The International Journal of Business Management and Technology, Volume 7 Issue 4July-August 2023
ISSN: 2581-3889

Research Article Open Access

Common misconceptions in basic statistics

José M. Pereira[1], Teresa Abreu[2], Ricardo Gonçalves[2], Mário Basto[2]

[1]*CICF - Research Center on Accounting and Taxation, Polytechnic Institute of Cávado and Ave, Barcelos, Portugal*
[2]*Higher School of Technology, Polytechnic Institute of Cávado and Ave, Barcelos, Portugal*

Abstract: *Probability and statistics play a crucial role in the understanding and accurate interpretation of data. Consequently, they are utilized in virtually every area of scientific research and academic field. On the other hand, improper interpretation of statistical concepts or results can result in erroneous conclusions, which can lead to incorrect decision-making. Numerous authors have highlighted the fact that misunderstandings of probability and statistics can have dire repercussions when it comes to decision-making. These errors of judgment encompass a vast array of statistical topics, including numerous misconceptions regarding probability and statistics. In this study, nine distinct concepts are examined. To attain this objective, an online survey containing nine statements was made available, and respondents were asked to indicate whether they believed each statement to be true or false. The findings were not what one would have hoped for, which highlights the importance of improving statistical education.*

Keywords: Bayes's rule, Coefficient of determination, Cronbach's alpha, Multiple regression, Pearson's correlation, Suppressor variable

I. INTRODUCTION

The correct comprehension of probabilities and statistics plays a crucial role in the vast variety of scientific disciplines. In all fields of science, including physics, biology, economics, psychology, sociology, medicine, and others, statistical analysis and knowledge of probabilities are essential for advancing the understanding of the world ([1], [2], [3], [4], [5], [6], [7]).

Statistics is the field of study that enables the collection, analysis, and interpretation of data. It provides the tools necessary to summarize complex information, recognize patterns and trends, and derive meaningful conclusions from empirical evidence. Through the use of statistics, a deeper comprehension of the phenomena under study can be attained.

Probability and statistics are potent instruments. They permit quantifying uncertainty and estimating the likelihood that an event will occur. One can make predictions, make informed decisions, and test scientific hypotheses based on probabilities. Without a proper grasp of probabilities, it is not possible to make accurate predictions, draw correct conclusions, make rational decisions, and correctly interpret research results.

A proper grasp of probabilities and statistics is essential for empirical progress in all fields of study. Without these instruments, scientific discoveries would be centered on untested hypotheses, and the understanding of the world would be limited. Therefore, scientists and researchers from all fields must have a firm understanding of probabilities and statistics in order to promote trustworthy and accurate conclusions. A thorough comprehension is required in order to extract pertinent information from data and generate meaningful results. Evidence-based decision-making, as opposed to relying solely on intuition or guesswork, experience planning and analysis, forecasting and modeling, critical evaluation of studies and research, identification of biases and errors, detection of patterns and trends, among many other situations, require these skills.

The amount of available information is increasing, but not all of it is reliable. Possessing a solid knowledge of probabilities andstatistics enables one to evaluate scientific studies, opinion polls, market reports, and other data-driven analyses critically. Also, methodological limitations, statistical errors, data manipulation, and misleading conclusions can be identified.However, the general public has misconceptions about a lot of concepts [5], [6], [7]. This is evident in all fields, but especially in the health field [8], [9], [10].The purpose of this study is to evaluate how well the general public comprehends a few basic statistical and probabilistic notions.

II. METHODS

Nine different ideas are explored in this research. This was accomplished by distributing a web-based questionnaire with nine separate statements and asking respondents to identify whether they found each statement to be true or false. A third option enabled respondents to say 'I don't know'. The statements were taken or adapted from [5].

The statements are:

1. Even with a large sample size, a single outlier may be enough to significantly impact the value of Pearson's correlation coefficient. – *true*

Pearson's correlation coefficient is computed by dividing the covariance by the product of the standard deviations. As a result, it seems reasonable to suppose that outliers have little effect because their impact on the numerator and denominator might cancel each other out [11]. However, this does not happen, the Pearson's correlation coefficient may be highly sensitive to data outliers and is hence not robust against them. The authors in [11] demonstrate analytically and through simulations that the existence of outliers can have a huge impact on the coefficient, particularly when they are detectable in both variables at the same time.Therefore, before computing the coefficient, it is advised to perform a visual or statistical assessment for outliers. Otherwise, the results obtained may distort the strength and the direction of the association [5].

Figure 1 and Figure 2 that follow are taken and adapted from Huck (2009).Two hundred observations were collected. On the right side of each Figure an outlier was added. Each Figure clearly demonstrates the influence of the outlier on the value of the Pearson's correlation coefficient.

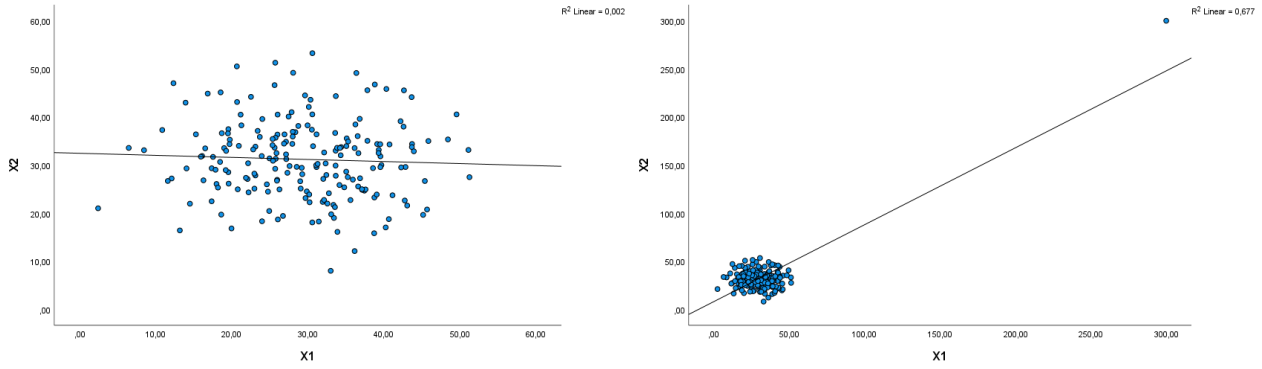

Figure 1. Left: $r_{x_1x_2} = -0.049$. Right: $r_{x_1x_2} = 0.823$.

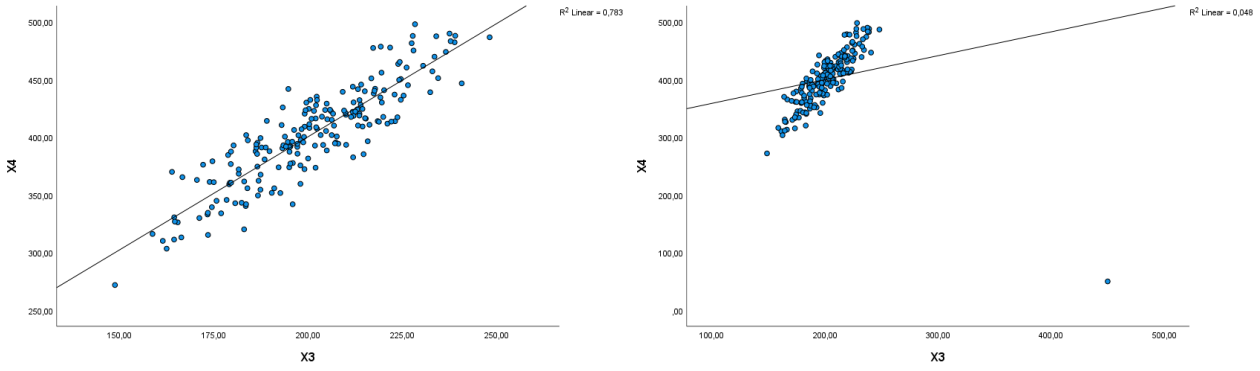

Figure 1. Left: $r_{x_3x_4} = 0.885$. Right:$r_{x_3x_4} = 0.218$.

2. Cronbach's alpha is a statistical measure of the reliability of a measurement instrument, and if it has a high value, it indicates that the items being measured are highly interrelated with one another, therefore confirming the allegation that the instrument measures only one dimension. – *false*

If an instrument is unidimensional, it will have a high Cronbach's alpha value. The opposite, however, is not true. The belief that Cronbach's alpha measures unidimensionality is inaccurate. A multidimensional scale can produce high values of Cronbach's alpha.

If the items on a measurement tool do not measure the same thing, two very different individuals with the same total score may appear to be similar, and a high Cronbach's alpha value can be attained [5]. A high value is necessary but not sufficient for unidimensionality.

3. If a fair coin is flipped n times, with n being an even number, the probability of obtaining as many heads as tails increases as n increases. – *false*

The number of heads (or tails) in n fair coin flips matches to a binomial distribution, $B(n, 0.5)$. The greater the number of times a coin is flipped, the greater the likelihood that the proportion of heads (or tails) approaches 0.5, or the number of heads (or tails) approaches its average value, $0.5n$. However, as n increases, the disparity between the number of heads and tails tends to grow. And, as n increases, the likelihood of getting exactly as many heads as tails decreases.

4. Among a random sample of 25 people, it is more likely that two or more individuals share the same date of birth, that no one shares the same date of birth. – *true*

If no one was born on February 29 and everyone was born with an equal chance on all other days of the year (if accounted for, these factors would have only a minimal impact on the calculated probability [5]), the likelihood that at least two people were born on the same day is:

$$1 - \frac{365 \cdot 364 \cdot 363 \cdot \cdots \cdot 341}{365^{25}} = 0.5687 = 56.87\%$$

This problem is often referred to as The Birthday Paradox. The sample size could be reduced to 23 individuals and the result would still hold.

5. If a person is diagnosed with an extremely rare and fatal disease by a routine screening procedure with a 99% accuracy rate, then the likelihood of death is quite high. – *false*

Consider the following events:

$T \rightarrow$ the screening procedure is positive

$D \rightarrow$ the person has the fatal disease

The text says that the sensitivity and specificity are equal to 0.99, which, translated into conditional probability language, are expressed, respectively, by:

$$P(T|D) = 0.99$$

$$P(\bar{T}|\bar{D}) = 0.99$$

The disease is also known to be extremely rare, therefore $P(D)$is quite low. The statement says that under these conditions,$P(D|T)$ is quite high, which is false.

Using Bayes's rule to calculate $P(D|T)$, one finds that this probability is given by:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} = \frac{0.99 \cdot P(D)}{0.99 \cdot P(D) + 0.01 \cdot [1 - P(D)]}$$

Assuming that the prevalence of the disease that is extremely rare is one in a thousand, or $P(D) = 0.001$, the probability of contracting the disease is 9.02%, which is a value very far from a likelihood of death quite high.

6. When there is a statistically significant correlation between two variables, it indicates that a strong relationship exists. – *false*

Some researchers tend to mistake statistical significance with practical significance. These are two distinct concepts. Both are essential for interpreting a study's results. Statistical significance refers to the existence of an effect in the

population, which is typically evaluated using the so-called p-value, whereas practical significance refers to the importance or relevance of the effect found in practice or in the real world.

7. If two 95% confidence intervals around two means partially overlap, there is no statistically significant difference between the two means at $\alpha = 0.05$. – *false*

It is a common misconception that if two 95% confidence intervals for two independent means overlap, there is no statistically significant difference between the two population means at$\alpha = 0.05$. However, this is not always the case.When two confidence intervals are being compared, each of them is built using information from a single sample only, excluding information from the other sample. On the other hand, a single confidence interval for the difference between two means is created using two samples data. This way, the inference is more precise since it can identify a significant difference that may not be observed when comparing two different confidence intervals.

8. If the coefficient of determinationR^2 is large, that does not imply necessarily that the regression model is good at explaining variance in the dependent variable. – *true*

The coefficient of determinationR^2 represents the proportion of the dependent variable's variability that can be accounted for by the independent variables and is unaffected by sample size. It appears that the model is better the higher its value. But this is not the case. R^2 rises as the number of independent variables increases, and this could lead one to conclude that the optimal model would have the greatest number of independent variables conceivable. This is false. This issue of fitting a model with an excessive number of independent variables is known as overfitting. The closer the number is to the sample size, the greater the severity of the problem and the higher the value of R^2. When the number of independent variables (not perfectly correlated) equals the number of observations, the value of R^2 reaches 1, meaning that, theoretically, 100 percent of the dependent variable can be explained by the independent variables, which is absurd. Many claim that there should be at least a 20:1 ratio between the number of cases and the number of independent variables.

9. An independent variable in multiple regression must not be included in the model if there is no correlation between it and the dependent variable, because this will not increase the coefficient of determinationR^2. –*false*

Identifying the best set of independent variables that, taken together, can explain variation in a dependent variable, is one of the objectives of multiple regression. When included to the model with other independent variables, a predictor variable that appears to be of little value by itself due to a low or zero correlation with the dependent variable, but which has a significant correlation with one or more of the existing predictors in the model, can actually boost the coefficient of determination R^2. This type of variable is referred to as a suppressor variable, because it is associated with one or more of the other independent variables, thereby reducing the error variance in those variables, making them work better and raising the coefficient of determinationR^2[5], [12]. When a suppressor variable is included in the model, the coefficients of the original predictors may change, and their associations with the dependent variable may even have their signs reversed. A suppressor variable modifies the existing link between an independent and dependent variables by strengthening its effect.

Importantly, the introduction of a suppressor variable can affect the interpretation of model coefficients. Taking into account the influence of the suppressor variable, the coefficients now reflect the partial relationships between the predictors and the dependent variable. Therefore, it is essential to interpret the coefficients with the suppressor variable present in mind.

Consider the following example (Table 1), where $y$is the dependent variable and $x_1$and x_2 are two independent variables.

Table 1. Example of multiple regression with a suppressor variable x_2

Case	y	x_1	x_2
1	-7.21433717	3.16745244	3.23661445
2	0.36224337	12.56070413	3.99484118
3	-6.40415948	14.33630717	5.57720982
4	1.84575556	9.10787456	2.79919486
5	-2.34561584	7.62777916	3.91072764
6	-2.79095476	17.29648231	5.52687923
7	-3.27254111	2.53812838	0.86794711

8	-4.63563483	7.23477809	2.80632489
9	-5.54578178	3.28923947	3.32015177
10	-3.97431288	2.49909973	2.29869754
11	-0.2766716	6.18219001	2.80903918
12	-9.68243928	2.26751121	3.96817416
13	-5.21017951	12.54368052	4.41753195
14	-0.84045798	4.54595722	2.1536441
15	-2.67207894	5.22402551	-0.00535812
16	5.76922829	20.57719243	3.61765302
17	-5.50489779	10.20776735	3.67160451
18	-5.57107581	4.37960713	2.72289329
19	5.55989216	24.84842032	6.65076949
20	2.75128652	7.26961027	3.06653662
21	0.87613536	6.85897648	3.03011809
22	-10.43113994	18.18673531	7.70624399
23	-4.1911633	2.95250774	1.91976777
24	-0.35845402	5.55045977	2.12736794
25	1.57501981	18.63355299	5.47802183
26	-4.51168595	17.91602437	4.91656365
27	-1.40286654	11.77654765	5.10865503
28	-0.25480191	10.75926376	3.82811486
29	0.56593228	19.2023676	8.90015516
30	-1.67865531	9.70607121	1.63218198

The Pearson correlations are:

$$r_{yx_1} = 0.386; r_{yx_2} = 0.000; r_{x_1x_2} = 0.770.$$

The coefficients of determination in the model with a single independent variable are:

$$R^2_{yx_1} = 0.386^2 = 0.149$$

$$R^2_{yx_2} = 0.000$$

With these data, it appears that the variable x_2 is unnecessary for the linear model which already includes the independent variable x_1.

However, the partial and semi-partial correlations (the former being expressed in relative terms and the latter in absolute terms) are:

$$r_{yx_1(x_2)} = 0.605; r'_{yx_1(x_2)} = 0.605$$

$$r_{yx_2(x_1)} = -0.505; r'_{yx_2(x_1)} = -0.466$$

The multiple coefficient of determination can be computed as either of the two methods listed below:

$$R^2_{yx_1x_2} = R^2_{yx_1} + (1 - R^2_{yx_1})r^2_{yx_2(x_1)} = R^2_{yx_1} + r'^2_{yx_2(x_1)} = 0.149 + (-0.466)^2 = 0.149 + 0.217 = 0.366$$

$$R^2_{yx_1x_2} = R^2_{yx_2} + (1 - R^2_{yx_2})r^2_{yx_1(x_2)} = R^2_{yx_2} + r'^2_{yx_1(x_2)} = 0.000 + 0.605^2 = 0.366$$

As can be observed, x_2 has a zero linear correlation with the dependent variable y. However, when x_2 is introduced into the linear regression model in which y is the dependent variable and x_1 is the independent variable, the correlation of x_2 with y, which was zero, increases in absolute value when this correlation is adjusted for the presence of the variable x_1 in the model, i.e., the absolute value partial and semi-partial correlations of $x_2$with y increase. In

addition, the absolute value of the partial and semi-partial correlations of x_1 with y increases relatively to the unadjusted correlation. This changes result in an increase of R^2 from 0.149 to 0.366.

III. Results

The questionnaire was accessible online from February 1 to April 14, a span of two months and fourteen days. The final sample comprised 127 individuals, consisting of 29 without a degree level, 66 with a degree level, and 32 with a master's degree or doctorate. There were 65 females and 62 males present. Everyone's age ranged from 20 to 69 years old.

1. Statement 1 (true). Even with a large sample size, a single outlier may be enough to significantly impact the value of Pearson's correlation coefficient.
   Figure 3 shows the distribution of responses by academic qualifications.

	Qualifications			
Statement 1	No degree level	Degree level	Master or PhD	Grand Total
Don't know	82,8%	68,2%	59,4%	69,3%
False	6,9%	7,6%	21,9%	11,0%
True	10,3%	24,2%	18,8%	19,7%

Figure 3. Distribution of responses to Statement 1 by academic qualifications (correct answers appear in green).

The majority answered that they do not know the answer. Those with a master's degree or doctorate were less likely to provide the correct response than those with a degree level, although there were no statistically significant differences in the responses of respondents with different academic backgrounds ($\chi^2(4) = 7.799, p = 0.099$).

2. Statement 2 (false). Cronbach's alpha is a statistical measure of the reliability of a measurement instrument, and if it has a high value, it indicates that the items being measured are highly interrelated with one another, therefore confirming the allegation that the instrument measures only one dimension.
   Figure 4 shows the distribution of responses by academic qualifications.

	Qualifications			
Statement 2	No degree level	Degree level	Master or PhD	Grand Total
Don't know	89,7%	87,9%	75,0%	85,0%
False		4,5%	6,3%	3,9%
True	10,3%	7,6%	18,8%	11,0%

Figure 4. Distribution of responses to Statement 2 by academic qualifications (correct answers appear in green).

Very few respondents provided the correct response, and none of those without at least a degree level did so. This assertion focused on an issue that many people are less familiar with: the Cronbach's alpha. Also, the vast majority of respondents indicated that they did not know. As a result, the poor results attained were somehow expected.

The differences in the responses of respondents with different academic backgrounds were not statistically significant ($\chi^2(4) = 4.610, p = 0.330$).

3. Statement 3 (false). If a fair coin is flipped n times, with n being an even number, the probability of obtaining as many heads as tails increases as n increases.
   Figure 5 shows the distribution of responses by academic qualifications.

Statement 3	Qualifications: No degree level	Degree level	Master or PhD	Grand Total
Don't know	34,5%	24,2%	9,4%	22,8%
False	27,6%	28,8%	37,5%	30,7%
True	37,9%	47,0%	53,1%	46,5%

Figure 5. Distribution of responses to Statement 3 by academic qualifications (correct answers appear in green).

This was the statement with the highest proportion of correct responses.

The differences in the responses of respondents with different academic backgrounds were not statistically significant ($\chi^2(4) = 5.735, p = 0.220$).

4. Statement 4 (true). Among a random sample of 25 people, it is more likely that two or more individuals share the same date of birth, that no one shares the same date of birth.
   Figure 6 shows the distribution of responses by academic qualifications.

Statement 4	Qualifications: No degree level	Degree level	Master or PhD	Grand Total
Don't know	24,1%	28,8%	15,6%	24,4%
False	51,7%	56,1%	65,6%	57,5%
True	24,1%	15,2%	18,8%	18,1%

Figure 6. Distribution of responses to Statement 4 by academic qualifications (correct answers appear in green).

When comparing the response rate of those who believe they know the correct answer, the illusion of knowledge is manifest in this statement.

The differences in the responses of respondents with different academic backgrounds were not statistically significant ($\chi^2(4) = 2.999, p = 0.558$).

5. Statement 5 (false). If a person is diagnosed with an extremely rare and fatal disease by a routine screening procedure with a 99% accuracy rate, then the likelihood of death is quite high.
   Figure 7 shows the distribution of responses by academic qualifications.

Statement 5	Qualifications: No degree level	Degree level	Master or PhD	Grand Total
Don't know	31,0%	22,7%	12,5%	22,0%
False	13,8%	21,2%	12,5%	17,3%
True	55,2%	56,1%	75,0%	60,6%

Figure 7. Distribution of responses to Statement 5 by academic qualifications (correct answers appear in green).

The illusion of knowledge is also patent in this statement.

The differences in the responses of respondents with different academic backgrounds were not statistically significant ($\chi^2(4) = 5.073, p = 0.280$).

6. Statement 6 (false). When there is a statistically significant correlation between two variables, it indicates that a strong relationship exists.
   Figure 8 shows the distribution of responses by academic qualifications.

Statement 6	Qualifications: No degree level	Degree level	Master or PhD	Grand Total
Don't know	37,9%	22,7%	3,1%	21,3%
False	6,9%	19,7%	12,5%	15,0%
True	55,2%	57,6%	84,4%	63,8%

Figure 8. Distribution of responses to Statement 6 by academic qualifications (correct answers appear in green).

This assertion displays a strong case of illusion of knowledge and exposes a persistent mistake that confuses statistical significance with real-world relevance.

In this case, thedifferences in the responses of respondents with different academic backgrounds were statistically significant ($\chi^2(4) = 14.051, p = 0.007$).

7. Statement 7 (false). If two 95% confidence intervals around two means partially overlap, there is no statistically significant difference between the two means at alpha = 0.05.
   Figure 9 shows the distribution of responses by academic qualifications.

Statement 7	Qualifications: No degree level	Degree level	Master or PhD	Grand Total
Don't know	86,2%	74,2%	46,9%	70,1%
False	6,9%	6,1%	6,3%	6,3%
True	6,9%	19,7%	46,9%	23,6%

Figure 9. Distribution of responses to Statement 7 by academic qualifications (correct answers appear in green).

A large majority of individuals did not know the answer. Also, very few responses are correct.

The differences in the responses of respondents with different academic backgrounds were statistically significant ($\chi^2(4) = 14.910, p = 0.005$).

8. Statement 8 (true). If the determination coefficient R^2 is large, that does not imply necessarily that the regression model is good at explaining variance in the dependent variable.
   Figure 10 shows the distribution of responses by academic qualifications.

Statement 8	Qualifications: No degree level	Degree level	Master or PhD	Grand Total
Don't know	89,7%	75,8%	53,1%	73,2%
False	10,3%	18,2%	12,5%	15,0%
True		6,1%	34,4%	11,8%

Figure 10. Distribution of responses to Statement 8 by academic qualifications (correct answers appear in green).

This was the statement in which individuals with a master's or doctorate degree provided the vast majority of accurate answers.

The differences in the responses of respondents with different academic backgrounds were statistically significant ($\chi^2(4) = 22.960, p < 0.001$).

9. Statement 9 (false). An independent variable in multiple regression must not be included in the model if there is no correlation between it and the dependent variable, because this will not increase the coefficient of determinationR^2.
   Figure 11 shows the distribution of responses by academic qualifications.

	Qualifications			
Statement 9	**No degree level**	**Degree level**	**Master or PhD**	**Grand Total**
Don't know	93,1%	80,3%	62,5%	78,7%
False	3,4%	4,5%	6,3%	4,7%
True	3,4%	15,2%	31,3%	16,5%

Figure 11. Distribution of responses to Statement 9 by academic qualifications (correct answers appear in green).

There was a very low percentage of correct responses. This statement was not trivial, which may explain the low percentage of correct answers.

The differences in the responses of respondents with different academic backgrounds were almost statistically significant ($\chi^2(4) = 9.385, p = 0.052$).

IV. Conclusion

More than 70% of respondents selected 'don't know' answering to statements 1, 2, 7, 8 and 9. For the remaining statements 3, 4, 5 and 6, this response was selected by between 21.3% and 24.4% of respondents. These were the statements regarding probability calculation and statistical significance.In statements 4, 5, and 6, the illusion of knowledge is apparent.

In almost all instances, the proportion of respondents who responded 'don't know' was greater among those with fewer academic credentials and decreased as credentials increased. The proportion of respondents with the correct response was not proportional to their qualifications, but rather was dependent on the specific statement.

Given that statistics are used in virtually all scientific fields and increasingly in people's daily lives, these findings highlight the need to strengthen statistical education so that information is correctly processed and interpreted, resulting in accurate, well-informed decisions.

References

[1] J. Pearl (2009),*Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2nd edition, 2009).

[2] D.S. Moore, G. P. McCabe, and Bruce A. Craig,*Introduction to the practice of statistics* (W.H. Freeman, 2014).

[3] R.L. Wasserstein, and N.A. Lazar, The ASA statement on p-values: Context, process, and purpose,*The American Statistician, 70(2),*2016, 129-133.

[4] M.H. Kutner, C.J. Nachtsheim, J. Neter, and W. Li (2004),*Applied Linear Statistical Models* (McGraw-Hill/Irwin, New York, 5th Edition, 2004).

[5] S. W. Huck,*Statistical misconceptions*(Psychology Press, Taylor & Francis Group, New York, 2009).

[6] P.I. Good,and J.W. Hardin,*Common Errors in Statistics (and How to Avoid Them)*(John Wiley & Sons, Inc., Hoboken, New Jersey, 2nd edition, 2006).

[7] M. Basto, T. Abreu, R. Gonçalves and J.M. Pereira, Example-based overview to statistical inference, *Advanced Mathematical Models & Applications, 8 (1)*, 2023, 5-13.

[8] C. Ferreira, T. Abreu, and M. Basto, p-value misconceptions in health professionals and patients,*Advances and Applications in Statistics, 63(1),* 2020, 75-83.

[9] C. Ferreira, T. Abreu, and M. Basto, Perception of transmitted risk in healthcare, *Journal of Public Health, 30*, 2022, 1245–1249.

[10] C. Ferreira, T. Abreu, R. Gonçalves, J.M. Pereira, and M. Basto, Survival and risk perceptions of healthcare professionals and patients,*Indonesian Journal of Social Sciences, 15(1),*2023, 1-7.

[11] Y. Kim, T.-H. Kim, and T.Ergün, The instability of the Pearson correlation coefficient in the presence of coincidental outliers, *Finance Research Letters, 13*, 2015, 243-257.

[12] K.K. Woolley, How Variables Uncorrelated with the Dependent Variable Can Actually Make Excellent Predictors: The Important Suppressor Variable Case, *Annual Meeting of the Southwest Educational Research Association*, Austin, TX, January 23-25, 1997.