Exercise Walkthrough: Heavy Planes

Justin Lanfermann

25. June 2025

Abstract

This document provides a detailed, step-by-step solution to the ”Heavy Planes” exercise. Each step is explained with reference to the concepts and definitions from the ”Discrete Probability Theory” script by Niki Kilbertus. The goal is to build a clear and understandable path from the problem statement to the final solution, highlighting the reasoning and the theoretical underpinnings at each stage.

Overview of the Exercise

We are analyzing a scenario involving the weight of 100 passengers on a plane with a maximum weight capacity of 8000 kg. We are given the mean and standard deviation of a passenger's weight. The exercise asks us to discuss modeling assumptions, compute the probability of exceeding the weight limit using the Central Limit Theorem [1], and perform a simple profitability analysis based on this probability.

(i) Modeling Passenger Weights as iid

The first step in any probabilistic modeling is to establish the assumptions about our random variables. Here, we consider the weights of 100 passengers, denoted by the random variables $(X_i)_{i=1}^{100}$. We are asked to argue for and against modeling them as independent and identically distributed (iid) [2].

Argument FOR iid assumption: Modeling the weights as iid is a common and powerful simplification.

- **Identically Distributed:** We can assume that each passenger is drawn from the same general population. Without further information, it's reasonable to believe that the statistical properties (like mean and variance) of passenger 1's weight are the same as for passenger 100.
- **Independent:** The weight of one randomly chosen passenger from a large population generally has no influence on the weight of another. This assumption is crucial for making calculations tractable, particularly for applying limit theorems. As stated in **Remark 1.75**, the iid assumption vastly simplifies mathematical analysis.

Argument AGAINST iid assumption: The iid assumption might not hold perfectly in reality.

- **Not Independent:** People often travel in groups (e.g., families, sports teams, tour groups). Members of a group may share genetics, lifestyle, or age, leading to correlated weights. For example, a flight carrying a national rugby team would have a very different weight distribution than one carrying a group of primary school children.

- **Not Identically Distributed:** The population of passengers might not be homogeneous. For instance, a flight route might predominantly serve a region with a certain demographic, or tickets could be sold with different baggage allowances, systematically affecting the total weight attributed to a passenger.

Conclusion: For this exercise, we will proceed with the given iid assumption, where $E[X_i] = 75$ kg and the standard deviation is $\sigma[X_i] = 20$ kg for all i. This means the variance is $\mathrm{var}[X_i] = (20)^2 = 400$ kg^2.

(ii) Computing the Exact Probability

We want to compute the exact probability that the total weight exceeds 8000 kg. Let $S_{100} = \sum_{i=1}^{100} X_i$ be the total weight of all 100 passengers. We are interested in the probability $P(S_{100} > 8000)$.

Why we can't compute the exact probability: To compute the exact probability $P(S_{100} > 8000)$, we would need to know the exact probability distribution of the sum S_{100}. To find the distribution of a sum of random variables, we generally need to know the full distribution of each individual random variable X_i, not just its mean [3] and variance [4].

The script provides the mean and variance, which are numerical summaries of the distribution (**Definition 2.1**, **Definition 2.5**). However, many different distributions can have the same mean and variance (e.g., a uniform distribution vs. a carefully constructed discrete distribution). Since the problem does not specify the probability density function (pdf) or probability mass function (pmf) of X_i (e.g., that they are normally distributed), we cannot determine the exact distribution of S_{100}.

(iii) Approximating the Probability with the CLT

Since we cannot compute the exact probability, we turn to an approximation. The **Central Limit Theorem (CLT, Theorem 2.64)** [1] is the perfect tool for this. It states that the sum of a large number of iid random variables will be approximately normally distributed, regardless of the original distribution of the variables (as long as they have a finite mean and variance).

Step 1: Define the sum and its parameters. We have the sum $S_{100} = \sum_{i=1}^{100} X_i$. We need its mean and variance.

- **Mean of the sum:** Using the linearity of expectation (**Proposition 2.4 (i)**), the expected value of the sum is the sum of the expected values.

$$E[S_{100}] = E\left[\sum_{i=1}^{100} X_i\right] = \sum_{i=1}^{100} E[X_i] = 100 \times 75 = 7500 \text{ kg}.$$

- **Variance of the sum:** Since the X_i are independent, the variance of the sum is the sum of the variances (**Proposition 2.8 (iv)**).

$$\mathrm{var}[S_{100}] = \mathrm{var}\left[\sum_{i=1}^{100} X_i\right] = \sum_{i=1}^{100} \mathrm{var}[X_i] = 100 \times 400 = 40000 \text{ kg}^2.$$

The standard deviation of the sum is $\sigma[S_{100}] = \sqrt{\mathrm{var}[S_{100}]} = \sqrt{40000} = 200$ kg.

Step 2: Apply the CLT. The CLT tells us that S_{100} is approximately normally distributed with mean 7500 and variance 40000.

$$S_{100} \approx \mathcal{N}(\mu = 7500, \sigma^2 = 40000).$$

To use standard tables or calculators, we standardize the random variable S_{100} to a standard normal random variable Z [5], where $Z \sim \mathcal{N}(0, 1)$.

$$Z = \frac{S_{100} - E[S_{100}]}{\sigma[S_{100}]} = \frac{S_{100} - 7500}{200}.$$

Step 3: Calculate the probability. We want to find $P(S_{100} > 8000)$. We transform this into a statement about Z.

$$\begin{aligned} P(S_{100} > 8000) &= P\left(\frac{S_{100} - 7500}{200} > \frac{8000 - 7500}{200}\right) \\ &= P\left(Z > \frac{500}{200}\right) \\ &= P(Z > 2.5). \end{aligned}$$

This probability can be expressed using the cumulative distribution function (CDF) of the standard normal distribution, denoted by $\Phi(z) = P(Z \leq z)$.

$$P(Z > 2.5) = 1 - P(Z \leq 2.5) = 1 - \Phi(2.5).$$

Using a calculator (like WolframAlpha), we find $\Phi(2.5) \approx 0.9938$.

$$P(S_{100} > 8000) \approx 1 - 0.9938 = 0.0062.$$

So, there is approximately a **0.62% chance** that the maximum weight capacity will be exceeded.

(iv) Profitability Analysis

We need to determine if the flight is profitable on average. This requires calculating the expected profit [3].

Step 1: Define the profit as a random variable. Let Π be the random variable for the profit of a single flight. The profit depends on whether the weight limit is exceeded.

- Revenue per flight is constant: 100 passengers × 3 Euros/passenger = 300 Euros.
- A cost of 3000 Euros is incurred only if $S_{100} > 8000$.

So, the profit Π can take two possible values:

- $\Pi = 300$ Euros, if $S_{100} \leq 8000$.
- $\Pi = 300 - 3000 = -2700$ Euros, if $S_{100} > 8000$.

Step 2: Calculate the expected profit. The expectation of a discrete random variable is the sum of its possible values, each weighted by its probability (**Definition 2.1 (i)**).

$$E[\Pi] = (300) \cdot P(S_{100} \leq 8000) + (-2700) \cdot P(S_{100} > 8000).$$

From part (iii), we have $P(S_{100} > 8000) \approx 0.0062$. Therefore, $P(S_{100} \leq 8000) = 1 - P(S_{100} > 8000) \approx 1 - 0.0062 = 0.9938$.

$$\begin{aligned} E[\Pi] &\approx (300 \times 0.9938) + (-2700 \times 0.0062) \\ &\approx 298.14 - 16.74 \\ &\approx 281.40. \end{aligned}$$

Since the expected profit $E[\Pi] \approx 281.40$ Euros is positive, the flight is **profitable on average**.

(v) [Optional] Improving the Model

The current model is a simplification. To maximize profits, an airline would build a more sophisticated model by relaxing the initial assumptions. This is an exercise in "thinking like a modeler."

Refining the Model:

- **Passenger Load:** Instead of assuming a fully booked flight (100 passengers), model the number of passengers N as a random variable. This could be influenced by season, day of the week, and ticket price. An overbooking strategy could be introduced, where more than 100 tickets are sold, assuming a certain no-show rate.

- **Weight Distribution:** Instead of a single iid model, use a mixture model for passenger weights to account for different demographics (e.g., adult male, adult female, child). The distribution could also be conditioned on the travel class, as business class passengers might have different weight/luggage characteristics. This moves away from the simple 'iid' assumption.

- **Revenue Model:** Ticket prices are not constant. Revenue should be modeled as the sum of variable ticket prices, plus ancillary revenue from baggage fees, seat selection, etc. Baggage fees are a key lever for controlling both revenue and weight.

- **Cost Model:** The cost of excess weight might not be a single fixed value. It could depend on the amount of extra fuel needed, potential delays, and other operational factors.

Optimization Goal: The goal would be to maximize the expected profit, $E[\Pi]$, by making strategic decisions based on the refined model. This could involve:

- **Dynamic Pricing:** Adjusting ticket prices based on demand forecasts and current booking levels.

- **Optimal Overbooking Level:** Finding the sweet spot for selling more tickets than seats to maximize occupancy without having to bump too many passengers.

- **Fueling Policy:** Base the amount of jet fuel on a more accurate forecast of total passenger and cargo weight for the specific flight, rather than a fixed upper limit. This could save fuel costs on lighter-than-average flights.

These refinements transform the problem from a simple probability calculation into a complex optimization problem, typical of the work done in operations research and business analytics (TUM-BWL).

Further Explanations

[1] The Central Limit Theorem (CLT)

The CLT (**Theorem 2.64**) is a cornerstone of probability theory. It states that, for a sufficiently large number of independent and identically distributed (iid) random variables, their sum (or average) will be approximately normally distributed, *regardless* of the distribution from which the individual variables were drawn. This is why the normal (or Gaussian) distribution appears so often in nature and statistics—it often arises from the sum of many small, independent effects.

[2] Independent and Identically Distributed (iid)

This is a fundamental assumption in many statistical models (**Definition 1.74**).

- **Identically Distributed:** All random variables in the sequence $(X_1, X_2, \dots)$ are drawn from the exact same probability distribution. They have the same mean, variance, and overall "shape."

- **Independent:** The outcome of one random variable does not influence the outcome of any other (**Definition 1.72**). Mathematically, the joint probability distribution is the product of the individual (marginal) distributions.

[3] Expectation (Expected Value)

The expectation of a random variable (**Definition 2.1**), denoted $E[X]$, is its long-run average value. It is a weighted average of all possible values the random variable can take, where the weights are the probabilities of those values. For a discrete variable, $E[X] = \sum_x x \cdot P(X = x)$. It represents the "center of mass" of the distribution.

[4] Variance and Standard Deviation

Variance, $\text{var}[X]$ (**Definition 2.5**), measures the "spread" or "dispersion" of a probability distribution. It is the expected value of the squared deviation from the mean: $\text{var}[X] = E[(X - E[X])^2]$. A small variance means the values tend to be close to the mean; a large variance means they are spread out.

- **Standard Deviation**, $\sigma[X]$, is simply the square root of the variance, $\sigma[X] = \sqrt{\text{var}[X]}$. Its advantage is that it is in the same units as the random variable itself, making it more interpretable than variance.

[5] Standard Normal Distribution ($\mathcal{N}(0, 1)$) and its CDF $\Phi(z)$

The **Normal Distribution** is the familiar "bell curve." A **Standard Normal Distribution** is a special case with a mean of 0 and a variance of 1. Any normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ can be converted to a standard normal variable Z by the transformation $Z = (X - \mu)/\sigma$.

- The **Cumulative Distribution Function (CDF)**, $\Phi(z)$, gives the probability that a standard normal variable Z will take a value less than or equal to z, i.e., $\Phi(z) = P(Z \leq z)$. This corresponds to the area under the bell curve to the left of z. There is no simple closed-form formula for $\Phi(z)$, so its values are typically found using statistical tables or software.