Probabilistic Reasoning, Statistical Attitudes, and AI's Role in Analyzing Students' Explanations

Marija Kaplar[1[0000-0002-0920-8276]] and Aleksandra Stevanović [2[0000-0001-5408-608X]]

[1] Faculty of Technical Sciences, University of Novi Sad, Serbia,
[2] Faculty of Information Technology, Belgrade Metropolitan University, Serbia,
marija.kaplar@uns.ac.rs,
aleksandra.stevanovic@metropolitan.ac.rs

Abstract. In contemporary society, the ability to understand and interpret data has become a fundamental skill. Regardless of STEM or non-STEM fields, all individuals is expected the minimum level of stochastic literacy. This study examined Non-STEM students' base-rate reasoning, their attitudes toward statistics, and the potential of artificial intelligence (AI) to support the analysis of students' explanations. There participated 105 students, from the University of Novi Sad (non-STEM: Law, Economics) during regular classes. They solved a base-rate task, provided written explanations, and completed a attitude questionnaire. Responses were coded into six categories reflecting correctness and explanation type. Overall, 41% of students answered the task correctly, with no significant differences by faculty or gender. However, almost half of the participants chose the same wrong option. Students of the Faculty of Economics were more likely to provide explanations (70%) compared to law students (25%). When considering only responses with explanations, a significant gender difference was found, with male students more likely to provide correct probability-based reasoning, while female students more often relied on equiprobability explanations. Attitude measures showed generally positive orientations although self-efficacy was weaker. AI-based (ChatGPT) coding of explanations gives results comparable to human classification and shows its potential for identifying misconceptions, albeit with certain limitations. Findings emphasize the need to strengthen probabilistic reasoning and statistical self-efficacy among Non-STEM students, while also pointing to AI's potential as a research and pedagogical tool.

Keywords: probabilistic reasoning, base-rate reasoning, attitudes toward statistics, non-STEM students, AI-assisted coding of explanations

1 Introduction

In contemporary society, possessing a basic level of stochastic literacy, encompassing knowledge of probability and statistics, is essential not only for professional purposes but also for personal decision-making (Kapler et al, 2021; 16]. The ability to correctly interpret and evaluate data is increasingly important across diverse fields, extending

Research Paper
DOI: https://doi.org/10.46793/eLearning2025.291K
Part of ISBN: 978-86-89755-37-4

beyond traditionally STEM-oriented professions. Even for students in non-STEM disciplines, foundational competencies in probability and statistics are crucial for informed reasoning and critical evaluation of information [16, 19].

The present study examined the extent to which non-STEM students are able to solve a task that can be regarded as a measure of basic stochastic literacy at the elementary level. In addition to selecting an answer, students were asked to provide written explanations in order to gain deeper insight into their reasoning processes. This task is also widely recognized as a diagnostic measure of the base rate neglect misconception (also referred to as base rate bias or insensitivity to prior probabilities) occurs when a base set or some other important information is neglected. Instead of considering the base rates, conclusions are typically drawn from the descriptive details presented in a task, even though such information is insufficient for making an accurate judgment. A common misconception is that if the description of a case appears to fit one category more closely, then the likelihood of belonging to that category must be higher (Kapler et al., 2021; 14, 15]. When a description does not clearly correspond to any category, individuals are often prone to the equiprobability bias**,** the belief that all possible outcomes are equally likely. In these cases, the base rate is again ignored, leading to erroneous probabilistic reasoning [10].

Beyond assessing students' susceptibility to base rate neglect, the study also explored their attitudes toward statistics, including the perceived importance of statistics, self-assessed knowledge, trust in data and media, and motivation to improve their statistical competence. Finally, in the third part of the study, we investigated the potential of artificial intelligence (ChatGPT) to support the analysis of students' explanations and to evaluate its possible role as a tool for identifying and addressing misconceptions related to probabilistic reasoning. Based on these considerations, the present study addresses three research questions concerning students' base-rate reasoning, their attitudes toward statistics, and the potential role of AI in analyzing students' explanations.

1.1 Research Questions

RQ1. To what extent do non-STEM students correctly solve a probability task, and how prevalent are base-rate neglect and the equiprobability bias in their written explanations?

RQ2. What levels and patterns characterize students' attitudes toward statistics across the four subscales: Value/Utility, Self-efficacy, Skepticism, and Motivation to Learn?

RQ3. Can an AI-based tool (ChatGPT) reliably code students' explanations against the human coding scheme?

All analyses were conducted with group comparisons by *faculty* (Law vs. Economics) and *gender***.** Because non-STEM programs are heterogeneous, meaningful differences may exist within this population; in our study, Law and Students of Economy follow substantially different curricula, so faculty-based contrasts were explicitly examined.

2 Literature Review

Decision-making in both daily practices and broader social and political contexts often relies on data analysis, highlighting the importance of data interpretation for fostering informed and active citizenship [5, 17, 20]. In many domains such as for example public health and climate change, misinterpretations of data can lead to severe and far-reaching consequences [8, 23]. Across many professional fields, the competence to accurately analyze and interpret data is regarded as essential, and professional associations, governmental bodies, and scholarly work alike stress the importance of cultivating these competencies [1, 7]. In this regard, familiarity with the fundamental concepts of probability and statistics is necessary for citizens, regardless of their professional background to participate actively in society and make sound decisions.

Although critical data literacy is fostered at all stages of education, higher education students, as future leaders and decision-makers, are expected to develop stronger knowledge and greater confidence in applying these skills. Previous research has explored misconceptions related both to probability and to statistics. Some of the results of studies conducted in fields such as banking, investment, auditing, management, insurance, and healthcare suggest that even highly educated adults often face difficulties when interpreting information and making judgments under uncertainty [3, 4, 11, 15].

Research by Khazanov and Prado [18] indicated that nearly four out of five university students enrolled in disciplines such as accounting, business, liberal arts, and mental health exhibited difficulties with probabilistic reasoning, particularly showing susceptibility to the equiprobability bias. Similarly, Hirsch and O'Donnell [12] reported that approximately three-quarters of students, predominantly from psychology but also across other fields, struggled with probability misconceptions. In such cases, individuals often resort to intuitive strategies or heuristics, which, while common, frequently lead to systematic errors in reasoning [13].

In the Serbian context, base rate neglect has been examined among STEM students, specifically engineering students, where around 35% were found to be prone to this type of misconception [17]. The study by Kaplar et al. [17] offers a useful point of reference for the present work because it employed the same base-rate task under equivalent testing conditions and procedures, allowing direct comparability of findings. Research on Serbian Non-STEM students to date has focused primarily on their interpretation of visually presented data, most notably the accurate reading of the arithmetic mean and standard deviation (Luzanin et al., 2022). In a cross-cultural design comparing Serbian STEM and Non-STEM students with STEM and Non-STEM students from Kent State University (USA), Luzanin et al. (2022) reported that Serbian Non-STEM students performed worse in interpreting the arithmetic mean than both Serbian STEM students and their STEM and Non-STEM peers from the U.S. Taken together, these studies motivated an examination of how Non-STEM students tackle probability-based tasks that index elements of stochastic literacy.

The overarching aim of such investigations is to identify potential weaknesses in general education and to highlight curricular areas where targeted improvement is warranted. This is particularly important given that previous research has shown that, even when students achieve high performance on probability tests, they often lack a deeper understanding of fundamental concepts, especially when making judgments about uncertain events in real-life contexts [9, 24]. At the same time, numerous studies highlight that carefully designed courses in probability and statistics can play an essential role in addressing these difficulties [6, 10, 11, 21]. These studies recommend emphasizing students' prior experiences, incorporating experimental and frequency-based approaches, and fostering intuitive understanding of probability concepts. Such strategies have been shown to reduce the likelihood of misconceptions and underscore the importance of instructional methods in shaping statistical reasoning.

Students' explanations provide valuable insights into their reasoning, yet analyzing large volumes of written responses is time-consuming, which is why artificial intelligence is increasingly being considered as a supportive tool in this process. Previous studies show that large language models can make a significant contribution to data analysis, but their role remains primarily supportive. McClure et al. [22] emphasize that the accuracy of AI coding is often comparable to, and in some cases higher than, human performance in deductive tasks, especially when the instructions are clearly defined. However, researchers' experiences also highlight certain limitations: Yan et al. [26] point out that AI can accelerate analysis, but its lack of reliability and deeper contextual understanding requires human validation. Williams [25] further warn about ethical and methodological challenges, such as hallucinations and the absence of theoretical grounding in interpretation. The common conclusion across these studies is that AI can enhance the speed and scope of data processing, but the human researcher remains essential to ensure accuracy and meaningful analysis.

3 Methodology

3.1 Sample

The sample consisted of undergraduate students from the University of Novi Sad (Table 1). Data were collected during regular class sessions. All students from the Faculty of Economics reported having passed at least one exam in probability and statistics at the time of testing. Participation in the study was voluntary, and all participants were informed about the aims of the research and provided their consent prior to participation.

Table 1. Sample distribution by gender and faculty

Faculty	Female	Male	Total
Faculty of Economics – Subotica	28	9	37
Law Faculty – Novi Sad	39	29	68
Total	67	38	105

3.2 Procedure

The study was conducted using a structured test consisting of several components. First, students provided demographic information. They were then asked to solve a probability task designed to assess base rate neglect and elements of statistical reasoning**.** In order to gain deeper insight into students' reasoning processes, participants were instructed to provide written explanations for their answers. The test further included a questionnaire on students' attitudes toward statistics, measured through Likert-scale items. Finally, students' open-ended responses were analyzed using an AI-based tool to explore the extent to which artificial intelligence can support the identification and understanding of misconceptions related to base rate neglect.

Statistical literacy through base rate neglect

As part of the study, participating non-STEM students were presented with a probability task (Fig. 1) that required them to reason about base rates in order to reach the correct solution. This task can be considered a component of statistical literacy, as it involves applying basic probabilistic reasoning to everyday contexts. At the same time, it is commonly used in the literature on cognitive biases as a diagnostic measure of the base rate neglect misconception, which occurs when individuals give insufficient weight to prior probabilities in favor of more salient but less relevant information.

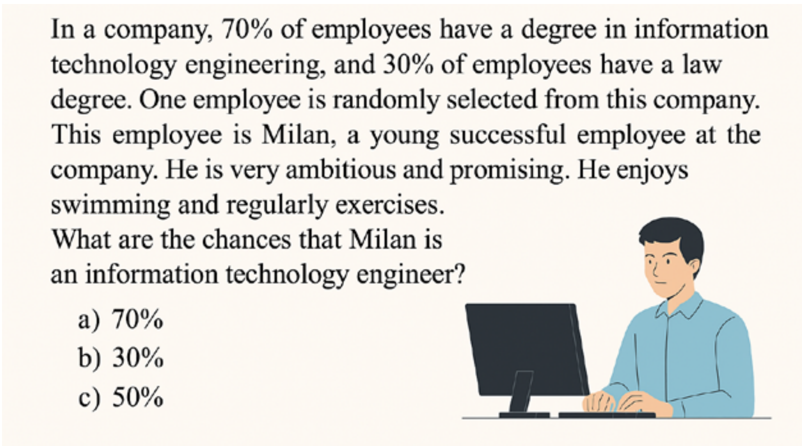

Fig. 1. Example of a probability problem illustrating base rate neglect and elements of statistical literacy *(Figure generated with an AI-based tool)*

In the present study, the task was administered before the attitude questionnaire, allowing us to examine both the correctness of students' responses and the types of justifications they provided. Focusing on a non-STEM student population provided an opportunity to explore statistical reasoning and misconceptions in a group with typically less formal training in mathematics and statistics, but whose professional competence increasingly requires data-informed decision-making.

Coding of Students' Explanations

Students' responses to the base rate task were categorized using a coding scheme that distinguished between accuracy of the answer and the presence and type of explanation. The first digit indicates whether the answer itself was correct (1) or incorrect (0). The second digit specifies the type of explanation:

- **0** – no explanation provided, regardless of correctness of the answer.
- **1** – explanation consistent with a key misconception (equiprobability bias: reasoning that two outcomes must be equally likely simply because they are two).
- **2** – alternative explanation that is incorrect or not based on probabilistic/statistical principles.
- **1 (with correct answer)** – explanation based on probability/statistical principles.
- **2 (with correct answer)** – explanation accompanying a correct response but not grounded in probability/statistical reasoning.

Accordingly, the six possible categories were:

- **00** – Incorrect answer without explanation.
- **01** – Incorrect answer with equiprobability bias explanation.
- **02** – Incorrect answer with alternative incorrect explanation.
- **10** – Correct answer without explanation.
- **11** – Correct answer with probability/statistics-based explanation.
- **12** – Correct answer with explanation not based on probability/statistics.

This scheme allowed us to differentiate between accuracy, presence of reasoning, and the type of reasoning students employed, thereby providing deeper insight into the nature of misconceptions such as base rate neglect and equiprobability bias.

Attitudes toward statistics

Attitudes were assessed with 9 items on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Guided by content and reliability screening, we formed four subscales and computed scale scores as the mean of available items (minimum half of items present). Item 2 was reverse-coded (6 − response) so that higher scores reflect greater skepticism. Higher scores indicate more of the construct. The items used in this questionnaire were previously validated and applied in two studies [16, 19]. The subscales shown in Table 2 were tested.

Table 2. Attitude toward statistics subscales with items and reliability coefficients

Subscale	Items (abbrev.)	Cronbach's α
Value/Utility (VU)	STATS_IMPORTANCE — "It is important to possess basic knowledge of statistics, regardless of profession." CRITICAL_EVALUATION — "The statistical knowledge I possess helps me critically evaluate research findings." PROFESSION_IMPORTANCE — "Statistics is essential for the profession I have chosen." PERSONAL_USAGE — "I use my knowledge of statistics to better interpret data."	0.72
Self-efficacy (SE)	STATS_EDUCATION — "During my education, I studied enough statistics." KNOWLEDGE_SATISFACTION — "My knowledge of statistics is satisfactory."	0.71
Skepticism (SK)	MEDIA_TRUST_R — "I trust research reported in the media." *(reverse-coded)* MISUSE — "Statistics are often misused."	0.46
Motivation to Learn	IMPROVE_STATS — "I would like to improve my knowledge of statistics.	—

Note: MEDIA is reverse-coded (Item 2; 6 − response).

AI-Supported analysis of student reasoning in base rate tasks

In the third part of the study, we focused on students' written explanations accompanying their answers to the base rate neglect task. These justifications were subjected to qualitative coding in order to categorize types of reasoning. The coding scheme distinguished between correct probability-based reasoning and several types of misconceptions (e.g., equiprobability bias, alternative incorrect explanations, or responses without justification).

To complement the manual categorization, students' responses were also processed using an AI-based tool, with the aim of examining whether artificial intelligence can reliably identify reasoning patterns and misconceptions. This dual approach allowed for both human-coded and AI-supported perspectives, providing a richer understanding of students' reasoning processes and exploring the potential of AI to support educational research in addressing misconceptions such as base rate neglect.

4 Results and discussion

4.1 Statistical literacy through base rate neglect

Distribution of responses to the base rate task by faculty is shown in Figure 2. Overall, 41% of students provided the correct solution to the base rate task, with 43% of students of the Faculty of Economics and 40% of law students responding accurately (Table 3). Across both faculties, approximately half of the students (around 50%) selected the 50% response option. This pattern suggests that nearly every second student was inclined toward the equiprobability bias, a misconception reflecting the belief that when two outcomes are possible, they must be equally likely [10].

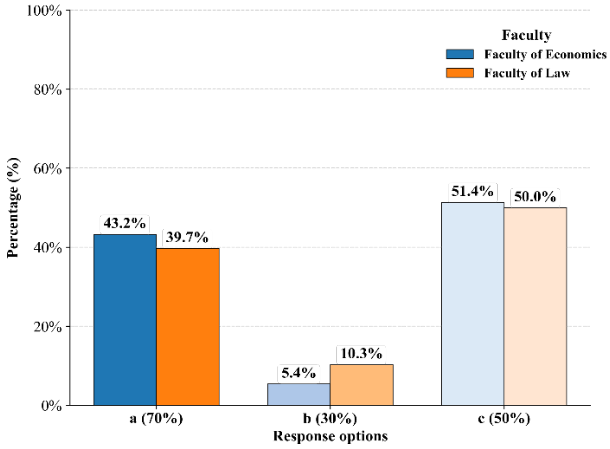

Fig. 2. Distribution of responses to the base rate task by faculty

A chi-square test revealed no statistically significant differences in accuracy between the two faculties, $\chi^2 = 0.12$, $p = 0.724$. Likewise, gender differences were not significant, $\chi^2 = 0.15$, $p = .698$. These findings are comparable to the results reported by Kaplar et al. [17], where approximately 65% of engineering students correctly solved the same problem. Taken together, the results suggest that while misconceptions are present among both STEM and non-STEM students, they tend to be somewhat more pronounced among non-STEM students.

Table 3. Correct and incorrect responses by faculty and gender with chi-square results

Category	Group	Correct	Incorrect	Total	χ^2	p
Faculty	Economics	16	21	37	0.12	0.724
	Law	27	41	68		
Gender	Female	26	41	67	0.15	0.698
	Male	17	21	38		

It might have been expected that students of the Faculty of Economics would perform better on this type of task, given that all of them had completed at least one course in probability and statistics. However, this was not the case. This finding aligns with previous research suggesting that misconceptions may persist regardless of formal education or the number of completed courses [9, 24]. The presence of such misconceptions is often rooted in incorrect intuitive reasoning rather than a simple lack of knowledge. Consequently, addressing these issues requires targeted instructional strategies and carefully designed practice integrated throughout the educational process [6, 10, 11, 21].

4.2 Analysis of students' explanations

Providing explanations in problem-solving tasks can serve as an indicator of how deeply students engage with the task and the extent of their genuine involvement. Prior research [16] has shown that students who offer explanations tend to achieve higher test scores. Providing explanations also was discussed by Attali et al. [2] who argue that the ability to articulate reasoning reflects a willingness to thoughtfully engage with the task, whereas a correct response alone may sometimes result from mere guessing. In the current study only about 40% of non-STEM students provided an explanation for their response to the base rate task (Table 4), which is lower compared to STEM students in earlier studies, where approximately 56% offered justifications (Kapler et al., 2021).

In the present study, the largest proportion of students (40%) fell into category 00, which represents incorrect answers without any explanation (Table 4). Of the total sample, 21% were able to select the correct response and provide a correct justification, whereas in a comparable study with STEM students this proportion was higher, at 32% [16]. Approximately 20% of participants were classified in category 01, where the explanations clearly reflected base rate neglect and the associated equiprobability bias. Illustrative responses from this category included:

"The chances are 50% because the worker could be either an IT engineer or a lawyer."
"There are two options, so the probability must be 50–50%."
"Since the worker was chosen randomly, the chance is equal for both professions."

According to previous literature, such explanations provide strong evidence for the presence of base rate neglect, specifically the equiprobability bias [10]. One possible reason for the persistence of this misconception may even be prior education, since many introductory courses in probability and statistics often rely on examples with equally likely outcomes, such as coin tosses or dice rolls [6, 10, 11, 21]. Taken together, these findings suggest that nearly half of the participants in our study may potentially hold the equiprobability bias, with approximately 20% providing explanations that explicitly confirm this misconception.

Table 4. Distribution of student' answers and explanations by faculty and by gender (cells show n and %; includes 00 & 10 and aggregates).

Type	Group	00	01	02	10	11	12	Without explanation (00+10)	With explanation (01+02+ 11+12)	N (total)

Faculty	Economics	8 (21.6%)	12 (32.4%)	1 (2.7%)	3 (8.1%)	12 (32.4%)	1 (2.7%)	11 (29.7%)	26 (70.3%)	37
	Law	33 (48.5%)	7 (10.3%)	1 (1.5%)	18 (26.5%)	9 (13.2%)	0 (0.0%)	51 (75.0%)	17 (25.0%)	68
Gender	Female	21 (31.3%)	18 (26.9%)	2 (3.0%)	12 (17.9%)	13 (19.4%)	1 (1.5%)	33 (49.3%)	34 (50.7%)	67
	Male	20 (52.6%)	1 (2.6%)	0 (0.0%)	9 (23.7%)	8 (21.1%)	0 (0.0%)	29 (76.3%)	9 (23.7%)	38

Note. Percentages are row-wise within each group over all responses. 'Without explanation' = 00+10; 'With explanation' = 01+02+11+12.

Additionally, students from the Faculty of Law were more often classified in categories 00 and 10 (answers without explanation), while students of the Faculty of Economics more frequently provided justifications. A chi-square test confirmed a significant faculty difference ($\chi^2 = 20.31$, $p < .001$) about 70% of students of economics gave an explanation, compared to only one quarter of law students. This shaped the category distribution, with students of economics equally divided between 11 (correct with explanation) and 01 (incorrect with equiprobability bias) at 32.4% each, whereas law students most often fell into 00 (incorrect without explanation) (Fig. 3).

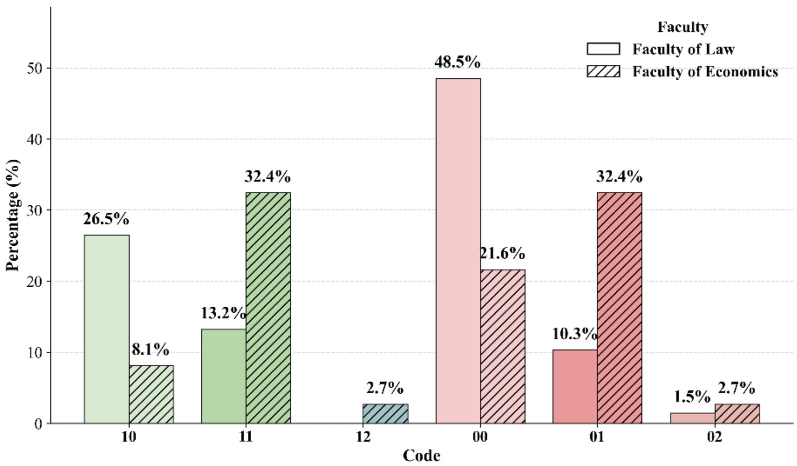

Fig. 3. Distribution of students' explanations across response categories by faculty

Gender differences were also significant, $\chi^2 = 7.34$, $p = .0068$, where roughly half of female students provided an explanation, compared to less than one quarter of males (Figure 4). When only responses with explanations were considered, a significant gender difference emerged (Figure 5). Female students were more frequently classified in category 01 (incorrect answer with equiprobability bias explanation), whereas male students were more often placed in category 11 (correct answer with correct explanation). Male students more frequently provided correct answers with explanations, and this difference was confirmed as statistically significant ($\chi^2 = 4.43$, $p = .035$) when responses with explanations were considered.

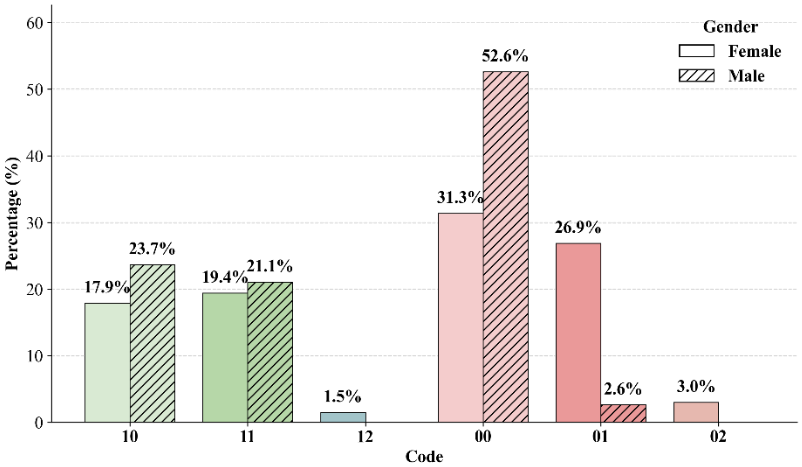

Fig. 4. Distribution of students' explanations across response categories by gender

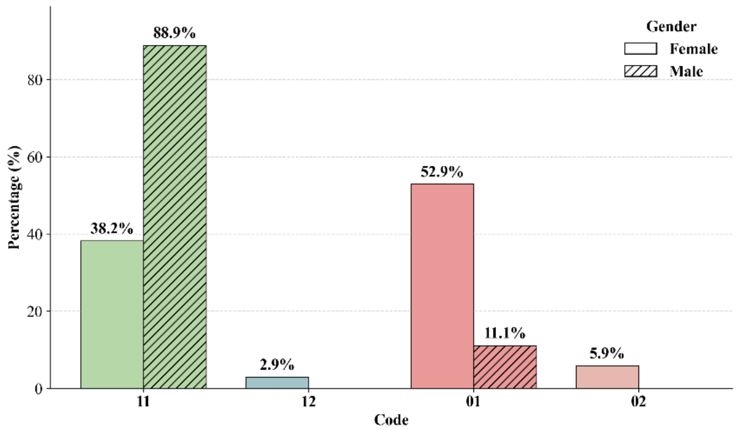

Fig. 5. Distribution of students' explanations across response categories by gender (excluding categories 00 and 10).

These findings indicate that, although no statistically significant differences emerged in the overall choice of options (a, b, or c), a closer analysis of students' explanations reveals meaningful distinctions. Students of economics were more inclined to provide justifications for their answers and, compared to their peers from the Faculty of Law, significantly more often selected the correct option accompanied by a valid explanation (code 11). In terms of gender, female students more frequently provided explanations overall, but male students were more likely to give correct probability-based explanations.

Sections 4.1 and 4.2 together provide the answer to RQ1: about 40% of non-STEM students solved the base-rate task correctly, while nearly half showed the equiprobability bias, with faculty and gender differing mainly in how students explained their answers.

4.3 Attitudes toward statistics

Table 5 presents the descriptive statistics for all items across the four subscales of attitudes toward statistics. Overall, the mean values suggest moderately positive attitudes, with the highest means observed for items reflecting the value and utility of statistics, as well as students' motivation to improve their knowledge and concerns about statistical misuse. Items related to self-efficacy yielded somewhat lower means, indicating that many students do not fully perceive their current knowledge as satisfactory.

Table 5. Descriptive overview of students' responses to attitude items toward statistics

Subscale	Item (abbr.)	N	Mean	SD	Mode	Mean (bar)
Value/Utility						
	STATS_IMPORTANCE	105	3,94	0,88	4	3,94
	CRITICAL_EVALUATION	105	3,26	0,97	3	3,26
	PROFESSION_IMPORTANCE	105	3,1	1,13	3	3,1
	PERSONAL_USAGE	105	3,18	1	3	3,18
Self-efficacy						
	STATISTIC_EDUCATION	105	2,94	1,28	2	2,94
	KNOWLEDGE_SATISFACTION	105	3,02	0,99	3	3,02
Skepticism						
	MEDIA	105	2,48	0,89	3	2,48
	MISUSAGE	105	3,62	0,91	4	3,62
Motivation						
	MOTIVATION	105	3,6	1,14	3	3,6

When compared to STEM students, no substantial difference was observed for value and utility of statistics (Mean = 3.99; Mode = 4), where responses were nearly identical [16]. Regarding students' motivation to improve their statistical knowledge, non-STEM students expressed somewhat lower interest than their STEM peers (Mean = 3.99; Mode = 5), although they still demonstrated a moderate willingness to enhance their skills. Finally, with respect to the belief that statistics are often misused, both STEM (Mean = 3.92; Mode = 5) and non-STEM students expressed similarly high levels of agreement, suggesting a shared perception that misuse of statistics is widespread.

When considering the mode values, additional nuances emerge. For several items (e.g., Importance of statistics and Misusage of statistics), the mode was 4, indicating that the most frequent response leaned toward agreement, even when the overall mean was somewhat lower. By contrast, items such as Statistical education had a mode of 2, showing that the most common response clustered around disagreement.

Interestingly, when comparing STEM and non-STEM students, only minor differences were observed for items related to Statistical education (STEM_Education_Mean = 2.12; Education_Mode = 1) and Media trust (STEM_Media_Mean = 2.36; Media_Mode = 3) [16]. Taken together, these findings indicate that both groups similarly rate their statistical knowledge as insufficient and report limited exposure to statistics during their education. At the same time, both STEM and non-STEM students demonstrated low trust in research reported by the media, suggesting awareness of potential misuse of statistics, while still recognizing the broader importance of statistical knowledge.

Table 6. Comparison of Law and Economics Students' Attitudes Toward Statistics

Item	Subscale	Mean- Faculty of Low	Mean- Faculty of Economic	U	p-value
PROFESSION_IMPORTANCE	Value/Utility	2,66	3,92	487	<0.001
STATISTIC_EDUCATION	Self-efficacy	2,54	3,68	614	<0.001
KNOWLEDGE_SATISFACTION	Self-efficacy	2,85	3,32	901,5	0,0127
MEDIA	Skepticism	2,29	2,81	886	0,0077
MISUSAGE	Skepticism	3,79	3,3	1595,5	0,0167
IMPR	Motivation	3,46	3,86	974,5	0,0491

Mann–Whitney U tests were conducted for nine items to examine potential differences between students of the Faculty of Law in Novi Sad and the Faculty of Economics in Subotica. As shown in Table 6, statistically significant differences emerged for six of these items. Students of the Faculty of Economics reported higher scores on Profession Importance, Statistic Education, Knowledge Satisfaction, Media trust, and Motivation to improve, whereas students of the Faculty of Law expressed stronger agreement with the statement that statistics is often misused. For the remaining three items, no significant differences were observed.

When comparing responses by gender, Mann–Whitney U tests revealed no statistically significant differences across any of the items. In summary, the findings suggest that students from the Faculty of Economics generally demonstrate more positive attitudes toward the value, personal relevance, and acquisition of statistical knowledge, whereas law students are more inclined to perceive statistics as frequently misused. Importantly, no gender-based differences were detected, indicating that these attitudes are more strongly shaped by academic context than by gender.

Findings for RQ2 show that students' attitudes toward statistics vary across the four subscales, with value and utility rated relatively high, while skepticism and low self-efficacy remain notable concerns.

4.4 AI in Analyzing Student Reasoning

To examine the consistency between human and AI coding, Cohen's kappa was calculated, yielding a value of $\kappa = 0.85$. This value indicates a high level of agreement, while the percentage agreement was 88.6%. The findings suggest a stable correspondence between the two coding approaches, although some discrepancies remain that require further analysis and consideration.

Two systematic patterns of disagreement emerged. In nine cases, the researcher coded the responses as 01 – equiprobability bias, while the AI coded them as 02 – alternative incorrect explanation. These were answers implying equal likelihood, such as "it either is or isn't" or "everyone has an equal chance." In three cases, the researcher coded the responses as 11 – probability/statistics-based explanation, while the AI coded them as 12 – correct but non-statistical explanation. These were verbal justifications such as "since there are more engineers, the chance is higher," which represent statistical reasoning but were not recognized as such by the AI unless accompanied by explicit numerical references. Importantly, there were no disagreements in categories without explanations (00 and 10), where coding fully aligned.

Based on these findings, revised prompts were introduced. The new version (AI coding v2) was designed to recognize verbal probabilistic explanations as 11 and to capture equiprobability bias even when not explicitly expressed as "50%." This broader interpretation, however, reduced agreement with the researcher's coding to about 40 percent (63 disagreements out of 105). The reason was again systematic: the AI more often classified ambiguous responses as equiprobability (01 instead of 02) and treated qualitative references to base rates as statistical (11 instead of 12). There were also no errors in answers without explanations (00 and 10), where the coding was consistent.

Overall, the comparison shows that the initial version of AI coding produced high alignment with human coders, with disagreements limited to clear patterns. The revised version broadened the recognition rules, but at the cost of significantly reducing agreement. These findings suggest that AI can be reliable in identifying straightforward cases, but that adjustments in how verbal reasoning is interpreted can substantially alter the results. Finding the right balance between inclusiveness and precision remains a key methodological challenge.

Results for RQ3 indicate that ChatGPT can code clear cases consistently with human raters, but agreement declines when explanations are more complex, suggesting AI has supportive but not substitutive potential.

5 Conclusion

This study gives a better understanding of how non-STEM students think about statistics. It looked at how well they can solve a basic probability problem, how they feel about statistics, and how artificial intelligence (AI) can help analyze their thinking. The results showed that less than half of the students solved the base-rate problem correctly, and many made a common mistake called the equiprobability bias — the belief that all outcomes are equally likely, even when they are not.

When responses were considered collectively, no differences emerged across faculties in the choice of answers. However, a closer examination revealed important contrasts. Students of the Faculty of Economics were more likely to provide written explanations, whereas students of the Faculty of Law more often submitted answers without elaboration. Female students were statistically more likely to provide explanations than their male peers. Among responses with explanations, further gender differences were evident: male students more frequently employed correct probability-based reasoning, while female students more often relied on equiprobability reasoning.

In general, students expressed positive attitudes toward statistics, particularly regarding its value and usefulness. However, many reported low confidence in their own knowledge and indicated that they had not studied statistics sufficiently during their prior education. On the other hand, they showed moderate to high motivation to improve their statistical knowledge. Students were also aware of the frequent misuse of data and, to a considerable extent, expressed a lack of trust in research findings reported in the media. These findings point to an opportunity for developing educational strategies that can enhance statistical knowledge among non-STEM students.

Finally, this study found that AI tools like ChatGPT can help in identifying misconceptions such as base rate neglect, but their accuracy depends on how clearly students express their reasoning and how well the prompts are written. Human oversight is still necessary to ensure reliable interpretation. These findings suggest that AI can support, but not replace, human coders in detecting reasoning errors.

The study is limited by the limited sample size and the descriptive methodology, which restrict the generalizability of the findings. Future research should include larger and more diverse groups of students and apply deeper statistical or mixed-method analyses to strengthen conclusions.

References

1. AlMuraie, E., Algarni, N. A., & Alahmad, N. Sh. (2021). Upper-secondary school science teachers' perceptions of the integrating mechanisms and importance of stem education. *Journal of Baltic Science Education*, 20(4), 546-557.https://doi.org/10.33225/jbse/21.20.546
2. Attali, Y., Laitusis, C., & Stone, E. (2016). Differences in reaction to immediate feedback and opportunity to revise answers for multiple-choice and open-ended questions. *Educational and Psychological Measurement*, 76(5), 787–802. https://doi.org/10.1177/0013164415612548.
3. Bílek, J., Nedoma, J., & Jirásek, M. (2018). Representativeness heuristics: A literature review of its impacts on the quality of decision-making. *Scientific papers of the University of Pardubice*. Series D, Faculty of Economics & Administration, 25(43), 29–38 Retrieved from https://www.scimagojr.com/journalsearch.php?q=21100390414&tip=sid&clean=0.
4. Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: Experimental study. *British Medical Journal*, 333(7562), 284–286. https://doi.org/10.1136/bmj.38884.663102.AE.
5. Carvalho, C., & Solomon, Y. (2012). Supporting statistical literacy: What do culturally relevant/realistic tasks show us about the nature of pupil engagement with statistics? *International Journal of Educational Research*, 55, 57–65. https://doi.org/10.1016/j.ijer.2012.06.006
6. Delmas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2), 28–58.
7. Dönmez, I., Idin, S., & Gürbüz, S. (2022). Determining lower-secondary students' STEM motivation: A profile from Turkey. Journal of Baltic Science Education, 21(1), 38. https://doi.org/10.33225/jbse/22.21.38
8. Faghmous, J. H., & Kumar, V. (2014). A big data guide to understanding climate change: The case for theory-guided data science. *Big Data*, 2(3), 155-163. http://doi.org/10.1089/big.2014.0026
9. Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, 75(3), 372–396. https://doi.org/10.1111/j.1751-5823.2007.00029.x.
10. Gauvrit, N., & Morsanyi, K. (2014). The equiprobability bias from a mathematical and psychological perspective. Advances in Cognitive Psychology, 10(4), 119–130. https://doi.org/10.5709/acp-0163-9.
11. Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S.(2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, Supplement, 8(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x.
12. Hirsch, L. S., & O'Donnell, A. M. (2001). Representativeness in statistical reasoning:Identifying and assessing misconceptions. *Journal of Statistics Education*, 9(2). https://doi.org/10.1080/10691898.2001.11910655.
13. Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristic and biases: The psychology of intuitive judgment*, (pp. 49–81). New York: Cambridge University Press.

14.Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. https://doi.org/10.1037/h0034747.
15.Kang, M., & Park, M. J. (2019). Employees' judgment and decision making in the banking industry: The perspective of heuristics and biases. *International Journal of Bank Marketing*, 37(1), 382–400. https://doi.org/10.1108/IJBM-04-2 018-0111.
16.Kaplar, M. (2022). *Recognizing misconceptions in working with data as a basis to enhance mathematical literacy* (Publication No. 29111073) [Doctoral dissertation, University of Novi Sad]. ProQuest Dissertations and Theses Global.
17.Kaplar, M., Lužanin, Z. & Verbić, S. (2021). Evidence of probability misconception in engineering students —why even an inaccurate explanation is better than no explanation. *International Journal of STEM Education*, 8, Article 18. https://doi.org/10.1186/s40594-021-00279-y
18.Khazanov, L., & Prado, L. (2010). Correcting students' misconceptions about probability in an introductory college statistics course. *ALM Internationa Journal*, 5(1), 23–35.
19.Lužanin, Z., Kaplar, M., & Hrubik-Vulanovic, T. (2022). Undergraduate STEM and non-STEM students' interpretation of mean in an infographic. *Journal of Baltic Science Education*, 21(4), 638-650. https://doi.org/10.33225/jbse/22.21.638
20.Maass, K., Doorman, M., Jonker, V., & Wijers, M. (2019). Promoting active citizenship in mathematics teaching. *ZDM*, 51(6), 991–1003. https://doi.org/10.1007/s11858-019-01048-6
21.Masel, J., Humphrey, P. T., Blackburn, B., & Levine, J. A. (2015). Evidence-based medicine as a tool for undergraduate probability and statistics education. CBE Life Sciences Education, 14(4), 1–10. https://doi.org/10.1187/cbe.15-04-0079.
22.McClure, J., Smyslova, D., Hall, A., & Jiang, S. (2024). Deductive Coding's Role in AI vs. Human Performance. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 809-813).
23.Park, S., Bekemeier, B., Flaxman, A. D. (2021). Understanding data use and preference of data visualization for public health professionals: A qualitative study. *Public Health Nursing*, 8(4), 531-541. https://doi.org/10.1111/phn.12863
24.Smyrnaiou, Z., Georgakopoulou, E., & Sotiriou, S. (2020). Promoting a mixed- design model of scientific creativity through digital storytelling—the CCQ model for creativity. *International Journal of STEM Education*, 7(1), 1–22. https://doi.org/10.1186/s40594-020-00223-6.
25.Williams, R. T. (2024). Paradigm shifts: exploring AI's influence on qualitative inquiry and analysis. *Frontiers in Research Metrics and Analytics*, *9*, 1331589.
26.Yan, L., Echeverria, V., Fernandez-Nieto, G. M., Jin, Y., Swiecki, Z., Zhao, L., Gašević, D., & Martinez-Maldonado, R. (2024, May). Human-ai collaboration in thematic analysis using chatgpt: A user study and design recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-7).