

# Deriving Realistic Examples of Statistics Problems for Life Science Majors Using LLM's

Jacob Andros, Mohamed Aburweis

Texas A&M University

## Stat 302: Statistical Methods for the Life Sciences

- Intro statistics class for undergrads majoring in **Agriculture, Animal Science, Anthropology, Biology, Ecology, and Nutrition.**
- Most common complaint in student evaluations: “The course did not feel very applicable to my field of study.”

## The Challenge?

- We are statisticians, not experts in agriculture, biology, ecology, etc.
- Rewriting problems in a life sciences context requires extensive time/research on our part, or collaborations with colleagues in other departments.
- Fortunately, LLM's like ChatGPT can be used to streamline the process!

## Example problem from **current course slides**:

- A survey asked **200 residents** of College Station if they use Verizon for their cell phone carrier. **38%** of the respondents said they use Verizon. At  $\alpha = .05$ , do these data provide convincing evidence that the proportion of College Station residents who use Verizon is **different from 1/3**?

## Rewritten example directly related to **ecology**:

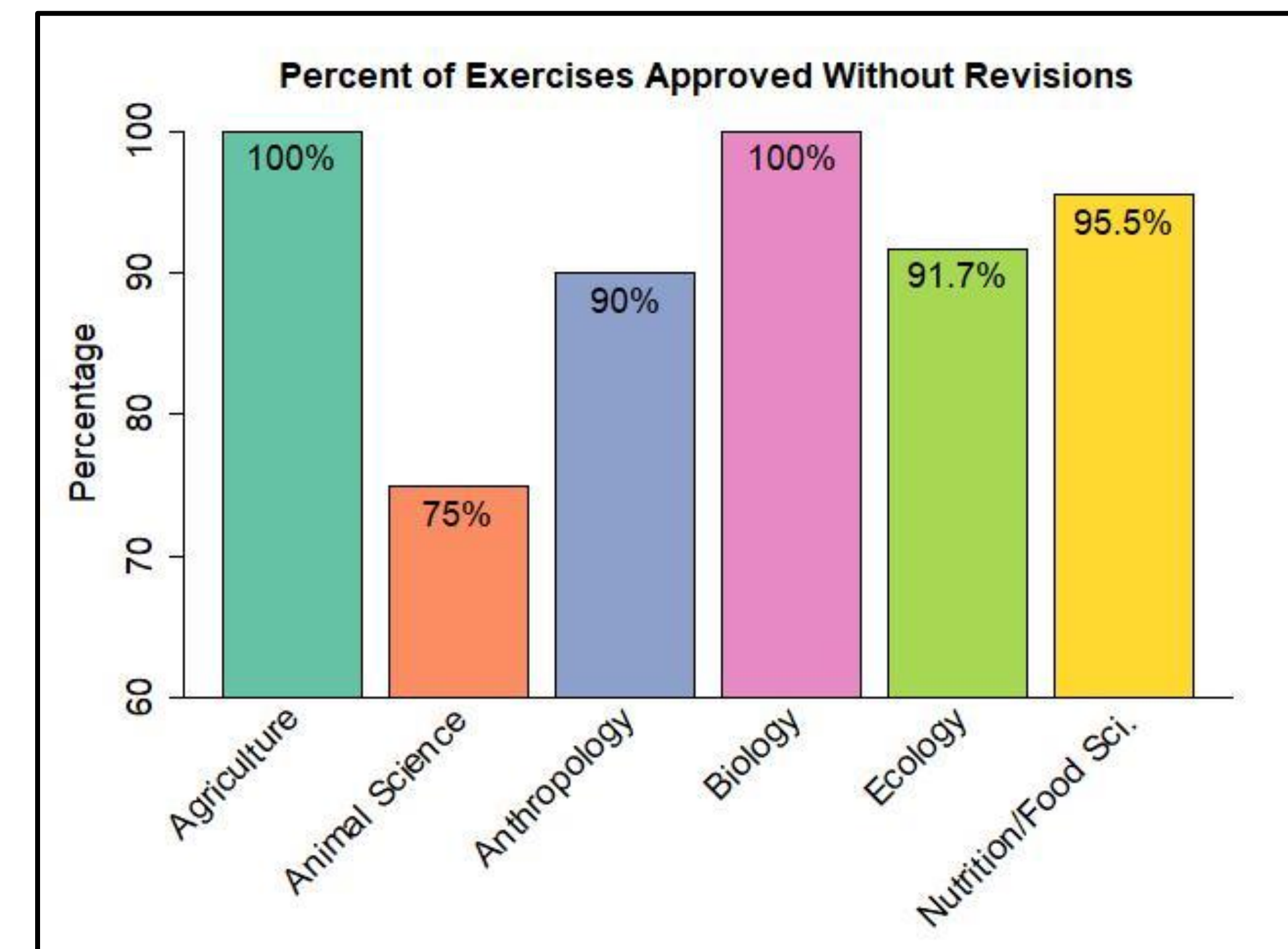
- Researchers surveyed **200 trees** in a national forest to determine whether they were infested with a particular species of bark beetle. They found that **38%** of the sampled trees showed signs of infestation. At  $\alpha = .05$ , do these data provide convincing evidence that the proportion of trees infested from the entire forest is **different from 1/3**?

## Data Collection

- The 6 life science fields were each randomly assigned 3 course chapters.
- We generated 8-10 exercises for each field and chapter using ChatGPT's 4o model.
- Exercises were then reviewed by professors in those departments for clarity and correctness.

## Most Common Revisions

- Unit of measurement (4) – Animal Sci. and Ecology
- Word change (3) – Anthropology and Ecology
- Number change (2) – Animal Sci. and Nutrition
- Threw out question (1) - Anthropology



## Example Prompt and Exercise

- Prompt: Give a **simple** example of how to calculate a **one-sample t interval** in the context of **ecology**.
- Exercise: An ecologist collects a random sample of **n = 10** water samples from a lake and measure the nitrogen levels. The sample has mean nitrate concentration of **3.2 mg/L** and a sample standard deviation of **0.5 mg/L**. Calculate a **95%** confidence interval for the mean nitrate concentration level in the lake.
- Evaluation: This checks out – freshwater typically has a nitrate concentration between 1 and 4 mg/L.



# Two Important “Tuning Knobs”:

## Difficulty Level (Simple vs. Complex)

Prompt: Give a (**simple/complex**), conceptual question about correlation in the context of biology.

Simple: A biologist measures the **body length** and **lung capacity** of a group of reptiles and calculates a **correlation coefficient of  $r = 0.85$** . What does this value tell you about the relationship between body length and lung capacity, assuming that it is indeed a linear relationship?

Complex: Researchers studying a population of mammals find a **strong negative correlation ( $-0.82$ )** between **daily activity level** and **body fat percentage**. However, a follow-up experiment shows **that increasing activity level does not** consistently lead to reductions in body fat. How can you explain the discrepancy between the strong negative correlation and results from the experiment?

## Approach (Conceptual vs Computational)

Prompt: Give a simple, (**conceptual/computational**) question about confidence intervals for proportions in the context of agriculture.

Conceptual: A farmer randomly samples **200 plants** and find that **40 of them** (20%) show signs of pest damage. Using this sample, the **95% confidence interval** for the true proportion of all corn plants affected by the pest is (**0.14, 0.26**). How should they interpret this 95% confidence interval in context?

Computational: A farmer wants to estimate the proportion of her apple trees that are producing high-quality fruit. She randomly samples **200** trees and finds that **153** of them are producing high-quality fruit. Construct a **95% confidence interval** for the true proportion of apple trees producing high-quality fruit.

# Limitations and Follow-up Prompts:

## Common Mistakes and Limitations:

- Frequently restates the field (“in the context of \_\_\_”) at the beginning of each question. (“An animal science study finds that...”, “An ecologist is studying...”).
- Does not always use feasibly small numbers for probability exercises involving the binomial distribution (assuming students are using a calculator and not R).
- Sometimes misuses causal terms in hypothesis test questions.
- Frequently uses only multiples of 5 and 10 for the numbers in computational problems. The keyword “complex” sometimes fixes this, but not always.

Original Prompt: Give a simple, computational question about performing a chi-square test for independence in the context of ecology.

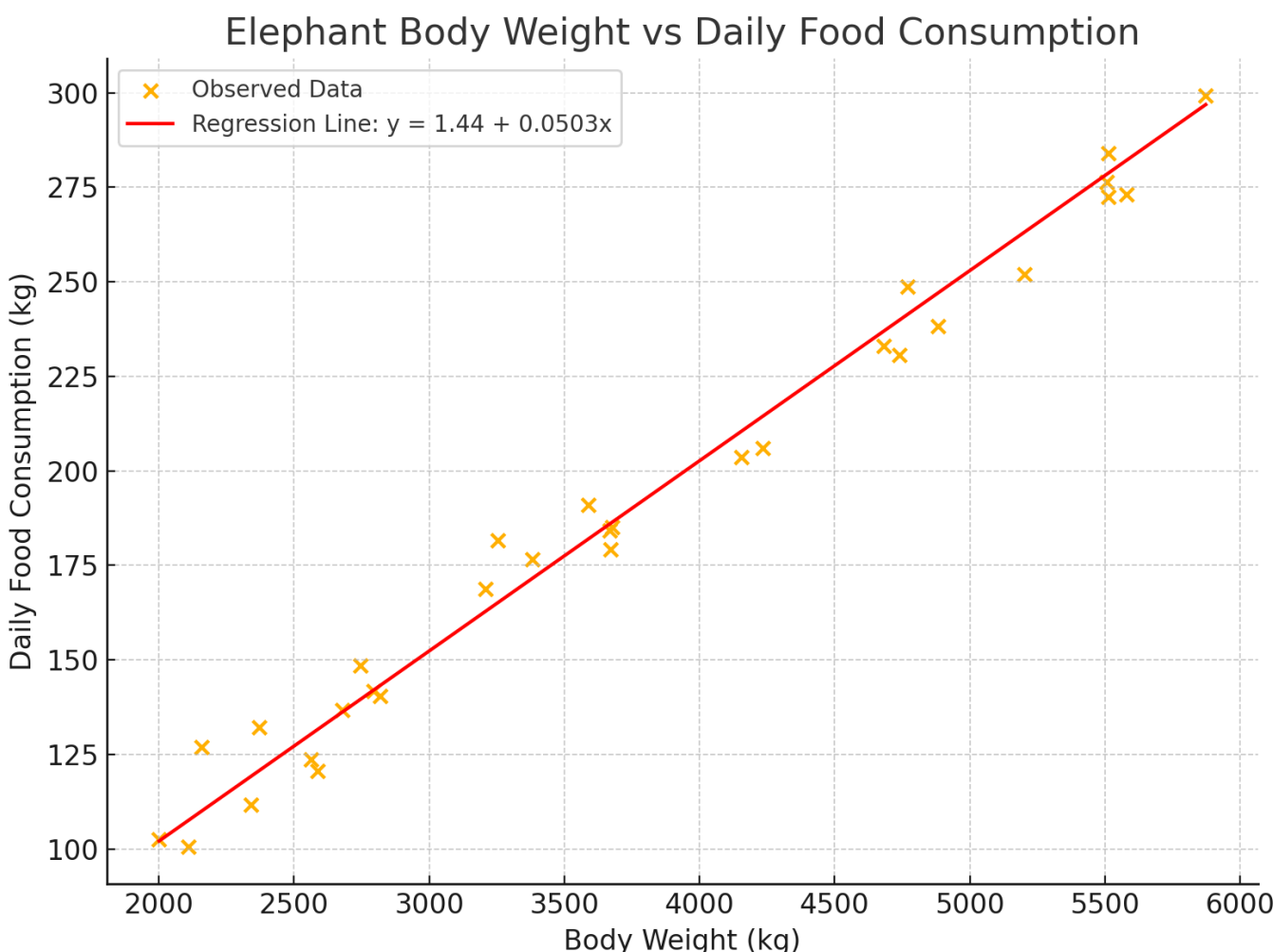
Follow-up Prompt: Can you have the table use data points that aren’t just multiples of 5?

	Forest	Grassland	Total
Deer	20	30	50
Rabbits	25	25	50
Total	45	55	100

	Forest	Grassland	Total
Deer	19	32	51
Rabbits	26	23	49
Total	45	55	100

## Other Benefits: Create Visualizations

- Topic: Types of outliers in regression
- Prompt: A zoologist is studying how much food is consumed per day (kg) by their elephants of various body weights (kg). They calculate the regression equation:  $\text{Daily Consumption} = 1.44 + 0.0503 \times (\text{Body Weight})$ , which is shown in the scatter plot. Create this scatter plot with the regression line.



# Next Steps

- Create a comprehensive question bank for the Stat 302 course, with exercises from all six fields present in each chapter.
- Gather student feedback to determine how relevant they feel the course is to their majors.
- Evaluate ChatGPT’s (and other LLM’s) ability to create exercises corresponding to the six hierarchical levels of learning.