Overview

Not all data points play fair. This module breaks down how missing values and outliers can twist your results, and shows how to handle them wisely. Learn how to spot skewed data, understand what it means, and decide what to do next.

Types of Missing Data

Type	Description	Example
MCAR	Missing completely at random (no pattern)	System glitch deletes lab results
MAR	Missing due to measurable factors	Men skip health survey questions

Handling Methods:

- Investigate causes.
- Use imputation (e.g., fill gaps with median/mean).
- Collect additional data.
- Report transparency on missingness.

Outliers

Type	Description	Example
Univariate	Extreme in one variable	$10,000 charge in a $1,000 budget
Multivariate	Extreme combination of variables	Low-income area with high internet access

Detection Tools:

- **Visualization:** Box plots, scatter plots.
- **Statistical Methods:**
    - **Z-score:** $z = \frac{x-mean}{std}$(values > 3 or <-3 are outliers)

- **Interquartile Range (IQR):** $Outliers\ Q1\ -\ 1.5\ x\ IQR$ or $Q3\ +\ 1.5\ x\ IQR$

Key Risks of Skewed Data

Issue	**Impact**
Missing Data	Bias, weak conclusions, incomplete insights
Ignored Outliers	Flawed decisions or missed critical trends
Misrepresented Data	Unfair policies or misleading analysis

Handling

- Keep if meaningful (e.g., rare trends).
- Remove/correct if errors (e.g., typos).

Best Practices

- **For Missing Data:** Prioritize understanding why data is missing (MCAR vs. MNAR).
- **For Outliers:** Context is key—decide retention/removal based on cause.
- **Transparency:** Document and report how missing data/outliers were addressed.

Real-World Example

- **Missing Data:** A temperature dataset with NaN values filled using imputation.
- **Outliers:** A contractor's $100 million charge flagged as extreme in budget analysis.

Key Takeaways

- **Missing data ≠ Random:** Classify and address based on type.
- **Outliers ≠ Noise:** Investigate before discarding.
- **Ethical Handling:** Ensure fairness and accuracy in analysis.