

# Introduction to Statistical Software

## Introduction to Statistical Software

Key concepts, common tools, and best practices for getting started

---

## 1. Why Use Statistical Software?

| Benefit          | Explanation                                                                               |
|------------------|-------------------------------------------------------------------------------------------|
| Speed            | Automates calculations that would be tedious by hand.                                     |
| Reproducibility  | Scripts and command logs allow others to repeat your analysis exactly.                    |
| Scalability      | Handles large datasets (thousands to millions of rows) without manual effort.             |
| Visualization    | Generates high-quality graphs and interactive dashboards.                                 |
| Advanced Methods | Provides built-in functions for regression, multivariate analysis, machine learning, etc. |

---

## 2. Major Families of Statistical Software

| Category                    | Popular Packages                                    | Typical Users                                            | Strengths                                                             |
|-----------------------------|-----------------------------------------------------|----------------------------------------------------------|-----------------------------------------------------------------------|
| General-purpose programming | R, Python (pandas, statsmodels, scikit-learn)       | Data scientists, researchers, analysts                   | Open source, extensive libraries, strong community support.           |
| Point-and-click GUI         | SPSS, SAS Enterprise Guide, JMP, Stata (GUI mode)   | Social scientists, health researchers, business analysts | Intuitive menus, built-in documentation, good for non-programmers.    |
| Enterprise analytics        | SAS, IBM SPSS Modeler, Microsoft Azure ML           | Large organizations, regulated industries                | Robust security, integration with data warehouses, support contracts. |
| Web-based / cloud           | Google Colab, RStudio Cloud, Databricks, JupyterHub | Teams collaborating remotely                             | No local installation, easy sharing, scalable compute resources.      |

# Introduction to Statistical Software

| Category          | Popular Packages                                                               | Typical Users                 | Strengths                                            |
|-------------------|--------------------------------------------------------------------------------|-------------------------------|------------------------------------------------------|
| Specialized tools | Minitab (quality control),<br>Epi Info (epidemiology),<br>MATLAB (engineering) | Domain-specific professionals | Tailored functions and templates for niche analyses. |

---

## 3. Core Concepts Common to All Packages

### 1. Data Import & Export

- Formats: CSV, Excel, JSON, SQL databases, SAS/SPSS/Stata files.
- Functions: `read.csv()` (R), `pandas.read_excel()` (Python), `GET FILE` (SPSS).

### 2. Data Cleaning & Transformation

- Missing-value handling, recoding, type conversion, reshaping (wide → long).
- Typical verbs: `filter()`, `mutate()`, `pivot_longer()` (R tidyverse) / `df.loc[]`, `df.melt()` (Python).

### 3. Exploratory Data Analysis (EDA)

- Summary statistics (`mean()`, `describe()`), frequency tables, cross-tabulations.
- Visual checks: histograms, boxplots, scatterplots, QQ-plots.

### 4. Statistical Modeling

- Linear regression, logistic regression, ANOVA, mixed-effects models, survival analysis.
- Syntax patterns: `lm(y ~ x1 + x2, data = df)` (R) vs. `sm.OLS(y, X).fit()` (Python).

### 5. Diagnostics & Validation

- Residual analysis, multicollinearity checks (VIF), goodness-of-fit metrics, cross-validation.

### 6. Reporting & Visualization

- Publication-ready graphics (ggplot2, seaborn, plotly).

# Introduction to Statistical Software

- Automated report generation: R Markdown, Jupyter notebooks, SAS ODS.
- 

## 4. Getting Started: A Minimal Workflow (R Example)

# 1. Load packages

```
library(tidyverse)    # data wrangling + ggplot2
```