Week 2

2024-09-23

Preliminary

R packages

The ***tidyverse*** package is a collection of R packages that are designed for data science.

```
library(tidyverse)
```

It includes

- dplyr (e.g., select, filter, mutate, arrange, summrize)
- tidyr (e.g., pivot_longer, pivot_wider, unite, complete)
- ggplot2 (e.g., ggplot, geom_point)
- purrr (e.g. map, map_int)
- magrittr (e.g., %>%)
- and many other

Visit https://www.tidyverse.org/packages/ to learn more.

Data frame and examples

In certain analysis tasks we need to deal with tabular data. Tabular data set represented by an R data frame. An example being used here is the Palmer penguins data set.

```
library(palmerpenguins)
```

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

1. Data wrangling

Data wrangling is a process of transforming data from one form to another in preparation for another downstream task (e.g., visualisation, modelling)

Understand the use of pipe operator (%>%). In the above code, we have used the pipe operator. Another example is given below.

```
# an example about the pipe operation

f1 <- function(a){ return (a+1) }
f2 <- function(b){ return (b+2) }
f3 <- function(c){ return (c+3) }

x <- 0

f1(x) # 0+1 = 1
f2(f1(x)) # 0+1+2 = 3
f3(f2(f1(x))) # 0+1+2+3 = 6

x %>% f1 # 1
x %>% f1 %>% f2 # 3
x %>% f1 %>% f2 %>% f3 # 6
```

Remark: "%>%" operator is given by the magrittr package.

Question (optional): the default behavior of "%>%" is to place the left hand side as the first argument of the function on the right hand side, when the function has multiple arguments. What if you want to the left hand side as the second argument? e.g., f2(1, f1(x))

Basic data wrangling operations using R:Selecting columns (select), filtering rows (filter), create new columns (mutate), rename columns (rename), sort rows (arrange), summarize (summarize), joining multiple data frames (join)

```
# try to understand what each of the following does


# select columns
penguinsv2 <- select(penguins, species, bill_length_mm,
                     body_mass_g, flipper_length_mm )
select(penguins, -species, -bill_length_mm, -body_mass_g)

# filter rows
filter(penguinsv2, species=='Gentoo')
filter(penguinsv2, species=='Gentoo' & body_mass_g>5000)

# create columns
penguinsv2 %>%
  mutate(flipper_bill_ratio=flipper_length_mm/bill_length_mm)

# rename columns
penguinsv2 %>% rename(f_l_m = flipper_length_mm)

# sort
penguinsv2 %>% arrange(bill_length_mm)
penguinsv2 %>% arrange(desc(bill_length_mm))
```

```r
# summarise
penguinsv2 %>%
  summarize(num_rows=n(), avg_weight_kg=mean(body_mass_g/1000, na.rm=TRUE),
            avg_flipper_bill_ratio=mean(flipper_length_mm/bill_length_mm,
                                        na.rm=TRUE))
penguinsv2 %>%
  group_by(species) %>%
  summarize(num_rows=n(), avg_weight_kg=mean(body_mass_g/1000, na.rm=TRUE),
            avg_flipper_bill_ratio=mean(flipper_length_mm/bill_length_mm,
                                        na.rm=TRUE))

penguinsv2 %>% summarize(across(everything(), ~sum(is.na(.x))))
# remark: ~sum(is.na(.x))) defines a function
```

Additional reading (optional): https://dplyr.tidyverse.org/

Understand the different types of join operations (inner join, outer join, left join, out join).

2. To answer questions about propoluation using data, we need probability theory.

The relationship between population and sample

In the last section, we discussed how to use R programming to extract key information from the sample, for example, the average body mass of the penguin sample. In statistics, "sample" is a term relative to the concept of population.

1. **Sample**: We refer to the data set as a data sample or just sample.

2. **Population**: In statistics, population refers to the entire group of individuals from which your sample is drawn.

So we're able to compute sample quantity based on the data. However, it is more often the case that our true interests lie in the associated population quantity, rather than the sample quantity themselves. An example is given below.

Answering questions with data: An illustrative example

In this example, we consider the penguins data as we discussed last week.

Suppose that we are interested in a research question about whether the average weight of Adelie penguins is less than that of Chinstrap penguins. This question concerns the whole set of Adelie penguins and Chinstrap penguins. So We want to compare the population quantities (i.e., the population average) of the two penguins species.

To answer the question, one would measure the body weights of all individual Adelie penguins and take the average, and then do the same for the Chinstrap penguins.However, in realistic, it is simply not

Figure 1: The population of penguins

possible to collect data for each individual penguins. Such a problem appears in many situations when we try to answer questions about population:

- 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

Since it is not practical to obtain data for every possible penguin, to answer the above research question, we can alternatively collect a data set associated with a subset of penguins (and we often refer the data set to as a sample). By looking at the the sample, we learn about the properties of the whole population (the set of all possible penguins).

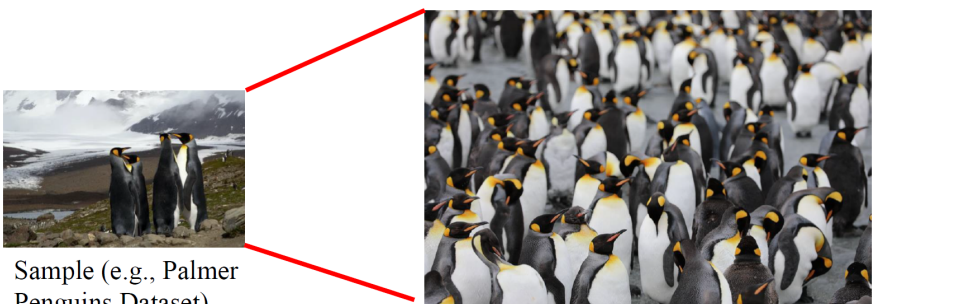


Figure 2: sample

With the sample, one of the natural questions that we may ask is: What can we say about the the population average of the two penguins species based on the sample that we have? This is relevant to a key statistical concept about inference.

1. **Inference**: The process of using data analysis to infer properties or quantities of the population.

Making inference about the underlying populations based on the samples lies at the heart of statistics, and we will cover more about this in the future lectures.

3. Population, random samples

One fundamental problem that is widely encountered in making inference is the problem of variability.

The problem of variability

Your sample is random (typically a small subset of your population)!

Example:

- Assume that the population is: 0, 1, $\cdots$, 1000
- the average of the population is therefore 500

To estimate the average of the population, we randomly draw a sample of size 10 and compute the average of the sample.

```
S <- sample(0:1000, 10)
mean(S)
```

```
## [1] 384.4
```

If we draw another sample, we get a different result

```r
S <- sample(0:1000, 10)
mean(S)
```

```
## [1] 359.1
```

This is the problem of variability.

The problem of variability arises because of the fact that samples are inherently variable. We need to take stochastic variation into account.

We need ***probability theory*** to study the stochastic variation and understand how a finite sample reflects a larger population of interest.

In this and for following weeks, we will cover basic concepts in probability theory.

Some fundamental concepts in probability theory.

The concepts of random experiments, event, and sample space

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

a). has a well-defined set of possible outcomes;

b). could (at least in principle) be repeated arbitrarily many times.

Examples: A coin flip for a coin (outcomes: Heads up, tails up)

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

Example: heads up; tails up; heads up or tails up

It is important to note that an event is different from a outcome (although an event may contain a single outcome).

An event A is said to occur if A contains the (actual) outcome of the random experiment.

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

Example: heads up, tails up

The sample space is also an event, and it can be viewed as "the event with largest size"

Discussion and Task:

Let us consider the 2024 US presidential election as a random experiment.

- Give an example of a single outcome (any outcome would be fine)
- Give an example of a single event (any event would be fine)
- What is the sample space?

You can discuss with your neighbors. Try to use as much mathematical language as you can.

4. Elementary set theory

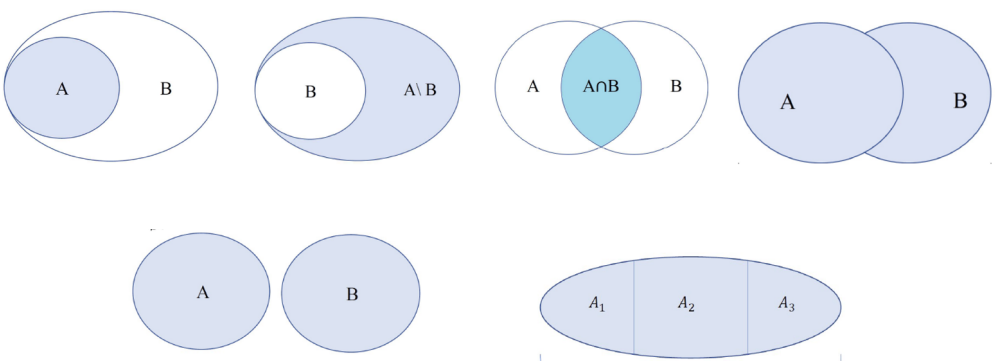
We have used the concept of set above (sample space, events).

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

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

Relationships among sets

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



Using set theory to describe events & sample space

Recall that

- An **event** is a set of possible outcomes of an experiment
- A **sample space** is the set of all possible outcomes of interest for a random experiment

We can use set to represent events and sample spaces

A is a subset of B
$\iff$ $x \in A$ then we have $x \in B$
$\iff$ A occurs then B must occur (B implies A)

Example: rolling a dice (sample space is $\{1, 2, \cdots, 6\}$)
Event $A = \{2, 3\},\ B = \{2, 3, 4\}$.
A occurs (the outcome is either 2 or 3), then B occurs.

A intersection B
$\iff$ $\{x : x \in A \text{ and } x \in B\}$
$\iff$ both A and B occurs

A union B
$\iff$ $\{x : x \in A \text{ or } x \in B\}$
$\iff$ either A or B occurs

Complement of B in A
$\iff$ $\{x : x \in A \text{ and } x \notin B\}$
$\iff$ A occurs but B does not occurs

Indicator function

A set can be described by an indicator function

For a set A, we can associate it with a binary function $\mathbf{1}_A : \Omega \to \{0, 1\}$ by

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

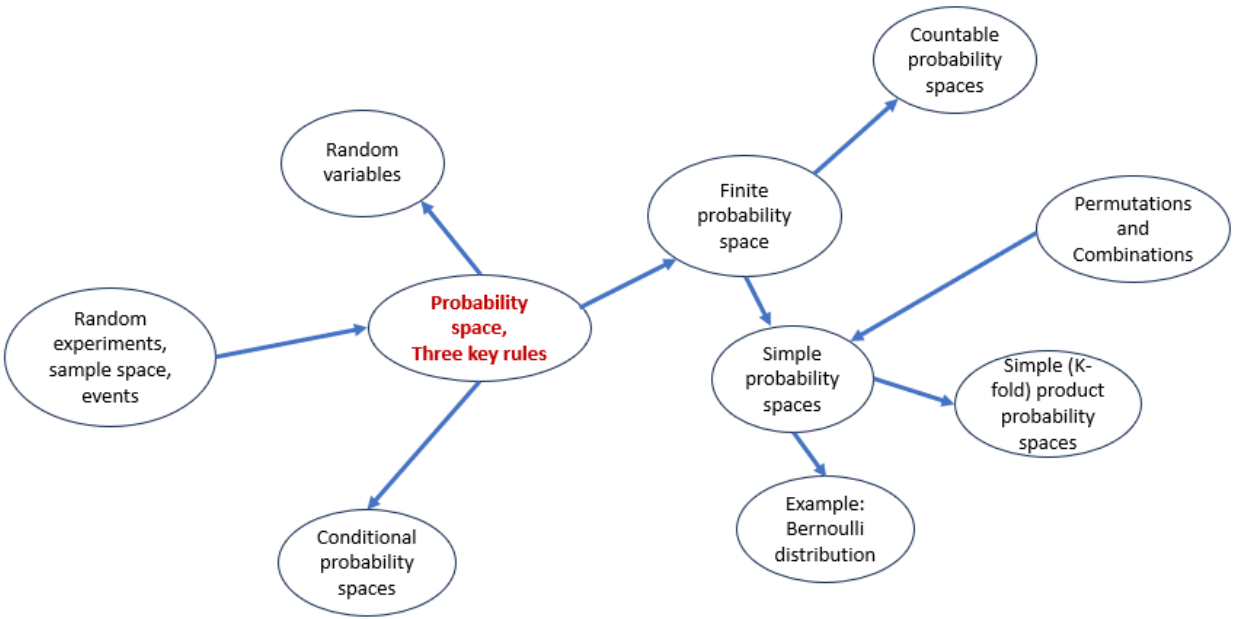
The function $\mathbf{1}_A$ is referred to as the indicator function of A.

Perform set operations with operations of indicator functions.

- If $A \subseteq B$, then $\mathbf{1}_A(\omega) \leq \mathbf{1}_B(\omega)$ for all $\omega \in \Omega$
- We have $\mathbf{1}_{A \cap B}(\omega) = \mathbf{1}_A(\omega) \cdot \mathbf{1}_B(\omega)$ for all $\omega \in \Omega$
- We have $\mathbf{1}_{A \cup B}(\omega) = \max(\mathbf{1}_A(\omega), \mathbf{1}_B(\omega))$ for all $\omega \in \Omega$
- We have $\mathbf{1}_{A \setminus B}(\omega) = \mathbf{1}_A(\omega) \cdot (1 - \mathbf{1}_B(\omega))$ for all $\omega \in \Omega$

5. The formal concepts of probability

Built on the above concepts, we introduce the formal concept of probability.



The formal definition of probability is given by three key rules, which are also called the laws of probability.

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

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

A probability space must satisfy the three key rules above. Other properties of probability can be derived from the three key rules.

Consequences and properties of probability

Consequence 1: The empty set has zero probability:

$$\mathbb{P}(\emptyset) = 0$$

Consequence 2: Monotonicity property: If $A, B \in \mathcal{E}$ are events and $A \subseteq B$, then

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

Consequence 3: Probabilities are between 0 and 1: For any event $A \in \mathcal{E}$, we have

$$0 \leq \mathbb{P}(A) \leq 1.$$

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

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

Consequence 5: Union of a finite sequence of disjoint events: For pairwise disjoint events $A_1, A_2, \cdots, A_n$,

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

Consequence 6: Probability of a complement: For any $S \subseteq \Omega$, its complement $S^c := \Omega \backslash S$ satisfies

$$\mathbb{P}(S^c) = 1 - \mathbb{P}(S).$$

Consequence 7: Probability of union and intersection of events: For events $A \subseteq \Omega$ and $B \subseteq \Omega$,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Goals:

Understand the concepts of probability through examples (verify that the example satisfies three key rules), e.g., the examples in Lecture 6.

Understand the consequences of the three key rules and why the three key rules in the definition of probability are inherently consistent.

(Optional) σ-algebra

In the definition of probability, we denote by $\mathcal{E}$ a well-behaved collection of events. What does the term "well-behaved" mean?

To make the definition of probability rigorous we require the concept of a σ algebra.

A σ-algebra is a collection $\mathcal{E}$ consisting of subsets $A \subseteq \Omega$ satisfying:

1. The set $\Omega \in \mathcal{E}$.
2. If $A \in \mathcal{E}$, then $\Omega \backslash A \in \mathcal{E}$.
3. If there is a countable sequence $A_1, A_2, A_3, \cdots$, and $A_i \in \mathcal{E}$ for all i, then

$$\cup_{i=1}^{\infty} A_i \in \mathcal{E}$$

Then we can revise the definition of probability as follows.

Definition (probability). Given a sample space Ω along with a ~~well-behaved collection of events~~ σ-algebra of event $\mathcal{E}$, 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

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

Please note that the third property in the definition of σ-algebra ensures that $\cup_{i=1}^{\infty} A_i$ is in $\mathcal{E}$, hence $\mathbb{P}(\cup_{i=1}^{\infty} A_i)$ is well-defined.