Basics of Statistics in Python

Lecture 3 – Monday September 1, 2025

Outline

- Working with Data in Python
- Summarizing Data
- Probability Distributions
- Measures of Central Tendency
- Measures of Dispersion
- Comparing Groups and Distributions

Outline

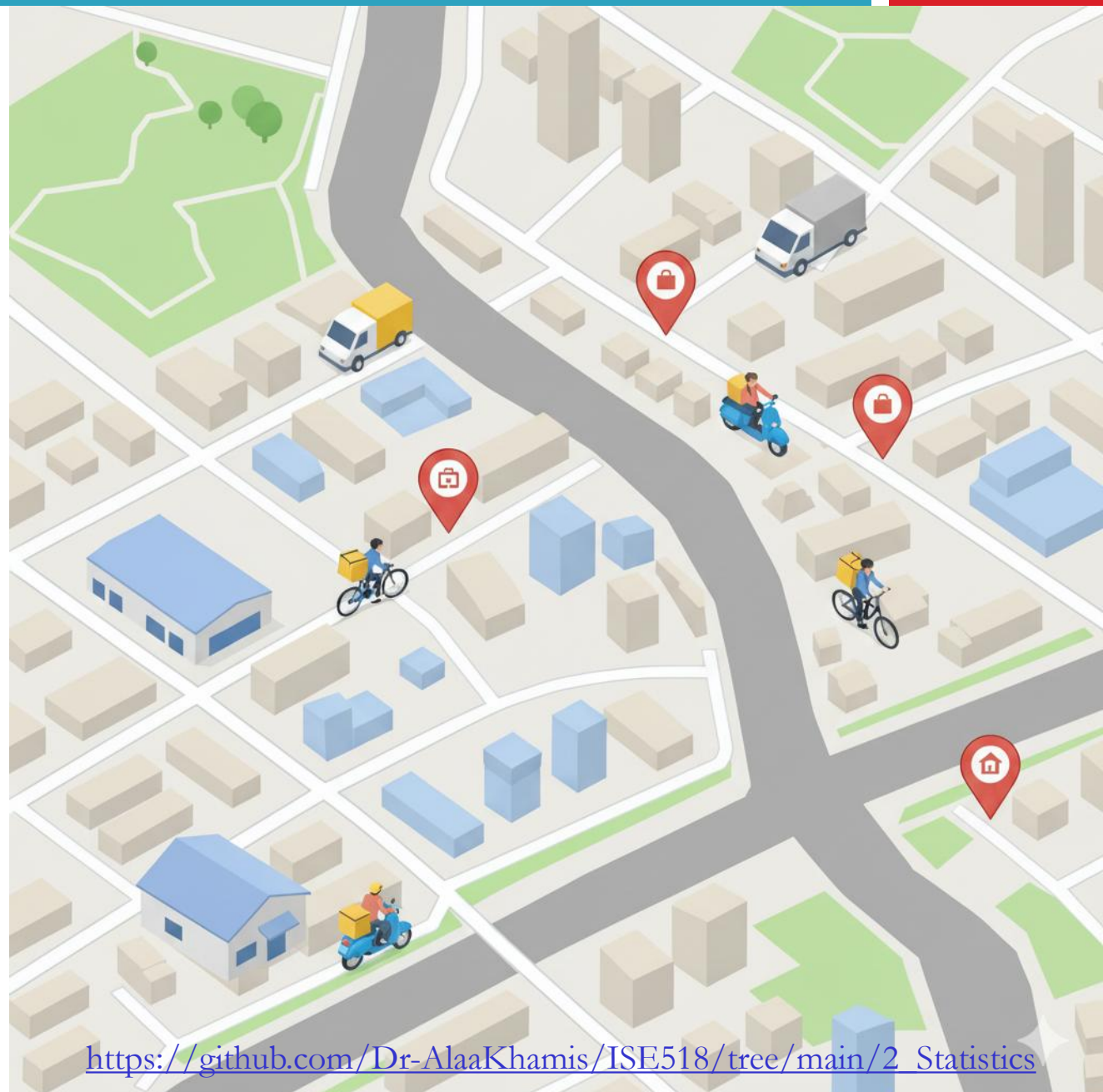
- **Working with Data in Python**
- Summarizing Data
- Probability Distributions
- Measures of Central Tendency
- Measures of Dispersion
- Comparing Groups and Distributions

Working with Data in Python

- Data: Last-mile delivery
  - order_id
  - region_id,
  - city,
  - courier_id,
  - lng,
  - lat,
  - aoi_id,
  - aoi_type,
  - accept_time,
  - accept_gps_time,
  - accept_gps_lng,
  - accept_gps_lat,
  - delivery_time,
  - delivery_gps_time,
  - delivery_gps_lng,
  - delivery_gps_lat,
  - ds

LaDe: The First Comprehensive Last-mile Delivery Dataset from Industry

For more information about last-mile delivery, read: Alaa Khamis, "Last-Mile Delivery: Definition and Trends", Medium, 2021.



https://github.com/Dr-AlaaKhamis/ISE518/tree/main/2_Statistics

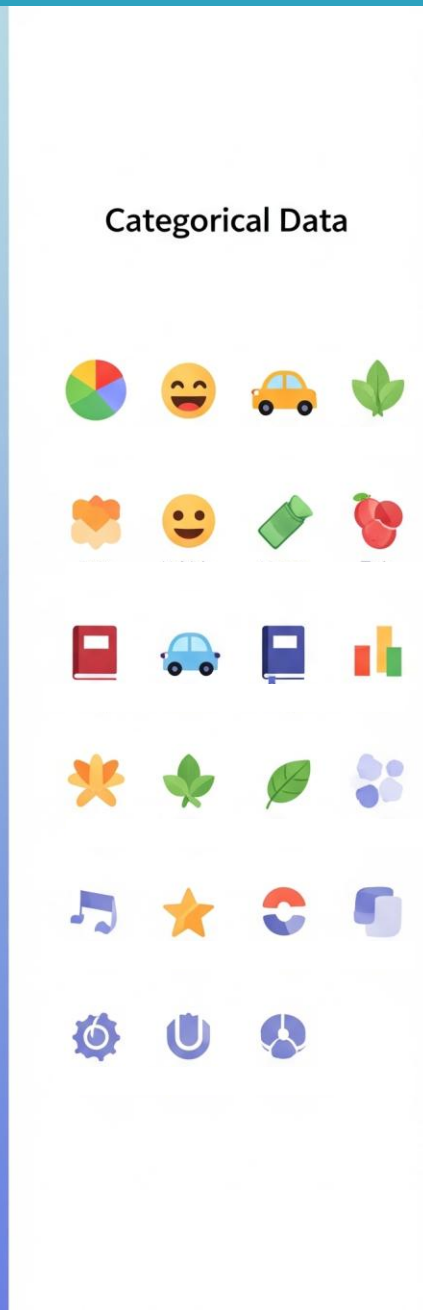
Working with Data in Python

- Data: Bottling plant
  - event_date: date of failure
  - line: production line (A, B, C)
  - asset_id: conveyor motor id
  - failure_mode: Mechanical, Electrical, Misalignment
  - time_to_failure_days: days since last failure (proxy for MTBF)
  - repair_time_hours: hours to repair (MTTR component)
  - downtime_hours: total downtime per event

Working with Data in Python

- Tabular data (CSV, Pandas DataFrame)
- Types of fields: numerical, categorical, date/time

https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/2_Statistics/Intro_Stats_RM_Section_1.ipynb



Outline

- Working with Data in Python
- **Summarizing Data**
- Probability Distributions
- Measures of Central Tendency
- Measures of Dispersion
- Comparing Groups and Distributions

Summarizing Data

- Frequency distribution of failures and repairs
- Visualizing with histograms (numerical data, e.g., time-to-failure)
- Visualizing with bar charts or pie charts (categorical data, e.g., failure modes)
- Hands-on: Create a histogram and pie chart for maintenance data

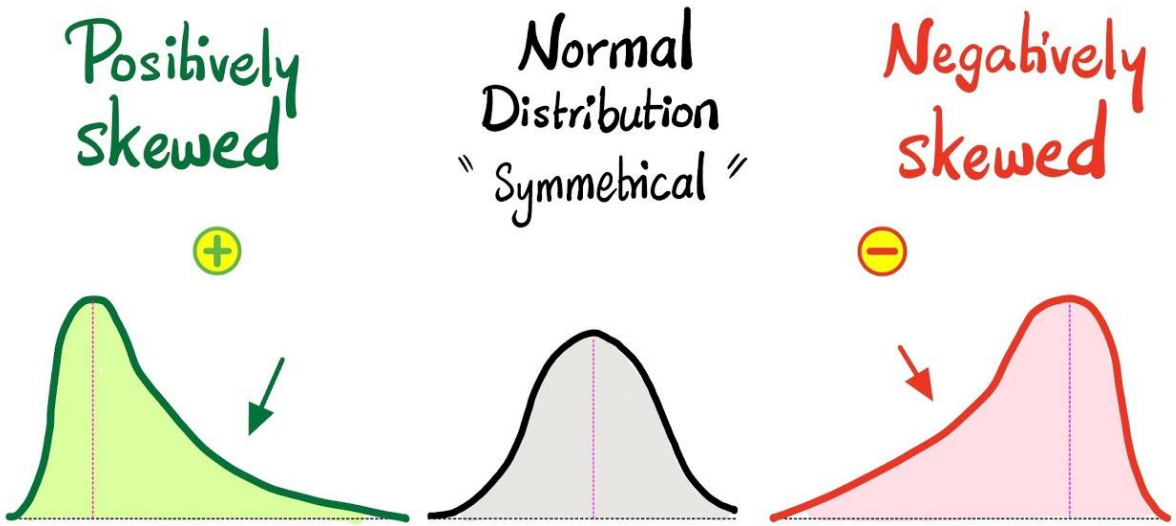
https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/2_Statistics/Intro_Stats_RM_Section_2.ipynb

Outline

- Working with Data in Python
- Summarizing Data
- **Probability Distributions**
- Measures of Central Tendency
- Measures of Dispersion
- Comparing Groups and Distributions

Probability Distributions

- Concept of distribution in reliability data
- Normal distribution and its relevance
- Example: Compare a normally distributed simulated dataset vs. skewed failure-time data



https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/2_Statistics/Intro_Stats_RM_Section_3.ipynb

Outline

- Working with Data in Python
- Summarizing Data
- Probability Distributions
- **Measures of Central Tendency**
- Measures of Dispersion
- Comparing Groups and Distributions

Measures of Central Tendency

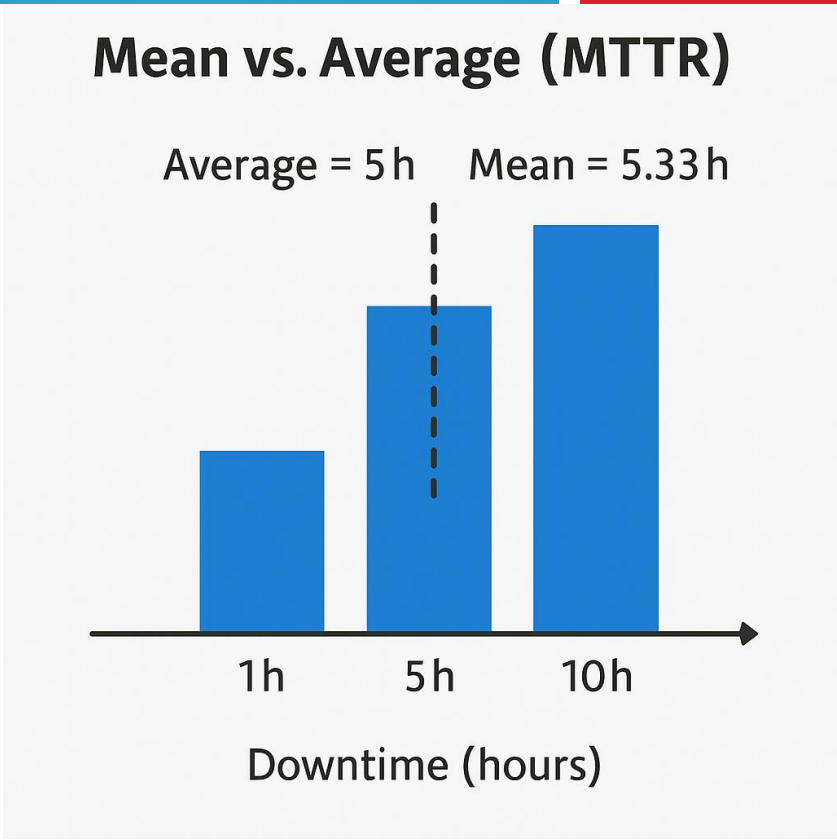
- **Mean**

"mean" and "average" are often used interchangeably, but there's a subtle difference in nuance and context. In statistics and predictive maintenance, it's safer to use "mean".

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n}$$

where

- x_i = individual time measurements (e.g., repair times or time between failures)
- n = total number of observations



Example: Suppose a conveyor motor had the following repair times after failures (in hours):

$x = [2,3,4,5,6]$. $Mean\ MTTR = \frac{2+3+4+5+6}{5} = \frac{20}{5} = 4\ hours$. This means, on average, the motor takes 4 hours to be repaired after a failure.

Measures of Central Tendency

- **Median (Middle value)**
  - If n is odd: $\text{Median} = x_{(n+1)/2}$
  - If n is even: $\text{Median} = \frac{x_{(n+1)/2} + x_{((n/2)+1)}}{2}$

where x_i values are sorted in ascending order.

Example: Repair times (sorted):[2,3,4,5,6]

Number of observations, n=5 (odd)

Median = 3rd value = 4 hours

Interpretation: Half of the repairs are completed in less than 4 hours, and half take longer.

Measures of Central Tendency

What are the mean and median income of the passengers in the bus? ☺

Measures of Central Tendency

- **Mode (Most frequent value)**

The value that occurs most frequently in the dataset.

Example: Repair times: [2,3,4,4,5,6,4]

The number 4 occurs 3 times, more than any other value

Mode MTTR = 4 hours

Interpretation: The most common repair duration for the motor is 4 hours.

- Hands-on: Calculate mean time to repair (MTTR) and median time between failures

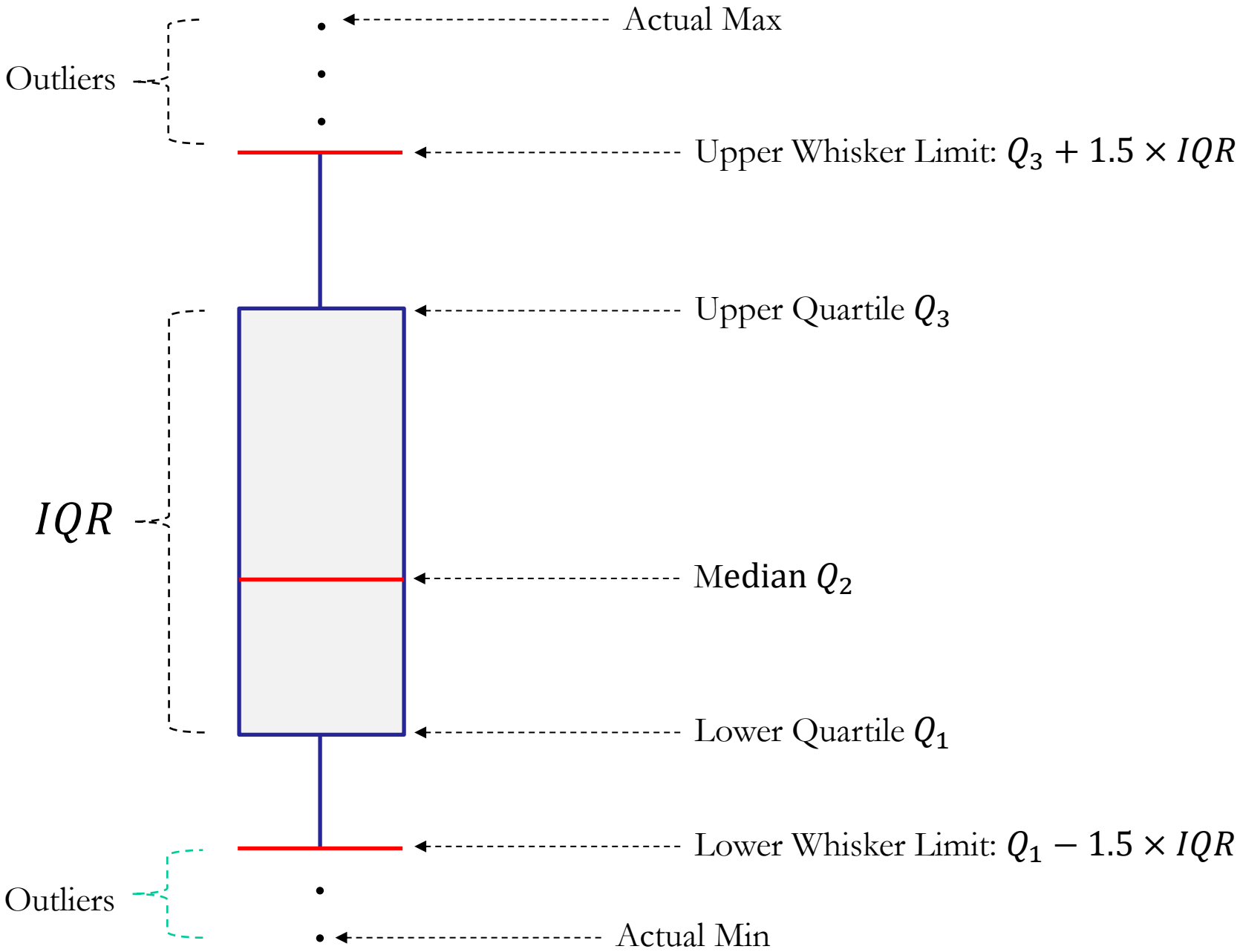
https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/2_Statistics/Intro_Stats_RM_Section_4.ipynb

Outline

- Working with Data in Python
- Summarizing Data
- Probability Distributions
- Measures of Central Tendency
- **Measures of Dispersion**
- Comparing Groups and Distributions

Measures of Dispersion

- **Box plot**



Measures of Dispersion

- **Range**

The difference between the maximum and minimum values in a dataset.

$$Range = x_{max} - x_{min}$$

Example:

Repair times (hours): [2,3,4,5,6]

Range=6−2=4 hours

Interpretation: The repairs vary by 4 hours from the shortest to the longest.

Measures of Dispersion

- **Interquartile Range (IQR)**

The range of the middle 50% of the data (between 25th percentile Q_1 and 75th percentile Q_3).

$$IQR = Q_3 - Q_1$$

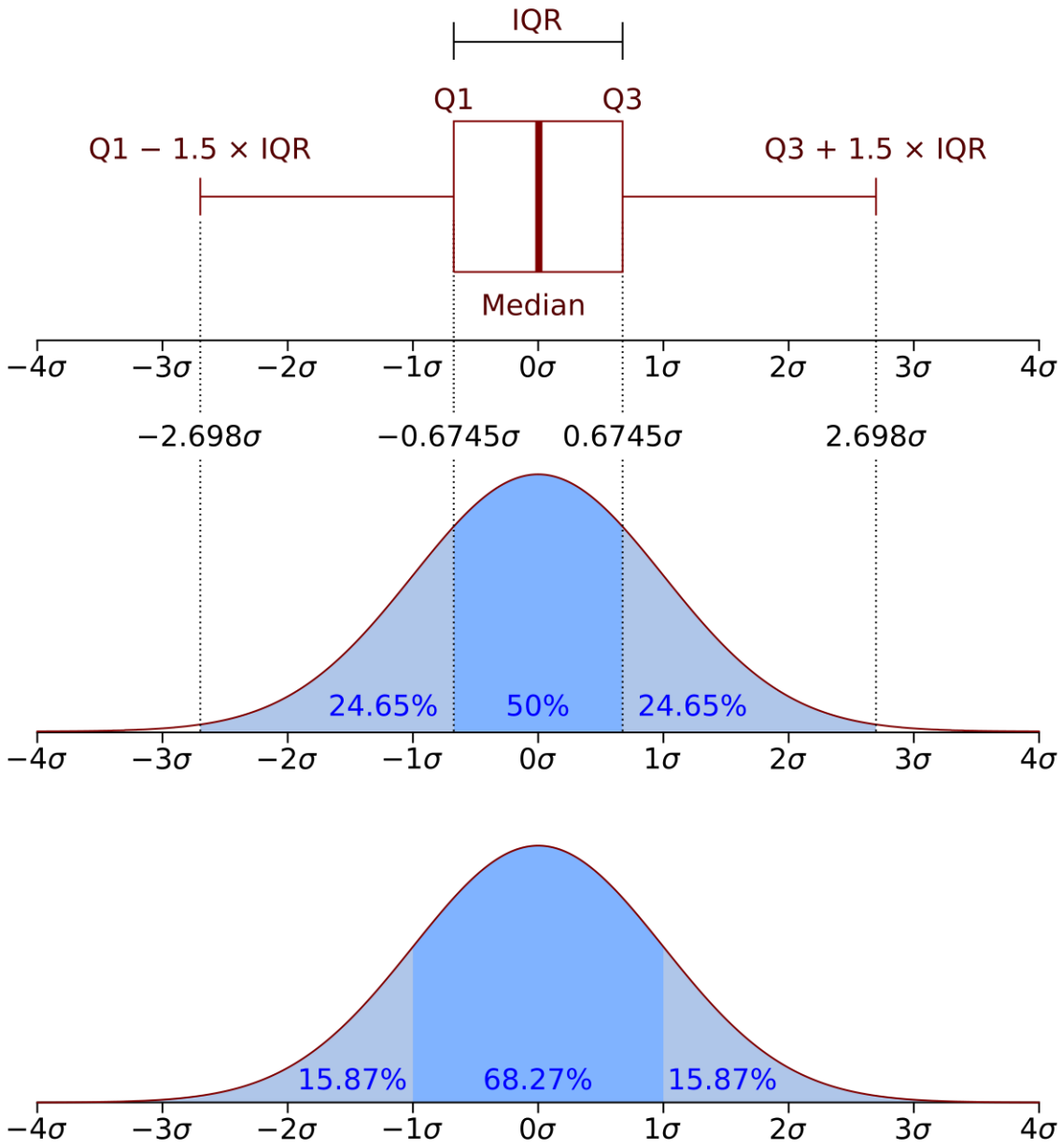
Example:

Repair times (hours): [2,3,4,5,6]

Q_1=2.5 (25th percentile)

Q_3=5 (75th percentile)

Interpretation: The central 50% of repair times span 2.5 hours, showing typical variation without extreme values.



Measures of Dispersion

- **Variance**

Measures how far each observation is from the mean (average squared deviation).

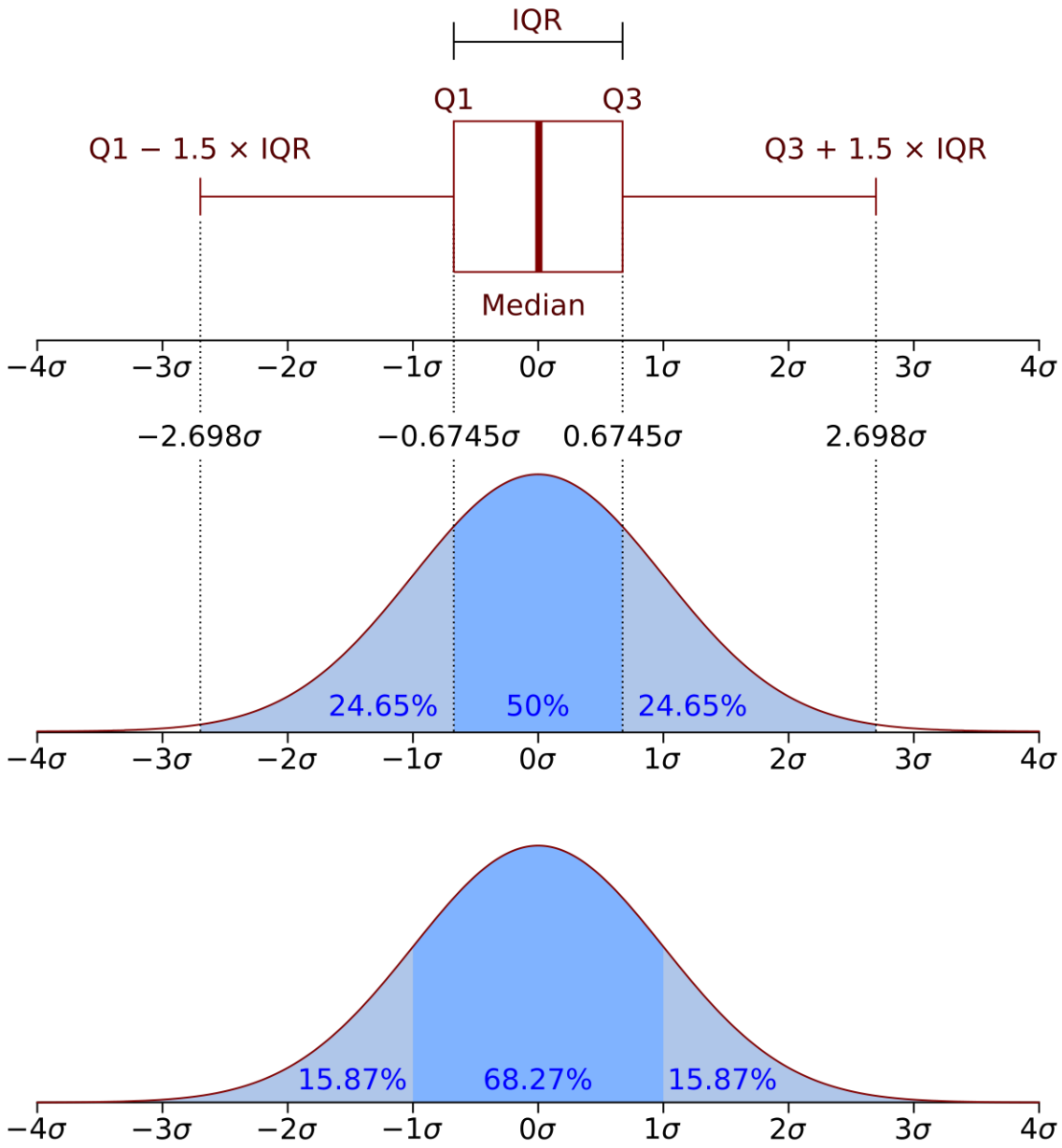
$$Variance\ (\sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

where x_i are the repair times, $\bar{x}$ is the mean, and n is the number of observations.

Example:

Repair times (hours): [2,3,4,5,6], $\bar{x}$ =4

$$\sigma^2 = \frac{4 + 1 + 0 + 1 + 4}{5} = 2\ hours^2$$



Measures of Dispersion

- **Standard Deviation**

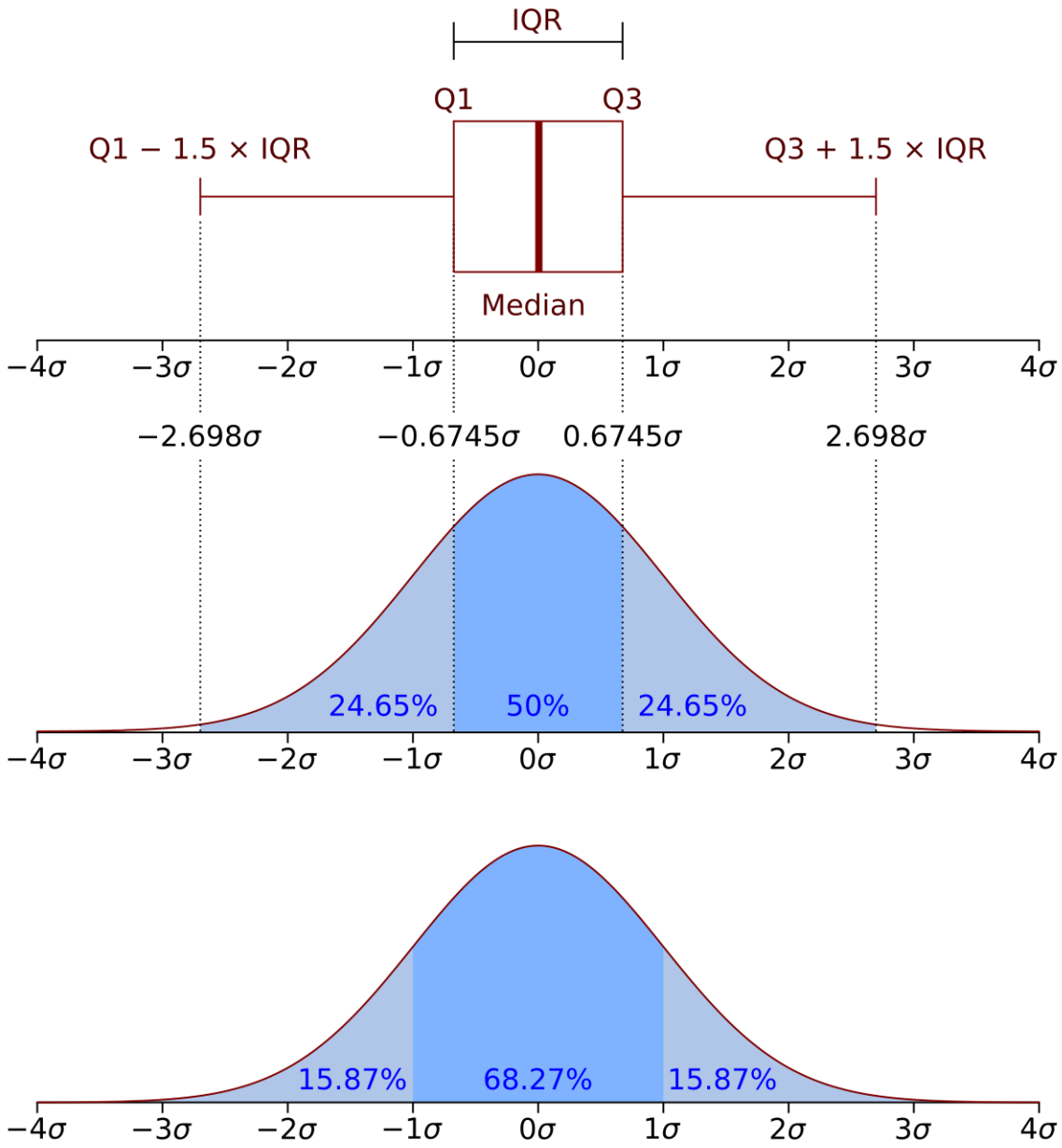
The square root of variance; measures the typical deviation from the mean in the original units.

$$SD = \sqrt{\sigma^2}$$

Example:

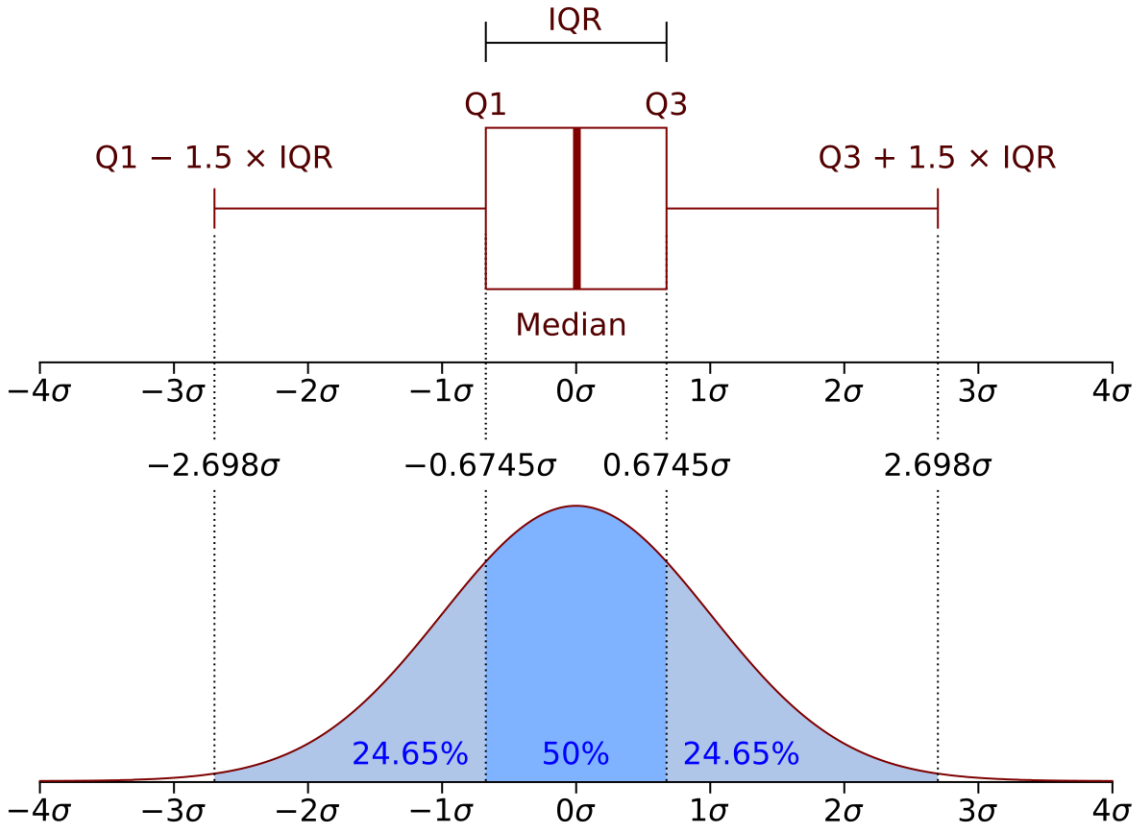
SD= $\sqrt{2}$=1.414 hours

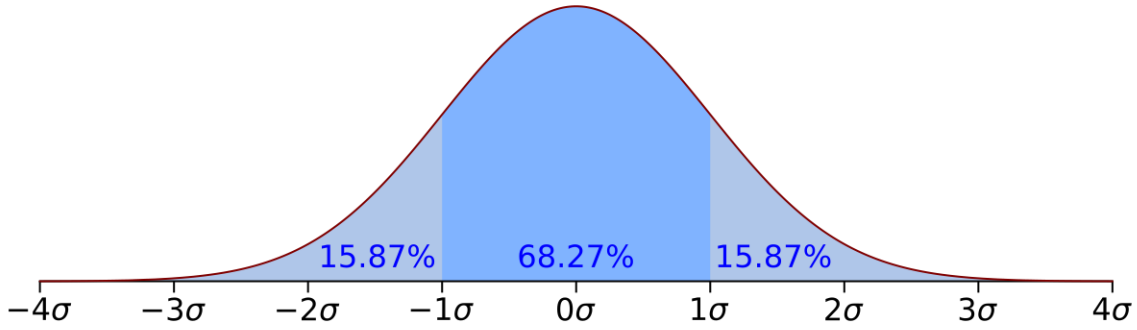
Most repairs deviate from the mean by approximately 1.4 hours, giving a sense of consistency in maintenance time.



Measures of Dispersion

Metric	What it Tells You (MTTR / MTBF context)
Range	Full spread of repair or failure times
IQR	Typical variability without outliers
Variance	How widely repair/failure times are spread around the mean
SD	Average deviation in original units, easier to interpret than variance





- Hands-on: Plot and compare time-to-failure data from two machine types

https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/2_Statistics/Intro_Stats_RM_Section_5.ipynb

Outline

- Working with Data in Python
- Summarizing Data
- Probability Distributions
- Measures of Central Tendency
- Measures of Dispersion
- **Comparing Groups and Distributions**

Comparing Groups and Distributions

- **Hypothesis Testing:** A method to decide if a claim about a population is likely true based on sample data.
- **P-Value:** The probability of observing the data (or more extreme) if the null hypothesis is true. Small p-values suggest rejecting the null hypothesis.
- **Shapiro-Wilk Test:** Checks if a data sample is normally distributed. Null hypothesis: the data follows a normal distribution.
- **Student's t-test:** Compares the means of two independent samples. Null hypothesis: the sample means are equal.
- **Mann-Whitney U Test:** Compares the distributions of two independent samples without assuming normality. Null hypothesis: the distributions (or means) are equal.

Comparing Groups and Distributions

Population

Sample

Null vs. Alternative Hypothesis

Null Hypothesis

H_0

A statement about a population parameter.

We test the likelihood of this statement being true in order to decide whether to accept or reject our alternative hypothesis.

Can include =, ≤, or ≥ sign.

Alternative Hypothesis

H_a

A statement that directly contradicts the null hypothesis.

We determine whether or not to accept or reject this statement based on the likelihood of the null (opposite) hypothesis being true.

Can include a ≠, >, or < sign.

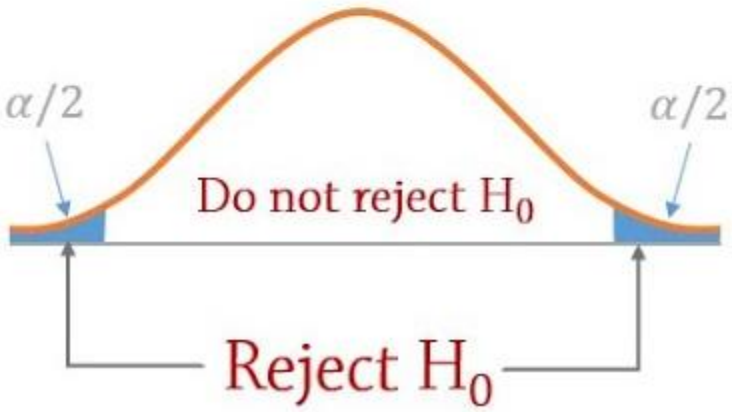
ThoughtCo.

Hypothesis Testing

Two-tailed

$H_0: \mu = 23$

$H_a: \mu \neq 23$

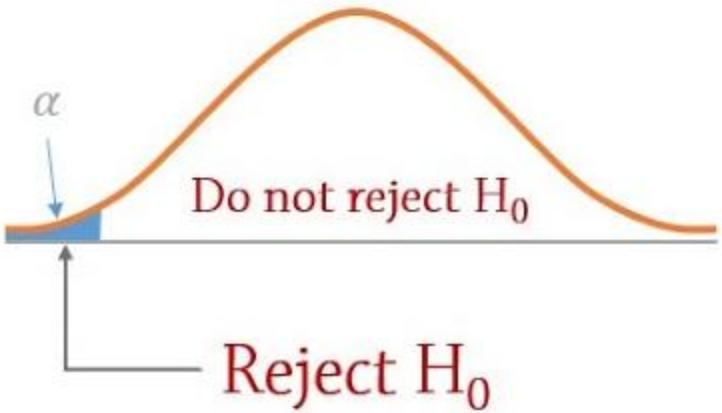


One-tailed

Left-tailed

$H_0: \mu \geq 23$

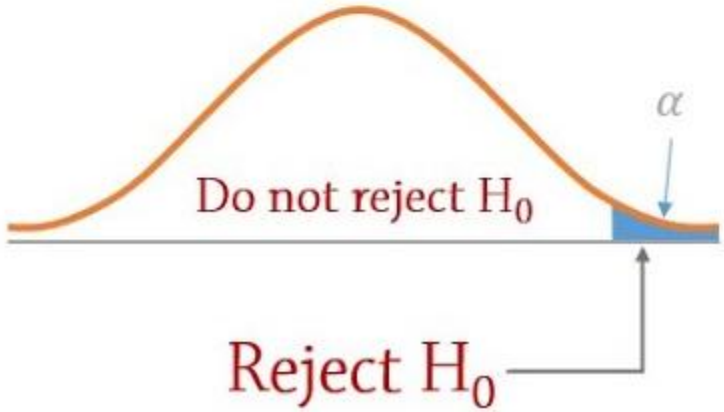
$H_a: \mu < 23$



Right-tailed

$H_0: \mu \leq 23$

$H_a: \mu > 23$



Comparing Groups and Distributions

- Hands-on: Test whether downtime differs significantly between two production lines

https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/2_Statistics/Intro_Stats_RM_Section_6.ipynb