Introduction to probability theory

Statistical Computing and Empirical Methods
Unit EMATM0061, Data Science MSc

Rihuan Ke
rihuan.ke@bristol.ac.uk

Teaching Block 1, 2022

University of BRISTOL

What we will cover today

We will discuss the key role of probability theory in understanding populations from data samples

We will introduce the formal concept of probability

We will derive several important consequences of the rules of probability

Understanding populations from samples

We attempt to answer such questions by looking at data.

Our data sets are samples from a much larger population of penguins.

Sample (e.g., Palmer Penguins Dataset)

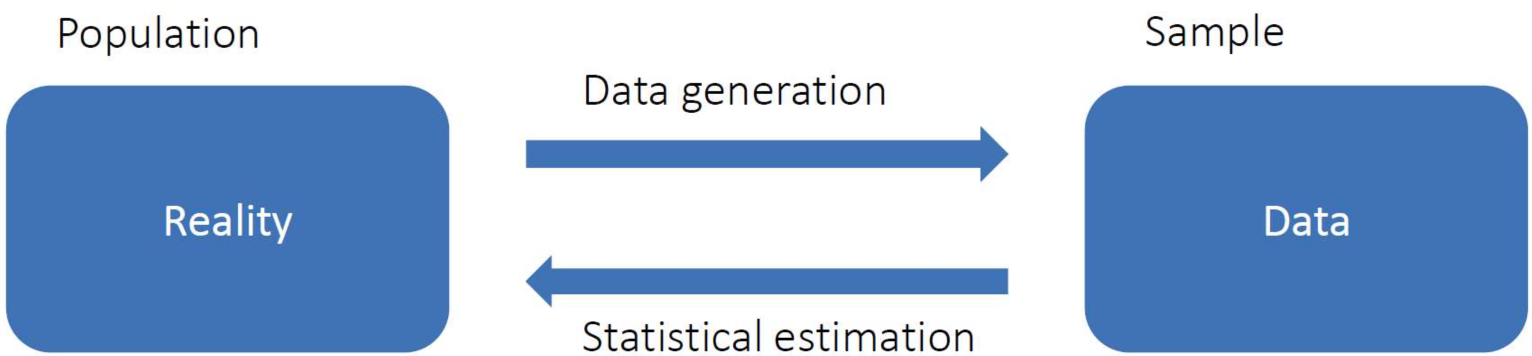
Population (e.g., all penguins in the areas of interest)

Statistical estimation and probability

The problem of variability:

- We can’t weigh every penguin in an entire species
- We can’t try a new marketing idea on all possible customers
- We can’t test a new medication on all patients’ current and future

We must think about how a finite sample reflects a larger population of interest (statistical estimation)



To model the data generation process we will require some probability theory!

Random experiments, events and sample spaces

A random experiment is a procedure (real or imagined) which:

1. has a well-defined set of possible outcomes;
2. could (at least in principle) be repeated arbitrarily many times.

An event is a set (i.e. a collection) of possible outcomes of an experiment

A sample space is the set of all possible outcomes of interest for a random experiment

What is probability?

We often make statements about the probability, likelihood or chance of different events.

"Given how cloudy it is, there's a high likelihood it will rain."

"There is a good chance that the level of inflation will fall due to the rise in interest rates."

"Bristol City Football Club probably won't win the Football Association Challenge cup this year."

We need probability theory to make such statements precise so we can reason about them quantitatively.

How to define probability?

A formal concept of probability can be built on a few rules.

Toy example / intuition (Experiment: Roll a dice)	Motivation →	The laws of probability
Events: e.g., $\{1, 2\}$, $\{3\}$, ...		Events A
Sample space: $\{1, 2, 3, 4, 5, 6\}$		Sample space Ω
probability of $\{1, 2\}$ is $1/3 \geq 0$		Rule 1: $\mathbb{P}(A) \geq 0$ for any event A
probability of $\{1, 2, 3, 4, 5, 6\}$ is 1		Rule 2: $\mathbb{P}(\Omega) = 1$ for sample space Ω
probability of $\{1, 2, 3\}$ = probability of $\{1\}$ + probability of $\{2, 3\}$		Rule 3: For pairwise disjoint events $A_1, A_2, \cdots$, we have $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Definition: Probability

Rules 1, 2, and 3 characterise what probability is.

Definition: Probability

Given a sample space Ω along with a well-behaved collection of events $\mathcal{E}$, a probability $\mathbb{P}$ is a function which assigns a number $\mathbb{P}(A)$ to each event $A \in \mathcal{E}$, and satisfies rules 1, 2, and 3:

Rule 1: $\mathbb{P}(A) \geq 0$ for any event A

Rule 2: $\mathbb{P}(\Omega) = 1$ for sample space Ω

Rule 3: For pairwise disjoint events $A_1, A_2, \cdots$, we have

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

These rules are known as the Kolmogorov axioms after the great mathematician Andrey Kolmogorov who formalized them in 1933

Example 1

Recall: Key elements include sample space Ω, the set of events $\mathcal{E}$, the function of probability $\mathbb{P}$

Example 1. Consider the rolls of a fair dice.

Sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$
Set of events $\mathcal{E} = \{A \subseteq \Omega\}$
Probability $\mathbb{P}(A) = \frac{|A|}{6}$ for any $A \in \mathcal{E}$

Rule 1: $\mathbb{P}(A) \geq 0$ ✓
Rule 2: $\mathbb{P}(\Omega) = 1$ ✓
Rule 3: $\mathbb{P}(A \cup B) = \frac{|A \cup B|}{6} = \frac{|A| + |B|}{6} = \mathbb{P}(A) + \mathbb{P}(B)$ ✓

Example 2

Recall: Key elements include sample space Ω, the set of events $\mathcal{E}$, the function of probability $\mathbb{P}$

Example 2. A customer in the dealership either buys a car (1) or doesn't buy a car (0)

Sample space $\Omega = \{0, 1\}$
Set of events $\mathcal{E} = \{A \subseteq \Omega\} = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$
Probability $\mathbb{P}(\emptyset) = 0,\ \mathbb{P}(\{0\}) = 1 - p,\ \mathbb{P}(\{1\}) = p,\ \mathbb{P}(\{0, 1\}) = 1$ (where $0 \leq p \leq 1$)

Rule 1: $\mathbb{P}(A) \geq 0$ ✓
Rule 2: $\mathbb{P}(\Omega) = 1$ ✓
Rule 3: $\mathbb{P}(\{0, 1\}) = \mathbb{P}(\{0\}) + \mathbb{P}(\{1\}), \mathbb{P}(\{0\} \cup \emptyset) = \mathbb{P}(\{0\}) + \mathbb{P}(\{\emptyset\}), \cdots$ ✓

What are the other desirable properties of probability?

Apart from the properties specified by Rules 1, 2, and 3, which are used to define probability, we have also other intuitively plausible properties, such as

- $\mathbb{P}(\emptyset) = 0$
- If $A, B \in \mathcal{E}$ are events and $A \subseteq B$ (i.e., B implies A), then $\mathbb{P}(A) \subseteq \mathbb{P}(B)$.
- For any event $A \in \mathcal{E}$, we have $0 \leq \mathbb{P}(A) \leq 1$.
- Given any sequence of events $S_1, S_2, \cdots$, we have $\mathbb{P}(\cup_{i=1}^{\infty} S_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(S_i)$.

These properties can be derived from the three rules.

Consequence 1 (the empty set has zero probability)

Recall that Rule 1: $\mathbb{P}(A) \geq 0$; Rule 2: $P(\Omega) = 1$; Rule 3: $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Consequence 1: $\mathbb{P}(\emptyset) = 0$

Proof: $\mathbb{P}(\Omega) = \mathbb{P}(\Omega \cup \emptyset) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset)$ (by Rule 3). Therefore $\mathbb{P}(\emptyset) = 0$.

Consequence 2 (monotonicity property of probability)

Recall that Rule 1: $\mathbb{P}(A) \geq 0$; Rule 2: $P(\Omega) = 1$; Rule 3: $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Consequence 2: If $A, B \in \mathcal{E}$ are events and $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Proof: Clearly, B and $A \backslash B$ are disjoint. So

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(A \cup (B \backslash A)) \\ &= \mathbb{P}(A) + \mathbb{P}(B \backslash A) \quad \text{(by Rule 3)} \\ &\geq \mathbb{P}(A) \end{aligned}$$

Consequence 3 (probabilities are between 0 and 1)

Recall that Rule 1: $\mathbb{P}(A) \geq 0$; Rule 2: $P(\Omega) = 1$; Rule 3: $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Consequence 3: For any event $A \in \mathcal{E}$, we have $0 \leq \mathbb{P}(A) \leq 1$.

Proof:
Firstly, we have $\mathbb{P}(A) \geq 0$ (by Rule 1).
Secondly, since $A \subseteq \Omega$

$$\begin{aligned} \mathbb{P}(A) &\leq \mathbb{P}(\Omega), \quad \text{(by consequence 2)} \\ &= 1 \quad \text{(by Rule 2)} \end{aligned}$$

Consequence 4 (the union bound)

Recall that Rule 1: $\mathbb{P}(A) \geq 0$; Rule 2: $P(\Omega) = 1$; Rule 3: $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Recall Consequence 2: $\mathbb{P}(A) \leq \mathbb{P}(B)$ if $A \subseteq B$.

Consequence 4: Given any sequence of events $S_1, S_2, \cdots$, we have

$$\mathbb{P}(\cup_{i=1}^{\infty} S_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(S_i).$$

Proof: Define a sequence of sets $A_1 = S_1$, $A_2 = S_2 \backslash S_1$, $A_3 = S_3 \backslash (S_1 \cup S_2)$, and

$$A_i := S_i \backslash (S_1 \cup S_2 \cdots \cup S_{i-1}) = S_i \backslash (\cup_{j<i} S_j) \quad \text{for } i = 4, 5, \cdots$$

Step 1 (to show that $A_1, A_2, \cdots$ are pairwise dijoint): For $i_0 < i_1$, we have

$$A_{i_1} \cap A_{i_0} \subseteq \{S_{i_1} \backslash (\cup_{j<i_1} S_j)\} \cap S_{i_0} = \emptyset.$$

So A_{i_0} and A_{i_1} are disjoint.

Step 2: $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} \{S_i \backslash (\cup_{j<i} S_j)\} = \cup_{i=1}^{\infty} S_i$.

Step 3. By Rule 3 and Consequence 2,

$$\mathbb{P}(\cup_{i=1}^{\infty} S_i) = \mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(S_i).$$

The laws of probability and their consequences

Definition: Probability

Given a sample space Ω along with a well-behaved collection of events $\mathcal{E}$, a probability $\mathbb{P}$ is a function which assigns a number $\mathbb{P}(A)$ to each event $A \in \mathcal{E}$, and satisfies rules 1, 2, and 3:

Rule 1: $\mathbb{P}(A) \geq 0$ for any event A

Rule 2: $\mathbb{P}(\Omega) = 1$ for sample space Ω

Rule 3: For pairwise disjoint events $A_1, A_2, \cdots$, we have

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Consequence 1: $\mathbb{P}(\emptyset) = 0$

Consequence 2: If $A, B \in \mathcal{E}$ are events and $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Consequence 3: For any event $A \in \mathcal{E}$, we have $0 \leq \mathbb{P}(A) \leq 1$.

Consequence 4:
Given any sequence of events $S_1, S_2, \cdots$, we have $\mathbb{P}(\cup_{i=1}^{\infty} S_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(S_i)$.

What have we covered?

We introduced the formal concept of probability as governed by Kolmogorov’s axioms.

We derived several important consequences of these rules

- The empty set has zero probability
- Probability is monotonic
- The probability of an event is always between zero and one
- The union bound.

Thanks for listening!

Dr. Rihuan Ke
rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science