Data Types: The Foundation of Method

¶ 1.1

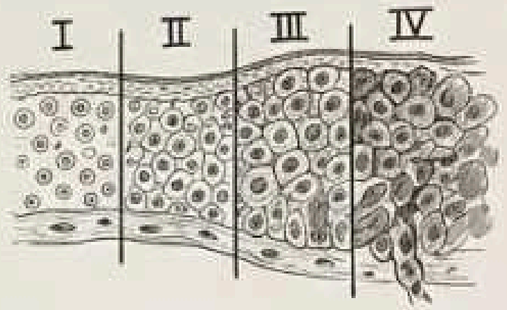
Data

- Categorical
  - Unordered



Places patients into distinct groups.

  - Ordered



- Numerical







Measured on a continuous scale.

The data type decides which methods we can use.

Gemini Notebook

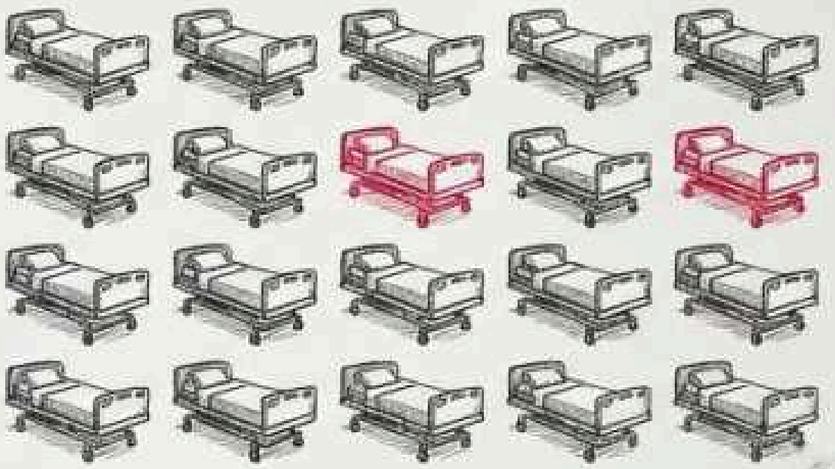
Summarizing Categorical Data: Proportions

¶ 1.2

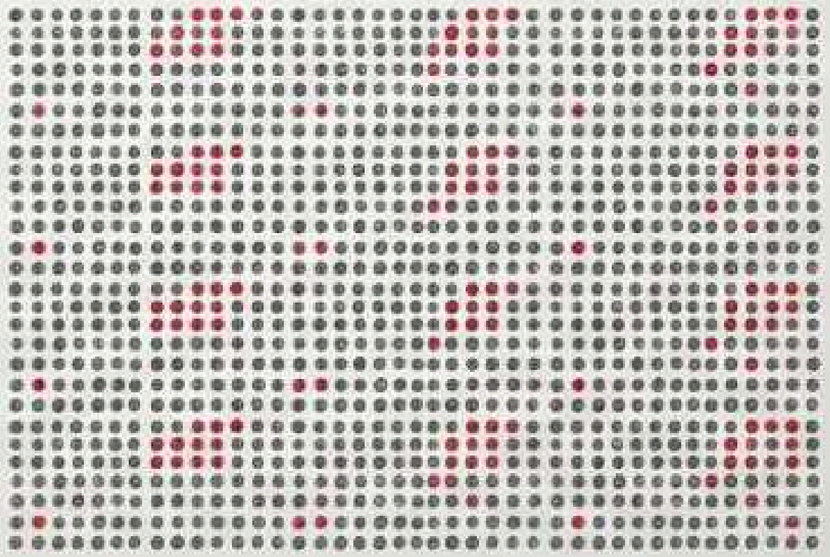
$$\text{Proportion} = \frac{\text{Number in Group}}{\text{Total (Denominator)}}$$

Example: $\frac{30}{200} = 0.15\ (15\%)$

Always report the denominator.



15% of 20 patients = Highly Uncertain



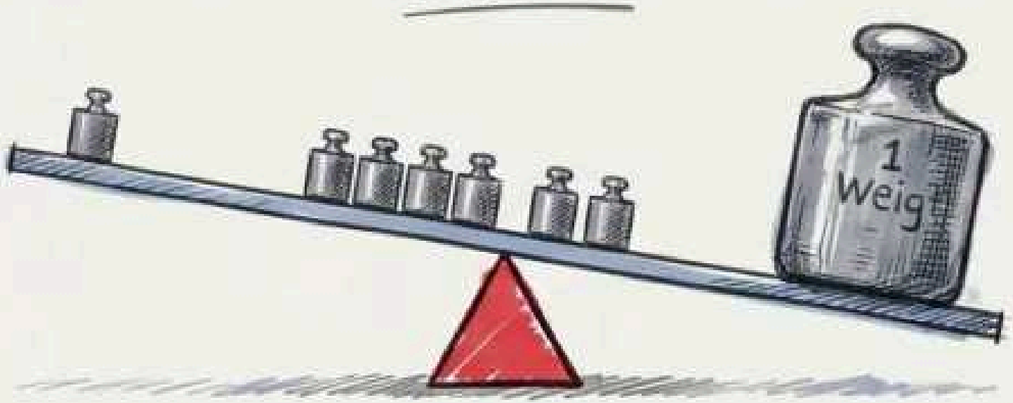
15% of 2,000 patients = Strong Statistical Confidence

Gemini Notebook

Describing Numerical Data: Centre

¶ 1.3

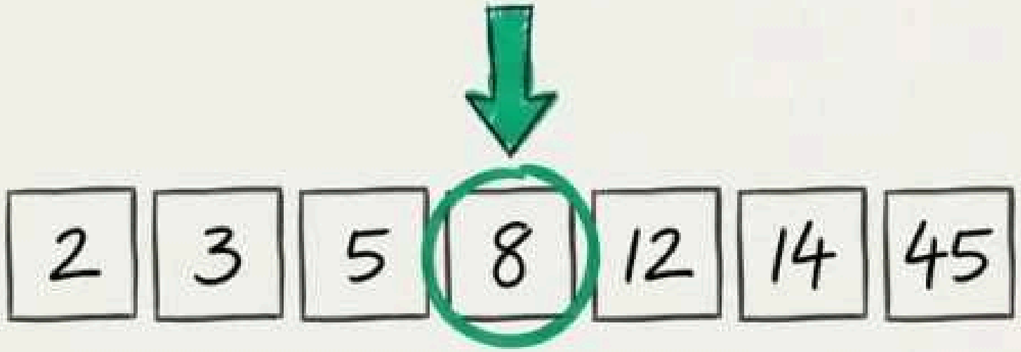
Mean



$$\bar{x} = \frac{\sum x_i}{n}$$

- Sum of values divided by count. Vulnerable to extremes.

Median



- The middle value of sorted data. Best for typical cases.

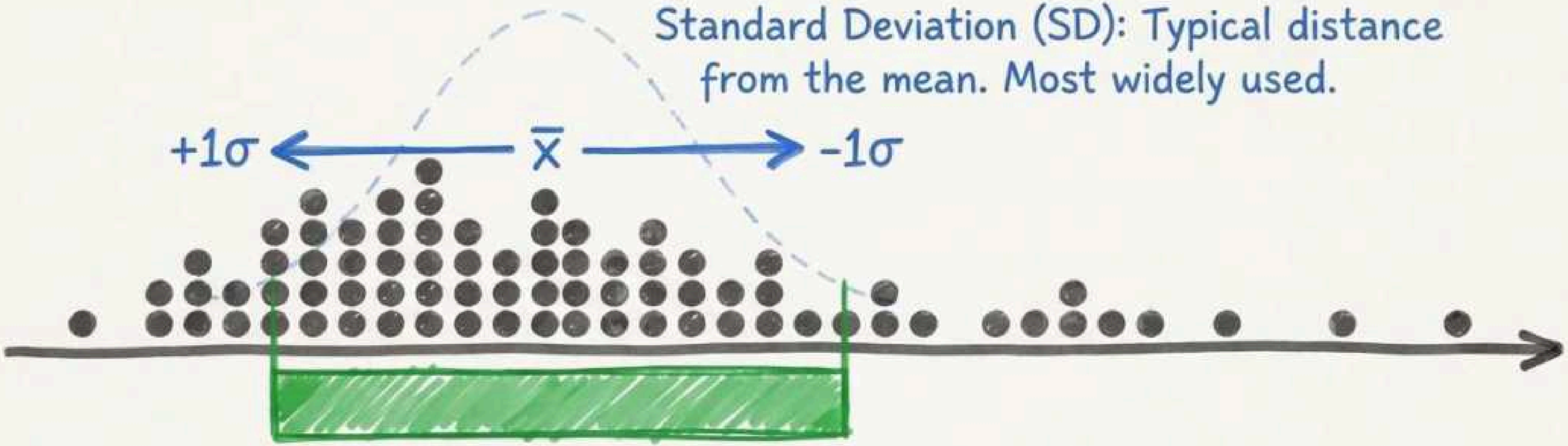
When a few values are massive (e.g., hospital stays, incomes), the mean is pulled upward; the median stays anchored to the typical case.

Gemini Notebook

Describing Numerical Data: Spread

¶ 1.4

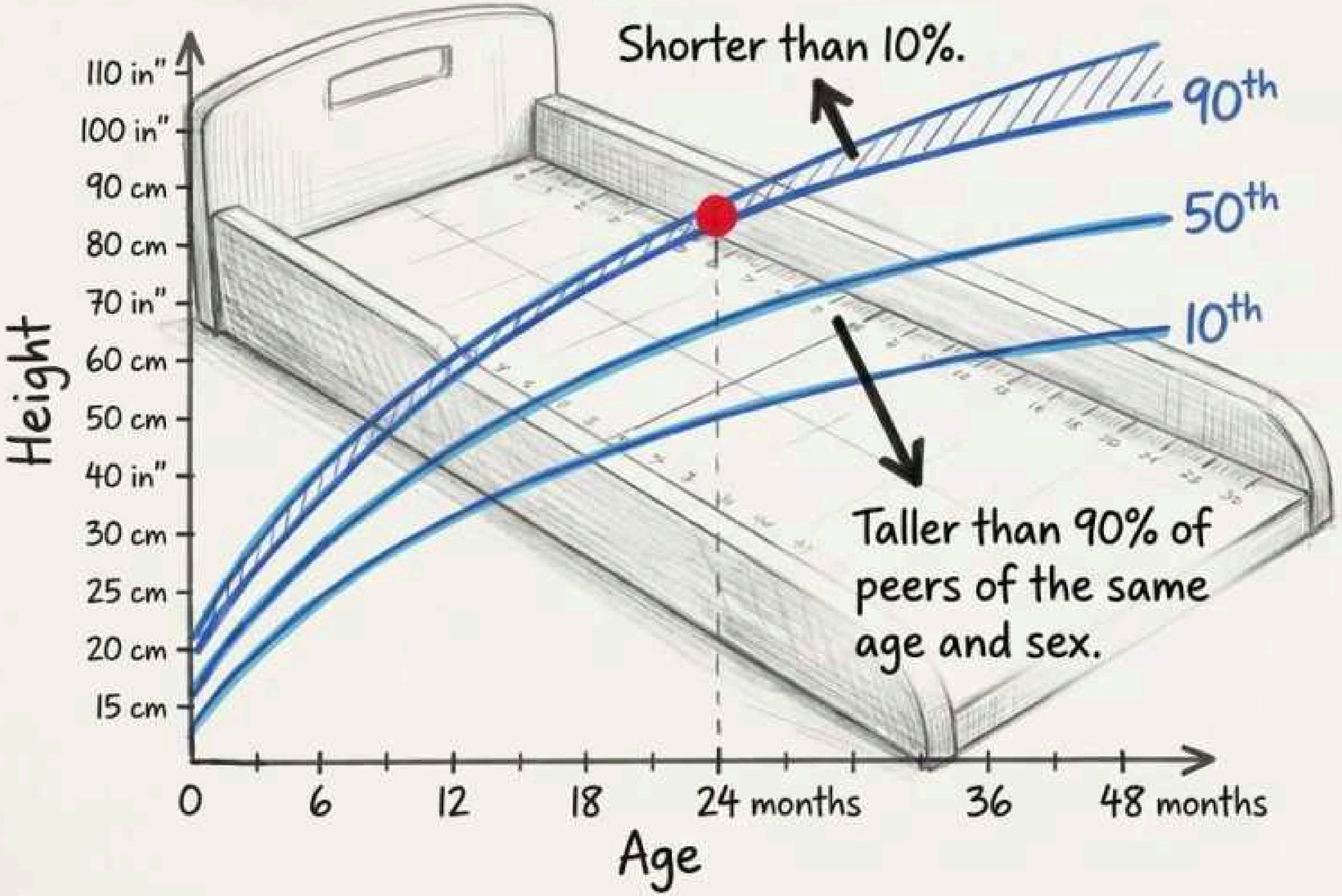
Range: Runs smallest to largest. Depends on just 2 observations.



Gemini Notebook

Contextualizing with Percentiles

¶ 1.5



Definition & Applications

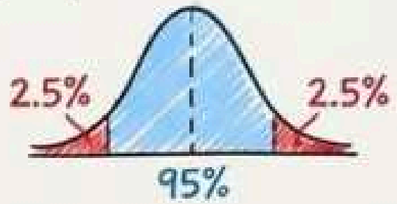
- Percentiles describe where a value sits within a distribution.
- Forms the foundation of growth charts, lab reference ranges, and clinical scoring.

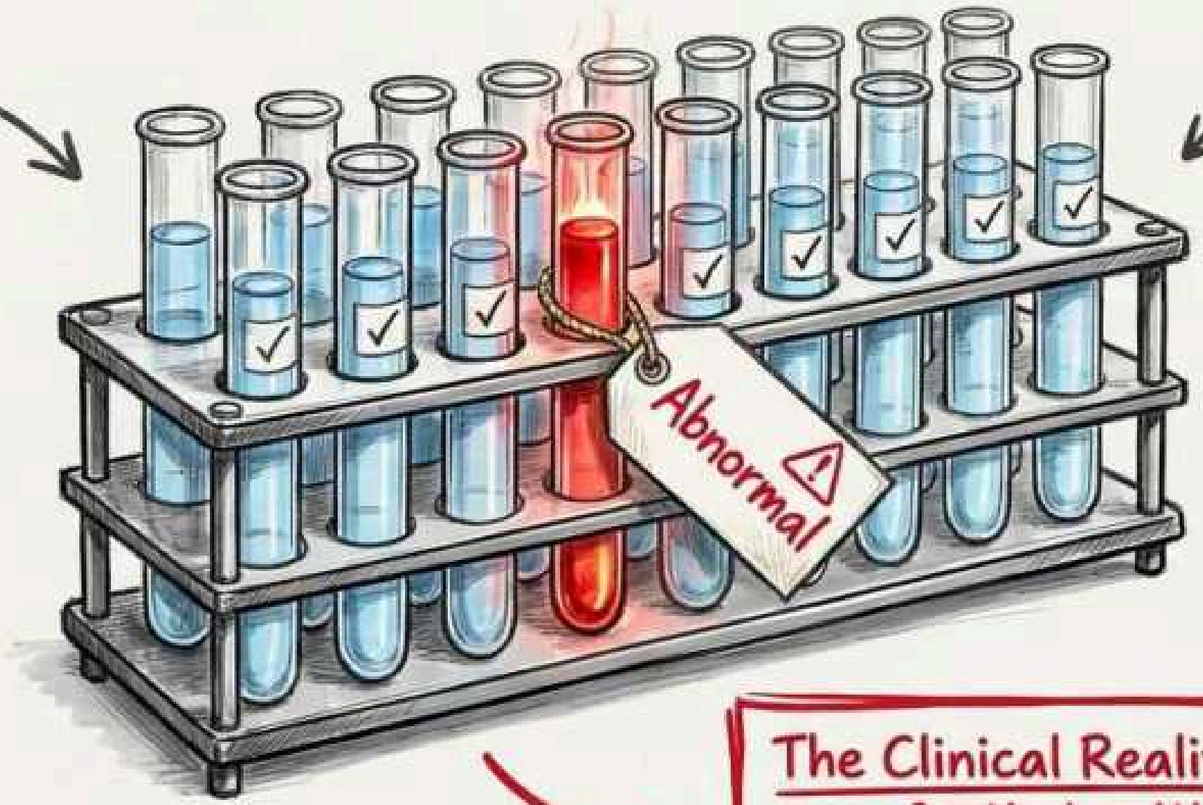
Gemini Notebook

The Danger of Multiple Testing

¶ 1.6

The 95% Rule:
A "normal range" includes the middle 95% of healthy people.





The Math:
1 in 20 (5%) healthy people fall outside the range by pure chance on a single test.

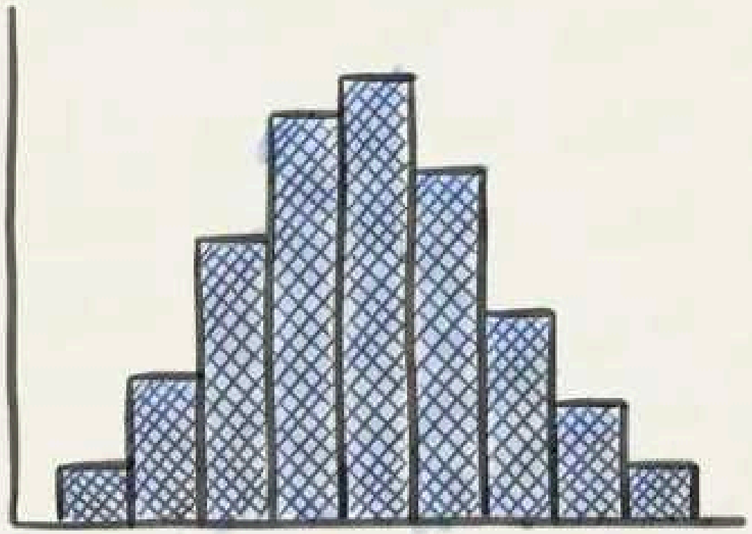
$$\frac{1}{20} = 5\%$$

The Clinical Reality: Order 20 tests on a perfectly healthy patient, and it is more likely than not (>50%) that at least one result will flag as abnormal.

Gemini Notebook

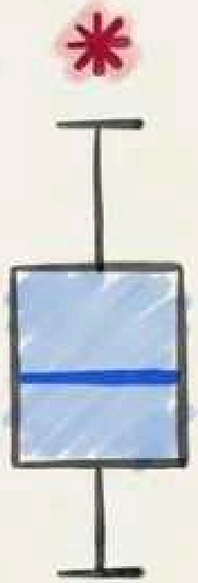
Graphing the Data

¶ 1.7



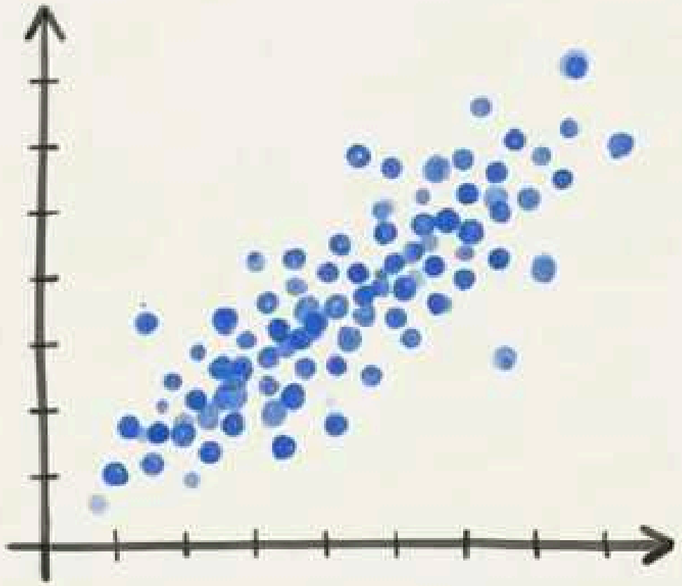
Histogram

Shows the shape of numerical data.



Box Plot

Summarizes centre, spread, and unusual values at a glance.



Scatter Plot

Exposes relationships between two numerical variables.

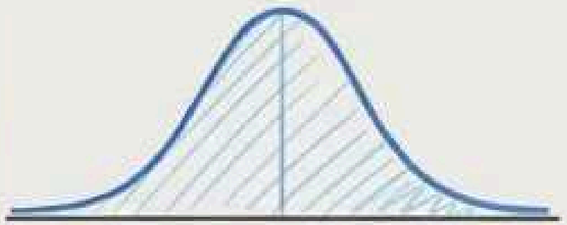
Golden Rule: Looking at the data visually before calculating anything is one of the most important habits in statistics.

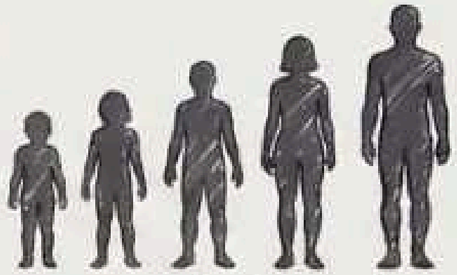
Gemini Notebook

Why Distribution Shape Matters

¶ 1.8

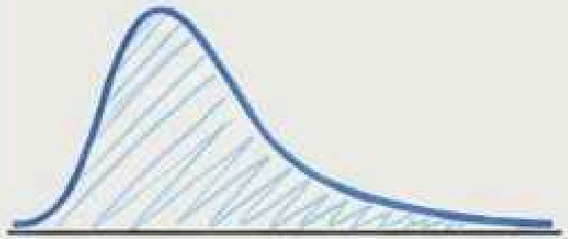
Symmetrical





Mean and median align. Common in biological measurements.

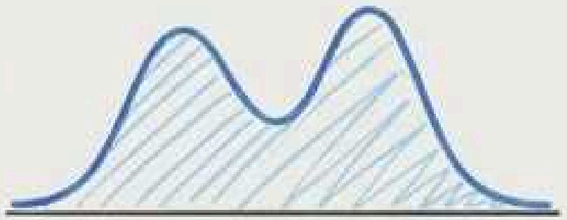
Skewed



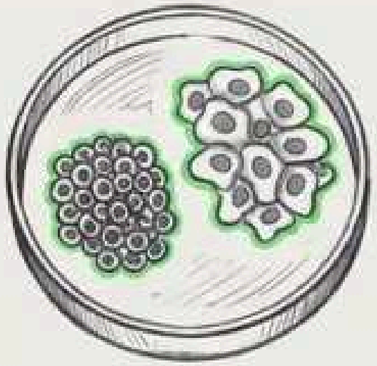


Long tail pulls the mean. Use median/IQR. (e.g., length of hospital stays)

Bimodal



Two peaks indicate two different groups have been mixed together.

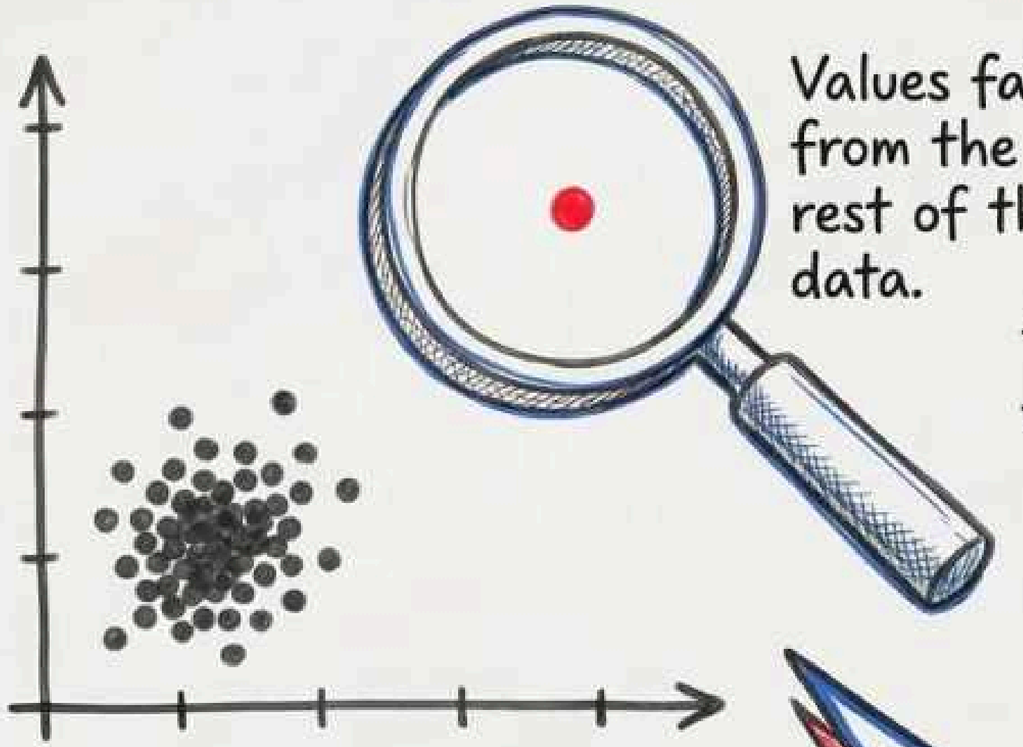


Warning: Choosing the wrong summary for the shape misleads the reader.

Gemini Notebook

Handling Outliers

¶ 1.9



Values far from the rest of the data.

Branch 1: Error
e.g., a misplaced decimal point.

→ Action: Correct the error.

Branch 2: Genuine
e.g., a patient with an extreme, valid response.

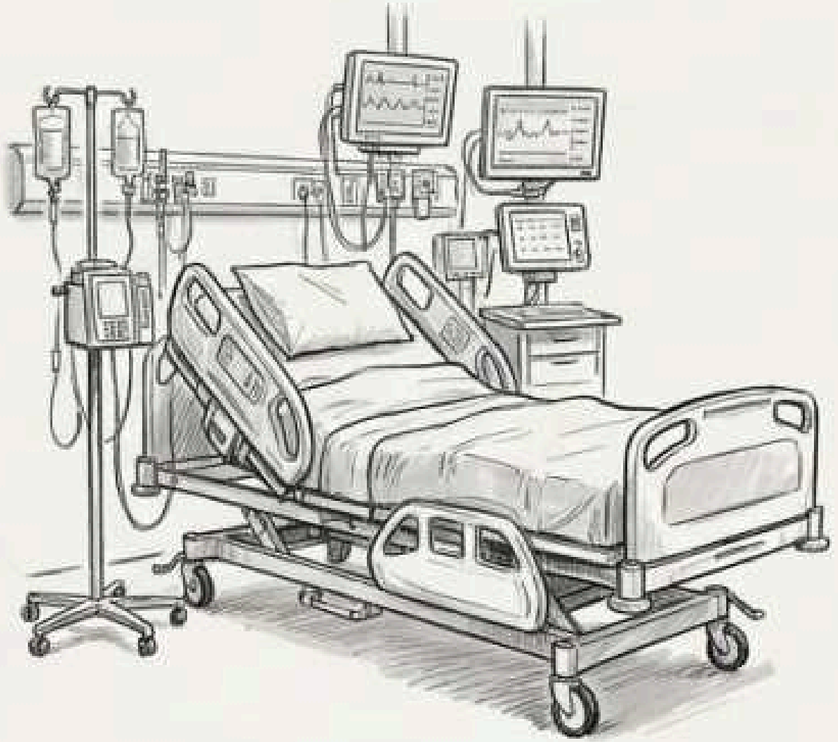
→ Action: Investigate.

The Golden Rule: Never delete outliers simply because they are inconvenient. Run the analysis with and without them to see if conclusions change.

Gemini Notebook

Population Context: Who Was Measured?

¶ 1.10



Hospital Patients

≠



General Population

- Core Principle: Every summary is entirely dependent on the sample population.
- Example: Mean blood pressure drawn from hospital patients will differ from one taken in the general public.
- Crucial Question: Before trusting any description, ask where the data came from and whom they represent.

Gemini Notebook

Synthesis: Describing Data

¶ 1.11

Describing data means choosing the right summaries for the type and shape of the variable.

1. Identify Type (Categorical vs. Numerical)

2. Observe Shape (Use Graphs)

3. Choose Summary (Mean vs. Median)

4. Contextualize (Denominator & Population)



Looking Ahead: Next Lecture: The language of chance. Moving from descriptive summaries to beyond the data we have.

Gemini Notebook