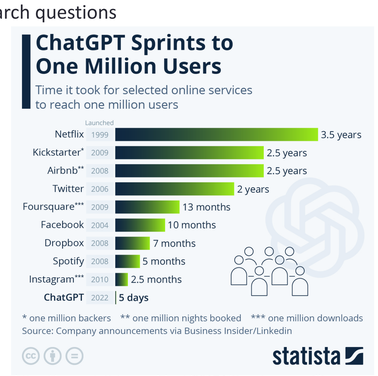
# APPLIED STATISTICS WITH CHATGPT: BOONS AND BANES

John R. Stevens and Maha Moussa

## BACKGROUND

- AI chatbots use machine learning and natural language processing to respond to requests in a conversational user interface
- A currently popular option (publicly launched Nov 2022) is ChatGPT
  - GPT = "generative pre-trained transformer", to generate human-like language responses
- Other AI chatbots: Google Bard, Microsoft Bing AI, IBM Watson Assistant, Jasper, Perplexity, YouChat, GitHub Copilot, Amazon CodeWhisperer; a newer/full version ($) of ChatGPT
- Consider: how can applied statisticians use these tools to
  - Collaborate with researchers in various fields
  - Approach statistical research questions
  - Train students




https://www.statista.com/chart/29174/time-to-one-million-users/

## EX. 1: POWER ANALYSIS IN R

**R code and general objective given to ChatGPT:**

```
> grp <- rep(c(1:4),each=15)
> y <- c(2,3,3,2,5,1,6,3,2,1,6,2,2,5,5,
+        7,8,9,6,5,8,9,7,9,7,6,9,2,1,9,
+        1,5,3,2,3,5,3,2,2,5,6,3,2,1,2,
+        5,3,2,1,3,2,1,5,3,2,5,3,2,1,2)
```

These data are from a pilot study, and I want to run a larger experiment. What sample size will I need if I want to have 80% power to detect a difference of at least 1.0 between groups 3 and 4?

**ChatGPT generated R code and output**

- [Initially] Assuming normal data
- [When prompted] For non-normal data
- [When prompted] In a Bayesian framework

ChatGPT's predicted R output was **grossly incorrect**

**R code was deeply flawed** and referred to a phantom function for which ChatGPT nevertheless showed output (but finally acknowledged function wasn't available)

**When asked for help choosing** between the parametric, non-parametric, and Bayesian approaches, ChatGPT:

- Was appropriately **cautious and thoughtful**
- **Discouraged me from bowing to pressure** from my advisor
- Encouraged me to **consult with a professional statistician**

## EX. 2: REPEATED MEASURES IN SAS

| Trt | | 1: S+C | | | 2: S+W | | | 3: G+C | | | 4: G+W | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Barrel | | 1 | … | 4 | 5 | … | 8 | 9 | … | 12 | 13 | … | 16 |
| Day | 1 | | | | | | | | | | | | |
| | … | | | | | | | | | | | | |
| | 14 | | | | | | | | | | | | |

**Study description given to ChatGPT:**

Excessive phosphorous in liquid animal waste is often a concern for regions surrounding large animal operations such as dairy farms. A study is conducted with research questions (i) "Can steel slag (a byproduct of steel production) be used to remove phosphorous from liquid animal waste?" and (ii) "Is the phosphorous-removal property of steel slag affected by temperature?" Four treatments are considered, based on a combination of steel slag presence or absence, and warm or cold temperature. When steel slag is absent, inert gravel is used. The four treatments can be summarized as follows: 1: steel slag, cold temperature; 2: steel slag, warm temperature; 3: gravel, cold temperature; 4: gravel, warm temperature. For each treatment, the steel slag or gravel is put into a large barrel, and the corresponding temperature is constantly applied to the barrel. Four replicates of each treatment are used, for a total of sixteen barrels randomly assigned to the four treatments. A hose is run to the top of each barrel, and through each hose the same chemical mixture is passed slowly into the barrel, with a constant (and known) phosphorous concentration. The mixture runs slowly down through each barrel and is allowed to pass freely through the bottom of the barrel into an outlet hose. Each barrel's outlet hose is tapped every 24 hours for two weeks, beginning at the time the mixture first flows into the barrel. Each time the hose is tapped, a small sample of outlet mixture is collected in a vial, and the vial's phosphorous concentration is measured.

**Based on this description, ChatGPT correctly identified:**

- study goals
- sample size
- potential limitations of the study
- experimental units
- Trt x Day as a fixed term

**ChatGPT struggled with:**

- Study design identification:

(prompted) (prompted)
CRD → split-plot → repeated measures

finally chose repeated measures as best, with explanation

When asked what conclusions could be different if I analyzed data as a split-plot instead of as a repeated measures design, **ChatGPT gave a correct and insightful answer**

When asked to generate a sample data set and SAS code illustrating different conclusions, **ChatGPT generated code**, and claimed to report significance test results and interaction plot code (but not results).

- Consistency of suggested code and identified design

Code didn't run correctly.

**ChatGPT can't run code**, and so kept coming up with proposed coding solutions that didn't work (and that also **misrepresented the design, despite my repeated corrections**).

ChatGPT acknowledged its limitations on generating code and **could not guarantee its correctness**.

## GENERAL (INITIAL) CONCLUSIONS

**ChatGPT is like an over-eager child**

- Energetic but in need of structure
- Frequently and enthusiastically wrong
- Responds well to feedback and will admit errors
  (but often repeats those same errors, like it has no developed character or soul)

Image by pressfoto on Freepik

**ChatGPT: miscellaneous boons and banes (pros and cons)**

- Can generate potential code, but can NOT run code – it only tries to predict the output (which is frequently gibberish)
- Can help find public data online (up to its current knowledge cutoff of September 2021)
- Can help statistical researchers discover alternative approaches
- To make it useful, you need to ask the right questions and don't automatically trust everything it gives you – it's like Wikipedia 20+ years ago, not a bad starting place, but definitely a bad stopping place: ask, but verify!

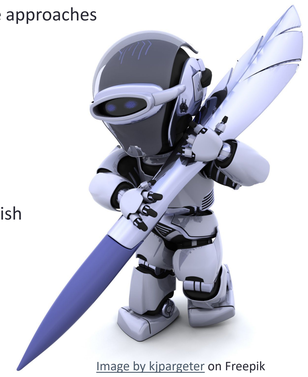


Image by kjpargeter on Freepik

- After some coaching, ChatGPT did learn to distinguish experimental designs (but would get hung up on split plot vs repeated measures), and even then it still tried to give incorrect SAS code when trying to debug – it didn't seem to care about the history within the conversation
- Could potentially mislead researchers and students with statistical recommendations, even if it can also give insight and thoughtful perspectives, and when pressed it will admit its limitations
- Could potentially cause researchers and students to (unknowingly) plagiarize – users "own" the output from ChatGPT, but that output may be lifted from other sources (which ChatGPT could sometimes identify when specifically asked)
- Lacks true creativity
- None of these conclusions are super surprising, but students and collaborators may not have the same statistical maturity and perspective to know how to verify and question ChatGPT results (especially with coding and references)

## CHATGPT AVAILABILITY

Free account at http://chat.openai.com