# BIOINFORMATICS & COMPUTATIONAL GENETICS MSc
# PROBABILITY AND STATISTICS

## INTRODUCTION

The random variation associated with "measurement" procedures in a scientific analysis requires a framework in which the **uncertainty** and **variability** that are inherent in the procedure can be handled. The key goal of Probability and Statistical modelling are to establish a mathematical framework within which *random* variation (due to, for example, experimental error or natural variation) can be quantified so that *systematic* variation (arising due to potentially important biological differences) can be studied.

Broadly, the "*Scientific Process*" involves several different stages:

$$
\begin{array}{cl}
\text{THEORETICAL MODELLING} & \longrightarrow \text{MATHEMATICAL} \\
\downarrow & \qquad\text{/PROBABILISTIC MODELLING} \\
\text{PREDICTION} & \\
\downarrow & \\
\text{EXPERIMENTATION/OBSERVATION} & \\
\downarrow & \\
\text{VALIDATION} &
\end{array}
$$

*Mathematical/Probabilistic modelling* facilitates PREDICTION; *Statistical Analysis* provides the means of validation of predicted behaviour.

To explain the variation in observed data, we need to introduce the concept of a *probability distribution*. Essentially we need to be able to model, or specify, or compute the "chance" of observing the data that we collect or expect to collect. This will then allow us to assess how likely the data were to occur by chance alone, that is, how "suprising" the observed data are in light of an assumed theoretical model.

For example, consider two nucleotide sequences of the same length that we wish to assess for similarity:

Sequence 1 $\quad ATAGTAGATACGCACCGAGGA$

Sequence 2 $\quad ATCTTAGATAGGCACTGAGGA$

How can we assess sequence similarity formally ? The number of discordant positions is 4, but how informative is that summary measure. ? Perhaps we need to assess the chance, for example, that a point mutation $A \to C$ occurs (as in the discordant position 3).in unit evolutionary time. Perhaps the chance of observing a sub-sequence

$$ATCTTA$$

rather than

$$ATAGTA$$

(in positions 1-6) in important. Is the hidden (or *latent*) structure in the sequence, corresponding to whether the sequence originates from a coding region or otherwise, important ? Can we even infer the hidden structure in light of the data we have observed ?

These questions can only really be answered when we have an understanding of randomness and variation. The framework that we will use to pose and answer such questions formally is given to us by *probability theory*.

# Chapter 1. Basic Probability Concepts

## 1.1 EXPERIMENTS AND EVENTS

An **experiment** is any procedure

(a) with a well-defined **set** of possible outcomes - the **sample space**, $S$.
(b) whose **actual** outcome is not known in advance.

A **sample outcome**, $s$, is precisely one of the possible outcomes of the experiment.

The **sample space**, $S$, is the entire set of possible outcomes.

**SIMPLE EXAMPLES**:

(a) Coin tossing: $S = \{H, T\}$.

(b) Dice : $S = \{1, 2, 3, 4, 5, 6\}$.

(c) Proportions: $S = \{x : 0 \le x \le 1\}$

(d) Time measurement: $S = \{x : x > 0\} = \mathbb{R}^+$

(e) Temperature measurement: $S = \{x : a \le x \le b\} \subseteq \mathbb{R}$

In biological sequence analysis, the experiment may involve the observation of a nucleotide or protein sequence, so that the sample space $S$ may comprise all sequences (of bases/amino acids) up to a given length, and a sample outcome would be a particular observed sequence.

There are two basic types of experiment, namely

**COUNTING**

and

**MEASUREMENT**

- we shall see that these two types lead to two distinct ways of specifying probability distributions.

The collection of sample outcomes is a **set** (a collection of items) , so we write

$$s \in S$$

if *$s$ is a member of the set $S$*.

**DEFINITION**

An **event** $E$ is a set of the possible outcomes of the experiment, that is $E$ is a **subset** of $S$, $E \subseteq S$, $E$ **occurs** if the actual outcome is in this set.

NOTE: the sets $S$ and $E$ can be either be written as a list of items, for example,

$$E = \{s_1, s_2, ..., s_n, ...\}$$

which may a finite or infinite list, or can only be represented by a continuum of outcomes, for example

$$E = \{x : 0.6 < x \le 2.3\}$$

Events are manipulated using **set theory** notation; if $E$, $F$ are two events, $E, F \subseteq S$,

| | | |
|---|---|---|
| Union | $E \cup F$ | "$E$ **or** $F$ **or both** occurs" |
| Intersection | $E \cap F$ | "$E$ **and** $F$ occur" |
| Complement | $E'$ | "$E$ **does not** occur" |

We can interpret the events $E \cup F$, $E \cap F$, and $E^{'}$ in terms of collections of sample outcomes, and use **Venn Diagrams** to represent these concepts.

Another representation for this two event situation is given by the following table:

| | $E$ | $E'$ | Sum |
|---|---|---|---|
| $F$ | $\mathrm{P}(E \cap F)$ | $\mathrm{P}(E' \cap F)$ | $\mathrm{P}(F)$ |
| $F'$ | $\mathrm{P}(E \cap F')$ | $\mathrm{P}(E' \cap F')$ | $\mathrm{P}(F')$ |
| Sum | $\mathrm{P}(E)$ | $\mathrm{P}(E')$ | |

so that, summing in the columns

$$\begin{aligned} \mathrm{P}(E \cap F) + \mathrm{P}(E \cap F') &= P(E) \\ \mathrm{P}(E' \cap F) + \mathrm{P}(E' \cap F') &= P(E') \end{aligned}$$

and summing in the rows

$$\begin{aligned} P(E \cap F) + P(E' \cap F) &= P(F) \\ P(E \cap F') + P(E' \cap F') &= P(F') \end{aligned}$$

Special cases of events:

| | | |
|---|---|---|
| THE IMPOSSIBLE EVENT | $\emptyset$ | the empty set, the collection of sample outcomes with zero elements |
| THE CERTAIN EVENT | $\Omega$ | the collection of all sample outcomes |

**DEFINITION**

Events $E$ and $F$ are **mutually exclusive** if

$$E \cap F = \emptyset$$

that is, the collections of sample outcomes $E$ and $F$ have no element in common.

## 1.2 RESULTS IN EVENT MANIPULATION

For events $E$, $F$, and $G$, the following equations can be used to simplify complex expressions:

$$\begin{aligned} E \cap (F \cup G) &\equiv (E \cap F) \cup (E \cap G) \\ E \cup (F \cap G) &\equiv (E \cup F) \cap (E \cup G) \\ (E \cup F)' &\equiv E' \cap F' \\ (E \cap F)' &\equiv E' \cup F' \end{aligned}$$

**DEFINITION**

Events $E_1, ..., E_k$ form a **partition** of event $F \subseteq S$ if

(a) $E_i \cap E_j = \emptyset$ for all $i$ and $j$ (b) $\displaystyle\bigcup_{i=1}^{k} E_i = E_1 \cup E_2 \cup ... \cup E_k = F$.

We are interested in mutually exclusive events and partitions because when we carry out probability calculations we will essentially be counting or enumerating sample outcomes; to ease this counting operation, it is desirable to deal with collections of outcomes that are completely distinct or **disjoint**.

## 1.3 THE RULES OF PROBABILITY

The probability function P(.) is a set function that assigns weight to collections of sample outcomes. We can consider assigning probability to an event by adopting

| | |
|---|---|
| CLASSICAL APPROACH | consider equally likely outcomes |
| FREQUENTIST APPROACH | consider long-run relative frequencies |
| SUBJECTIVE APPROACH | consider your personal degree of belief |

It is legitimate to use any justification where appropriate or plausible

**MATHEMATICAL PROPERTIES OF PROBABILITY - THE PROBABILITY AXIOMS**

It is sufficient to require that the probability function P(.) must satisfy the following properties.

For any events $E$ and $F$ in sample space $S$,

(1) $0 \leq \mathrm{P}(E) \leq 1$

(2) $\mathrm{P}(\Omega) = 1$

(3) If $E \cap F = \emptyset$, then $\mathrm{P}(E \cup F) = \mathrm{P}(E) + \mathrm{P}(F)$

From the axioms, we can immediately prove the following results:

$P(E') = 1 - P(E)$, $P(\emptyset) = 0$

If $E_1, ..., E_k$ are events such that $E_i \cap E_j = \emptyset$ for all $i, j$, then

$$P\left(\bigcup_{i=1}^{k} E_i\right) = P(E_1) + P(E_2) + ... + P(E_k).$$

If $E \cap F \neq \emptyset$, then $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

**EXAMPLE CALCULATION** Examination Pass Rates

The examination performance of students in a year of eight hundred students is to be studied: a student either chooses an essay paper or a multiple choice test. The pass figures and rates are given in the table below:

| | PASS | FAIL | PASS RATE |
|---|---|---|---|
| FEMALE | 200 | 200 | 0.5 |
| MALE | 240 | 160 | 0.6 |

The result of this study is clear: the pass rate for MALES is higher than that for FEMALES.

Further investigation revealed a more complex result: for the essay paper, the results were as follows;

| | PASS | FAIL | PASS RATE |
|---|---|---|---|
| FEMALE | 120 | 180 | 0.4 |
| MALE | 30 | 70 | 0.3 |

so for the essay paper, the pass rate for FEMALES is higher than that for MALES.

For the multiple choice test, the results were as follows;

| | PASS | FAIL | PASS RATE |
|---|---|---|---|
| FEMALE | 80 | 20 | 0.8 |
| MALE | 210 | 90 | 0.7 |

so for the multiple choice paper, the pass rate for FEMALES is higher than that for MALES.

Hence we conclude that FEMALES have a higher pass rate on the essay paper, and FEMALES have a higher pass rate on the multiple choice test, but MALES have a higher pass rate overall.

This apparent contradiction can be resolved by careful use of the probability definitions. First introduce notation; let $E$ be the event that the student chooses an essay, $F$ be the event that the student is female, and $G$ be the event that the student passes the selected paper.

## 1.4 CONDITIONAL PROBABILITY

**DEFINITION**

For two events $E$ and $F$ with $\mathrm{P}(F) > 0$, the **conditional probability** that $E$ occurs, **given** that $F$ occurs, is written $\mathrm{P}(E|F)$, and is defined by

$$\mathrm{P}(E|F) = \frac{\mathrm{P}(E \cap F)}{\mathrm{P}(F)} \qquad \text{so that} \qquad \mathrm{P}(E \cap F) = \mathrm{P}(E|F)\mathrm{P}(F)$$

It is easy to show that this new probability operator P( . | . ) satisfies the probability axioms.

[In the exam results problem, what we really have specified are conditional probabilities. From the pooled table, we have

$$\mathrm{P}(G|F) = 0.5 \qquad \mathrm{P}(G|F^{'}) = 0.6,$$

from the essay results table, we have

$$\mathrm{P}(G|E \cap F) = 0.4 \qquad \mathrm{P}(G|E \cap F^{'}) = 0.3,$$

and from the multiple choice table, we have

$$\mathrm{P}(G|E^{'} \cap F) = 0.8 \qquad \mathrm{P}(G|E^{'} \cap F^{'}) = 0.7$$

and so interpretation is more complicated than originally thought.]

The probability of the **intersection** of events $E_1, ..., E_k$ is given by the **chain rule**

$$\mathrm{P}(E_1 \cap ... \cap E_k) \ = \ \mathrm{P}(E_1)\mathrm{P}(E_2|E_1)\mathrm{P}(E_3|E_1 \cap E_2)...\mathrm{P}(E_k|E_1 \cap E_2 \cap ... \cap E_{k-1})$$

**Special Case: Independence**

Events $F$ and $F$ are **independent** if

$$\mathrm{P}(E|F) \ = \ \mathrm{P}(E) \text{ so that } \mathrm{P}(E \cap F) \ = \ \mathrm{P}(E)\mathrm{P}(F)$$

and so if $E_1, ..., E_k$ are independent events, then

$$\mathrm{P}(E_1 \cap ... \cap E_k) = \prod_{i=1}^{k} \mathrm{P}(E_i) = \mathrm{P}(E_1)...\mathrm{P}(E_k)$$

A simple way to think about joint and conditional probability is via a probability tree:

The chain rule construction is particularly important in biological sequence analysis; consider one of the sequences from page 1

$$ATAGTAGATACGCACCGAGGA$$

If we wish to assess the probability of seeing such a sequence, we might let

$$\mathrm{P}(A) = p_A \qquad \mathrm{P}(C) = p_C \qquad \mathrm{P}(G) = p_G \qquad \mathrm{P}(T) = p_T$$

for some suitable probabilities satisfying

$$0 \le p_A, p_C, p_G, p_T \le 1 \qquad p_A + p_C + p_G + p_T = 1$$

and assume independence so that

$$\mathrm{P}(ATAGTAGATACGCACCGAGGA) = p_A \times p_T \times p_A \times ... \times p_G \times p_A$$

which simplifies to

$$\mathrm{P}(ATAGTAGATACGCACCGAGGA) = p_A^8 p_C^4 p_G^6 p_T^3$$

However, the assumption of independence may not be correct; perhaps knowledge about a base being in one position influences the probability of the base in the next position. In this case, we would have to write (in general)

$$\begin{aligned}\mathrm{P}(ATAGTAGATACGCACCGAGGA) = \; & \mathrm{P}(A) \times \mathrm{P}(T|A) \times \mathrm{P}(A|AT) \times \mathrm{P}(G|ATA) \times ... \\ & ... \times \mathrm{P}(A|ATAGTAGATACGCACCGAGG)\end{aligned}$$

Finally, our estimate (or specified value) for $p_A, p_C, p_G, p_T$ may change due to the hidden structure of the underlying genomic segment; that is, whether the segment is from a codon or otherwise; for example

$$\begin{aligned}\mathrm{P}(A|\mathrm{Exon}) &= p_A^{(E)} \\ \mathrm{P}(A|\mathrm{Intron}) &= p_A^{(I)}\end{aligned}$$

where it is not necessarily the case that $p_A = p_A^{(E)} = p_A^{(I)}$.

## 1.5 THE TOTAL PROBABILITY RESULT

If events $E_1, ..., E_k$ form a partition of event $F \subseteq S$, and event $G \subseteq S$ is such that $\mathrm{P}(G) > 0$, then

$$\begin{aligned}\mathrm{P}(F) &= \sum_{i=1}^{k} \mathrm{P}(F|E_i)\mathrm{P}(E_i) \\ \mathrm{P}(F|G) &= \sum_{i=1}^{k} \mathrm{P}(F|E_i \cap G)\mathrm{P}(E_i|G)\end{aligned}$$

The results follows because we have by assumption that

$$F = \bigcup_{i=1}^{k} (E_i \cap F) \Longrightarrow \mathrm{P}(F) = \sum_{i=1}^{k} \mathrm{P}(E_i \cap F) = \sum_{i=1}^{k} \mathrm{P}(F|E_i)\mathrm{P}(E_i)$$

by probability axiom (3), as the collection $E_1 \cap F, ..., E_k \cap F$ are mutually exclusive.

## 1.6 BAYES RULE

For events $E$ and $F$ such that $\mathrm{P}(E), \mathrm{P}(F) > 0$,

$$\mathrm{P}(E|F) = \frac{\mathrm{P}(F|E)\mathrm{P}(E)}{\mathrm{P}(F)}$$

If events $E_1, ..., E_k$ form a partition of $S$, with $\mathrm{P}(E_i) > 0$ for all $i$, then then

$$\mathrm{P}(E_i|F) = \frac{\mathrm{P}(F|E_i)\mathrm{P}(E_i)}{\mathrm{P}(F)} = \frac{\mathrm{P}(F|E_i)\mathrm{P}(E_i)}{\sum_{j=1}^{k}\mathrm{P}(F|E_j)\mathrm{P}(E_j)}$$

Note that this result follows immediately from the conditional probability definition that

$$\mathrm{P}(E \cap F) = \mathrm{P}(E|F)\mathrm{P}(F) \qquad \text{and} \qquad \mathrm{P}(E \cap F) = \mathrm{P}(F|E)\mathrm{P}(E)$$

and hence equating the right hand sides of the two equations we have

$$\mathrm{P}(E|F)\mathrm{P}(F) = \mathrm{P}(F|E)\mathrm{P}(E)$$

and hence the result follows.

Note that in the second part of the theorem,

$$\mathrm{P}(E_i|F) = \frac{\mathrm{P}(F|E_i)\mathrm{P}(E_i)}{\mathrm{P}(F)} = \frac{\mathrm{P}(F|E_i)}{\mathrm{P}(F)}\,\mathrm{P}(E_i)$$

so the probabilities $\mathrm{P}(E_i)$ are re-scaled to $\mathrm{P}(E_i|F)$ by conditioning on $F$. Note that

$$\sum_{i=1}^{k} \mathrm{P}(E_i|F) = 1$$

This theorem is very important because, in general,

$$\mathrm{P}(E|F) \neq \mathrm{P}(F|E)$$

and it is crucial to condition on the correct event in a conditional probability calculation.

**EXAMPLE** Lie-detector test.

In an attempt to achieve a criminal conviction, a lie-detector test is used to determine the guilt of a suspect. Let $G$ be the event that the suspect is guilty, and let $T$ be the event that the suspect fails the test.

The test is regarded as a good way of determining guilt, because laboratory testing indicate that the detection rates are high; for example it is known that

$$\text{P[ Suspect Fails Test | Suspect is Guilty ]} = \mathrm{P}(T|G) = 0.95 = 1 - \alpha, \text{ say}$$

$$\text{P[ Suspect Passes Test | Suspect is Not Guilty ]} = \mathrm{P}(T'|G') = 0.99 = \beta, \text{ say.}$$

Suppose that the suspect fails the test. What can be concluded ? The probability of real interest is $\mathrm{P}(G|T)$; we do not have this probability but can compute it using Bayes Theorem.

For example, we have

$$\mathrm{P}(G|T) = \frac{\mathrm{P}(T|G)\mathrm{P}(G)}{\mathrm{P}(T)}$$

where $\mathrm{P}(G)$ is not yet specified, but $\mathrm{P}(T)$ can be computed using the Theorem of Total probability, that is,

$$\mathrm{P}(T) = \mathrm{P}(T|G)\mathrm{P}(G) + \mathrm{P}(T|G')\mathrm{P}(G')$$

so that

$$\mathrm{P}(G|T) = \frac{\mathrm{P}(T|G)\mathrm{P}(G)}{\mathrm{P}(T|G)\mathrm{P}(G) + \mathrm{P}(T|G')\mathrm{P}(G')}$$

Clearly, the probability $\mathrm{P}(G)$, the probability that the suspect is guilty *before* the test is carried out, plays a crucial role. Suppose, that $\mathrm{P}(G) = p = 0.005$, so that only 1 in 200 suspects taking the test are guilty. Then

$$\mathrm{P}(T) = 0.95 \times 0.005 + 0.01 \times 0.995 = 0.0147$$

so that

$$\mathrm{P}(G|T) = \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} = 0.323$$

which is still relatively small. So, as a result of the lie-detector test being failed, the probability of guilt of the suspect has increased from 0.005 to 0.323.

More extreme examples can be found by altering the values of $\alpha$, $\beta$ and $p$.