# Bioinformatics MSc
# Mathematics, Probability and Statistics

D. A. Stephens
Department of Mathematics
Imperial College London

November 2004

# Contents

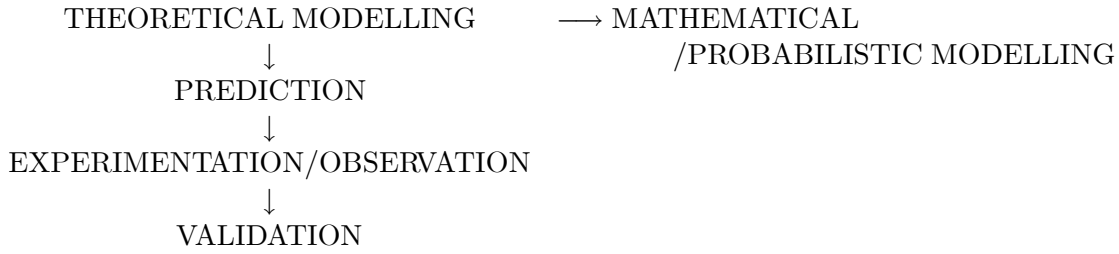
# CHAPTER 1

# PROBABILITY THEORY

## 1.1 INTRODUCTION AND MOTIVATION

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 is to establish a mathematical framework within which *random* variation (due, for example, to 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:




*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 "surprising" 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 $ATAGTAGATACGCACCGAGGA$

Sequence 2 $ATCTTAGATAGGCACTGAGGA$

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 $A \rightarrow C$ 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.*

## 1.2 BASIC PROBABILITY CONCEPTS

### 1.2.1 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: $S = \{x : 0 \leq x \leq 1\}$
(d) Time measurement: $S = \{x : x > 0\} = \mathbb{R}^+$
(e) Temperature measurement: $S = \{x : a \leq x \leq b\} \subseteq \mathbb{R}$

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 1.2.1** 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

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

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.

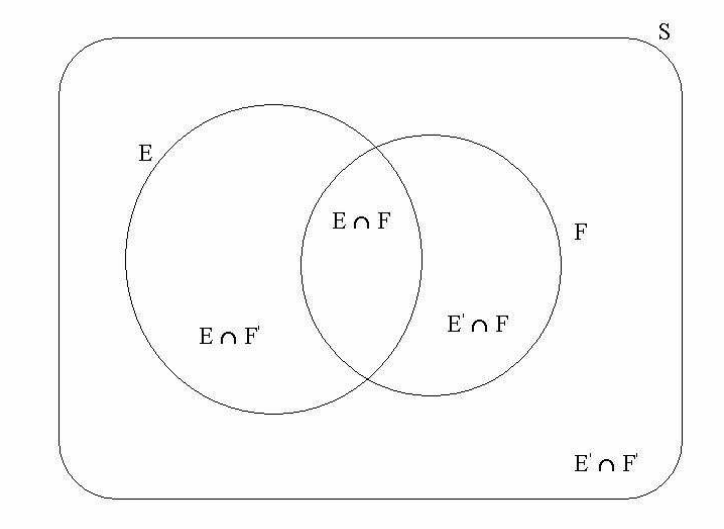



Figure 1.1: Venn Diagram

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

| | $E$ | $E'$ | Union |
|---|---|---|---|
| $F$ | $(E \cap F)$ | $(E' \cap F)$ | $F$ |
| $F'$ | $(E \cap F')$ | $(E' \cap F')$ | $F'$ |
| Union | $E$ | $E'$ | |

so that, taking unions in the columns

$$(E \cap F) \cup (E \cap F') = E$$

$$(E' \cap F) \cup (E' \cap F') = E'$$

and in the rows

$$(E \cap F) \cup (E' \cap F) = F$$

$$(E \cap F') \cup (E' \cap F') = F'$$

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 1.2.2** 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.

Mutually exclusive events are very important in probability and statistics, as they allow complicated events to be simplified in such a way as to allow straightforward probability calculations to be made.

**Definition 1.2.3** 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) $\bigcup_{i=1}^{k} E_i = E_1 \cup E_2 \cup ... \cup E_k = F$.

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**.

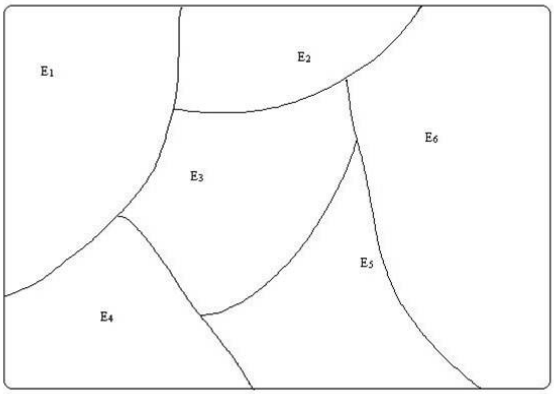



Figure 1.2: Partition of $\Omega$

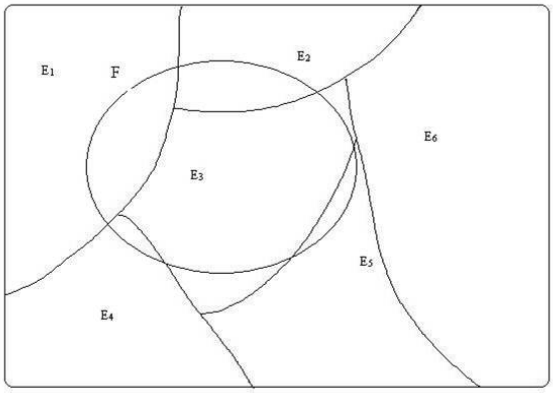



Figure 1.3: Partition of $F \subset \Omega$

## 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. In fact, it is sufficient to require that the probability function $P(.)$ must satisfy the following mathematical properties. For any events $E$ and $F$ in sample space $S$,

(1) $0 \leq P(E) \leq 1$

(2) $P(\Omega) = 1$

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

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 $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

For the $2 \times 2$ table above, we have the following:

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

so that, summing in the columns

$$P(E \cap F) + P(E \cap F') = P(E)$$

$$P(E' \cap F) + P(E' \cap F') = P(E')$$

and summing in the rows

$$P(E \cap F) + P(E' \cap F) = P(F)$$

$$P(E \cap F') + P(E' \cap F') = P(F')$$

**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 AND INDEPENDENCE

**Definition 1.4.1** For two events $E$ and $F$ with $P(F) > 0$, the **conditional probability** that $E$ occurs, **given** that $F$ occurs, is written $P(E|F)$, and is defined by

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

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

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

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

**Definition 1.4.2** Events $F$ and $F$ are **independent** if

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

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

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

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

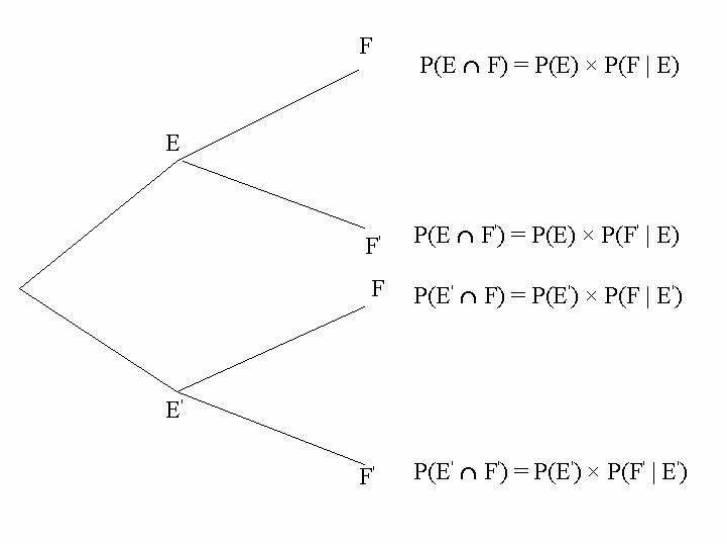



Figure 1.4: Probability Tree for the Theorem of Total Probability

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

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

for some suitable probabilities satisfying

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

and assume independence so that

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

which simplifies to

$$P(ATAGTAGATACGCACCGAGGA) = p_A^8 p_C^4 p_G^6 p_T^3$$

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)

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

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

$$P(A|Exon) \quad = p_A^{(E)}$$

$$P(A|Intron) \quad = p_A^{(I)}$$

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

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

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

from the essay results table, we have

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

and from the multiple choice table, we have

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

and so interpretation is more complicated than originally thought.]

## 1.5 THE TOTAL PROBABILITY RESULT

If events $E_1, ..., E_k$ form a partition of $S$, and event $F \subseteq S$, then

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

The results follows because we have by assumption that

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

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 $P(E)$, $P(F) > 0$,

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

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

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

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

$$P(E \cap F) = P(E|F)P(F) \qquad and \qquad P(E \cap F) = P(F|E)P(E)$$

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

$$P(E|F)P(F) = P(F|E)P(E)$$

and hence the result follows. Note also that in the second part of the theorem,

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

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

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

This theorem is very important because, in general,

$$P(E|F) \neq P(F|E)$$

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

**Example 1.6.1 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

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

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

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

For example, we have

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

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

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

so that

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

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

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

so that

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

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 $p$.

# CHAPTER 2

# PROBABILITY DISTRIBUTIONS

## 2.1 MOTIVATION

The probability definitions, rules, and theorems given previously are all framed in terms of events in a sample space. For example, for an experiment with possible sample outcomes denoted by the *sample space* $S$, an *event* $E$ was defined as any collection of sample outcomes, that is, any subset of the set $S$.

$$\begin{array}{lllll} \text{EXPERIMENT} & \longrightarrow \text{SAMPLE OUTCOMES} & \longrightarrow \text{EVENTS} & \longrightarrow \text{PROBABILITIES} \\ & S = \{s_1, s_2, ...\} & \longrightarrow E \subseteq S & \longrightarrow P(E) \end{array}$$

In this framework, it is necessary to consider each experiment with its associated sample space separately - the nature of sample space $S$ is typically different for different experiments.

**EXAMPLE 1**: Count the number of days in February which have zero precipitation.
**SAMPLE SPACE** $S = \{0, 1, 2, ..., 28\}$.Let $E_i =$ "$i$ days have zero precipitation"; $E_0, ..., E_{28}$ partition $S$.

**EXAMPLE 2**: Count the number of goals in a football match.
**SAMPLE SPACE**: $S = \{0, 1, 2, 3, ...\}$. Let $E_i =$ "$i$ goals in the match"; $E_0, E_1, E_2, ...$ partition $S$

In both of these examples, we need a formula to specify each $P(E_i) = p_i$.

**EXAMPLE 3** Measure the operating temperature of an experimental process.
**SAMPLE SPACE**: $S = \{\ x\ :\ x > T_{min}\ \}$.

Here it is difficult to express

$$P[\text{ "Measurement is } \ x \text{ " }]$$

but possible to think about

$$P[\text{ "Measurement is } \ \leq x \text{ " }] = F(x), \text{ say,}$$

and now we seek a formula for $F(x)$.

A general notation useful for all such examples can be obtained by considering a sample space that is **equivalent** to $S$ for a general experiment, but whose form is more familiar. For example, for a general sample space $S$, if it were possible to associate a subset of the **integer** or **real** number systems, $\mathbb{X}$ say, with $S$, then attention could be restricted to considering events in $\mathbb{X}$, whose structure

is more convenient, as then *$S$ are collections of sample outcomes events in $\mathbb{X}$ are intervals of the real numbers* For example, consider an experiment involving counting the number of breakdowns of a production line in a given month. The experimental sample space $S$ is therefore the collection of sample outcomes $s_0, s_1, s_2, ...$ where $s_i$ is the outcome "there were $i$ breakdowns"; events in $S$ are collections of the $s_i$s. Then a useful equivalent sample space is the set $\mathbb{X} = \{0, 1, 2, ...\}$, and events in $\mathbb{X}$ are collections of non-negative integers. Formally, therefore, we seek a function or **map** from $S$ to $\mathbb{X}$. This map is known as a **random variable**.

## 2.2 RANDOM VARIABLES

A **random variable** $X$ is a function from experimental sample space $S$ to some set of real numbers $\mathbb{X}$ that maps $s \in S$ to a unique $x \in \mathbb{X}$

$$\begin{array}{rcl} X: & S \longrightarrow \mathbb{X} \subseteq \mathbb{R} \\ & s \longmapsto x \end{array}$$

**Interpretation** A random variable is a way of describing the outcome of an experiment in terms of real numbers.

| | RANDOM VARIABLE | TO BE SPECIFIED |
|---|---|---|
| **EXAMPLE 1** | $X =$"No. days in Feb. with zero precipitation" | $P[\ X = x\ ]$ for $x = 0, 1, 2, ..., 28$ |
| **EXAMPLE 2** | $X =$"No. goals in a football match" | $P[\ X = x\ ]$ for $x = 0, 1, 2, 3, ...$ |
| **EXAMPLE 3** | $X =$"the measured operating temperature" | $P[\ X \leq x\ ]$ for $x > T_{min}$. |

Therefore $X$ is merely the count/number/measured value corresponding to the outcome of the experiment.

Depending on the type of experiment being carried out, there are two possible forms for the set of values that $X$ can take:

A random variable is **DISCRETE** if the set $\mathbb{X}$ is of the form

$$\mathbb{X} = \{x_1, x_2, ..., x_n\} \text{ or } \mathbb{X} = \{x_1, x_2, ...\},$$

that is, a finite or infinite set of **distinct** values $x_1, x_2, ..., x_n, ....$ Discrete random variables are used to describe the outcomes of experiments that involve **counting** or **classification**.

A random variable is **CONTINUOUS** if the set $\mathbb{X}$ is of the form

$$\mathbb{X} = \bigcup_i \{x : a_i \leq x \leq b_i\}$$

for real numbers $a_i, b_i$, that is, the union of **intervals** in $\mathbb{R}$. Continuous random variables are used to describe the outcomes of experiments that involve **measurement**.

## 2.3 PROBABILITY DISTRIBUTIONS

A **probability distribution** is a function that assigns probabilities to the possible values of a random variable. When specifying a probability distribution for a random variable, two aspects need to be considered. First, the range of the random variable (that is, the values of the random variable which have positive probability) must be specified. Secondly, the method via which the probabilities are assigned to different values in the range must be specified; typically this is achieved by means of a function or formula.
In summary, we need to find a function or formula via which

$$P[\ X = x\ ] \qquad \text{or} \qquad P[\ X \leq x\ ]$$

can be calculated for each $x$ in a suitable range $\mathbb{X}$. The functions used to specify these probabilities are just real-valued functions of a single real argument, similar to polynomial, exponential, logarithmic or trigonometric functions such as (for example)

$$\begin{aligned} f(x) &= 6x^3 - 3x^2 + 2x - 5 \\ f(x) &= e^x \\ f(x) &= \sin(x) + 2x\cos(2x) \end{aligned}$$

and so on. However, the fundamental rules of probability mean that the functions specifying $P[\ X = x\ ]$ or $P[\ X \leq x\ ]$ must exhibit certain properties. As we shall see below, the properties of these functions, and how they are manipulated mathematically, depend crucially on the nature os the random variable.

## 2.4 DISCRETE PROBABILITY DISTRIBUTIONS

For discrete random variables there are two routes via which the probability distribution can be specified.

### 2.4.1 PROBABILITY MASS FUNCTION

The probability distribution of a *discrete* random variable $X$ is described by the **probability mass function (pmf)** $f_X$, specified by

$$f_X(x) = P[X = x] \qquad \text{for } x \in \mathbb{X} = \{x_1, x_2, ..., x_n, ...\}$$

Because of the probability axioms, the function $f_X$ must exhibit the following properties:

$$(i) f_X(x_i) \geq 0 \quad \text{for all } i \qquad\qquad (ii) \sum_i f_X(x_i) = 1.$$

### 2.4.2 DISCRETE CUMULATIVE DISTRIBUTION FUNCTION

The **cumulative distribution function** or **cdf**, $F_X$, is defined by

$$F_X(x) = P[X \leq x] \qquad for\ x \in \mathbb{R}$$

The cdf $F_X$ must exhibit the following properties:

(i) $\lim_{x \longrightarrow -\infty} F_X(x) = 0$

(ii) $\lim_{x \longrightarrow \infty} F_X(x) = 1$

(iii) $\lim_{h \to 0^+} F_X(x+h) = F_X(x)$ [i.e. $F_X$ is continuous from the right]

(iv) $a < b \Longrightarrow F_X(a) \leq F_X(b)$ [i.e. $F_X$ is non-decreasing]

(v) $P[a < X \leq b] = F_X(b) - F_X(a)$

The cumulative distribution function defined in this way is a "**step function**".

The functions $f_X$ and/or $F_X$ can be used to describe the **probability distribution** of random variable $X$.

**EXAMPLE** An electrical circuit comprises six fuses.

let $X =$"number of fuses that fail within one month". Then

$$\mathbb{X} = \{0, 1, 2, 3, 4, 5, 6\}$$

To specify the probability distribution of $X$, can use the mass function $f_X$ or the cdf $F_X$. For example,

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $f_X(x)$ | $\frac{1}{16}$ | $\frac{2}{16}$ | $\frac{4}{16}$ | $\frac{4}{16}$ | $\frac{2}{16}$ | $\frac{2}{16}$ | $\frac{1}{16}$ |
| $F_X(x)$ | $\frac{1}{16}$ | $\frac{3}{16}$ | $\frac{7}{16}$ | $\frac{11}{16}$ | $\frac{13}{16}$ | $\frac{15}{16}$ | $\frac{16}{16}$ |

as $F_X(0) = P[\ X \leq 0\ ] = P[\ X = 0\ ] = f_X(0)$, $F_X(1) = P[\ X \leq 1\ ] = P[\ X = 0\ ] + P[\ X = 1\ ] = f_X(0) + f_X(1)$, and so on. Note also that, for example,

$$P[\ X \leq 2.5\ ] \equiv P[\ X \leq 2\ ]$$

as the random variable $X$ only takes values $0, 1, 2, 3, 4, 5, 6$.

**EXAMPLE** A computer is prone to crashes.

Suppose that $P[$ "Computer crashes on any given day" $] = \theta$, for some $0 \leq \theta \leq 1$, independently of crashes on any other day.

Let $X =$"number of days until the first crash". Then

$$\mathbb{X} = \{1, 2, 3, ...\}$$

To specify the probability distribution of $X$, can use the mass function $f_X$ or the cdf $F_X$. Now,

$$f_X(x) = P[\ X = x\ ] = (1-\theta)^{x-1}\theta$$

for $x = 1, 2, 3, ...$ (if the first crash occurs on day $x$, then we must have a sequence of $x-1$ crash-free days, followed by a crash on day $x$). Also

$$F_X(x) = P[\ X \leq x\ ] = P[\ X = 1\ ] + P[\ X = 2\ ] + ... + P[\ X = x\ ] = 1 - (1-\theta)^x$$

as the terms in the summation are merely a geometric progression with first term $\theta$ and common term $1-\theta$.

### 2.4.3 RELATIONSHIP BETWEEN $f_X$ AND $F_X$

The fundamental relationship between $f_X$ and $F_X$ is obtained by noting that if $x_1 \leq x_2 \leq ... \leq x_n \leq ...$, then

$$P[X \leq x_i] = P[X = x_1] + ... + P[X = x_i],$$

so that

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i)\ ,$$

and

$$f_X(x_1) = F_X(x_1) \qquad\qquad f_X(x_i) = F_X(x_i) - F_X(x_{i-1})$$

$$f_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \quad \text{for } i \geq 2$$

so $P[c_1 < X \leq c_2] = F_X(c_2) - F_X(c_1)$ for any real numbers $c_1 < c_2$.

Hence, in the discrete case, we can calculate $F_X$ from $f_X$ by **summation**, and calculate $f_X$ from $F_X$ by **differencing**.

## 2.5 SPECIAL DISCRETE PROBABILITY DISTRIBUTIONS

Discrete probability models are used to model the outcomes of counting experiments. Depending on the experimental situation, it is often possible to justify the use of one of a class of "Special" discrete probability distributions. These are listed in this chapter, and are all motivated from the central concept of a *binary* or 0-1 trial, where the random variable concerned has range consisting of only two values with associated probabilities $\theta$ and $1-\theta$ respectively; typically we think of the possible outcomes as "successes" and "failures". All of the distributions in this section are derived by making different modelling assumptions about sequences of 0-1 trials.

| | |
|---|---|
| Single 0-1 trial - count number of 1s | $\Longrightarrow$ BERNOULLI DISTRIBUTION |
| $n$ independent 0-1 trials - count number of 1s | $\Longrightarrow$ BINOMIAL DISTRIBUTION |
| Sequence of independent 0-1 trials - count number of trials until first 1 | $\Longrightarrow$ GEOMETRIC DISTRIBUTION |
| Sequence of independent 0-1 trials - count number of trials until $n^{th}$ 1 is observed | $\Longrightarrow$ NEGATIVE BINOMIAL DISTRIBUTION |
| Limiting case of binomial distribution | $\Longrightarrow$ POISSON DISTRIBUTION |

**Definition 2.5.1 THE DISCRETE UNIFORM DISTRIBUTION**:

This model gives **equal** probability to each of the $N$ possible values of the random variable, and

$$f_X(x) = \frac{1}{N} \qquad for 1 \leq x \leq N$$

and zero otherwise

**Definition 2.5.2 THE BERNOULLI DISTRIBUTION**

This model gives probabilities for the number of successes in a single 0-1 trial where the probability of success is $\theta$.

$$f_X(x) = \theta^x(1-\theta)^{1-x} \qquad x \in \{0,1\}$$

**Definition 2.5.3 THE BINOMIAL DISTRIBUTION**

This model gives probabilities for the number of successes in $n$ 0-1 trials, where the probability of success is $\theta$

$$f_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} \qquad x \in \{0,1,...,n\}$$

where $n! = n \times (n-1) \times (n-2) \times ... \times 3 \times 2 \times 1$.

Note that the sum of $n$ independent and identically distributed (i.i.d.) $Bernoulli(\theta)$ random variables has a Binomial distribution, that is,

$$If\ X_1,...,X_n \sim Bernoulli(\theta) \text{ i.i.d., then } X = \sum_{i=1}^{n} X_i \sim Binomial(n,\theta)$$

**Definition 2.5.4 THE GEOMETRIC DISTRIBUTION**

A model for the number of 0-1 trials until the first success is obtained

$$f_X(x) = (1-\theta)^{x-1}\theta \qquad x \in \{1, 2, ...\}$$

where the probability of success on each trial is $\theta$. a **discrete waiting-time** model. Sometimes this model is specified alternately for the set $\{0, 1, 2, ...\}$ by

$$f_X(x) = (1-\theta)^{x}\theta \qquad x \in \{0, 1, 2, ...\}$$

and sometimes in the parameterization $\phi = 1 - \theta$ as

$$f_X(x) = \phi^x (1-\phi) \qquad x \in \{0, 1, 2, ...\}$$

**Definition 2.5.5 THE NEGATIVE BINOMIAL DISTRIBUTION**

A model for the number of 0-1 trials until the $n$th success is obtained, where the probability of success on each trial is $\theta$.

$$f_X(x) = \binom{x-1}{n-1}\theta^n(1-\theta)^{x-n} \qquad x \in \{n, n+1, n+2, ...\}.$$

The sum of $n$ i.i.d. $Geometric(\theta)$ random variables has a Negative Binomial distribution, that is,

$$\text{If } X_1, ..., X_n \sim Geometric(\theta) \qquad \text{with } X_1, ..., X_n \text{i.i.d, then} \qquad X = \sum_{i=1}^{n} X_i \sim NegBin(n, \theta)$$

that is, the number of trials until the $n$th 1 is the sum of the number of trials until the first 1, plus the number of trials between the first and second 1, etc. For this reason, the negative binomial distribution is also known as the **GENERALIZED GEOMETRIC** distribution.

**Definition 2.5.6 THE POISSON DISTRIBUTION**

A limiting case of the binomial distribution;

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x \in \{0, 1, 2, ...\}$$

That is, if we write $\theta = \lambda/n$, and then consider a limiting case as $n \longrightarrow \infty$, then

$$f_X(x) = \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\frac{n!}{(n-x)!}\left(\frac{1}{n-\lambda}\right)^x \longrightarrow \frac{\lambda^x}{x!}e^{-\lambda}$$

as $n \longrightarrow \infty$ with $\lambda$ constant. That is, we have that

$$Binomial(n, \theta) \longrightarrow Poisson(\lambda)$$

.in this limiting case. The Poisson model is appropriate for count data, where the number of events (accidents, breakdowns etc) that occur in a given time period are being counted.

It is also related to the **Poisson process**; consider a sequence of events occurring **independently** and **at random** in time at a **constant rate** $\lambda$ per unit time. Let $X(t)$ be the random variable defined for $t > 0$ by

"$X(t) = x$" if and only if "$x$ events occur in time interval $[0, t)$".

Then $X(t) \sim Poisson(\lambda t)$, so that

$$f_{X(t)}(x) = P[X(t) = x] = \frac{e^{-(\lambda t)}(\lambda t)^x}{x!} \qquad x \in \{0, 1, 2, ...\}.$$

## 2.6 CONTINUOUS PROBABILITY DISTRIBUTIONS

### 2.6.1 CONTINUOUS CUMULATIVE DISTRIBUTION FUNCTION

The probability distribution of a *continuous* random variable $X$ is defined by the continuous **cumulative distribution function** or **c.d.f.**, $F_X$, specified by

$$F_X(x) = P[X \leq x] \qquad \text{for all } x \in \mathbb{X}$$

that is, an identical definition to the discrete case.

The continuous cdf $F_X$ must exhibit the same properties: as for the discrete cdf, except

(iii) $\lim_{h \to 0} F_X(x+h) = F_X(x)$ [i.e. $F_X$ is continuous]

### 2.6.2 PROBABILITY DENSITY FUNCTION

The **probability density function**, or **pdf**, $f_X$, is defined by

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

so that, by a fundamental calculus result,

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

The pdf $f_X$ must exhibit the following properties:

$$(i) f_X(x) \geq 0\ for x \in \mathbb{X} \qquad (ii) \int_{\mathbb{X}} f_X(x)\ dx = 1.$$

In the continuous case, we calculate $F_X$ from $f_X$ by **integration**, and $f_X$ from $F_X$ by **differentiation**.

In both discrete and continuous cases, $F_X(x)$ id defined for all $x \in \mathbb{R}$, and $f_X(x)$ also defined for all $x$ but may be zero for some values of $x$. Also, if $X$ is continuous, we have

$$P[\ a \leq X \leq b\ ] = F_X(b) - F_X(a) \longrightarrow 0$$

as $b \longrightarrow a$. Hence, for each $x$, we must have

$$P[\ X = x\ ] = 0$$

if $X$ is continuous. Therefore must use $F_X$ to specify the probability distribution initially, although it is often easier to think of the "shape" of the distribution via the pdf $f_X$. Any function that satisfies the properties for a pdf can be used to construct a probability distribution. Note that, for a continuous random variable

$$f_X(x) \neq P[X = x].$$

## 2.7 SPECIAL CONTINUOUS PROBABILITY DISTRIBUTIONS

Here is a list of probability models are used in standard modelling situations. Unlike the discrete case, there are not really any explicit links between most of them, although some connections can be made by means of "transformation" from one variable to another.

### Definition 2.7.1 THE CONTINUOUS UNIFORM DISTRIBUTION

A model with **constant** probability density on a region,

$$f_X(x) = \frac{1}{b-a} \qquad a < x < b$$

the cumulative distribution function (cdf) is also straightforward

$$F_X(x) = \frac{x-a}{b-a} \qquad a < x < b$$

### Definition 2.7.2 THE EXPONENTIAL DISTRIBUTION

A **continuous** waiting-time model

$$f_X(x) = \lambda e^{-\lambda x} \qquad x \in \mathbb{R}^+$$

The cdf for the exponential distribution can be calculated easily;

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt = \int_0^x \lambda e^{-\lambda t}\ dt = 1 - e^{-\lambda x} \qquad x \geq 0.$$

and note that

$$P[\ X > x\ ] = 1 - P[\ X \leq x\ ] = 1 - F_X(x) = e^{-\lambda x}$$

which may give some motivation for using the Exponential model in practice.

One important property of the Exponential distribution is the **lack of memory** property which is defined as follows; consider the **conditional** probability that random variable $X$ is greater than $x_0 + x$, given that $X$ is greater than $x_0$; such a conditional probability is important if we are trying to assess, say, the probability that, if a component functions without failure for a time $x_0$, it continues to function without failure until $x_0 + x$. From the conditional probability definition, we have

$$\begin{aligned} P\left[X > x_0 + x | X > x_0\right] &= \frac{P\left[(X > x_0 + x) \cap (X > x_0)\right]}{P\left[X > x_0\right]} \\ &= \frac{P\left[X > x_0 + x\right]}{P\left[X > x_0\right]} \\ &= \frac{e^{-\lambda(x_0+x)}}{e^{-\lambda x_0}} = e^{-\lambda x} \\ &= P\left[X > x\right] \end{aligned}$$

so that the component has "forgotten" the time period up to $x_0$.

**Definition 2.7.3 THE GAMMA DISTRIBUTION**

An extension to the Exponential model

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \in \mathbb{R}^+$$

where $\Gamma(\alpha)$ is a "special function" known as the **Gamma Function**. It can be computed for any $\alpha > 0$.

The Gamma distribution is another **continuous** waiting-time model. It can be shown the sum of i.i.d. Exponential random variables has a Gamma distribution, that is, if $X_1, X_2, ..., X_n$ are independent and identically distributed $Exponential(\lambda)$ random variables, then

$$X = \sum_{i=1}^{n} X_i \sim Gamma(n, \lambda)$$

Notes :

(1) If $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.; If $\alpha = 1, 2, ...$, $\Gamma(\alpha) = (\alpha - 1)!$.

(2) $\Gamma(1/2) = \sqrt{\pi}$.

(3) If $\alpha = 1, 2, ...$, then the $Gamma(\alpha/2, 1/2)$ distribution is known as the **Chi-squared distribution** with $\alpha$ **degrees of freedom**, denoted $\chi^2_\alpha$.

**Definition 2.7.4 THE NORMAL DISTRIBUTION**

A probability model that reflects observed (**empirical**) behaviour of data samples; this distribution is often observed in practice.

$$f_X(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \qquad x \in \mathbb{R}.$$

The pdf is symmetric about $\mu$, and hence $\mu$ is controls the *location* of the distribution and $\sigma^2$ controls the *spread* or *scale* of the distribution.

Notes :

(1) The Normal density function is justified by the **Central Limit Theorem**.

(2) Special case: $\mu = 0, \sigma^2 = 1$ - the **standard** or **unit** normal distribution. In this case, the density function is denoted $\phi(x)$, and the cdf is denoted $\Phi(x)$ so that

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt = \int_{-\infty}^{x} \left(\frac{1}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}t^2\right\}\, dt.$$

This integral can only be calculated numerically.

(3) If $X \sim N(0,1)$, and $Y = \sigma X + \mu$, then $Y \sim N(\mu, \sigma^2)$.

(4) If $X \sim N(0,1)$, and $Y = X^2$, then $Y \sim Gamma(1/2, 1/2) = \chi^2_1$.

(5) If $X \sim N(0,1)$ and $Y \sim \chi^2_\alpha$ are independent random variables, then random variable $T$, defined by

$$T = \frac{X}{\sqrt{Y/\alpha}}$$

has a **Student-t distribution** with $\alpha$ **degrees of freedom**. The Student-t distribution plays an important role in certain statistical testing procedures.

## 2.8 EXPECTATION AND VARIANCE

For a **discrete** random variable $X$ taking values in set $\mathbb{X}$ with mass function $f_X$, the **expectation** of $X$ is defined by

$$E_{f_X}[\ X\ ] = \sum_{x \in \mathbb{X}} x f_X(x) \equiv \sum_{x=-\infty}^{\infty} x f_X(x)$$

For a **continuous** random variable $X$ with pdf $f_X$, the expectation of $X$ is defined by

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

The **variance** of $X$ is defined by

$$Var_{f_X}[X] = E_{f_X}[(X - E_{f_X}[X])^2] = E_{f_X}\left[X^2\right] - \left\{E_{f_X}[X]\right\}^2.$$

**Interpretation :** The expectation and variance of a probability distribution can be used to aid description, or to characterize the distribution; the EXPECTATION is a measure of **location** (that is, the "centre of mass" of the probability distribution. The VARIANCE is a measure of **scale** or **spread** of the distribution (how widely the probability is distributed) .

**EXAMPLE** Suppose that $X$ is a discrete Poisson random variable taking values on $\mathbb{X} = \{0, 1, 2, ...\}$ with pdf

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0, 1, 2, ...$$

and zero otherwise. Then

$$E_{f_X}[\ X\ ] \ = \sum_{x=-\infty}^{\infty} x f_X(x) = \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

using the power series expansion definition for the exponential function

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + ...$$

**EXAMPLE** Suppose that $X$ is a continuous random variable taking values on $\mathbb{X} = \mathbb{R}^+$ with pdf

$$f_X(x) = \frac{2}{(1+x)^3} \qquad x > 0.$$

Then, integrating by parts.

$$\begin{aligned} E_{f_X}[\ X\ ] \ &= \ \int_{-\infty}^{\infty} x f_X(x)\ dx = \int_0^{\infty} \frac{2x}{(1+x)^3}\ dx = \left[-\frac{x}{(1+x)^2}\right]_0^{\infty} + \int_0^{\infty} \frac{1}{(1+x)^2}\ dx \\ &= \ 0 - \left[-\frac{1}{1+x}\right]_0^{\infty} = 1 \end{aligned}$$

### 2.8.1 EXPECTATIONS OF SUMS OF RANDOM VARIABLES:

Suppose that $X_1$ and $X_2$ are independent random variables, and $a_1$ and $a_2$ are constants. Then if $Y = a_1X_1 + a_2X_2$, it can be shown that

$$E_{f_Y}[Y] = a_1E_{f_{X_1}}[X_1] + a_2E_{f_{X_2}}[X_2]$$

$$Var_{f_Y}[Y] = a_1^2Var_{f_{X_1}}[X_1] + a_2^2Var_{f_{X_2}}[X_2]$$

so that, in particular (when $a_1 = a_2 = 1$) we have

$$E_{f_Y}[Y] = E_{f_{X_1}}[X_1] + E_{f_{X_2}}[X_2]$$

$$Var_{f_Y}[Y] = Var_{f_{X_1}}[X_1] + Var_{f_{X_2}}[X_2]$$

so we have a simple additive property for expectations and variances. Note also that if $a_1 = 1, a_2 = -1$, then

$$E_{f_Y}[Y] = E_{f_{X_1}}[X_1] - E_{f_{X_2}}[X_2]$$

$$Var_{f_Y}[Y] = Var_{f_{X_1}}[X_1] + Var_{f_{X_2}}[X_2]$$

Sums of random variables crop up naturally in many statistical calculations. Often we are interested in a random variable $Y$ that is defined as the sum of some other **independent and identically distributed** (i.i.d) random variables, $X_1, ..., X_n$. If

$$Y = \sum_{i=1}^{n} X_i \qquad with \qquad E_{f_{X_i}}[X_i] = \mu \quad and \quad Var_{f_{X_i}}[X_i] = \sigma^2$$

we have

$$E_{f_Y}[Y] = \sum_{i=1}^{n} E_{f_{X_i}}[X_i] = \sum_{i=1}^{n} \mu = n\mu \qquad and \qquad Var_{f_Y}[Y] = \sum_{i=1}^{n} Var_{f_{X_i}}[X_i] = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

and also, if

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{is the } \textbf{sample mean} \text{ random variable}$$

then, using the properties listed above

$$E_{f_{\overline{X}}}\left[\overline{X}\right] = \frac{1}{n}E_{f_Y}[Y] = \frac{1}{n}n\mu = \mu \qquad and \qquad Var_{f_Y}[Y] = \frac{1}{n^2}Var_{f_Y}[Y] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

### 2.8.2 EXPECTATIONS OF A FUNCTION OF A RANDOM VARIABLE

Suppose that $X$ is a random variable, and $g(.)$ is some function. Then we can define the expectation of $g(X)$ (that is, the expectation of a function of a random variable) by

$$E_{f_X}[\ g(X)\ ] = \begin{cases} \sum_{x=-\infty}^{\infty} g(x)f_X(x) & \text{DISCRETE CASE} \\ \int_{-\infty}^{\infty} g(x)f_X(x)\ dx & \text{CONTINUOUS CASE} \end{cases}$$

For example, if $X$ is a continuous random variable, and $g(x) = \exp\{-x\}$ then

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

Note that $Y = g(X)$ is also a random variable whose probability distribution we can calculate from the probability distribution of $X$.

### 2.8.3 RESULTS FOR STANDARD DISTRIBUTIONS

The expectations and variances for the special distributions described in previous sections are as follows:

- **DISCRETE DISTRIBUTIONS**

| | Parameters | EXPECTATION | VARIANCE |
|---|---|---|---|
| $Bernoulli(\theta)$ | $\theta$ | $\theta$ | $\theta(1-\theta)$ |
| $Binomial(n,\theta)$ | $n,\theta$ | $n\theta$ | $n\theta(1-\theta)$ |
| $Poisson(\lambda)$ | $\lambda$ | $\lambda$ | $\lambda$ |
| $Geometric(\theta)$ | $\theta$ | $\frac{1}{\theta}$ | $\frac{(1-\theta)}{\theta^2}$ |
| $NegBinomial(n,\theta)$ | $n,\theta$ | $\frac{n}{\theta}$ | $\frac{n(1-\theta)}{\theta^2}$ |

- **CONTINUOUS DISTRIBUTIONS**

| | Parameters | EXPECTATION | VARIANCE |
|---|---|---|---|
| $Uniform(a,b)$ | $a,b$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| $Exponential(\lambda)$ | $\lambda$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| $Gamma(\alpha,\beta)$ | $\alpha,\beta$ | $\frac{\alpha}{\beta}$ | $\frac{\alpha}{\beta^2}$ |
| $Beta(\alpha,\beta)$ | $\alpha,\beta$ | $\frac{\alpha}{\alpha+\beta}$ | $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ |
| $Normal(\mu,\sigma^2)$ | $\mu,\sigma^2$ | $\mu$ | $\sigma^2$ |

### 2.8.4 ENTROPY

For a random variable $X$ with mass or density function $f_X$, the **entropy** of the distribution is defined by

$$H_{f_X}[X] = E_{f_X}[-\log f_X(X)] = \begin{cases} -\sum_x \log[f_X(x)]\ f_X(x) & \text{DISCRETE CASE} \\ -\int \log[f_X(x)]\ f_X(x)dx & \text{CONTINUOUS CASE} \end{cases}$$

where log in this case can mean logarithm to any base; typically, $\log_2$ or ln (natural log) are used. One interpretation of the entropy of a distribution is that it measures the "evenness" of the distribution, that is, a distribution with high entropy assigns approximately equal probability to each value of the random variable.

**EXAMPLES** Consider the discrete uniform distribution

$$f_X(x) = \frac{1}{N} \qquad x = 1, 2, ..., N$$

Then

$$H_{f_X}[X] = -\sum_x \log[f_X(x)]\ f_X(x) = -\sum_{x=1}^{N} \log\left[\frac{1}{N}\right] \times \frac{1}{N} = \log N$$

In fact, the uniform distribution is the distribution for which entropy is maximized. Now consider the discrete distribution, $f_X(x) = 1$ when $x = x_0$, and zero otherwise (so that $\mathrm{P}[X = x_0] = 1$, and thus $X$ is **certain** to take the value $x_0$). Then

$$H_{f_X}[X] = \log[f_X(x_0)] \times f_X(x_0) = 0$$

These two examples illustrate another interpretation for the entropy as an overall measure of *uncertainty*. In the first example, there is the **maximum** possible uncertainty, whereas in the second example the uncertainty is at a **minimum**.

### 2.8.5 RELATIVE ENTROPY

It is also possible to compare two distributions using an entropy measure. Consider two discrete distributions with mass functions $f_0$ and $f_1$. Then the **relative entropy** of $f_1$ with respect to $f_0$ , $H_{f_0||f_1}$, and the **relative entropy** of $f_0$ with respect to $f_1$ , $H_{f_1||f_0}$, are defined by

$$\begin{aligned} H_{f_0||f_1}[X] &= E_{f_0}\left[\log\left\{\frac{f_0(X)}{f_1(X)}\right\}\right] = \sum_x \log\left\{\frac{f_0(x)}{f_1(x)}\right\}\ f_0(x) \\ H_{f_1||f_0}[X] &= E_{f_1}\left[\log\left\{\frac{f_1(X)}{f_0(X)}\right\}\right] = \sum_x \log\left\{\frac{f_1(x)}{f_0(x)}\right\}\ f_1(x) \end{aligned}$$

where the sum extends over values of $x$ for which both $f_0$ and $f_1$ are non-zero. It is also possible to obtain an overall measure of the difference in entropy terms between the two distributions as the sum of these two measures.

$$H_{f_0,f_1}[X] = H_{f_0||f_1}[X] + H_{f_1||f_0}[X] = \sum_x \log\left\{\frac{f_0(x)}{f_1(x)}\right\}\ f_0(x) + \sum_x \log\left\{\frac{f_1(x)}{f_0(x)}\right\}\ f_1(x)$$

It can be shown that $H_{f_0||f_1}[X], H_{f_1||f_0}[X]$ and hence $H_{f_0,f_1}[X]$ are all non-negative. Furthermore, we can define the **support** for $x$ in favour of $f_0$ over $f_1$, denoted $S_{0,1}(x)$ by

$$S_{0,1}(x) = \log\left\{\frac{f_0(x)}{f_1(x)}\right\}$$

with the equivalent definition for $S_{1,0}(x)$ (where $S_{1,0}(x) = -S_{0,1}(x)$) Using this definition, we see that $S_{0,1}(X)$ is a random variable, and using the general definition of expectation we have that the expectation of $S_{0,1}(x)$ is

$$\sum_x S_{0,1}(x)\ f_0(x) = \sum_x \log\left\{\frac{f_0(x)}{f_1(x)}\right\}\ f_0(x) = H_{f_0||f_1}[X]$$

## 2.9 TRANSFORMATIONS OF RANDOM VARIABLES

Consider a discrete or continuous random variable $X$ with range $\mathbb{X}$ and probability distribution described by mass/pdf $f_X$, or cdf $F_X$. Suppose $g$ is a function. Then $Y = g(X)$ is also a random variable as $Y$ and typically we wish to derive the probability distribution of random variable $Y$; in order to do this, we might consider the inverse transformation $g^{-1}$ from $\mathbb{Y}$ to $\mathbb{X}$ Consider first the cdf of $Y$, $F_Y$, evaluated at a point $y \in$ . We have

$$F_Y(y) = \mathrm{P}[\ Y \le y\ ] = \ \mathrm{P}[\ g(X) \le y\ ] = \begin{cases} \displaystyle\sum_{x \in A_y} f_X(x) & \text{if } X \text{ is discrete} \\ \\ \displaystyle\int_{A_y} f_X(x)\ dx & \text{if } X \text{ is continuous} \end{cases}$$

where $A_y = \{\ x \in \mathbb{X}\ :\ g(x) \le y\}$. Often, the set is $A_y$ is easy to identify for a given $y$, and this becomes our main objective in the calculation.

**EXAMPLE** Suppose that $g(x) = \exp x$. Then

$$F_Y(y) = P[\ Y \le y\ ] = \ P[\ \exp X \le y\ ] = P[\ X \le \log y\ ] = F_X(\log y)$$

so that $A_y = \{\ x \in \mathbb{X}\ :\ \exp x \le y\} = \{\ x \in \mathbb{X}\ :\ x \le \log y\}$

**EXAMPLE** Suppose that $g(x) = ax + b$. Then

$$F_Y(y) = P[\ Y \le y\ ] = \ P[\ aX + b \le y\ ] = P\left[X \le \frac{y-b}{a}\right] = F_X\left(\frac{y-b}{a}\right)$$

so that $A_y = \{\ x \in \mathbb{X}\ :\ ax + b \le y\} = \{\ x \in \mathbb{X}\ :\ x \le (y-b)/a\}\,.$

**EXAMPLE** Suppose that $g(x) = x^2$. Then

$$F_Y(y) = P[\ Y \le y\ ] = \ P\left[X^2 \le y\right] = P\left[-\sqrt{y} \le X \le \sqrt{y}\right] = F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right)$$

so that $A_y = \left\{\ x \in \mathbb{X}\ :\ x^2 \le y\right\} = \left\{\ x \in \mathbb{X}\ :\ \sqrt{y} \le x \le \sqrt{y}\right\}$

We may be interested in the mass or density function of the newly formed variable $Y$; in that case we could take the cdf formed above and use it to calculate the mass function/pdf. For example, if $Y = aX + b$ when $a > 0$ then

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right) \Longrightarrow f_Y(y) = \frac{d}{dy}\left\{F_X\left(\frac{y-b}{a}\right)\right\} = \frac{d}{dy}\left\{\left(\frac{y-b}{a}\right)\right\} f_Y\left(\frac{y-b}{a}\right) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

using the chain rule for differentiation that says

$$\frac{d}{dx}\{g(h(x)\} = h'(x)g'(h(x)) \qquad whereg'(x) = \frac{dg(x)}{dx} \qquad h'(x) = \frac{dh(x)}{dx}$$

In the discrete case, it may be easier to consider the mass function directly rather than the cdf. However, for a particular type of transformations, namely 1-1 transformations, it is possible to produce a general transformation result that allows direct calculation of the distribution of the transformed variable.

### 2.9.1 LOCATION/SCALE TRANSFORMATIONS

A particular type of 1-1 transformation is a *location/scale* transformation; this transformation take the form

$$Y = \mu + \lambda X$$

where $\mu$ and $\lambda > 0$ are two real-parameters. The pdf of the location/scale transformed variable $Y$ was derived above using first principles

$$F_Y(y) = P[Y \leq y] = P[\mu + \lambda X \leq y] = P\left[X \leq \frac{y-\mu}{\lambda}\right] = F_X\left(\frac{y-\mu}{\lambda}\right)$$

and therefore, by differentiation

$$f_Y(y) = \frac{d}{dy}\left\{F_X\left(\frac{y-\mu}{\lambda}\right)\right\} = \frac{1}{\lambda} f_X\left(\frac{y-b}{\lambda}\right)$$

Sometimes we are interested in a *scale* transformation only where $\mu = 0$ in the transformation above.

### 2.9.2 TRANSFORMATION CONNECTIONS BETWEEN DISTRIBUTIONS

Some of the continuous distributions that we have studied are directly connected by transformations

| Distribution of $X$ | Transformation | Distribution of $Y$ |
|---|---|---|
| $X \sim Uniform(0,1)$ | $Y = -\frac{1}{\lambda}\log X$ | $Y \sim Exponential(\lambda)$ |
| $X \sim Gamma(\alpha,1)$ | $Y = X/\beta$ | $Y \sim Gamma(\alpha,\beta)$ |
| $X \sim Normal(0,1)$ | $Y = \mu + \sigma X$ | $Y \sim Normal(\mu,\sigma^2)$ |
| $X \sim Normal(0,1)$ | $Y = X^2$ | $Y \sim Gamma\left(\frac{1}{2},\frac{1}{2}\right) \equiv \chi^2_1$ |

## 2.10 JOINT PROBABILITY DISTRIBUTIONS

Consider a vector of $k$ random variables, $\mathbf{X} = (X_1, ..., X_k)$, (representing the outcomes of $k$ different experiments carried out once each, or of one experiment carried out $k$ times). The probability distribution of $\mathbf{X}$ is described by a **joint** probability mass or density function.

e.g. Consider the particular case $k = 2$, $\mathbf{X} = (X_1, X_2)$. Then the following functions are used to specify the probability distribution of $\mathbf{X}$;

### 2.10.1 JOINT PROBABILITY MASS/DENSITY FUNCTIONS

The joint mass/density function is denoted

$$f_{X_1,X_2}(x_1, x_2)$$

that is, a function of **two** variables.

- This function assigns probability to the joint space of outcomes

- in the discrete case,

$$f_{X_1,X_2}(x_1, x_2) = P[(X_1 = x_1) \cap (X_2 \cap x_2)]$$

- which implies that we need

(i) $f_{X_1,X_2}(x_1, x_2) \geq 0$ for all possible outcomes $x_1, x_2$.

(ii) $\sum\sum f_{X_1,X_2}(x_1, x_2) = 1$ or $\int\int f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 = 1$

where the double summation/integration is over all possible values of $(x_1, x_2)$.

Typically, such a specification is represented by a **probability table**; for example for **discrete** random variables $X_1$ and $X_2$, we could have

| | | $X_1$ | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| $X_2$ | 1 | 0.100 | 0.200 | 0.000 | 0.000 |
| | 2 | 0.200 | 0.250 | 0.050 | 0.000 |
| | 3 | 0.000 | 0.050 | 0.050 | 0.025 |
| | 4 | 0.000 | 0.000 | 0.025 | 0.050 |

where the entry in column $i$, row $j$ is

$$f_{X_1,X_2}(i, j) = P[X_1 = i) \cap (X_2 = j)] = P[X_1 = i, X_2 = j],$$

Here we only study joint distributions for two variables, but the extension to more than two variables is straightforward.

### 2.10.2 MARGINAL MASS/DENSITY FUNCTIONS

The joint mass function automatically defines the probability distribution of the individual random variables. For example, if $k = 2$, then we have the two marginal mass/density functions are $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. In the discrete and continuous cases respectively

$$f_{X_1}(x_1) = \sum_{x_2} f_{X_1,X_2}(x_1, x_2) \qquad f_{X_2}(x_2) = \sum_{x_1} f_{X_1,X_2}(x_1, x_2)$$

$$f_{X_1}(x_1) = \int f_{X_1,X_2}(x_1, x_2)\, dx_2 \qquad f_{X_2}(x_2) = \int f_{X_1,X_2}(x_1, x_2)\, dx_1$$

so the marginal mass/density function for random variable $X_1$ is obtained by summing/integrating out the joint mass/density function for $X_1$ and $X_2$ over all possible values of random variable $X_2$. In the discrete case

$$P[X_1 = x_1] = \sum_{x_2} P[(X_1 = x_1) \cap (X_2 = x_2)]$$

which is a result that is justified by the Theorem of Total Probability.

In the table above, the marginal mass functions can be computed easily

| | | $X_1$ | | | | $\downarrow$ |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | $f_{X_2}(x_2)$ |
| $X_2$ | 1 | 0.100 | 0.200 | 0.000 | 0.000 | **0.300** |
| | 2 | 0.200 | 0.250 | 0.050 | 0.000 | **0.500** |
| | 3 | 0.000 | 0.050 | 0.050 | 0.025 | **0.125** |
| | 4 | 0.000 | 0.000 | 0.025 | 0.050 | **0.075** |
| $\rightarrow$ | $f_{X_1}(x_1)$ | **0.300** | **0.500** | **0.125** | **0.075** | |

so that the marginal mass functions are formed by the column and row sums respectively. In this case, it turns out that $f_{X_1}(x) = f_{X_2}(x)$, for each $x = 1, 2, 3, 4$, but this will not always be the case.

### 2.10.3 CONDITIONAL MASS/DENSITY FUNCTIONS

In the discrete two variable case, consider the probability

$$P[\ X_1 = x_1 \mid X_2 = x_2\ ]$$

that is, the conditional probability distribution of $X_1$, given that $X_2 = x_2$. This conditional distribution is easily computed from the conditional probability definition, that is

$$P[X_1 = x_1 | X_2 = x_2] = \frac{P[\ X_1 = x_1, X_2 = x_2\ ]}{P[\ X_2 = x_2\ ]} = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

that is, proportional to the $x_2$ row of the table.

By extending these concepts, we may define the conditional probability distributions for both variables in the discrete and continuous cases; The two conditional mass/density functions are $f_{X_1|X_2}(x_1|x_2)$ and $f_{X_2|X_1}(x_2|x_1)$

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} \qquad f_{X_2}(x_2) > 0$$

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} \qquad f_{X_1}(x_1) > 0$$

In the discrete case, this result becomes

$$f_{X_1|X_2}(x_1|x_2) = P[X_1 = x_1|X_2 = x_2] = \frac{P[(X_1 = x_1) \cap (X_2 = x_2)]}{P[X_2 = x_2]}$$

if $P[X_2 = x_2] > 0$, which is justified by the definition of conditional probability.

For example, consider the table

| | | $X_1$ | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | $f_{X_2}(x_2)$ |
| $X_2$ | 1 | 0.100 | **0.200** | 0.000 | 0.000 | 0.300 |
| | 2 | 0.200 | **0.250** | 0.050 | 0.000 | 0.500 |
| | 3 | 0.000 | **0.050** | 0.050 | 0.025 | 0.125 |
| | 4 | 0.000 | **0.000** | 0.025 | 0.050 | 0.075 |
| | $f_{X_1}(x_1)$ | 0.300 | **0.500** | 0.125 | 0.075 | |

The highlighted column gives the conditional mass function for $X_2$ given that $X_1 = 2$; from the definition,

$$f_{X_2|X_1}(1|2) = \frac{0.200}{0.500} = 0.400 \qquad f_{X_2|X_1}(1|2) = \frac{0.250}{0.500} = 0.500$$

$$f_{X_2|X_1}(3|2) = \frac{0.050}{0.500} = 0.100 \qquad f_{X_2|X_1}(4|2) = \frac{0.000}{0.500} = 0.000$$

Note that

$$\sum_{x=1}^{4} f_{X_2|X_1}(x|2) = 0.400 + 0.500 + 0.100 + 0.000 = 1$$

which we must have for a conditional mass function.

Note that, in general, the conditional mass functions will be **different** for different values of the conditioning variable.

**SUMMARY**

Suppose that $X_1$ and $X_2$ are discrete random variables that take values $\{1, 2, ..., n\}$ and $\{1, 2, ..., m\}$ respectively. Then the joint mass function can be displayed as a table with $n$ columns and $m$ rows, where

- the $(i, j)$th cell contains $P[(X_1 = i) \cap (X_2 = j)]$
- the marginal mass function for $X_1$ is given by the **column** totals
- the marginal mass function for $X_2$ is given by the **row** totals
- the conditional mass function for $X_1$ given $X_2 = j$ is given by the $j$th row divided by the sum of the $j$th row
- the conditional mass function for $X_2$ given $X_1 = i$ is given by the $i$th column divided by the sum of the $i$th column

for $i = 1, ..., n$ and $j = 1, ..., m$.

**CONTINUOUS EXAMPLE**

If the joint density of continuous variables $X_1$ and $X_2$ is given by

$$f_{X_1,X_2}(x_1, x_2) = x_2^2 e^{-x_2(1+x_1)}$$

for $x_1, x_2 \geq 0$ and zero otherwise. It can be shown that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_1 dx_2 = 1$$

and that the marginal pdf for $X_1$ is given by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_2 = \int_{0}^{\infty} x_2^2 e^{-x_2(1+x_1)}\ dx_2 = \frac{2}{(1 + x_1)^3}$$

for $x_1 \geq 0$, and zero otherwise.

### 2.10.4 INDEPENDENCE

Random variables $X_1$ and $X_2$ are **independent** if

(i) the joint mass/density function of $X_1$ and $X_2$ factorizes into the product of the two marginal pdfs, that is,

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

(ii) the range of $X_1$ does not conflict/influence/depend on the range of $X_2$ (and *vice versa*).

The concept of independence for random variables is closely related to the concept of independence for events.

### 2.10.5 THE MULTINOMIAL DISTRIBUTION

The **multinomial** distribution is a multivariate generalization of the **binomial** distribution. One interpretation of the binomial distribution is that it is the probability distribution for the random variable that counts the number of "successes" in a sequence of Bernoulli experiments. Let us label the "successes" (1) as **Type I** outcomes and "failures" (0) as **Type II** outcomes. Suppose as usual that the probability of Type I outcomes is $\theta$ (so $0 \leq \theta \leq 1$) and hence the probability of Type II outcomes in the urn is $1-\theta$. If $n$ experiments are carried out, and $X$ is the random variable corresponding to the number of Type I outcomes, then $X \sim Binomial(n, \theta)$

$$f_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad x \in \{0, 1, 2, ..., n\}$$

Now consider a generalization; suppose that there are $k+1$ types of outcomes ($k = 1, 2, ...$), with

$$\text{"Probability of type } i \text{ outcome"} \;=\; \theta_i$$

for $i = 1, ..., k+1$. Let $X_i$ be the random variable corresponding to the number of type $i$ outcomes in $n$ repeats of the experiment, for $i = 1, ..., k$. Then the joint distribution of vector $\mathbf{X} = (X_1, ..., X_k)$ is given by

$$f_{X_1,...,X_k}(x_1, ..., x_k) = \frac{n!}{x_1!...x_k!x_{k+1}!}\theta_1^{x_1}....\theta_k^{x_k}\theta_{k+1}{}^{x_{k+1}} = \binom{n}{x_1, x_2, ..., x_k}\prod_{i=1}^{k+1}\theta_i^{x_i}$$

where $0 \leq \theta_i \leq 1$ for all $i$, and $\theta_1 + ... + \theta_k + \theta_{k+1} = 1$, and where $x_{k+1}$ is defined by $x_{k+1} = n - (x_1 + ... + x_k)$. This is the mass function for the **MULTINOMIAL DISTRIBUTION** which reduces to the binomial if $k = 1$. It can also be shown that the marginal distribution of $X_i$ is given by

$$X_i \sim Binomial(n, \theta_i).$$

**EXAMPLE** Consider the sequence

$$ATAGTAGATACGCACCGAGGA$$

For the probability of seeing such a sequence, we let

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

for some suitable probabilities satisfying

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

and assume independence so that

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

which simplifies to

$$P(ATAGTAGATACGCACCGAGGA) = p_A^8 p_C^4 p_G^6 p_T^3$$

However, for if we merely wish to identify the probability that in a sequence of 21 bases, we observe 8 $A$, 4 $C$, 6 $G$ and 3 $T$ in **any order** then the multinomial distribution is used, that is the probability is given by

$$\binom{21}{8, 4, 6, 3}p_A^8 p_C^4 p_G^6 p_T^3 = \frac{21!}{8! \times 4! \times 6! \times 3!}p_A^8 p_C^4 p_G^6 p_T^3$$

## 2.11 COVARIANCE AND CORRELATION

**Definition 2.11.1 COVARIANCE**
The **covariance** of two random variables $X_1$ and $X_2$ is denoted $Cov_{f_{X_1,X_2}}[X_1, X_2]$, and is defined by

$$Cov_{f_{X_1,X_2}}[X_1, X_2] = E_{f_{X_1,X_2}}[(X_1 - \mu_1)(X_2 - \mu_2)] = E_{f_{X_1,X_2}}[X_1X_2] - \mu_1\mu_2$$

where

$$E_{f_{X_1,X_2}}[X_1X_2] = \begin{cases} \sum_{x_2}\sum_{x_1} x_1x_2 f_{X_1,X_2}(x_1, x_2) & X_1 \text{and } X_2 \textbf{discrete} \\ \int\int x_1x_2 f_{X_1,X_2}(x_1, x_2)\, dx_1dx_2 & X_1 \text{and } X_2 \textbf{continuous} \end{cases}$$

is the expectation of the function $g(x_1, x_2) = x_1x_2$ with respect to the joint probability function $f_{X_1,X_2}$, and where $\mu_i = E_{f_{X_i}}[X_i]$ is the expectation of $X_i$, for $i = 1, 2$.

**Definition 2.11.2 CORRELATION**
The **correlation** of $X_1$ and $X_2$ is denoted $Corr_{f_{X_1,X_2}}[X_1, X_2]$, and is defined by

$$Corr_{f_{X_1,X_2}}[X_1, X_2] = \frac{Cov_{f_{X_1,X_2}}[X_1, X_2]}{\sqrt{Var_{f_{X_1}}[X_1]\ Var_{f_{X_2}}[X_2]}}$$

If

$$Cov_{f_{X_1,X_2}}[X_1, X_2] = Corr_{f_{X_1,X_2}}[X_1, X_2] = 0$$

then variables $X_1$ and $X_2$ are **uncorrelated**. Note that if random variables $X_1$ and $X_2$ are **independent** then

$$Cov_{f_{X_1,X_2}}[X_1, X_2] = E_{f_{X_1,X_2}}[X_1X_2] - E_{f_{X_1}}[X_1]E_{f_{X_2}}[X_2] = E_{f_{X_1}}[X_1]E_{f_{X_2}}[X_2] - E_{f_{X_1}}[X_1]E_{f_{X_2}}[X_2] = 0$$

and so $X_1$ and $X_2$ are also uncorrelated (note that the converse does not necessarily hold).

Key interpretation

**COVARIANCE AND CORRELATION ARE MEASURES OF THE DEGREE OF ASSOCIATION BETWEEN VARIABLES**

that is, two variables for which the correlation is **large** in magnitude are strongly associated, whereas variables that have low correlation are **weakly** associated.

### 2.11.1 PROPERTIES OF COVARIANCE AND CORRELATION

(i) For random variables $X_1$ and $X_2$, with (marginal) expectations $\mu_1$ and $\mu_2$ respectively, and (marginal) variances $\sigma_1^2$ and $\sigma_2^2$ respectively, if random variables $Z_1$ and $Z_2$ are defined $Z_1 = (X_1 - \mu_1)/\sigma_1$ and $Z_2 = (X_2 - \mu_2)/\sigma_2$ so that $Z_1$ and $Z_2$ are *standardized* variables. Then

$$Corr_{f_{X_1,X_2}}[X_1, X_2] = Cov_{f_{Z_1,Z_2}}[Z_1, Z_2].$$

(ii) The extension to $k$ variables: covariances can only be calculated for *pairs* of random variables, but if $k$ variables have a joint probability structure it is possible to construct a $k \times k$ *matrix*, $\mathbf{C}$ say, of covariance values, whose $(i, j)$th element is

$$Cov_{f_{X_i,X_j}}[X_i, X_j] = Cov_{f_{X_i,X_j}}[X_i, X_j]$$

for $i, j = 1, .., k$, (so $\mathbf{C}$ is *symmetric)* that captures the complete covariance structure in the joint distribution. If $i = j$,

$$Cov_{f_{X_i,X_i}}[X_i, X_i] \equiv Var_{f_{X_i}}[X_i]$$

The matrix $\mathbf{C}$ is referred to as the **variance-covariance** matrix.

(iii) If random variable $X$ is defined by

$$X = \sum_{i=1}^{k} a_i X_i$$

for random variables $X_1, ..., X_k$ and constants $a_1, ..., a_k$, then

$$\begin{aligned} E_{f_X}[X] &= \sum_{i=1}^{k} a_i E_{f_{X_i}}[X_i] \\ Var_{f_X}[X] &= \sum_{i=1}^{k} a_i^2 Var_{f_{X_i}}[X_i] + 2\sum_{i=1}^{k}\sum_{j=1}^{i-1} a_i a_j Cov_{f_{X_i,X_j}}[X_i, X_j] \end{aligned}$$

(iv) Combining (i) and (iii) when $k = 2$, and defining standardized variables $Z_1$ and $Z_2$,

$$\begin{aligned} 0 \leq Var_{f_{Z_1,Z_2}}[Z_1 \pm Z_2] &= Var_{f_{Z_1}}[Z_1] + Var_{f_{Z_2}}[Z_2] \pm 2\ Cov_{f_{Z_1,Z_2}}[Z_1, Z_2] \\ &= 1 + 1 \pm 2\ Corr_{f_{X_1,X_2}}[X_1, X_2] \\ &= 2(1 \pm Corr_{f_{X_1,X_2}}[X_1, X_2]) \end{aligned}$$

and hence we have the key result that

$$-1 \leq Corr_{f_{X_1,X_2}}[X_1, X_2] \leq 1.$$

that is, the correlation is **bounded** between -1 and 1. We will see later how to compute covariance and correlation for sample data; there is a close relationship between theoretical and sample covariances and correlations.

## 2.12 EXTREME VALUES

### 2.12.1 ORDER STATISTICS, MAXIMA AND MINIMA

**Definition 2.12.1 ORDER STATISTICS**
For $n$ random variables $X_1, ...., X_n$, the **order statistics**, $Y_1, ...., Y_n$, are defined by

$$Y_i = X_{(i)} - \text{“the } i^{th}\text{smallest value in } X_1, ...., X_n\text{”}$$

for $i = 1, ..., n$, so that

$$Y_1 = X_{(1)} = \min\{X_1, ..., X_n\} \qquad Y_n = X_{(n)} = \max\{X_1, ..., X_n\}$$

For $n$ independent, identically distributed random variables $X_1, ..., X_n$, with marginal density function $f_X$, there are two main results to consider; it can be shown that the joint density function of the order statistics $Y_1, ...., Y_n$ is given by

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = n! f_X(y_1)...f_X(y_n) \qquad y_1 < ... < y_n$$

and that the marginal pdf of the $j$th order statistic $Y_j$ for $j = 1, ..., n$ has the form

$$f_{Y_j}(y_j) = \frac{n!}{(j-1)!(n-j)!} \{F_X(y_j)\}^{j-1} \{1 - F_X(y_j)\}^{n-j} f_X(y_j)$$

To derive the marginal pdf of the **maximum** $Y_n$, first consider the marginal cdf of $Y_n$;

$$\begin{aligned} F_{Y_n}(y_n) &= P[Y_n \le y_n] = P[\max\{X_1, ..., X_n\} \le y_n] = P[X_1 \le y_n, X_2 \le y_n, ..., X_n \le y_n] \\ &= \prod_{i=1}^{n} P[X_i \le y_n] = \prod_{i=1}^{n} \{F_X(y_n)\} \\ &= \{F_X(y_n)\}^n \end{aligned}$$

and so

$$f_{Y_n}(y_n) = n \{F_X(y_n)\}^{n-1} f_X(y_n) \qquad \text{differentiating using the chain rule}$$

By a similar calculation, we can find the marginal pdf/cdf for the **minimum** $Y_1$;

$$\begin{aligned} F_{Y_1}(y_1) &= P[Y_1 \le y_1] = 1 - P[Y_1 > y_1] = 1 - P[\min\{X_1, ..., X_n\} > y_1] \\ &= 1 - P[X_1 > y_1, X_2 > y_1, ..., X_n > y_1] \\ &= 1 - \prod_{i=1}^{n} P[X_i > y_1] = 1 - \prod_{i=1}^{n} \{1 - F_X(y_1)\} \\ &= 1 - \{1 - F_X(y_1)\}^n \end{aligned}$$

and so

$$f_{Y_1}(y_1) = n \{1 - F_X(y_1)\}^{n-1} f_X(y_1) \qquad \text{differentiating using the chain rule}$$

Hence

$$F_{Y_1}(y_1) = 1 - \{1 - F_X(y_1)\}^n \qquad f_{Y_1}(y_1) = n \{1 - F_X(y_1)\}^{n-1} f_X(y_1)$$

### 2.12.2 GENERAL EXTREME VALUE THEORY

A special probabilistic theory has been developed for extreme observations (i.e. maxima or minima) derived from sequences of independent (and, indeed dependent) sequences of random variables $X_1, ...., X_n$. The theory is based on the following general "large sample" or asymptotic results. Results for **maxima** are given here, and the results for minima follow by considering the maxima of the sequence with their sign changed, that is, the sequence

$$(-X_1), ...., (-X_n)$$

**Theorem 2.12.1** *Let $M_n$ be the **maximum** of a sample of n independent and identically distributed random variables $X_1, ...., X_n$. Then there exist real constants $c_n > 0$ and $d_n$ such that the distribution of the random variable*

$$Z_n = \frac{M_n - d_n}{c_n}$$

***converges in distribution** to one of three continuous probability distributions as $n \to \infty$; we write*

$$Z_n \xrightarrow{d} Z \sim F_Z$$

*where $F_Z$ is either*

*(I) the **Frechet-type** distribution with cdf*

$$F_Z(z) = \begin{cases} 0 & z < \mu \\ \exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{-\alpha}\right\} & z \geq \mu \end{cases}$$

*for parameters $\alpha, \sigma > 0$ and $\mu \in \mathbb{R}$*

*(II) the **Weibull-type** distribution with cdf*

$$F_Z(z) = \begin{cases} \exp\left\{-\left(\frac{\mu-z}{\sigma}\right)^{\alpha}\right\} & z \leq \mu \\ 1 & z > \mu \end{cases}$$

*for parameters $\alpha, \sigma > 0$ and $\mu \in \mathbb{R}$.*

*(III) the **Gumbel-type** distribution with cdf*

$$F_Z(z) = \exp\left\{-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right\} \qquad -\infty < z < \infty$$

*for parameters $\sigma > 0$ and $\mu \in \mathbb{R}$.*

Type (I) and Type (II) distributions are transformed versions of Type (III) distributions; the transformations are

$$\log(Z-\mu) \qquad and \qquad -\log(\mu-Z)$$

respectively. In addition, the distribution of the variable $(-Z)$ has, in each case an extreme value distribution.

**Definition 2.12.2 GENERALIZED EXTREME VALUE DISTRIBUTION**

The three distributions above can be incorporated into a single probability distribution, the *Generalized Extreme Value (GEV)* distribution.by allowing the parameters to take their specific values or limiting forms; this cdf takes the form

$$F_Z(z) = \exp\left\{-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}\right\} \qquad 1+\alpha\left(\frac{z-\mu}{\sigma}\right) > 0$$

for parameters $-\infty < \mu, \alpha < \infty$ and $\sigma > 0$. We have that

(I)$\xi = 1/\alpha > 0$ gives the Frechet-type distribution

(II) $\xi = -1/\alpha < 0$ gives the Weibull-type distribution

(III) $\xi \to 0$ gives the Gumbel-type distribution

The GEV distribution describes the probability distribution of maximum order statistics. We now consider **threshold-excedance** distributions; that is, the distribution of observed values beyond a certain high threshold value. Let $X$ be a random variable with cdf $F_X$, and for some fixed $u$, let

$$Y = (X-u)\, I_{X>u}$$

that is, $Y = X - u$ if $X > u$. By straightforward calculation

$$F_Y(y;u) = P[Y \le y;u] = P[X \le u+y | X > u] = \frac{F_X(u+y) - F_X(u)}{1-F_X(u)} \qquad y > 0$$

The distribution of $Y$ as $u$ approaches some upper endpoint is known as the *Generalized Pareto Distribution.*

**Definition 2.12.3 GENERALIZED PARETO DISTRIBUTION**

The *Generalized Pareto Distribution (GPD)* for random variable $Y$, given some threshold $u$, takes the form

$$F_Y(y;u) = 1 - \left(1+\xi\frac{y}{\sigma_u}\right)^{-1/\xi} \qquad y > 0$$

that is, the parameter $\sigma_u$ depends on the threshold $u$, and the parameter $\xi$ does not

For events occurring in time or space, the **number** $N$ of events that exceed a threshold $u$ in any time interval $t$, $X(t)$, is often adequately modelled using a Poisson distribution with parameter $\lambda t$; we say that the events occur at rate $\lambda$. Given that $N \ge 1$, the excedances themselves are distributed according to the GPD model, and the largest excedance is well modelled using a GEV distribution.

# CHAPTER 3

# STATISTICAL ANALYSIS

Statistical analysis involves the informal/formal comparison of hypothetical or predicted behaviour with experimental results. For example, we wish to be able to compare the predicted outcomes of an experiment, and the corresponding probability model, with a data histogram. We will use both *qualitative* and *quantitative* approaches.

## 3.1 GENERAL FRAMEWORK, NOTATION AND OBJECTIVES

Suppose that an experiment or **trial** is to be repeated $n$ times under identical conditions. Let $X_i$ be the random variable corresponding to the outcome of the $i$th trial, and suppose that each of the $n$ random variables $X_1, ..., X_n$ takes values in sample space $\mathbb{X}$. Often, assumptions can reasonably be made about the experimental conditions that lead to simplifications of the joint probability model for the random variables. Essentially, the assumption of identical experimental conditions for each of the $n$ trials implies that the random variables corresponding to the trial outcomes are **identically distributed**, that is, in the usual notation, the (marginal) mass/density function of $X_i$ is denoted $f(x)$ dropping the subscript on the function $f$. Another common assumption is that the random variables $X_1, ..., X_n$ are **independent**. Thus $X_1, ..., X_n$ are usually treated as **i.i.d.** random variables.

In practice, it is commonly assumed that $f$ takes one of the familiar forms (*Binomial*, *Poisson*, *Exponential*, *Normal* etc.). Thus $f$ depends on one or more parameters ($\theta,\ \lambda,\ (\mu, \sigma)$ etc.). The role of these parameters could be indicated by re-writing the function $f(x)$ as

$$f(x) \equiv f(x; \theta) \qquad x \in \mathbb{X} \tag{$*$}$$

where $\theta$ here is a **parameter**, which may possibly be vector-valued.

It is important here to specify precisely the range of values which this parameter can take; in a Poisson model, we have parameter $\lambda > 0$, and in a Normal model, we have parameters $\mu \in \mathbb{R},\ \sigma \in \mathbb{R}^+$. In the general case represented by (*) above, we have parameter $\theta \in \Theta$ where $\Theta$ is some subset of $\mathbb{R}^d$ and $d = 1, 2$, say, is the number of parameters. We refer to $\Theta$ as the **parameter space**. In practice, of course, parameter $\theta$ is **unknown** during the experiment.

### 3.1.1 OBJECTIVES OF A STATISTICAL ANALYSIS

After the experiment has been carried out, a sample of **observed data** will have been obtained. Suppose that we have observed outcomes $x_1, ..., x_n$ on the $n$ trials (that is, we have observed $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$), termed a **random sample**. This sample can be used to answer qualitative and quantitative questions about the nature of the experiment being carried out. The objectives of a statistical analysis can be summarized as follows. We want to, for example,

- **SUMMARY : Describe** and **summarize** the sample $\{x_1, ..., x_n\}$ in such a way that allows a specific probability model to be proposed.
- **INFERENCE : Deduce** and **make inference about** the parameter(s) of the probability model $\theta$.
- **TESTING : Test** whether $\theta$ is "**significantly**" larger/smaller/different from some specified value.
- **GOODNESS OF FIT : Test** whether the probability model encapsulated in the mass/density function $f$, and the other model assumptions are **adequate** to explain the experimental results.

The first objective can be viewed as an **exploratory** data analysis exercise - it is crucially important to understand whether a proposed probability distribution is suitable for modelling the observed data, otherwise the subsequent formal inference procedures (estimation, hypothesis testing, model checking) cannot be used.

## 3.2 EXPLORATORY DATA ANALYSIS

We wish first to produce summaries of the data in order to convey general trends or features that are present in the sample. Secondly, in order to propose an appropriate probability model, we seek to **match** features in the observed data to features of one of the conventional (Poisson, Exponential, Normal) probability distributions that may be used in more formal analysis. The four principal features that we need to assess in the data sample are

1. The **location**, or the "average value" in the sample.
2. The **mode**, or "most likely" value or interval observed in the sample.
3. The **scale** or **spread** in the sample.
4. The **skewness** or **asymmetry** in the sample.

These features of the sample are important because we can relate them **directly** to features of probability distributions.

### 3.2.1 NUMERICAL SUMMARIES

The following quantities are useful numerical summary quantities

- Sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample variance: either ($S^2$ or $s^2$ may be used)

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Sample quantiles: suppose that the sample has been sorted into ascending order and relabelled $x_{(1)} < ... < x_{(n)}$. Then the $p$th quantile, $0 < p < 100$,is given by

$$x^{(p)} = x_{(k)}$$

where $k$ is the nearest integer to $pn/100$. Special cases include

| | | |
|---|---|---|
| Median | $m$ | $= x^{(50)}$, the $50^{th}$quantile |
| Lower quartile | $q_{25}$ | $= x^{(25)}$, the $25^{th}$quantile |
| Upper quartile | $q_{75}$ | $= x^{(75)}$, the $75^{th}$quantile |
| Inter-quartile range | $IQR$ | $= q_{75} - q_{25}$ |
| Sample minimum | $x_{\min}$ | $= x_{(1)}$ |
| Sample maximum | $x_{\max}$ | $= x_{(n)}$ |
| Sample range | $R$ | $= x_{(n)} - x_{(1)}$ |

- Sample skewness

$$A = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Sample kurtosis

$$K = \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2}$$

### 3.2.2 LINKING SAMPLE STATISTICS AND PROBABILITY MODELS.

Consider the discrete probability distribution defined on the set of observed sample outcomes $\{x_1, ..., x_n\}$, by placing equal probability $1/n$ on each value, that is, the probability distribution specified by mass function denoted $f_{(n)}$

$$f_{(n)}(x) = \frac{1}{n} \qquad x \in \{x_1, ..., x_n\}.$$

Then the expectation and variance of this probability distribution are given by

$$E_{f(n)}[X] = \sum_{i=1}^{n} x_i f_{(n)}(x_i) = \sum_{i=1}^{n} x_i \left\{\frac{1}{n}\right\} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \qquad Var_{f(n)}[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = S^2$$

that is, the sample mean. Similarly, the variance of this probability distribution is equal to sample variance. In fact, each of the summary statistics listed above can be viewed as a feature of the probability distribution described by mass function $f_{(n)}$.

Now, consider this probability distribution as $n$ increases to infinity. Then the sample mass function $f_{(n)}$ tends to a function $f$ which can be regarded as the "true" mass/density function, and the sample mean, variance, percentiles etc. tend to the true mean, variance, percentiles of the distribution from which the data are generated. In practice, of course, $n$ is always finite, and thus the true distribution, true mean etc., cannot be known exactly. Therefore, we approximate the true distribution by an appropriately chosen distribution (Poisson, Exponential, Normal etc.) with parameters chosen to correspond to the observed sample properties.

### 3.2.3 GRAPHICAL SUMMARIES

The most common graphical summary technique is the **histogram**. Typically, the sample space $\mathbb{X}$ is divided into a number of subsets $\mathbb{X}_1, ..., \mathbb{X}_H$, and the frequency with which a data value in the sample is observed to lie in subset $h = 1, ..., H$ is noted. This procedure leads to a set of counts $n_1, ..., n_H$ (where $n_1 + ... + n_H = n$) which are then plotted on a graph as a set of bars, where the $h$th bar has height $n_h$ and occupies the region of $\mathbb{X}$ corresponding to $\mathbb{X}_h$.

The histogram again aims to approximate the "true" probability distribution generating the data by the observed sample distribution. It illustrates graphically the concepts of location, mode, spread and skewness and general shape features that have been recognized as important features of probability distributions.

### 3.2.4 OUTLIERS

Sometimes, for example due to slight variation in experimental conditions, one or two values in the sample may be much larger or much smaller in magnitude than the remainder of the sample. Such observations are termed **outliers** and must be treated with care, as they can distort the impression given by some of the summary statistics. For example, the sample mean and variance are extremely sensitive to the presence of outliers in the sample. Other summary statistics, for example those based on sample percentiles, are less sensitive to outliers. Outliers can usually be identified by inspection of the raw data, or from careful plotting of histograms.

## 3.3 PARAMETER ESTIMATION

It is often of interest to draw inference from data regarding the parameters of the proposed probability distribution; recall that many aspects of the standard distributions studied are controlled by the distribution parameters. It is therefore important to find a simple and yet general technique for parameter estimation

### 3.3.1 MAXIMUM LIKELIHOOD ESTIMATION

Maximum Likelihood Estimation is a systematic technique for estimating parameters in a probability model from a data. Suppose a sample $x_1, ..., x_n$ has been obtained from a probability model specified by mass or density function $f(x; \theta)$ depending on parameter(s) $\theta$ lying in parameter space $\Theta$. The **maximum likelihood estimate** or **m.l.e.** is produced as follows;

**STEP 1** Write down the **likelihood function**, $L(\theta)$, where

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

that is, the product of the $n$ mass/density function terms (where the $i$th term is the mass/density function evaluated at $x_i$) viewed as a function of $\theta$.

**STEP 2** Take the natural log of the likelihood, and collect terms involving $\theta$.

**STEP 3** Find the value of $\theta \in \Theta$, $\hat{\theta}$, for which $\log L(\theta)$ is maximized, for example by differentiation. If $\theta$ is a single parameter, find $\hat{\theta}$ by solving

$$\frac{d}{d\theta} \{\log L(\theta)\} = 0$$

in the parameter space $\Theta$. If $\theta$ is vector-valued, say $\theta = (\theta_1, ..., \theta_d)$, then find $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_d)$ by simultaneously solving the $d$ equations given by

$$\frac{\partial}{\partial \theta_j} \{\log L(\theta)\} = 0 \qquad j = 1, ..., d$$

in parameter space $\Theta$.

Note that, if parameter space $\Theta$ is a bounded interval, then the maximum likelihood estimate may lie on the boundary of $\Theta$.

**STEP 4** Check that the estimate $\hat{\theta}$ obtained in STEP 3 truly corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\log L(\theta)$ with respect to $\theta$. If

$$\frac{d^2}{d\theta^2} \{\log L(\theta)\} < 0$$

at $\theta = \hat{\theta}$, then $\hat{\theta}$ is confirmed as the m.l.e. of $\theta$ (other techniques may be used to verify that the likelihood is maximized at $\hat{\theta}$).

This procedure is a systematic way of producing parameter estimates from sample data and a probability model; it can be shown that such an approach produces estimates that have good properties. After they have been obtained, the estimates can be used to carry out *prediction* of behaviour for future samples.

**EXAMPLE** A sample $x_1, ..., x_n$ is modelled by a Poisson distribution with parameter denoted $\lambda$

$$f(x;\theta) \equiv f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \qquad x = 0, 1, 2, ...$$

for some $\lambda > 0$.

**STEP 1** Calculate the likelihood function $L(\lambda)$. For $\lambda > 0$,

$$L(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} \right\} = \frac{\lambda^{x_1+...+x_n}}{x_1!....x_n!}e^{-n\lambda}$$

**STEP 2** Calculate the log-likelihood $\log L(\lambda)$.

$$\log L(\lambda) = \sum_{i=1}^{n} x_i \log\lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

**STEP 3** Differentiate $\log L(\lambda)$ with respect to $\lambda$, and equate the derivative to zero.

$$\frac{d}{d\lambda}\{\log L(\lambda)\} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \Longrightarrow \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

**STEP 4** Check that the second derivative of $\log L(\lambda)$ with respect to $\lambda$ is negative at $\lambda = \hat{\lambda}$.

$$\frac{d^2}{d\lambda^2}\{\log L(\lambda)\} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0 \ at \lambda = \hat{\lambda}$$

### 3.3.2 METHOD OF MOMENTS ESTIMATION

Suppose that $X_1, ..., X_n$ is a random sample from a probability distribution with mass/density function $f_X$ that depends on vector parameter $\theta$ of dimension $k$, and suppose that a sample $x_1, ..., x_n$ has been observed. Let the $j$th **theoretical** moment of $f_X$ be denoted $\mu_j$, that is, let

$$\mu_j = E_{f_X}\left[X^j\right]$$

and let the $j$th **sample** moment, denoted $m_j$ be defined for $j = 1, ..., k$ by

$$m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

Then $m_j$ is an **estimate** of $\mu_j$, and

$$M_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

is an **estimator** of $\mu_j$. This method of estimation involves **matching the theoretical moments to the sample moments**, giving (in most cases) $k$ equations in the $k$ elements of vector $\theta$ which may be solved simultaneously to find the parameter estimates. Intuitively, and recalling the Weak Law of Large Numbers, it is reasonable to suppose that there is a close relationship between the theoretical properties of a probability distribution, and large sample derived estimates; for example, we know that, for large $n$, the sample mean converges in probability to the theoretical expectation.

## 3.4 SAMPLING DISTRIBUTIONS

Maximum likelihood can be used systematically to produce estimates from sample data. Consider the following example; if a sample of data $x_1, ..., x_n$ are believed to have a Normal distribution with parameters $\mu$ and $\sigma^2$, then the maximum likelihood estimates based on the sample are given by

$$\widehat{\mu} = \overline{x} \qquad \widehat{\sigma}^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

If five samples of eight observations are collected, however, we might get five different sample means

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $\overline{x}$ |
|---|---|---|---|---|---|---|---|---|
| 10.4 | 11.2 | 9.8 | 10.2 | 10.5 | 8.9 | 11.0 | 10.3 | 10.29 |
| 9.7 | 12.2 | 10.4 | 11.1 | 10.3 | 10.2 | 10.4 | 11.1 | 10.66 |
| 12.1 | 7.9 | 8.6 | 9.6 | 11.0 | 11.1 | 8.8 | 11.7 | 10.10 |
| 10.0 | 9.2 | 11.1 | 10.8 | 9.1 | 12.3 | 10.3 | 9.7 | 10.31 |
| 9.2 | 9.7 | 10.8 | 10.3 | 8.9 | 10.1 | 9.7 | 10.4 | 9.89 |

and so the estimate $\hat{\mu}$ of $\mu$ is different each time. We attempt to understand how $\overline{x}$ varies by calculating the **probability distribution** of the corresponding **estimator**, $\overline{X}$.

The estimator $\overline{X}$ is a **random variable**, the value of which is **unknown** *before* the experiment is carried out. As a random variable, $\overline{X}$ has a probability distribution, known as the **sampling distribution**. The form of this distribution can often be calculated, and used to understand how $\overline{X}$ varies. In the case where the sample data have a Normal distribution, the following theorem gives the sampling distributions of the maximum likelihood estimators;

**THEOREM** If $X_1, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, then

$$(I) \qquad \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$(II) \qquad \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{nS^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

$(III)$ $\overline{X}$and $S^2$are **independent** random variables.

**Interpretation :** This theorem tells us how the sample mean and variance will behave if the original random sample is assumed to come from a Normal distribution. In particular, it tells us that

$$E\left[\overline{X}\right] = \mu \qquad E[S^2] = \frac{n-1}{n}\sigma^2 \qquad E[s^2] = \sigma^2$$

If we believe that $X_1, ..., X_{10}$ are i.i.d random variables from a Normal distribution with parameters $\mu = 10.0$ and $\sigma^2 = 25$, then $\overline{X}$ has a Normal distribution with parameters $\mu = 10.0$ and $\sigma^2 = 25/10 = 2.5$.

The result will be used to facilitate formal tests about model parameters. For example, given a sample of experimental, we wish to answer **specific** questions about parameters in a proposed probability model.

## 3.5 HYPOTHESIS TESTING

Given a sample $x_1, ..., x_n$ from a probability model $f(x;\theta)$ depending on parameter $\theta$, we can produce an estimate $\hat{\theta}$ of $\theta$, and in some circumstances understand how $\hat{\theta}$ varies for repeated samples. Now we might want to test, say, whether or not there is evidence from the sample that true (but unobserved) value of $\theta$ is not equal to a specified value. To do this, we use estimate of $\theta$, and the corresponding estimator and its sampling distribution, to quantify this evidence.

In particular, we concentrate on data samples that we can presume to have a normal distribution, and utilize the Theorem from the previous section. We will look at two situations, namely **one sample** and **two sample** experiments. Suppose that $X_1, ..., X_n \sim N\left(\mu, \sigma^2\right)$ (one sample) and $X_1, ..., X_n \sim N(\mu_X, \sigma_X^2)$, $Y_1, ..., Y_n \sim N(\mu_Y, \sigma_Y^2)$ (two sample)

- **ONE SAMPLE** Possible tests of interest: $\mu = c_1, \sigma = c_2$
- **TWO SAMPLE** Possible tests of interest: $\mu_X = \mu_Y, \sigma_X = \sigma_Y$

### 3.5.1 TESTS FOR NORMAL DATA I - THE Z-TEST ($\sigma$ KNOWN)

If $X_1, ..., X_n \sim N(\mu, \sigma^2)$ are the i.i.d. outcome random variables of $n$ experimental trials, then

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad and \qquad \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

with $\overline{X}$ and $S^2$ statistically independent. Suppose we want to test the **hypothesis** that $\mu = c$, for some specified constant $c$, (where, for example, $c = 20.0$) is a plausible model; more specifically, we want to test the hypothesis $H_0 : \mu = c$ against the hypothesis $H_1 : \mu \neq c$, that is, we want to test whether $H_0$ is true, or whether $H_1$ is true. Now, we know that, in the case of a Normal sample, the distribution of the estimator $\overline{X}$ is Normal, and

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Longrightarrow Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N\left(0, 1\right)$$

where $Z$ is a **random variable**. Now, when we have observed the data sample, we can calculate $\overline{X}$, and therefore we have a way of testing whether $\mu = c$ is a plausible model; we calculate $\overline{x}$ from $x_1, ..., x_n$, and then calculate

$$z = \frac{\overline{x} - c}{\sigma/\sqrt{n}}$$

If $H_0$ is true, and $\mu = c$, then the **observed** $z$ should be an observation from an $N(0, 1)$ distribution (as $Z \sim N(0, 1)$), that is, it should be near zero with high probability. In fact, $z$ should lie between -1.96 and 1.96 with probability $1 - \alpha = 0.95$, say, as

$$P[-1.96 \leq Z < 1.96] = \Phi(1.96) - \Phi(-1.96) = 0.975 - 0.025 = 0.95$$

.If we observe $z$ to be outside of this range, then there is evidence that $H_0$ is **not true**.

Alternatively, we could calculate the probability $p$ of observing a $z$ value that is **more extreme** than the $z$ we did observe; this probability is given by

$$p = \begin{cases} 2\Phi(z) & z < 0 \\ 2(1 - \Phi(z)) & z \geq 0 \end{cases}$$

If $p$ is very small, say $p \leq \alpha = 0.05$, then again. there is evidence that $H_0$ is **not true**. In summary, we need to assess whether $z$ is a **surprising** observation from an $N(0, 1)$ distribution - if it is, then we can **reject** $H_0$.

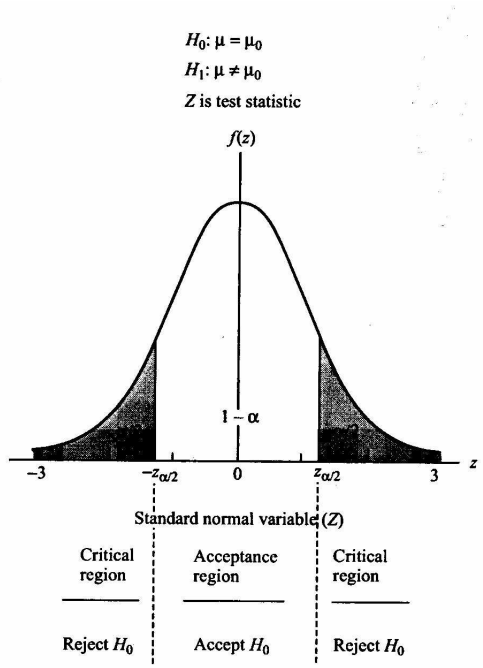



Figure 3.1: CRITICAL REGIONS IN A Z-TEST (taken from *Schaum's ELEMENTS OF STATISTICS II, Bernstein & Bernstein)*

### 3.5.2 HYPOTHESIS TESTING TERMINOLOGY

There are five crucial components to a hypothesis test, namely

- **TEST STATISTIC**
- **NULL DISTRIBUTION**
- **SIGNIFICANCE LEVEL**, denoted $\alpha$
- **P-VALUE**, denoted $p$.
- **CRITICAL VALUE(S)**

In the Normal example given above, we have that

- $z$ is the **test statistic**
- The distribution of random variable $Z$ if $H_0$ is true is the **null distribution**
- $\alpha = 0.05$ is the **significance level** of the test (we could use $\alpha = 0.01$ if we require a "stronger" test)
- $p$ is the **p-value** of the test statistic under the null distribution
- The solution $C_R$ of $\Phi(C_R) = 1 - \alpha/2$ ($C_R = 1.96$ above) gives the **critical values** of the test $\pm C_R$.

**EXAMPLE :** A sample of size 10 has sample mean $\overline{x} = 19.7$. To test the hypothesis

$$H_0 : \mu = 20.0$$
$$H_1 : \mu \neq 20.0$$

under the assumption that the data follow a Normal distribution with $\sigma = 1.0$. We have

$$z = \frac{19.7 - 20.0}{1/\sqrt{10}} = -0.95$$

which lies between the critical values $\pm 1.96$, and therefore we have no reason to reject $H_0$. Also, the p-value is given by $p = 2\Phi(-0.95) = 0.342$, which is greater than $\alpha = 0.05$, which confirms that we have no reason to reject $H_0$.

### 3.5.3 TESTS FOR NORMAL DATA II - THE T-TEST ($\sigma$ UNKNOWN)

In practice, we will often want to test hypotheses about $\mu$ when $\sigma$ is unknown. We cannot perform the Z-test, as this requires knowledge of $\sigma$ to calculate the $z$ statistic.

We proceed as follows; recall that we know the sampling distributions of $\overline{X}$ and $s^2$, and that the two estimators are statistically independent. Now, from the properties of the Normal distribution, if we have independent random variables $Z \sim N(0, 1)$ and $Y \sim \chi^2_\nu$, then we know that random variable $T$ defined by

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

has a Student-$t$ distribution with $\nu$ degrees of freedom. Using this result, and recalling the sampling distributions of $\overline{X}$ and $s^2$, we see that

$$T = \frac{\left(\overline{X} - \mu\right) / \left(\sigma/\sqrt{n}\right)}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2(n-1)}}} = \frac{\left(\overline{X} - \mu\right)}{s/\sqrt{n}} \sim t_{n-1}$$

and $T$ has a Student-$t$ distribution with $n - 1$ degrees of freedom, denoted $St(n-1)$. Thus we can repeat the procedure used in the $\sigma$ known case, but use the sampling distribution of $T$ rather than that of $Z$ to assess whether the test statistic is "surprising" or not. Specifically, we calculate

$$t = \frac{(\overline{x} - \mu)}{s/\sqrt{n}}$$

and find the critical values for a $\alpha = 0.05$ significance test by finding the ordinates corresponding to the 0.025 and 0.975 percentiles of a Student-$t$ distribution, $St(n-1)$ (rather than a $N(0, 1)$) distribution.

**EXAMPLE :** A sample of size 10 has sample mean $\overline{x} = 19.7$. and $s^2 = 0.78^2$. To test

$$H_0 : \mu = 20.0$$
$$H_1 : \mu \neq 20.0$$

under the assumption that the data follow a Normal distribution with $\sigma$ unknown. We have test statistic $t$ given by

$$t = \frac{19.7 - 20.0}{0.78/\sqrt{10}} = -1.22.$$

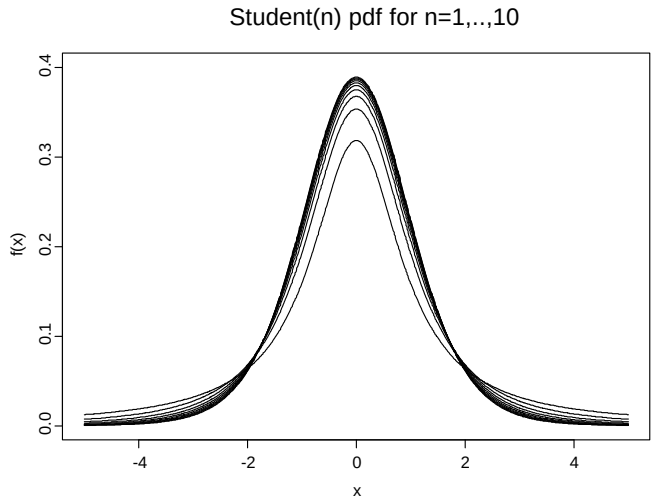



Figure 3.2: Student-t distribution for different values of the degrees of freedom.

The upper critical value $C_R$ is obtained by solving $F_{St(n-1)}(C_R) = 0.975$, where $F_{St(n-1)}$ is the c.d.f. of a Student-$t$ distribution with $n-1$ degrees of freedom; here $n = 10$, so we can use the statistical tables to find $C_R = 2.262$, and not that, as Student-$t$ distributions are symmetric the lower critical value is $-C_R$. Thus $t$ lies between the critical values, and therefore we have no reason to reject $H_0$. The p-value is given by

$$p = \begin{cases} 2F_{St(n-1)}(t) & t < 0 \\ 2(1 - F_{St(n-1)}(t)) & t \geq 0 \end{cases}$$

so here, $p = 2F_{St(n-1)}(-1.22)$ which we can find to give $p = 0.253$; this confirms that we have no reason to reject $H_0$.

### 3.5.4 TESTS FOR NORMAL DATA III - TESTING $\sigma$.

The Z-test and T-test are both tests for the parameter $\mu$. Suppose that we wish to test a hypothesis about $\sigma$, for example

$$H_0 : \sigma^2 = c$$
$$H_1 : \sigma^2 \neq c$$

We construct a test based on the estimate of variance, $s_2$. In particular, we saw in a previous Theorem that the random variable $Q$, defined by

$$Q = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

if the data have an $N(\mu, \sigma^2)$ distribution. Hence if we define test statistic $q$ by

$$q = \frac{(n-1)s^2}{c}$$

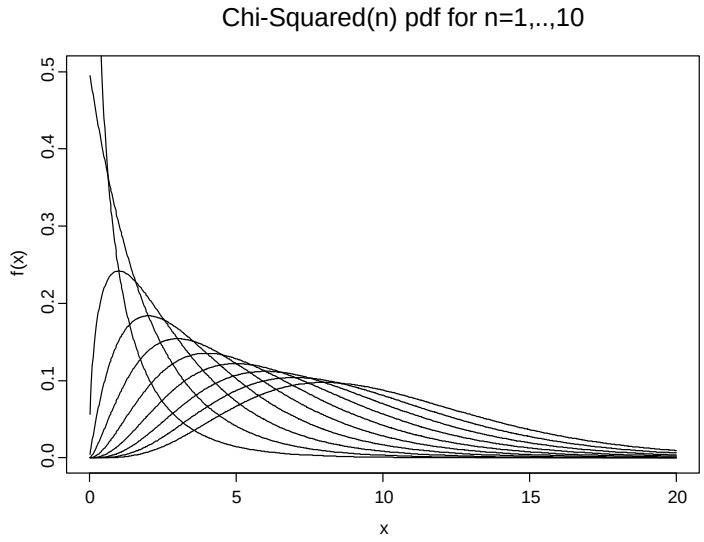



Figure 3.3: Chisquared distribution for different values of the degrees of freedom.

then we can compare $q$ with the critical values derived from a $\chi^2_{n-1}$ distribution; we look for the 0.025 and 0.975 quantiles - note that the Chi-squared distribution is not symmetric, so we need two distinct critical values.

In the above example, to test

$$H_0 : \sigma^2 = 1.0$$
$$H_1 : \sigma^2 \neq 1.0$$

we compute test statistic

$$q = \frac{(n-1)s^2}{c} = \frac{90.78^2}{1.0} = 5.43.75$$

and compare with

$$C_{R_1} = F_{\chi^2_{n-1}}(0.025) \Longrightarrow C_{R_1} = 2.700 \qquad C_{R_2} = F_{\chi^2_{n-1}}(0.975) \Longrightarrow C_{R_2} = 19.022$$

so $q$ is not a surprising observation from a $\chi^2_{n-1}$ distribution, and hence we cannot reject $H_0$.

### 3.5.5 TWO SAMPLE TESTS

It is straightforward to extend the ideas from the previous sections to two sample situations where we wish to compare the distributions underlying two data samples. Typically, we consider sample one, $x_1, ..., x_{n_X}$, from a $N(\mu_X, \sigma^2_X)$ distribution, and sample two, $y_1, ..., y_{n_Y}$, independently from a $N(\mu_Y, \sigma^2_Y)$ distribution, and test the equality of the parameters in the two models. Suppose that the sample mean and sample variance for samples one and two are denoted $(\overline{x}, s^2_X)$ and $(\bar{y}, s^2_Y)$ respectively.

1. First, consider testing the hypothesis

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

when $\sigma_X = \sigma_Y = \sigma$ is known. Now, we have from the sampling distributions theorem we have

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma^2}{n_X}\right) \qquad \overline{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{n_Y}\right) \Longrightarrow \overline{X} - \overline{Y} \sim N\left(0, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}\right)$$

and hence

$$Z = \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim N(0,1)$$

giving us a test statistic $z$ defined by

$$z = \frac{\overline{x} - \overline{y}}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

which we can compare with the standard normal distribution; if $z$ is a surprising observation from $N(0,1)$, and lies outside of the critical region, then we can reject $H_0$. This procedure is the **Two Sample Z-Test**.

2. If $\sigma_X = \sigma_Y = \sigma$ is unknown, we parallel the one sample T-test by replacing $\sigma$ by an estimate in the two sample Z-test. First, we obtain an estimate of $\sigma$ by "pooling" the two samples; our estimate is the **pooled estimate**, $s_P^2$, defined by

$$s_P^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$$

which we then use to form the test statistic $t$ defined by

$$t = \frac{\overline{x} - \overline{y}}{s_P\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

It can be shown that, if $H_0$ is true then $t$ should be an observation from a Student-$t$ distribution with $n_X + n_Y - 2$ degrees of freedom. Hence we can derive the critical values from the tables of the Student-$t$ distribution.

3. If $\sigma_X \neq \sigma_Y$, but both parameters are known, we can use a similar approach to the one above to derive test statistic $z$ defined by

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

which has an $N(0,1)$ distribution if $H_0$ is true.

4. If $\sigma_X \neq \sigma_Y$, but both parameters are unknown, we can use a similar approach to the one above to derive test statistic $t$ defined by

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

for which the distribution if $H_0$ is true is not analytically available, but can be adequately approximated by a $Student\,(m)$ distribution, where

$$m = \frac{(w_X + w_Y)^2}{\left(\frac{w_X^2}{n_X - 1} + \frac{w_Y^2}{n_Y - 1}\right)}$$

where

$$w_X = \frac{s_X^2}{n_X} \qquad w_Y = \frac{s_Y^2}{n_Y}$$

Clearly, the choice of test depends on whether $\sigma_X = \sigma_Y$ or otherwise; we may test this hypothesis formally; to test

$$H_0 : \sigma_X = \sigma_Y$$
$$H_1 : \sigma_X \neq \sigma_Y$$

we compute the test statistic $q = s_X^2/s_Y^2$, which has a null distribution known as the **Fisher** or $F$ distribution with $(n_X - 1, n_Y - 1)$ degrees of freedom; this distribution can be denoted $F(n_X - 1, n_Y - 1)$, and its quantiles are tabulated. Hence we can look up the 0.025 and 0.975 quantiles of this distribution (the $F$ distribution is not symmetric), and hence define the critical region; informally, if the test statistic $q$ is very small or very large, then it is a surprising observation from the $F$ distribution and hence we reject the hypothesis of equal variances.

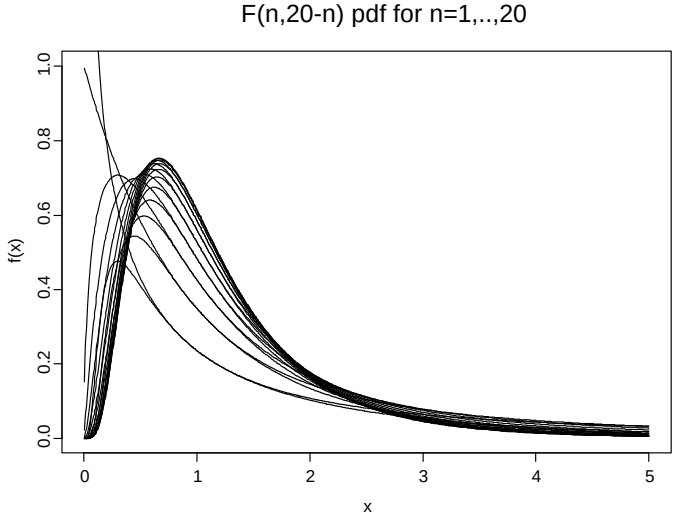



Figure 3.4 : F distribution for different values of the degrees of freedom.

### 3.5.6 ONE-SIDED AND TWO-SIDED TESTS

So far we have considered hypothesis tests of the form

$$\begin{aligned} H_0 &: \mu = c \\ H_1 &: \mu \neq c \end{aligned}$$

which is referred to as a **two-sided test**, that is, the alternative hypothesis is supported by an extreme test statistic in **either** tail of the distribution. We may also consider a **one-sided test** of the form

$$\begin{array}{ll} H_0 : \mu = c & \\ H_1 : \mu > c & \end{array} \quad or \quad \begin{array}{ll} H_0 : \mu = c & \\ H_1 : \mu < c & \end{array} .$$

Such a test proceeds exactly as the two-sided test, except that a significant result can only occur in the right (or left) tail of the null distribution, and there is a single critical value, placed, for example, at the 0.95 (or 0.05) probability point of the null distribution.

### 3.5.7 CONFIDENCE INTERVALS

The procedures above allow us to test specific hypothesis about the parameters of probability models. We may complement such tests by reporting a **confidence interval**, which is an interval in which we believe the "true" parameter lies with high probability. Essentially, we use the sampling distribution to derive such intervals. For example, in a one sample Z-test, we saw that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

that is, that, for critical values $\pm C_R$ in the test at the 5 % significance level

$$P\left[-C_R \leq Z \leq C_R\right] = P\left[-C_R \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq C_R\right] = 0.95$$

Now, from tables we have $C_R = 1.96$, so re-arranging this expression we obtain

$$P\left[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

from which we deduce a **95 % Confidence Interval** for $\mu$ based on the sample mean $\overline{x}$ of

$$\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

We can derive other confidence intervals (corresponding to different significance levels in the equivalent tests) by looking up the appropriate values of the critical values. The general approach for construction of confidence interval for generic parameter $\theta$ proceeds as follows. From the modelling assumptions, we derive a **pivotal quantity**, that is, a statistic, $T_{PQ}$, say, (usually the test statistic random variable) that depends on $\theta$, but whose sampling distribution is "parameter-free" (that is, does not depend on $\theta$). We then look up the critical values $C_{R_1}$ and $C_{R_2}$, such that

$$P\left[C_{R_1} \leq T_{PQ} \leq C_{R_2}\right] = 1 - \alpha$$

where $\alpha$ is the significance level of the corresponding test. We then rearrange this expression to the form

$$P\left[c_1 \leq \theta \leq c_2\right] = 1 - \alpha$$

where $c_1$ and $c_2$ are functions of $C_{R_1}$ and $C_{R_2}$ respectively. Then a $1-\alpha$ % Confidence Interval for $\theta$ is $[c_1, c_2]$.

## SUMMARY

For the tests discussed in previous sections, the calculation of the form of the confidence intervals is straightforward: in each case, $C_{R_1}$ and $C_{R_2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of the pivotal quantity.

### ONE SAMPLE TESTS

| Test | PivotalQuantity $T_{PQ}$ | Null Distn. | Parameter | Confidence Interval |
|---|---|---|---|---|
| $Z-TEST$ | $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ | $N(0,1)$ | $\mu$ | $\overline{x} \pm C_R\ \sigma/\sqrt{n}$ |
| $T-TEST$ | $T = \dfrac{\overline{X}-\mu}{s/\sqrt{n}}$ | $St(n-1)$ | $\mu$ | $\overline{x} \pm C_R\ s/\sqrt{n}$ |
| $Q-TEST$ | $Q = \dfrac{(n-1)s^2}{\sigma^2}$ | $\chi^2_{n-1}$ | $\sigma^2$ | $\left[\dfrac{(n-1)s^2}{C_{R_2}} : \dfrac{(n-1)s^2}{C_{R_1}}\right]$ |

### TWO SAMPLE TESTS

| Test | PivotalQuantity $T_{PQ}$ | Null Distn. | Parameter | Confidence Interval |
|---|---|---|---|---|
| $Z-TEST(1)$ | $Z = \dfrac{(\overline{X}-\mu_X)-(\overline{Y}-\mu_Y)}{\sigma\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}}$ | $N(0,1)$ | $\mu_X-\mu_Y$ | $(\overline{x}-\overline{y}) \pm C_R\ \sigma\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}$ |
| $T-TEST$ | $T = \dfrac{(\overline{X}-\mu_X)-(\overline{Y}-\mu_Y)}{s_P\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}}$ | $St(n_X+n_Y-2)$ | $\mu_X-\mu_Y$ | $(\overline{x}-\overline{y}) \pm C_R\ s_P\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}$ |
| $Z-TEST(2)$ | $Z = \dfrac{(\overline{X}-\mu_X)-(\overline{Y}-\mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_X}+\dfrac{\sigma_Y^2}{n_Y}}}$ | $N(0,1)$ | $\mu_X-\mu_Y$ | $(\overline{x}-\overline{y}) \pm C_R\sqrt{\dfrac{\sigma_X^2}{n_X}+\dfrac{\sigma_Y^2}{n_Y}}$ |
| $Q-TEST$ | $Q = \dfrac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$ | $F(n_X-1, n_Y-1)$ | $\dfrac{\sigma_X^2}{\sigma_Y^2}$ | $\left[\dfrac{s_X^2}{C_{R_2}s_Y^2} : \dfrac{s_X^2}{C_{R_1}s_Y^2}\right]$ |

## 3.6 MODEL TESTING AND VALIDATION

Techniques used for estimation and hypothesis testing allow specific and quantitative questions about the parameters in a probability model to be posed and resolved on the basis of a collection of sample data $x_1, ..., x_n$. However, the question as to the validity of the assumed probability model (for example, Binomial, Poisson, Exponential, Normal etc.) has yet to be addressed.

### 3.6.1 PROBABILITY PLOTS

The probability plotting technique involves comparing *predicted* and *observed* behaviour by comparing quantiles of the proposed probability distribution with sample quantiles. Suppose that a sample of data of size $n$ are to be modelled using a proposed probability model with c.d.f. $F_X$ which possibly depends on unknown parameter(s) $\theta$. The sample data are first sorted into ascending order, and then the $i$th datum, $x_i$, corresponds to the $100i/(n+1)$th quantile of the sample. Now, the equivalent *hypothetical* quantile of the distribution, $q_i$ is found as the solution of

$$F_X(q_i) = \frac{i}{n+1} \qquad i = 1, ..., n.$$

If the model encapsulated in $F_X$ is an acceptable model for the sample data, then for large $n$, $x_i \approx q_i$, so a plot of $\{(q_i, x_i) : i = 1, ..., n\}$ should be a straight line through the origin with slope 1. Hence the validity of $F_X$ as a model for the sample data can be assessed through such a plot.

**EXAMPLE** For the *Exponential*(1) model, $F_X$ is given by

$$F_X(x) = 1 - e^{-x} \qquad x \geq 0$$

so the probability plot consists of examining $\{(q_i, x_i) : i = 1, ..., n\}$ where

$$1 - e^{-q_i} = \frac{i}{n+1} \Longrightarrow q_i = -\log\left\{1 - \frac{i}{n+1}\right\}$$

**EXAMPLE** For the $N(0,1)$ model, $F_X \equiv \Phi$ is only available numerically (for example via statistical tables). Here the probability plot consists of examining $\{(q_i, x_i) : i = 1, ..., n\}$ where

$$\Phi(q_i) = \frac{i}{n+1} \Longrightarrow q_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

**EXAMPLE** For the *Exponential*($\lambda$) model, we plot $\{(q_i, x_i) : i = 1, ..., n\}$ where

$$F_X(q_i) = 1 - e^{-\lambda q_i} = \frac{i}{n+1} \Longrightarrow q_i = -\frac{1}{\lambda}\log\left\{1 - \frac{i}{n+1}\right\}.$$

Hence, if we define $q_i^*$ by

$$q_i^* = -\log\left\{1 - \frac{i}{n+1}\right\}$$

then if the model is correct, a plot of $\{(q_i^*, x_i) : i = 1, ..., n\}$ should be approximately a straight line through the origin with slope $1/\lambda$; hence $\lambda$ can be estimated from this plot by using linear regression.

**EXAMPLE** For the $N(\mu, \sigma^2)$ model, is again only available numerically (for example via statistical tables). Here the probability plot consists of examining $\{(q_i, x_i) : i = 1, ..., n\}$ where

$$F_X(q_i) = \Phi\left(\frac{q_i - \mu}{\sigma}\right) = \frac{i}{n+1} \Longrightarrow q_i = \mu + \sigma\Phi^{-1}\left(\frac{i}{n+1}\right).$$

Hence, if we define $q_i^*$ by

$$q_i^* = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

then if the model is correct, a plot of $\{(q_i^*, x_i) : i = 1, ..., n\}$ should be approximately a straight line with intercept $\mu$ and slope $\sigma$; hence $\mu, \sigma$ can again be estimated from this plot by using linear regression.

### 3.6.2 THE CHI-SQUARED GOODNESS-OF-FIT TEST

The problem of testing a hypothesis as to whether a data sample $x_1, ..., x_n$ is well-modelled by a specified probability distribution can be approached from a "goodness-of-fit" perspective.

Suppose that the data are recorded as the number of observations, $O_i$, say in a sample of size $n$ that fall into each of $k$ categories or "bins". Suppose that under the hypothesized model with mass/density function $f_X$ or c.d.f. $F_X$, the data follow a specific probability distribution specified by probabilities $\{p_i : i = 1, ..., k\}$. These probabilities can be calculated directly from $f_X$ or $F_X$, possibly after parameters in the model have been estimated using maximum likelihood. Then, if the hypothesized model is correct, $E_i = np_i$ observations would be expected to fall into category $i$. An intuitively sensible measure of the **goodness-of-fit** of the data to the hypothesized distribution is given by the **chi-squared statistic**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

A formal hypothesis test of model adequacy can be carried out in the usual framework; here the chi-squared statistic is the test statistic, and the null distribution (the distribution of the test statistic if the hypothesis is TRUE) is approximately a **chi-squared distribution** with $k - d - 1$ degrees of freedom, where $d$ is the number of parameters in $f_X$ or $F_X$ that were estimated in order calculate the probabilities $p_1, ..., p_k$.

**EXAMPLE : Testing the fit of a Poisson distribution**

An early experiment into the properties of radioactive materials involved counting the number of alpha particles emitted from a radioactive source in 2612 consecutive 7.5 second intervals. A total of 10126 particles were counted, and the observed frequencies for each of the numbers of counts (per 7.5s) from 0 to 12 were recorded.

| Count | $O_i$ | $p_i$ | $E_i$ | $(O_i - E_i)^2/E_i$ |
|---|---|---|---|---|
| 0 | 57 | 0.021 | 54 | 0.167 |
| 1 | 204 | 0.080 | 210 | 0.171 |
| 2 | 383 | 0.156 | 407 | 1.415 |
| 3 | 525 | 0.201 | 525 | 0.000 |
| 4 | 532 | 0.195 | 510 | 0.949 |
| 5 | 408 | 0.151 | 395 | 0.428 |
| 6 | 273 | 0.098 | 255 | 1.271 |
| 7 | 139 | 0.054 | 141 | 0.028 |
| 8 | 49 | 0.026 | 68 | 5.309 |
| 9 | 27 | 0.011 | 30 | 0.300 |
| 10 | 10 | 0.004 | 11 | 0.091 |
| 11 | 4 | 0.002 | 4 | 0.000 |
| 12 | 2 | 0.000 | 1 | 1.000 |
| >12 | 0 | 0.001 | 1 | 1.000 |
| Total | 2612 | 1.000 | 2612 | 12.129 |

To test the hypothesis that the data follow a Poisson distribution, a chi-squared test can be performed. First, we estimate Poisson parameter $\lambda$ by its m.l.e., which is $\hat{\lambda} = \overline{x} = 10126/2612 = 3.877$. Secondly, we calculate probabilities $p_i$ using the Poisson formula. Thirdly, we calculate the theoretical (expected) frequencies $E_i = np_i$ for each category. Finally, we calculate the $\chi^2$ statistic as the sum of the (standardized) squared differences between observed and expected frequencies.

In this case, $\chi^2 = 12.129$. To complete the test we find that the 95th percentile of a Chi-squared distribution with $k-1-1 = 12$ degrees of freedom is 21.03. This implies that the $\chi^2$ statistic would only be surprising at a significance level of 0.05 if it was larger than 21.03. Here, as $\chi^2 = 12.129$, and therefore not surprising. Hence there is no evidence to indicate that the data are not from a Poisson distribution.

Clearly, the categorization is arbitrary, and several of the categories in example 1 could be combined. As a general rule, the categories should be chosen so that there is at least **five** observed counts in each.

Hence, to carry out a Chi-squared goodness of fit test, we use the following logic. If a given hypothesis is true, it can be shown that the chi-squared statistic $\chi^2$ for a sample of data has a particular Chi-squared distribution. If $\chi^2$ takes a value that is surprising or unlikely under that probability distribution (for example if its value lies in the extreme right-hand tail and is larger, say, than the 95th percentile of the distribution) it is very likely that the hypothesis is false and should be rejected.

## 3.7 HYPOTHESIS TESTING EXTENSIONS

In this section, we examine extensions to the hypothesis testing methods described above: although the methods differ slightly, the general approach remains the unchanged. We proceed using the same strategy, outlined below, For a data sample $x_1, ..., x_n$, with corresponding random variables $X_1, ..., X_n$, we

1. consider a pair of competing **hypotheses**, $H_0$ and $H_1$
2. define a suitable test statistic $T = T(X_1, ..., X_n)$ (that is, some function of the original random variables; this will define the **test statistic**), and a related pivotal random variable $T_{PQ} = T_{PQ}(X)$
3. **assume that $H_0$ is true**, and compute the sampling distribution of $T$, $f_T$ or $F_T$; this is the **null distribution**
4. compute the **observed** value of $T$, $t = T(x_1, ..., x_n)$; this is the **test statistic**
5. assess whether $t$ is a surprising observation from the distribution $f_T$. If it **is** surprising, we have evidence to **reject** $H_0$; if it is not surprising, we **cannot reject** $H_0$

This strategy can be applied to more complicated normal examples, and also non-normal and non-parametric testing situations. It is a general strategy for assessing the statistical evidence for or against a hypothesis.

### 3.7.1 ANALYSIS OF VARIANCE

The first extension we consider still presumes a normality assumption for the data, but extends the ideas from $Z$ and $T$ tests, which compare at most two samples, to allow for the analysis of any number of samples. **Analysis of variance** or **ANOVA** is used to display the sources of variability in a collection of data samples. The ANOVA F-test compares variability **between** samples with the variability **within** samples.

**ONE-WAY ANOVA**

The T-test can be extended to allow a test for differences between more than two data samples. Suppose there are $K$ samples of sizes $n_1, ..., n_K$ from different populations. The model can be represented as follows: let $y_{kj}$ be the $j$th observation in the $k$th sample, then

$$y_{kj} = \mu_k + \varepsilon_{kj}$$

for $k = 1, ..., K$, and $\varepsilon_{kj} \sim N\left(0, \sigma^2\right)$. This model assumes that

$$Y_{kj} \sim N\left(\mu_k, \sigma^2\right)$$

and that the expectations for the different samples are different. We can view the data as a table comprising $K$ columns, with each column corresponding to a sample.

To test the hypothesis that each population has the same mean, that is, the hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = ... = \mu_K \\ H_1 &: not H_0 \end{aligned}$$

an **Analysis of Variance (ANOVA)** F-test may be carried out.

To carry out a test of the hypothesis, the following **ANOVA table** should be completed;

| Source | D.F. | Sum of squares | Mean square | $F$ |
|---|---|---|---|---|
| Between Samples | $K-1$ | $FSS$ | $FSS/(K-1)$ | $\frac{FSS/(K-1)}{RSS/(n-K)}$ |
| Within Samples | $n-K$ | $RSS$ | $RSS/(n-K)$ | |
| Total | $n-1$ | $TSS$ | | |

where $n = n_1 + ... + n_K$, and

$$TSS = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj} - \overline{y}_{..})^2 \qquad RSS = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj} - \overline{y}_{k})^2 \qquad FSS = \sum_{k=1}^{K} n_k(\overline{y}_{k} - \overline{y}_{..})^2$$

where TSS is the **total** sum-of-squares (i.e. total deviation from the overall data mean $\overline{y}_{.}$)RSS is the **residual** sum-of-squares (i.e. sum of deviations from individual sample means $\overline{y}_k$, $k = 1, ..., K$) and FSS is the **fitted** sum-of-squares (i.e. weighted sum of deviations of sample means from the overall data mean, with weights equal to number of data points in the individual samples) Note:that

$$TSS = FSS + RSS$$

If the $F$ statistic is calculated in this way, and compared with an F distribution with parameters $K-1$, $n-K$, the hypothesis that all the individual samples have the same mean can be tested.

**EXAMPLE** Three genomic segments were used to studied in order to discover whether the distances (in kB) between successive occurrences of a particular motif were substantially different. Several measurements were taken using for each segment;

| | Method | | |
|---|---|---|---|
| | SEGMENT A | SEGMENT B | SEGMENT C |
| | 42.7 | 44.9 | 41.9 |
| | 45.6 | 48.3 | 44.2 |
| | 43.1 | 46.2 | 40.5 |
| | 41.6 | | 43.7 |
| | | | 41.0 |
| Mean | 43.25 | 46.47 | 42.26 |
| Variance | 2.86 | 2.94 | 2.66 |

For these data, the ANOVA table is as follows;

| Source | D.F. | Sum of squares | Mean square | $F$ |
|---|---|---|---|---|
| SEGMENTS | 2 | 34.1005 | 17.0503 | 6.11 |
| Residual | 9 | 25.1087 | 2.7899 | |
| Total | 11 | 59.2092 | | |

and the $F$ statistic must be compared with an $F_{2,9}$ distribution. For a significance test at the 0.05 level, $F$ must be compared with the 95th percentile (in a **one-sided** test) of the $F_{2,9}$ distribution. This value is 4.26. Therefore, the $F$ statistic **is** surprising, given the hypothesized model, and therefore there is evidence to reject the hypothesis that the segments are identical.

**TWO-WAY ANOVA**

One-way ANOVA can be used to test whether the underlying means of several groups of observations are equal Now consider the following data collection situation Suppose there are $K$ treatments, and $L$ groups of observations that are believed to have different responses, that all treatments are administered to all groups, and measurement samples of size $n$ are made for each of the $K \times L$ combinations of treatments $\times$ groups. The experiment can be represented as follows: let $y_{klj}$ be the $j$th observation in the $k$th treatment on the $l$th group, then

$$y_{klj} = \mu_k + \delta_l + \varepsilon_{klj}$$

for $k = 1, ..., K$, $l = 1, ..., L$, and again $\varepsilon_{klj} \sim N\left(0, \sigma^2\right)$. This model assumes that $Y_{kj} \sim N\left(\mu_k + \delta_l, \sigma^2\right)$ and that the expectations for the different samples are different. We can view the data as a 3 dimensional-table comprising $K$ columns and $L$ rows, with $n$ observations for each column $\times$ row combination, corresponding to a sample.

It is possible to test the hypothesis that each **treatment**, and/or that each **group** has the same mean, that is, the two null hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = ... = \mu_K \\ H_0 &: \delta_1 = \delta_2 = ... = \delta_K \end{aligned}$$

against the alternative $H_1$ :not $H_0$ in each case. For these tests, a **Two-way Analysis of Variance (ANOVA)** F-test may be carried out. The Two-Way ANOVA table is computed as follows

| Source | D.F. | Sum of squares | Mean square | $F$ |
|---|---|---|---|---|
| TREATMENTS | $K-1$ | $FSS_1$ | $FSS_1/(K-1)$ | $\frac{FSS_1/(K-1)}{RSS/(R+1)}$ |
| GROUPS | $L-1$ | $FSS_2$ | $FSS_2/(L-1)$ | $\frac{FSS_2/(L-1)}{RSS/(R+1)}$ |
| Residual | $R+1$ | $RSS$ | $RSS/(R+1)$ | |
| Total | $N-1$ | $TSS$ | | |

where $N = K \times L \times n$, $R = N - L - K$. and again

$$TSS = FSS_1 + FSS_2 + RSS.$$

In the table below, there are $K = 6$ Treatments, and $L = 3$ Groups, and $n = 1$

| | I | II | III | GROUP totals |
|---|---|---|---|---|
| 1 | 0.96 | 0.94 | 0.98 | 2.88 |
| 2 | 0.96 | 0.98 | 1.01 | 2.95 |
| 3 | 0.85 | 0.87 | 0.86 | 2.58 |
| 4 | 0.86 | 0.84 | 0.90 | 2.60 |
| 5 | 0.86 | 0.87 | 0.89 | 2.62 |
| 6 | 0.89 | 0.93 | 0.92 | 2.74 |
| TREATMENT totals | 5.38 | 5.43 | 5.56 | 16.37 |

There are two natural hypotheses to test; first, do the TREATMENTS differ, and second, do the GROUPS differ ?

**Two-way analysis of variance,** with the rows and columns representing two source of variation can be used to analyze such data. Two-way analysis of variance studies the variability due to

- the GROUPS effect, and
- the TREATMENTS effect,

and calibrates them against the average level of variability in the data overall.. For example, for the data above we have the following two-way ANOVA table

| Source | D.F. | Sum of squares | Mean square | $F$ |
|---|---|---|---|---|
| TREATMENT | 5 | 0.040828 | 0.0081656 | 31.54 |
| GROUP | 2 | 0.002878 | 0.001439 | 5.57 |
| Residual | 10 | 0.002589 | 0.0002589 | |
| Total | 17 | 0.046295 | | |

The two $F$ statistics can be interpreted as follows

- The first $F$ statistic ($F = 31.54$) is the test statistic for the test of equal means in the **rows**, that is, that there is no difference between **TREATMENTS**. This statistic must be compared with an $F_{5,10}$ distribution (the two degrees of freedom being the entries in the degrees of freedom column in the specimens and residual rows of the ANOVA table). The 95th percentile of the $F_{5,10}$ distribution is 3.33, and thus the test statistic is **more extreme** than this critical value, and thus the hypothesis that each specimen has the same mean can be **rejected**.
- The second $F$ statistic, ($F = 5.57$), is the test statistic for the test of equal means in the **columns**, that is, that there is no difference between **GROUPS**. This statistic must be compared with an $F_{2,10}$ distribution (the two degrees of freedom being the entries in the degrees of freedom column in the methods and residual rows of the ANOVA table). The 95th percentile of the $F_{2,10}$ distribution is 4.10, and thus the test statistic is **more extreme** than this critical value, and thus the hypothesis that each method has the same mean can be **rejected**.

If replicate data are available, it is possible also to fit an **interaction**, that is, to discover whether the pattern of variability is significantly different amongst the different TREATMENTS or GROUPS.

ANOVA $F$ tests allow the comparison of **between group** and **within group** variability

- significant **between group** variability indicates a systematic difference between the groups
- if the standardized ratio of **between group variability to within group variability** is large, then there is evidence of a systematic variation.
- in effect, ANOVA compares estimates of **variance components.**

### 3.7.2 NON-NORMAL DATA: COUNTS AND PROPORTIONS

A common form of non-normal data arise when the counts of numbers of "successes" or "failures" that arise in a fixed number of trials. In this case, the Binomial distribution model is appropriate;

- in **one sample** testing, we model the number of successes, $X$, by assuming $X \sim Binomial(n, \theta)$ and test hypotheses about $\theta$,
- in **two sample** testing, we assume that the number of successes in the two samples are random variables $X_1$ and $X_2$, where $X_1 \sim Binomial(n_1, \theta_1)$ and $X_2 \sim Binomial(n_2, \theta_2)$, for example test $H_0 : \theta_1 = \theta_2$ against some alternative hypothesis ($\theta_1 \neq \theta_2, \theta_1 > \theta_2$ or $\theta_1 < \theta_2$)

**ONE-SAMPLE TESTING:**

In the one sample case, two alternative approaches can be adopted. The first is a so-called **exact** test, where the distribution of the chosen test statistic under $H_0 : \theta = c$ is computed exactly, giving exact critical values and $p$-values. The second is a approximate test based on a Normal approximation to the binomial distribution. For the **exact** test, we note that, **if $H_0$ is true**, and $\theta = c$, then $X \sim Binomial(n, c)$ so the critical values in a two-sided test can be computed directly by inspection of the $Binomial(n, c)$ c.d.f; that is

$$F_{BIN}\left(C_{R_1}; n, \theta = c\right) = 0.025 \qquad C_{R_2} = F_{BIN}\left(0.975; n, \theta = c\right)$$

where $F_{BIN}\left(-; n, \theta\right)$ is the c.d.f. of the $Binomial(n, \theta)$ distribution

$$F_{BIN}\left(x; n, \theta\right) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} \theta^i \left(1-\theta\right)^{n-i}$$

and $\lfloor x \rfloor$ is the smallest whole number not greater than $x$. For the **approximate** test, we use the fact that

$$X \sim Binomial(n, \theta) \approx Normal\left(n\theta, n\theta\left(1-\theta\right)\right)$$

and hence random variable $Z$

$$Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

is approximately distributed as $Normal(0, 1)$. For the approximate test of $H_0 : \theta = c$, we therefore use the test statistic

$$z = \frac{x - nc}{\sqrt{nc(1-c)}}$$

($x$ is the observed count) and compare this with the standard normal c.d.f.

**TWO SAMPLE TESTING:**

For a two sample test of $H_0 : \theta_1 = \theta_2$, we use a similar normal approximation to the one-sample case. If $H_0$ is true, then there is a common probability $\theta$ determining the success frequency in both samples, and the maximum likelihood estimate of $\theta$ is

$$\widehat{\theta} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{x}{n}, say$$

and thus it can be shown that the test statistic.

$$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\frac{(n_1+n_2)}{n_1 n_2}\left(\frac{x_1+x_2}{n_1+n_2}\right)\left(1-\frac{x_1+x_2}{n_1+n_2}\right)}}$$

has an approximate standard Normal distribution.

### 3.7.3 CONTINGENCY TABLES AND THE CHI-SQUARED TEST

**Contingency tables** are constructed when a sample of data of size $n$ are cross-classified according to $D$ factors, with factor $d$ having $k_d$ categories, for $d = 1, ..., D$. The cross-classification can be represented by a $D$-way table of $k_1 \times k_2 \times ... \times k_D$ "cells", with each cell containing a fraction of the original data. Such a table when $D = 2$, $k_1 = 4$ and $k_2 = 6$ is illustrated below

| | | COLUMN | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| ROW | 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{14}$ | $n_{15}$ | $n_{16}$ | $n_{1.}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{24}$ | $n_{25}$ | $n_{26}$ | $n_{2.}$ |
| | 3 | $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_{34}$ | $n_{35}$ | $n_{36}$ | $n_{3.}$ |
| | 4 | $n_{41}$ | $n_{42}$ | $n_{43}$ | $n_{44}$ | $n_{45}$ | $n_{46}$ | $n_{4.}$ |
| | Total | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | $n_{.4}$ | $n_{.5}$ | $n_{.6}$ | $n$ |

where $\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} n_{ij} = n$.It is often of interest to test whether row classification is **independent** of column classification, as this would indicate **independence** between row and column factors. An approximate test of this hypothesis can be carried out using a **Chi-Squared Goodness-of-Fit** statistic; if the independence model is correct, the expected cell frequencies $\hat{n}_{ij}$ can be calculated as

$$\hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n} \qquad i = 1, ..., k_1,\ j = 1, ..., k_2$$

where $n_{i.}$ is the *total* of cell counts in row $i$ and $n_{.j}$ is the *total* of cell counts in column $j$, and that, under independence, the $\chi^2$ test statistic

$$\chi^2 = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

has an approximate chi-squared distribution with $(k_1 - 1)(k_2 - 1)$ degrees of freedom.

Another approximate test is based on a **Likelihood Ratio** (**LR**) statistic

$$LR = 2\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}}$$

This statistic also has an approximate Chi-squared distribution $\chi^2_{(k_1-1)(k_2-1)}$distribution, again given that $H_0$ **is true.**

We will see further analysis of count data in section (3.7.3) below, and in Chapter 4

### 3.7.4 $2 \times 2$ TABLES

When $k_1 = k_2 = 2$, the contingency table reduces to a two-way binary classification

| | | COLUMN | | |
|---|---|---|---|---|
| | | 1 | 2 | Total |
| ROW | 1 | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| | Total | $n_{.1}$ | $n_{.2}$ | $n$ |

In this case we can obtain some more explicit tests: one is again an **exact test**, the other is based on a normal approximation. The chi-squared test described above is feasible, but other tests may also be constructed:

- **FISHER'S EXACT TEST FOR INDEPENDENCE**

  Suppose we wish to test for **independence** between the row and column variables of a contingency table. When the data consist of two categorical variables, a contingency table can be constructed reflecting the number of occurrences of each factor combination. **Fisher's exact test** assesses whether the classification according to one factor is independent of the classification according to the other, that is the test is of the null hypothesis $H_0$ that the factors are independent, against the general alternative, **under the assumption that the row and column totals are fixed**.

  The data for such a table comprises the row and column totals $(n_{1.}, n_{2.}, n_{.1}, n_{.2})$ and the cell entries $(n_{11}, n_{12}, n_{21}, n_{22})$. The test statistic can be defined as the upper left cell entry $n_{11}$; for the null distribution, we compute the probability of the observing **all possible tables** with these row and column totals.. Under $H_0$ this distribution is **hypergeometric** and the probability of observing the table $(n_{11}, n_{12}, n_{21}, n_{22})$ is

$$p(n_{11}) = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n}{n_{.1}}} = \frac{n_{1.}!n_{.1}!n_{2.}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

  where $n! = 1 \times 2 \times 3 \times .. \times (n-1) \times n$.

  For the $p$-value, we need to assess the whether or not the observed table is surprising under this null distribution; suppose we observe $n_{11} = x$, then we can compare $p(x)$ with all $p(y)$ for all feasible $y$, that is $y$ in the range $\max\{0, n_{1.} - (n - n_{.1})\} \leq y \leq \min\{n, n_{.1}\}$. We are thus calculating the null distribution **exactly** given the null distribution assumptions and the row and column totals; if the observed test statistic lies in the tail of the distribution, we can reject the null hypothesis of independent factors.

- **McNEMAR'S TEST FOR SYMMETRY IN PAIRED SAMPLES**

  In a $2 \times 2$ table representing paired data (where observations are, for example, matched in terms of medical history or genotype, or phenotype) the usual chi-squared test is not appropriate, and **McNemar's test** can instead be used. Consider the following table for a total of $n$ matched pairs of observations, in which each individual in the pair has been classified (or randomized to class) A or B, with one A one B in each pair, and then the

outcome (disease status, survival status) recorded.

| | | A | | |
|---|---|---|---|---|
| | | YES | NO | Total |
| | YES | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| B | NO | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| | Total | $n_{.1}$ | $n_{.2}$ | $n$ |

that is, $n_{11}$ pairs were observed for which both A and B classified individuals had disease/survival status YES, whereas $n_{12}$ pairs were observed for which the A individual had status NO, but the B individual had status YES, and so on.

An appropriate test statistic here for a test of symmetry or "discordancy" in these results (that is, whether the two classifications are significantly different in terms of outcome) is

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

which effectively measures how different the off-diagonal entries in the table are. This statistic is an adjusted Chi-squared statistic, and has a $\chi^2_1$ distribution under the null hypothesis that there is no asymmetry. Again a one-tailed test is carried out: "surprising" values of the test statistic are large.

### 3.7.5 NON-PARAMETRIC TESTS

The standard test for the equality of expectations of two samples is the two-sample T-test. This test is predicated on the assumption of normality of the underlying distributions. In many cases, such an assumption is inappropriate, possible due to distributional asymmetry or the presence of outliers, and thus other tests of the hypothesis of equality of population locations must be developed.

Some of the standard non-parametric tests used in statistical analysis are described below: we concentrate on two-sample tests for the most part. All of these tests can be found in good statistics packages.

- **THE MANN-WHITNEY-WILCOXON TEST**

  Consider two samples $x_1, ..., x_{n_1}$ and $y_1, ..., y_{n_2}$. The **Mann-Whitney-Wilcoxon** test proceeds as follows; first, sort the pooled sample into ascending order. Add up the ranks of the data from sample one to get $u_1$ say. Repeat for sample two to get $u_2$. Note that

  $$u_1 + u_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

  The Mann-Whitney-Wilcoxon statistic is $u_1$. It can be shown that, under the hypothesis that the data are from populations with the equal medians, then $u_1$ has an approximate normal distribution with mean and variance

  $$\frac{n_1(n_1 + n_2 + 1)}{2} \qquad \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

  This is the non-parametric alternative to the two sample t-test.

- **THE KOLMOGOROV-SMIRNOV TEST**

  The two-sample Kolmogorov-Smirnov test is a non-parametric test for comparing two samples via their **empirical cumulative distribution function**. For data $x_1, ..., x_n$, the empirical c.d.f. is the function $\widehat{F}$ defined at $x$ by

  $$\widehat{F}(x) = \frac{c(x)}{n} \qquad\qquad c(x) = \text{“}Numberofdata \leq x\text{”}$$

  Thus, for two samples, we have two empirical c.d.f.s $\widehat{F}_1(x)$ and $\widehat{F}_2(x)$. The (two-sided) **Kolmogorov-Smirnov** test of the hypothesis that the two samples come from the same underlying distribution, $H_0 : F_1 = F_2$, is based on the statistic

  $$T = \max_x \left| \widehat{F}_1(x) - \widehat{F}_2(x) \right|.$$

  It is easy to show that $0 \leq T \leq 1$, but the null distribution of $T$ is not available in closed form. Fortunately, the $p$-value probability in the test for test statistic $t$, $p = \mathrm{P}[T > t]$ can be obtained for various different sample sizes using statistical tables or packages.

- **THE CHI-SQUARED GOODNESS-OF-FIT TEST**

  It is often required to test whether a sample can be well modelled using a particular distribution; the **chi-squared goodness-of-fit test** is the most commonly used test. It is a non-parametric test for which the null distribution can be well approximated by a Chi-squared distribution. It is studied in more detail in the *Model Validation* section below.

- **THE SHAPIRO-WILK TEST FOR NORMALITY**

  It is often required to test whether a sample of data are normally distributed; if they are, then many of the tests described above can be utilized. The **Shapiro-Wilk** test can be used to test this hypothesis; the test statistic is commonly denoted $W$, and critical and $p$- values from its null distribution are available from tables or statistics packages.

- **THE KRUSKAL-WALLIS TEST**

  The **Kruskal-Wallis rank test** is a nonparametric alternative to a one-way analysis of variance. The null hypothesis is that the true location parameter is the same in each of the samples. The alternative hypothesis is that at least one of the samples has a different location. Unlike one-way ANOVA, this test does not require normality

- **THE FRIEDMAN RANK SUM TEST**

  The **Friedman rank sum test** is a nonparametric alternative to a two-way analysis of variance. It is appropriate for data arising from an experiment in which exactly one observation was collected from each experimental unit, or group, under each treatment. The elements of the samples are assumed to consist of a treatment effect, plus a group effect, plus independent and identically distributed residual errors

### 3.7.6 EXACT TESTS

In the above sections, we have seen Chi-squared tests being used to test hypotheses about data. These tests involved the construction of a chi-squared statistic, of the form

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where $O_i$ and $E_i$ are the observed and expected/predicted counts in each of $k$ (cross) classifications or categories (see section (3.6.2) for further details).

The distribution of the test-statistic is typically approximated by a Chi-squared distribution with an appropriately chosen degrees of freedom. This approximation is good when the sample size is large, but not good when the table is "sparse", with some low (expected) cell entries (under the null hypothesis). The approximation breaks down for small sample sizes due to the inappropriateness of the Normal approximation referred to in section (3.7.2)

We have also seen two examples of **Exact Tests:** the exact binomial test in section (3.7.2) and Fisher's Exact Test in section (3.7.4). For these tests, we proceeded as follows, mimicking the general hypothesis strategy outlined at the start of the section.

1. Write down a null hypothesis $H_0$ and a suitable alternative hypothesis $H_1$
2. Construct a test statistic $T$ deemed appropriate for the hypothesis under study
3. Compute the null distribution of $T$, that is the sampling distribution of $T$ if $H_0$ is true, $f_T$
4. Compare the observed value of $T$, $t = T(x)$ for sample data $x = (x_1, ..., x_n)$ with the null distribution and assess whether the observed test statistic is a surprising observation from $f_T$; if it is reject $H_0$

Step 3 is crucial: for some tests (for example, one and two sample tests based on the Normal distribution assumption), it is possible to find $f_T$ analytically for appropriate choices of $T$ in Step 2. For others, such as the chi-squared goodness of fit and related tests, $f_T$ is only available approximately. However, the null distribution (and hence the critical regions and $p$-value) can, in theory, **always** be found : it is the probability distribution of the statistic $T$ under the model restriction imposed by the null hypothesis.

**EXAMPLE** Suppose a data sample are collected and believed to be from a $Poisson\,(\lambda)$ distribution, and we wish to test $H_0 : \lambda = 2$. We might regard the sample mean statistic $T = \overline{X}$ as an appropriate test statistic. Then

$$F_T(t;\lambda) = P[T \leq t;\lambda] = P[\overline{X} \leq t;\lambda] = P\left[\sum_{i=1}^{n} X_i \leq nt;\lambda\right] = P\left[Y \leq nt;\lambda\right]$$

where $Y = \sum_{i=1}^{n} X_i$. But elementary probability theory tells us that $Y \sim Poisson(n\lambda)$, so if $H_0$ is true, we have the null distribution c.d.f. as

$$F_T(t;\lambda = 2) = \sum_{x=0}^{\lfloor nt \rfloor} \frac{e^{-2}2^x}{x!}$$

and thus the critical values for the test, and the $p$-value, available numerically.

The main difficulty with exact tests is in the computation of $F_T(t;\lambda)$; this is rarely analytically possible. However, we will see in a later section that this null c.d.f. can be approximated using simulation methods, or **Permutation Tests.**

## 3.8 POWER AND SAMPLE SIZE

In a statistical analysis, it seems reasonable that the **sample size** required to test specific null and alternative hypotheses depends on how small of a difference we are trying to detect, how much variability is inherent in our data, and how certain we want to be of our results. In this section we will quantify the role of each of these aspects in experimental design.

In a classical test of $H_0$ (null hypothesis) versus $H_1$ (alternative hypothesis), there are four possible outcomes, two of which are erroneous:

1. Do not reject $H_0$ when is $H_0$ true.
2. Reject $H_0$ when $H_0$ is false.
3. Reject $H_0$ when $H_0$ is true (**Type I error**).
4. Do not reject $H_0$ when $H_0$ is false (**Type II error**).

Recall that to construct a test, the distribution of the test statistic under $H_0$ is used to find a **critical region** which will ensure the probability of committing a type I error does not exceed some predetermined **significance level**, $\alpha$. The **power**, $\beta$, of the test is its ability to **correctly reject the null hypothesis**, or

$$\beta = 1 - P(TypeII\ Error),$$

which is based on the distribution of the test statistic under $H_1$ .The required sample size is then a function of

- The null and alternative hypotheses;
- The target $\alpha$;
- The desired power to detect $H_1$ ;
- The variability within the population(s) under study.

Our objective here is to find a relationship between the above factors and the sample size that enables us to select a sample size consistent with the desired $\alpha$ and $\beta$.

### 3.8.1 POWER CALCULATIONS FOR NORMAL SAMPLES

In a **one-sample test** of a normal mean, to complete a power/sample size calculation, we first specify the model and hypotheses; we have $X_1, ..., X_n$ as a set of random variables relating to the observed data $x_1, ..., x_n$, and an assumption that

$$X_i \sim N(\mu, \sigma^2)$$

for $i = 1, ..., n$. If $\sigma^2$ is known, to perform a two-sided test of equality the hypotheses would be as follows:

$$\begin{aligned} H_0 &: \mu = c_0 \\ H_1 &: \mu = c_1 \end{aligned}$$

The maximum likelihood estimate of $\mu$ is the sample mean, which is normally distributed, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. The test statistic is

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

and under $H_0$,

$$Z = \frac{\overline{X} - c_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

We reject $H_0$ at significance level $\alpha$ if the $z$ statistic is more extreme than the critical values of the test are

$$c_0 \pm C_R \frac{\sigma}{\sqrt{n}} \qquad\qquad C_R = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Now, if $H_1$ is true, $X \sim N(c_1, \sigma^2)$, and hence

$$Z = \frac{\overline{X} - c_0}{\sigma/\sqrt{n}} \sim N\left(\frac{c_1 - c_0}{\sigma/\sqrt{n}}, 1\right).$$

so the probability that $z$ lies in the critical region is

$$1 - \beta = \Phi\left(\frac{\sqrt{n}\,(c_0 - c_1)}{\sigma} - C_R\right) + \Phi\left(\frac{\sqrt{n}\,(c_1 - c_0)}{\sigma} - C_R\right)$$

Thus for fixed $\alpha, c_0, c_1$ and $n$, we can compute the power. Similar calculations are available for other of the normal distribution-based tests.

In fact. the power equation can be rearranged to be explicit in one of the other parameters if $\beta$ is regarded as fixed. For example, if $\alpha, \beta, c_0$ and $c_1$ are fixed, we can rearrange to get a sample size calculation to test for fixed difference $\Delta = c_1 - c_0$

$$n = \frac{\sigma^2\left(C_R + \Phi^{-1}(1-\beta)\right)^2}{(c_1 - c_0)^2}$$

### 3.8.2 EXTENSIONS: SIMULATION STUDIES

Extension of the power and sample size calculations to more general testing scenarios, such as for two sample data, or ANOVA, or non-normal data is reasonably straightforward. However, for general or complex tests, the calculation can often only proceed by **simulation**. We would begin as usual by considering the null hypothesis, the test statistic and the null distribution to derive the critical values. Then, the probability of correctly rejecting the null hypothesis for a specified alternative by simulating data from the model under the alternative, and recording the proportion of times that the null hypothesis is correctly rejected.

We discuss simulation-based statistical calculations in section (3.10) below.

## 3.9 MULTIPLE TESTING

The multiple testing corrections are used when several independent statistical tests are being performed simultaneously. It is necessary because, while a given significance level $\alpha$ may be appropriate for each individual comparison, it is not for the set of all comparisons

Recall first that the significance level $\alpha$ can be interpreted as the maximum allowable false positive rate in the test of $H_0$, that is the value of

$$\alpha = P(\text{Test rejects } H_0 | H_0 \text{is TRUE})$$

The idea of a multiple testing correction is to control the rate of false positive results when many significance tests are being carried out. In order to avoid a lot of spurious positives, when the null hypothesis is rejected when it is actually true, the $\alpha$ value needs to be lowered to account for the number of comparisons being performed.

### 3.9.1 THE BONFERRONI AND OTHER CORRECTIONS

The simplest and most conservative approach is the Bonferroni correction, which sets the significance level for the entire set of $k$ comparisons, $\alpha_B(k)$, by taking the common significance level $\alpha$ for each comparison and letting

$$\alpha_B(k) = \frac{\alpha}{k}$$

If we have $k$ independent significance tests $T_1, ..., T_k$ for hypotheses $H_0^{(1)}, ...H_0^{(k)}$ at the $\alpha$ level, the probability $p$ that we will **not** reject any of the hypotheses **if they are all true** is merely

$$P\left(\left(\bigcap_{i-1}^{k} \text{Test } T_i \text{does not reject } H_0^{(i)}\right) | \text{All } H_0^{(i)} \text{are TRUE}\right) =$$

$$= \prod_{i=1}^{k} P\left(\text{Test } T_i \text{does not reject } H_0^{(i)} | H_0^{(i)} \text{is TRUE}\right)$$

$$= \prod_{i=1}^{k} (1-\alpha) = (1-\alpha)^k$$

Thus the **actual** significance level for the **series** of tests is $1-(1-\alpha)^k$. For example, with $\alpha = 0.05$ and $k = 10$ we get $p = 0.95^{10} \approx 0.60$. This means, however that

$$P\left(\text{At least one test } T_i \text{ decrees rejection of its } \mathrm{H}_0^{(i)} \ | \textbf{All } H_0^{(i)} \text{are } \textbf{TRUE}\right) = 1 - (1-\alpha)^k = 0.4$$

so that we now have a probability of 0.40 that one of these 10 tests will turn out significant, and one of the $H_0$ will be falsely rejected.

In order to guarantee that the overall significance test is still at the level, we have to adapt the common significance level $\alpha'$of the individual tests. This results in the following relation between the overall significance level $\alpha$ and the individual significance levels $\alpha'$ :

$$\left(1-\alpha'\right)^k = 1-\alpha$$

so that the Bonferroni correction $\alpha_B(k)$ is defined by

$$\alpha_B(k) = \alpha' = 1 - (1-\alpha)^{1/k} \approx \frac{\alpha}{k}$$

Thus, given $k$ tests , and we compare the individual test $p$-values are less than or equal to $\alpha_B(k)$, then the experiment-wide $p$-value is less than or equal to $\alpha$. Another justification for this result follows from a probability result called the *Bonferroni inequality*

$$P\left(E_1 \cup E_2 \cup ... \cup E_k\right) \leq \sum_{i=1}^{k} P(E_i)$$

The Bonferroni correction is a conservative correction, in that it is overly stringent in reducing the test-by-test significance level $\alpha$. This is not the only correction that could be used; the package SPLUS has other options, and an extensive list is given below: in section (3.9.2).

### 3.9.2 THE FALSE DISCOVERY RATE

A general framework for multiple testing assessments and corrections can be constructed by considering a $2 \times 2$ classification of test outcomes as follows: consider $k$ simultaneous hypothesis tests, and their outcomes

| | | OBSERVED RESULT | | |
|---|---|---|---|---|
| | | $H_0$ NOT REJECTED | $H_0$ REJECTED | Total |
| | $H_0$ TRUE | $N_{00}$ | $N_{10}$ | $N_{.0}$ |
| TRUTH | $H_0$ NOT TRUE | $N_{01}$ | $N_{11}$ | $N_{.1}$ |
| | Total | $k-R$ | $R$ | $k$ |

Note that the row classification is not observed; in fact, in this table only $k$ and $R$ are observed.

Typically, correction methods concentrate on the $N_{10}$ entry, that is,

$$P\left[N_{10} > 0 | \textbf{All } H_0^{(i)} \text{are TRUE}\right] \leq \alpha$$

for appropriately chosen significance level $\alpha$. This type of control is termed **familywise** control, and the quantity $\alpha$ is termed the Type I, or **Familywise error rate (FWER)**.

An alternative approach is to control the expected level of **false discoveries**, or the **False Discovery Rate (FDR)**. The $FDR$ is defined by

$$FDR = \begin{cases} \frac{N_{10}}{R} & R > 0 \\ 0 & R = 0 \end{cases}$$

A standard procedure, the **BENJAMINI-HOCHBERG (BH)** procedure, adjusts the $k$ $p$-values that result from the tests in a sequential fashion such that the expected false discovery rate is bounded above by $\alpha$, for any suitably chosen $\alpha$ in the range $(0, 1)$. The BH procedure is described below:

1. Compute the $p$-values in the $k$ original hypothesis tests: $p_1, ..., p_k$, and sort them into ascending order so that

$$p_{(1)} < p_{(2)} < ... < p_{(k)}$$

with corresponding (order statistic) random variables

$$P_{(1)} < P_{(2)} < ... < P_{(k)}$$

2. Set $0 < \alpha < 1$.

3. Define $p_{(0)} = 0$ and

$$R_{BH} = \max \left\{ 0 \leq i \leq k : p_{(i)} \leq \frac{i}{k} \alpha \right\}$$

4. Reject $H_0^{(j)}$ for each test $j$ where

$$p_j \leq p_{R_{BH}}$$

This procedure guarantees that the expected FDR is bounded above by $\alpha$, a result that holds for independent tests (where the samples themselves are independent) and also for some samples that are not independent. An adjusted procedure can be used for false negative, or Type II results, or the **False Non Discovery Rate (FNDR)**

$$FNDR = \begin{cases} \dfrac{N_{01}}{k - R} & R > 0 \\ \\ 0 & R = 0 \end{cases}$$

Therefore a variety of corrections can be made. Let $p_i$ be the $p$-value derived from the $i$th hypothesis test, then the following $p$-value thresholds may be used

| | | | |
|---|---|---|---|
| IDENTITY | $\alpha_I(k)$ | $= \alpha$ | |
| BONFERRONI | $\alpha_B(k)$ | $= \alpha/k$ | |
| THRESHOLD | $\alpha_T(k)$ | $= t$ | |
| FIRST $r$ | $\alpha_{T_r}(k)$ | $= p_{(r)}$ | the $r$th largest $p$-value |
| BH | $\alpha_{BH}(k)$ | $= p_{R_{BH}}$ | |

### 3.9.3 STEP-DOWN AND STEP-UP ADJUSTMENT PROCEDURES

The final type of $p$-value adjustments that are utilized are **Step** procedures, usually either **Step-Down** or **Step-Up** procedures, which are methods to control the Familywise Error rate FWER and False Discovery rate FDR respectively. The recipes are reasonably straightforward to implement computationally, and include the Benjamini and Hochberg procedures as special cases.

In addition to the notation introduced in the previous two subsections, for the $k$ hypothesis tests under consideration, let $t_j$ be the $j$th test statistic, and let $t_{(j)}$ denote the $j$th largest absolute value from $t_1, ...t_k$, with corresponding random variables $T_j$ and $T_{(j)}$. The adjusted $p$-values $p_j^*$ can be defined in the following ways

- Bonferroni single step

$$p_j^* = \min \{kp_j, 1\}$$

- Holm step-down

$$p^*_{(j)} = \min_{1\leq i\leq j} \left\{ \min \left\{ (k-i+1)\, p_{(i)}, 1 \right\} \right\}$$

- Bonferroni step-down

$$p^*_{(j)} = \min_{j\leq i\leq k} \left\{ \min \left\{ (k-i+1)\, p_{(i)}, 1 \right\} \right\}$$

- Westfall & Young step-down minP

$$p^*_{(j)} = \max_{1\leq i\leq j} \left\{ P\left[ \max \left\{ P_{(i)}, P_{(i+1)}, ..., P_{(k)} \right\} \leq p_{(i)} | \text{At least one } H_0 \text{NOT TRUE} \right] \right\}$$

- Westfall & Young step-down maxT

$$p^*_{(j)} = \max_{1\leq i\leq j} \left\{ P\left[ \max \left\{ \left|T_{(i)}\right|, \left|T_{(i+1)}\right|, ..., \left|T_{(k)}\right| \right\} \geq \left|t_{(i)}\right| | \text{At least one } H_0 \text{NOT TRUE} \right] \right\}$$

- Benjamini & Hochberg step-up

$$p^*_{(j)} = \min_{j\leq i\leq k} \left\{ \min \left\{ \left( \frac{k}{i} \right) p_{(i)}, 1 \right\} \right\}$$

- Benjamini & Yuketeli conservative step-up

$$p^*_{(j)} = \min_{j\leq i\leq k} \left\{ \min \left\{ k \frac{s(k)}{i}, 1 \right\} \right\} \qquad \text{where } s(k) = \sum_{i=1}^{k} \frac{1}{i}$$

These adjustments to observed $p-$values are all attempts to preserve the integrity of the tests in large multiple testing situations. The final two Westfall and Young procedures can often only be computed in a simulation study.

Note that these methods do not alter the ordering of the test results from "most significant" to "least significant"; it may be sensible, therefore to fix on a number of results to report.

## 3.10 PERMUTATION TESTS AND RESAMPLING METHODS

For most of the hypothesis tests above, we start with the assumptions and work forward to derive the sampling distribution of the test statistic under the null hypothesis. For **permutation tests,** we will reverse the procedure, since the sampling distribution involves the permutations which give the procedure its name and are the key theoretical issue in understanding the test. For **resampling** or **bootstrap** methods , we will resample the original data uniformly and randomly so as to explore the variability of a test statistic.

### 3.10.1 PERMUTATION TESTS

A permutation is a reordering of the numbers $1, ..., n$. For example, (1, 2, 3, 4, 5, 6), (1, 3, 2, 4, 5, 6), (4, 5, 2, 6, 1, 3) (3, 2, 1, 6, 4, 5) are all permutations of the numbers 1 through 6 (note that this includes the standard order in first line). There are $n! = 1 \times 2 \times 3 \times ... \times n$ permutations of $n$ objects.

The central idea of permutation tests refers to rearrangements of the data. The null hypothesis of the test specifies that **the permutations are all equally likely**. The sampling distribution of the test statistic under the null hypothesis is computed by forming all (or many) of the permutations, calculating the test statistic for each and considering these values all equally likely.

Consider the following two group example, where we want to test for any significant difference between the groups.

$$\begin{aligned} \text{Group 1} &: 55, 58, 60 \\ \text{Group 2} &: 12, 22, 34 \end{aligned}$$

Here are the steps we will follow to use a permutation test to analyze the differences between the two groups. For the original order the sum for Group 1 is 173. In this example, if the groups were truly equal (**and the null hypothesis was true**) then randomly moving the observations among the groups would make no difference in the sum for Group 1. Some of the sums would be a little larger than the original sum and some would be a bit smaller. For the six observations there are 720 permutations of which there are 20 distinct combinations for which we can compute the sum of Group 1.

| ORDER | GROUP1 | GROUP2 | SUM | ORDER | GROUP1 | GROUP2 | SUM |
|---|---|---|---|---|---|---|---|
| 1 | 55, 58, 60 | 12, 22, 34 | 173 | 11 | 12, 22, 60 | 55, 58, 34 | 94 |
| 2 | 55, 58, 12 | 60, 22, 34 | 125 | 12 | 12, 58, 22 | 55, 60, 34 | 92 |
| 3 | 55, 58, 22 | 12, 60, 34 | 135 | 13 | 55, 12, 22 | 12, 55, 58 | 89 |
| 4 | 55, 58, 34 | 12, 22, 34 | 148 | 14 | 12, 34, 60 | 55, 58, 34 | 106 |
| 5 | 55, 12, 60 | 58, 22, 34 | 127 | 15 | 12, 58, 34 | 55, 22, 60 | 104 |
| 6 | 55, 22, 60 | 12, 58, 34 | 137 | 16 | 55, 12, 34 | 12, 58, 60 | 101 |
| 7 | 55, 34, 60 | 12, 22, 58 | 149 | 17 | 22, 34, 60 | 55, 58, 34 | 116 |
| 8 | 12, 58, 60 | 55, 22, 34 | 130 | 18 | 22, 58, 34 | 55, 22, 60 | 114 |
| 9 | 22, 58, 60 | 12, 55, 34 | 140 | 19 | 55, 22, 34 | 12, 58, 60 | 111 |
| 10 | 34, 58, 60 | 12, 22, 55 | 152 | 20 | 12, 22, 34 | 55, 58, 60 | 68 |

Of these 20 different orderings only **one** has a Group 1 sum that greater than or equal to the Group 1 sum from our original ordering. Therefore the probability that a sum this large or larger would occur by chance alone is $1/20 = 0.05$ and can be considered to be statistically significant.

### 3.10.2 MONTE CARLO METHODS

In this case the permutation yielded an **exact test** because we were able to enumerate all of the possible combinations. In larger examples it will not be possible to list every permutation so we will have to take a large number of random orderings, sampled uniformly from the permutation distribution. A general **Monte Carlo** strategy for two sample testing is outlined below:

1. For two sample tests for samples of size $n_1$ and $n_2$, compute the value of the test statistic for the observed sample $t^*$
2. Randomly select one of the $(n_1 + n_2)!$ permutations, re-arrange the data according to this permutation, allocate the first $n_1$ to pseudo-sample 1 and the remaining $n_2$ to pseudo-sample 2, and then compute the test statistic $t_1$
3. Repeat 2. $N$ times to obtain a random sample of $t_1, t_2, ..., t_N$ of test statistics from the TRUE null distribution.
4. Compute the $p$-value by reporting $\frac{\text{Number of } t_1, t_2, ..., t_N \text{more extreme than } t^*}{N}$

### 3.10.3 RESAMPLING METHODS AND THE BOOTSTRAP

In statistical analysis, we usually interested in obtaining estimates of a parameter via some statistic, and also an estimate of the variability or uncertainty attached to this point estimate, and a confidence interval for the true value of the parameter.

Traditionally, researchers have relied on normal approximations to obtain standard errors and confidence intervals. These techniques are valid only if the statistic, or some known transformation of it, is asymptotically normally distributed. If the normality assumption does not hold, then the traditional methods should not be used to obtain confidence intervals. A major motivation for the traditional reliance on normal-theory methods has been computational tractability, computational methods remove the reliance on asymptotic theory to estimate the distribution of a statistic.

**Resampling techniques** such as the **bootstrap** and **jackknife** provide estimates of the standard error, confidence intervals, and distributions for any statistic. The fundamental assumption of bootstrapping is that the observed data are representative of the underlying population. By resampling observations from the observed data, the process of sampling observations from the population is mimicked. The key techniques are

- **THE BOOTSTRAP:** In bootstrap resampling, B new samples, each of the same size as the observed data, are drawn with replacement from the observed data. The statistic is first calculated using the observed data and then recalculated using each of the new samples, yielding a bootstrap distribution. The resulting replicates are used to calculate the bootstrap estimates of bias, mean, and standard error for the statistic.
- **THE JACKKNIFE:** In **jackknife** resampling, a statistic is calculated for the $n$ possible samples of size $n-1$, each with one observation left out. The default sample size is $n-1$, but more than one observation may be removed. Jackknife estimates of bias, mean, and standard error are available and are calculated differently than the equivalent bootstrap statistics.

Using the bootstrap and jackknife procedures, all informative summaries (mean, variance, quantiles etc) for the sample-based estimates' sampling distribution can be approximated.

## 3.11 REGRESSION ANALYSIS AND THE LINEAR MODEL

Suppose that we have $n$ measurements of two variables $X$ and $Y$, that is, a sample of pairs of observations $\{(x_i, y_i) : i = 1, ..., n\}$, and it is believed that there is a **linear** relationship between $X$ and $Y$. Suppose that we regard $X$ as a **controlled** variable, that is, we can control the values of $X$ at which $Y$ is measured. Our aim is to try and **predict** $Y$ for a given value of $X$, and thus we have to build a probability model for $Y$ conditional on $X = x$ that incorporates the linear dependence.

### 3.11.1 TERMINOLOGY

$Y$ is the **response** or **dependent** variable
$X$ is the **covariate** or **independent** variable

A simple relationship between $Y$ and $X$ is the **linear regression model**, where

$$E[Y|X = x] = \alpha + \beta x,$$

that is, conditional on $X = x$, the expected or "predicted" value of $Y$ is given by $\alpha + \beta x$, where $\alpha$ and $\beta$ are unknown parameters; in other words, we model the relationship between $Y$ and $X$ as a straight line with **intercept** $\alpha$ and **slope** $\beta$. For data $\{(x_i, y_i) : i = 1, ..., n\}$, the objective is to estimate the unknown parameters $\alpha$ and $\beta$. A simple estimation technique, is **least-squares estimation**.

### 3.11.2 LEAST-SQUARES ESTIMATION

Suppose that a sample, $\{(x_i, y_i) : i = 1, ..., n\}$, is believed to follow a linear regression model, $E[Y|X = x] = \alpha + \beta x$. For fixed values of $\alpha$ and $\beta$, let $y_i^{(P)}$ denote the expected value of $Y$ conditional on $X = x_i$, that is

$$y_i^{(P)} = \alpha + \beta x_i$$

Now define error terms $e_i$, $i = 1, ..., n$ by

$$e_i = y_i - y_i^{(P)} = y_i - \alpha - \beta x_i$$

that is, $e_i$ is the vertical discrepancy between the **observed** and **expected** values of $Y$. The objective in least-squares estimation is find a "line of best fit", and this is achieved by inspecting the squares of the error terms $e_i$, and choosing $\alpha$ and $\beta$ such that the sum of the squared errors is **minimized**; we aim to find the straight line model for which the total error is smallest.

Let $S(\alpha, \beta)$ denote the error in fitting a linear regression model with parameters $\alpha$ and $\beta$. Then

$$S(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i^{(P)})^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

To calculate the least-squares estimates, we have to minimize $S(\alpha, \beta)$ as a function of $\alpha$ and $\beta$. This can be achieved in the usual way by taking partial derivatives with respect to the two parameters, and equating the partial derivatives to zero simultaneously.

$$(1) \frac{\partial}{\partial \alpha} \{S(\alpha, \beta)\} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$

$$(2) \frac{\partial}{\partial \beta} \{S(\alpha, \beta)\} = -2 \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0$$

Solving (1), we obtain an equation for the least-squares estimates $\widehat{\alpha}$ and $\widehat{\beta}$

$$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \widehat{\beta}\overline{x}.$$

Solving (2) in the same way, and combining the last two equations, and solving for $\widehat{\beta}$ gives

$$\widehat{\beta} = n\frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left\{\sum_{i=1}^{n} x_i\right\}^2} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2} \qquad \Longrightarrow \qquad \widehat{\alpha} = \frac{\sum_{i=1}^{n} x_i y_i - \widehat{\beta}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} = \bar{y} - \widehat{\beta}\overline{x}$$

$$S_x = \sum_{i=1}^{n} x_i \qquad S_y = \sum_{i=1}^{n} y_i \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Therefore it is possible to produce estimates of parameters in a linear regression model using least-squares, without any specific reference to probability models. In fact, the least-squares approach is very closely related to maximum likelihood estimation for a specific probability model.

The **correlation coefficient**, $r$, measures the degree of association between $X$ and $Y$ variables and is given by

$$r = \frac{nS_{xy} - S_x S_y}{\sqrt{(nS_{xx} - S_x^2)(nS_{yy} - S_y^2)}}$$

and therefore is quite closely related to $\widehat{\beta}$.

### 3.11.3 LEAST-SQUARES AS MAXIMUM LIKELIHOOD ESTIMATION

Suppose that $X$ and $Y$ follow a linear regression model

$$E[Y|X = x] = \alpha + \beta x,$$

and recall that the error terms $e_i$ were defined

$$e_i = y_i - \alpha - \beta x_i.$$

Now, $e_i$ is the vertical discrepancy between observed and expected behaviour, and thus $e_i$ could be interpreted as the observed version of a **random variable**, say $\epsilon_i$, which represents the random uncertainty involved in measuring $Y$ for a given $X$. A plausible probability model might therefore be that the random variables $\epsilon_i$, $i = 1, ...n$, were independent and identically distributed, and

$$\epsilon_i \sim N(0, \sigma^2),$$

for some error variance parameter $\sigma^2$. Implicit in this assumption is that the distribution of the random error in measuring $Y$ does not depend on the value of $X$ at which the measurement is made. This distributional assumption about the error terms leads to a probability model for the variable $Y$. As we can write

$$Y = \alpha + \beta X + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, then given on $X = x_i$, we have the conditional distribution $Y_i$ as

$$Y_i|X = x_i \sim N(\alpha + \beta x_i, \sigma^2),$$

where random variables $Y_i$ and $Y_j$ are **independent** (as $\epsilon_i$ and $\epsilon_j$ are independent). On the basis of this probability model, we can derive a likelihood function, and hence derive maximum likelihood estimates. For example, we have the likelihood $L(\theta) = L(\alpha, \beta, \sigma^2)$ defined as the product of the $n$ conditional density terms derived as the conditional density of the observed $y_i$ given $x_i$,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i; x_i, \theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right\} \end{aligned}$$

The maximum likelihood estimates of $\alpha$ and $\beta$, and error variance $\sigma^2$, are obtained as the values at which $L(\alpha, \beta, \sigma^2)$ is maximized. But, $L(\alpha, \beta, \sigma^2)$ is maximized when the term in the exponent, that is

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

is minimized. But this is **precisely** the least-squares criterion described above, and thus the m.l.e s of $\alpha$ and $\beta$ assuming a Normal error model are **exactly equivalent** to the least-squares estimates.

### 3.11.4 ESTIMATES OF ERROR VARIANCE AND RESIDUALS

In addition to the estimates of $\alpha$ and $\beta$, we can also obtain the maximum likelihood estimate of $\sigma^2$,

$$\hat{\widehat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2 = S^2$$

Often, a **corrected** estimate, $s^2$, of the error variance is used, defined by

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned}$$

where $\hat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$ is the **fitted value** of $Y$ at $X = x_i$. Note also that, having fitted a model with parameters $\widehat{\alpha}$ and $\widehat{\beta}$, we can calculate the error in fit at each data point, or **residual**, denoted $e_i, i = 1, ..., n$, where $e_i = y_i - \hat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i$.

### 3.11.5 PREDICTION FOR A NEW COVARIATE VALUE

Suppose that, having fitted a model, and obtained estimates $\widehat{\alpha}$ and $\widehat{\beta}$ using maximum likelihood or least-squares, we want to predict the $Y$ value for a new value $x^*$ of covariate $X$. By considering the nature of the regression model, we obtain the predicted value $y^*$ as

$$y^* = \widehat{\alpha} + \widehat{\beta}x^*$$

### 3.11.6 STANDARD ERRORS OF ESTIMATORS AND T-STATISTICS

We need to be able to understand how the estimators corresponding to $\widehat{\alpha}$ and $\widehat{\beta}$ behave, and by how much the estimate is likely to vary. This can be partially achieved by inspection of the **standard errors** of estimates, that is, the square-root of the variance in the sampling distribution of the corresponding estimator. It can be shown that

$$s.e.(\widehat{\alpha}) = s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} \qquad s.e.(\widehat{\beta}) = s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}}$$

where $s$ is the square-root of the corrected estimate of the error variance. It is good statistical practice to report standard errors whenever estimates are reported. The standard error of a parameter also allows a test of the hypothesis "parameter is equal to zero". The test is carried out by calculation of the **t-statistic**, that is, the ratio of a parameter estimate to its standard error. The t-statistic must be compared with the 0.025 and 0.975 percentiles of a Student-$t$ distribution with $n-2$ degrees of freedom as described below.

### 3.11.7 HYPOTHESIS TESTS AND CONFIDENCE INTERVALS

We may carry out hypothesis tests for the parameters in a linear regression model; as usual we need to be able to understand the sampling distributions of the corresponding estimators. In the linear regression model, the sampling distributions of the estimators of $\alpha$ and $\beta$ have **Student-$t$ distributions** with $n-2$ degrees of freedom, hence we use the test statistics

$$t_\alpha = \frac{\widehat{\alpha} - c}{s.e.(\widehat{\alpha})} \qquad t_\beta = \frac{\widehat{\beta} - c}{s.e.(\widehat{\beta})}$$

to test the null hypothesis that the parameter is equal to $c$.

Typically, we use a test at the 5 % significance level, so the appropriate critical values are the 0.025 and 0.975 quantiles of a $St(n-2)$ distribution. It is also useful to report, for each parameter, a confidence interval in which we think the **true** parameter value (that we have estimated by $\widehat{\alpha}$ or $\widehat{\beta}$) lies with high probability. It can be shown that the 95% confidence intervals are given by

$$\alpha : \widehat{\alpha} \pm t_{n-2}(0.975)s.e.(\widehat{\alpha}) \qquad \beta : \widehat{\beta} \pm t_{n-2}(0.975)s.e.(\widehat{\beta})$$

where $t_{n-2}(0.975)$ is the 97.5th percentile of a Student-$t$ distribution with $n-2$ degrees of freedom.

The confidence intervals are useful because they provide an alternative method for carrying out hypothesis tests. For example, if we want to test the hypothesis that $\alpha = c$, say, we simply note whether the 95% confidence interval contains $c$. If it does, the hypothesis can be accepted; if not the hypothesis should be rejected, as the confidence interval provides evidence that $\alpha \neq c$.

We may carry out a hypothesis test to carry out whether there is significant correlation between two variables. We denote by $\rho$ the true correlation; then to test the hypothesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

we use the test statistic

$$t_r = r\sqrt{\frac{n-2}{1-r^2}}$$

which we compare with the null distribution which is Student-$t$ with $n-2$ degrees of freedom. If $|t_r| > t_{n-2}(0.975)$, then we can conclude that the true correlation $\rho$ is significantly different from zero. An alternative test of the hypothesis is given by the **Fisher $z$ statistic**

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

which has a null distribution that is $N(0, 1/(n-3))$. Hence, if $|\sqrt{n-3}z_r| > \Phi^{-1}(0.975) = 1.96$, then we can conclude that the true correlation $\rho$ is significantly different from zero.

### 3.11.8 MULTIPLE LINEAR REGRESSION

In everything that is described above, we have used a model in which we predicted a response $Y$ from a single covariate $X$. This simple model can be extended to the case where $Y$ is modelled as a function of $p$ covariates $X_1, ..., X_p$, that is, we have the conditional expectation of $Y$ given by

$$E[Y|X_1 = x_1, ..., X_p = x_p] = \alpha + \beta_1 x_1 + ... + \beta_p x_p$$

,so that the observation model is given by

$$Y_i|X_1 = x_{i1}, ..., X_p = x_{ip} \sim N(\alpha + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \sigma^2).$$

Again, we can use maximum likelihood estimation to obtain estimates of the parameters in the model, that is, parameter vector $(\alpha, \beta_1, ..., \beta_p, \sigma^2)$, but the details are slightly more complex, as we have to solve $p+1$ equations simultaneously. The procedure is simplified if we write the parameters as a single vector, and perform matrix manipulation and calculus to obtain the estimates.

### 3.11.9 WORKED EXAMPLE

The following data are believed to follow a linear regression model;

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 0.54 | 2.03 | 3.15 | 3.96 | 6.25 | 8.17 | 11.08 | 12.44 | 14.04 | 14.34 | 18.71 | 19.90 |
| $y$ | 11.37 | 11.21 | 11.61 | 8.26 | 14.08 | 16.25 | 11.00 | 14.94 | 16.91 | 15.78 | 21.26 | 20.25 |

We want to calculate estimates of $\alpha$ and $\beta$ from these data. First, we calculate the summary statistics;

$$S_x = \sum_{i=1}^{n} x_i = 118.63 \qquad S_y = \sum_{i=1}^{n} y_i = 172.92 \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 = 1598.6 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i = 1930.9$$

with $n = 12$ which leads to parameter estimates

$$\widehat{\beta} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2} = \frac{12 \times 1930.9 - 118.63 \times 172.92}{12 \times 1598.6 - (118.63)^2} = 0.5201$$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\overline{x} = 14.410 - 0.5201 \times 9.8842 = 9.269$$

This fit leads to the following fitted values and residuals;

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 0.54 | 2.03 | 3.15 | 3.96 | 6.25 | 8.17 | 11.08 | 12.44 | 14.04 | 14.34 | 18.71 | 19.90 |
| $y$ | 11.37 | 11.21 | 11.61 | 8.26 | 14.08 | 16.25 | 11.00 | 14.94 | 16.91 | 15.78 | 21.26 | 20.25 |
| $\hat{y}$ | 9.55 | 10.33 | 11.95 | 12.37 | 12.52 | 13.52 | 15.03 | 15.73 | 16.57 | 16.73 | 19.00 | 19.62 |
| $e$ | 1.82 | 0.88 | −0.34 | −4.11 | 1.56 | 2.73 | −4.03 | −0.80 | 0.34 | −0.95 | 2.26 | 0.63 |

The **corrected variance estimate**, $s^2$, is given by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 3.438 \qquad \Longrightarrow \qquad s = 2.332$$

The **standard errors** for the two parameters are given by

$$s.e.(\widehat{\alpha}) = s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} = 1.304$$

$$s.e.(\widehat{\beta}) = s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}} = 0.113$$

The **t-statistics** for the two parameters are given by

$$t_\alpha = \frac{\widehat{\alpha}}{s.e.(\widehat{\alpha})} = \frac{9.269}{1.304} = 7.109$$

$$t_\beta = \frac{\widehat{\beta}}{s.e.(\widehat{\beta})} = \frac{0.520}{0.113} = 4.604.$$

The 0.975 percentile of a Student-$t$ distribution with $n - 2 = 10$ degrees of freedom is found from tables to be 2.228. Both t-statistics are more extreme than this critical value, and hence it can be concluded that both parameters are significantly different from zero.

To calculate the **confidence intervals** for the two parameters. we need to use the 0.975 percentile of a $St(10)$ distribution. >From above, we have that $St(10)(0.975) = 2.228$, and so the confidence intervals are given by

$$\alpha : \widehat{\alpha} \pm t_{n-2}(0.975)s.e.(\widehat{\alpha}) = 9.269 \pm 2.228 \times 1.304 = (6.364 : 12.174)$$

$$\beta : \widehat{\beta} \pm t_{n-2}(0.975)s.e.(\widehat{\beta}) = 0.5201 \pm 2.228 \times 0.113 = (0.268 : 0.772)$$

so that, informally, we are 95% certain that the true value of $\alpha$ lies in the interval (6.724 : 12.174), and that the true value of $\beta$ lies in the interval (0.268 : 0.772). This amounts to evidence that, for example, $\alpha \neq 0$ (as the confidence interval for $\alpha$ does not contain 0), and evidence that $\beta \neq 1$ (as the confidence interval for $\beta$ does not contain 1).

## 3.12 GENERALIZING THE LINEAR MODEL

In the previous section we concentrated on a simple model where response $Y$ is modelled as a function of $p$ covariates $X_1, ..., X_p$, and the conditional expectation of $Y$ given by

$$E[Y|X_1 = x_1, ..., X_p = x_p] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

so that the observation model is given by

$$Y_i|X_1 = x_{i1}, ..., X_p = x_{ip} \sim N(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \sigma^2). \tag{3.1}$$

In this section, we demonstrate how this model can be represented in matrix form, and demonstrate that many of the simple models and tests studied previously in section (**??**) can be viewed as special cases of the more general class of **Linear Models.**

In addition, we demonstrate how the linear model can be extended to allow for the modelling of data that are not normally distributed; often we collect discrete data for example, or data where the normality assumption (3.1) is not appropriate.

### 3.12.1 REGRESSION AS A LINEAR MODEL

Equation (3.1) can be expressed in vector/matrix form as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

| | | |
|---|---|---|
| $\mathbf{Y}$ | $= [Y_1, Y_2, ..., Y_n]^T$ | an $n \times 1$column vector(the RESPONSE) |
| $\mathbf{X}_i$ | $= [1, x_{i1}, x_{i2}, ..., x_{ip}]$ | a $1 \times (p+1)$row vector |
| $\mathbf{X}$ | $= [1, \mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p]^T$ | an $n \times (p+1)$matrix (the DESIGN MATRIX) |
| $\boldsymbol{\beta}$ | $= \left[\beta_0, \beta_1, ..., \beta_p\right]^T$ | a $(p+1) \times 1$column vector(the PARAMETER VECTOR) |
| $\boldsymbol{\epsilon}$ | $= [\epsilon_1, \epsilon_1, ..., \epsilon_p]^T$ | an $n \times 1$column vector(the RANDOM ERRORS) |

This form of the regression model illustrates the fact that the model is **linear** in $\boldsymbol{\beta}$ (that is, the elements of $\boldsymbol{\beta}$ appear in their untransformed form). This is important as it allows particularly straightforward calculation of parameter estimates and standard errors, and also makes clear that some of the other models that we have already studied, such as ANOVA models, also fall into the linear model class.

It is reasonably straightforward to show that the least-squares/maximum likelihood estimates of $\boldsymbol{\beta}$ for any linear model take the form:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y \qquad \widehat{\boldsymbol{\sigma}}^2 = \frac{1}{n-p}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

where $\mathbf{X}^T$ is the **transpose** of matrix $\mathbf{X}$ : the $(i, j+1)$ element of $\mathbf{X}$ is $x_{ij}$ for $j = 1, 2, ..., p$, which is the $(j+1, i)$ element of $\mathbf{X}^T$. The $p \times p$ **variance-covariance** matrix is

$$\widehat{\boldsymbol{\sigma}}^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

and the diagonal elements of this matrix give the squared standard errors for the estimates and hence quantify uncertainty. A goodness-of-fit measure that records the adequacy of the model in representing that data is the log-likelihood value evaluated at the maximum likelihood estimates

$$-2\log L\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2\right)=n\log\widehat{\boldsymbol{\sigma}}^2+\frac{1}{\widehat{\boldsymbol{\sigma}}^2}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

Note that, here, the entries in the design matrix are merely the raw values for the $p$ predictors and the $n$ data points. However, these entries can be replaced by any functions of the predictor values, such as polynomial or non-linear functions of the $x_{ij}$, for example

$$x_{ij}^2, x_{ij}^3, ... \qquad\qquad g_{ij}(x_{ij})=e^{x_{ij}}$$

The most important feature is that is the model is still linear in $\boldsymbol{\beta}$.

### 3.12.2 THE EXTENDED LINEAR MODEL

The linear model formulation can also be extended in the following way; in vector/matrix form as follows:

$$\mathbf{Y}=\mathbf{g}\left(\mathbf{X}\right)\boldsymbol{\beta}+\boldsymbol{\epsilon} \tag{3.2}$$

where

| | | |
|---|---|---|
| $\mathbf{Y}$ | $=[Y_1,Y_2,...,Y_n]^T$ | an $n\times 1$column vector(the RESPONSE) |
| $\mathbf{g}\left(\mathbf{X}_i\right)$ | $=[1,g_1\left(\mathbf{x}_i\right),g_2\left(\mathbf{x}_i\right),...,g_p\left(\mathbf{x}_i\right)]$ | a $1\times(p+1)$row vector |
| $\mathbf{X}$ | $=[1,\mathbf{g}\left(\mathbf{X}_1\right),\mathbf{g}\left(\mathbf{X}_2\right),...,\mathbf{g}\left(\mathbf{X}_p\right)]^T$ | an $n\times(p+1)$matrix (the DESIGN MATRIX) |
| $\boldsymbol{\beta}$ | $=\left[\beta_0,\beta_1,...,\beta_p\right]^T$ | a $(p+1)\times 1$column vector(the PARAMETER VECTOR) |
| $\boldsymbol{\epsilon}$ | $=[\epsilon_1,\epsilon_1,...,\epsilon_p]^T$ | an $n\times 1$column vector(the RANDOM ERRORS) |

that is, the original predictors $\mathbf{X}$ are incorporated in transformed form. That is, for $i=1,...,n$ and $j=1,...,p$, we have that $g_j\left(\mathbf{x}_i\right)$ is a scalar function of vector $\mathbf{x}_i$; for example

$$g_j\left(\mathbf{x}_i\right) = x_{i1}+x_{i2}$$

$$g_j\left(\mathbf{x}_i\right) = \sqrt{x_{i1}x_{i2}}$$

and so on. This general representation is very flexible, but in statistical terms is very straightforward. The equation in (3.2) ensures that this model is still a linear model in the parameters $\boldsymbol{\beta}$. The $g_j$ functions are often called **basis** functions. This class of model incorporates

- **Fourier representations** (trigonometric basis functions)
- **splines** (polynomial basis functions)
- **wavelets** (localized, orthogonal basis functions)

### 3.12.3 GENERALIZED LINEAR MODELS

A **Generalized Linear Model (GLM)** is an extension of the linear model that allows for non-normal data to be modelled. For example, we could have that $Y_i$ is distributed as a discrete variable, where either Bernoulli, binomial, multinomial, negative binomial or Poisson models may be appropriate, or that $Y_i$ is a continuous variable having a Gamma, or Lognormal, or some other non-normal distribution.

The regression model, where the distribution of $Y$ is deemed to be dependent on the predictors or covariates, is an attractive idea that can be extended to this non-normal situation. In the linear model, the responses are modelled via their **expectation,** conditional on the values of the predictor variables. This idea is retained in the GLM setting, where the model is completed by key a component of the GLM is the **link function** $g$, and in fact we have that if $\mu = E[Y]$

$$g(\mu) = \mathbf{X}\boldsymbol{\beta}$$

where the term $\mathbf{X}\boldsymbol{\beta}$ is the **linear predictor**. The function $g$ maps the expectation $\mu$, which may be a parameter on a bounded range, to the whole real line. It must be a 1-1 function (one value of $\mu$ maps to one and only one real value via $g$) that has a well-defined inverse function $g^{-1}$.

**EXAMPLE** In the case of *Binomial* data, each individual data point $Y_i$ is $Bernoulli(\theta)$ distributed, so that

$$\mu = E[Y] = \theta$$

where $0 \leq \theta \leq 1$. Hence suitable link functions must be mappings from the range $[0, 1]$ to $\mathbb{R}$. Such functions include

- the **logistic** link $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- the **probit** link $g(\mu) = \Phi^{-1}(\mu)$
- the **complementary log-log** link $g(\mu) = -\log(-\log(\mu))$

**EXAMPLE** In the case of *Poisson* data, where each individual data point $Y_i$ is $Poisson(\lambda)$ distributed, so that

$$\mu = E[Y] = \lambda$$

where $\lambda > 0$. Hence suitable link functions must be mappings from the range $(0, \infty)$ to $\mathbb{R}$. One such function is the **log link**

$$g(\mu) = \log(\mu)$$

Inference for GLMs can be carried out using similar techniques to those studied already, such as the maximum likelihood procedure. Usually, the maximum likelihood estimates are obtained by numerical maximization; GLM estimation functions are readily available in most statistics packages such as SPLUS. The results of a GLM fit are of a similar form to those for the ordinary linear model, that is, including

- a set of parameter estimates $\widehat{\boldsymbol{\beta}}$ and standard errors $s.e.\left(\widehat{\beta}\right)$,
- a set of linear predictions $\mathbf{X}\widehat{\boldsymbol{\beta}}$ and fitted values $\widehat{y} = g^{-1}\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$
- a goodness of fit measure $-2\log L\left(\widehat{\boldsymbol{\beta}}\right)$

## 3.13 CLASSIFICATION

**Classification** is a common statistical task in which the objective is to allocate or categorize an object to one of a number of **classes** or categories on the basis of a set of predictor or covariate measurements. Typically, the **predictor** measurements relate to two or more variables, and the **response** is univariate, and often a **nominal** variable, or label. Specifically, we aim to the observed variability in a response variable $Y$ via consideration of predictors $X = (X_1, ..., X_K)$. The principal difference between classification and conventional regression is that the response variable is a nominal categorical variable, that is, for data item $i$

$$Y_i \in \{0, 1, 2, ...K\}$$

so that the value of $Y_i$ is a **label** rather than a numerical value, where the label represents the **group** or **class** to which that item belongs.

We again wish to use the **predictor** information in $X$ to allocate $Y$ to one of the classes There are two main goals:

- to partition the observations into two or more labelled classes. The emphasis is on **deriving a rule** that can be used to **optimally assign** a new object to the labeled classes.
  - This is the process of **CLASSIFICATION**
- to describe either graphically or algebraically, the different features of observations from several known collections. We attempt to find **discriminants** whose numerical values are such that the collections are separated as much as possible.
  - This is the process of **DISCRIMINATION**

Both are special cases of what is termed **MULTIVARIATE ANALYSIS**

Typically, the exercise of classification will be **predictive**, that is,

- we have a set of data available where both the **response** and **predictor** information is known
  - these data are the **training** data
- we also have a set of data where only the **predictor** information is known, and the **response** is to be predicted
  - these data are the **test** data
- often we will carry out an exercise of **model-building** and **model-testing** on a given data set by extracting a **training set**, building a model using the training data, whilst holding back a proportion (the **test set**) for model-testing.

### 3.13.1 CLASSIFICATION FOR TWO CLASSES ($K = 2$)

Let $f_1(x)$ and $f_2(x)$ be the probability functions associated with a (vector) random variable $X$ for two populations 1 and 2. An object with measurements $x$ must be assigned to either class 1 or class 2. Let $\mathbb{X}$ denote the sample space. Let $\mathcal{R}_1$ be that set of $x$ values for which we classify objects into class 1 and $\mathcal{R}_2 \equiv \mathbb{X} \backslash \mathcal{R}_1$ be the remaining $x$ values, for which we classify objects into class 2.

The **conditional probability**, $P(2|1)$, of classifying an object into class 2 when, in fact, it is from class1 is:

$$P(2|1) = \int_{\mathcal{R}_2} f_1(x)\, dx.$$

Similarly, the conditional probability, $P(1|2)$, of classifying an object into class 1 when, in fact, it is from class 2 is:

$$P(1|2) = \int_{\mathcal{R}_1} f_2(x)\, dx$$

Let $p_1$ be the *prior* probability of being in class 1 and $p_2$ be the *prior* probability of 2, where $p_1 + p_2 = 1$. Then,

$$\begin{aligned} P(\text{Object correctly classified as class 1}) &= P(1|1)p_1 \\ P(\text{Object misclassified as class 1}) &= P(1|2)p_2 \\ P(\text{Object correctly classified as class 2}) &= P(2|2)p_2 \\ P(\text{Object misclassified as class 2}) &= P(2|1)p_1 \end{aligned}$$

Now suppose that the *costs* of misclassification of a class 2 object as a class 1 object, and vice versa are, respectively.$c(1|2)$ and $c(2|1)$. Then the expected cost of misclassification is therefore

$$c(2|1)P(2|1)p_1 + c(1|2)\, P(1|2)p_2.$$

The idea is to choose the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ so that this expected cost is minimized. This can be achieved by comparing the predictive probability density functions at each point $x$

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \frac{p_1}{p_2} \geq \frac{c(1|2)}{c(2|1)} \right\} \qquad \mathcal{R}_2 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \frac{p_1}{p_2} < \frac{c(1|2)}{c(2|1)} \right\}$$

or by minimizing the total probability of misclassification

$$p_1 \int_{\mathcal{R}_2} f_1(x)\, dx + p_2 \int_{\mathcal{R}_1} f_2(x)\, dx$$

If $p_1 = p_2$, then

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \right\}$$

and if $c(1|2) = c(2|1)$, equivalently

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \right\}$$

and finally if $p_1 = p_2$ and $c(1|2) = c(2|1)$ then

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq 1 \right\} \equiv \{ x : f_1(x) \geq f_2(x) \}$$

### 3.13.2 CLASSIFICATION FOR TWO NORMAL SAMPLES

Suppose that we have two (multivariate) normal classes (in $d$ dimensions), that is where

- **class 1**: $X \sim N_d(\mu_1, \Sigma_1)$

$$f_1(x) = \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\Sigma_1|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right\}$$

- **class 2**: $X \sim N_d(\mu_2, \Sigma_2)$

$$f_2(x) = \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\Sigma_2|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)\right\}$$

We sometimes assume that $\Sigma_1 = \Sigma_2 = \Sigma$ (*homogeneity of variances).* Using the previous formula, we identify the following **classification rule**; we allocate an observation with predictor variable $x_0$ to class 1 if

$$(\mu_1 - \mu_2)^T \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2) \geq \log\left[\frac{c(1|2)\, p_2}{c(2|1)\, p_1}\right]. \tag{3.3}$$

More generally, if $\Sigma_1 \neq \Sigma_2$, we allocate an observation with predictor variable $x_0$ to class 1 if

$$-\frac{1}{2} x_0^T \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right) x_0 + \left(\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}\right) x_0 - k \geq \log\left[\frac{c(1|2)\, p_2}{c(2|1)\, p_1}\right] \tag{3.4}$$

where

$$k = \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\left(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_2^T \Sigma_2^{-1} \mu_2\right)$$

The parameters $\mu_1, \mu_2$ and $\Sigma$, $\Sigma_1$ and $\Sigma_2$ may be estimated from training data.

- if the covariance matrices are presumed **equal** then we have a total of

$$2d + \frac{1}{2} d(d+1)$$

parameters to estimate

- if the covariance matrices are presumed **unequal** then we have a total of

$$2d + d(d+1)$$

parameters to estimate Thus with limited data in $d$ dimensions, we may be limited in the type of analysis can be done. In fact, we may have to further restrict the type of covariance structure that we may assume; for example, we might have to restrict attention to

- **diagonal** covariance matrices ($2d$ parameters in total),
- or an assumption of **sphericity** ($2(d+1)$ parameters in total)

Despite their simplicity, such models often work well in practice.

### 3.13.3 DISCRIMINATION

**Discriminant analysis** works in a very similar fashion; from equations (3.3) and (3.4) we note that the boundary between regions $\mathcal{R}_1$ and $\mathcal{R}_2$ takes one of two forms

- **Equal covariances**: we have a **straight line/plane** defined by an equation of the form

$$A_1 x + a_0$$

  where $A_1$ is a $d \times d$ matrix

- **Unequal covariances**: we have a **quadratic surface** defined by an equation of the form

$$x^T B_2 x + B_1 x + b_0$$

  where $B_1$ and $B_2$ are $d \times d$ matrices.

### 3.13.4 ASSESSMENT OF CLASSIFICATION ACCURACY

The performance of a classification rule can be achieved in a number of ways: we can examine

- the **within-sample** classification error: the proportion of elements in the training sample that are misclassified by the rule
- the **leave-one-out** classification error: the proportion of elements in the training sample when the model is built (that is, the parameters are estimated) on a training sample that omits a single data point, and then attempts to classify that point on the trained model
- an $m$-**fold cross-validation** : the data are split into $m$ subsamples of equal size, and one is selected at random to act as a **pseudo-test** sample. The remaining data are used as **training** data to build the model, and the prediction accuracy on the pseudo-test sample is computed. This procedure is repeated for all possible splits, and the prediction accuracy computed as a average of the accuracies over all of the splits.
- accuracy using **bootstrap resampling** to achieve the **cross-validation** based estimates of accuracy from above.

The theory behind the assessment of classification accuracy is complex.

### 3.13.5 ROC CURVES

**Receiver Operating Characteristic (ROC)** curves can also be used to compare the classification performance classifiers. We consider the results of a particular classifier for two populations, say one population with a disease, the other population without the disease. Suppose that a single characteristic, $x$, is to be used to classify individuals.

The classification procedures above reduce to a simple rule; we classify an individual to class 1 if

$$x < t_0$$

for some threshold $t_0$, and to class 2 otherwise. We then consider the following quantities:

- **Sensitivity**: probability that a test result will be positive when the disease is present (true positive rate, expressed as a percentage).
- **Specificity**: probability that a test result will be negative when the disease is not present (true negative rate, expressed as a percentage).
- **Positive likelihood ratio**: ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease

$$\frac{TruePositiveRate}{FalsePositiveRate}$$

- **Negative likelihood ratio:** ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease

$$\frac{\text{False Negative Rate}}{\text{True Negative Rate}}$$

- **Positive predictive value**: probability that the disease is present when the test is positive (expressed as a percentage).
- **Negative predictive value:** probability that the disease is not present when the test is negative (expressed as a percentage).

| | | Disease Class | | |
|---|---|---|---|---|
| | | 1 | 2 | Total |
| Predicted Class | 1 | $a$ | $c$ | $a+c$ |
| | 2 | $b$ | $d$ | $b+d$ |
| | Total | $a+b$ | $c+d$ | $a+b+c+d$ |

- Sensitivity:/Specificity:

$$Sensitivity : \frac{a}{a+b} \qquad Specificity : \frac{d}{c+d}$$

- Likelihood Ratios

$$PLR = \frac{Sensitivity}{1 - Specificity} \qquad NLR = \frac{1 - Sensitivity}{Specificity}$$

- Predictive Values

$$PPV = \frac{a}{a+c} \qquad NPV = \frac{d}{b+d}$$

As the classifier producing the predicted class depends on the threshold $t_0$, we can produce a plot of how these quantities change as $t_0$ changes.

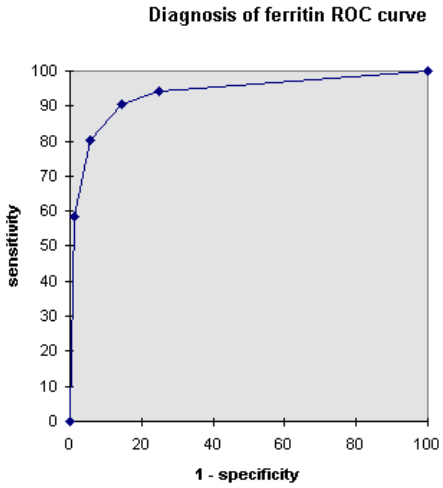
If we plot

$$\begin{aligned} x(t_0) &: 1 - Specificity\ att_0 \\ y(t_0) &: Sensitivity\ att_0 \end{aligned}$$

then we obtain an **ROC curve**;

- for a good classifier would rise steeply and then flatten off ; such a curve would have a large area underneath it on the unit square (the domain of $(x(t_0), y(t_0))$)

- for a poor classifier would be have an ROC curve near the line $y = x$.




**ROC Curve**

### 3.13.6 GENERAL CLASSIFICATION SCHEMES

The general exercise of classification can be seen as a exercise in **regression modelling** for a nominal categorical variable. Previously, we studied **regression**, and more briefly **generalized linear regression**.

- For a **binary response,** or a **two-class** problem, we can use **logistic** or **binary** regression

- For a **multinomial response,** or a **multi-class** problem we can use **multinomial** regression

Because of this regression context, we can use all the previous tools for analysis in regression models that we have used previously.

It is common to view classification of objects in a GLM framework;

### 3.13.7 SUPERVISED AND UNSUPERVISED CLASSIFICATION

An important distinction can be drawn between classification problems in which **training data**, that is response and predictor pairs for known cases, are available, which are referred to as **supervised** learning problems, and problems where no such training data are available, and all inferences about substructure within the data must be extracted from the **test data** alone, possibly only with some background or prior knowledge.

## 3.14 PRINCIPAL COMPONENTS & PARTIAL LEAST SQUARES

**Principal Components Analysis (PCA)** can be used to reveal the underlying variance structure of a data set and a device for data visualization for exploratory data analysis. In its role at reducing rank of a data matrix, PCA can be used to estimate how "true variates" or sources of variability there are in a multivariate data set. **Partial Least Squares (PLS)** gives a similar data reduction decomposition, but also incorporates (the required) dependence between predictor and response present in a regression context.

### 3.14.1 PRINCIPAL COMPONENTS ANALYSIS

Principal Components Analysis (PCA) is a technique used in high dimensional data analysis, for example in the analysis of microarray data. It is used to **reduce the dimensionality** of a data set or data matrix. Broadly, it describes the data set in terms of its components of variance. Each **principal component** describes a **percentage of the total variance** of a data set, and computes **loadings** or **weights** that each variate contributes to this variance. For example, the first principal component of a data set describes the dimension which accounts for the greatest amount of variance of the data set. The **coefficients** of the principal components quantify the loading or weight of each variate to that amount of variance.

The mathematical assumptions behind PCA include multivariate normality of the underlying observations. It is not strictly a regression model, as it only analyzes the **predictor** variables, or, at least, treats all variables equivalently. As a technique, however, it does often contribute in regression or classification analysis because of its data reduction properties.

#### Mathematical Construction

In PCA the data matrix is typically arranged with observations in rows, and different predictors in columns. In a classification context, we might wish to see how much information the predictor variables contained. Suppose that the $N \times p$ data matrix $\mathbf{X}$ is so arranged, but also that $\mathbf{X}$ is **centred**, so that the mean within a column is zero - this is achieved by taking the raw predictor data matrix, and subtracting from each element in a column that columns' sample mean

In linear regression, the matrix $\mathbf{X}$ was referred to as the design matrix, and used to estimate parameters in the regression model using the formula for response vector $\mathbf{y}$.

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{3.5}$$

with prediction

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Note that if $\mathbf{X}$ is the centred matrix as defined, we have that

$$\mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{N}$$

is the **sample covariance matrix**. Now using standard matrix techniques, we may (uniquely) write $\mathbf{X}$ in the following way

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{3.6}$$

where $\mathbf{U}$ is $(N \times p)$, $\mathbf{V}$ is $(p \times p)$ such that

$$\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$$

for $p$-dimensional identity matrix $\mathbf{I}_p$ (that is, $\mathbf{U}$ and $\mathbf{V}$ are **orthogonal** matrices), and $\mathbf{D}$ is a $(p \times p)$ matrix with diagonal elements $d_1 \geq d_2 \geq ... \geq d_p \geq 0$ and zero elements elsewhere. The representation in (3.6) is termed the **singular value decomposition** (**SVD**) of $\mathbf{X}$. Note that, using this form, we have

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T = \mathbf{V}\mathbf{L}\mathbf{V}^T \tag{3.7}$$

This representation is called the **eigen decomposition;**.the diagonal elements of $\mathbf{L} = \mathbf{D}^2$ are

$$l_1 \geq l_2 ... \geq l_p \geq 0;$$

these are termed the **eigenvalues** of $\mathbf{X}^T\mathbf{X}$. The columns of the matrix $\mathbf{V}$ are termed the **eigenvectors** of $\mathbf{X}^T\mathbf{X}$, and the $j^{th}$ column, $v_j$ is the eigenvector associated with eigenvalue $l_j$.

The **principal components** of $\mathbf{X}^T\mathbf{X}$ are defined via the columns of $\mathbf{V}$, $\mathbf{v}_1, ..., \mathbf{v}_p$. The $j^{th}$ principal component is $\mathbf{z}_j$, defined by

$$\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = l_j\mathbf{u}_j$$

for normalized vector $\mathbf{u}_j$. The first principal component $\mathbf{z}_1$ has largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$; we have that

$$Var\left[\mathbf{z}_1\right] = \frac{l_1}{N}$$

Now, recall that in the SVD, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, that is the columns of $\mathbf{V}$ are orthogonal. Hence the principal components $\mathbf{z}_1, ..., \mathbf{z}_p$ are also orthogonal.

The **total variance explained** by the data is a straightforward function of the centered design matrix; it is the sum of the diagonal elements (or **trace)** of the matrix $\mathbf{S}$, given by

$$trace\left(\mathbf{S}\right) = \sum_{j=1}^{p}\left[\mathbf{S}\right]_{jj} = \frac{trace\left(\mathbf{X}^T\mathbf{X}\right)}{N} = \frac{trace\left(\mathbf{L}\right)}{N} = \frac{1}{N}\sum_{j=1}^{p} l_j$$

and hence the $j^{th}$ principal component accounts for a proportion

$$\frac{l_j}{\sum_{k=1}^{p} l_j} \tag{3.8}$$

of the total variance.

Using principal components, therefore, it is possible to find the "directions" of largest variability in terms of a **linear combination** of the columns of the design matrix; a linear combination of column vectors $\mathbf{x}_1, ..., \mathbf{x}_p$ is a vector $\mathbf{w}$ of the form

$$\mathbf{w} = \sum_{j=1}^{p} \pi_j\mathbf{x}_j$$

for coefficients (**loadings**) $\pi = (\pi_1, ..., \pi_p)$ - for the first principal component, $\pi = \mathbf{v}_1$.

**Statistical Properties**

It is of practical use to be able to see how many principal components are needed to explain the variability in the data. To do this we need to study the statistical properties of the elements of the decomposition used. If the predictor variables have a **multivariate normal distribution**, we have the following statistical result. Suppose that predictor vector $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$ have a multivariate normal distribution for $i = 1, ..., N$,

$$\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$$

where the $(p \times p)$ covariance matrix $\Sigma$ has eigen decomposition

$$\Sigma = \Gamma \Lambda \Gamma^T$$

for eigenvalue matrix $\Lambda = diag(\lambda_1, ..., \lambda_p)$ and eigenvector matrix $\Gamma = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p)$. Then the centred sample covariance matrix $\mathbf{S}$

$$\mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{N}$$

with eigen decomposition

$$\mathbf{S} = \mathbf{V}\mathbf{L}\mathbf{V}^T$$

for sample eigenvalue matrix $\mathbf{L} = diag(l_1, ..., l_p)$ and eigenvector matrix $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_p)$ is such that, approximately, as $N \to \infty$,

$$\mathbf{l} = (l_1, ..., l_p)^T \sim N\left(0, 2\Lambda^2\right)$$

that is, the sample eigenvalues are approximately independently normally distributed with variance

$$\frac{2\lambda_j^2}{N-1}$$

**Uses**

The main use of principal components decomposition is in **data reduction** or **feature extraction**. It is a method for looking for the main sources of variability in the predictor variables, and the argument follows that the first few principal components contain the majority of the explanatory power in the predictor variables. Thus, instead of using the original predictor variables in the linear (regression) model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_X + \boldsymbol{\epsilon}$$

we can use instead the principal components as predictors

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta}_Z + \boldsymbol{\epsilon}.$$

where $\mathbf{Z} = \mathbf{X}\mathbf{V}$, where $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ are the parameters vectors in the regression model, both of dimension $(p \times 1)$. The data compression or feature extraction arises if, instead of taking all $p$ of the principal components, we take only the first $k$, that is, we extract the first $k$ columns of matrix $\mathbf{Z}$, and reduce $\boldsymbol{\beta}_Z$ to being a $(k \times 1)$ vector. Choosing $k$ can be done by inspection of the "**scree**" plot of the successive scaled eigenvalues as in (3.8).

### 3.14.2 PARTIAL LEAST SQUARES

The principal components analysis outlined above is an excellent method for extracting the linear combinations of the input predictors that are the largest sources of variability. But, implicit in the PCA definition, is the constraint that no aspect of relationship between the predictors $\mathbf{X}$ and the response $Y$ is recognized. Hence, if the PCA is to be used as a feature extraction method for use in regression, there may well be a deficiency in the principal components as predictors themselves.

**Partial Least Squares (PLS)** is a related feature extraction procedure where the relationship between the predictors $\mathbf{X}$ and the response $Y$ is modelled explicitly. It does this by accounting for the correlation between response and prediction under the usual linear model formulation. An algorithm to construct the PLS components is given by (Hastie *et al.* 2001)[1].

1. Let $\mathbf{x}_j = (x_{j1}, ...x_{jn})^T$ be the $j^{th}$ column of the design matrix $\mathbf{X}$, appropriately centred (by subtracting the column mean $\overline{x}_j$) and scaled (by column sample standard deviation $s_j$) to have sample mean zero and sample variance one.
2. Set $\widehat{\mathbf{y}}^{(0)} = 1\overline{y}$ and $\mathbf{x}_j^{(0)} = \mathbf{x}_j$
3. For $m = 1, 2..., p$,
   - $\mathbf{z}_m = \sum_{j=1}^{p} \widehat{\phi}_{mj} \mathbf{x}_j^{(m-1)}$, where
   $$\widehat{\phi}_{mj} = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \sum_{i=1}^{n} v_{1i} v_{2i}$$
   is the **inner product** operator
   - $\widehat{\theta}_m$ is defined by
   $$\widehat{\theta}_m = \frac{\langle \mathbf{z}_m, \mathbf{y} \rangle}{\langle \mathbf{z}_m, \mathbf{z}_m \rangle}$$
   - $\widehat{\mathbf{y}}^{(m)}$ is defined by
   $$\widehat{\mathbf{y}}^{(m)} = \widehat{\mathbf{y}}^{(m-1)} + \widehat{\theta}_m \mathbf{z}_m$$
   - $\mathbf{x}_j^{(m)}$ is defined by
   $$\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - \left[ \frac{\left\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \right\rangle}{\langle \mathbf{z}_m, \mathbf{z}_m \rangle} \right] \mathbf{z}_m$$
   so that, for each $j = 1, ..., p$, $\mathbf{x}_j^{(m)}$ is **orthogonal** to $\mathbf{z}_m$.
4. Record the sequence of fitted vectors $\widehat{\mathbf{y}}^{(1)}, \widehat{\mathbf{y}}^{(2)}, ..., \widehat{\mathbf{y}}^{(p)}$.

[1] *The Elements of Statistical Learning: Data Mining Inference and Prediction*, p68 Hastie, Tibshirani and Freedman, Springer Series in Statistics, 2001.

5. Evaluate the PLS coefficients as

$$\widehat{\beta}_j^{(PLS)} = \sum_{k=1}^{m} \widehat{\phi}_{kj} \widehat{\theta}_k$$

and the $m^{th}$ PLS direction is

$$\mathbf{z}_m = \sum_{k=1}^{p} \widehat{\phi}_{mk} \mathbf{x}_k$$

In the construction of each PLS direction $\mathbf{z}_m$ the predictors are weighted by the strength of their univariate impact on $\mathbf{y}$. The algorithm first regresses $\mathbf{y}$ first on $\mathbf{z}_1$, giving coefficient $\widehat{\theta}_1$, and then orthogonalizes $\mathbf{x}_1, ..., \mathbf{x}_p$ to $\mathbf{z}_1$, then proceeds to regress $\mathbf{y}$ first on $\mathbf{z}_2$ on these orthogonalized vectors, and so on. After $M \leq p$ steps, the vectors

$$\mathbf{z}_1, ..., \mathbf{z}_M$$

have been produced, and can be used as the inputs in a regression type model, to give the

$$\widehat{\beta}_1^{(PLS)}, ..., \widehat{\beta}_M^{(PLS)}$$

# CHAPTER 4

# STATISTICAL MODELS AND METHODS IN BIOINFORMATICS

## 4.1 STRUCTURAL BIOINFORMATICS: BIOLOGICAL SEQUENCE ANALYSIS

The statistical analysis of biological sequences utilizes the ideas previously introduced in the context of general probability theory, such as conditional probability, entropy, covariance and order statistics, and statistical tools such as maximum likelihood estimation, hypothesis testing and $p$-value calculation. In many ways, biological sequences provide a relatively simple data analysis environment, in that the raw data themselves are usually discrete items. The main difficulty lies in the scale of the analysis as genomic/proteomic data sets are typically vast.

A key distinction that we will make is between **observable** and **unobservable** quantities:

- **observable** variables usually consist of nucleotides for DNA sequences and amino-acid residues for protein sequences, that is, quantities that we can "measure" or observe at all points in the sequence. These quantities are usually observed without error, but occasionally can be subject to uncertainty (such the uncertainty arising from base-calling algorithms). Occasionally, other, larger scale observable quantities might be available, such as DNA motifs, or protein secondary structure.
- **unobservable** variables correspond to hidden or **latent** structure that is not observable, such as CpG island categories, regulatory regions, introns/exons, protein secondary and higher-order structures. These variables are the main focus of our interest.

## 4.2 STOCHASTIC PROCESSES

A **stochastic process**, denoted here by $\{X_t\}$ or $X(t)$, is a sequence of discrete or continuous random variables indexed by a time parameter $t$ that itself may be **discrete** (so that $t = 0, \pm 1, \pm 2, ...$ say) or **continuous** (so that $t \geq 0$ say). For example, a nucleotide sequence can be thought of as a simple stochastic process where time variable $t$ indexes base position, and random variables $X_1, X_2, X_3, ...$ are discrete random variables taking values on the integers $\{1, 2, 3, 4\}$ corresponding to $\{A, C, G, T\}$ say, that correspond to the bases in consecutive positions. We can treat the sequence of variables $\{X_1, X_2, X_3, ...\}$ merely as a collection of random variables and specify and study their joint distribution. Usually, the most simple type of stochastic process is a sequence of *discrete* variables observed in **discrete** time, as then we can merely use the chain rule to write down the joint probability mass function

$$P[X_1 = x_1, X_1 = x_2, X_1 = x_3, ...] = P[X_1 = x_1] \times P[X_2 = x_2 | X_1 = x_1] \times P[X_3 = x_3 | X_1 = x_1, X_2 = x_2] ...$$

Such a model will form the basis of much of the statistical analysis of biological sequences, as it allows us to build up the probability distribution for an observed sequence.

### 4.2.1 THE POISSON PROCESS

Consider a sequence of events occurring in *continuous* time according to some probabilistic rules that are to be defined, and consider random variables $\{X(t), t \geq 0\}$ that count the **number** of events that occur in the intervals $(0, t]$ for $t \geq 0$, so that $X(t)$ takes values $0, 1, 2, ....$

Suppose that

1. $X(0) = 0$

2. For all $0 < s \leq t$, $h > 0$, and non-negative integers $n$ and $m$,

$$P[X(t+h) - X(t) = n | X(s) = m] = P[X(t+h) - X(t) = n]$$

that is, the numbers of events occurring in disjoint intervals are probabilistically **independent**.

3. For $\triangle t > 0$ small,

$$P[X(t + \triangle t) - X(t) = 1] \approx \lambda \triangle t$$

for some $\lambda > 0$, that is, the probability of **exactly one** event occurring in the small interval $(t, t + \triangle t]$ is, for small $\triangle t$, **proportional to the length of the interval**, $\triangle t$. The parameter $\lambda$ is known as the **rate** parameter.

4. For $\triangle t > 0$ small,

$$P[X(t + \triangle t) - X(t) \geq 2] \approx 0$$

for some $\lambda > 0$, that is, the probability of **more than one** event occurring in a small interval $(t, t + \triangle t]$ is essentially zero.

It can be shown that the sequence of discrete variables $\{X(t), t \geq 0\}$ each follow a discrete **Poisson** distribution, that is, if $P_n(t)$ =P[Precisely $n$ events occur in $(0, t]$] it can be shown that

$$P_n(t) = P[X(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \qquad n = 0, 1, 2, ...$$

The sequence $\{X(t), t \geq 0\}$ form a **homogeneous Poisson Process with rate** $\lambda$ (that is, a Poisson process with constant rate $\lambda$)

### 4.2.2 DISTRIBUTIONAL RESULTS FOR TO THE POISSON PROCESS

We have studied the Poisson process in previous sections; the relevant distributional results can be summarized as follows:

1. For $t \geq 0$, $X(t) \sim Poisson(\lambda t)$

2. If $T_1, T_2, T_3, ...$ define the inter-event times of events occurring as a Poisson process, that is, for $n = 1, 2, 3, ...$

$$T_n = \text{“time between } (n-1)^{st} \text{ and } n^{th} \text{ event}''$$

then $T_1, T_2, T_3, ...$ are a sequence of **independent** and **identically distributed** random variables with

$$T_n \sim Exponential(\lambda)$$

3. If $Y_1, Y_2, Y_3, ...$ define the times of events occurring as a Poisson process, that is, for $n = 1, 2, 3, ...Y_n =$"time at which $n$th event occurs", then $Y_1, Y_2, Y_3, ...$ are a sequence of random variables with

$$Y_n \sim Gamma(n, \lambda)$$

4. Consider the interval of length $L$, $(0, L]$, and suppose that $k$ events occur (according to the Poisson process rules) in that interval. If $V_1, V_2, ..., V_k$ define the (ordered) event times of the $k$ events occurring in that interval, then, given $L$ and $k$, $V_1, V_2, ..., V_k$ are the **order statistics** derived from an **independent** and **identically distributed** random sample $U_1, U_2, ..., U_k$ where

$$U_i \sim Uniform(0, L)$$

The homogeneous Poisson process is the standard model for *discrete* events that occur in *continuous* time. It can be thought of as a limiting case of an independent Bernoulli process (a model for *discrete* events occurring in *discrete* time); let

$$X_t \sim Bernouilli\,(\theta) \qquad t = 1, 2, 3,$$

be an i.i.d. sequence: a realization of this sequence might look like

$$00101010001001010000100100001.....$$

For such a process, we have seen (in chapter 2) that

- the number of 1s that occur in any finite and disjoint subsequences are independent random variables
- the number of 1s that occur in any finite subsequence of $n$ time points is a $Binomial\,(n, \theta)$ random variable
- the numbers of trials between successive 1s are i.i.d. $Geometric\,(\theta)$ random variables.

Now consider the a large sequence where $\theta$ is small; the Bernoulli sequence approximates a continuous time sequence, where the events (i.e. the 1s) occur at a constant rate of $\lambda = n\theta$ per $n$ trials, and the Poisson process requirements are met.

The homogeneous Poisson process is a common model that is used often in many scientific fields. For example, in genetics, the Poisson process is used to model the occurrences of crossings-over in meiosis. In sections below we will see how the Poisson process model can be used to represent occurrences of motifs in a biological sequence.

### 4.2.3 MARKOV CHAINS

Markov chains are a special type of *discrete* valued, *discrete* time stochastic process that play a central role in the modelling of biological sequences. Consider a sequence of discrete random variables $\{X_t\}$ indexed in discrete time by $t = 1, 2, 3, ...,$ and each having a range of possible values (or **states**) $\mathbb{X} = \{1, 2, 3, ..., S\}$ say (with $S$ finite).

Suppose that the joint distribution of the $\{X_1, X_2, X_3, ...\}$ is specified (using chain-rule ideas) entirely via the one-step ahead conditional probabilities

$$P\left[X_{t+1} = x_{t+1} | X_1 = x_1, ..., X_t = x_t\right] = P\left[X_{t+1} = x_{t+1} | X_t = x_t\right]$$

(known as the **Markov** or **memoryless property)** so that

$$P\left[X_{t+1} = j | X_t = i\right] = \Pr\left(\text{State } i\text{at time } t \rightarrow \text{State } j\text{at time } t+1\right) = p_{ij}$$

for $i, j \in \{1, 2, 3, ..., S\}$, say, that **does not depend on** $t$. The probabilistic specification can be encapsulated in the $S \times S$ matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1S} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{S1} & p_{S2} & p_{S3} & \cdots & p_{SS} \end{bmatrix}$$

which is called the **transition matrix** where the element in row $i$ and column $j$ defines the probability of moving from state $i$ to state $j$. Note that the row totals must equal 1.

The sequence of random variables described by the matrix $P$ form a **Markov Chain**. Thinking back to the chain rule, in order to complete the specification, a probability specification for the initial state random variable $X_1$ is required; we can denote this discrete probability distribution by row vector of probabilities $\pi^{(1)} = \left(\pi_1^{(1)}, \pi_2^{(1)}, ...\pi_S^{(1)}\right)$. To compute the (marginal) probability of random variable $X_t$ taking the value $i$, we can use matrix algebra and an iterative calculation as follows; let $\pi^{(t)} = \left(\pi_1^{(t)}, \pi_2^{(t)}, ...\pi_S^{(t)}\right)$ denote the probability distribution of $X_t$. First, using the Theorem of Total Probability (chapter 1), conditioning on the different possible values of $X_{t-1}$

$$P\left[X_t = j\right] = \sum_{i=1}^{S} P\left[X_t = j | X_{t-1} = i\right] P\left[X_{t-1} = i\right]$$

which can be re-written

$$\pi_j^{(t)} = \sum_{i=1}^{S} p_{ij} \pi_i^{(t-1)}$$

or in matrix form

$$\pi^{(t)} = \pi^{(t-1)} P$$

Using this definition recursively, we have

$$\pi^{(t)} = \pi^{(t-1)} P = \pi^{(t-2)} P^2 = \pi^{(t-3)} P^3 = ... = \pi^{(2)} P^{t-2} = \pi^{(1)} P^{t-1}$$

which gives a mechanism for computing the marginal probability after $t$ steps.

### 4.2.4 THE RANDOM WALK WITH ABSORBING STATES.

Suppose that the Markov chain is defined on the states $\mathbb{X} = \{0, 1, 2, 3, ..., S\}$

$$P\left[X_{t+1} = j | X_t = i\right] = p_{ij} = \begin{cases} p & j = i + 1 \\ 1 - p & j = i - 1 \end{cases}$$

and zero otherwise, unless $i = 0$ when

$$P\left[X_{t+1} = 0 | X_t = 0\right] = 1 \qquad P\left[X_{t+1} > 0 | X_t = 0\right] = 0$$

so that the chain gets "stuck" in state 0, or unless $i = S$ when

$$P\left[X_{t+1} = S | X_t = S\right] = 1 \qquad P\left[X_{t+1} < S | X_t = S\right] = 0$$

so the chain gets "stuck" in state $S$. Here the states 0 and $S$ are termed **absorbing** states. This type of Markov process is termed a **random walk.**

### 4.2.5 STATIONARY DISTRIBUTIONS OF MARKOV CHAINS

If the Markov chain does not have any absorbing states, or absorbing subsets of states, then it is of interest to try to determine what happens to the marginal probability distributions of the variables $\{X_t\}$ as $t$ becomes large. It can be shown that, in most cases, the Markov chain eventually "settles down", that is, that the marginal distribution stabilizes. That is, for the Markov chain with transition matrix $P$, there exists a **stationary** or **equilibrium** probability distribution $\pi = (\pi_1, \pi_2, ..., \pi_S)$ that satisfies

$$\pi = \pi P$$

so that, $\pi^{(t)} \to \pi$, or $\left(\pi_1^{(t)}, \pi_2^{(t)}, ..., \pi_S^{(t)}\right) \to (\pi_1, \pi_2, ..., \pi_S)$ as $t \to \infty$. This equilibrium distribution can be computed algebraically, or using numerical computation.

**EXAMPLE:** Consider the Markov chain with transition matrix for $S = 4$

$$P = \begin{bmatrix} 0.6 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.5 & 0.1 \\ 0.1 & 0.3 & 0.1 & 0.5 \end{bmatrix}$$

Then, the equilibrium distribution $\pi$ can be obtained by solving the system of equations

$$\begin{aligned} 0.6\pi_1 + 0.1\pi_2 + 0.2\pi_3 + 0.1\pi_4 &= \pi_1 \\ 0.1\pi_1 + 0.7\pi_2 + 0.1\pi_3 + 0.1\pi_4 &= \pi_2 \\ 0.2\pi_1 + 0.2\pi_2 + 0.5\pi_3 + 0.1\pi_4 &= \pi_3 \\ \pi_1 + \pi_2 + \pi_3 + \pi_4 &= 1 \end{aligned}$$

Note that the last row of $P$ is not used in the calculation, but is replaced by the probability distribution constraint that the probabilities must sum to 1. This system can be solved (using SPLUS or MAPLE) to find

$$\pi = (\pi_1, \pi_2, ..., \pi_S) = (0.2414, 0.3851, 0.2069, 0.1667)$$

This stationary distribution can be obtained easily by computing the $n$-step ahead transition matrix $P_n = P^n = P \times P \times ... \times P$ in the limit as $n \to \infty$; some SPLUS code to do this is below

```
p_matrix(c(0.6,0.1,0.2,0.1,0.1,0.7,0.1,0.1,0.2,0.2,0.5,0.1,0.1,0.3,0.1,0.5),nrow=4,byrow=T
p.power_p
for(i in 1:50){
 p.power_p.power %*% p }
print(p.power)
```

which gives the result

```
          [,1]      [,2]      [,3]      [,4]
[1,] 0.2413793 0.3850575 0.2068966 0.1666667
[2,] 0.2413793 0.3850575 0.2068966 0.1666667
[3,] 0.2413793 0.3850575 0.2068966 0.1666667
[4,] 0.2413793 0.3850575 0.2068966 0.1666667
```

### 4.2.6 MARKOV MODELS FOR DNA SEQUENCES

The Markov process models described in the previous section can be used to model DNA sequences. Genomic sequences are comprised of coding and non-coding regions, and small-scale features such as splice sites, and thus some generalization of the models is required. However, if a homogeneous segment can be identified, a Markov model might seem plausible. Here the $S = 4$ states are the nucleotide codes $(A, C, G, T)$ and the transition matrix (to be identified numerically) is

$$P = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

The S-Plus Exercise 1 was concerned with estimating the parameters in this matrix; in fact Maximum Likelihood Estimation gives a formal justification for the technique used there. However, there is no reason to believe that the assumption of homogeneity will hold across large-scale genomic regions.

### 4.2.7 HIDDEN MARKOV MODELS FOR DNA SEQUENCES

Here we note that Markov process models can be used to model the *hidden* or *latent* structure in DNA sequences that determines whether a segment is a coding region, intron, splice site, C+G islands etc. So far we have only modelled the observed nucleotide sequence using the Markov structure. However, we can couple the Markov model for the observed structure with a **Hidden Markov Model (HMM)** for the latent structure.

Suppose that a nucleotide position can be classified as being in a **region** of one of a number, $H$. Then, we might assume that, within a homogeneous region labelled by $h$ the observed nucleotide sequence follows a Markov chain with transition matrix $P_h$ To complete the specification, we assume that there is a sequence in parallel to the nucleotide sequence, comprising region label random variables $H_1, H_2, ...$ which itself follows a Markov chain governed by transition matrix $P_\theta$ where for $i, j \in \{1, 2, 3, ..., H\}$

$$P\left[H_{t+1} = j | H_t = i\right] = \Pr\left(\text{Region type } i \text{at time } t \rightarrow \text{Region type } j \text{at time } t+1\right) = \theta_{ij}$$

We will study the HMM specification in more detail in later sections.

## 4.3 TESTS FOR A SINGLE DNA SEQUENCE

The discovery of structure in a DNA sequence is one of the primary goals of Bioinformatics. This structure can take the form of nucleotide frequency changes, changes in dependence (as in the Markov process construction), the appearance of large scale and small scale structures. We will now study several types of analysis. First, we study some general but important results concerning Order Statistics; these play a role in both the analysis of single sequences, and alignment theory.

### 4.3.1 EXTREME ORDER STATISTICS IN BIOINFORMATICS

Recall that the extreme **order statistics** derived from an independent and identically distributed collection of random variables $X_1, ..., X_n$ are defined by

$$Y_1 = X_{(1)} = \min\{X_1, ..., X_n\} \qquad Y_n = X_{(n)} = \max\{X_1, ..., X_n\}$$

and have marginal cdfs and mass functions/pdfs given by

$$\begin{aligned}
MAXIMUM \qquad f_{Y_n}(y_n) &= n\{F_X(y_n)\}^{n-1} f_X(y_n) \\
F_{Y_n}(y_n) &= \{F_X(y_n)\}^n \\
MINIMUM \qquad f_{Y_1}(y_1) &= n\{1 - F_X(y_1)\}^{n-1} f_X(y_1) \\
F_{Y_1}(y_1) &= 1 - \{1 - F_X(y_1)\}^n
\end{aligned}$$

where $f_X$ and $F_X$ define the distribution of the original variables. For some of the distributions we have studied, the distributions of the extreme order statistics also have simple forms.

**EXAMPLE:** If $X_1, ..., X_n \sim Uniform(0, 1)$, then for $0 \leq x \leq 1$

$$f_X(x) = 1 \qquad F_X(x) = x$$

and hence for the extreme order statistics

$$\begin{aligned}
MAXIMUM \qquad f_{Y_n}(y_n) &= n\{F_X(y_n)\}^{n-1} f_X(y_n) = ny_n^{n-1} \\
F_{Y_n}(y_n) &= y_n^n \\
MINIMUM \qquad f_{Y_1}(y_1) &= n\{1 - y_1\}^{n-1} \\
F_{Y_1}(y_1) &= 1 - \{1 - y_1\}^n
\end{aligned}$$

By extension, if $X_1, ..., X_n \sim Uniform(0, L)$, then for $0 \leq x \leq L$

$$f_X(x) = \frac{1}{L} \qquad F_X(x) = \frac{x}{L}$$

and hence for the extreme order statistics

$$\begin{aligned} MAXIMUM \qquad f_{Y_n}(y_n) &= n\left(\frac{y_n}{L}\right)^{n-1}\frac{1}{L} = \frac{ny_1^{n-1}}{L^n} \\ F_{Y_n}(y_n) &= \left(\frac{y_n}{L}\right)^n \\ MINIMUM \qquad f_{Y_1}(y_1) &= n\left\{1-\frac{y_1}{L}\right\}^{n-1}\frac{1}{L} = \frac{n(L-y_1)^{n-1}}{L^n} \\ F_{Y_1}(y_1) &= 1-\left\{1-\frac{y_1}{L}\right\}^n = 1-\left\{\frac{L-y_1}{L}\right\}^n \end{aligned}$$

**EXAMPLE:** The lifetime (until degradation) of cellular proteins or RNA molecules can be well modelled by an *Exponential* distribution (Ewens and Grant, p43). Suppose $n$ such molecules are to be studied, and their respective lifetimes represented by random variables $X_1, ..., X_n$, regarded as independent and identically distributed. Now, if $X_1, ..., X_n \sim Exponential(\lambda)$, then for $x > 0$

$$f_X(x) = \lambda e^{-\lambda x} \qquad F_X(x) = 1 - e^{-\lambda x}$$

and hence for the extreme order statistics

$$\begin{aligned} MAXIMUM \qquad f_{Y_n}(y_n) &= n\left\{1-e^{-\lambda y_n}\right\}^{n-1}\lambda e^{-\lambda y_n} = n\lambda e^{-\lambda y_n}\left\{1-e^{-\lambda y_n}\right\}^{n-1} \\ F_{Y_n}(y_n) &= \left\{1-e^{-\lambda y_n}\right\}^n \\ MINIMUM \qquad f_{Y_1}(y_1) &= n\left\{1-\left\{1-e^{-\lambda y_1}\right\}\right\}^{n-1}\lambda e^{-\lambda y_1} = n\lambda e^{-n\lambda y_1} \\ F_{Y_1}(y_1) &= 1-\left\{1-\left\{1-e^{-\lambda y_1}\right\}\right\}^n = 1-e^{-n\lambda y_1} \end{aligned}$$

These final results indicate that

$$Y_1 \sim Exponential(n\lambda)$$

and hence, by previous work

$$E_{f_{Y_1}}[Y_1] = \frac{1}{n\lambda} \qquad Var_{f_{Y_1}}[Y_1] = \frac{1}{n^2\lambda^2}$$

It can be shown (with some work) that, to a reasonable approximation

$$E_{f_{Y_n}}[Y_n] \approx \frac{\gamma + \log n}{\lambda} \qquad Var_{f_{Y_n}}[Y_n] \approx \frac{\pi^2}{6\lambda^2}$$

where $\gamma = 0.577216$.

**EXAMPLE:** For discrete random variables $X_1, ..., X_n \sim Geometric(\theta)$, then for $x = 1, 2, 3, ...$

$$f_X(x) = (1-\theta)^{x-1}\theta \qquad F_X(x) = 1 - (1-\theta)^x$$

For convenience we adjust the distribution (as in the SPLUS package) so that for $x = 0, 1, 2, .$

$$f_X(x) = \phi^x(1-\phi) \qquad F_X(x) = 1 - \phi^{x+1}$$

where $\phi = 1 - \theta$. In this adjusted distribution, for the extreme order statistics, the cdfs are given by

$$\begin{aligned} MAXIMUM \qquad F_{Y_n}(y_n) &= \left\{1 - \phi^{y_n+1}\right\}^n \\ MINIMUM \qquad F_{Y_1}(y_1) &= 1 - \left\{1 - \left\{1 - \phi^{y_1+1}\right\}\right\}^n = 1 - \phi^{n(y_1+1)} \end{aligned}$$

and hence, for $Y_n$ we have

$$\begin{aligned} P[Y_n \leq y_n] &= \left\{1 - \phi^{y_n+1}\right\}^n \\ P[Y_n \geq y_n] &= 1 - P[Y_n > y_n - 1] = 1 - F_{Y_n}(y_n - 1) = 1 - \{1 - \phi^{y_n}\}^n \end{aligned}$$

and thus

$$P[Y_n = y_n] = P[Y_n \leq y_n] - P[Y_n \leq y_n - 1] = \left\{1 - \phi^{y_n+1}\right\}^n - \{1 - \phi^{y_n}\}^n$$

Now, $\phi$ is a probability, so we can re-parameterize by writing $\phi = e^{-\lambda}$ for some $\lambda > 0$ and hence

$$\begin{aligned} P[Y_n \leq y_n] &= \left\{1 - e^{-\lambda(y_n+1)}\right\}^n \\ P[Y_n \geq y_n] &= 1 - \left\{1 - e^{-\lambda y_n}\right\}^n \\ P[Y_n = y_n] &= \left\{1 - e^{-\lambda(y_n+1)}\right\}^n - \left\{1 - e^{-\lambda y_n}\right\}^n \end{aligned}$$

which are similar formulae to the *Exponential* case above. It can be shown (again after some work) that

$$E_{f_{Y_n}}[Y_n] \approx \frac{\gamma + \log n}{\lambda} - \frac{1}{2} \qquad Var_{f_{Y_n}}[Y_n] \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12}$$

that is, very similar to the results for the Geometric distribution above.

**SOME APPROXIMATIONS**

For large $n$, some further approximations can be made: If we let

$$\mu_n = E_{f_{Y_n}}[Y_n] \qquad \sigma_n^2 = Var_{f_{Y_n}}[Y_n]$$

then it can be shown that

$$P[Y_n \leq y_n] \approx \exp\left\{-\exp\left\{-\left[\frac{\pi(y_n - \mu_n)}{\sigma_n\sqrt{6}} + \gamma\right]\right\}\right\}$$

so that in the *Exponential* case, where

$$\mu_n \approx \frac{\gamma + \log n}{\lambda} \qquad \sigma_{\max}^2 \approx \frac{\pi^2}{6\lambda^2}$$

then

$$P[Y_n \leq y_n] \approx \exp\{-n\exp\{-\lambda y_n\}\}$$

which can be compared with the exact result above

$$P\left[Y_n \leq y_n\right] = \left\{1 - e^{-\lambda y_n}\right\}^n$$

In the discrete *Geometric* case, it can be using similar approaches that for large $n$

$$\exp\left\{-nC\exp\left\{-\lambda y_n\right\}\right\} \leq P\left[Y_n \leq y_n\right] \leq \exp\left\{-nC\exp\left\{-\lambda(y_n+1)\right\}\right\}$$

and therefore that

$$1-\exp\left\{-nC\exp\left\{-\lambda y_n\right\}\right\} \leq P\left[Y_n \geq y_n\right] \leq 1-\exp\left\{-nC\exp\left\{-\lambda(y_n-1)\right\}\right\}$$

where the constant $C$ is a constant to be defined. These results will provide the probabilistic basis for sequence analysis via BLAST.

**EXAMPLE: Dependent random variables**
Suppose that $n$ points are selected uniformly on the interval $(0,1)$. These points define $n+1$ intervals of random lengths say $U_1, U_2, ..., U_{n+1}$ for which

$$U_1 + U_2 + ... + U_{n+1} = 1$$

The random variables $U_1, U_2, ..., U_{n+1}$ are **not independent** so the theory derived above is not applicable. It can be shown that if

$$U_{\min} = \min\left\{U_1, U_2, ..., U_{n+1}\right\}$$

then

$$P\left[U_{\min} \leq u\right] = 1-(1-(n+1)u)^n \qquad\qquad 0 < u < \frac{1}{n+1}$$

so that

$$F_{U_{\min}}(u) \quad = 1-(1-(n+1)u)^n$$

$$F_{U_{\min}}(u) \quad = n(n+1)(1-(n+1)u)^{n-1}$$

These results are important in the analysis of ***r*-scans**; if we wish to test that the occurrences of a particular nucleotide pattern or "word" occur uniformly in a genomic segment of length $L$, then we could use

### 4.3.2 LONG SEQUENCES OF NUCLEOTIDE REPEATS

Suppose that interest lies in detecting whether a nucleotide is present in a genomic segment in long repeat stretches. The statistical objective is to process a segment of length $N$ say, and to test the hypothesis $H_0$ that the sequence is **random and independent** against the hypothesis $H_1$ that there is evidence of non-randomness or a lack of independence

From the start of the segment, reading "left-to-right" along the nucleotide sequence we can gather data on *run-lengths.* Suppose that the nucleotide of interest, $A$ say, is coded 1 (success) and the other nucleotides are coded 0 (failure). Then, the sequence

$$CGAGAAGATATAAATTCAAATA$$

codes as

$$00101101010111000111101$$

giving run lengths of successes of

$$0, 0, 1, 0, 2, 0, 1, 0, 3, 0, 0, 0, 3, 0, 1$$

Now, if

$$H_0 : \text{Sequence is realization of a random and independent sampling process}$$

then under $H_0$ it can be shown that the distribution of run-lengths should follow an adjusted $Geometric\,(1 - p_A)$ distribution, that is, if $X_i$ is defined the run-length of run $i$, then

$$f_{X_i}(x) = (1 - p_A)p_A^x \qquad\qquad x = 0, 1, 2, ...$$

where $p_A$ is a hypothesized marginal probability of nucleotide $A$. Furthermore, **under** $H_0$, the collection of run-length random variables, $X_1, ..., X_n$ derived from a sequence are independent and identically distributed.

Now suppose that

$$Y_n = \max\{X_1, ..., X_n\}$$

then using the extreme value theory results it can be shown that (under $H_0$)

$$F_{Y_n}(y - 1) = P\left[Y_n < y\right] = \left(1 - p_A^y\right)^n \qquad \Longrightarrow \qquad P\left[Y_n \geq y\right] = 1 - \left(1 - p_A^y\right)^n$$

Hence, for a formal significance test of the hypothesis $H_0$, we may use the observed version of $Y_n$ (that is, the sample maximum) as the test statistic, and compute a $p$-value $p$

$$p = P\left[Y_n \geq y_n\right] = 1 - \left(1 - p_A^{y_n}\right)^n$$

Now, note that $n$ must be specified before this $p$-value can be computed (effectively, we need to choose $n$ large enough) A recommended choice is $n \approx (1 - p_A)\,N$, giving that

$$p \approx 1 - \left(1 - p_A^{y_n}\right)^{(1-p_A)N}$$

which using an exponential approximation gives

$$p \approx 1 - \exp\{-(1 - p_A)Np^{y_n}\}$$

Hence, for a test at the $\alpha = 0.05$ significance level, we must check whether the computed $p$ is less than $\alpha$. If it is, we reject the hypothesis $H_0$ that the sequence is random and independent.

### 4.3.3 R-SCANS

In r-scan analysis, interest lies in detecting short nucleotide "words" in a long genomic segment of length $L$. If $L$ is very large, then the locations of the occurrences of the words, can be regarded as points in the (continuous) interval $(0, L)$, or, without loss of generality, the interval $(0, 1)$. Suppose that the word is detected a total of $k$ times. An appropriate test of the hypothesis $H_0$ that the locations are uniformly distributed in $(0, 1)$ can be based on the test statistic random variable $Y_{k+1}$ where

$$Y_{k+1} = \text{"the maximum inter-location length"}$$

(note that the $k$ points segment the genomic interval into $k + 1$ regions of lengths $U_1, U_2, ..., U_{k+1}$. Then

$$Y_{k+1} = max\left\{U_1, U_2, ..., U_{k+1}\right\}$$

Again, to construct the significance test and calculate a $p$-value, we seek the distribution of this maximum order statistic. However, previous techniques cannot be used, as the random variables $U_1, U_2, ..., U_{k+1}$ are **not independent.** With much work, it can be shown that, as $k$ becomes large, **under** $H_0$,

$$p = P\left[Y_{k+1} \geq y\right] \approx 1 - \exp\left\{-(k+2)e^{-(k+2)y}\right\}$$

which may be re-written

$$p = P\left[Y_{k+1} \geq \frac{\log(k+2) + y}{(k+2)}\right] \approx 1 - \exp\left\{-e^{-y}\right\}$$

This may be generalized to enable tests based on other test statistics to be carried out, such as those based on "$r$-scan" values (maxima of the sums of $r$ adjacent inter-point intervals. Finally, as an alternative test, we could instead use the **minimum** order statistic $Y_1$ Again, after much work, it can be demonstrated that **under** $H_0$ the $p$-value is given (approximately) by

$$p = P\left[Y_1 \leq y\right] \approx 1 - \exp\left\{-y(k+2)^2\right\}$$

## 4.4 ANALYSIS OF MULTIPLE BIOLOGICAL SEQUENCES

Perhaps the most important aspect of the analysis of biological sequences is the search for sequence similarity, as this may give insight into the evolutionary relationships (homologies) between sequences, or into the biological function associated with a particular genomic region or protein sequence. The an initial analysis might be based on **frequency** ("first-order") comparison (of nucleotides/amino acid residues), whereas more sophisticated analyses will concentrate on studying series **dependence** or **association** within the sequences (in a "second-order" comparison).

### 4.4.1 MARGINAL FREQUENCY ANALYSIS

The most rudimentary assessment of sequence similarity that can be carried out is one based on the relative marginal frequencies of the nucleotides or amino acids within each sequence. This can be done within the context of a **Goodness-of-Fit test.** Consider the table of observed frequencies corresponding to two nucleotide sequences

| | Nucleotide | | | | |
|---|---|---|---|---|---|
| | $A$ | $C$ | $G$ | $T$ | Total |
| Sequence 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{14}$ | $n_{1.}$ |
| Sequence 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{24}$ | $n_{2.}$ |
| Total | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | $n_{.4}$ | $n$ |

We wish to test the hypothesis $H_0$ that the two sequences have the same (marginal) nucleotide probabilities, $p_A = p_1, p_C = p_2, p_G = p_3$ and $p_T = p_4$. To carry out the test, we first compute the estimates of these nucleotide probabilities (using **maximum likelihood estimation**) under the assumption that $H_0$ **is true**; it turns out that the **estimates** and **expected** or **fitted values** if $H_0$ **is true** are given by

$$\hat{p}_1 = \frac{n_{.1}}{n} \quad \hat{p}_2 = \frac{n_{.2}}{n} \quad \hat{p}_3 = \frac{n_{.3}}{n} \quad \hat{p}_4 = \frac{n_{.4}}{n} \quad \Longrightarrow \quad \hat{n}_{ij} = n_{i.}\hat{p}_j = \frac{n_{i.}n_{.j}}{n} \quad i = 1, 2,\ j = 1, 2, 3, 4$$

There are two test statistics that may be used; the first is the **Chi-squared statistic or the Likelihood Ratio (LR)** statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \qquad LR = 2\sum_{i=1}^{2}\sum_{j=1}^{4} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}}$$

Both of these statistics have an approximate **Chi-squared distribution** $\chi^2_{(r-1)(c-1)} = \chi^2_3$ distribution, again given that $H_0$ **is true.** Typically, a significance level of $\alpha = 0.05$ is used for this test, and the critical value in a one-tailed test of the hypothesis is at the 0.95 point of this distribution, that is, at 7.81.

**EXAMPLE:** Consider the two sequences detailed in the following table:

| | Nucleotide | | | | |
|---|---|---|---|---|---|
| | $A$ | $C$ | $G$ | $T$ | Total |
| Sequence 1 | 273 | 233 | 258 | 236 | 1000 |
| Sequence 2 | 281 | 246 | 244 | 229 | 1000 |
| Total | 554 | 479 | 502 | 465 | 2000 |

For these sequences, the test statistics are computed to be $\chi^2 = 0.964$ and $LR = 0.964$, and thus there is **no evidence to reject** $H_0$ that the nucleotide probabilities are the same for the two sequences.

### 4.4.2 PROBABILISTIC ASSESSMENT OF ALIGNMENTS

In the search for, for example, sequence homology, a key objective is to assess the degree of similarity between two aligned biological sequences. As an illustration, it is important to be able to assess the similarity between two DNA sequences

$$GGAGACTGTAGACAGCTAATGCTATA$$
$$GAACGCCCTAGCCACGAGCCCTTATC$$

that match exactly in positions 1,3,6,9,10,11,13,14,23,24 and 25, allowing for the fact that evolutionary forces (substitution/mutation, deletion and insertion) will disrupt the exact matching of truly homologous sequences. In fact, is important to allow for partially aligned sequences

$$CGGGTA--TCCAA$$
$$CCC-TAGGTCCCA$$

where the symbol $-$ indicates a wild-card position. Alignments are categorized as

- **global -** when entire sequences are aligned
- **local** - when only subsequences are aligned
- **gapped/ungapped** - when the wild-cards $-$ are allowed or otherwise
- **pairwise/multiple** - when two or more than two sequences are processed.

**EXAMPLE: GLOBAL UNGAPPED PAIRWISE ALIGNMENT**

For a probabilistic assessment of the degree of global alignment of two sequences, we use the theory of maximal runs in binary sequences developed above. For the two ungapped sequences, write a 1 ("success") if there is a match at a given position, and a 0 ("failure") otherwise, so that the entire match binary sequence is

$$00100100111011000000001110$$

with runs of successes of lengths 1, 1, 3, 2 and 3, with zero run lengths otherwise. Applying the previous theory, if we consider hypotheses $H_0$ that the sequences are both randomly generated, the collection of run-length random variables, $X_1, ..., X_n$ derived from the two sequences are independent and identically distributed as $Geometric(p)$ random variables, where

$$p_m = p_A^2 + p_C^2 + p_G^2 + p_T^2$$

Again, considering the random variable $Y_n = \max\{X_1, ..., X_n\}$ as the test statistic, then under $H_0$

$$F_{Y_n}(y) = P[Y_n < y] = (1 - p_m^y)^n$$

so that we have the tail probability calculation

$$P[Y_n \geq y] = 1 - (1 - p_m^y)^n$$

For a formal significance test of the hypothesis $H_0$, therefore, we may again use the maximum run length as the test statistic, and compute a $p$-value, $p$

$$p = P[Y_n \geq y_n] = 1 - (1 - p_m^{y_n})^n$$

For two sequences of length $N$, we have that is $n \approx (1 - p_m)\, N$, giving that

$$\begin{aligned} p &\approx 1 - (1 - p_m^{y_n})^{(1-p_m)N} \\ &\approx 1 - \exp\{-(1 - p_m) N p^{y_n}\} \end{aligned}$$

For a test at the $\alpha = 0.05$ significance level, we must check whether the computed $p$-value is less than $\alpha$. If it is, we reject the hypothesis $H_0$ that the sequence is random and independent. For the sequences above, $y_n = 3$, and if one of the assumptions of $H_0$ is that $p_A = p_C = p_G = p_T = \frac{1}{4}$ so that $p_m = \frac{1}{4}$, then the approximate $p$-value is

$$1 - \left(1 - \left(\frac{1}{4}\right)^3\right)^{26\times(3/4)} = 0.264$$

which is **not** less than $\alpha = 0.05$, and so there is **no evidence to reject** $H_0$, and hence there is no evidence of alignment (or homology). In fact, using this approximation, we would need a run length

$$y_{CRIT} \geq \frac{\log\left(1 - (1 - 0.05)^{(1/26)\times(4/3)}\right)}{\log\left(\frac{1}{4}\right)} = 4.29$$

to reject the null model.

**EXAMPLE: LOCAL UNGAPPED PAIRWISE ALIGNMENT**

For a probabilistic assessment of the degree of local alignment of two sequences, we carry out a similar hypothesis test, using match statistics as test statistics. Instead of considering runs of matches ("successes"), we consider runs of partial matches. as follows. Let random variable $X$ be defined as the length of a subsequence of the original sequence that contains **at most** $k$ **mismatches** ("failures"). The length $X$ of the subsequence is the number of positions up to but not including the $(k+1)$st mis-match; under $H_0$

$$H_0 : \text{Alignment at random without homology}$$

the probability distribution of such a variable is the **Negative Binomial** (generalized geometric) distribution , and the probability distribution (cdf) of $X$ is essentially given by the negative binomial mass function as

$$F_X(x) = P[X \leq x] = \sum_{j=k}^{x} \binom{j}{k} p_m^{j-k} (1 - p_m)^{k+1} \qquad x = k, k+1, k+2, ...$$

For a test statistic, we will again consider a maximum order statistic $Y_n = \max\{X_1, ..., X_n\}$ derived from the analysis of $n$ subsequences.

It is clear from the above formulae that both the definition of the test statistic, the calculation of the null distribution, and $p$-value etc, is complex, and will be more complicated if the model for sequence generation is generalized to be more realistic. However, the principles of statistical hypothesis testing can be applied quite generally.

### 4.4.3 MONTE CARLO SIMULATION

Unfortunately, the usual order statistics computations and approximations cannot be applied in general here, as the analysis would involve the study of **overlapping** sequences, giving variables $X_1, X_2, X_3...$ etc that are **not independent.** In fact, the simplest way to proceed in the calculation of $p$-values is to use **simulation** to generate a look up table of probabilities with which the observed maximum run-length can be compared. Broadly, the simulation proceeds as follows:

1. Generate two sequences of length $N$ under the model implied by $H_0$, that is, two independent sequences with probabilities $p_A, p_C, p_G, p_T$ for each nucleotide.

2. For different values of $k$ ($k = 0, 1, 2, 3, ...$) trawl the sequences to find the longest (contiguous) subsequence which contains at most $k$ mismatches. Record the length $y$ of this longest subsequence.

3. Return to (1) and repeat a large number (1000000 say) of times and form a two-way table of the frequencies with which the longest subsequence containing $k$ mismatches was of length $y$ ($y = k, k + 1, k + 2, ...$)

4. Convert the frequencies into probabilities by dividing by the number of times the simulation is repeated.

This **Monte Carlo simulation** procedure often gives a probability distribution of sufficient accuracy to allow a $p$-value to be computed.

### 4.4.4 ALIGNMENT ALGORITHMS

It is possible to align a pair of sequences without explicit reference to probabilistic rules. The general approach to deterministic alignment is to utilize a **scoring rule** that quantifies the degree of alignment between the sequences. One simple scoring rule scores the alignment by computing the difference between the number of matches and the number of mismatches and wild-cards. More sophisticated algorithms score matches according to the similarity between the characters at a putative match position.

Two common "Dynamic Programming" algorithms are used to perform deterministic alignments; they are

- The Needleman - Wunsch algorithm
- The Smith - Waterman algorithm

Both algorithms can be modified to incorporate gaps in alignments and to penalize them appropriately. The Smith-Waterman algorithm is regarded as a superior algorithm as it computes the optimal alignment between two sequences in a shorter computational time. Neither algorithm is explicitly probabilistic in nature, but can be regarded (loosely) as a form of maximum probability/maximum likelihood procedure.

## 4.5 PROTEINS AND PROTEIN SUBSTITUTION MATRICES

For the alignment of protein sequences, special **scoring rules** that correctly quantify the degree of mismatch between two amino acid residues that may lie in corresponding positions in the two sequences. The scoring rule used **is** probabilistic in nature; it quantifies and utilizes the relative probabilities that two residues $x$ and $y$ say appear in corresponding positions due to actual homology and substitution compared with random chance positioning.

Let $p_x$ and $p_y$ denote, respectively, the probabilities that residues $x$ and $y$ appear in a given position in a protein sequence, and let $p_{xy}$ denote the probability that $x$ and $y$ appear in that position due to homology and substitution. Consider the quantities

$$e_{xy} = \frac{2p_x p_y}{p_{xy}} \qquad (\text{if } x \neq y) \qquad e_{xy} = \frac{p_x p_y}{p_{xy}} \qquad (\text{if } x = y) \qquad s_{xy} = -2\log_2 e_{xy}$$

where $e_{xy}$ is the **relative likelihood** of chance alignment over homology and $s_{xy}$ is a **scoring function** that is **positive** if homology is more likely than chance alignment.

The scores $s_{xy}$ can be used as the basis for protein sequence alignment. Typically, the probabilities $p_x, p_y$ and $p_{xy}$ that appear in the equations above are either estimated using **maximum likelihood** from protein sequence data, or derived from look-up tables such as the **Point Accepted Mutation (PAM)** that is computed using maximum likelihood in a **Markov Chain** model of the evolutionary processes involved, or **Block Substitution Matrices (BLOSUM)**

**EXAMPLE : THE BLOSUM50 MATRIX**

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0 | -1 | -4 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 | -2 | -4 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 | 0 | 3 | -1 | -4 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 | -1 | -2 | -1 | -4 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 | 0 | 0 | -1 | -4 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 | -4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 0 | 1 | -1 | -4 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 | -3 | -1 | -1 | -4 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 | -3 | -3 | -1 | -4 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 | -2 | -1 | -2 | -4 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 | -1 | -1 | 0 | -4 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 | -3 | -2 | -1 | -4 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 | -3 | -2 | -1 | -4 |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -4 |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |

## 4.6 THE STATISTICS OF BLAST AND PSI-BLAST

As a method to provide an assessment of sequence homology, the BLAST/PSI-BLAST approaches and technologies are perhaps the most widely used in biological sequence analysis. Originally designed for DNA sequences but more recently adapted to deal with protein sequences, the BLAST approach provides a simple heuristic means of quantifying the degree of alignment between two or more sequences that may be implemented routinely with reference to an established database of known sequences.

The essence of a BLAST analysis lies in a heuristic statistical hypothesis test; by carefully setting/estimating a few fundamental parameters, the null distribution of an alignment test statistic can be evaluated, and the $p$-value in a test of the null hypothesis of no homology can be evaluated and the appropriate conclusions drawn. We begin this section by describing the key components of the test, before studying the mathematical background in more detail.

### 4.6.1 BLAST: THE BASIC COMPONENTS

If we are to view the BLAST calculation as a hypothesis test, then we need to identify the five key components of such a test appropriate to biological sequence alignment. Recalling Chapter 3 material, we have that a hypothesis test has five components, namely

- the **TEST STATISTIC**
- the **NULL DISTRIBUTION**
- the **SIGNIFICANCE LEVEL**, denoted $\alpha$
- the **P-VALUE**, denoted $p$.
- the **CRITICAL VALUE(S).**

For BLAST, we have

- the test statistic is an **alignment score** $S$
- the null distribution is an **extreme value distribution**, of a similar form to those derived in section 4.3.1
- $\alpha = 0.05$, with corrections for multiple testing
- the $p$-value is computed in terms of the **E-VALUE** which, for two sequences of length $n$ and $m$ is defined as

$$E = Kmn \exp\{-\lambda S\} \qquad p = 1 - e^{-E}$$

  $E$ is the expected number of "high scoring segment" pairs of sequences, and $K$ and $\lambda$ are parameters to be specified or estimated from appropriate sequence databases

- Critical values in the test are evaluated from the null distribution in the usual way

### 4.6.2 STOCHASTIC PROCESS MODELS FOR SEQUENCE ALIGNMENT

To assess whether a given sequence alignment constitutes evidence for homology, it is critically important to be able to assess the statistical significance of the alignment, that is, in short, what degree of alignment can be expected by chance alone For a statistical analysis to be carried out (in a formal **hypothesis testing** framework), as in previous sections, we are required to compute the distribution of some test statistic under the assumption that a null hypothesis $H_0$ is true. In practice, chance alignments could arise from non-homologous sequences, or from related but structurally altered (shuffled) sequences, or from sequences that are generated randomly based upon a DNA or protein sequence model. For the required hypothesis test, statistical results are usually only available analytically (and even then often only approximately) using the last of these definitions, whereas empirical results, based on for example Monte Carlo simulation, may use any of the definitions.

Assessment of the statistical significance of the alignment of two biological sequences is based on properties of a discrete state **stochastic** (**Markov**) **process** similar to the Markov chains introduced in Section 4.1. Consider first two DNA sequences of equal length, and the positions at which they match/align (coded 1) similar to the notation of previous sections:

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G$ | $G$ | $A$ | $G$ | $A$ | $C$ | $T$ | $G$ | $T$ | $A$ | $G$ | $A$ | $C$ | $A$ | $G$ | $C$ | $T$ | $A$ | $A$ | $T$ | $G$ | $C$ | $T$ | $A$ | $T$ | $A$ |
| $G$ | $A$ | $A$ | $C$ | $G$ | $C$ | $C$ | $C$ | $T$ | $A$ | $G$ | $C$ | $C$ | $A$ | $C$ | $G$ | $A$ | $G$ | $C$ | $C$ | $C$ | $T$ | $T$ | $A$ | $T$ | $C$ |
| 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |

Now, suppose that when the Match sequence is read from left to right a Match is given an alignment "score" of +1, whereas a non Match is given a score of -1, and a running or cumulative score is recorded for each position

| Sequence 1 | $G$ | $G$ | $A$ | $G$ | $A$ | $C$ | $T$ | $G$ | $T$ | $\cdots$ | $G$ | $C$ | $T$ | $A$ | $T$ | $A$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 2 | $G$ | $A$ | $A$ | $C$ | $G$ | $C$ | $C$ | $T$ | $A$ | $\cdots$ | $C$ | $T$ | $T$ | $T$ | $T$ | $C$ |
| Match | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | $\cdots$ | 0 | 0 | 1 | 1 | 1 | 0 |
| Score | $+1$ | $-1$ | $+1$ | $-1$ | $-1$ | $+1$ | $-1$ | $-1$ | $+1$ | $\cdots$ | $-1$ | $-1$ | $+1$ | $+1$ | $+1$ | $-1$ |
| Cumulative | 1 | 0 | 1 | 0 | $-1$ | 0 | $-1$ | $-2$ | $-1$ | $\cdots$ | $-5$ | $-6$ | $-5$ | $-4$ | $-3$ | $-4$ |

If $X_i$ is the discrete random variable recording the match score at position $i$ so that

$$X_i = \begin{cases} +1 & \text{Match} \\ -1 & \text{non Match} \end{cases}$$

and $S_i$ is the discrete random variable recording the cumulative match score at position $i$ then by definition

$$S_i = \sum_{j=1}^{i} X_j = S_{i-1} + X_i$$

and hence the sequence of random variables $S_1, S_2, S_3, ...$ form a Markov process that is in fact a **random walk**.on the integers (note that this random walk does not have any absorbing states). For the two sequences above, the complete **observed** sequence $s_1, s_2, s_3, ..., s_{25}, s_{26}$ is given by

$$\begin{array}{l} CUMULATIVE\ SCORE \\ 1, 0, 1, 0, -1, 0, -1, -2, -1, 0, 1, 0, 1, 2, 1, 0, -1, -2, -3, -4, -5, -6, -5, -4, -3, -4 \end{array}$$

Note that such a sequence of random variables can be defined and observed whatever scoring method is used to associate alignment scores with positions in the sequence; this is of vital importance when alignment of protein sequences is considered.

We wish to use the observed sequence $s_1, s_2, s_3, ...$ to quantify the degree of alignment between the sequences; this is achieved as follows. A exact **local** alignment between the sequences is a subsequence where the two sequences match exactly. In the sequences above, the exact local alignments are observed at positions 1,3,6,9-11,13-14 and 23-25. Next consider the **ladder** points, that is, those positions in the sequence at which the cumulative score is lower than any previous point; the ladder points

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *LADDERPOINTS* | 0 | 5 | 8 | 19 | 20 | 21 | 22 |
| SCORE | 0 | $-1$ | $-2$ | $-3$ | $-4$ | $-5$ | $-6$ |

Finally, consider the successive sections of the walk between the ladder points. In particular, consider the **excursions** of the random walk, that is, the successive differences between the **maximum** cumulative score for that subsection and the score at the previous ladder point.

| SUBSECTION | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Begins at Ladder Point | 0 | 5 | 8 | 19 | 20 | 21 | 22 | |
| Ladder Point Score | 0 | $-1$ | $-2$ | $-3$ | $-4$ | $-5$ | $-6$ | (1) |
| Ends at | 4 | 7 | 18 | 19 | 20 | 21 | 26 | |
| Maximum subsection score | 1 | 0 | 2 | $-3$ | $-4$ | $-5$ | $-3$ | (2) |
| Achieved at position(s) | 1, 3 | 6 | 14 | 19 | 20 | 21 | 25 | |
| Excursion | 1 | 1 | 4 | 0 | 0 | 0 | 3 | (2)-(1) |

The alignment, ladder points, maximum subsection scores and excursions can be displayed graphically, as in Figure (4.1)

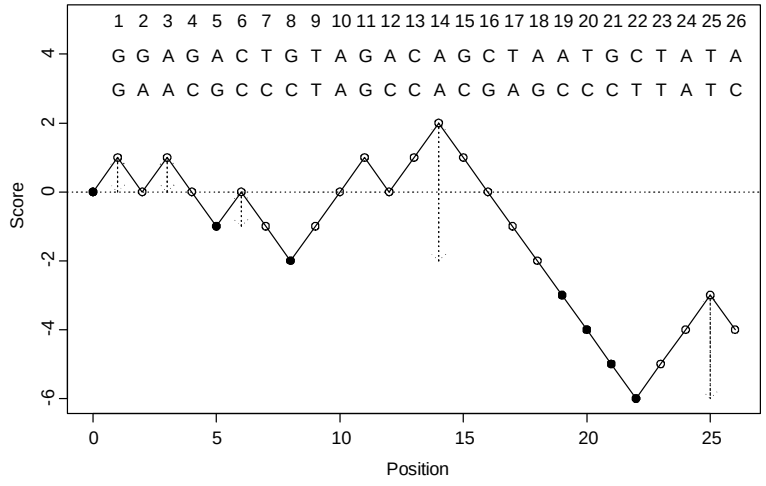



Figure 4.1: Ladder points in BLAST Analysis

The excursions measure the degree of local alignment between the two sequences; reading from left to right along the alignment, within each subsection (between the ladder points) the magnitude of

the excursion measures the cumulative evidence that the sequences are evolutionarily related *within a localized window.* Consider, for example subsection 3 that starts at the ladder point at position 8 (cumulative score -2) and extends to position 18 (before the next ladder point at position 19). There is a close (but not exact) alignment between the first 7 positions, and the degree of support for true alignment peaks at position 14 (cumulative score +4), before declining. Note that, due to the Markovian nature of the underlying process, the sequence subsections between the ladder points have identical probabilistic properties.

In biological sequence analysis, the objective is to optimize and quantify the statistical significance between two arbitrary sequences. Some generalizations to the stochastic process model described above are necessary (to allow for gapped and multiple alignments, analysis of sequences of different lengths etc), but in principle the method is straightforward. Practically, when, for example interrogating a protein database for matches to a potentially novel sequence, it is often sufficient to specify some **threshold** cumulative score value that indicates an acceptable alignment. The choice of such a threshold is clearly quite arbitrary, but some guidance as to what constitutes a sensible threshold value may be obtained by studying the probabilistic properties of the stochastic process models described above.

### 4.6.3 APPLICATIONS OF RANDOM WALK THEORY TO BLAST

The theory derived in Appendix A can be used to calibrate the results obtained from BLAST/PSIBLAST analyses. The objective is to compute the constants $C, \theta$ and $A$ for use in equation (A.7) in the particular case of DNA/protein sequence alignment. Throughout, the null hypothesis $H_0$ at the centre of the alignment assessment will be that the two (or more) sequences are unrelated in evolutionary terms, and also that the nucleotides/amino acids that comprise the sequence

For two sequences of length $N$, let

$$p_j = P\,[\text{Nucleotide/Amino-acid character at any position in the sequence}]$$

so that the probability of observing the (ordered) pair of characters $(j,k)$ (i.e. for sequence 1, then sequence 2) at a given position under $H_0$ is merely $p_j \times p_k$, whereas under the alternative hypothesis of an evolutionary relationship, the probability of observing the ordered pair is $q\,(j,k)$ say.

Now, recall the discussion of **substitution matrices** above, where for the alignment of two characters $x$ and $y$ in a given position we defined a **likelihood ratio** chance alignment over homology, $e_{xy}$ and a **score** $s_{xy}$, given by

$$e_{xy} = \begin{cases} \dfrac{2p_x p_y}{p_{xy}} & x \neq y \\ \dfrac{p_x p_y}{p_{xy}} & x = y \end{cases} \qquad s_{xy} = -2\log_2 e_{xy} \tag{4.1}$$

where, ultimately, the scores comprise the substitution matrix; essentially we replace $x$ by $j$ and $y$ by $k$ in the development below.

Now, we wish to compute some key constants $C, \theta$ and $A$ that appear below in equations (A.5-(A.8) to apply the random walk theory. It transpires that the scores correspond closely to the step sizes, that is, the amount by which the cumulative score can change in the random walk at any position, that appear in the random walk theory above. Let the step size associated with a character pair $(j,k)$ be $S\,(j,k)$. From (A.5), under $H_0$ where the probability of observing a step of size $S\,(j,k)$ is $p_j p_k$ , we have that $\theta$ satisfies

$$\sum_{j,k} p_j p_k e^{\theta S(j,k)} = 1 \tag{4.2}$$

In most BLAST implementations, the constant $\theta$ is denoted $\lambda$; there is no analytic solution to (4.2), but $\theta$ (or, from now on, $\lambda$) can be obtained numerically. Now, from the above considerations, we have for the score function

$$S(j,k) = \frac{1}{\lambda}\log\frac{q(j,k)}{p_j p_k} \qquad \Longleftrightarrow \qquad q(j,k) = p_j p_k e^{\lambda S(j,k)} \tag{4.3}$$

This quantity $q(j,k)$ is defined via substitution matrix, or by point mutation arguments; if $m_{jk}^{(n)}$ is the probability that character $j$ mutates to $k$ in $n$ evolutionary time periods, then it can be shown that

$$S(j,k) = \frac{1}{\lambda}\log\frac{m_{jk}^{(n)}}{p_k} \qquad\qquad q(j,k) = p_j m_{jk}^{(n)}$$

This score function can be considered from an Entropy or Relative Entropy perspective (recall Chapter 2); from (4.2) and (4.3) we have that the $q(j,k)$ form a probability distribution (in fact, the probability distribution of character pairs under the **alternative** hypothesis $H_1$ that the sequences **are** evolutionarily related) and that the score $S(j,k)$ is the **support** for $H_1$ over $H_0$. Hence, under $H_1$, the expected score is

$$H = \sum_{j,k} q(j,k)\log\frac{q(j,k)}{p_j p_k} = \sum_{j,k} q(j,k)\lambda S(j,k) = \lambda\sum_{j,k} q(j,k)S(j,k) = \lambda E_q[S] \tag{4.4}$$

which is merely the **relative entropy** of the two distributions. Here $\mathrm{E}_q[S]$ is the **expected score** under the alternative hypothesis. For high-scoring segments (i.e. where the degree of alignment is high), we therefore have that the expected score is $H/\lambda$.

The **degree of alignment** between two sequences can be quantified using statistical significance testing. If the maximum alignment statistic $Y_{\max}$ is obtained, then for any $y$

$$1 - e^{-Ke^{-\lambda y}} \leq P\left[Y_{\max} > \frac{1}{\lambda}\log N + y\right] \leq 1 - e^{-Ke^{-\lambda(y-1)}} \tag{4.5}$$

where $N$ is the sequence length and $K = Ce^{-\lambda}/A$, where $C$ and $A$ are the constants defined previously. The central probability can be re-written

$$P\left[Y_{\max} > \frac{1}{\lambda}\log N + x\right] = P\left[\lambda Y_{\max} - \log N > y\right]$$

which motivates the definition of the **normalized score**

$$S' = \lambda Y_{\max} - \log(NK) \tag{4.6}$$

that allows (4.5) to be re-written $\exp\left\{-e^{\lambda}e^{-s}\right\} \leq P[S' \leq s] \leq \exp\left\{-e^{-s}\right\}$, and thus the $p$-value associated with any normalized score $s' = \lambda y_{\max} - \log(NK)$ is

$$p \approx 1 - \exp\left\{-e^{-s'}\right\} \tag{4.7}$$

The **bit score** reported by BLAST is defined by

$$\frac{\lambda Y_{\max} - \log K}{\log 2}$$

Another quantity reported by BLAST is the expected number of high-scoring excursions, $E'$, defined by

$$E' = NKe^{-\lambda y_{\max}} \tag{4.8}$$

that is the **expected** number of excursions that (would) **exceed** the observed maximum excursion for aligned sequences of length $N$. It is easily seen that

$$S' = -\log E' \qquad \therefore p \approx 1 - \exp\left\{-E'\right\} \Rightarrow E' = -\log(1-p) \approx p \text{ if } p\text{is small} \tag{4.9}$$

thus we have a further approximation to the $p$-value in terms of $E'$ (that is merely a function of the observed data an modelling assumptions).

### 4.6.4 THE KARLIN-ALTSCHUL SUM STATISTIC

Another useful test statistic is the Karlin-Altschul Sum statistic which considers not only the largest observed excursion of the alignment score random walk, but instead the $r$ largest excursions, for some $r = 1, 2, 3, ...$ to be specified. Denote by $Y_1, Y_2, ..., Y_r$ the $r$ largest excursions where $Y_1 \geq Y_2 \geq ... \geq Y_r$, and consider the corresponding normalized scores

$$S_i = \lambda Y_i - \log(NK) \qquad i = 1, ..., r$$

Then let

$$T_r = S_1 + ... + S_r = \sum_{i=1}^{r} S_i$$

be the sum of these normalized scores; $T_r$ is the Karlin-Altschul Sum statistic. It can be shown that

$$P\left[T_r \geq t\right] \approx \frac{e^{-t} t^{r-1}}{r!(r-1)!} \tag{4.10}$$

where the approximation holds for $t > r(r+1)$. Equation (4.10) gives a means of calculation of a $p$-value to quantify the statistical significance of the alignment.

### 4.6.5 UNALIGNED SEQUENCES AND SEQUENCES OF DIFFERENT LENGTHS

So far we have only considered the alignment of ungapped sequences of the same length. Finally, we consider more general and thus more practical situations where the sequences are arbitrary, that is, not aligned and of different lengths . Our objective will be to apply the theory of previous sections to find an optimal alignment amongst all possible (ungapped,overlapping) local alignments.

For two sequences of length $N_1$ and $N_2$, there are $N_1 + N_2 - 1$ possible local alignments in which there is an overlapping segment between the two sequences. For each of these alignments, a random walk construction can be used to quantify the degree of alignment as above. The total number of (distinct) character comparisons is $N_1 \times N_2$ (all characters from one sequence compared with all characters from the other at some stage), and it is this quantity that is used to generalize the theory of previous sections. In fact, the normalized score in (4.6) becomes

$$S' = \lambda Y_{\max} - \log(N_1 N_2 K)$$

and this newly defined score can be used in (4.7) to get an approximate $p$-value. Now, as some of the $N_1 + N_2 - 1$ alignments are for very short sequences, a correction is made for edge effects; in fact the edge-effect corrected normalized score and expected number of high scoring excursions are

$$S' = \lambda Y_{\max} - \log\left(N_1' N_2' K\right) \qquad E' = N_1' N_2' K e^{-\lambda y_{\max}}$$

where $N_i' = N_i - \lambda Y_{\max}/H$, for $i = 1, 2$, and where $H$ is the entropy quantity that appears in (4.4). For the Karlin-Altschul Sum statistic, the correction is different, with

$$(r-1) + \frac{\lambda}{H}\left(1 - \frac{r+1}{r} f\right) \sum_{i=1}^{r} Y_i$$

being subtracted from $N_1$ and $N_2$ where $f$ is a fixed **overlap adjustment factor** taken to be around $0.1 - 0.2$.

### 4.6.6 CORRECTIONS FOR MULTIPLE TESTING

An issue to be considered here is that of **multiple testing** (section 3.9); if a sequence of different values of $r = 1, 2, 3, ...$ are used then a sequence of hypothesis tests is implied, and because of this a numerical correction to the $p$-values calculated must be made. For $r = 1$, one such correction implies that

$$p \approx 1 - e^{-E} \qquad where \qquad E = 2N_1' N_2' K e^{-\lambda y_{\max}}$$

whereas for $r = 2, 3, 4, ...$ the corrected $p$-value is the value given by (4.7) divided by a factor $(1 - \pi)\pi^{r-1}$.

### 4.6.7 BLAST FOR DATABASE QUERIES

Perhaps the most common use of BLAST is to interrogate databases for matches/homologies for a potentially novel sequence. In principle, this database search should check alignment against all sequences stored in the database, in order to identify high scoring sequences or segments. With such a procedure, however, adjustments to the statistical quantities derived above need to made in advance of a formal statistical test.

Suppose that, using the alignment methods described above a collection of $p$-values are obtained for all the sequences in a database for alignment with some test sequence. Suppose that the highest alignment score obtained is $\upsilon$ for some segment of length $N_2$. Then (using a *Poisson approximation*) the probability that, in the database search there is at least one segment that scores at least $\upsilon$ is

$$1 - e^{-E} \qquad where \qquad E = 2N_1' N_2' K e^{-\lambda \upsilon}$$

Then, if the total length of the database interrogated is $D$ then the **expected number** of segments that score at least $\upsilon$ is approximately

$$E_D = \frac{D}{N_2}\left(1 - e^{-E}\right)$$

as the entire database is a factor of $D/N_2$ times longer than the sequence that gave the highest alignment score. Hence an appropriate approximation to the required $p$-value is

$$p_D = 1 - e^{-E_D}$$

For tests involving the sum statistics of section 4.5.4, a similar correction to the expected value and $p$-value is obtained.

### 4.6.8 A TYPICAL BLAST EXAMPLE

Using BLASTP (with default setting $E = 10$, BLOSUM 62 matrix), on the following protein sequence (NCBI accession number P40582)

```
  1 mslpiikvhw ldhsrafrll wlldhlnley eivpykrdan frappelkki hplgrsplle
 61 vqdretgkkk ilaesgfifq yvlqhfdhsh vlmsedadia dqinyylfyv egslqpplmi
121 efilskvkds gmpfpisyla rkvadkisqa yssgevknqf dfvegeiskn ngylvdgkls
181 gadilmsfpl qmaferkfaa pedypaiskw lktitseesy aaskekaral gsnf
```

we obtain the following output:

- a clickable, graphical display of alignments
- a list of the aligned sequences and the $E$-scores
- detailed information on each aligned sequence (including the target sequence)
- a character by character description of matches, wild card matches, and insertions
- a final summary of the analysis

```
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 24,178,712
Number of Sequences: 1578346
Number of extensions: 994651
Number of successful extensions: 2068
Number of sequences better than 10.0: 21
Number of HSP's better than 10.0 without gapping: 6
Number of HSP's successfully gapped in prelim test: 15
Number of HSP's that attempted gapping in prelim test: 2040
Number of HSP's gapped (non-prelim): 22
length of query: 234
length of database: 517,180,703
effective HSP length: 121
effective length of query: 113
effective length of database: 326,200,837
effective search space: 36860694581
effective search space used: 36860694581
T: 11
A: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.8 bits)
S2: 71 (32.0 bits)
```

## 4.7 HIDDEN MARKOV MODELS

A brief discussion of the role of Hidden Markov Models (HMMs) was given in section (4.2.7); here we study the statistical aspects (estimation, hypothesis testing etc.) in relation to biological sequences in more detail. Recall that previously, the character (nucleotide/amino acid) in a given position can be classified was be classified as being part of a **region** of one of a number, $\mathbb{H}$, of types, and that within a homogeneous region labelled by $k$ the observed nucleotide sequence follows a Markov chain with transition matrix $P_k$ , and finally that there is a **latent** sequence in parallel to the **observed** sequence, comprising region label random variables $H_1, H_2, ...$ which itself follows a Markov chain governed by transition matrix $P_\theta$ where

$$P\left[H_{t+1} = j | H_t = i\right] = \Pr\left(\text{Region type } i \text{at time } t \rightarrow \text{Region type } j \text{at time } t+1\right) = \theta_{ij} \qquad i, j \in \mathbb{H} \tag{4.11}$$

which may be represented as

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Observed sequence | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
| Latent sequence | $H_1$ | $H_2$ | $H_3$ | $H_4$ | $H_5$ | $H_6$ | $H_7$ | $H_8$ | $H_9$ | $H_{10}$ |

For illustration consider a nucleotide sequence, with $X_i \in \{A, C, G, T\}$, near the boundary of a coding/non-coding region. Let $\mathbb{H} = \{0, 1\}$ be the set of available region types, where 0 classifies a non-coding region and 1 classifies a coding region. If the coding region begins at nucleotide 7, then a possible configuration would be

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Observed sequence | $A$ | $C$ | $T$ | $C$ | $G$ | $A$ | $A$ | $C$ | $C$ | $G$ |
| Latent sequence | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

so that the realized latent sequence is

$$h = (h_1, h_2, ..., h_{10}) = (0, 0, 0, 0, 0, 0, 1, 1, 1, 1)$$

However, of course, in practice the latent sequence **is not observed** and therefore the statistical analysis issues centre on inference about (estimation of) the latent sequence $\{H_t, t \geq 1\}$ and the parameters in the Markov transition matrices $P_h, h \in \mathbb{H}$, and $P_\theta$.

### 4.7.1 LIKELIHOOD INFERENCE FOR HIDDEN MARKOV MODELS

The statistical analysis of biological sequences via HMMs is based on probabilistic ideas used to construct **likelihood** functions. First we study the analysis in full generality before restricting attention to specific cases. For an observed sequence of length $n$, denote by $x = (x_1, x_2, ..., x_n)$ and $h = (h_1, h_2, ..., h_{10})$ the observed and realized latent sequences respectively. Suppose, for that the observed sequence, $x_t \in \mathbb{X} = \{1, 2, ..., n_X\}$, and suppose that for the latent sequence $h_t \in \mathbb{H} = \{0, 1, 2, ..., n_H\}$ is the set of available region types.

For $k \in \mathbb{H}$, let

$$P_k = \begin{bmatrix} p_{11}^{(k)} & p_{12}^{(k)} & \cdots & p_{1n_X}^{(k)} \\ p_{21}^{(k)} & p_{22}^{(k)} & \cdots & p_{2n_X}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n_X1}^{(k)} & p_{n_X2}^{(k)} & \cdots & p_{n_Xn_X}^{(k)} \end{bmatrix}$$

be the $n_X \times n_X$ Markov transition matrix for region type $k$, and let

$$P_\theta = \begin{bmatrix} \theta_{00} & \theta_{00} & \cdots & \theta_{0n_H} \\ \theta_{10} & \theta_{11} & \cdots & \theta_{1n_H} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{n_H 0} & \theta_{n_H 1} & \cdots & \theta_{n_H n_H} \end{bmatrix}$$

be the $(n_H + 1) \times (n_H + 1)$ Markov transition matrix between regions. Recall that, in each case, the rows of these matrices are conditional probability distributions and therefore must sum to one. Typically, we will be considering $n_X = 4$ (for DNA sequences) and $n_X = 20$ for protein sequences, and $n_H$ up to 5.

Given the latent sequence $h$ and using the notation $\mathcal{P} = (P_0, P_1, ..., P_{n_H}, P_\theta)$, the likelihood function derived from the observed data $x$,can be defined in the spirit of earlier sections, and using the chain-rule for probabilities as

$$\begin{aligned} L(h_1, h_2, ..., h_n, \mathcal{P}) &= f(x_1, x_2, ..., x_n; h_1, h_2, ..., h_n, \mathcal{P}) \\ &= f(x_1; h_1, \mathcal{P}) \\ &\quad \times f(x_2; x_1, h_1, h_2, \mathcal{P}) \\ &\quad \times f(x_3; x_1, x_2, h_1, h_2, \mathcal{P}) \\ &\quad \times \dots \\ &\quad \times f(x_n; x_1, x_2, ..., x_{n-1}, h_1, h_2, ..., h_n, \mathcal{P}) \end{aligned} \tag{4.12}$$

Now, because of the Markov assumption for the observed data, the conditional probability expressions can be simplified. Specifically, for $t = 2, 3, ..., n$

$$f(x_t; x_1, x_2, ..., x_{t-1}, h_1, h_2, ..., h_t, \mathcal{P}) = f(x_t; x_{t-1}, h_{t-1}, h_t, \mathcal{P})$$

as the observation in position $t$ conditional on previous values is dependent only on the observation in position $t-1$. Furthermore, if $h_t = h_{t-1} = k$ say (that is, there is no change in region type between position $t-1$ and position $t$), then

$$f(x_t; x_{t-1}, h_{t-1}, h_t, \mathcal{P}) = f(x_t; x_{t-1}, h_{t-1}, h_t, P_k)$$

where

$$f(x_t; x_{t-1}, h_{t-1}, h_t, P_k) = p_{ij}^{(k)} \qquad if x_{t-1} = i \; and x_t = j$$

is the probability of a transition between states $i$ and $j$ within region type $k$ between positions $t-1$ and $t$. If $h_t \neq h_{t-1}$, say $h_{t-1} = k_1$ and $h_t = k_2$ with $k_1 \neq k_2$ then it is assumed that

$$f(x_t; x_{t-1}, h_{t-1}, h_t, \mathcal{P}) = f(x_t; h_t, P_{k_2})$$

where

$$f(x_t; h_t, P_{k_2}) = p_j^{(k_2)} \qquad if x_t = j$$

and $p_j^{(k_2)}$ is the **stationary or equilibrium probability** of state $j$ in region type $k_2$ (see section 4.2.3). Note that, for any $k$, the equilibrium distribution

$$p^{(k)} = \left(p_1^{(k)}, p_2^{(k)}, ..., p_{n_X}^{(k)}\right)$$

is defined entirely by the transition matrix $P_k$.

**Example 4.7.1** In the examples above, we have

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Observed sequence | $A$ | $C$ | $T$ | $C$ | $G$ | $A$ | $A$ | $C$ | $C$ | $G$ |
| Coded | 1 | 2 | 4 | 2 | 3 | 1 | 1 | 2 | 2 | 3 |
| Latent sequence | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

and thus $n_X = 4, n_H = 1$. The likelihood is thus

$$\begin{aligned} L\left(h_1, h_2, ..., h_n, \mathcal{P}\right) &= f(x_1, x_2, ..., x_{10}; h_1, h_2, ..., h_{10}, \mathcal{P}) \\ &= f(x_1; h_1, \mathcal{P}) \times f(x_2; x_1, h_1, h_2, \mathcal{P}) \times f(x_3; x_2, h_2, h_3, \mathcal{P}) \times \ldots \times f(x_{10}; x_9, h_9, h_{10}, \mathcal{P}) \\ &= f(1; 0, \mathcal{P}) \times f(2; 1, 0, 0, \mathcal{P}) \times f(4; 2, 0, 0, \mathcal{P}) \times \ldots \times f(3; 2, h_9, h_{10}, \mathcal{P}) \\ &= \underbrace{\underbrace{p_1^{(0)}}_{Pos1} \times \underbrace{p_{12}^{(0)}}_{Pos2} \times \underbrace{p_{24}^{(0)}}_{Pos3} \times \underbrace{p_{42}^{(0)}}_{Pos4} \times \underbrace{p_{23}^{(0)}}_{Pos5} \times \underbrace{p_{31}^{(0)}}_{Pos6}}_{Regiontype0} \times \underbrace{\underbrace{p_1^{(1)}}_{Pos7} \times \underbrace{p_{12}^{(1)}}_{Pos8} \times \underbrace{p_{22}^{(1)}}_{Pos9} \times \underbrace{p_{23}^{(1)}}_{Pos10}}_{Regiontype1} \end{aligned}$$

Previously, such a likelihood has formed the basis for statistical inference. Using **maximum likelihood estimation** we could estimate the unobserved parameters $(h_1, h_2, ..., h_n, \mathcal{P})$ by choosing those values at which $L\left(h_1, h_2, ..., h_n, \mathcal{P}\right)$ is maximized, that is, we choose

$$\left(\hat{h}_1, \hat{h}_2, ..., \hat{h}_n, \hat{\mathcal{P}}\right) = \arg\max L\left(h_1, h_2, ..., h_n, \mathcal{P}\right)$$

where, recall,

$$L\left(h_1, h_2, ..., h_n, \mathcal{P}\right) = f\left(x|h, \mathcal{P}\right)$$

The inference problem now is twofold. We wish to

(a) report the **most probable states** $h = (h_1, h_2, ..., h_n)$ in light of the data $x$

(b) estimate the parameters in $\mathcal{P} = (P_0, P_1, ..., P_{n_H}, P_\theta)$

The estimation of the most probable states is complicated by the structure in this latent sequence. Remember that the Markov assumption means that the joint distribution of random variables $H = (H_1, H_2, ..., H_n)$ should be written (using the chain rule) as

$$\begin{aligned} f(h_1, h_2, h_3, ..., h_n) &= f\left(h_1\right) \times f\left(h_2|h_1\right) \times f\left(h_3|h_1, h_2\right) \times ... \times f\left(h_n|h_1, h_2, h_3, ..., h_{n-1}\right) \\ &= f\left(h_1\right) \times f\left(h_2|h_1\right) \times f\left(h_3|h_2\right) \times ... \times f\left(h_n|h_{n-1}\right) \end{aligned} \tag{4.13}$$

and this dependence structure should be taken into account. Recall also that the joint distribution of vector $H$ depends on transition matrix $P_\theta$, terms in (4.13) will be either **transition probabilities** $\theta_{ij}$ or **equilibrium probabilities** $\theta_i$ derived from $P_\theta$.

For HMMs, likelihood based inference is carried out via **Bayes Rule** that allows the **posterior probability** of the states in the latent sequence to be computed. The key quantity is the joint conditional probability of the hidden states, given the observed sequence, that is $p\,(h|x)$ where

$$f\,(h|x) = \frac{f(x|h)f(h)}{f(x)} \tag{4.14}$$

suppressing the dependence on the other parameters, where the first term in the numerator comes from (4.12) and the second term comes from (4.13) The denominator is the joint (unconditional) probability of observing the data sequence $x$ that can be computed via the **Total Probability** result as

$$f(x) = \sum_h f(x|h)f(h) \tag{4.15}$$

where the summation is over all possible state vector configurations. Inference will require efficient computational methods as the summation in (4.15) and the maximizations that are required both involve large numbers of terms

### 4.7.2 COMPUTATIONAL METHODS FOR HMMS

The computational aspect of likelihood based inference can be broken down into three sub-problems.

(i) Compute the **conditional likelihood** in (4.15) given transition probabilities $\mathcal{P}$

$$f(x|\mathcal{P}) = \sum_h f(x|h,\mathcal{P})f(h|\mathcal{P})$$

(ii) Find the state vector that maximizes the joint conditional probability in (4.14), that is,

$$\hat{h} = \arg\max f\,(h|x)$$

(iii) Find the maximum likelihood estimates of the parameters in $\mathcal{P}$.

The **doubly-Markov** model described above, that is, with a Markov structure in the observed data and a Markov structure in the unobserved states is a model that requires much computational effort. Typically, a simplification is made, in that the matrices $P_0, P_1, ..., P_{n_H}$ are assumed to be diagonal, that is, we may write for $k = 0, 1, ..., n_H$,

$$P_k = \begin{bmatrix} p_1^{(k)} & 0 & \cdots & 0 \\ 0 & p_2^{(k)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{n_X}^{(k)} \end{bmatrix}$$

so that the observed data are **conditionally independent** given the unobserved states, so that there is no dependence between characters in adjacent positions in the sequence. In this case the likelihood is formed (in the example) as

$$\begin{aligned} L\left(h_1, h_2, ..., h_n, \mathcal{P}\right) &= f(x_1, x_2, ..., x_{10}; h_1, h_2, ..., h_{10}, \mathcal{P}) \\ &= f(x_1; h_1, \mathcal{P}) \times f(x_2; h_2, \mathcal{P}) \times f(x_3; h_3, \mathcal{P}) \times \ldots \times f(x_{10}; h_{10}, \mathcal{P}) \\ &= \underbrace{\underbrace{p_1^{(0)}}_{Pos1} \times \underbrace{p_2^{(0)}}_{Pos2} \times \underbrace{p_4^{(0)}}_{Pos3} \times \underbrace{p_2^{(0)}}_{Pos4} \times \underbrace{p_3^{(0)}}_{Pos5} \times \underbrace{p_1^{(0)}}_{Pos6}}_{Regiontype0} \times \underbrace{\underbrace{p_1^{(1)}}_{Pos7} \times \underbrace{p_2^{(1)}}_{Pos8} \times \underbrace{p_2^{(1)}}_{Pos9} \times \underbrace{p_2^{(1)}}_{Pos10}}_{Regiontype1} \end{aligned}$$

Even if this assumption is made, the computational task is still considerable. For example, for task (i), it is possible merely to list all possible state vector configurations that appear as the summand in (4.15), and to sum out over them. However, this is a calculation requiring a large number of computations; for a sequence of length $n$, the direct calculation requires

$$2n \times (n_H + 1)^n$$

and for a sequence of length 10 with $n_H = 1$ as in example 1, this number is $20 \times 2^{10} = 20480$, but this number increases quickly as the sequence length/number of region type increases. For example:

| | | $n_H$ | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| $n$ | 10 | $2.048 \times 10^4$ | $1.181 \times 10^6$ | $2.097 \times 10^7$ | $1.953 \times 10^8$ |
| | 100 | $2.535 \times 10^{32}$ | $2.103 \times 10^{50}$ | $3.321 \times 10^{62}$ | $4.158 \times 10^{72}$ |
| | 500 | $3.273 \times 10^{153}$ | $3.636 \times 10^{241}$ | $1.072 \times 10^{304}$ | $3.055 \times 10^{352}$ |

so even for moderate-sized problems the number of computations is large. Thus, instead of direct calculation, the **Forward** and **Backward** algorithms are used. See section B for full details

For the next stage of the inferential process, it is required to compute the most likely sequence of unobserved states, given the observed data, that is

$$\hat{h} = \arg\max f\left(h|x\right) = \arg\max \frac{f(x|h)f(h)}{f(x)} = \arg\max f(x|h)f(h) = \arg\max f(x, h)$$

This is achieved via the **Viterbi Algorithm** The final stage involves **parameter estimation**, and the **Baum-Welch Algorithm.;** see section B for full details.

## 4.8 FUNCTIONAL BIOINFORMATICS: GENE EXPRESSION ANALYSIS VIA MICROARRAYS

A common biological problem is to detect differential expression of a gene in two or more tissue or cell types, as any differences may contribute to the understanding of the cellular organization (pathways, regulatory networks), or may provide a mechanism for discrimination between future unlabelled samples. Microarray experiments have made the study of gene expression routine; instantaneous measurements of mRNA levels for large numbers of different genes can be obtained for different tissue or cell types in a matter of hours. The most important aspects of a statistical analysis of gene expression data are, therefore twofold; the analysis should be readily implementable for large data sets (large numbers of genes, and/or large numbers of samples), and should give representative, robust and reliable results over a wide range of experiments.

### 4.8.1 MICROARRAY DATA: THE TWO TYPES OF ARRAY

- **cDNA microarrays**: In cDNA microarray competitive hybridization experiments, the mRNA levels of a genes in a target sample are compared to the mRNA level of a control sample by attaching fluorescent tags (usually red and green respectively for the two samples) and measuring the relative fluorescence in the two channels. Thus, in a test sample (containing equal amounts of target and control material), differential expression **relative** to the control is either in terms of *up-regulation* or *down-regulation* of the genes in the target sample. Any genes that are up-regulated in the target compared to the control and hence that have larger amounts of the relevant mRNA, will fluoresce as predominantly red, and any that are down-regulated will fluoresce green. Absence of differences in regulation will give equal amounts of red and green, giving a yellow fluor. Relative expression is measured on the log scale

$$y = \log \frac{x_{TARGET}}{x_{CONTROL}} = \log \frac{x_R}{x_G} \tag{4.16}$$

  where $x_R$ and $x_G$ are the fluorescence levels in the RED and GREEN channels respectively.

- **Oligonucleotide arrays:** The basic concept oligonucleotide arrays is that the array is produced to interrogate specific target mRNAs or genes.by means of a number of oligo probes usually of length no longer than 25 bases; typically 10-15 probes are used to hybridize to a specific mRNA, with each oligo probe designed to target a specific segment of the mRNA sequence. Hybridization occurs between oligos and test DNA in the usual way. The novel aspect of the oligonucleotide array is the means by which the **absolute** level of the target mRNA is determined; each *perfect match* (PM) probe is paired with a *mismatch* (MM) probe that is identical to the prefect match probe **except** for the nucleotide in the centre of the probe, for which a mismatch nucleotide is substituted, as indicated in the diagram below.

$$\begin{aligned} PM &: ATGTATACTATT\boxed{A}TGCCTAGAGTAC \\ MM &: ATGTATACTATT\boxed{C}TGCCTAGAGTAC \end{aligned}$$

  The logic is that the target mRNA, which has been fluorescently tagged, will bind perfectly to the PM oligo, and not bind at all to the MM oligo, and hence the absolute amount of the target mRNA present can be obtained as the **difference**

$$x_{PM} - x_{MM}$$

  where $x_{PM}$ and $x_{MM}$ are the fluorescence measurements of for the PM and MM oligos respectively.

In a microarray experiment, therefore, we typically have access to expression/expression profile data, possibly for a number of replicate experiments, for each of a (usually large) number of genes or expressed sequence tags (ESTs) or probes. We will denote by

$$y_{ijk} \qquad i = 1, ..., N,\ j = 1, ....n_i, k = 1, ..., T$$

the expression data for each of $N$ genes, with $n_i$ replicate observations of a time series of length $T$ for gene $i$. The hybridization experiments are carried out under strict protocols, and every effort is made to regularize the production procedures, from the preparation stage through to imaging. Typically, replicate experiments are carried out; the same array gene/oligo set are used to investigate the portions of the same test sample.

### 4.8.2 STATISTICAL ANALYSIS OF MICROARRAY DATA

Conventional statistical analysis techniques and principles (hypothesis testing, significance testing, estimation, simulation methods/Monte Carlo procedures) can be used in the analysis of microarray data. The principal biological objectives of a typical microarray analysis are:

1. **Detection of differential expression**: up- or down-regulation of genes in particular experimental contexts, or in particular tissue samples, or cell lines at a given time instant.

2. **Understanding of temporal aspects of gene regulation**: the representation and modelling of patterns of changes in gene regulation over time.

3. **Discovery of gene clusters:** the partitioning of large sets of genes into smaller sets that have common patterns of regulation.

4. **Inference for gene networks/biological pathways:** the analysis of co-regulation of genes, and inference about the biological processes involving many genes concurrently.

There are typically several key issues and models that arise in the analysis of microarray data; we have previously studied these techniques in a general statistical context.

- **array normalization:** arrays are often imaged under slightly different experimental conditions, and therefore the data are often very different even from replicate to replicate. This is a systematic experimental effect, and therefore needs to be adjusted for in the analysis of differential expression. A misdiagnosis of differential expression may be made purely due to this systematic experimental effect.

- **measurement error**: the reported (relative) gene expression levels models are only in fact proxies for the true level gene expression in the sample. This requires a further level of variability to be incorporated into the model.

- **modelling**: the sources of variability present in the data can be explained using conventional statistical tools of linear and non-linear models. In addition, it may be necessary also to used *mixed* regression models, where gene specific *random-effects* terms are incorporated into the model. For example, a common linear mixed model for non time-course data is as follows: for gene $i$ under condition $j$, in replicate (array) $l$, we have that

$$y_{ij}^{(l)} = \alpha^{(l)} + \gamma_{ij} Z_{ij} + \varepsilon_{ij}^{(l)}$$

where $\alpha^{(l)}$ is an array effect, $\gamma_{ij}^{(l)}$ is a gene specific (random) effect for gene $i$ under condition $j$, $Z_{ij}$ is an indicator variable determining whether the $i^{th}$ gene is in fact differentially expressed under the $j^{th}$ condition, and $\varepsilon_{ij}^{(l)}$ is an uncorrelated random effect.

- **multivariate analysis**: the covariability of response measurements, in time course experiments, or between $PM$ and $MM$ measurements for an oligonucleotide array experiment, is best handled using multivariate modelling.

- **testing:** one- and two-sample hypothesis testing techniques, based on parametric and non-parametric testing procedures can be used in the assessment of the presence of differential expression. For detecting more complex (patterns of) differential expression, in more general structured models, the tools of analysis of variance (ANOVA) can be used to identify the chief sources of variability.

- **multiple testing/False discovery**: in microarray analysis, a classical statistical analysis using significance testing needs to take into account the fact that a very largs number of tests are carried out. Hence significance levels of tests must be chosen to maintain a required *family-wise error rate*, and to control the *false discovery rate*.

- **classification**: the genetic information contained in a gene expression profile derived from microarray experiments for, say, an individual tissue or tumour type may be sufficient to enable the construction of a *classification rule* that will enable subsequent classification of new tissue or tumour samples.

- **cluster analysis:** the discovery of subsets of larger sets of genes that have common patterns of regulation.can be achieved using the statistical techniques of *cluster analysis* (see section 4.9).

- **computer-intensive inference**: for many testing and estimation procedures needed for microarray data analysis, simulation-based methods (bootstrap estimation, Monte Carlo and permutation tests, Monte Carlo and Markov chain Monte Carlo) are often necessary to enable the appropriate calibration of the inferences being made. This is especially true when complex and *hierarchical* or *multi-level* models are used to represent the different sources of variability in the data.

- **data compression/feature extraction:** the methods of principal components analysis and extended linear modelling via *basis functions* can be used to extract the most pertinent features of the large microarray data sets.

- **experimental design**: statistical experimental design can assist in determining the number of replicates, the number of samples, the choice of time points at which the array data are collected and many other aspects of microarray experiments. In addition, power and sample size assessments can inform the experimenter as to the statistical worth of the microarray experiments that have been carried out.

Typically, data derived from both types of microarray highly noise and artefact corrupted. The statistical analysis of such data is therefore quite a challenging process. In many cases, the replicate experiments are very variable. The other main difficulty that arises in the statistical analysis of microarray data is the dimensionality; a vast number of gene expression measurements are available, usually only on a relatively small number of individual observations or samples, and thus it is hard to establish any general distributional models for the expression of a single gene.

# 4.9 CLUSTER ANALYSIS OF MICROARRAY DATA

## 4.9.1 CLUSTER ANALYSIS

**Cluster analysis** is the searching for **groups** (**clusters**) in the data, in such a way that objects belonging to the same cluster resemble each other, whereas objects in different clusters are dissimilar. In two or three dimensions, clusters can be visualized. With more than three dimensions, or in the case of dissimilarity data (see below), we need some kind of analytical assistance. Generally speaking, clustering algorithms fall into two categories:

1. **Partitioning Algorithms:** A partitioning algorithm describes a method that divides the data set into $k$ clusters, where the integer $k$ needs to specified. Typically, you run the algorithm for a range of $k$-values. For each $k$, the algorithm carries out the clustering and also yields a quality index which allows you to select the ìbestî value of $k$ afterwards.

2. **Hierarchical Algorithms:**. A hierarchical algorithm yields an entire hierarchy of clusterings for the given data set. *Agglomerative methods* start with the situation where each object in the data set forms its own cluster, and then successively merges clusters until only one large cluster (the entire data set) remains. *Divisive methods* start by considering the whole data set as one cluster, and then splits up clusters until each object is separated.

Data sets for clustering of $N$ observations can have either of the following structures:

- an $N \times p$ **data** matrix, where rows contain the different observations, and columns contain the different variables.

- an $N \times N$ **dissimilarity** matrix, whose $(i,j)^{th}$ element is $d_{ij}$, the **distance** or **dissimilarity** between observations $i$ and $j$ that has the properties

  - $d_{ii} = 0$
  - $d_{ij} \geq 0$
  - $d_{ji} = d_{ij}$

- Typical data distance measures between two data points $i$ and $j$ with measurement vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ include

  - the *Euclidean* distance for continuous measurements

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)}$$

  - the *Manhattan* distance for continuous or discrete measurements

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}| = \sum_{k=1}^{p} \sqrt{(x_{ik} - x_{jk})^2}$$

For **ordinal** (categorical) or **nominal (**label) data, other dissimilarities can be defined.

### 4.9.2 PARTITIONING METHODS

Partitioning methods are based on specifying an initial number of groups, and iteratively reallocating observations between groups until some equilibrium is attained. Several different algorithms are available

1. The **k-Means** algorithm: In the $k$-means algorithm the observations are classified as belonging to one of $k$ groups. Group membership is determined by calculating the **centroid** for each group (the multidimensional version of the mean) and assigning each observation to the group with the closest centroid. The $k$-means algorithm alternates between calculating the centroids based on the current group memberships, and reassigning observations to groups based on the new centroids. Centroids are calculated using least-squares, and observations are assigned to the closest centroid based on least-squares. This assignment is performed in an iterative fashion, either from a starting allocation or configuration, or from a set of starting centroids.

2. **Partitioning around medoids (PAM):** The PAM method uses **medoids** rather than **centroids** (that is, medians rather than means in each dimension. This approach increases robustness relative to the least squares approach given above.

### 4.9.3 HIERARCHICAL CLUSTERING

Hierarchical Clustering procedures can be carried out in two ways

- **Heuristic Criteria** The basic hierarchical agglomeration algorithm starts with each object in a group of its own. At each iteration it merges two groups to form a new group; the merger chosen is the one that leads to the smallest increase in the sum of **within-group sums of squares**. The number of iterations is equal to the number of objects minus one, and at the end all the objects are together in a single group. This is known as *Ward's method*, the *sum of squares method*, or the *trace method*. The hierarchical agglomeration algorithm can be used with criteria other than the sum of squares criterion, such as the *average*, *single or complete linkage* methods described below.

- **Model-Based Criteria** Model-based clustering is based on the assumption that the data are generated by a mixture of underlying probability distributions. Specifically, it is assumed that the population of interest consists of $k$different subpopulations, and that the density of an observation from the th subpopulation is for some unknown vector of parameters.

Hence, hierarchical clustering is a method of organizing a set of objects into sets of using a similarity/discrepancy measure or by some overall potential function. Agglomerative clustering initially places each of the $N$ items in its own cluster. At the first level, two objects are to be clustered together, and the pair is selected such that the potential function increases by the largest amount, leaving $N-1$ clusters, one with two members, the remaining $N-2$ each with one. At the next level, the optimal configuration of $N-2$ clusters is found, by joining two of the existing clusters. This process continuous until a single cluster remains containing all $N$ items.

In conventional hierarchical clustering, the method of agglomeration or combining clusters is determined by the distance between the clusters themselves, and there are several available choices. For merging two clusters $C_i$ and $C_j$, with $N_1$ and $N_2$ elements respectively, the following criteria can be used

- In *average* (or *average linkage)* clustering, the two clusters that have the smallest *average distance between the points in one cluster and the points in the other*

$$d(C_i, C_j) = \frac{1}{N_1 N_2} \sum_{k \in C_i, l \in C_2} d_{kl}$$

are merged

- In *connected (single linkage, nearest-neighbour)* clustering, the two clusters that have the smallest *distance between a point in the first cluster and a point in the second cluster*

$$d(C_i, C_j) = \min_{k \in C_i, l \in C_2} d_{kl}$$

are merged.

- In *compact (complete linkage, furthest-neighbour)* clustering, the two clusters that have the largest *distance between a point in the first cluster and a point in the second cluster*

$$d(C_i, C_j) = \max_{k \in C_i, l \in C_2} d_{kl}$$

are merged.

### 4.9.4 MODEL-BASED HIERARCHICAL CLUSTERING

Another approach to hierarchical clustering is **model-based clustering**, which is based on the assumption that the data are generated by a mixture of $K$ underlying probability distributions. Given data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$, let

$$\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_N)$$

denote the cluster labels, where

$$\gamma_i = k$$

if the $i^{th}$ data point comes from the $k^{th}$ subpopulation. In the classification procedure, the **maximum likelihood** procedure (section 3.3.1) is used to choose the parameters in the model.

Commonly, the assumption is made that the data in the different subpopulations follow multivariate normal distributions, with mean $\boldsymbol{\mu}_k$ and covariance matrix $\Sigma_k$ for cluster $k$ If

$$\Sigma_k = \sigma^2 I_p \qquad I_p = diag(1, ..., 1), \ ap \times p \ matrix.$$

then maximizing the likelihood is the same as minimizing the sum of within-group sums of squares that underlies Ward's method. Thus, Ward's method corresponds to the situation where clusters are **hyperspherical** with the **same variance**. If clusters are not of this kind, (for example, if they are thin and elongated), Ward's method tends to break them up into hyperspherical blobs.

Other forms of $\Sigma_k$ yield clustering methods that are appropriate in different situations. The key to specifying this is the **eigen** decomposition of $\Sigma_k$, given by **eigenvalues** $\lambda_1, ..., \lambda_p$ and **eigenvectors** $\mathbf{v}_1, ..., \mathbf{v}_p$, as in **Principal Components Analysis** (section 3.14.1, equation (3.7)) The **eigenvectors** of $\Sigma_k$, specify the orientation of the $k^{th}$ cluster, the largest **eigenvalue** $\lambda_1$ specifies

its variance or size, and the ratios of the other eigenvalues to the largest one specify its shape. Further, if

$$\Sigma_k = \sigma_k^2 I_p$$

the criterion corresponds to hyperspherical clusters of different sizes; this is known as the *Spherical* criterion.

Another criterion results from constraining only the **shape** to be the same across clusters.. This is achieved by fixing the eigenvalue ratios

$$\alpha_j = \frac{\lambda_j}{\lambda_1} \qquad j = 2, 3, ..., p$$

across clusters; common choices for the specification are

- $\alpha_j = 0.2$ giving **ellipsoidal** clusters
- $\alpha_j = 0.01$ gives almost **linear**, **concentrated** clusters
- $\boldsymbol{\alpha_j = 1}$ giving the **spherical** model

### 4.9.5 MODEL-BASED ANALYSIS OF GENE EXPRESSION PROFILES

The clustering problem for vector-valued observations can be formulated using models used to represent the gene expression patterns via the extended linear model (section 3.12.2), that is, linear models and non-linear basis functions. Generically we wish to capture the behaviour of the gene expression ratio $y$ as a function of time $t$ and measurement error. The basis of our modelling strategy would be to use models that capture the characteristic behaviour of expression profiles that we can expect to observe due to different forms of regulation. A regression framework and model is often adopted

$$y_t = f(\beta, t) + \varepsilon_t.$$

More specifically, we model $y$ using a linear model

$$y_t = X_t\beta + \varepsilon_t$$

where $X_t$ is (in general) a $1 \times p$ vector of specified functions of $t$, and $\beta$ is a $p \times 1$ parameter vector. In vector representation, where $y = (y_1, ..., y_T)$, we have

$$y = X\beta + \varepsilon \tag{4.17}$$

and a classical linear model. The precise form of design matrix $X$ is at the moment left unspecified. Typically we take the random error terms $\{\varepsilon_t\}$ as independent and identically distributed Normal variables with variance $\sigma^2$, implying that the conditional distribution of the responses $Y$ is multivariate normal

$$Y|X, \beta, \sigma^2 \sim N\left(X\beta, \sigma^2 I_T\right) \tag{4.18}$$

where now $X$ is $T \times p$ where $I_T$ is the $T \times T$ identity matrix. For this model, the maximum likelihood/ordinary least squares estimates of $\beta$ and $\sigma^2$ are

$$\widehat{\beta}_{ML} = \left(X^T X\right)^{-1} X^T y \qquad \widehat{\sigma}^2 = \frac{1}{T-p}\left(y - \widehat{y}\right)^T \left(y - \widehat{y}\right)$$

for fitted values $\widehat{y} = X\widehat{\beta}_{ML} = X\left(X^T X\right)^{-1} X^T y$. as seen in

### 4.9.6 BAYESIAN ANALYSIS IN MODEL-BASED CLUSTERING

In a **Bayesian analysis** of the model in (4.17) a joint **prior** distribution $p\left(\beta,\sigma^2\right)$ is specified for $\left(\beta,\sigma^2\right)$, and a **posterior** distribution conditional on the observed data is computed for the parameters. The calculation proceeds using Bayes Rule (section 1.6), and is given by

$$p\left(\beta,\sigma^2|y,x\right)=\frac{L\left(y;x,\beta,\sigma^2\right)p\left(\beta,\sigma^2\right)}{\int L\left(y;x,\beta,\sigma^2\right)p\left(\beta,\sigma^2\right)\ d\beta\ d\sigma^2}$$

where $L\left(y;x,\beta,\sigma^2\right)$ is the **likelihood** function from section 3.3.1. In the linear model context, typically, a so-called **conjugate prior** specification is used where

$$p\left(\beta|\sigma^2\right)\equiv N\left(v,\sigma^2V\right)\qquad\qquad p\left(\sigma^2\right)\equiv InverseGamma\left(\frac{\alpha}{2},\frac{\gamma}{2}\right)\tag{4.19}$$

($v$ is $p\times 1$, $V$ is $p\times p$ positive definite and symmetric, all other parameters are scalars) and using this prior standard Bayesian calculations show that conditional on the data

$$p\left(\beta|y,\sigma^2\right)\equiv N\left(v^*,\sigma^2V^*\right)\qquad\qquad p\left(\sigma^2|y\right)\equiv InverseGamma\left(\frac{T+\alpha}{2},\frac{c+\gamma}{2}\right)\tag{4.20}$$

where

$$V^*=\left(X^TX+V^{-1}\right)^{-1}\qquad v^*=\left(X^TX+V^{-1}\right)^{-1}\left(X^Ty+V^{-1}v\right)\tag{4.21}$$

$$c=y^Ty+v^TV^{-1}v-\left(X^Ty+V^{-1}v\right)^T\left(X^TX+V^{-1}\right)^{-1}\left(X^Ty+V^{-1}v\right)$$

In regression modelling, it is usual to consider a centred parameterization for $\beta$ so that $v=0$, giving

$$v^*=\left(X^TX+V^{-1}\right)^{-1}X^Ty$$

$$c=y^Ty-y^TX^T\left(XX+V^{-1}\right)^{-1}X^Ty=y^T\left(I_T-X\left(X^TX+V^{-1}\right)^{-1}X^T\right)y$$

IThe critical quantity in a Bayesian clustering procedure is the **marginal likelihood** or **prior predictive distribution** for the data in light of the model.

$$p(y)=\int p\left(y|\beta,\sigma^2\right)p\left(\beta|\sigma^2\right)p\left(\sigma^2\right)d\beta d\sigma^2.\tag{4.22}$$

Combining (3.3.1) and (4.19) gives that

$$p(y)=\left(\frac{1}{\pi}\right)^{T/2}\frac{\gamma^{\alpha/2}\Gamma\left(\frac{T+\alpha}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\frac{|V^*|^{1/2}}{|V|^{1/2}}\frac{1}{\{c+\gamma\}^{(T+\alpha)/2}}\tag{4.23}$$

For a collection of data sequences $y_1,...,y_N$ (4.23) can be evaluated and used as the basis of a **dissimilarity measure** as an input into a hierarchical clustering procedure. The marginal likelihood in (4.23) can easily be re-expressed for clustered data.

### 4.9.7 CHOOSING THE NUMBER OF CLUSTERS

A hierarchical clustering procedure gives the sequence by which the clusters are merged (in agglomerative clustering) or split (in divisive clustering) according the model or distance measure used, but does not give an indication for the number of clusters that are present in the data (under the model specification). This is obviously an important consideration.

One advantage of the model-based approach to clustering is that it allows the use of statistical model assessment procedures to assist in the choice of the number of clusters. A common method is to use approximate **Bayes factors** to compare models of different orders (i.e. models with different numbers of clusters). This method gives a systematic means of selecting the parameterization of the model, the clustering method, and also the number of clusters. The **Bayes factor** is the posterior odds for one model against the other assuming neither is favored *a priori*. Two methods based on the Bayes factor have been used.

- The **Approximate Weight of Evidence** (**AWE**) This is a heuristically derived approximation to twice the log Bayes factor
- The **Bayesian Information Criterion** (**BIC**) A more reliable approximation to twice the log Bayes factor called the *Bayesian Information Criterion*, which, for model $M$ is given by

$$BIC_M = 2\log L_M + const \approx 2\log L_M\left(\widehat{\theta}\right) - d_M \log N$$

  where $L_M$ is the Bayesian marginal likelihood from (4.22), $L_M\left(\widehat{\theta}\right)$ is the maximized log likelihood of the data for the model $M$, and $d_M$ is the number of parameters estimated in the model. The number of clusters is not considered a parameter for the purposes of computing the BIC. The **larger** the value of the BIC, the **stronger** the evidence for the model.

### 4.9.8 DISPLAYING THE RESULTS OF A CLUSTERING PROCEDURE

The principal display plot for a clustering analysis is the *dendrogram*,which plots all of the individual data linked by means of a binary "tree". Such a plot is displayed below. The data comprise a set of gene expression profiles, with expression in challenged cells being measured relative to unchallenged cells, over a time course of five measurements made over a number of hours. The total number of genes in this experiment is 10585. Two dendrograms for different clustering procedures (average and compact linkage) are displayed for a subset of 500 randomly selected genes

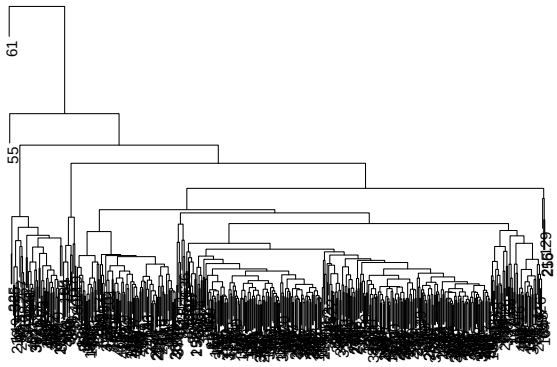



Figure 4.2: Average Linkage

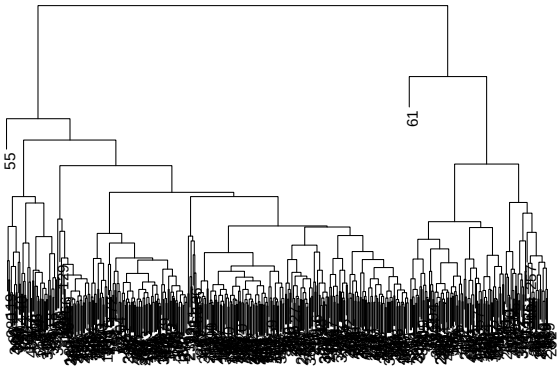



Figure 4.3 Compact linkage

The distance up the tree, or "height" can be used in determining the number of clusters that are present in the data. Another useful plot is the BIC plot for the hierarchical clustering, as this gives an indication of how many clusters are present in the data. Figure 4.4 gives a typical BIC plot for successive numbers of clusters between 0 and 200.

### 4.9.9 CLASSIFICATION VIA MODEL-BASED CLUSTERING

Any clustering procedure can be used as the first step in the construction of classification rules. Suppose that it, on the basis of an appropriate decision procedure,. it is known that there are $C$ clusters, and that a set of existing expression profiles $y_1, ..., y_N$ have been allocated in turn to the clusters. Let $z_1, ..., z_N$ be the cluster allocation labels for the profiles. Now, suppose further that the $C$ clusters can be decomposed further into two subsets of sizes $C_0$ and $C_1$, where the subsets represent perhaps clusters having some common, known biological function or genomic origin. For example, in a cDNA microarray, it might be known that the clones are distinguishable in terms of the organism from which they were derived. A new objective could be to allocate a novel gene and expression profile to one of the subsets, and one of the clusters within that subset. ,

Let $y_{ijk}$ denote, for $i = 0, 1$, $j = 1, 2, ..., C_i$, $k = 1, 2, ..., N_{ij}$ denote the $k$th profile in cluster $j$ in subset $i$. Let $y^*$ denote a new profile to be classified, and $\xi^*$ be the binary classification-to-subset, and $z^*$ the classification-to-cluster variable for $y^*$. Then, by Bayes Rule, for $i = 1, 2$,

$$P\left[\xi^* = i | y^*, y, z\right] \propto p\left(y^* | \xi^* = i, y, z\right) P\left[\xi^* = i | y, z\right] \tag{4.24}$$

The two terms in (4.24) are to be determined on the basis of the clustering output.

# CHAPTER A

# STOCHASTIC PROCESSES AND RANDOM WALKS

## A.1 PROPERTIES OF SIMPLE RANDOM WALKS

The simple **random walk with absorbing states** gives some insight into how the BLAST cumulative score random walk behaves. We consider the behaviour in detail here Consider the homogeneous random walk $Y_n$ say that moves in steps of $+1$ with probability $p$ and $-1$ with probability $q = 1 - p$, but that is absorbed at when it enters states $a$ or $b$ $(a < b)$; suppose that $p \neq q$. For any value $h$ with $a \leq h \leq b$, let

$$w_h = P\left[\text{Random walk is eventually absorbed at } b \text{ after starting at } h\right]$$

$$u_h = P\left[\text{Random walk is eventually absorbed at } a \text{ after starting at } h\right] = 1 - w_h$$

so that $w_a = u_b = 0, w_b = u_a = 1$. It can be shown that

$$w_h = \frac{e^{\theta h} - e^{\theta a}}{e^{\theta b} - e^{\theta a}} \quad \& \quad u_h = \frac{e^{\theta b} - e^{\theta h}}{e^{\theta b} - e^{\theta a}} \qquad \text{where} \qquad \theta = \log\left(\frac{q}{p}\right). \tag{A.1}$$

The number of steps taken until the random walk is absorbed at either $a$ or $b$ (the **absorption time**) is also of interest. This absorption time is a discrete random variable whose probability distribution depends on the starting position $h$, and also on $a, b, p$ and $q$. Denoting by $N_h$ the absorption time random variable, we would ideally like to derive the probability distribution of $N_h$, but unfortunately this is not straightforward. Instead, we compute the expected value of $N_h$, which is more readily available. It can be shown that

$$m_h = E_{f_{N_h}}\left[N_h\right] = \frac{w_h\left(b - h\right) + u_h\left(a - h\right)}{p - q} \tag{A.2}$$

We can now consider limiting cases of $w_h, u_h$ and $m_h$ of particular relevance to biological sequence analysis and the BLAST stochastic process model described in the previous section; the case where the lower absorbing state $a$ is $-1$ and the upper absorbing state $b$ is some threshold value $y \geq 1$. In this case, for $h = 0$, we have from (A.1)

$$w_0 = \frac{1 - e^{-\theta}}{e^{\theta y} - e^{-\theta}}$$

which is the probability that the random walk reaches and is absorbed at $y$ prior to being absorbed at $-1$. It can be shown that this probability is approximately equal to $\left(1 - e^{-\theta}\right) e^{-\theta y}$. But this is the probability that the random walk is absorbed at threshold $y$; hence if $Y$ is the **maximum** state reached by the same random walk **without** the upper absorbing state, then we must have that

$$P\left[Y \geq y\right] \approx \left(1 - e^{-\theta}\right) e^{-\theta y} = C e^{-\theta y} \tag{A.3}$$

say, where $C = \left(1 - e^{-\theta}\right)$. Also, from (A.2), we have (as $b \to \infty$ so that by definition $w_0 \to 0$ and $u_0 \to 1$) that

$$m_0 = E_{f_{N_0}}[N_0] = \frac{bw_0 - u_0}{p - q} \to \frac{1}{q - p} = A, \ say. \tag{A.4}$$

The case $a = -1$ is relevant to previous discussions of the BLAST stochastic process $\{S_n\}$ if we consider the subsections determined by the ladder points; a random walk starting at the most recent ladder point carries out an excursion that comes to a halt when the next ladder point is reached. Thus, if the most recent ladder point is at position $i$ and has cumulative score $s_i$, then the new random walk $\{S'_n\}$ defined by

$$S'_{n-i} = S_n - s_i \qquad n \geq i$$

starts at $h = 0$ and is comes to a halt/is absorbed at $a = -1$ when the next ladder point is reached.

## A.2 SIMPLE RANDOM WALK GENERALIZATIONS

For a more general analytic perspective, consider a homogeneous Markov chain with a range of step sizes and with different associated probabilities

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| STEP SIZE | $-c$ | $-c+1$ | $\cdots$ | $0$ | $\cdots$ | $d-1$ | $d+1$ |
| PROBABILITY | $p_{-c}$ | $p_{-c+1}$ | $\cdots$ | $p_0$ | $\cdots$ | $p_{d-1}$ | $p_d$ |

with the assumptions

(i) $p_{-c}, p_d > 0$

(ii) The **expected step size** $m_{STEP}$ is **negative**, that is

$$m_{STEP} = \sum_{j=-c}^{d} j p_j < 0$$

(iii) The step sizes that have non-zero probability also have no common divisor other than 1.

- such a chain forms the basis for PSIBLAST analysis of protein sequences.

Again, assuming that such a random walk starts at $h = 0$, and is absorbed at state $a = -1$ or reaches threshold $b = y \geq 1$, it is straightforward to see that the random walk will come to a halt at one of the states $-c, -c+1, ..., -1, y, y+1, ...., y+d-1$, and if

$$P_k = P\left[\text{Random walk halts in state } k\right]$$

then it can be shown that

$$\sum_{k=-c}^{-1} P_k e^{k\theta} + \sum_{k=y}^{y+d-1} P_k e^{k\theta} = 1$$

for some unique positive quantity $\theta$ that satisfies

$$\sum_{j=-c}^{d} p_j e^{j\theta} = 1 \tag{A.5}$$

Equation (A.5) can be used to compute $\theta$ for a given set of $p_j$. Clearly, $P_k$ for $k = -c, -c+1, ..., -1, y, y+1, ...., y+d-1$ depends on the threshold value $y$, but the limiting probability $R_j = \lim_{y\to\infty} P_j$ can be defined and computed, and used to derive the limiting expected absorption time

$$A = -\frac{1}{m_{STEP}} \sum_{j=1}^{c} jR_{-j} \tag{A.6}$$

This general Markov random walk has properties that are essentially the same as for the simple random walk described above. For example, it can be shown that if $Y$ is the value at the **maximum** state reached by the walk then analogously to (A.4) above, by considering an absorbing state at threshold $y$ and the probability of being absorbed at $y$, we have

$$P[Y \geq y] \approx Ce^{-\theta y} \tag{A.7}$$

where $\theta$ satisfies (A.5) and $C$ is given by

$$C = \frac{\overline{Q}\left(1 - \sum_{j=1}^{c} R_{-j} e^{-j\theta}\right)}{(1 - e^{-\theta}) \sum_{k=1}^{d} kQ_k e^{k\theta}} \tag{A.8}$$

The quantities $Q_1, Q_2, ..., Q_d$ are probabilities defined by the behaviour of an unrestricted general random walk with the step sizes defined above and with a negative expected step size (and hence a downward drift) In fact, for $k \geq 0$

$$Q_k = P[\text{Unrestricted random walk passes through state} k]$$

so that by construction $\sum_{k=1}^{d} Q_k < 1$ (as the random walk may never visit a positive state) and also

$$\sum_{k=1}^{d} Q_k e^{k\theta} = 1.$$

Finally $\overline{Q}$ is the probability that the random walk never reaches a positive state, that is

$$\overline{Q} = 1 - Q_1 - Q_2 - ... - Q_k$$

Although these expressions are complicated, and require detailed computation, the most important facts are contained in the formula in (A.7) and the knowledge that the constants $C, \theta$ and $A$ can be computed.

# CHAPTER B

# ALGORITHMS FOR HMMs

## B.1 THE FORWARD ALGORITHM

Define the **forwards variables** $\alpha(t,i)$ for $t = 1,2,3,...,n$ and $i \in \mathbb{H}$ by

$$\alpha(t,i) = P\left[X_1 = x_1, ..., X_t = x_t, H_t = h_t = i\right]$$

that is, the joint probability of the observing the actual data up to position $t$ and having the region type at position $t$ equal to $i$. Now, if, for all $i$, the values of $\alpha(n,i)$ are known, then as the terms in (4.15) can be rewritten

$$f(x|h)f(h) = f(x,h) = P\left[X_1 = x_1, ..., X_t = x_n, H_1 = h_1, ..., H_t = h_n\right]$$

it can be shown that

$$f(x) = \sum_{i=0}^{n_H} \alpha(n,i)$$

Our objective will be to compute $\alpha(1,i), \alpha(2,i), \alpha(3,i), ..., \alpha(n,i)$ recursively for each $i \in \mathbb{H}$. First we initialize by setting

$$\alpha(1,i) = \theta_i p_{x_1}^{(i)} \tag{B.1}$$

and then define, each $t$,

$$\begin{aligned} \alpha(t+1,i) &= P\left[X_1 = x_1, ..., X_{t+1} = x_{t+1}, H_{t+1} = h_{t+1} = i\right] \\ &= \sum_{j=0}^{n_H} P\left[X_1 = x_1, ..., X_{t+1} = x_{t+1}, H_{t+1} = h_{t+1} = i, H_t = h_t = j\right] \end{aligned} \tag{B.2}$$

using the Total Probability rule, partitioning with respect to the state in position $t$. However, we have using conditional probability arguments that the summand can be rewritten

$$\begin{aligned} P\left[X_{t+1} = x_{t+1} | X_1 = x_1, ..., X_t = x_t, H_{t+1} = h_{t+1} = i, H_t = h_t = j\right] \times \\ P\left[X_1 = x_1, ..., X_t = x_t, H_{t+1} = h_{t+1} = i, H_t = h_t = j\right] \end{aligned}$$

which can be further simplified as, by assumption the first term is merely

$$P\left[X_{t+1} = x_{t+1} | H_{t+1} = h_{t+1} = i\right] = p_{x_{t+1}}^{(i)} \tag{B.3}$$

and also the second term is

$$P\left[H_{t+1} = h_{t+1} = i | X_1 = x_1, ..., X_t = x_t, H_t = h_t = j\right] P\left[X_1 = x_1, ..., X_t = x_t, H_t = h_t = j\right]$$

where

$$P\left[H_{t+1}=h_{t+1}=i|X_1=x_1,...,X_t=x_t,H_t=h_t=j\right]=P\left[H_{t+1}=h_{t+1}=i|H_t=h_t=j\right]=\theta_{ji} \tag{B.4}$$

and

$$P\left[X_1=x_1,...,X_t=x_t,H_t=h_t=j\right]=\alpha\left(t,j\right) \tag{B.5}$$

Hence combining (B.2)-(B.5) gives

$$\alpha(t+1,i)=\sum_{j=0}^{n_H}p_{x_{t+1}}^{(i)}\theta_{ji}\alpha\left(t,j\right) \tag{B.6}$$

and so we have a recursion formula. In fact (B.1) and (B.6) combined give a method of computing the (conditional) likelihood

$$f(x|\mathcal{P})=\sum_h f(x|h,\mathcal{P})f(h|\mathcal{P}) \tag{B.7}$$

required for (i) that can be completed in $n\times n_H^2$ steps. This number is relatively small compared to $2n\times\left(n_H+1\right)^n$.

## B.2 THE BACKWARD ALGORITHM

The algorithm described above can be alternately implemented in reverse time. Let

$$\beta(t,i)=P\left[X_{t+1}=x_{t+1},...,X_n=x_n|H_t=h_t=i\right]$$

Then we have similar recursion formulae

$$\beta(n-1,i)=\sum_{j=0}^{n_H}\theta_{ij}p_{x_n}^{(j)}$$

and

$$\beta(t-1,i)=\sum_{j=0}^{n_H}p_{x_t}^{(j)}\theta_{ij}\beta(t,j)$$

and thus another means of computing (B.7).

## B.3 THE VITERBI ALGORITHM

The Viterbi algorithm is a **dynamic programming** algorithm for computing $\hat{h}$, that is, the most likely sequence of unobserved states, given the observed data

$$\hat{h}=\arg\max f\left(h|x\right)=\arg\max\frac{f(x|h)f(h)}{f(x)}=\arg\max f(x|h)f(h)=\arg\max f(x,h)$$

It proceeds as follows: first, define

$$\delta_1(i) = P[H_1 = h_1 = i, X_1 = x_1]$$

and

$$\delta_t(i) = \max_{h_1,...,h_{t-1}} P[H_1 = h_1, ..., H_{t-1} = h_{t-1}, H_t = h_t = i, X_1 = x_1, ..., X_t = x_t]$$

so that $\delta_t(i)$ is the maximum probability, over all possible routes, of ending up in unobserved state $i$ at time $t$. Then

$$\max_i \delta_n(i) = \max_{h_1,...,h_n} P[H_1 = h_1, ..., H_n = h_n = i, X_1 = x_1, ..., X_n = x_n]$$

is the maximum probability, over all possible routes, of ending in unobserved state $i$ at time $n$. Secondly, compute the $\delta$s recursively; for each $i$ define

$$\delta_1(i) = \theta_i p_{x_1}^{(i)}$$

and for $t = 2, 3, ..., n$, and $0 \leq j \leq n_H$

$$\delta_t(j) = \max_i \delta_{t-1}(i)\, \theta_{ij} p_{x_t}^{(j)}$$

Finally, let

$$\hat{h}_n = \arg\max_i \delta_n(i)$$

and for $t = n-1, n-2, ..., 2, 1$ define

$$\hat{h}_t = \arg\max_i \delta_t(i)\, \theta_{i\hat{h}_{t+1}}$$

so that $\hat{h}_t$ for each $t$ is the state that maximizes the joint probability. Eventually we have a computed a vector

$$\hat{h} = \left(\hat{h}_1, ..., \hat{h}_n\right) = \arg\max f(x, h)$$

that is required for step (ii).

## B.4 THE BAUM-WELCH ALGORITHM

Normally to estimate parameters in probability models given a data sample, we would utilize a formal estimation procedure such as maximum likelihood. Often this is a difficult problem for HMMs; the likelihood and parameter space are complex and high-dimensional. The **Baum-Welch** algorithm provides a useful means of producing parameter estimates that are at least intuitively appealing and appropriate in the majority of cases, if not theoretically optimal.

The parameters to be estimated in the HMM are the **transition** probabilities $\theta_{ij}$

$$\theta_{ij} = \Pr(\text{Region type } i \text{at time } t \rightarrow \text{Region type } j \text{at time } t+1) = P[H_{t+1} = j | H_t = i] \qquad i, j \in \mathbb{H},$$

the **emission** probabilities in (4.11), written more simply as

$$p_j^{(i)} = P\left[X_t = j | H_t = i\right] \qquad i \in \mathbb{H}, j \in \mathbb{X}$$

(that is, the probability of observing character $j$ in region type $i$), and for full generality, the **marginal** probabilities.

$$\pi_i = \Pr\left(Regiontypei\ atposition1\right) \qquad i \in \mathbb{H}$$

In order to estimate these parameters, we need a set of **training samples** or sequences $\mathbf{Y}^{(D)} = Y^{(1)}, Y^{(2)}...$ with corresponding unobserved state sequences $\mathbf{Q}^{(D)} = Q_1, Q_2, ...$and for sample $d$

$$Y_d = \left(y_1^{(d)}, y_2^{(d)}, ...\right) \qquad\qquad Q_d = \left(q_1^{(d)}, q_2^{(d)}, ...\right)$$

The Baum-Welch approach is an iterative procedure for estimating these parameters using the training samples that proceeds, at each step, conditional on the observed sequence $X = x$. The algorithm proceeds in the following steps

I **Initialization:** choose initial values for $\theta_{ij}, p_j^{(i)}$ and $\pi_i$ from some appropriate probability distribution, or from prior knowledge of the modelling situation

II **Re-estimation:**for a parameter update, set

$$\begin{aligned} \widehat{\pi}_i &= E\left[N_i^{(1)} | \mathbf{Y}^{(D)}\right] \\ \widehat{\theta}_{ij} &= \frac{E\left[N_{ij} | \mathbf{Y}^{(D)}\right]}{E\left[N_i | \mathbf{Y}^{(D)}\right]} \\ \widehat{p}_j^{(i)} &= \frac{E\left[N_i\left(j\right) | \mathbf{Y}^{(D)}\right]}{E\left[N_i | \mathbf{Y}^{(D)}\right]} \end{aligned} \tag{B.8}$$

where the expectations of following random variables

| | |
|---|---|
| $N_i^{(1)}$ | is the number of times region type $i$ appears in position 1 |
| $N_i$ | is the number of occurrences of region type $i$ |
| $N_{ij}$ | is the number of transitions from region type $i$ to region type $j$ |
| $N_i(j)$ | is the number of times character $j$ appears in region type $i$ |

are conditional expectations given the training data sequence.

III **Computation:** Let

$$\xi_t^{(d)}\left(i, j\right) = P\left[q_t^{(d)} = i, q_{t+1}^{(d)} = j | \mathbf{Y}\right]$$

for $i, j \in \mathbb{H}$, where the superscript $(d)$ indicates calculation from the training data sample $d$.. From the conditional probability definition, this expression can be re-written

$$\xi_t^{(d)}\left(i, j\right) = \frac{P\left[q_t^{(d)} = i, q_{t+1}^{(d)} = j, \mathbf{Y}\right]}{P\left[\mathbf{Y}\right]} \tag{B.9}$$

where the denominator can be computed using the Forward or Backward algorithm above, and the numerator can be calculated by using the Forwards and Backwards variables $\alpha(.,.)$ and $\beta(.,.)$ of the previous algorithms

$$P\left[q_t^{(d)} = i, q_{t+1}^{(d)} = j, \mathbf{Y}\right] = \alpha(t,i)\,\theta_{ij} p_{y_{t+1}}^{(d,j)} \beta(t+1,j)$$

where $p_{y_{t+1}}^{(d,j)}$ is the probability of observing character $y_{t+1}$in from in position $t+1$ in region type $j$ in the training data sample $d$. Let

$$I_t^{(d)}(i) = \begin{cases} 1 & if q_t^{(d)} = i \\ 0 & otherwise \end{cases}$$

be an indicator random variable. Then the number of times region type $i$ is observed in the training sample is

$$\sum_d \sum_t I_t^{(d)}(i)$$

(recall $d$ indexes training sample sequences) and the expected number of times is

$$\sum_d \sum_t E\left[I_t^{(d)}(i)|Q^{(d)}\right] = \sum_d \sum_t P\left[I_t^{(d)}(i) = 1|Q^{(d)}\right] = \sum_d \sum_t P\left[q_t^{(d)} = i|Q^{(d)}\right]$$

Using the Theorem of Total Probability and (B.9)

$$P\left[q_t^{(d)} = i|Q^{(d)}\right] = \sum_{j=1}^{n_H} \xi_t^{(d)}(i,j)$$

and hence the expected number of times region type $i$ is observed in the training sample is

$$\sum_d \sum_t \sum_{j=1}^{n_H} \xi_t^{(d)}(i,j). \tag{B.10}$$

Similarly, the expected number of transitions from region type $i$ to region type $j$ is

$$\sum_d \sum_t \xi_t^{(d)}(i,j) \tag{B.11}$$

These formulae can be substituted into (B.8) to compute the iterative procedure. The only remaining quantity to be estimated is

$$E\left[N_i(j)\,|\mathbf{Y}^{(D)}\right]$$

that appears in the numerator in the final iterative formula for $\widehat{p}_j^{(i)}$. This is estimated in a similar fashion to the other quantities; let

$$I_t^{(d)}(i,j) = \begin{cases} 1 & \text{if } q_t^{(d)} = i \text{and } Y_t^{(d)} = j \\ 0 & \text{otherwise} \end{cases}$$

be the indicator variable that is equal to one if, for training sample $d$, character $j$ occurs in region type $i$ at position $t$. Then

$$E\left[N_i(j)\,|\mathbf{Y}^{(D)}\right] = \sum_d \sum_t \sum_{Y_t^{(d)}=j} \sum_{j=1}^{n_H} \xi_t^{(d)}(i,j)$$

which completes the computation.