Imperial College London

12th October 2018

# Probability theory

https://bitbucket.org/mfumagal/statistical_inference

Matteo Fumagalli

# Intended Learning Outcomes

By the end of this session, you will be able to:

- Describe the principles of set theory and set operations
- Illustrate the axiomatic foundations of probability theory and appropriate counting methods
- Identify dependence and indepedence of events
- Show the utility of distribution functions for random variables
- Demonstrate how to implement basic probability calculus in R

# Set theory

Draw conclusions about a population of objects after an experiment. What are the possible outcomes of the experiment?

**Sample space**

The set $S$ of all possible outcomes of a particular experiment is called the *sample space* for the experiment.

# Set theory

Examples of experiments:

- tossing a con: $S = \{H, T\}$
- GCSE scores of randomly selected pupils:

# Set theory

Examples of experiments:

- tossing a con: $S = \{H, T\}$
- GCSE scores of randomly selected pupils: $S = \{0, 1, 2, ..., 8, 9\}$.
- reaction time to a stimulus:

# Set theory

Examples of experiments:

- tossing a con: $S = \{H, T\}$
- GCSE scores of randomly selected pupils: $S = \{0, 1, 2, ..., 8, 9\}$.
- reaction time to a stimulus: $S = (0, \infty)$.

Imperial College London

# Set theory

jupyter notebook:

- Imagine that our experiment consists on observing the nucleotidic sequence of a particular gene of interest. What is the sample space $S_G =?$
- Now suppose that we are interested in making inference on the amino acidic sequence of a protein. What is the sample space $S_P =?$
- Suppose that our observations consist in the divergence between orthologous genes of arbitrary length. What is the sample space for such divergence $S_D =?$

# Sets and events

One the sample space has been defined (e.g. countable?), we consider collections of possible outcomes of an experiments.

**Event**

An *event* is any collection of possible outcomes of an experiment, that is, any subset of $S$, including $S$ itself.

Let $A$ ben an event, a subset of $S$. We say that the event $A$ occurs if the outcome of the experiment is in the set $A$.

$$A \subset B \iff x \in A \implies x \in B$$

$$A = B \iff A \subset B \text{ and } B \subset A$$

# Set operations

Given any two events $A$ and $B$:

- Union: the union of $A$ and $B$ is the set of elements that belong to either $A$ or $B$ or both:

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

# Set operations

Given any two events $A$ and $B$:

- Union: the union of $A$ and $B$ is the set of elements that belong to either $A$ or $B$ or both:

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

- Intersection: the intersection of $A$ and $B$ is the set of elements that belong to both $A$ and $B$:

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

# Set operations

Given any two events $A$ and $B$:

- Union: the union of $A$ and $B$ is the set of elements that belong to either $A$ or $B$ or both:

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

- Intersection: the intersection of $A$ and $B$ is the set of elements that belong to both $A$ and $B$:

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

- Complementation: the complement of $A$ is the set of all elements that are not in $A$:

$$A^c = \{x : x \notin A\}$$

# Set operations

jupyter notebook:

Consider an experiment of drawing a nucleotidic base at random. The sample space is $S = \{A, C, G, T\}$ and some possible events are $A = \{G, C, A\}$ and $B = \{T, G\}$.

What is the union of $A$ and $B$? What is the intersect between $A$ and $B$? What is the complement of $A$? What is the complement of the union of $A$ and $B$?

Use Venn diagrams and R!

# Properties of Set operations

- Commutativity

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

- Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

# Properties of Set operations

- Distributive laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- DeMorgan's laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

# Partition

Two events $A$ and $B$ are *disjoint* (or mutually exclusive) if $A \cap B = \emptyset$.
The events $A_1, A_2, \ldots$ are pairwise disjoint if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

If $A_1, A_2, \ldots$ are pairwise disjoint and $\bigcup_{i=0}^{\infty} A_i = S$, then the collection $A_1, A_2, \ldots$ forms a *partition* of S.

# Axiomatic foundations

For each event $A$ in the sample space $S$ we associate a number between 0 and 1 that we call the probability of A, denoted as $P(A)$.

A collection of subsets of $S$ is called a *sigma algebra* (or Borel field), denoted by $\mathcal{B}$, if it satisfies the following properties:

1. $\emptyset \in \mathcal{B}$
2. if $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$
3. if $A_1, A_2, ... \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$

# Axiomatic foundations

jupyter notebook:

- Is the collection of sets $\{\emptyset, S\}$ a sigma algebra with $S$?
- If $S$ has $n$ elements, there are $2^n$ sets in $\mathcal{B}$. What is $\mathcal{B}$ if $S = \{1, 2, 3\}$?

# Probability function

Given a sample space $S$ and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function $P$ with domain $\mathcal{B}$ that satisfies:

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
2. $P(S) = 1$
3. If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then $P(\bigcup_{i=1}^{\infty}) = \sum_{i=1}^{\infty} P(A_i)$

Any function $P$ that satisfies these Axioms of Probability is called a probability function.

# Probability function

Let $S = \{s_1, s_2, ..., s_n\}$ a finite and/or countable set.
Let $\mathcal{B}$ be any sigma algebra of subsets of $S$.
Let $p_1, p_2, ..., p_n$ nonnegative numbers that sum to 1.
For any $A \in \mathcal{B}$, define $P(A)$ by

$$P(A) = \sum_{\{i: s_i \in A\}} p_i$$

Then $P$ is a probability function on $\mathcal{B}$.

# Calculus

If $P$ is a probability function and $A$ is any set in $\mathcal{B}$, then

1. $P(\emptyset) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$

prove nr.3?

# Calculus

If $P$ is a probability function and $A$ and $B$ are any sets in $\mathcal{B}$, then

1. $P(B \cap A^c) = P(B) - P(A \cap B)$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. if $A \subset B$, then $P(A) \leq P(B)$

# Calculus

Since $P(A \cup B) \leq 1$, then

$$P(A \cap B) \geq P(A) + P(B) - 1$$

This is a special case of the *Bonferroni's Inequality*: to bound the probability of a simultaneous event from the probabilities of the individual events.

# Calculus

If $P$ is a probability function, then

1. $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ for any partition $C_1, C_2, ...$ of $S$
2. $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ for any sets $A_1, A_2, ...$

This property is called *Boole's Inequality*, a general version of the Bonferroni's Inequality.

Imperial College London

# Counting

Methods of counting are used to construct probability assignments on finite sample spaces.

## Fundamental Theorem of Counting

If a job consists of $k$ separate tasks, the $i-th$ of which can be done in $n_i$ ways, $i = 1, ..., k$, then the entire job can be done in $n_1 \times n_2 \times ... \times n_k$ ways.

Imperial College London

# Counting

Let's assume that one protein domain is comprised of 6 amino acids. To be able to calculate the probability of having exactly one sequence we first must count how many different combinations of 6 amino acids can be observed.

Is that enough information to calculate such probabilities? What else do we need?

Imperial College London

# Counting

| | without replacement | with replacement |
|---|---|---|
| ordered | | |
| unordered | | |

# Ordered, without replacement

The first amino acid can be selected in ... ways, the second in ...

# Ordered, without replacement

The first amino acid can be selected in ... ways, the second in ...

For a positive integer $n$, $n!$ (read $n$ factorial) is the product of all the positive integers less than or equal to $n$, that is, $n! = n \times (n-1) \times (n-2) \times ... \times 2 \times 1$. We define $0! = 1$.

# Ordered, with replacement

All amino acids can be selected in ... ways ...

# Unordered, without replacement

Since we know the number of ways when the ordering must be taken into account, ..., we need to ... this number by all the possible ways that 6 amino acid can be ordered, which is ...

# Unordered, without replacement

Since we know the number of ways when the ordering must be taken into account, ..., we need to ... this number by all the possible ways that 6 amino acid can be ordered, which is ...

For nonnegative integers $n$ and $r$, where $n \geq r$, we define the symbol $\binom{n}{r}$, read $n$ choose $r$, as

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

These numbers are also called *binomial coefficients*.

# Unordered, with replacement

Imagine to assign 6 amino acids on the 20 possible values. We can reduce the ways of counting noting that we just need to keep track of the arrangement and that the two boundaries of the 22 interval bounds (since we can 20 bins) are not relevant. Therefore, we have $\frac{25!}{6!19!}$ ways.
$\binom{n+r-1}{r}$

jupyter notebook: fill in the table of counting methods.

# Enumerating outcomes

Suppose $S = \{s_1, ..., s_n\}$ is a finite sample space. If all the outcomes are equally likely, then $P(\{s_i\}) = 1/N$ for every outcome $s_i$.
If so,

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} \frac{1}{N} = \frac{\text{nr. elements in } A}{\text{nr. elements in } S}$$

# Conditional probability

If $A$ and $B$ are events in $S$, and $P(B) > 0$, then the *conditional probability of $A$ given $B$*, written $P(A|B)$, is

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

# Conditional probability

If $A$ and $B$ are events in $S$, and $P(B) > 0$, then the *conditional probability of $A$ given $B$*, written $P(A|B)$, is

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

The original sample space $S$ has been **updated** to $B$, so that $P(B|B) = 1$.
If $A$ and $B$ are disjoints, then $P(A \cap B) = 0$ and $P(A|B) = P(B|A) = 0$.

# Independence

Two events, $A$ and $B$, are statistically independent if

$$P(A \cap B) = P(A)P(B)$$

If the occurrence of $B$ has no effect on the probability of $A$, then $P(A|B) = P(A)$.

# Independence

If $A$ and $B$ are independent events, then the following pairs are also independent:

- $A$ and $B^c$
- $A^c$ and $B$
- $A^c$ and $B^c$

# Independence

A collection of events $A_1, A_2, ..., A_n$ are *mutually independent* if for any subcollection $A_{i_1}, ..., A_{i_k}$, we have

$$P(\bigcap_{j=1}^{k} A_{i_j}) = \prod_{j=1}^{k} P(A_{i_j})$$

Simultaneous independence requires a strong definition.

# Random variables

Assume we are interested in the GC content of a nucleotidic sequence $\{A, C, G, T\}$ of 50 base pairs.
If we record "1" for a GC base and "0" otherwise (A or T), what is the sample space for this experiment?

# Random variables

Assume we are interested in the GC content of a nucleotidic sequence $\{A, C, G, T\}$ of 50 base pairs.
If we record "1" for a GC base and "0" otherwise (A or T), what is the sample space for this experiment?
$S$ has $2^{50}$ elements!

# Random variables

Assume we are interested in the GC content of a nucleotidic sequence $\{A, C, G, T\}$ of 50 base pairs. We record "1" for a GC base and "0" otherwise (A or T),
If we define $X =$ number of 1s recorded out of 50, what is the sample space for $X$?

# Random variables

Assume we are interested in the GC content of a nucleotidic sequence $\{A, C, G, T\}$ of 50 base pairs. We record "1" for a GC base and "0" otherwise (A or T),
If we define $X =$ number of 1s recorded out of 50, what is the sample space for $X$?
$S$ for $X$ is the set of integers $\{0, 1, 2, ..., 50\}$!

# Random variables

When we define $X$, we created a mapping (a function) from the original sample space to a new sample space.

A *random variable* is a function from a sample space $S$ into the real numbers.

In defining a random variable, we also define a new sample space. Can our probability function on the original sample space be used for the random variable?

# Random variables

Suppose we have a sample space $S = \{s_1, ..., s_n\}$ with a probability function $P$ and we define a random variable $X$ with range $\mathcal{X} = \{x_1, ..., x_m\}$.

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\})$$

If $\mathcal{X}$ is uncountable?

# Random variables

Suppose we have a sample space $S = \{s_1, ..., s_n\}$ with a probability function $P$ and we define a random variable $X$ with range $\mathcal{X} = \{x_1, ..., x_m\}$.

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\})$$

If $\mathcal{X}$ is uncountable?

$$P_X(X \in A) = P(\{s \in S : X(s) \in A\})$$

# Distribution functions

With every random variable $X$, we associate a function called the cumulative distribution function of $X$.

The *cumulative distribution function* (or cdf) of a random variable $X$, denoted by $F_X(x)$ is defined by

$$F_X(x) = P_X(X \leq x), \text{ for all } x$$

Experiment: roll a die, what is $F_X(x)$?

# Cumulative distribution function

jupyter notebook:

Consider the experiment of sampling three nucleotidic bases and let $X$ be the number of GC bases observed. What is the cdf of $X$?

# Cumulative distribution function

jupyter notebook:

Consider the experiment of sampling three nucleotidic bases and let $X$ be the number of GC bases observed. What is the cdf of $X$?

$$F_X(x) = 0, \text{ if } -\infty < x < 0$$
$$F_X(x) = ?, \text{ if }$$
$$\dots$$

# Cumulative distribution function

The function $F_X(x)$ is a cdf if and only if:

1. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$
2. $F_X(x)$ is a nondecreasing function of $x$
3. $F_X(x)$ is right-continuous

Imperial College London

# Cumulative distribution function

Whether a cdf is continuous or not corresponds to the associated random variable being continuous or not.

A random variable $X$ is *continuous* if $F_X(x)$ is a continuous function of $x$. A random variable $X$ is *discrete* if $F_X(x)$ is a step function of $x$.

# Cumulative distribution function

The random variables $X$ and $Y$ are *identically distributed* if, for every set $A \in \mathcal{B}^1$ (the smallest algebra sigma), $P(X \in A) = P(Y \in A)$.

If $X$ and $Y$ are identically distributed, $F_X(x) = F_Y(x)$ for every $X$. $F_X(x)$ completely determines the probability distribution of a random variable $X$.

# Density and mass functions

The *probability mass function (pmf)* of a discrete random variable $X$ is given by

$$f_X(x) = P(X = x) \text{ for all } x$$

# Probability mass function

From the pmf, we can calculate probabilities of events, as

$$P(a \leq X \leq b) = \sum_{k=a}^{b} f_X(k)$$

for positive integers $a$ and $b$, with $a \leq b$.

What happens if $F_X(x)$ is continuos?

# Probability density function

The *probability density function*, (or pdf), $f_X(x)$, of a continuous random variable $X$ is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \text{ for all } x$$

# Probability density function

The *probability density function*, (or pdf), $f_X(x)$, of a continuous random variable $X$ is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \text{ for all } x$$

$$\frac{d}{dx}F_X(x) = f_X(x)$$

$$P(X \leq x) = F_X(x) =$$

# Probability density function

The *probability density function*, (or pdf), $f_X(x)$, of a continuous random variable $X$ is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \text{ for all } x$$

$$\frac{d}{dx}F_X(x) = f_X(x)$$

$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_x(t)dt$$

# Probability mass/density function

A function $f_X(x)$ is a pdf (or pmf) of a random variable $X$ if and only if

1. $f_X(x) \geq 0$ for all $x$
2. $\sum_x f_X(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (pdf)

# Probability distributions

- Discrete probability distributions include: the uniform discrete, Bernoulli, binomial, geometric and Poisson distributions.
- Univariate continuous probability distributions include the uniform continuous, exponential, Normal, Chi-squared ($\chi^2$), log-normal, Gamma.

# Discrete uniform distribution

For a fair dice all of the six faces have the same probability

$$P(X = i) = \frac{1}{6} \qquad \text{for all } i \in \{1, 2, 3, 4, 5, 6\}.$$

We say that $X$ follows a **discrete uniform distribution** on $\Omega = \{1, 2, 3, 4, 5, 6\}$.

More generally, the discrete uniform distribution is a probability

distribution where a finite number of values are equally likely to be observed; every one of $n$ values has equal probability $1/n$.

Remark: There exists a continuous version of the uniform distribution.

What is the pmf (or pdf?) and cdf?

# Expectations and variance

The expected value of a continuous random variable $X$ is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

The variance of a continuous random variable is

$$Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

# Bernoulli distribution

A **Bernoulli trial** is an experiment that has two outcomes: the sample space is $\Omega = \{\text{success,fail}\}$ generally replaced by $\Omega = \{1, 0\}$.

If $X$ is a random variable following a Bernoulli distribution

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p.$$

A classical example of a Bernoulli experiment is a single toss of a coin. The coin might come up heads with probability $p$ and tails with probability $1 - p$. The experiment is called *fair* if $p = 0.5$.

The expected value of a Bernoulli random variable $X$ is:

$$E[X] = 1 \times p + 0 \times (1 - p) = p$$

The variance of a Bernoulli random variable $X$ is:

$$Var[X] = E[(X - E[X])^2] = (1-p)^2 \times p + (0-p)^2 \times (1-p) = p(1-p)$$

# Binomial distribution

Therefore, the probability of obtaining $k$ successes out of $n$ Bernoulli trials is

$$P(Y = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

We say that the random variable $Y$ follows a **Binomial distribution with parameters** $n$ **and** $p$, denoted by $\mathcal{B}(n, p)$ .

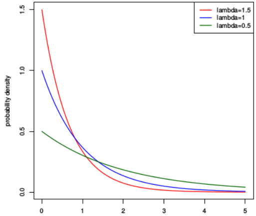
$E[Y] = np$
$Var[Y] = np(1-p)$
The variance is a measure of spread and it increases with $n$ and decreases as $p$ approaches 0 or 1. For a given $n$, the variance is maximized when $p = 0.5$.

Imperial College London

# Continuous - exponential distribution

If $X$ is an exponential random variable then its density function is given by

$$p_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



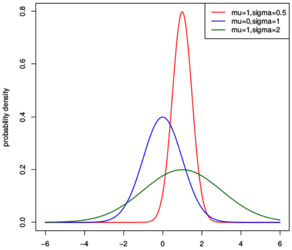

$E[Y] = 1/\lambda$
$Var[Y] = 1/\lambda^2$

# Continuous - Normal distribution

The probability density function of a normal (or Gaussian) random variable is

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

which depends on a location parameter, $\mu$, and a scale parameter, $\sigma$. If $X \sim \mathcal{N}(\mu, \sigma^2)$, $E[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.




- symmetric, bell shape curve
- points of inflection at $\mu - \sigma$ and $\mu + \sigma$
- mean $\mu$ is also the median and the mode

Imperial College London

# Expectations and variance

- If $X$ is a continuous random variable with pdf $f_X(x)$, then for any real-valued function $g$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

- If $a$ and $b$ are constants, $E[aX + b] = aE[X] + b$ and $Var[aX + b] = a^2 Var[X]$