AI Innovation Project: Evaluating ChatGPT's Effectiveness

Project Overview

This project investigates how effectively **ChatGPT** supports statistical learning and data analysis tasks, especially in:

- Understanding theoretical concepts.
- Writing and explaining R codes.
- Assisting with imputation and model fitting.
- Clarifying certain statistical concepts.
- Identifying when user guidance or human judgment is essential.

Representative Examples

Example 1: MLE in a Multinomial Model

Setup:

We observe counts from a multinomial distribution:

$$(x_1, x_2, x_3) = (5, 2, 7), \quad P = \left(\frac{\theta}{2}, \frac{\theta}{2}, 1-\theta\right), \quad \theta \in (0, 1)$$

Goal: Find the Maximum Likelihood Estimator (MLE) of θ

ChatGPT's Output:

- Correctly formulates the **likelihood function**:

$$L(\theta) \propto \left(\frac{\theta}{2}\right)^7 (1-\theta)^7$$

- Derives the **log-likelihood**:

$$\ell(\theta) = 7\log(\theta) + 7\log(1-\theta) + \text{const}$$

- Takes the derivative and solves:

$$\frac{d\ell}{d\theta} = \frac{7}{\theta} - \frac{7}{1-\theta} = 0 \quad \Rightarrow \hat{\theta} = \frac{1}{2}$$

*ChatGPT handles symbolic differentiation and verification **efficiently**.

Example 2: Conditional Expectation of Indicator in Poisson Model

Setup:

Let $X_1, X_2, \ldots, X_n \sim \text{Poisson}(\lambda)$ be independent.
Define the total count:

$$T = \sum_{i=1}^{n} X_i$$

Suppose, we are interested in computing:

$$\mathbb{E}[I(X_1 = 0) \mid T = t]$$

ChatGPT's Derivation:

Applies the **definition of conditional probability**:

$$\mathbb{E}[I(X_1 = 0) \mid T = t] = \frac{P(X_1 = 0,\ \sum_{i=2}^{n} X_i = t)}{P(T = t)}$$

Using **independence and the convolution property** of Poisson random variables:

- $X_1 \sim \text{Poisson}(\lambda)$
- $\sum_{i=2}^{n} X_i \sim \text{Poisson}((n-1)\lambda)$
- $T \sim \text{Poisson}(n\lambda)$

So:

$$\mathbb{E}[I(X_1 = 0) \mid T = t] = \frac{e^{-\lambda} \cdot \frac{((n-1)\lambda)^t}{t!}}{e^{-n\lambda} \cdot \frac{(n\lambda)^t}{t!}} = \left(\frac{n-1}{n}\right)^t$$

ChatGPT recognizes and correctly simplifies this using ***known distributions*** *and conditional probability identities.*

While the previous examples demonstrated how ChatGPT handles **symbolic manipulation** and **conditional expectations** with ease, not all problems lend themselves to such direct computation.

We now turn to a **deeper inferential questions** that test estimation theory:

Can We Find an Unbiased Estimator for $\frac{1}{\lambda}$ in the Poisson Model?

Background: What Is an Unbiased Estimator?

Let $X \sim f(x \mid \theta)$ be a random variable whose distribution depends on a parameter θ. A statistic $\hat{\theta}(X)$ is said to be an **unbiased estimator** of a function $\tau(\theta)$ if:

$$\mathbb{E}_\theta[\hat{\theta}(X)] = \tau(\theta) \quad \text{for all } \theta$$

Unbiasedness is a **desirable property** in estimation because it ensures, on average, that the estimator neither overestimates nor underestimates the target.

Problem Setup

Let $X \sim \text{Poisson}(\lambda)$, and suppose we only observe a **single value of X**.

We ask:
> *Can we find a function $g(X)$ such that $\mathbb{E}[g(X)] = \frac{1}{\lambda}$?*

This means we are looking for an **unbiased estimator of $\frac{1}{\lambda}$** based on a single X

ChatGPT's Strategy (Inefficient)

ChatGPT's attempted solution goes as follows:

- Guesses a function:

$$g(x) = \frac{1}{x} \cdot \mathbf{1}_{\{x \geq 1\}}, \quad g(0) = 0$$

- Computes the expectation:

$$\mathbb{E}[g(X)] = \sum_{x=1}^{\infty} \frac{1}{x} \cdot \frac{e^{-\lambda} \lambda^x}{x!}$$

- Concludes that this sum does **not** simplify to $\frac{1}{\lambda}$

Limitation:
This is a **trial-and-error approach** and does not prove whether any such function *g can or cannot* exist. It only shows that this particular guess fails.

Mathematically Rigorous Approach (User-Suggested)

Let's assume a **function** $g(x)$ exists such that:

$$\mathbb{E}[g(X)] = \frac{1}{\lambda}$$

◆ Step 1: Unbiasedness Condition

We start from the definition of unbiasedness:

$$\sum_{x=0}^{\infty} g(x) \cdot \frac{e^{-\lambda} \lambda^x}{x!} = \frac{1}{\lambda}$$

Multiply both sides by e^{λ}:

$$\sum_{x=0}^{\infty} \frac{g(x)}{x!} \lambda^x = \frac{e^{\lambda}}{\lambda}$$

This means the **left-hand side** is a power series in λ, which defines an **entire function** (analytic for all $\lambda \in \mathbb{R}$).

◆ Step 2: Expand the Right-Hand Side

We now expand the right-hand side using known series:

$$\frac{e^\lambda}{\lambda} = \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{x!}$$

Let $y = x - 1$. Then:

$$\frac{e^\lambda}{\lambda} = \sum_{y=0}^{\infty} \frac{\lambda^y}{(y+1)!}$$

This can be further simplified as:

$$\sum_{y=0}^{\infty} \frac{1}{y+1} \cdot \frac{\lambda^y}{y!}$$

So now we compare:

$$\sum_{x=0}^{\infty} \frac{g(x)}{x!}\lambda^x \quad \text{and} \quad \sum_{y=0}^{\infty} \frac{1}{y+1} \cdot \frac{\lambda^y}{y!}$$

Matching the coefficients of λ^x, we must have:

$$\frac{g(x)}{x!} = \frac{1}{x+1} \cdot \frac{1}{x!} \Rightarrow g(x) = \frac{1}{x+1}$$

This leads to the function:

$$g(x) = \frac{1}{x+1}$$

There is **no function** $g(x)$ such that:

$$\mathbb{E}[g(X)] = \frac{1}{\lambda}, \quad X \sim \text{Poisson}(\lambda)$$

This contradiction **proves the nonexistence of an unbiased estimator** for $\frac{1}{\lambda}$ in the Poisson model.

Orthogonality Characterization and UMVUE Existence

Background: What Is a UMVUE?

An estimator $\hat{\theta}(X)$ of a parameter θ is said to be:

- **Unbiased** if $\mathbb{E}_\theta[\hat{\theta}(X)] = \theta$
- **Uniformly Minimum Variance Unbiased Estimator (UMVUE)** if it is unbiased and has the **lowest possible variance** among all unbiased estimators for every value of θ

What Is Orthogonality Characterization?

The **orthogonality condition** establishes necessary and sufficient condition for an unbiased estimator to be qualified as the UMVUE.

An unbiased estimator $\hat{\theta}(X)$ is the UMVUE **if and only if** it is **uncorrelated** with every **unbiased estimator of 0**

How Orthogonality Helps

To test if an unbiased estimator $\hat{\theta}$ is UMVUE, we must:

1. **Construct an unbiased estimator of 0**, say U
2. Check whether:

$$\mathbb{E}_\theta[\hat{\theta} \cdot U] = 0 \quad (\text{for all } \theta)$$

   If this **fails**, then $\hat{\theta}$ is **not** the UMVUE.

This is the **orthogonality characterization** approach.

ChatGPT’s Limitation

When asked:

> *"Does the UMVUE of θ exist for $X_i \sim \mathrm{Uniform}(\theta, \theta+1)$?"*

ChatGPT states that:

- $T = (X_{(1)}, X_{(n)})$ is sufficient but **not complete**
- The Lehmann–Scheffé theorem therefore **does not apply**

However, **ChatGPT does not know to check whether a nonzero unbiased estimator of 0 exists**. That is:

- It **does not explore contradiction-based reasoning**, where:
  - If one **assumes** a UMVUE exists,
  - Then applies orthogonality characterization,
  - And ends up with a **logical contradiction**, implying **no such UMVUE exists**

This type of proof is subtle and typically requires **human insight or manual derivation**.

Summary Takeaway

Concept	Summary
UMVUE	Unbiased estimator with minimum variance for all θ
Lehmann–Scheffé	Guarantees uniqueness if complete sufficient statistic exists
Orthogonality Characterization	established necessary and sufficient condition for the UMVUE
ChatGPT's role	Good at basic identification, not at deep counterexamples/ deep logical derivation

Handling Rigorous Proofs – Karlin–Rubin Theorem

The **Karlin–Rubin Theorem** is a landmark result in the theory of hypothesis testing. It identifies conditions under which a **one-sided hypothesis test** is **uniformly most powerful (UMP)** — meaning no other test of the same size (type I error rate) has greater power **at all alternative parameter values**.

Understanding the theorem requires familiarity with key concepts from **exponential family theory**, **sufficiency**, and **monotone likelihood ratios (MLR)**. These are central in **estimation theory**

What Is a UMP Test?

In hypothesis testing, a **Uniformly Most Powerful (UMP) test** is one that **maximizes power** (probability of correctly rejecting the null hypothesis) **for every value of the alternative**, while **maintaining the desired type I error rate** under the null hypothesis.

Formally, a test ϕ is UMP of size α for testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$, if:

- $\mathbb{E}_\theta[\phi(X)] \leq \alpha$ for all $\theta \leq \theta_0$
- $\mathbb{E}_\theta[\phi(X)] \geq \mathbb{E}_\theta[\psi(X)]$ for **any other test** ψ of size α, and all $\theta > \theta_0$.

Karlin–Rubin theorem provides a **constructive method** to build them under certain regularity conditions.

The Karlin–Rubin Theorem

Let $X \sim f(x;\theta)$, where the density belongs to a **one-parameter exponential family**:

$$f(x;\theta) = h(x)\exp[\eta(\theta)T(x) - A(\theta)]$$

Assume:

- $T(X)$ is a **sufficient statistic** for θ (it retains all information about the parameter that is present in the data)
- The family has a **monotone likelihood ratio (MLR)** in $T(x)$

Then, for testing:

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0,$$

the test of the form:

$$\phi(x) = \begin{cases} 1, & T(x) > c \\ \gamma, & T(x) = c \\ 0, & T(x) < c \end{cases}$$

is **Uniformly Most Powerful (UMP)** of level α.

ChatGPT is generally good at restating this result

ChatGPT Limitation: Proof-Level Detail

Despite knowing the theorem's conditions and conclusion, ChatGPT often **misses key analytic steps** required for a full proof. In particular, it skips the step involving **distribution comparison and derivative analysis** — essential for demonstrating **power monotonicity** and **type I error control**.

Crucial Missing Step: Comparing Distribution Functions

Let

$$G(t) = F_T(t \mid \theta_1) - F_T(t \mid \theta_0), \quad \text{with } \theta_1 > \theta_0$$

Differentiating gives:

$$G'(t) = f_T(t \mid \theta_1) - f_T(t \mid \theta_0) = f_T(t \mid \theta_0)\left(\frac{f_T(t \mid \theta_1)}{f_T(t \mid \theta_0)} - 1\right)$$

- If the likelihood ratio $\frac{f_T(t|\theta_1)}{f_T(t|\theta_0)}$ is **increasing in** t (MLR condition),
- Then $G'(t)$ changes sign **at most once**, from **negative to positive**

Also, both $F_T(t|\theta)$ approach 0 as $t \to -\infty$, and 1 as $t \to \infty$, so:

$$\lim_{t \to \pm\infty} G(t) = 0 \Rightarrow G(t) < 0$$

Which implies:

$$F_T(t \mid \theta_1) < F_T(t \mid \theta_0) \quad \Rightarrow \quad P(T > t \mid \theta_1) > P(T > t \mid \theta_0)$$

This shows **type I error rate** is controlled for all null parameter points at alpha level.

Summary: ChatGPT vs Human Derivation

While ChatGPT can accurately restate the Karlin–Rubin theorem and provide a basic intuitive explanation, it often struggles with delivering full analytic proofs unless prompted step-by-step.

Rigorous statistical theorems like Karlin–Rubin require careful logical reasoning, detailed logical connecting arguments; areas where human expertise remains crucial. ChatGPT typically misses the deeper structure, such as proving power monotonicity or verifying Type I error control for all null values.

🔄 Transitioning from Theoretical Foundations to Real-World Data

We now shift our focus to a **practical, real-world application**: examining and handling **missing data in the NHANES dataset**. This case study tests ChatGPT's capabilities in **data wrangling, visualization, EDA, missingness classification, and survey logic interpretation**

NHANES Data Analysis

Background and Study Objective

This is survey data collected by the US National Center for Health Statistics (NCHS) which has conducted a series of health and nutrition surveys since the early 1960's. Since 1999 approximately 5,000 individuals of all ages are interviewed in their homes every year and complete the health examination component of the survey. The health examination is conducted in a mobile examination centre (MEC).

Variable Domains and Types

The dataset contains a broad range of variables grouped into meaningful categories:

- **Demographics**:
  `Age` , `Gender` , `Race3` , `Education` , `HHIncomeMid` , `Poverty` , `HomeOwn` , `Work`
- **Physical Measurements**:
  `BMI` , `BPSysAve` , `BPDiaAve` , `Height` , `Weight`
- **Health Variables**:
  `Diabetes` , `HealthGen` , `TotChol` , `Depressed` , `SleepTrouble` , `DirectChol`
- **Lifestyle Variables**:
  `PhysActive` , `Alcohol12PlusYr` , `SmokeNow` , `Smoke100` , `HardDrugs` , `Marijuana`

Selected Variable Descriptions

Variable	Description
`Age`	Age in years at the time of screening (80+ grouped as 80)
`Gender`	Biological sex: Male or Female
`Race3`	Race category: Mexican, Hispanic, White, Black, Asian, Other
`Education`	Highest level of education attained for age ≥ 20
`HHIncomeMid`	Midpoint of income category
`HomeOwn`	Indicates whether participant owns, rents, or has another arrangement
`Work`	Employment status
`BMI`	Body Mass Index (kg/m²), calculated from height and weight
`BPSysAve`	**Average systolic blood pressure**, across multiple measurements
`BPDiaAve`	**Average diastolic blood pressure**, across multiple measurements
`HealthGen`	Self-reported general health (Excellent → Poor)
`Diabetes`	Participant told by doctor they have diabetes (Yes/No)
`TotChol`	Total cholesterol level in mmol/L
`PhysActive`	Whether participant engages in moderate/vigorous activity (Yes/No)
`Alcohol12PlusYr`	Has consumed ≥12 alcoholic drinks in a year (Yes/No)
`Smoke100`	Smoked ≥100 cigarettes in lifetime (Yes/No)
`SmokeNow`	**Currently smokes** (Yes/No), but only asked if `Smoke100` = Yes

Variable	Description
`HardDrugs`	Has tried heroin, cocaine, meth, etc. (Yes/No)

1. Understanding Missing Data Mechanisms

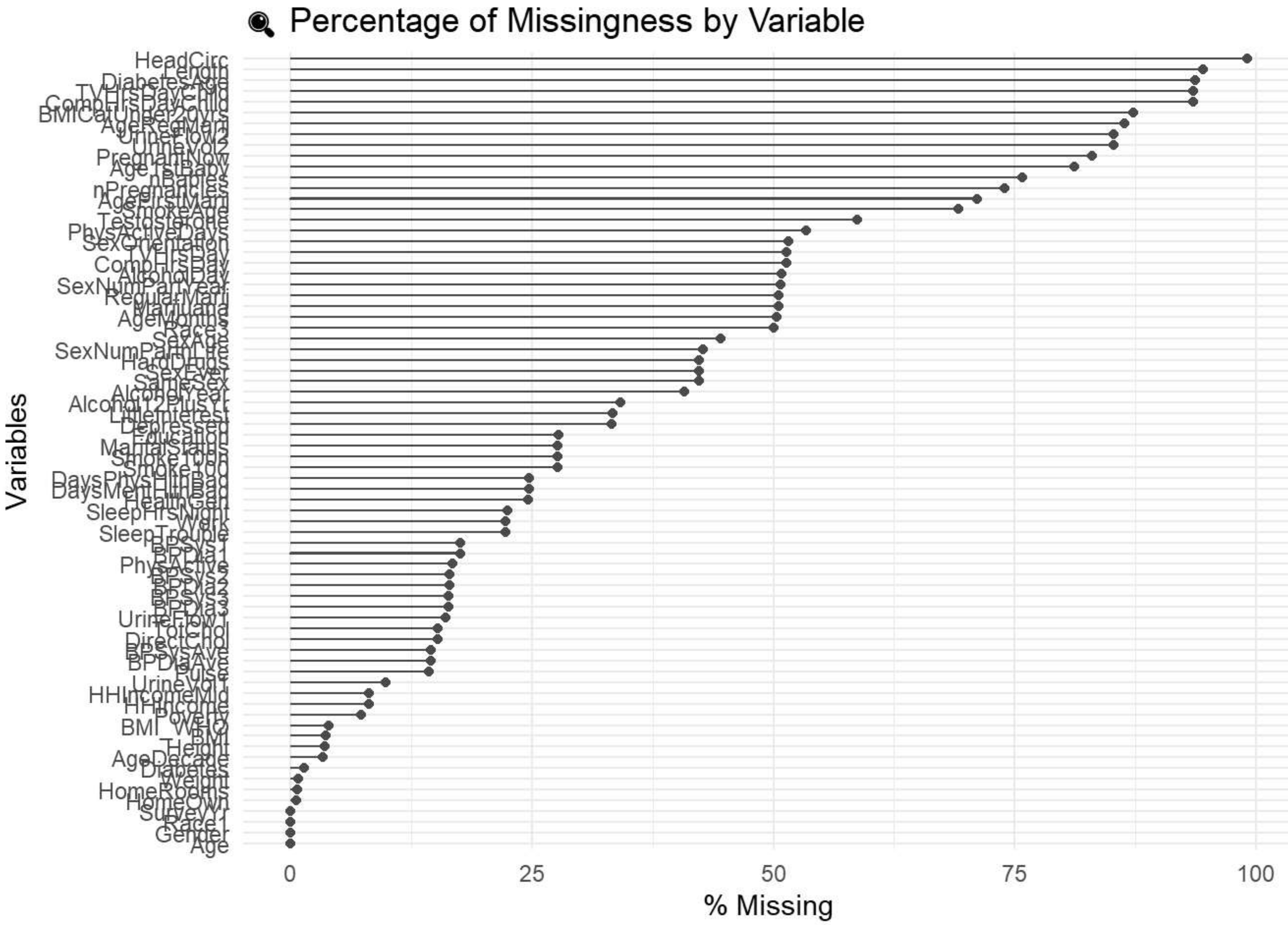
The three classical missing data mechanisms are:

Mechanism	Definition	Example
MCAR (Missing Completely at Random)	Missingness is independent of both observed and unobserved data	Data lost due to random equipment failure
MAR (Missing At Random)	Missingness depends only on observed data	Blood test missing only for older individuals
NMAR (Not Missing At Random)	Missingness depends on the unobserved (missing) values themselves	People with high income choosing not to report it

ChatGPT is good at summarizing these definitions, providing theoretical explanations with appropriate examples.

However, *ChatGPT struggles to correctly classify real-world missingness from raw data without explicit user context.*

2. Exploring Missing Data pattern in NHANES's variables



```r
# 📋 Table: Top Variables with Missingness (Excluding Smoke100)
missing_summary <- sapply(nhanes_data[, names(nhanes_data) != "Smoke100"], function(x) sum(is.na(x)))
missing_perc <- round(100 * missing_summary / nrow(nhanes_data), 2)

missing_table <- data.frame(
  Variable = names(missing_summary),
  Missing_Count = missing_summary,
  Missing_Percent = missing_perc
)

kable(missing_table %>%
        arrange(desc(Missing_Percent)) %>%
        head(8),
      caption = "Top 8 Variables with Missingness in the NHANES Subset (Excluding Smoke100)",
      col.names = c("Variable", "Missing Count", "Missing (%)"),
      format = "html") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE)
```

Top 8 Variables with Missingness in the NHANES Subset (Excluding Smoke100)

	Variable	Missing Count	Missing (%)
HeadCirc	HeadCirc	9912	99.12
Length	Length	9457	94.57
DiabetesAge	DiabetesAge	9371	93.71
TVHrsDayChild	TVHrsDayChild	9347	93.47
CompHrsDayChild	CompHrsDayChild	9347	93.47
BMICatUnder20yrs	BMICatUnder20yrs	8726	87.26
AgeRegMarij	AgeRegMarij	8634	86.34
UrineFlow2	UrineFlow2	8524	85.24

These are examples where **ChatGPT performs very well** — generating reproducible R code for exploratory data analysis (EDA), including:

- **Computing missing counts and percentages**
- **Creating diagnostic plots** (`gg_miss_var()` , `md.pattern()` , `aggr()` , etc.)
- **Automating common EDA tasks** with tidyverse syntax

Such **purely computational steps**; summarizing structure, and visualizations,are ideal for ChatGPT, especially when the user provides the dataset and goals clearly.

3. Overview of Missingness

`SmokeNow` and `Smoke100` have notable missingness. In particular, `SmokeNow` contains many NAs because its value depends on whether the respondent has ever smoked 100 cigarettes (i.e., `Smoke100 == "Yes"`). For those who answered “No” to `Smoke100` , `SmokeNow` is not applicable; thus, the NAs are **structural**, not missing at random.

Combined systolic blood pressure (`BPSysAve`), `HardDrugs` , `HealthGen` , `Alcohol12PlusYr` , `TotChol` , `Race3` , `Education` , `SleepTrouble` , `Depressed` , and `Work` variables also have substantial missingness.

Several variables, such as `HomeOwn` , `HHIncomeMid` , `Diabetes` , and `BMI` , have relatively few missing values.

`SmokeNow` is a special case: its missingness is **not at random** because it is **conditional** on the value of `Smoke100` .

ChatGPT’s Limitation

On its own, ChatGPT:

- Fails to ask whether a skip pattern exists.
- May misclassify the missingness as MAR or MCAR.
- Needs prompting to check conditional skip logic.

Critique of ChatGPT’s EDA on NHANES

While ChatGPT generates **visually clean and syntactically correct R code** for exploratory data analysis, several **key issues** arise in the context of real-world datasets like **NHANES**.

Issue 1: Incorrect Variable Names

ChatGPT frequently references incorrect or non-existent variable names when performing EDA. For example, it suggested:

```
select(age, bmi, systolic_bp, diastolic_bp)
```

However, in the actual **NHANES** dataset:

- `systolic_bp` and `diastolic_bp` **do not exist**
- The correct variables are:
  - `BPSysAve` = *average systolic blood pressure*
  - `BPDiaAve` = *average diastolic blood pressure*

Issue 2: Lack of variables Screening or variables type checking

ChatGPT does **not perform variable validation** before plotting. It may attempt to:

- Include non-existent variables like `systolic_bp`, `diastolic_bp`.
- Plot categorical variables without checking if they're factors.
- Generate visualizations without confirming data compatibility.

This leads to **code errors** or misleading visualizations if used without manual checks.

Interpretation vs Computational Gap

When prompted with conceptual questions such as:

> "Is it fair to include missing education levels as a separate category in boxplots?"

ChatGPT gives a **generic** answer:

> "Yes, it can make sense and be interpretable."

However, it fails to:

- Address whether the **missingness is informative**
- Give interpretive diagnostics unless specified to do so.

Key Weakness

ChatGPT excels at generating **clean, syntactically correct code** and visual outputs.
However, it lacks the ability to:

- **Diagnose** the nature of missingness
- **Infer** underlying structure or logic
- Make **statistical judgments** without explicit user input

This underscores the need for **human oversight** in real-world data analysis.