**Concept Interview: P-values & AI created content**

**Hannah N is currently taking statistics, just recently took the final exam this past week. In Hannah’s class, they use the book and video that uses SBI but the professor has their own self-created materials, not the provided investigations.**

|  |
| --- |
| Thank you so much for being willing to participate in this interview. In this interview, we are interested in **your thinking**. You are the expert in your own thinking, so it is impossible to be wrong! Helping instructors understand how students think about and talk about statistics is going to help many students learn statistics better.  You are welcome to use any resources you like, including notes, applets, paper, or anything else that would help you. We ask you not to use the Internet at large, for the sake of time, but if there are any web-based resources you would like to use, you are welcome to. |

Can people reliably tell the difference between AI created content and human recreated content?

Artificial intelligence (AI) can be used to write, create art, and compose music. As AI becomes more common, determining whether humans can distinguish between AI-created content and human created content becomes increasingly important.

1. If you were a researcher trying to answer this research question, how would you design a study?

Student said she would randomly select people as best she could, have a form to sign up, it would be voluntary, as many people as she could. She would break it into different groups, by which she meant she would have some questions about music, some about text, some about art and ask the participant whether they thought it was created by AI or a human.

Some tools available on the Internet allow a person to test how well they can distinguish AI-created content from human-created content. These tools are set up like a quiz. In each question, the person looks at an object, such as photographs, art, or a piece of writing. They then guess whether the object was created by AI or not. Here are some examples of questions.

Example 1: Images Example 2: Text

A screenshot of a computer screen

Description automatically generatedA screenshot of a cat

Description automatically generated

1. Can you explain this scenario in your own words?

Student felt this was similar to their plan.

1. If someone was not able to distinguish between AI created content and human created content, how well do you think they would do on a set of questions like these? How many would they get right? How many would they get wrong?

Student felt that the person would get the majority wrong, but not all wrong. There might be a handful that were correct. Out of 10, the person might get 2 right if they were not able to tell the difference between AI and human created content. Unless they were good at guessing. Good at guessing meant that either they used a strategy, like always guess c, which lined up with the quiz, or else they had some kind of good ability to look at answer choices and know which one was more likely to be right, even if they did not know the content.

1. A professor tried a set of questions like these about identifying whether a human or AI had written a given block of text. She got 4 out of 4 correct. What are your thoughts about these results?

The student had these theories:

* The professor had already seen the text or images beforehand
* The professor was really good at guessing
* The professor really was good at detecting whether things were created by AI, they were tech savy.
* Maybe all of the above could be true
* It depends upon the images, some are more obvious than others.

If there were 10 questions and the professor got all 10 correct, then the student felt there has to be something going on – maybe the professor was good at detecting, or maybe there was a study sheet, or maybe the images were obvious.

1. If you had to model the idea of randomly guessing on a set of 4 questions like the example questions, how would you do it? You could use spinners, dice, coins, the applets you use in class, or other methods.

The student would use coin flips because coins are 50/50 and the questions are 50/50, too. There are only 2 answers, human or AI, and there’s only heads or tails for a coin flip.

She would flip the coin 4 times because there are 4 questions, but then wondered if she should flip 5 times for good luck, but settled on 4 for the 4 questions.

1. If we were investigating the question of whether this professor could detect AI written text, what would the null hypothesis be, in your own words? What do you think the alternative hypothesis would be?

Student reported that the null would be Ho = 4 and Ha < 4 because people would get fewer than 4 questions correct.

When asked if there was a 10-question quiz what the hypotheses would be, student reported that the null would be Ho = 10 and Ha < 10. People would not get all of them right.

Just a note: in an earlier interview, Hannah N did not use proportions when trying to decide how to make two groups that were balanced in terms of gender. She used an alternating strategy, putting 1 man in each group, and then 1 woman in each group, until all of the participants were assigned to a group. Others in math ed have found that there can be issues with understanding continuous quantities and how students will go back to whole number thinking, so maybe this is an example of that. Hannah started with the symbolics, “Ho is equal to” instead of starting with the contextual “People are guessing…” which is in contrast to Keelie’s interview, where she did not talk about the quantitative or symbolic but only the contextual after getting a definition.

Researcher did not clarify the hypotheses issue, since clinical interviews are about capturing. (Student had already completed final exam.) Appreciate thoughts from colleagues on whether students’ thinking should be addressed here, because with a correction the student might have been better able to show their thinking later? Because ideas were not conflicting? But then maybe that is not authentic.

1. A student used the Rossman Chance applets to simulate this situation. The student reasoned that if the professor were guessing whether the block of text was written by either AI or by a human, there were only two possibilities, so it would be equivalent to tossing a coin. If there were four questions, that would be like tossing four coins. The student used a computer to toss four coins and recorded the number of heads in each toss of 4 coins. The graph of the first 100 groups of 4 coin tosses is below.

A graph with numbers and lines

Description automatically generated

How might you use this graph to think about how strong the evidence is that the professor can reliably determine if a block of text is written by AI or a human?

This question needs some work for clarity.

Student said that most people got 1-3, so she would reject the null hypothesis because most people got less than four. (I do believe that she meant most people would get less than four *by blind guessing,* as represented by the coin flips*.)* She noted that 1, 2, 3, are higher/more frequent than 4.

Student said that the evidence is not that strong that the professor who scored 4/4 can detect whether content was created by AI because “0 is close to getting all 4 right,” meaning that 0 heads out of 4 coin tosses is as frequent as 4 heads out of 4 coin tosses. I think this is her logic: If it’s equally likely that you can get all four wrong by blindly guessing as getting all four right by blindly guessing, then there’s not a lot of evidence to show that the professor isn’t blindly guessing. The student seemed to indicate that the professor *could* be able to detect AI, but the evidence isn’t strong enough for us to know.

1. The one-sided p-value in this instance is 0.0628. How do you make sense of that p-value? What does that p-value mean to you?

Student said: The p value is greater than 0.05, so we don’t reject it, so the null is correct and the alternative is wrong. Wouldn’t have strong enough evidence to reject the null, not enough to support the alternative. Saw this as the opposite of what she had said earlier.

Researcher wondering: sometimes students say something like the null is correct and not the alternative (which is not how hypothesis testing works), but it seems like this is an intermediary thought or maybe a short cut, because then they do go on to qualify with language about “could” and “not enough evidence.” Maybe they are just adopting instructor’s language, but it did feel like in this interview that the student was explaining the room for doubt and strength of evidence, and the “this one is right, this one is wrong,” could be kind of like a short-hand as they think?

1. Suppose the professor had gotten 8 out of 8 possible questions correct. How would that change your strength of evidence or your conclusion?

Student said: This wouldn’t really change their conclusion. If it was put on the graph, it would be on the lower side compared to everything else. (Lower frequency?) Felt like there would not be enough evidence, maybe if he had 20 questions that would be stronger. Maybe he’s really good at guessing or if the questions are easy or if he created the questions.

Other types of quizzes have questions where the person has to choose the real image from a selection of images that includes AI editing, like the question below.

A collage of a child

Description automatically generated

A collage of a person

Description automatically generated

1. How is this situation different from the earlier situation?

Student noted that there are 6 options instead of 2, so now there is less of a chance of guessing correctly. It is more difficult to choose the correct image. Probably there will be more people getting incorrect answers than correct answers.

1. If there were a series of questions like these, what would the null and alternative hypotheses be?

Ho = 6 and Ha < 6 because there are 6 images

One of the quizzes has four questions like this. On these questions, the professor got 3 out of 4 correct.

1. How would you simulate this situation?

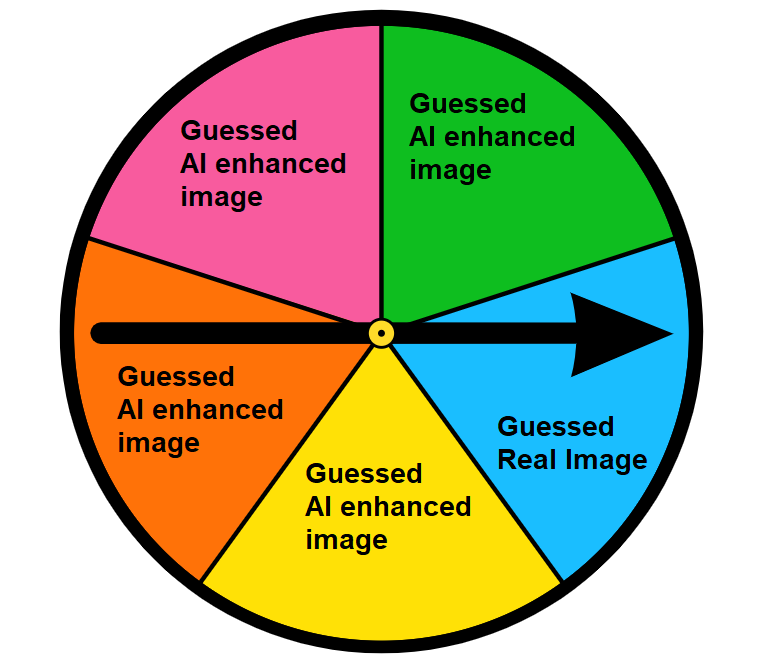
Student reported that they would use a die because a die has 6 faces and there are 6 images to choose from. They reported that they would roll the die 4 times to represent the 4 questions. Student had a little bit of a hard time determining which numbers would represent which outcomes, but eventually settled on 6 would be guessing which image was the real image correctly, because it is difficult to roll a 6, and then all of the other numbers on the die matching to choosing the AI images as the original image.

1. How do you think the p-value would compare to the p-value you found earlier?

The student believed that the p value would be smaller because it would be harder to get a lot of questions correct just by guessing.

Note: put a box in here to have all of the details about the two different situations in one place beside each other

1. One student decided to spin a spinner with six sections four times to simulate this situation. One section of the spinner was labeled as guessing the correct image, as shown below. Within each set of four spins, the student counted the number of times the spinner landed on the section representing guessing the correct image. The graph of the first 100 simulations is shown below.

A graph of a number of sucess

Description automatically generated

1. How could you use this graph to think about the strength of evidence about whether the professor was guessing?

In this case, the student felt that there was very strong evidence that the professor could detect the difference between AI edited and real images because getting 3 out of 4 correct happened so relatively rarely, compared to the frequency of getting 0, 1, or 2 of the questions correct by guessing.

1. Can you estimate the p-value from this graph?

Said that the p value was very small. Said that the alternative hypothesis was less than 6 correct, which is what happened, so you would reject the null hypothesis.

1. What does this p-value mean to you?
2. Thank you for helping us design these prompts to help us learn about student thinking. Is there anything in the prompt that was confusing or could have been clearer?

Source: <https://www.tidio.com/blog/ai-test/>