ISEC Lisboa

Project (2025/2026)

ESCOLA DE GESTÃO, ENGENHARIA E AERONÁUTICA

Name:	**Student Number:**
Name:	**Student Number:**
Name:	**Student Number:**
Name:	**Student Number:**
Name:	**Student Number:**

The project consists of 14 questions for a total of 20 points (100 %).

Introduction

Your group will act as data analysts for a fictional airline consulting firm. Collect real flight data from FlightRadar24 and use it to analyze flight punctuality at a Portuguese airport. Base all statistical work on your collected data, show all calculations and code(s) used to develop it.

Part I – Building the Sample

(5%) **1.** 16 flights are pre-filled in Table 1 (Appendix). Add **14 flights** from LPPT across **at least 3 different days** – exactly **7 AM and 7 PM** – to complete the 30-flight sample. Record the flight number, airline, and departure/arrival delays. Then comment briefly on the sample's representativeness and any potential bias introduced by your choice of days or airlines.

(2%) **2.** Describe briefly the **sampling method** used to collect the 14 flights your group added (e.g. convenience sampling, systematic sampling). What are the potential **limitations** of this method for drawing conclusions about all flights at this airport? In your answer, reflect on whether a different sampling strategy would have yielded more reliable conclusions, and why.

Part II – Descriptive Statistics & Point Estimation

3. Using your sample of 30 flights, calculate the following point estimators. Show the formula and the result for each one.

(2%) (a) The sample mean of departure delay $\bar{x}$ (in minutes). Does this value suggest that, on average, flights at your chosen airport tend to depart late, on time, or even early? What operational factors could explain this?

(2%) (b) The median of departure delay $\widetilde{x}$. Compare the median with the mean: are they close to each other or noticeably different? What does that difference (or lack thereof) tell you about the shape of the delay distribution at this airport?

(2%) (c) The mode of departure delay (if it exists). Comment on whether it is a useful measure for this data and what it would mean in practice if many flights shared the same delay value.

(3%) (d) The sample standard deviation s of departure delay. Interpret this value in the context of airport operations: does it indicate that delays are consistent and predictable, or highly variable? What consequences could high variability have for passengers and ground crews?

(3%) (e) The coefficient of variation $CV = \dfrac{s}{|\bar{x}|} \times 100\%$, defined only when $\bar{x} \neq 0$. What does this value indicate about the variability of delays relative to the average? Would you consider delays at this airport to be stable or unstable based on this measure? Justify your answer. **Note:** If your sample mean $\bar{x}$ is zero or very close to zero, explain why the CV is not a meaningful measure in that case and rely instead on the standard deviation alone to characterise variability.

(4%) (f) The (Pearson) correlation coefficient between departure delay and arrival delay. Comment on the sign, magnitude, and practical meaning of the result. Does it make operational sense that departure and arrival delays would be correlated in this way? Are there situations where a flight could absorb a departure delay and still arrive on time?

(2%) (g) The sample proportion $\hat{p}$ of flights considered **delayed** (departure delay > 15 min). Comment on whether this proportion seems high or low for a typical airport, and what it might imply for passenger satisfaction and airline performance metrics.

(4%) (h) Identify any **outliers** using the criterion $|x_i - \bar{x}| > 2s$. List all flights that meet this criterion. For each outlier identified, discuss whether it should be kept or removed from the analysis, justifying your decision based on what the extreme value might represent operationally (e.g. severe weather, mechanical failure, or a data recording error).

(3%) (i) Represent the collected data using charts such as histograms of departure and arrival delays. Describe what the shape of each histogram reveals about the distribution of delays. Are delays concentrated around a central value, or spread widely? Do the histograms support or contradict the numerical measures calculated in the previous parts?

Part III – Probability

4. Using your sample of 30 flights, classify each flight in a contingency table according to the group (AM/PM) and punctuality status (delayed: departure delay > 15 min).

(2%) (a) Complete the contingency table below with the absolute frequencies from your sample:

	Delayed	**On time**	**Total**
AM			15
PM			15
Total			30

(b) Using relative frequencies as probability estimates, calculate:

(2%) i. The joint probability $\hat{P}(\text{AM} \cap \text{delayed})$.

(2%) ii. The conditional probability $\hat{P}(\text{delayed} \mid \text{AM})$.

(2%) iii. The conditional probability $\hat{P}(\text{delayed} \mid \text{PM})$.

(3%) (c) Based on the conditional probabilities calculated, does the time of day (AM vs. PM) appear to influence the probability of a flight being delayed? Justify briefly. Consider also whether there are real-world operational reasons why AM and PM flights might behave differently in terms of punctuality.

5. The airport estimates that each delayed flight incurs an additional **€1500** in ground handling and gate occupation costs. Let X be the number of delayed flights in 100 flights, where $\hat{p}$ is the sample proportion obtained in Question 3g. Let $C = 1500X$ be the **total additional operational cost** (in euros) incurred by the airport.

(2%) (a) Compute the expected total additional cost $E[C]$ for the 100 flights.

(2%) (b) Compute the standard deviation $\mathrm{SD}(C)$ of the additional cost.

(2%) (c) Interpret both values. What operational decisions could the airport make based on these estimates? Comment on what the standard deviation of the cost implies for financial planning and whether relying solely on the expected value would be prudent.

Part IV – Interval Estimation

(5%) **6.** Construct a **95% confidence interval** for the population mean departure delay μ, using the t distribution with $n - 1$ degrees of freedom. Justify why you use t instead of z. After computing the interval, comment on its width: does it give the airport manager a precise or imprecise picture of the true average delay, and what would be needed to obtain a narrower interval?

(5%) **7.** Construct a **95% confidence interval** for the population variance of departure delays σ^2, using the chi-square distribution with $n-1$ degrees of freedom and critical values $\chi^2_{0.975,\,29}$ and $\chi^2_{0.025,\,29}$. Comment on the asymmetry of the interval and what the upper bound implies about the worst-case variability in delays that the airport should plan for.

(5%) **8.** Construct a **95% confidence interval** for the population proportion of delayed flights p, using $z_{0.025} = 1.96$. Comment on whether the interval includes values that would be operationally concerning for the airport, and what proportion of delays would be considered acceptable in a real airline performance context.

(5%) **9.** Interpret the three confidence intervals in plain language. What do they collectively tell a real airport manager about the reliability and punctuality of operations? Based on these intervals, would you advise the airport manager to take corrective action, and why?

Part V – Hypothesis Testing

10. Using your 30 observations, you will fit two candidate distributions to the departure delay data and determine which one is more appropriate.

(3%) (a) Test whether departure delays follow a normal distribution with mean $\mu = \bar{x}$ and standard deviation $\sigma = s$ using the Kolmogorov–Smirnov test at significance level $\alpha = 0.05$ (critical value $D_{0.05} \approx 0.242$ for $n = 30$). State the conclusion. Comment on whether the result surprises you given the shape of the histograms in Part II and what it implies about using normal-based statistical methods on this data.

(3%) (b) Discard negative observations (early departures) and zero-delay observations, and let m be the number of remaining non-negative delays. Estimate the rate parameter as $\hat{\lambda} = 1/\bar{x}^+$, where $\bar{x}^+$ is the mean of those m observations. Apply the Kolmogorov–Smirnov test to verify whether the negative exponential distribution fits these non-negative delays. Compute $D_{0.05}$, for your m. State the conclusion.

(c) Based on the results of parts (a) and (b), select the distribution that best fits your data. Using that distribution, calculate:

(1%) i. The probability that a flight departs with more than 30 minutes of delay.

(1%) ii. The probability that a flight departs between 0 and 20 minutes late.

(1%) iii. The probability that a flight departs *early* (i.e. departure delay < 0). Note that if you selected the exponential distribution, this probability is zero by definition – comment on what that implies about the model's limitations in representing early departures.

(2%) (d) Compare the two distributions as candidate models for departure delays. Discuss which one is more appropriate given the test results and the nature of delay data in practice. Do the probabilities computed in part (c) seem realistic? What are the limitations of the chosen model, and what would it mean for the airport if decision-makers relied on the wrong distribution to estimate the frequency of severe delays?

(5%) **11.** The airport authority claims that the **average departure delay is at most 15 minutes**. Using your sample data, test this claim at significance level $\alpha = 0.05$. State the conclusion and comment on its practical implications: if the claim is rejected, what actions should the airport authority consider? If it is not rejected, does that mean delays are acceptable?

(5%) **12.** Using the proportion $\hat{p}$ from your sample, test whether **more than 20% of flights are delayed** at $\alpha = 0.05$. State the conclusion and reflect on what this result means for passengers and for the airport's compliance with typical airline punctuality benchmarks (e.g. the industry standard that fewer than 20% of flights should be delayed).

(5%) **13.** Before comparing means, test whether the AM and PM groups have **equal population variances** using an F-test at $\alpha = 0.05$. Comment on what a large difference in variances between AM and PM flights implies about the predictability of delays in each period.

(5%) **14.** Divide the 30 flights into two groups: **AM** and **PM**. Let $\bar{x}_1$, s_1 refer to the AM group and $\bar{x}_2$, s_2 refer to the PM group (departure delays, in minutes). Using the result of the F-test from the previous question, apply either the pooled t-test or Welch's t-test. Test at level $\alpha = 0.05$ whether AM and PM flights have **different mean departure delays**. State the conclusion and discuss whether the result is consistent with the conditional probabilities found in Part III. What operational explanation could justify a difference (or lack of difference) between AM and PM delay patterns?

A Data Collection Table

30-Flight Sample

Table 1: Sample of 30 flights. Rows 1–16 are pre-filled as illustrative examples (note: flight numbers and delay values are fictional); your group must complete rows 17–30 with 7 AM and 7 PM flights collected from FlightRadar24.

#	Flight	Airline	Date	Dep. delay (min)	Arr. delay (min)	Group (AM/PM)
1	TP1234	TAP Air Portugal	2026-05-25	-5	-3	AM
2	FR4421	Ryanair	2026-05-27	8	6	AM
3	U24418	easyJet	2026-05-29	22	19	AM
4	TP1780	TAP Air Portugal	2026-05-25	-2	-4	AM
5	IB3454	Iberia	2026-05-31	35	31	AM
6	FR4398	Ryanair	2026-06-02	0	2	AM
7	LH1803	Lufthansa	2026-05-28	94	88	AM
8	U24602	easyJet	2026-06-04	-8	-6	AM
9	TP1455	TAP Air Portugal	2026-05-26	12	10	PM
10	FR4812	Ryanair	2026-05-30	-4	-2	PM
11	VY8834	Vueling	2026-06-01	41	37	PM
12	TP1672	TAP Air Portugal	2026-05-27	6	8	PM
13	U24756	easyJet	2026-06-03	19	17	PM
14	IB3478	Iberia	2026-05-29	33	29	PM
15	TP1567	TAP Air Portugal	2026-06-05	4	3	PM
16	TP1890	TAP Air Portugal	2026-06-06	3	2	PM
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						