Computational Statistics and Data Analysis (MVComp2)

Solutions to exercise 1

Lecturer Tristan Bereau

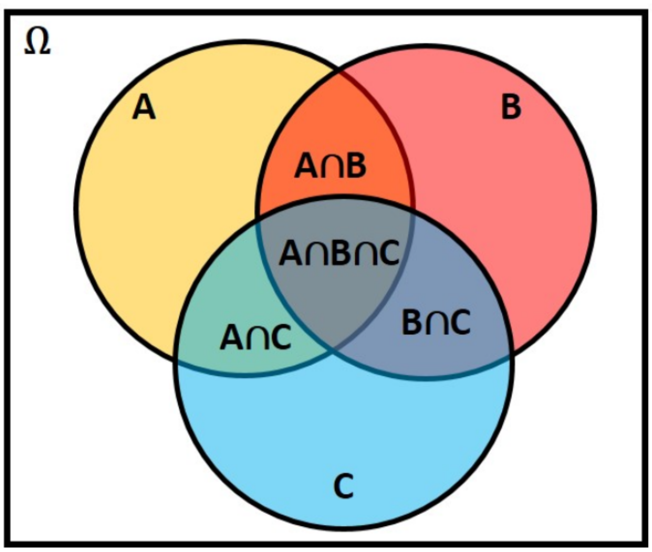
Semester Wi23/24
Due Oct. 27, 2023

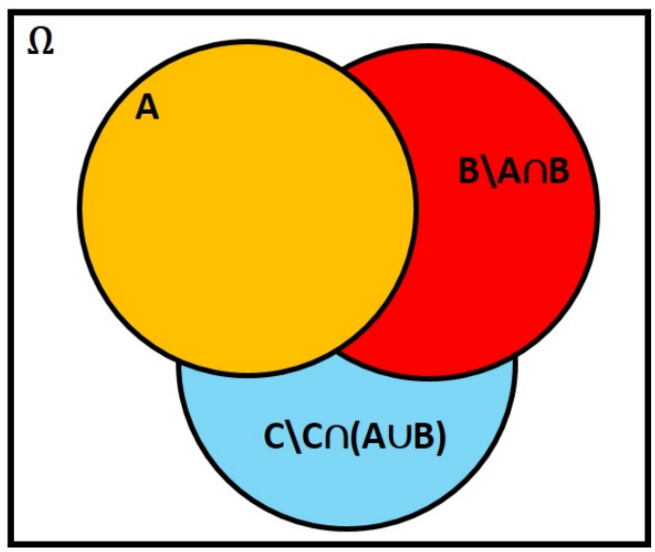
1 Set theory (3 points)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Prove that for arbitrary events $A, B, C \subseteq \Omega$ the following holds:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

1.1 Solution





$$\begin{aligned}
P(A \cup B \cup C) &= P(A \cup (B \setminus (A \cap B)) \cup (C \setminus (C \cap (A \cup B)))) \\
&= P(A) + P(B \setminus (A \cap B)) + P(C \setminus (C \cap (A \cup B))) \\
&= P(A) + (P(B) - P(A \cap B)) + P(C) - P((C \cap A) \cup (C \cap B)) \\
&= P(A) + P(B) + P(C) - P(A \cap B) - [P(C \cap A) + P(C \cap B) - P(A \cap B \cap C)] \\
&= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)
\end{aligned}$$

2 Conditional probabilities (3 points)

Two events, A and B, are such that $P(A) = 0.5$, $P(B) = 0.3$, and $P(A \cap B) = 0.1$. Find the following

(a) $P(A|B)$

(b) $P(A|A \cup B)$

(c) $P(A|A \cap B)$

(d) Are A and B independent?

2.1 Solution

(a) $P(A|B)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3}$$

(b) $P(A|A \cup B)$

$$\begin{aligned} P(A|A \cup B) &= \frac{P(A \cap (A \cup B))}{P(A \cup B)} \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) = 0.7 \\ P(A \cap (A \cup B)) &= P((A \cap A) \cup (A \cap B)) = P(A \cup (A \cap B)) = P(A) = 0.5 \\ \Rightarrow P(A|A \cup B) &= \frac{5}{7} \end{aligned}$$

(c) $P(A|A \cap B)$

$$\begin{aligned} P(A|A \cap B) &= \frac{P(A \cap (A \cap B))}{P(A \cap B)} \\ P(A \cap (A \cap B)) &= P((A \cap A) \cap (A \cap B)) = P(A \cap (A \cap B)) = P(A \cap B) \\ \Rightarrow P(A|A \cap B) &= \frac{P(A \cap (A \cap B))}{P(A \cap B)} = \frac{P(A \cap B)}{P(A \cap B)} = 1 \end{aligned}$$

(d) Are A and B independent?

No, because these two equations are not equal:

$$P(A|B)P(B) = 0.1, \qquad P(A)P(B) = 0.15$$

3 Event composition (2 points)

In a group of 100 students, 40 are taking a math class, 30 are taking a physics class, and 20 are taking both math and physics. You randomly select a student. Calculate the probability that the selected student is taking a math class, but not a physics class.

3.1 Solution

Define events:

- **M**: student takes a math class
- **P**: student takes a physics class

$$P(M \cap \bar{P}) = P(M) - P(M \cap P) = \frac{40}{100} - \frac{20}{100} = \frac{20}{100} = \frac{1}{5}$$

4 Coding exercise: Coupon collector's problem (2 points)

You are collecting a set of different coupons from a cereal box. Each box contains a coupon, and there are a total of N different coupons to collect. The coupons are equally likely to be obtained. Run and average over 100 simulations to estimate the number of boxes you need to buy to collect the entire set of N coupons.

Plot a graph showing how the expected number of boxes changes as a function of N, consider the range $1 \leq N \leq 50$.

4.1 Solution

```python
import random
import numpy as np
import matplotlib.pyplot as plt

def coupon_collector_simulation(num_coupons: int, num_simulations: int) -> float:
  def single_simulation() -> int:
    collected_coupons = set()
    boxes_bought = 0
    while len(collected_coupons) < num_coupons:
        new_coupon = random.randint(1, num_coupons)
        if new_coupon not in collected_coupons:
            collected_coupons.add(new_coupon)
        boxes_bought += 1
    return boxes_bought

  return np.mean(
    [single_simulation() for _ in range(num_simulations)]
  )

# Run the simulations
num_coupons = np.arange(1, 50)
num_simulations = 100
expected_boxes = [
  coupon_collector_simulation(num, num_simulations) for num in num_coupons
]

# Plot the results
plt.scatter(num_coupons, expected_boxes)
```

```
plt.xlabel("Number of Different Coupons")
plt.ylabel("Expected Number of Boxes")
plt.title("Coupon Collector's Problem")
plt.grid()
plt.show()
```