# CEE 260/MIE 273: Probability and Statistics in Civil Engineering
# M1a: Data and Sampling

**Prof. Oke**

UMass Amherst

College of Engineering

September 2, 2025

# Outline

1. Probability, Statistics and Uncertainty

2. Frequency and Histograms

3. Outlook

# Welcome

This course is designed to introduce you to probability and statistics and the role they play in engineering. Most of the applications will be based in the civil, environmental, industrial and mechanical engineering disciplines.

After today's class, please take the time to do the following:

- Fill out the class pre-survey at `https://forms.gle/rzvaVxdujGk1TQiX9`
- Install JupyterLab on your laptop computer. Most of you may find it easiest to install this via Anaconda. We will go over this in class in detail on Thursday.

# Probability and statistics

Probability vs. Statistics[1]:

- **Probability** *deals with predicting the likelihood of future events, whereas* **statistics** *involves the analysis of the frequency of past events.*
- **Probability** *is primarily a theoretical branch of mathematics that studies the consequences of mathematical definitions.* **Statistics** *is primarily an applied branch of mathematics that tries to make sense of observations in the real world.*

*...*

*In summary, probability theory enables us to find the consequences of a given ideal world, whereas, statistical theory permits us to measure the extent to which our world is ideal.*

[1] S. Skiena (2001). Calculated Bets: Computers, Gambling, and Mathematical Modeling to Win. Cambridge University Press

# Probability and statistics in engineering

In engineering, probability and statistics enable us to:

- Describe phenomena
- Develop robust design and decision making procedures
- Quantify uncertainty and risk

# Uncertainty in engineering

Helpful to consider 2 broad categories of uncertainty:

## Aleatory uncertainty

- Data-based
- Represents inherent variability of a process or phenomenon
- This variability can be portrayed via a histogram/frequency diagram or scattergram (in the case of two variables)

## Epistemic uncertainty

Scientific uncertainty in the model of a process

- Knowledge-based
- Due to imperfect representations (idealized models)
- Resulting in inaccurate estimates e.g. central values (mean, median), etc

# Example 1: The unknown die

*Two students are presented with the results of four previous rolls of an unseen die: 2, 3, 3 and 4. What is the model for this die?*

**Student A** relies on prior knowledge. Most dice have six faces that are equally likely. Student A constructs a model in which the aleatory uncertainty is given by a **uniform distribution** with values between 1 and 6.

**Student B** takes an empirical approach, developing a model based on the assumption that the die is five-faced and loaded such that it rolls 3 most often, 2 and 4 less often, and 1 and 5 least often.

# Example 1: The unknown die (cont.)

| | Probability | |
|---|---|---|
| Value | **Model A** | **Model B** |
| 1 | $1/6$ | 0.1 |
| 2 | $1/6$ | 0.2 |
| 3 | $1/6$ | 0.4 |
| 4 | $1/6$ | 0.2 |
| 5 | $1/6$ | 0.1 |
| 6 | $1/6$ | 0.0 |

Both models represent *epistemic uncertainty* in the properties of the die. With additional data (further rolls), the correctness of the two models can be accurately judged.

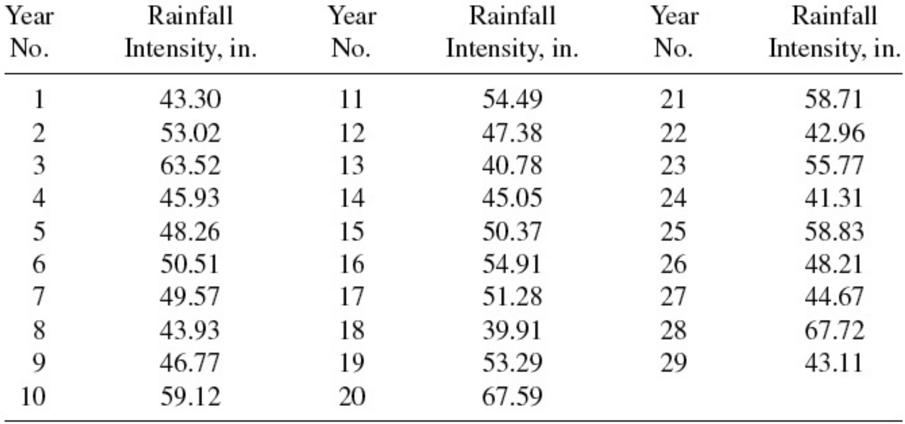
The *aleatory uncertainty* remains the same, but our quantification can improve with more information.

# Activity 1: Birthday Paradox

- Before we do anything else, I want everyone to make a prediction. We have 160 people in this class. What do you think is the probability that at least two people in this room share the exact same birthday (same month and day)?
- Write down a percentage from 0% to 100%. Be honest about your first instinct—there are no wrong answers here!

# Example 2: Rainfall intensity

Rainfall in a watershed area is recorded over a period of 29 years



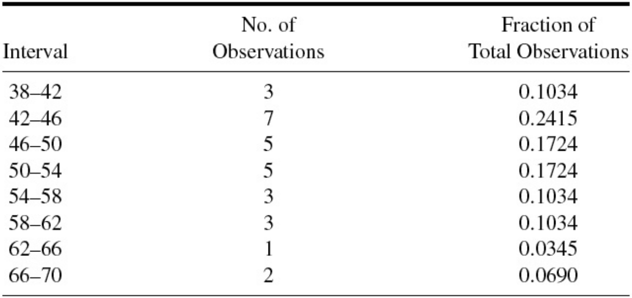
| Year No. | Rainfall Intensity, in. | Year No. | Rainfall Intensity, in. | Year No. | Rainfall Intensity, in. |
|---|---|---|---|---|---|
| 1 | 43.30 | 11 | 54.49 | 21 | 58.71 |
| 2 | 53.02 | 12 | 47.38 | 22 | 42.96 |
| 3 | 63.52 | 13 | 40.78 | 23 | 55.77 |
| 4 | 45.93 | 14 | 45.05 | 24 | 41.31 |
| 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 | | |

*Plot a histogram of the data*

# Example 2: Rainfall intensity (cont.)

Basic steps in histogram plotting:

- Find the range of the data
- Divide range into convenient number of intervals
- Count number of observations within in each interval

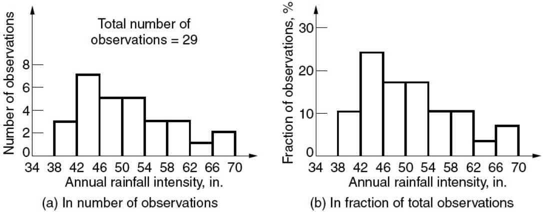


| Interval | No. of<br>Observations | Fraction of<br>Total Observations |
|---|---|---|
| 38–42 | 3 | 0.1034 |
| 42–46 | 7 | 0.2415 |
| 46–50 | 5 | 0.1724 |
| 50–54 | 5 | 0.1724 |
| 54–58 | 3 | 0.1034 |
| 58–62 | 3 | 0.1034 |
| 62–66 | 1 | 0.0345 |
| 66–70 | 2 | 0.0690 |

# Activity 2: Data About Us

# Example 2: Rainfall intensity (cont.)

- Plot counts (or fraction) on the *ordinate* ($y$) axis and intervals on the *abscissa* ($x$)

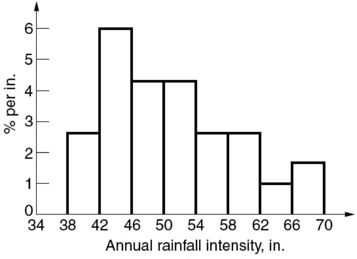


(a) In number of observations

(b) In fraction of total observations

# Example 2: Rainfall intensity (cont.)

For further analyses, comparing the empirical frequency to a theoretical frequency distribution is necessary. In such cases, we find the empirical frequency function by obtaining a frequency diagram with an area of 1.

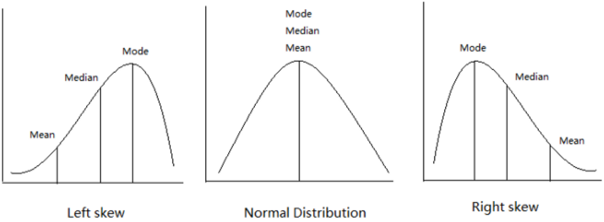


(c)

# Mean, median, mode: measures of central tendency

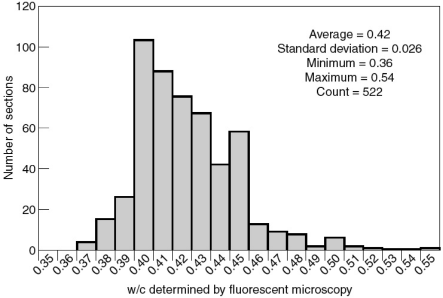
These indicate where the *center* of a distribution is located.

- **Mean**: average of a sample
- **Median**: central value in an ordered[2] sample (or average of central values in even-numbered sample)
- **Mode**: Most frequently occuring value



[2] Arranged from least to greatest

# Example 3: Water-cement ratio of concrete specimens



A lower ratio indicates greater stiffness and strength

## Questions

- What is the mode of this sample? $\boxed{0.39}$
- What is the *range* of this sample? Range $= 0.54 - 0.36 = 0.18$

# Example 4: Impact speeds of passenger car accidents

*Is this distribution left-skewed or right-skewed?*

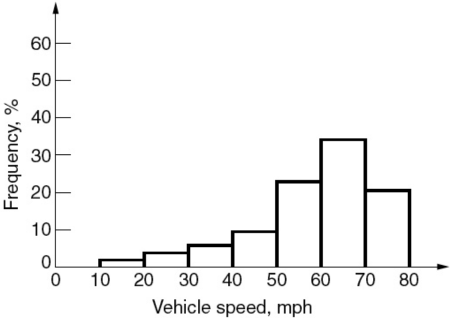
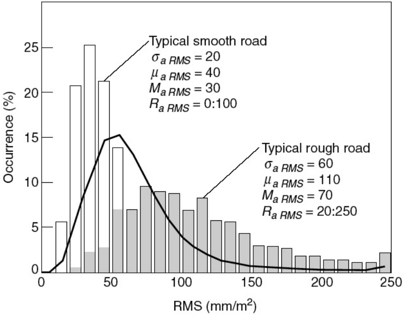


**Figure 1.18** Impact speeds of passenger car accidents

This distribution is left-skewed (long left tail; mean left of mode)

# Example 5: Road roughness profiles

Surface roughness is measured by the root mean square (RMS) of height deviations per given area. Here, we are given the distributions for two samples and fitted *mixture distribution*. $\sigma$ is the standard deviation, $\mu$ is the mean.

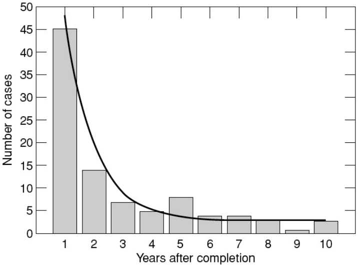


## Questions

- What does $M$ represent? Mode
- What does $R$ represent? Range

# Example 6: Dam failures in the United States

In order to predict the probability of a random variable, we can fit theoretical probability distributions to discrete data.



*What is the associated theoretical PDF of the histogram in the above figure?*

# Recap

- Course introduction (take survey, install Jupyterlab)
- Uncertainty in engineering
- Histograms and scattergrams

# Summary statistics: key definitions

Sample: finite set of $n$ observations

Sample mean: average of the sample; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample variance: measure of dispersion; $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Sample standard deviation (SSD): square root of sample variance

Sampling error: measure of how well $\bar{x}$ estimates population mean (standard deviation of sample mean)[3]; $s_{\bar{x}} = \frac{s_X}{\sqrt{n}}$

Coefficient of variation $\delta_X = \frac{s_X}{\bar{x}}$

[3] In measurement theory, the same formulation defines the standard error