

Statistical Consulting: Data, Questions, Analysis

<https://gemini.google.com/app/606071c0ef01e7da>

User prompt: How would I go about presenting multiple facets of a statistical consulting question to you: a dataset or spreadsheet; a sketch of research question and analysis hoped for (as a default); and a prompt or question from the consultant (me) about appropriate analyses, assumptions, etc. I am asking primarily how this information can be passed to you, eg in the form of uploads, attachments, or whatever.

Response: That is a fantastic way to approach a statistical consulting problem. Breaking down your query into those three specific components—the data, the research goals, and your specific technical questions—is exactly how a professional collaboration works.

Here is a practical guide on how you can best pass each of these elements to me in this interface, along with some tips to get the most out of our exchange.

1. The Dataset or Spreadsheet

Because we are interacting through a chat interface, you cannot directly upload a live, interactive Excel or .csv file. However, you can easily share your data in several highly effective ways:

- **For Small Datasets (Markdown Table):** If your data is small (under 30 rows), you can copy and paste it directly. I can read standard markdown tables, or even raw, tab-delimited text copied straight from Excel.
- **For Medium/Large Datasets (The "Head" and Structure):** You don't need to show me all 10,000 rows. Instead, paste:
 - The **first 5 to 10 rows** (to show the formatting).
 - A **variable dictionary** or list of columns with their data types (e.g., Age [Continuous], Treatment_Group [Categorical: A, B, Control], Post_Test_Score [Discrete, 1-100]).
 - The **sample size (N)** and any known missing data patterns.
- **For Complex Structures (Simulated Data):** If your data is proprietary or highly sensitive, you can write out a tiny, simulated "dummy" dataset with 5 rows that mirrors your actual data structure.

2. Sketch of Research Questions & Intended Analysis

This is the "map" of your study. To pass this to me, simply write a quick narrative or bulleted list outlining:

- **The Big Picture:** What is the real-world domain (e.g., clinical trial, marketing A/B test, ecological survey) and the ultimate goal?
- **The Variables' Roles:** Explicitly state your independent variables (predictors), dependent variables

(outcomes), and any covariates or confounding factors you need to control for.

- **The Study Design:** Is it a between-subjects design, within-subjects (repeated measures), longitudinal, or a randomized block design?
- **Your Default Plan:** If you have a starting hypothesis (e.g., *"I was planning on running a mixed-effects ANOVA"*), share it! It gives us a great baseline to evaluate.

3. Your Prompt/Questions (The Consultant's Query)

This is where you tell me exactly where you are feeling stuck or where you want a second opinion. You can frame your questions around:

- **Methodology Choice:** *"Given my sample size and non-normal distribution, is an ANOVA robust enough, or should I pivot to a non-parametric alternative like Kruskal-Wallis?"*
- **Assumption Checking:** *"What diagnostic plots or statistical tests should I run in R/Python to verify homoscedasticity for this specific model?"*
- **Coding & Implementation:** *"Can you write the R code using lme4 to specify this random-intercepts model, and explain how to interpret the output?"*
- **Power & Sample Size:** *"How do I justify my sample size post-hoc, or calculate the required sample size for a future iteration of this study?"*

An Example of How to Package It:

1. The Data Structure: I have a dataset of $N=120$ patients. Columns: ID (nominal), Group (Treatment vs. Placebo), Pre_Score (continuous, 0-50), Post_Score (continuous, 0-50).

2. Research Question: Does the Treatment group show a significantly greater improvement in scores compared to the Placebo group, controlling for baseline (Pre_Score)?

3. Consultant's Question: Should I run a repeated-measures ANOVA, or is an ANCOVA predicting Post_Score using Pre_Score as a covariate statistically superior here? Can you explain the trade-offs regarding statistical power?

Whenever you are ready, feel free to lay out your scenario using this framework, and we can dive right in!

*User prompt: If I included a dataset in *csv format in a portion of the input question/prompt, would you run one or more practice analyses as part of answering some of these consulting questions ?*

Response: Yes, absolutely! If you paste a dataset in CSV format (even a subset or a small representative sample) directly into your prompt, I can use an active Python environment to run real, practical analyses on it.

I can write, execute, and verify the Python code in real-time to give you exact statistical outputs. Here is what I can do with your pasted CSV data:

1. Exploratory Data Analysis (EDA) & Summary Stats

I can parse your CSV into a pandas DataFrame to calculate:

- Descriptive statistics (means, medians, standard deviations, range).
- Missing value checks and data distribution metrics (skewness, kurtosis).
- Frequency tables for categorical variables.

2. Assumption Checking

Before recommending a model, we can test the data's properties:

- **Normality:** Run Shapiro-Wilk or Kolmogorov-Smirnov tests.
- **Homoscedasticity:** Run Levene's or Breusch-Pagan tests.
- **Multicollinearity:** Calculate Variance Inflation Factors (VIF) or correlation matrices.

3. Running the Actual Models (Practice Analyses)

I can run the actual statistical models you are considering and output the regression tables, ANOVA summaries, or test statistics. For example:

- **Hypothesis Testing:** Run *t*-tests, Chi-Square tests, or ANOVA.
- **Regression Modeling:** Fit Ordinary Least Squares (OLS) linear regressions, logistic regressions, or mixed-effects models.
- **Non-Parametric Alternatives:** Run Wilcoxon rank-sum or Kruskal-Wallis tests if assumptions are violated.

4. Data Visualization

I can generate statistical plots (like box plots, Q-Q plots, residual plots, or scatter plots with trend lines), save them, and display them directly in our chat so you can visually inspect the distributions and model fit.

How to format it for the best results:

To make sure the Python environment reads your data perfectly, just format your CSV using a Markdown code block like this:

```
SubjectID,Group,PreTest,PostTest
1,Treatment,24,35
2,Treatment,28,42
3,Control,22,23
4,Control,25,27
...
```

Whenever you have a dataset ready—whether it's your actual raw data (if small) or a representative dummy dataset—just paste it in, tell me what analysis you are leaning toward, and we can run a full test drive together!
