总体 随机样本 初等集合论

Population, random samples and elementary set theory

Statistical Computing and Empirical Methods
Unit EMATM0061, Data Science MSc

Rihuan Ke
rihuan.ke@bristol.ac.uk

Teaching Block 1, 2024

What we will cover today

随机变异

We will introduce the fundamental problem of stochastic variability

随机实验, 样本空间和事件

We introduced the concepts of a random experiment, sample space and event

We introduced some fundamental concepts from elementary set theory:

- Intersections, unions, subsets, complements; cardinality, countable & uncountable infinities;

交集, 并集, 子集, 补集, 基数, 可数与不可数的无穷集合

We discussed how these set theoretic concepts can be used to reason about events.

推理事件

The problem of variability

Adelie penguins

Chinstrap penguins

Are Adelie penguins lighter than Chinstrap penguins?

The problem of variability

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)

The problem of variability

We can start by looking at the sample mean of both species

However, different samples can lead to different conclusions!

Sample 1:

											Mean
Adelie	4100	3050	3100	3800	3500	3350	3400	3550	4150	3625	3562
Chinstrap	3600	3650	4800	4400	3800	4400	3500	4500	3500	3300	3945

Sample 2:

											Mean
Adelie	3550	3550	3950	2925	4775	3900	3550	4000	3950	3300	3745
Chinstrap	2700	3325	3650	3950	3800	4300	4050	3900	3675	3700	3705

The underlying reason behind this is that these samples are inherently variable 本质上是有变异性的

We need to take stochastic (随机变异) variation into account

We need to introduce ideas from the theory of (概率论) probability

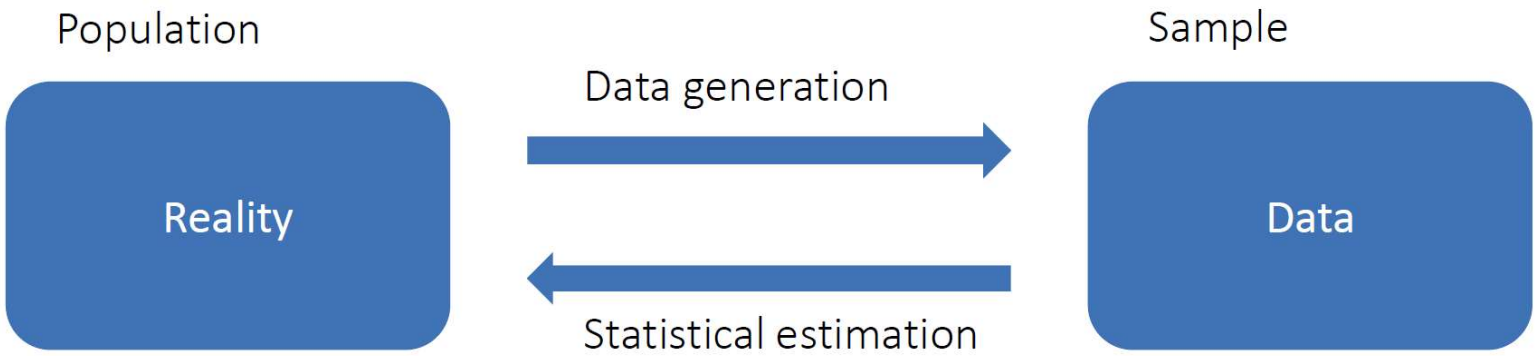
The problem of variability

大样本问题仍然存在

Even if we use a larger sample the problem of variability persists:

- 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

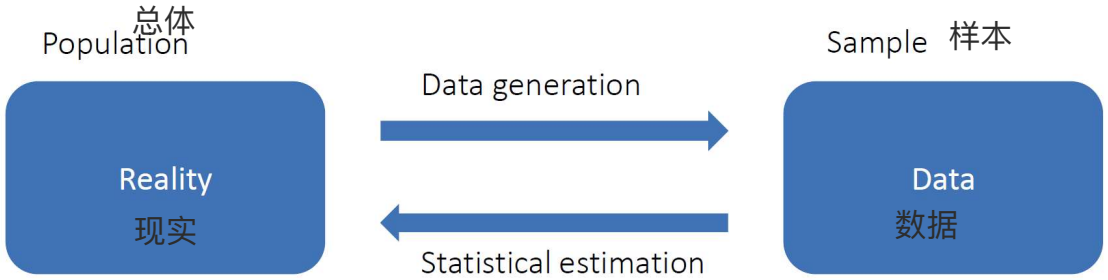
We must think about how a finite sample reflects a larger population of interest (statistical estimation) 统计估计



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

Key motivating questions

We must think about how a finite sample reflects a larger distribution (statistical estimation)



theta hat: 表示通过样本数据估计出的值

We will be exploring how to use statistics $\hat{\theta}$ to learn about population quantities θ

1. (Statistical estimation) How can we design an effective statistic $\hat{\theta}$ for a parameter of interest θ?theta: 代表总体的一个真实参数, 但通常我们无法直接知道它们的具体值
2. (Quantifying uncertainty) How can we quantify our uncertainty about the quantity θ?
3. (Hypotheses testing) How can we use statistics $\hat{\theta}$ to test hypotheses about θ?
4. (Prediction) How can we use our understanding of the population to make predictions about new data?

Probability theory

So we will need to learn probability in order to answer these questions

In the rest of this lecture, we will introduce the following fundamental concepts in probability theory:

1. Random experiments, sample space and event

2. Elementary set theory

In the next lecture, we will look at how these concepts allow us to formalize the fundamental rules of probability.

如何让我们形式化概率的基本规则

1. Random experiments

随机试验是一个过程

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

1. has a well-defined set of possible outcomes; 有一组明确的可能结果
2. could (at least in principle) be repeated arbitrarily many times.
   可以 (至少理论上)无限次重复

Examples.

1. A coin flip for a coin
2. Roll of a dice
3. A customer goes into a shop and decides whether to buy coffee or tea

Note: This is a very broad definition of experiments. There is no suggestion that the experiment is designed or controlled, although this connotation is typical in the natural sciences.

2. Events & sample space

事件(Event) 是一组即一个集合实验可能的结果

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

Random experiment	An event
1. A coin flip for a coin	Heads up or tails up
2. Rolling a dice	One or a few of {1,2,3,4,5,6}
3. A customer goes into a shop	Buy coffee or buy tea or both

是所有感兴趣的可能结果的集合

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

Random experiment	Sample space
1. A coin flip for a coin	Two ways of landing heads ” & tails ”
2. Rolling a dice	The whole set {1,2,3,4,5,6}
3. A customer goes into a shop	All possible purchases (and no purchases).

3. Elementary set theory – definition

The definitions of events and sample space are based on the concept of sets:

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

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

A set is just a collection of objects of interest (our interest is in sets of possible outcomes).

Examples:

1. The set $\mathbb{N}$ consists of all positive whole numbers;
2. The set $\mathbb{R}$ consists of all real numbers;
3. The set $[0, 1]$ consists of all real numbers between zero and one;
4. The empty set $\emptyset$ doesn't contain any objects.

Common set notations

1. We often use **curly braces** $\{\cdots\}$ (containing a list) to denote finite sets of objects.

Example: $\{1, 2, 3, 4, 5\}$ denotes the set of whole numbers less than or equal to five.

2. We write $x \in A$ to denote that x is an element of the set A

Example: We have $1 \in \{1, 2, 3, 4, 5\}$.

3. We write $x \notin A$ to denote x which is not an element of the set A.

Example: We have $6 \notin \{1, 2, 3, 4, 5\}$.

4. Given a set A and a property F we write $\{x \in A : F(x)\}$ for the set of all element x in the set A which satisfy the property F.

集合A中所有满足性质F的元素

Examples: If $A = \{1, 2, 3, 4, 5\}$, then $\{x \in A : x \text{ is odd}\} = \{1, 3, 5\}$ and $\{x \in A : x\text{is even}\} = \{2, 4\}$.

Finite & infinite sets

基数 元素数量

The cardinality of a set is just the number of elements

如果集合的基数是非负的
则A是有限集合

Finite set: A set A is finite if the cardinality of A is a non-negative integer i.e., $1, 2, 3, \cdots$

Examples of finite sets:

空集是基数等于0的集合

- The empty set $\emptyset$ is finite with cardinality zero;
- The set $B = \{2, 4, 6, 8, 10\}$ have cardinality 5

Infinite set: A set is infinite if it is not a finite set.

Examples of infinite sets

- The set $\mathbb{N}$ consisting of all natural numbers;
- The set $\mathbb{Q}$ consisting of all rational numbers;
- The set $\mathbb{R}$ consisting of all real numbers;

Countably & uncountably infinite sets

可数无限集合

Countably infinite set: An infinite set A is countably infinite if there exists an enumeration $a_1, a_2, \cdots, a_n, a_{n+1}, \cdots$ such that

$$A = \{a_1, a_2, \cdots, a_n, a_{n+1}, \cdots\} = \{a_n : n \in \mathbb{N}\}$$

Examples of countably infinite sets

- The set $\mathbb{N}$ consisting of all natural numbers
- The set of all even numbers $\{2, 4, 6, 8, \cdots\}$
- The set $\mathbb{Q} := \{\pm m/n : m, n \in \mathbb{N} \cup \{0\}\}$ of all rational numbers

不可数无限集合

Uncountably infinite set: A set A is uncountably infinite whenever A is infinite but not countably infinite.

Examples of uncountably infinite sets

可数无限集合 可以通过枚举自然数的方式列出所有元素
而不可数无限集合则不能用这种方式列举出来 其规模比可数无限集合更大

- The set consisting of all real numbers;
- Intervals $[a, b]$ for real numbers $a \neq b$;

4. Relationship among sets

Next, we will discuss the relationship among sets, and introduce the concepts of

- Equal sets
- Subsets
- Complement of sets
- Intersection of sets
- Union of sets
- Disjoint sets
- Partitions of sets

Equal sets

Definition: (Equal sets): Two sets A and B are equal if and only if they contain the same elements, that is,

$$\text{if } x \in A, \text{ then } x \in B, \text{ and, if } x \in B, \text{ then } x \in A$$

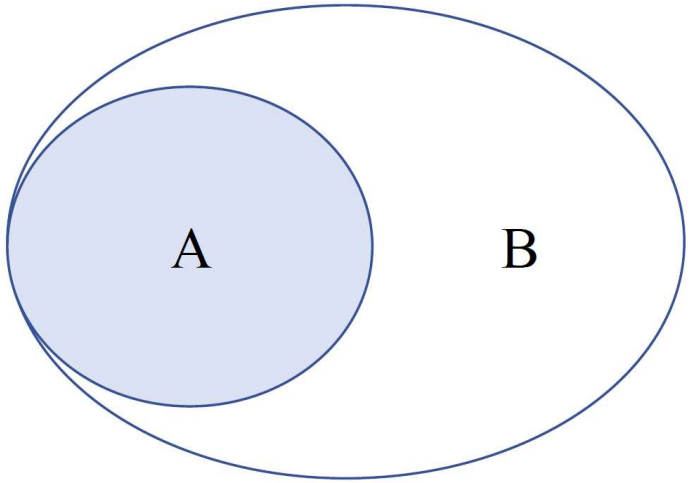
与顺序无关

Order of elements doesn't matter: $\{1, 2, 3\} = \{3, 2, 1\}$

与重复次数不重要

Multiplicity of elements doesn't matter: $\{1, 2, 2, 3\} = \{1, 2, 3\}$

Subsets



Subsets: We say that A is a subset of B if

every element of A is also an element of B.

If A is a subset of B, then we write $A \subseteq B$.

If A is a subset of B, then the event A implies the event B.

Example of subsets:

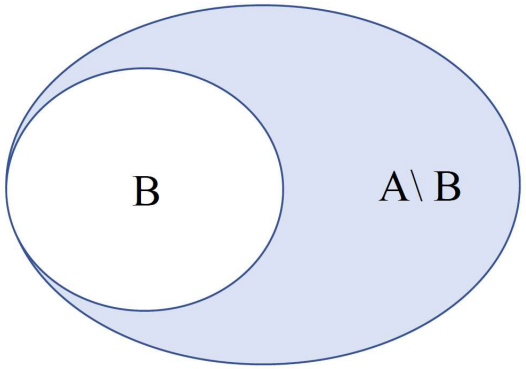
- Any set A is a subset of A
- $\{1, 2, 3\} \subseteq \{1, 2, 3, 4, 5\}$
- Given any event A in the sample space Ω, we have $A \subseteq \Omega$ (recall the sample space is the entire set of all possible outcomes)

Complement of a set

Complement of a set: The complement of B in A is

$$A\backslash B := \{x \in A \mid x \notin B\},$$

that is, the complement of B in A consists of all elements in A but not in B.



$A\backslash B$ is also referred to as the set difference between A and B.

Example: $\{1, 2, 3, 4, 5\}\backslash\{0, 1, 2, 3\} = \{4, 5\}$

事件的补集

Complement of an event: For an event A in the sample space Ω, $\Omega\backslash A$ is the complement event in which A does not occur.

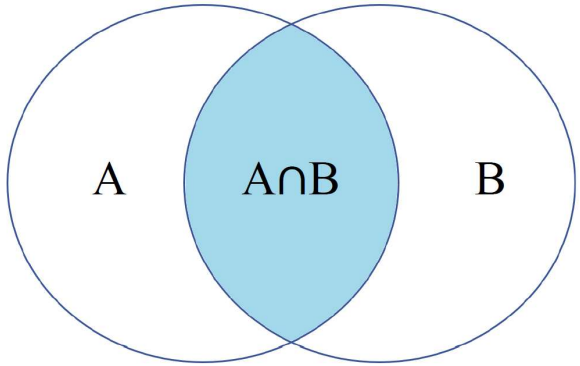
We also write the complement event as A^c.

Intersection and union of sets

Intersections of sets: The intersection between two sets A and B is

$$A \cap B = \{x : x \in A \text{ and } x \in B\},$$

that is, $A \cap B$ is the set of all elements in both A and B.

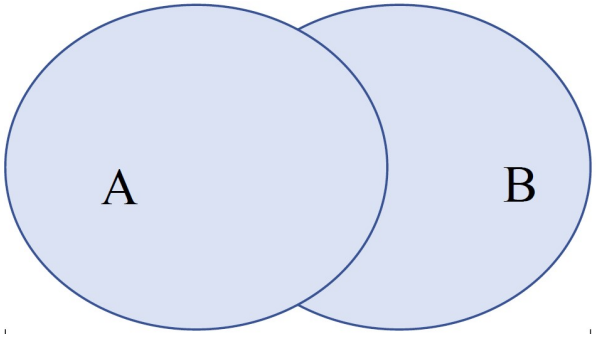


Example: $\{3\} = \{1, 2, 3\} \cap \{3, 4, 5\}$

Unions of sets: The Union between two sets A and B is

$$A \cup B = \{x : x \in A \text{ or } x \in B\},$$

that is, $A \cup B$ is the set of all elements in either A and B.



Example: $\{1, 2, 3, 4, 5\} = \{1, 2, 3\} \cup \{3, 4, 5\}$

Intersections and unions of many sets

We can have intersection and unions of many sets $A_1, \cdots, A_N \subseteq \Omega$:

所有 Ai 都包括的元素集合

x要同时属于A1, A2, A3...

$$\cap_{i=1}^{N} A_i := A_1 \cap \cdots \cap A_n := \{x \in \Omega : x \in A_i \text{ for all } i = 1, 2, \cdots, N\}$$

$$\cup_{i=1}^{N} A_i := A_1 \cup \cdots \cup A_n := \{x \in \Omega : x \in A_i \text{ for at least one } i = 1, 2, \cdots, N\}$$

至少在一个集合 Ai 中包含的元素集合

x至少属于一个Ai

对于无限多个集合

Given infinitely many sets $A_1, A_2, A_3, \cdots$

回想：符号的意义

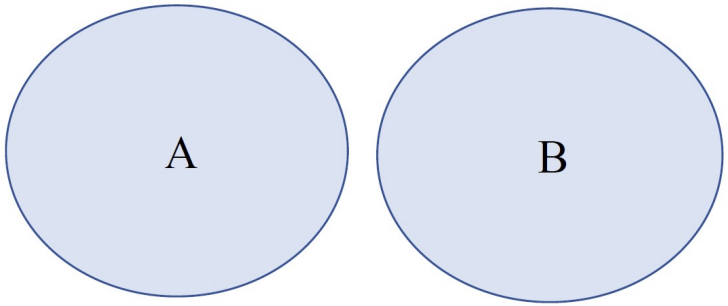
$$\cap_{n \in \mathbb{N}} A_i := \{x \in \Omega : x \in A_n \text{ for all } i \in \mathbb{N}\}$$

$$\cup_{n \in \mathbb{N}} A_i := \{x \in \Omega : x \in A_n \text{ for at least one } i \in \mathbb{N}\}$$

Disjoint sets and partitions

不相交集合

Disjoint sets: two sets A and B are disjoint if $A \cap B = \emptyset$.



Example: $\{1, 2, 3\}$ and $\{4, 5\}$ are disjoint.

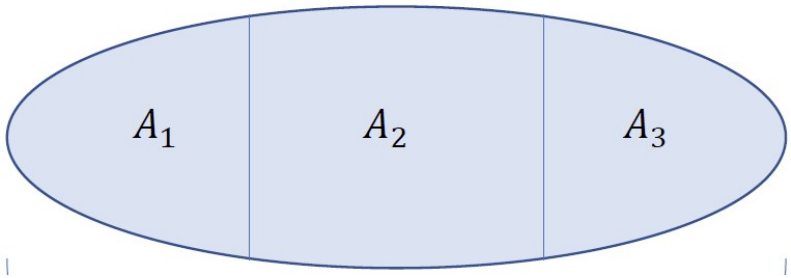
成对不相交

Pairwise disjoint: For sets $A_1, A_2, \cdots, A_k$ is pairwise disjoint if $A_i \cap A_j = \emptyset$ for any $i \neq j$. 任何两个不同的集合都不重叠

集合X的一个划分是将X分成若干个成对不相交的子集A1, A2..., 这些子集的并集等于X

Partition: A partition of set X is a family of $A_1, A_2, \cdots, A_k$ which are pair-wise disjoint and

$$X = \cup_{k \in \{1,2,\cdots,K\}} A_k$$



Example: $\{1, 2\}$, $\{3\}$, $\{4, 5\}$ are a partition of $\{1, 2, 3, 4, 5\}$.

Using set theory to describe events & sample space

We can use ideas from set theory to reason about events and sample spaces

- $A \subseteq B$ (subset) means that event A implies event B A蕴含暗示B A发生 B一定发生
- $A \cap B$ (the intersection) denotes the event in which both A and B occur A和B同时发生
- $A \cup B$ (the union) denotes the event in which at least one of A or B occur 至少发生A或者B, 即其中一个事件或两个事件同时发生的情况
- $A \backslash B$ (the compliment) denotes the event in which A occurs but B does not occur 表示A发生但B不发生的事件

5. Indicator functions 指示函数

指示函数是一个很简单的函数 用来表示一个事件是否发生 或者一个元素是否属于某个集合

Indicator function: Let $A \subseteq \Omega$ be a set (or event). We can associate A with a binary function $\mathbb{1}_A : \Omega \to \{0, 1\}$ by

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

The function $\mathbb{1}_A$ is referred to as the indicator function of A.

With the indicator function, we can represent set operations with functions:

- If $A \subseteq B$, then $\mathbb{1}_A(\omega) \leq \mathbb{1}_B(\omega)$ for all $\omega \in \Omega$
- We have $\mathbb{1}_{A \cap B}(\omega) = \mathbb{1}_A(\omega) \cdot \mathbb{1}_B(\omega)$ for all $\omega \in \Omega$ 只有元素w同时在集合A和B中, indicator为1 否则0
- We have $\mathbb{1}_{A \cup B}(\omega) = \max(\mathbb{1}_A(\omega), \mathbb{1}_B(\omega))$ for all $\omega \in \Omega$ 如果元素w属于A或B, indicator为1; 如果不属于任何一个 则为0
- We have $\mathbb{1}_{A \setminus B}(\omega) = \mathbb{1}_A(\omega) \cdot (1 - \mathbb{1}_B(\omega))$ for all $\omega \in \Omega$ 元素w在A不在B时, indicator为1

What have we covered?

We introduced the fundamental problem of stochastic variability

We introduced the concepts of a random experiment, sample space and event

We introduced some fundamental concepts from elementary set theory:

- Intersections, unions, subsets, complements, cardinality, countable & uncountable infinities;

We discussed how these set theoretic concepts can be used to reason about events

We explored the idea of using indication functions to represent set operations

University of BRISTOL

Thanks for listening!

Dr. Rihuan Ke

rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science