Statistics Project: Inference on Flight Performance

November 17, 2025

1 Context and Dataset

This project focuses on the statistical analysis of simulated operational data from Airbus. The dataset `Airbus_simu.csv` contains several flights with the following fields:

Variable	Description	Unit
`index`	Row index	–
`callsign`	Flight callsign	–
`flight_time`	Total flight time	seconds
`distance`	Distance flown	meters
`fuel`	Fuel consumed	kilograms

Each observation corresponds to a unique flight. Some flights may appear several times in the raw file, and duplicates must be removed. New variables should be derived for the analysis:

$$\begin{aligned}
\text{Flight time (h)} &= \frac{\text{flight_time (s)}}{3600}, \\
\text{Distance (km)} &= \frac{\text{distance (m)}}{1000}, \\
\text{Average speed (km/h)} &= \frac{\text{Distance (km)}}{\text{Flight time (h)}}, \\
\text{Fuel rate (kg/km)} &= \frac{\text{fuel (kg)}}{\text{Distance (km)}}.
\end{aligned}$$

For comparative analysis, consider two types of flights based on the flight time:

$$\text{“Short” if time} < 3 \text{ hours}, \quad \text{“Long” if time} > 6 \text{ hours}.$$

2 Project Objectives

The goal of this project is to apply the principles of estimation, confidence intervals, and hypothesis testing to aeronautical performance data. The analysis should rely on the Central Limit Theorem (CLT) or normality assumptions where applicable.

3 Preliminaries

3.1 Equality of mean test

Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be independent iid samples from two populations. The purpose of this section is to design a test with null hypothesis $H_0 : E[X_1] = E[Y_1]$ against the alternative hypothesis $H_1 : E[X_1] \neq E[Y_1]$.

1. Let S_X^2 (resp. S_Y^2) be the sample variance of the first (resp. second) sample. Assuming $n, m \geq 30$, justify that the distribution of:

$$\bar{X} - E[X_1] \text{ (resp. } \bar{Y} - E[Y_1]\text{)}$$

   is approximately $\mathcal{N}(0, V(X_1)/n)$ (resp. $\mathcal{N}(0, V(Y_1)/m)$) .

2. It can be proved that the sum of two independent random variables with respective distributions $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)$ is $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ distributed. Under the null hypothesis H_0, explain why:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

   is approximately $\mathcal{N}(0, 1)$ distributed.

3. Design a test with $1 - \alpha$ confidence level for H_0 against H_1.

3.2 Testing for normality

Do some research on "Shapiro-Wilk test" and report the result. Give your interpretation about this test. Find information about Q–Q plots and their use for normality testing.

3.3 Pearson's correlation

Find what is a linear model and what is Pearson's correlation. Summarize in a short paragraph your understanding of these notions.

4 Questions

1. **Data cleaning**: Delete data larger than 16 hours. Separate the original dataset into two datasets: short-haul and long-haul flights.

2. **Estimation of Mean Speed:** Estimate the mean cruise speed (in km/h) for short-haul and long-haul flights respectively. Construct a 95% confidence interval for each, using the Central Limit Theorem.

3. **Estimation of Fuel Efficiency:** Estimate the mean fuel consumption rate (in kg/km) and its 95% confidence interval for each of the two datasets. Comment on the precision of the estimate and its operational interpretation.

4. **Comparison of Short and Long Flights on the Speed:** Test whether there is a significant difference in average cruise speed between short and long flights:

$$H_0 : \mu_{\text{short}} = \mu_{\text{long}} \quad \text{vs.} \quad H_1 : \mu_{\text{short}} \neq \mu_{\text{long}}.$$

5. **Comparison of Short and Long Flights on the Fuel Rate:** Test whether longer flights have significantly different fuel efficiency compared with shorter ones:

$$H_0 : \mu_{\text{short}} = \mu_{\text{long}} \quad \text{vs.} \quad H_1 : \mu_{\text{short}} \neq \mu_{\text{long}}.$$

6. **Correlation Analysis:** Examine whether flight distance and average fuel consumption rate (kg/km) are significantly correlated:

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_1 : \rho \neq 0.$$

   Provide the Pearson correlation coefficient and interpret its magnitude and sign.

7. **Validation of Normality:** Verify whether the variables (flight time, distance, fuel rate) follow a normal distribution using the Shapiro–Wilk test and Q–Q plots. Discuss how the Central Limit Theorem justifies the use of normal-based inference even when normality is not perfect.

5 Understand the P-Values

Compute **p-values** for all hypothesis tests and interpret them in context. For example, one can use p-values to support conclusions:

- $p < 0.05$: reject H_0, conclude evidence for the alternative.
- $p \geq 0.05$: fail to reject H_0, conclude no significant evidence.
- For two-sample t-tests (speed or fuel rate), report:

$$t \text{ statistic}, \text{ degrees of freedom (df)}, \text{ and } p\text{-value}.$$

Interpret whether the difference between short and long flights is statistically significant at the 1%, 5%, 10% level.

- For correlation analysis, report:

$$r \text{ (Pearson correlation)}, t\text{-statistic}, \text{df}, p\text{-value}.$$

Comment on the strength and significance of the correlation.

6 Deliverables

- Written report with cleaned dataset, statistical results, confidence intervals, p-values, and interpretations.
- Power BI dashboard visualizing:
  - Mean speed and fuel rate with 95% confidence intervals.
  - Comparison of short vs long flights.
  - Scatter plot of distance vs fuel consumption rate
- R script including data cleaning, derived variables, estimation, CI computation, hypothesis testing, and export to Power BI.