RESEARCH ARTICLE

Using Generative AI to help with statistical test selection and analysis

Tom Goodale, Academic Achievement Team, Liverpool John Moores University, Liverpool, UK.
Email: t.r.goodale@ljmu.ac.uk

Abstract

One of the most common questions that students ask statistics advisors is ‘What test should I do?’ This paper explores the use of generative AI chatbots, specifically ChatGPT, as a tool to assist students, in particular those with limited experience in statistics, in selecting appropriate statistical tests for their analyses. Traditional methods, such as flowcharts and online test selectors, require at least a basic understanding of measurement scales and research design, which can be an issue for many students who have limited exposure to statistics on their courses. This research focuses on developing and refining prompts to guide ChatGPT in providing accurate and relevant statistical test recommendations. A hypothetical scenario was used to test the effectiveness of various prompts, ranging from simple, naïve questions to more sophisticated ones utilising specific prompt patterns, such as the ‘context manager’ and ‘flipped interaction.’ These patterns were selected to enhance the chatbot’s responses and ensure the relevance and accuracy of the test suggestions. The findings suggest that while AI chatbots like ChatGPT can be a valuable resource for students, their effectiveness is highly dependent on the quality of the prompts used. The paper concludes with a discussion on the potential of these AI tools in educational settings, acknowledging the limitations of current technology and suggesting directions for future research and development.

Keywords: statistical test selection, ChatGPT, generative AI, prompt engineering.

1. Introduction

Choosing the correct statistical test is a core task of any statistical analysis. Whilst experienced users of statistics will not often struggle with this task, those with less experience, especially students, often struggle with this. In order to help with this many textbook authors, such as Field (2018), include some sort of flowchart to help in the decision making process; a simple web search finds a plethora of flowcharts available online. Additionally, students may be pointed towards one of the free online statistical test selectors (Jackson-Wood, 2023; Sage, 2023; Dokuz Eylül University, 2023; Sheffield Hallam University, 2023), one of which was developed after much research (Suner, Karakülah and Dicle, 2014), or they may be pointed towards commercial ones such as on the Laerd statistics website (Laerd, 2023).

In order to use these flowcharts and online test selectors successfully, students need to know something about measurements scales, i.e. the distinction between nominal, ordinal and interval/ratio (Stevens, 1946). Additionally, many of the flowcharts and automated test selectors start with questions around the specific research hypothesis that is being tested – commonly a distinction is made between a test of association, a test of group differences, or prediction – although some flowcharts and test selectors instead start by asking what the measurement levels are. Thus, when teaching students how to use these tools it is important to teach both the measurements scales and, at least at a coarse level, something about study design. Many students do not have this necessary background and thus have to ask supervisors or statistical support staff to help them choose a test.

The increasing sophistication of generative AI chatbots such as ChatGPT potentially provides a powerful alternative to this traditional method of choosing a statistical test, and could, to a large extent, reduce the need for students to seek external help.

At the current time chatbots do, however, require some care to use, especially for naïve users. The quality of the response is highly dependent on the quality of the prompt, and even with the best prompts there is still the possibility of 'hallucinations' where the chatbot confidently provides information which is factually incorrect. This paper explores prompts and reports an effective prompt that naïve statistical users might use for test selection.

2. Chatbot Prompts

Prompts are at the heart of the interaction between a user and an AI chatbot. The simplest prompts ask a question which the chatbot answers by use of its pre-trained material. More sophisticated prompts go beyond using a single question and answer; they can introduce extra material to supplement the chatbot's own data, establish a tone and style for the chatbot's responses, set a context within which a question is set, or control the format of output. Recently White et al. (2023) have taken the idea of 'design patterns' from software engineering (Gamma, 1995) to categorise common 'prompt patterns.' These patterns, detailed in Table 1, offer a structured approach to interacting with chatbots such as ChatGPT.

Table 1. Prompt patterns from White et al. (2023).

Pattern Category	**Prompt Pattern**
Input Semantics	Meta Language Creation
Output Customization	Output Automater Persona Visualization Generator Recipe Template
Error Identification	Fact Check List Reflection
Prompt Improvement	Question Refinement Alternative Approaches Cognitive Verifier RefusalBreaker
Interaction	Flipped Interaction Game Play Infinite Generation
Context Control	Context Manager

Of particular interest to this paper are the 'context manager' pattern, which allows context to be set, and the 'flipped interaction' pattern, which asks the chatbot to query information from the user, and the 'persona' pattern, which helps to shape the interaction.

3. Developing a prompt for Statistical Test Selection

In order to develop prompts and test their effectiveness, the following hypothetical scenario was used:

- An experiment was conducted to test the effectiveness of five different fertiliser concentrations on crop height.
- This was performed via a 10x5 design whereby ten fields were each split into five plots. In each field the five plots were each assigned one of the five levels of fertiliser.
- 20 measurements of crop height were taken in each plot.

In this experiment the response variable is the crop height and there are two explanatory factors – the amount of fertiliser, which is the factor of interest, and the field, which should be considered a blocking factor. This experimental design should lead to the recommendation of a two-way ANOVA.

All tests were done with ChatGPT 3.5, which is freely available. The results may have been different with the more powerful ChatGPT 4.0, or with other AI chatbots. As I was running these tests I tried to answer as I would expect a student to do so from my experience but may not always have succeeded in this.

3.1 Naïve prompt

As a baseline test a naïve prompt of ‘What test do I need to do to examine the relationship between crop height and fertiliser’ was used. This is the sort of question I might expect to see from a student coming to me.

The response from ChatGPT was (see Supporting data, Example 1):

> To examine the relationship between crop height and fertilizer, you can conduct a statistical analysis using various methods, depending on the nature of your data and your research goals. Here are some common tests and approaches you can consider:
>
> **Correlation Analysis**:
>
> **Pearson's Correlation Coefficient**: Use this if both crop height and fertilizer are continuous variables, and you want to assess the strength and direction of the linear relationship between them.
>
> **Spearman's Rank Correlation**: Use this if your data is not normally distributed or if there might be non-linear relationships between the variables.
>
> **Regression Analysis**:
>
> **Simple Linear Regression**: This is appropriate if you want to predict crop height based on the amount of fertilizer used. It provides a linear equation that describes the relationship between the two variables.
>
> **Multiple Regression**: If you have more than one independent variable (e.g., multiple types of fertilizers or other factors influencing crop height), multiple regression can be used to analyze their combined effects.
>
> **Analysis of Variance (ANOVA)**:
>
> **One-Way ANOVA**: Use this when you have multiple groups of crops that received different levels or types of fertilizer, and you want to determine if there are statistically significant differences in crop height among the groups.

> **Two-Way ANOVA**: This is used when you have two independent variables (e.g., type of fertilizer and time of application) and you want to assess their individual and interactive effects on crop height.
>
> **Chi-Square Test**:
>
> If your data is categorical and you want to examine the relationship between discrete levels of fertilizer and crop height categories (e.g., tall, medium, short), you can use a chi-square test for independence.
>
> **Non-parametric Tests**:
>
> If your data violates assumptions of normality or homoscedasticity, consider non-parametric tests like the Mann-Whitney U test or Kruskal-Wallis test.
>
> **Time Series Analysis**:
>
> If you are collecting data on crop height and fertilizer application over time, you might use time series analysis techniques to examine trends, seasonality, and autocorrelation in the data.
>
> **Data Visualization**:
>
> Before conducting formal statistical tests, it's often a good idea to visualize your data using scatter plots, box plots, or other relevant plots to gain insights into the relationship between crop height and fertilizer.
>
> The choice of test depends on your specific research design, the type of data you have, and the research question you are trying to answer. It's essential to carefully plan your study, collect data systematically, and consult with a statistician or data analyst if you are unsure about the appropriate analysis method for your specific situation.

This is a very broad answer; it has introduced statistical concepts that the student may be unfamiliar with, and whilst the correct test is there, it has not singled it out. A student receiving such an answer might be overwhelmed and could potentially spend a lot of time learning about irrelevant analyses before narrowing it down to ANOVA and subsequently two-way ANOVA.

3.2 First attempt at a more sophisticated prompt

In order to overcome the shortcomings of the naïve prompt, three prompt patterns suggested themselves:

- To shape the interaction and make it more relevant the persona pattern was introduced via ‘Act as a statistics adviser’
- To reduce the complexity and assumed background the context manager pattern was introduced via ‘talking to an undergraduate student’ and ‘The student wants to know what statistical tests to do with their data.’
- The flipped interaction pattern was then used to force ChatGPT to ask questions of the user in order determine the test: ‘Ask questions one at a time until you can recommend a test and then provide your recommendation.’

Thus, the full prompt was

> Act as a statistics advisor talking to an undergraduate student with no knowledge of statistics. The student wants to know what statistical tests to do with their data. Ask questions one at a time until you can recommend a test and then provide your recommendation.

For reasons of brevity the full ChatGPT output is not reproduced here, but can be accessed online (see Supporting data, Example 2).

With this prompt ChatGPT asked a sequence of questions and provided the following recommendation (see Supporting data, Example 2):

> Great, if your crop height measurements are normally distributed within each group and have roughly equal variances, you can proceed with a one-way ANOVA test. This test will help you determine if there are statistically significant differences in crop heights between the different levels of fertilization.
>
> In your case, you have five different levels of fertilization (including the control group with no fertilizer). You can set up the null hypothesis (H0) as: "There is no significant difference in crop heights between the different levels of fertilization."
>
> And the alternative hypothesis (H1) as: "There is a significant difference in crop heights between at least two of the fertilization levels."
>
> Running a one-way ANOVA test will provide you with an F-statistic and a p-value. If the p-value is less than your chosen significance level (typically 0.05), you can reject the null hypothesis and conclude that there is a statistically significant difference in crop heights between at least two fertilization levels.
>
> If the ANOVA test indicates significance, you can perform post-hoc tests (e.g., Tukey's HSD) to identify which specific fertilization levels are different from each other.
>
> Remember to report the results, including the F-statistic, p-value, and any post-hoc test findings, in your research report or paper.

Whilst not the ideal test, as it doesn't consider the blocking effect of the fields, it is a concrete test and ChatGPT has also gone some way to suggest ways in which it could be reported.

3.3 An improved prompt

Whilst the previous prompt was a marked improvement on the naïve prompt, ChatGPTs occasionally asked several questions at once, which could lead to confusion and might not always be fully answered by students. In order to try to counteract this, the prompt was modified to:

> Act as a statistics advisor talking to an undergraduate student with no knowledge of statistics.
>
> I want to know what statistical tests to do with my data.
>
> Ask questions one at a time until you can recommend a test and then provide your recommendation. The questions must be short and explain any statistical terms used.
>
> Ask your first question.

This succeeded in its goal of forcing ChatGPT to ask shorter, more direct and specific questions (see Supporting data, Examples 4, 4c, 4d, 4e). In many cases, but by no means all, this also led to the recommendation of a 2-way ANOVA.

3.4 Tests with other scenarios

This prompt was also used for two other scenarios:

- Spring extension vs weight: ChatGPT suggested regression, which would be appropriate; with the information given in the scenario, however, correlation would also have been appropriate and simpler (see Supporting data, Example 5).
- The relationship between gender (nominal with 2 categories) and car colour (nominal with 2 categories). ChatGPT correctly identified a Chi-squared test of independence for this case, and

was also able to answer questions on assumption checking when prompted (see Supporting data, Example 6).

4. Conclusions

AI chatbots such as ChatGPT provide a powerful alternative to more traditional methods of statistical test selection such as flowcharts and interactive test selectors. With effective prompts they can be used by statistically naïve users to make sensible test selections, thus allowing these users to make a start on their analysis without having to speak to supervisors or statistics advisors. These recommendations, however, may not be the ideal tests, and may not use the full structure available in the data, and indeed may change from run to run. Ideally students should still have a certain level of statistical knowledge in order for them to judge the quality of the answer. One possible strategy they may choose to adopt is to run the prompt several times and then investigate the different tests that are recommended in order to better choose the best one; generative AI may also be of use for this step as it would be able to compare and contrast the tests for the given scenario.

Free generative AI chatbots such as ChatGPT 3.5 have the potential to give students swift access to statistical help at times when it may be difficult for them to access a statistics advisor, such as overnight, at weekends, or at particularly busy times of year. Whilst this advice is not perfect, they would allow students to progress their analysis and complete their assignments. As the sophistication of chatbots increases the quality of advice will almost certainly increase.

In the future the community could develop more sophisticated prompts for the free versions of chatbots. The recent release of ‘Custom GPTs’ by Open AI may also provide a way for the community to build a custom statistics advisor GPT which could be made available via a web application.

5. Supporting data

Example 1: https://chat.openai.com/share/28b6b6bb-b5d9-4cf3-8d6d-6b90e0a86dfa

Example 2: https://chat.openai.com/share/5e4b7330-b09a-4fc1-b3c2-caf0cc25cfa7

Example 4: https://chat.openai.com/share/8d95b63e-d9c1-4452-be04-ff3d1e74723d

Example 4c: https://chat.openai.com/share/cdb4b9be-8259-4e9c-91c8-6f2e43e2a6ed

Example 4d: https://chat.openai.com/share/f94af8e8-77e6-4e83-8034-30092d134904

Example 4e: https://chat.openai.com/share/61c23440-41b5-4a6c-ab9a-f76d139e88bc

Example 5: https://chat.openai.com/share/341a61bc-371c-4b0a-a018-42afb6e2d0f5

Example 6: https://chat.openai.com/share/204004fe-4cf0-4aad-aa7c-fcad99fc7525

6. References

Dokuz Eylül University (2023). *StatXFinder: a decision support tool for appropriate statistical test selection*. Available at: https://medinfo.deu.edu.tr/statxfinder/index.php [Accessed 13 November 2023].

Field, A.P. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Los Angeles: SAGE.

Gamma, E. (1995). *Design patterns: elements of reusable object-oriented software*. Addison-Wesley.

Jackson-Wood, M. (2023). *statistical test flowchart*. Available at: https://www.statsflowchart.co.uk/ [Accessed 13 November 2023].

Laerd (2023). *Laerd Statistics*. Available at: https://statistics.laerd.com/ [Accessed 13 November 2023].

Sage (2023). *Which Stats Test*. Available at: https://methods.sagepub.com/which-stats-test [Accessed 13 November 2023].

Sheffield Hallam University (2023). *Interactive test chooser*. Available at: https://maths.shu.ac.uk/mathshelp/WhichTestIndex.php [Accessed: 20 December 2023].

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103 (2684), 677-680. https://doi.org/10.1126/science.103.2684.677

Suner, A., Karakülah, G. and Dicle, O. (2014). Towards a web-based decision support tool for selecting appropriate statistical test in medical and biological sciences. *Studies in Health Technology and Informatics*, 205, pp.48-52.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. and Schmidt, D.C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv preprint*. https://doi.org/10.48550/arXiv.2302.11382