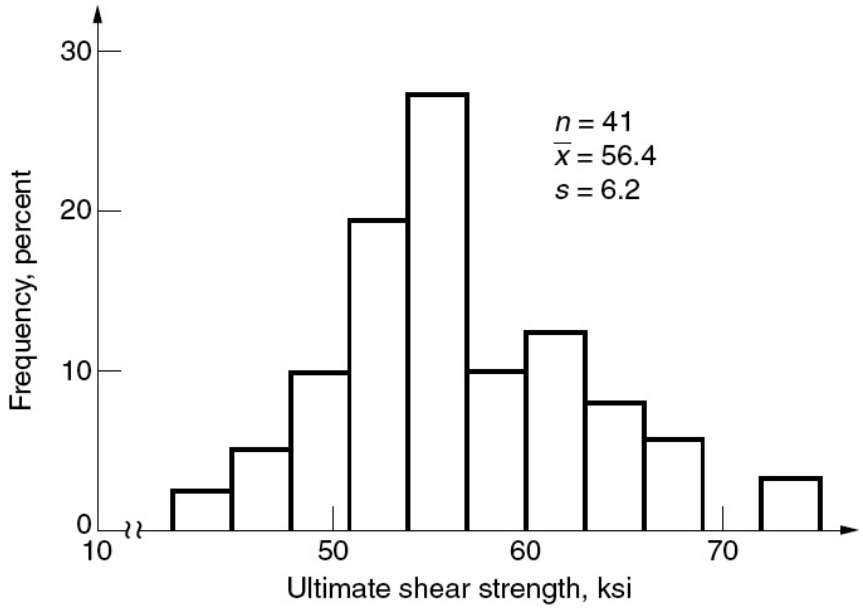
This is the square root of the variance:

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

Example 7: Ultimate shear strength of steel fillet welds

The ultimate shear strength of a material is its maximum allowable force per unit area prior to sliding failure. This is measured for a sample of fillet welds:



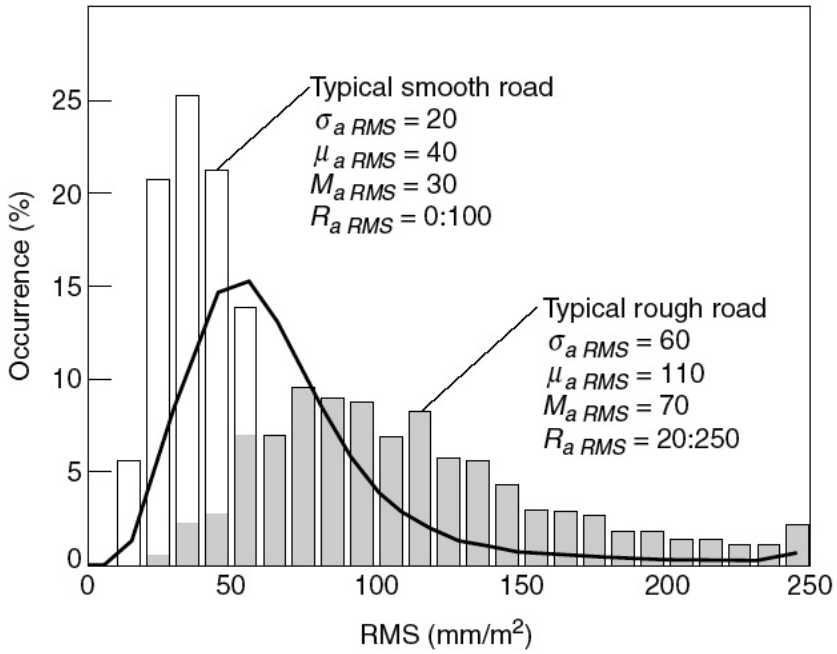


The **sample size** is 41. The sample mean is 56.4 ksi. The standard deviation is 6.2 ksi.

Note: the mean and standard deviation have the same units

Example 8: Comparing two distributions

Recall the surface roughness distributions from two samples:



Questions

- Which sample has the larger standard deviation?
- What are the implications?

Sampling error and coefficient of variation

The sample variance and standard deviation measure the aleatory uncertainty in the data.

Sampling error

Defined as the standard deviation of the sample mean:

$$s_{\overline{x}} = \frac{s_X}{\sqrt{n}} \tag{5}$$

Measures how well the sample mean $\overline{x}$ estimates the population mean. Also known as **standard error**.

Coefficient of variation

Measures dispersion relative to the mean.

$$\delta_X = \frac{s_X}{\overline{x}} \tag{6}$$

Example 9: Rainfall intensities summary statistics

Recall rainfall intensities:

TABLE 1.1 Rainfall Intensity Data Recorded over a Period of 29 Years

Year No.	Rainfall Intensity, in.	Year No.	Rainfall Intensity, in.	Year No.	Rainfall Intensity, in.
1	43.30	11	54.49	21	58.71
2	53.02	12	47.38	22	42.96
3	63.52	13	40.78	23	55.77
4	45.93	14	45.05	24	41.31
5	48.26	15	50.37	25	58.83
6	50.51	16	54.91	26	48.21
7	49.57	17	51.28	27	44.67
8	43.93	18	39.91	28	67.72
9	46.77	19	53.29	29	43.11
10	59.12	20	67.59		

$$\bar{x} = \frac{1}{29}(43.30 + \cdots + 43.11) = 50.70 \text{ in.}$$

$$s_X^2 = \frac{1}{29}\left[(43.30 - 50.70)^2 + \cdots + (43.11 - 50.70)^2\right] = 57.34$$

$$s_X = 7.57 \text{ in.}$$

$$s_{\bar{x}} = \frac{7.57}{\sqrt{29}} = 1.41 \text{ in.}$$

The sampling error $s_{\bar{x}}$ measures the epistemic uncertainty in estimating the average annual rainfall intensity.

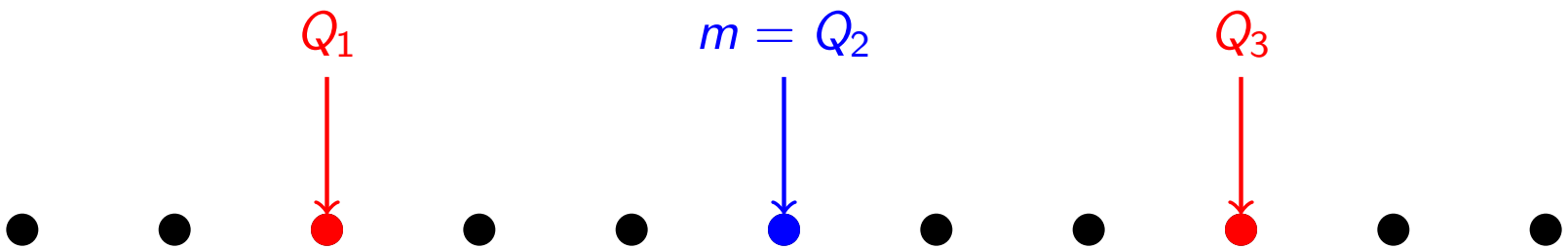
Quantiles

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

Questions

1. Quintiles ($QU_1, QU_2, \ldots$) partition a distribution into 5 equal groups.How many quintiles are there?
2. The second quartile Q_2 can be expressed as which percentile?[a]

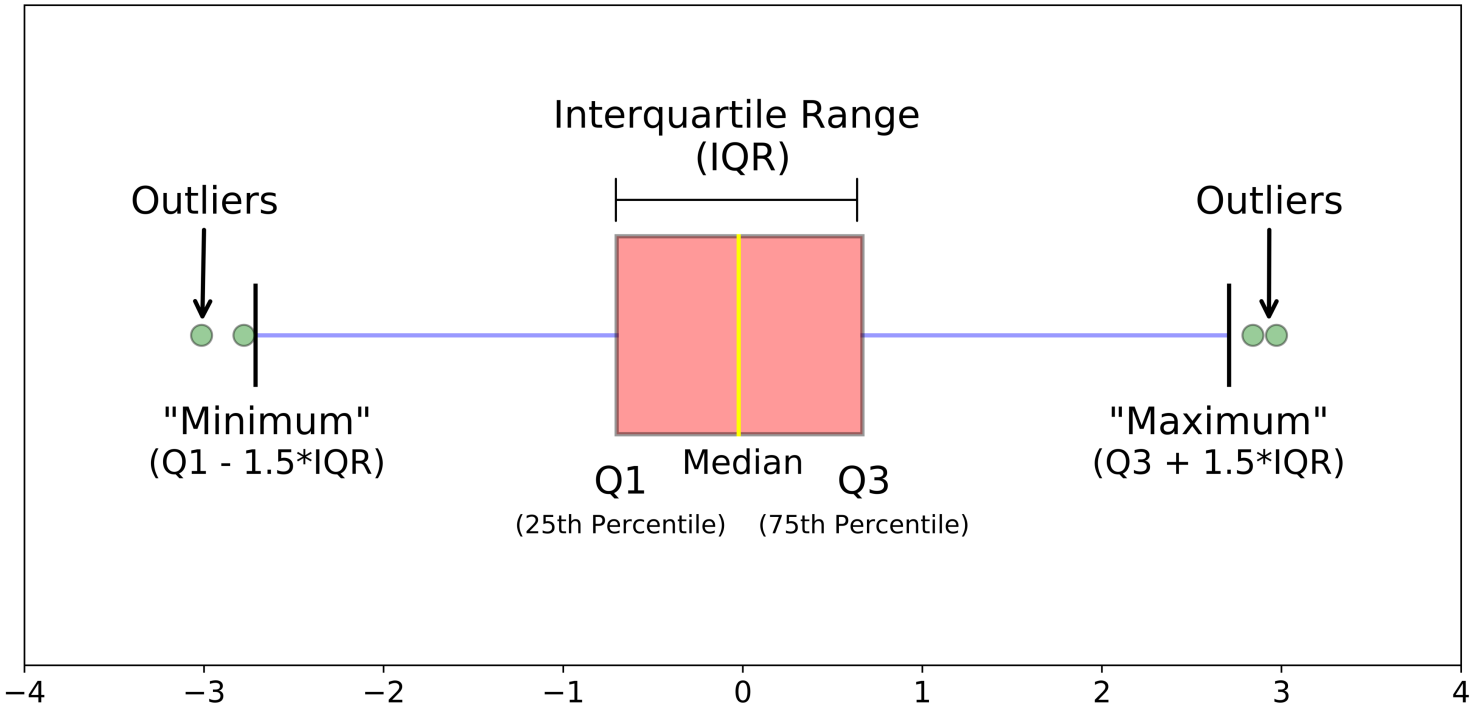
Answers: Q1: There are 4 quintiles; Q2: $Q_2 = P_{50}$ (fiftieth percentile)

[a] Recall that a quartile splits a sample into 4 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

Python

We will begin our introduction to Python via basic statistical analyses.
The platform will be Google Colab.

Summarizing data

Example 10: 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.

Recap

- Pre-survey review
- Python Introduction:
  - Summarizing data
  - Visualizing data
  - Histograms
  - Boxplots