On Using AI in Data Analysis

Prof. Amjad D. Al-Nasser
Yarmouk University

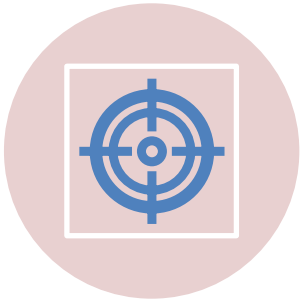
The 3rd International Conference on Artificial Intelligence in Sustainable Development

(RAISD 2025)

November 19, 2025
Bristol Hotel, Amman, Jordan

Why This Topic Matters

DATA IS LARGER AND MORE COMPLEX



NEED FOR FASTER, ACCURATE ANALYSIS

AI EXPANDS WHAT STATISTICS CAN DO

AI in data analysis refers to the application of intelligent algorithms to collect, clean, process, and interpret data

Machine Learning (ML): Enables computers to learn from data and improve their performance: regression, classification, and clustering

Deep Learning: A subset of ML that employs AI to analyze complex, high-dimensional data such as images, speech, or text

Natural Language Processing (NLP): Facilitates the analysis of textual data

Predictive Analytics: Uses AI algorithms to forecast future trends

Automated Data Preparation: AI tools can clean, normalize, and transform raw data to improve data quality

Advantages of Using AI in Data Analysis

Efficiency and Speed

Accuracy and Consistency

Automation of Repetitive Tasks

Discovery of Complex Patterns

Real-Time Analysis

AI Adoption and Usage Statistics

Metric	Statistic
Companies Using AI (at least one function)	**78% (as of 2024)**
AI Users Globally (est. 2025)	**378 million**
Professionals Using AI Daily	**74% of global AI users**
Data Scientists Using AI Daily	**45.1%**
AI Adoption as Top Priority for Companies	**83%**
Adults Using Generative AI Daily	**Approx. 1/3**

Common Statistical Techniques Requested and Used with AI

Descriptive Statistics

Hypothesis Testing

Regression Analysis

Clustering

Predictive Analytics

Time Series Analysis

Natural Language Processing (NLP)

Dimensionality Reduction (PCA)

Challenges

OVERFITTING

BIAS

INTERPRETABILITY ISSUES

ETHICAL & DATA QUALITY CONCERNS

AI platforms vs. Statistical software

Aspect	Traditional Statistical Software	AI Platforms
Data Assumptions	**Requires explicit assumptions** about data distribution (e.g., linearity), which can be rigorously tested.	**Automatically learns complex patterns** without strong prior assumptions about the data structure (non-linear relationships).
Data Quality/Bias	**More resilient to minor data issues**; bias introduced is often easier to trace and correct manually.	**Highly sensitive** to data quality and inherent biases in the training data (e.g., a hiring algorithm trained on male-dominated data excluded female candidates).
Generalizability	**Models can sometimes be less effective** when applied to data outside their specific context, especially with limited initial data.	**Excels at generalization** and making accurate, repeatable predictions on unseen data if properly trained and validated (e.g., using cross-validation).

AI platforms vs. Statistical software

Feature	Traditional Statistical Software	AI Platforms
Accuracy (Predictive Tasks)	Good for low-dimensional data	Higher, especially for complex/large data
Speed (Processing Large Data)	Slower, often manual setup required	Faster (up to **5x** for analytics tasks)
Scalability	Less scalable for "big data"	Highly scalable, designed for massive datasets
Interpretability	High (transparent models)	Low to moderate ("black box" models often require extra tools like SHAP for explanation)

Task	AI Platforms	Common Error
Descriptive statistics (means, SD, counts, proportions)	High concordance — ChatGPT-4 produced results matching SPSS/R in most cases reported.	Rare rounding / formatting differences; **sometimes omitted exact sample-weight handling**.
Independent / paired t-tests	Very high agreement (identical test statistics / p-values reported in multiple studies).	**Minimal differences** (rounding).
Simple linear regression	High concordance for coefficients, SEs, and p-values in studies.	Occasionally, **differences in default contrasts** or handling of categorical variables (reference levels).
One-way ANOVA (F statistic)	F-values generally consistent between ChatGPT-4 and SPSS/R.	Post-hoc (pairwise mean diffs, CIs, adjusted p-values) **showed discrepancies in several reports**.
Repeated-measures ANOVA	F statistics sometimes matched, but studies reported incorrect degrees-of-freedom or mis-handled within-subject structure.	**Wrong df, incorrect sphericity** corrections or missing Greenhouse-Geisser adjustments.
MANOVA / Multivariate tests	Less reliable; studies report inconsistent or incorrect outputs for MANOVA.	**Incorrect dfs, missing multivariate test statistics** / misreported test assumptions.

Task	AI Platforms	Common Error
Non-parametric tests (Mann-Whitney, Wilcoxon, Kruskal-Wallis)	Mixed results — some metrics matched, but differences reported in U/Z values, IQRs, or p-values.	Differences in ranking/tie handling and test variants (continuity correction).
Multivariable regression (GLMs, logistic regression)	Good agreement on coefficients and p-values for many examples; careful on variable coding and reference categories.	Differences arise from link function defaults, factor coding, or omitted interactions.
Complex models (multilevel/hierarchical models, SEM, survival analysis)	Not well-evaluated / not robust in current literature — studies warn ChatGPT-4 struggles with these.	**Often lacks detailed diagnostics**, misapplies model assumptions, or omits necessary steps.
Code generation (R / Python syntax)	Strong: ChatGPT-4 can generate runnable code for many common analyses; accelerates workflow.	**Code may need minor fixes**; default package choices or plotting options may differ from user's standards.
Interpretation / narrative explanation	Excellent: clear plain-language explanations of results and guidance.	**Explanations can be overconfident**; may omit caveats or assumption checks unless prompted.

Strengths

Traditional Statistics:	• Strong inference	• Clear interpretation
AI Platform	• High predictive accuracy	• Handles complex, large-scale data

Weaknesses

Traditional Statistics:	Limited with nonlinear/high-dimensional data	
AI Platform	• Risk of overfitting	• Less interpretable

When to Use Each

Traditional Statistics:	Explanation, small samples
AI Platform	• Prediction, large or unstructured data

Closing Message

AI —NOT REPLACES—THE STATISTICIAN.

FUTURE = STATISTICS + INTELLIGENT ALGORITHMS.

Selected References

- Ramcharitar, K. (2025). *Technology focus: Data science, analytics, and artificial intelligence*. Journal of Petroleum Technology, 77(1), 87–88. https://doi.org/10.2118/0125-0087-JPT

- Cao, Y. (2025). Using AI and big data analytics to support entrepreneurial decisions in the digital economy. *Sci Rep* 15, 36933. https://doi.org/10.1038/s41598-025-20871-4

- Vieriu, A. M., & Petrea, G. (2025). The Impact of Artificial Intelligence (AI) on Students' Academic Development. *Education Sciences*, *15*(3), 343. https://doi.org/10.3390/educsci15030343

Thank You

Questions?

Prof. Amjad D. Al-Nasser