School of Mathematical and Computer Sciences
Heriot-Watt University

NOTES ON

SURVIVAL MODELS

Takis Konstantopoulos

Spring 2006

These are brief notes on Survival Models for the MSc course in Actuarial Mathematics. I reserve the right to modify/expand the notes; any modifications will be published on my webpage:

`www.ma.hw.ac.uk/~takis`

PLURA FACIUNT HOMINES E CONSUETUDINE, QUAM E RATIONE
(Men do more from habit than from reason)
–Anon.

THERE ARE NO SUCH THINGS AS APPLIED SCIENCES,
ONLY APPLICATIONS OF SCIENCE
–Louis Pasteur

Contents

Preface

This course introduces some basic models used in the analysis of survival data. The goal is to decide, from often incomplete and censored data, how to estimate parameters, test hypotheses and intelligently deduce how to calculate life annuities and other matters of interest in life insurance.

The models can be quite sophisticated based, for example, on the concept of point processes. However, in this class, we shall only discuss the simplest of them. For instance, we shall discuss trivial Markov chains in continuous time with a small, finite, number of states, and model a lifetime as the time till the chain is absorbed by a specific state (the graveyard). The caveat is that we allow for time-inhomogeneous Markov chains for which explicit solutions of even the most trivial functionals are sometimes impossible to derive analytically.

The prerequisites are:
- A course in Probability.
- A course in Statistics.
- Standard Analysis, Calculus and Elementary (Linear) Differential Equations.
- Introductory Stochastic Processes concepts.
- Elementary Algebra and Arithmetic.
- Ability to think, read (in English) and count.
- Elementary (first-order) Logic.

While I will take it for granted that you know what a set, a function, a relation, a set operation, a random variable, and a probability is, I shall try to make the course self-consistent as regards the more "advanced" notions.

Regarding references: Life Insurance was a matter of concern long ago: The famous Leonhard Euler [7] had thought about the problem of how to pay annuities and provided some suggestions. He was not using probability. A very accessible book on survival models is that of Eland-Johnson and Johnson [6]. A more advanced text, using the language of point processes, is that of Fleming and Harrington [8]. A good book, discussing the whole range of the actuarial notation and applications, is the recent book by Błaszczyszyn and Rolski [3] (but it is only available in Polish). Another accessible book is that of Gerber and Cox [9]. Finally, to check your background on Probability and Statistics, measure it against an undergraduate book such as that of Williams [14]. You need to know what is in it; if necessary, review it rapidly.

The reason I wrote the notes is that I feel I cannot teach the subject, (or any other subject for that matter), without providing some rational explanation. Indeed, I strongly believe that there is no way to learn anything unless you understand it, at some level. Also, my constant suggestion is that "learning" by rote is a recipe for failure.* More to the point, it

*That, of course, depends on one's definition of failure vs success.

is rather obvious that there would be no need for a student to come to class and attend my lectures if he/she had to learn things by heart. He/she should take a manual and read it at his/her leisure. Indeed, there are many excellent manuals (cookbooks) regarding the subject of Statistics. These manuals tell you, step-by-step, how to apply a method for data analysis. Good manuals are [15], [16], both available online. But the point of teaching is to–at the minimum–understand what you are reading in the manual and the logic (or absence of it) behind it so that you can apply it intelligently.

These notes are quite elementary in terms of their mathematical content. So, I offer my apologies to the mathematically inclined for they will not find extreme rigour herein. However, they will find that the notes are not devoid of logic and some logical organisation and this has been my purpose. Those who wish to see the topic from a proper mathematical standpoint should consult, e.g., Fleming and Harrington [8] and Liptser and Shiryaev [12].

It is clear that taking discrete observations in time should not be seen as the estimation of random variables but of stochastic processes. Just a trivial example: Let T be the time somebody dies. To estimate T, we should consider the (trivial but useful) process $\mathbf{1}(T \leq t)$, $t \geq 0$. In this sense, the notion of hazard rate (or force of mortality in the actuarial lingo) μ_t should be replaced by the much more natural process $\int_0^{t \wedge T} \mu_s ds$: Indeed, on one hand, the integral of μ_t is more fundamental than μ_t itself (a version of the former exists for any random variable); on the other hand, it makes no sense to go beyond T, and hence the minimum between t and T in the above quantity is natural. Mathematically, $\int_0^{t \wedge T} \mu_s ds$ is the **compensator** of $\mathbf{1}(T \leq t)$, and using this trivial but useful observation makes things much more clear, both for physical AND mathematical reasons, than a completely static point of view. After all, these things have been around for decades...

I have followed the syllabus of the course and not ventured beyond its restricting boundaries. Had I written the notes without paying attention to the syllabus, I would have organised them differently, and then they would make much more sense.

The student will realise that, except for the first 3 chapters, the remaining of them are reviews of what he/she should have learned in Probability and Statistics: Markov chains is standard material in Probability and, in any case, the syllabus only requires discussion of, essentially, just definitions and elementary computations. Then, the Statistics parts of these notes (which could probably be 50% of them) are just standard, plain Statistics subjects as taught in standard courses.

I wrote these notes very quickly and they should not be considered complete. I therefore expect lots of typos... Any corrections would be appreciated.

My thanks to Iain Currie who gave me the notes he has been using. Some of the examples here are his examples.

Chapter 1

Introduction

This chapter introduces you to the notation I shall be using throughout the notes. You should familiarise yourselves with it, along with your revision of Probability and Statistics.

1.1 Overview

We shall be dealing with positive numbers representing lifetimes. Despite the morbid character of these notes, the subject is applicable to situations beyond life and death. For instance, a medical statistician may be interested in finding how long it takes for a new drug to cure a particular disease. She designs an experiment, administers the drug to volunteers* (the subjects) and takes measurements. Things are not as simple as they look. There are several problems. First there are psychological factors, i.e. a subject may, just by believing that he is taking a drug, feel better. The experiment must be designed so that some subjects take the drug, and some not, without them knowing it. Second, a subject may become bored and drop out from the study. At first sight, one may say, forget about the data corresponding to dropouts. But they do provide valuable information: if a subject has dropped out x units of time till he became part our experiment, then we know he has not been cured by that time. We have partial information about the random variable T, i.e. that a particular observation of the random variable resulted a value that is larger than x. So we do not throw away the data, but try to modify our estimators in order to take into account *partial observations* . Partial observations are referred to as *censoring* . The type of censoring just described is called right censoring. Another type is this: a subject enters our experiment at some unknown time A, prior to the start of our observation at time t_0. We observe him completely, i.e. until he is cured at time $A + T$. Thus we know that $A < t_0 < A + T$. The only thing we know about T is that $T > t_0 - A$, but A is unknown. This is called *left censoring.* A mixed situation is when we have left censoring and, in addition, the subject drops out at time $t_1 > t_0$. In this case we know that $A < t_0 < t_1 < A + T$, but A is unknown. So we have left censoring ($T > t_0 - A$) and right censoring ($T > x := t_1 - t_0$). We call this *interval censoring.*

*This is just an assumption. For instance, in Huntsville, Texas, drugs were administered to prisoners on the death row because they were going to die anyway.

The best language to describe Survival Models is the language of point processes. I am only occasionally going to use it, so let me describe what this object is. A *(simple) point process* is a random discrete set.[†] If we talk about a point process of real numbers, from the very fact that real numbers are ordered[‡] we can enumerate the points of a point process as $T_1 < T_2 < \cdots$. (We may also want to introduce points $T_1 > T_0 > T_{-1} > \cdots$.) Note that, on purpose, I do not specify the exact range of the index n because it may be random. Thus, describing the point process is equivalent to describing the random variables $\{T_n\}$ which is a "trivial"[§] matter.

Here is an example: let $U_1, \ldots, U_n$ be i.i.d. uniform random variables in the integral $[0,1] = \{x : 0 \le x \le 1\}$. Then the first point of the point process is $T_1 = \min(U_1, \ldots, U_n)$ and so on, till the last point, which is $T_n = \max(U_1, \ldots, U_n)$. In statistical notation, $T_k := U_{(k)}$, and the joint distribution of $\{T_k, 1 \le k \le n\}$ is easily found via the joint density

$$P(T_1 \in dt_1, \ldots, T_n \in dt_n) = \frac{1}{n!}\mathbf{1}(t_1 < \cdots < t_n)dt_1 \ldots dt_n.$$

(The reader is referred to Section 1.2 for the "dt" notation.) Indeed, there are $n!$ permutations of $(U_1, \ldots, U_n)$ and each permutation has exactly the same density: $P(U_1 \in dt_1, \ldots, U_n \in dt_1) = dt_1 \ldots dt_n$. The *counting measure* associated to any point process is the quantity

$$N(B) = \sum_n \mathbf{1}(T_n \in B),$$

where B is a set. In words, $N(B)$ is the number of points falling in B. The *counting process* associated to the point process is the function

$$N_t = N[0,t], \quad t \ge 0$$

(and if t also runs on negative numbers, we may let $N_t = -N(t,0)$, for $t < 0$). We shall only deal with the simplest of point processes, such as the Poisson process, or the point process of counts of a finite number of censored data.

The *force of mortality (FOM)* is the subject of Chapter 2. This is also known as the *hazard rate* μ_t and $\mu_t\delta$ gives the probability that a lifetime T expires in the δ-neighbourhood of t given that it has not expired up to t, when δ is small. It is only defined for *absolutely continuous* random variables. It is a useful concept because of its practicality, and because Markovian models are essentially defined through FOM's, among other reasons. But its real force lies in the fact that it gives a way to "predict" what a point process will do as it evolves in time. This is described in the Section 2.4.

Chapter 3 discusses the standard way to estimate a distribution function. This is called the Fundamental Theorem of Statistics. It rests on the Fundamental Theorem of Probability, also known as the (Strong) "Law" of Large Numbers.[¶] When we introduce censoring, we have to modify the empirical distribution function and this was done by Kaplan and Meier.

[†]A set of real numbers or vectors is called discrete if it is at most countable and has no limit points. Any finite set is discrete. The set of integers is discrete. The set $1, 1/2, 1/3, 1/4, \ldots$ is countable but has an accumulation point (the point 0), so it is not discrete.

[‡]Ordering means: $x < y$ iff $y - x$ is positive

[§]One man's triviality is another man's professionalism.

[¶]A stupid terminology: there are no laws in Mathematics. Indeed, this law is a theorem which, I hope, you have seen and proved hundreds of times so far, but, for the sake of completeness I prove it once more, in a very simple case.

Next, a parametric estimator is introduced, namely the Cox regression model. It is useful in cases where subjects belong to different types.

In Chapter 4 we introduce continuous time Markov chains with time-varying characteristics. Indeed, the rate of becoming ill changes with the age, so it makes sense, in practice, to seek time-varying models. Unfortunately, if a Markov chain has a cycle, it is, in generally, not tractable analytically. Thus, numerical schemes are devised. The numerical methods are not discussed here, but they are extremely classical. They can be found in any book that discusses time-varying linear systems, i.e. time-varying linear differential equations. There are thousands of research papers on the subject. In Section 4.4 we specialise to the classical homogeneous case. This is a subject you know from undergraduate classes, so the chapter only serves as a reminder. In particular, I remind you how to find means of quantities representing hitting times, e.g. the first time a subject dies or the first time a subject retires, and so on. Section 4.5 discusses Maximum Likelihood Estimators (MLE) for Markovian Models. In other words, we model a lifetime as a Markov chain with some unknown parameters, and, from observations, we estimate the parameters. We discuss properties of the estimators, such as asymptotic normality.

Chapter 5 discusses how to estimate when the only information available is counts. In other words, we only observe numbers of deaths (and not actual times) over specific intervals.

Chapter 6 describes various statistical tests that one performs to see whether certain continuous approximations to estimated data are good or not.

The last part of the notes offers discussion, and exercises to accompany the first part (which is the "theory" part[‖]). This is to be taken seriously too. We shall discuss it in class, but you need to try several of the problems by yourselves.

1.2 Conventions and notation

The *distribution function* of a real-valued random variable X is the function

$$F(x) := P(X \leq x), \quad x \in \mathbb{R}.$$

Note that it is increasing, right-continuous and $F(-\infty) = 0, F(+\infty) = 1$. The *survival function* (or *complementary distribution function*) is the function $1 - F(x) = P(X > x)$; it is denoted either by $\overline{F}(x)$ or by $S(x)$.

More generally, the distribution of a random variable X is any function that allows us to compute probabilities of the form $P(X \in B)$, where B is a nice set. The distribution function is such function. For instance, B may be the set of all numbers whose decimal expansion contains consecutive 0's in blocks of length which is a prime number. Another example is the set B containing all real numbers not exceeding 2.

We say that a sequence $X_1, X_2, \ldots$ of random variables is *i.i.d.* if any finite sub-collection is independent and if all random variables have the same distribution.

A random variable is *continuous* if F is continuous. A random variable is *absolutely continuous* if there is a function f such that $F(x) = \int_0^\infty f(y)dy$. Necessarily, an absolutely continuous random variable is continuous. The function f is called density. Example: Let $\xi_1, \xi_2, \ldots$ be i.i.d. with $P(\xi_1 = 0) = P(\xi_1 = 1) = 1$. Define $X := \sum_{n=1}^\infty \xi_n/2^n$. Then X takes

[‖] Although, I never understood what means theory and what means applications/examples

values between 0 and 1 and $P(X \leq x) = x$ for all $0 \leq x \leq 1$. Since $dP(X \leq x)/dx = 1$, we conclude that $f(x) = 1$ serves as a density, and so this X is absolutely continuous. Next define $Y = \sum_{n=1}^{\infty} 2\xi_n/3^n$. It can be shown that $P(Y \leq y)$ is a continuous function of y but there is no function f such that $P(Y \leq y) = \int_{-\infty}^{y} f(z)dz$, and so Y is continuous but not absolutely continuous.

Caution: In practice, people use a slippery language and say "continuous" when they actually mean "absolutely continuous". We shall be using the informal notation $P(X \in dx)$ to stand for the probability that X is between x and $x + dx$ where dx is "infinitesimal". So, $P(X \in dx) \approx f(x)dx$. The notation is useful when applying the total probability formula, for example, assuming that X, Y are independent (absolutely) continuous random variables,

$$\begin{aligned} P(X+Y \leq t) &= \int P(X+Y \leq t, X \in dx) = \int P(x+Y \leq t, X \in dx) \\ &= \int P(Y \leq t-x)P(X \in dx) = \int P(Y \leq t-x) f_X(x)dx, \end{aligned}$$

and so, differentiating with respect to t, we find

$$f_{X+Y}(t) = \int f_Y(t-x) f_X(x)dx.$$

This notation is more than just notation: it is a correct way of formally expressing something deeper concerning Integration Theory. We refer the student to her standard courses on the subject. Hereafter, we shall not bother explaining "technicalities".**

When a random variable is not continuous, it is called discrete. A discrete random variable X assumes at most countably many values $c_1, c_2, \ldots$, and so we may define the probabilities $P(X = c_i) = p_i$ directly. Of course, $\sum_i p_i = 1$. Let us denote by $\mathfrak{d}_c$ (the *Dirac distribution*) the distribution of a random variable that takes value c with probability 1. In other words,

$$\mathfrak{d}_c(B) = P(X \in B) = \mathbf{1}(c \in B) = \begin{cases} 1 \text{ if } c \in B \\ 0 \text{ if } c \notin B \end{cases},$$

where B is a set of real numbers. Then a random variable that takes values c_i with probability p_i has distribution

$$\sum_i p_i \mathfrak{d}_{c_i}.$$

Indeed, for any set B,

$$P(X \in B) = \sum_i p_i \mathbf{1}(c_i \in B) = \sum_i p_i \mathfrak{d}_{c_i}(B).$$

A random variable may be mixed, i.e. it may have a continuous and a discrete part.

Example 1.1. Let T be exponential with rate μ. Find the distribution of the random variable $X := T \wedge c$, for a given constant $c > 0$.
Solution: It is clear that X is neither discrete nor continuous; it is mixed. Indeed, $P(X = c) = P(T > c) = e^{-\mu c}$. On the other hand, if $0 < x < c$, we have $P(X \leq x) = P(T \leq x) =$

**But bear in mind that... one man's technicality is another man's professionalism.

$1 - e^{-\mu x}$. The latter faction has derivative $\mu e^{-\mu x}$. We can describe the situation by saying that

$$\begin{cases} \text{With probability} & p := e^{-\mu c} \text{ the r.v. } X \text{ takes value } c \\ \text{With probability} & 1-p \text{ the r.v. } X \text{ has density } (1-p)^{-1}\mu e^{-\mu x}\mathbf{1}(0 < x < c). \end{cases}$$

In other words, if B is a set,

$$\begin{aligned} P(X \in B) &= p\mathbf{1}(X \in B) + \int_{B\cap(0,c)} \mu e^{-\mu x} dx \\ &= p\mathfrak{d}_c(B) + (1-p)\int_{B\cap(0,c)} \frac{\mu e^{-\mu x}}{1-p} dx. \end{aligned}$$

Formally, we can write

$$P(X \in dx) = p\mathfrak{d}_c(dx) + \mu e^{-\mu x} dx, \quad 0 \le x \le c.$$

The reader is advised to keep this last notation in mind because I am going to use it later.

The notation $X \stackrel{\mathrm{d}}{=} Y$ means that the random variables X and Y have the same distribution (one says that they are equal in distribution). Occasionally, one writes $X \sim Y$. When F is a distribution function, $X \sim F$ means that F is the distribution function of the random variable X. The notation applies to vectors as well. So $(X, Y) \stackrel{\mathrm{d}}{=} (X', Y')$ means that $P(X \le x, Y \le y) = P(X \le x', Y \le y')$ for all x, y, so, obviously, $(X, Y) \stackrel{\mathrm{d}}{=} (X', Y')$ implies that $X \stackrel{\mathrm{d}}{=} X'$ and $Y \stackrel{\mathrm{d}}{=} Y'$ (but the converse is false).

Concerning random variables with positive values, the following are notations used in Actuarial Science . We let T be such a random variable, and F its distribution function.

1. $F_x(t) := P(T \le t + x \mid T > x) =: {}_tq_x$
2. f_x is the density of F_x (whenever it exists)
3. T_x is a random variable with distribution F_x (note that T_x cannot, in general, be defined as a deterministic function of T; rather, it is defined by means of its distribution). In other words: $T_x \stackrel{\mathrm{d}}{=} (T - x \mid T > x)$
4. $\overline{F_x}(t) = S_x(t) = P(T > t + x \mid T > x) =: {}_tp_x$
5. $q_x = {}_1q_x$ (the 1 here has no significance other than it designates a standard period such as one year[††]; by possibly changing units we can always make the standard period equal to 1)
6. $p_x = {}_1p_x$.

[††]Our confusion with units of time, such as years, months, days, etc., goes back to the Babylonians . You will have noticed that one year has 12 months, one month has 30 days (approximately), one hour has 60 minutes, one minute has 60 seconds. All these integers are divisors of 60. Just as we use (mostly) decimal, the Babylonians used a sexagesimal numbering system, that is, base 60 (schoolchildren , presumably, had to memorise a multiplication table with $60 \times 61/2 = 1830$ entries). There was a good reason for that: 60 has a large number of divisors: $1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60$ (and every number below 60 has at most 10 divisors) and, for want of accuracy, this was convenient. What is quite silly is that, about 4000 years later, we still use the same old-fashioned system when it comes to time units. If, therefore, a tradition takes at least so many years to break, then it is no wonder that the actuarial notation will still be with us for a while.

Chapter 2

Force of mortality

We introduce hazard rates as alternative descriptions for the law of an absolutely continuous random variable. We shall mostly use the actuarial term *force of mortality* in lieu of *hazard rate* . These are mathematically useful functions but also very convenient in applications (one talks of mortality rates). . We will learn how to handle these objects and how to approximate them when we partially know them. The real usefulness arises in connection with discrete-event phenomena occurring at discrete, a priori unknown, epochs of time. The force of mortality gives rise to the concept of the compensator which is a smoothed, predictable, counterpart of a counting process. We will learn how to compute compensators which will later turn out to be the correct tools for constructing statistical estimators. Their appeal lies in their intuitive meaning which makes is easy for us to understand what the estimators are about and why they work.

2.1 Positive random variables

In this section, we shall pay special attention to the random variable T_x that, as you will recall, is defined to have the distribution of $T - x$, given that $T > x$. Some warm-up facts/exercises:

Fact 2.1. *If T has survival function $S(t) = P(T > t)$ then T_x has survival function $S_x(t) = P(T_x > t) = S(x + t)/S(x)$, $t \geq 0$. If T has density f then T_x has density $f_x(t) = f(x + t)/S(x)$, $t \geq 0$. (Left as an exercise.)*

View the operation $Z \mapsto Z_x$ as an operation that changes the distribution of the random variable Z and produces the distribution of the random variable Z_x. We can substitute any random variable in place of Z. In particular, we can let $Z = T_t$, in which case $Z_x = T_{t,x}$ has the following meaning: it is in distribution equal to $T_t - x$ given that $T_t > x$.

Fact 2.2. *It holds that $T_{t,x} \stackrel{d}{=} T_{t+x}$ for any $t, x \geq 0$. (Exercise.)*

The force of mortality (FOM or hazard rate) of a positive random variable T with density

f and distribution function F is defined as the function

$$\mu_t := \frac{f(t)}{1 - F(t)}, \quad t \geq 0.$$

The FOM is clearly depicted in Figure 2.1. By interpreting $0/0$ as zero we may define μ_t

Figure 2.1: *The Force of Mortality*

for all $t \geq 0$. The physical meaning of μ_t is that

Proposition 2.1. *As $h \to 0$,*

$$P(T \leq t + h \mid T > t) = P(T_t \leq h) = \mu_t h + o(h).$$

In other words, μ_t is the value of the density of T_t at zero: $\mu_t = f_t(0)$.

Proof. Since $f(t) = -dS(t)/dt$, we have $S(t) - S(t + h) = f(t)h + o(h)$, as $h \to 0$, and so

$$P(T \leq t + h \mid T > t) = \frac{S(t) - S(t + h)}{S(t)} = \frac{f(t)h + o(h)}{S(t)} = \mu_t h + o(h),$$

where we used the fact that $o(h)$ divided by something that does not depend on h is still $o(h)$. □

The logarithmic derivative of a function g is, by definition, the derivative of $\log g$, i.e. it is equal to g'/g.

Fact 2.3. *Observe that μ_t is the negative of the logarithmic derivative of $S(t)$, i.e. $\mu_t = -S'(t)/S(t)$. (Exercise.)*

Proposition 2.2. *From the mortality rate of T we can recover its survival function and, consequently, its density:*

$$S(t) = \exp - \int_0^t \mu_u du$$
$$f(t) = S(t)\mu_t$$

Proof. From Exercise 2.3 we have that $-\mu_u = \frac{d}{du} \log S(u)$. Integrate this from 0 to t to get $-\int_0^t \mu_u du = \log S(t) - \log S(0)$. But $S(0) = P(T > 0)$. Since we assume that T has density, we must have $P(T = 0) = 0$ and so $P(T > 0) = 1$. So $S(0) = 1$ which has logarithm zero. Applying the exponential function on both sides we obtain the desired formula. The formula for $f(t)$ is found by $f(t) = -\frac{d}{dt} S(t)$. □

Note that $f(t) = S(t)\mu_t$ has a natural physical meaning. Think of T as a lifetime. First, $\mu_t dt$ is, approximately, the probability that the lifetime expires between t and $t + dt$, when dt is infinitesimal, given it has not yet (i.e. by time t) expired. Multiplying this by the probability $S(t)$ that the lifetime exceeds t we get, by the definition of conditional probability, the (unconditional) probability that the lifetime expires between t and $t + dt$; which is, by definition, $f(t)dt$.

Proposition 2.3. *Let T have force of mortality μ_t. Then T_x has force of mortality $\mu_{x,t} = \mu_{x+t}$, survival function $S_x(t) = \exp - \int_x^{x+t} \mu_u du$, and density $f_x(t) = S(x+t)\mu_{x+t}/S(x)$.*

Proof. From Exercise 2.1 we have that $T_{x,t} \stackrel{\mathrm{d}}{=} T_{x+t}$, and this establishes the first claim. The second claim follows from Fact (2.2) by replacing T by T_x:

$$S_x(t) = \exp - \int_0^t \mu_{x,u} du.$$

But we just showed that $\mu_{x,u} = \mu_{x+u}$. So

$$S_x(t) = \exp - \int_0^t \mu_{x+u} du = \exp - \int_x^{x+t} \mu_{x+v} dv,$$

where we made the change of variable $u \mapsto v = x + u$. The density of T_x is, again by Fact 2.2, $S_x(t)\mu_{x,t}$, and, since $S_x(t) = S(x+t)/S(x)$ (Exercise 2.1) and $\mu_{x.t} = \mu_{x+t}$, we have what we need. □

Fact 2.4 (Exercise in notation)**.** *By making use of the actuarial notation, show that:*

$$\begin{aligned}
{}_{s+t}p_x &= {}_tp_x \cdot {}_sp_{x+t} \\
f(t) &= {}_tp_0 \cdot \mu_t \\
f_x(t) &= {}_tp_{x+t} \cdot \mu_{x+t} \\
{}_tp_0 &= \exp - \int_0^t \mu_s ds \\
{}_tp_x &= \exp - \int_x^{x+t} \mu_s ds \\
{}_tq_x &= \int_0^t {}_sp_x \cdot \mu_{x+s} ds
\end{aligned}$$

Fact 2.5. *(i) Suppose that $S(t) = e^{-\lambda t}$, $t \geq 0$, for some positive constant λ. Then $\mu_t \equiv \lambda$. (ii) Suppose that T is uniform on the interval $[0, 1]$. Then $\mu_t = 1/(1-t)$, for $0 \leq t < 1$ (and zero otherwise). (iii) Suppose that $S(t) = t^{-\alpha}$, for $t \geq 1$, for some positive constant α. Then $\mu_t = \alpha/t$, for $t \geq 1$. (Exercise.)*

Comments: In case (i) we have an example of a random variable with constant force of mortality. In case (ii) we have increasing force of mortality (also known as increasing failure rate). In case (iii) we have decreasing force of mortality (also known as decreasing failure rate).

Fact 2.6. *There is a random variable with non-monotonic force of mortality. (Exercise.)*

Important remark: Are there random variables, other the one in (i), with constant force of mortality? The answer is NO. Because, if μ_t is constant, say equal to some positive constant λ, then, by Fact 2.2, $S(t)$ MUST be $e^{-\lambda t}$, as in case (i). Such a random variable, as you recall, called exponential. An exponential random variable also has the *memoryless property*:

$$T_t \stackrel{\mathrm{d}}{=} T, \quad \text{for all } t.$$

Fact 2.7. *Among positive continuous random variables, only the exponential has the memoryless property. Also show that the memoryless property can be written as ${}_tp_s = {}_tp_0$ for all $s, t \geq 0$. (Exercise.)*

2.2 Interpolation

Frequently, we may have some partial information about a lifetime T, i.e., for example, we may know ${}_hq_x = P(T < x + h \mid T > x)$, for a fixed x and h and may, from this information alone, find a "reasonable" formula for the distribution $P(T < x+t \mid T > h)$, for all $t \in [0, h]$. From this formula we will be able to compute quantities of interest, such as the probability that the lifetime expires within a subinterval $[x + a, x + b]$ of $[x, x + h]$. The following are three common practices: (We shall take, without loss of generality, $h = 1$.)

I. The uniform assumption

This states that T_x is uniform on $[0, 1]$ conditionally on $T_x < 1$. In other words,

$$P(T_x < t \mid T_x < 1) = t, \quad 0 \leq t \leq 1.$$

Proposition 2.4. *The uniform assumption holds iff*[*]

$$P(T_x < t) = tP(T_x < 1)$$

In actuarial notation, recall that $P(T_x < t)$ is denoted also as ${}_tq_x$ and $P(T_x < 1)$ as ${}_1q_x$. The latter is *also* denoted as q_x.

Fact 2.8 (Exercise in notation)**.** *Show Fact 2.4 and thus show that the uniform assumption is equivalent to*

$${}_tq_x = tq_x.$$

Now consider finding ${}_{b-a}q_{x+a} = P(T_x < b \mid T_x > a)$, which is the probability that the lifetime expires in the interval $[x + a, x + b]$, given it had not expired before $x + a$, for any $0 < a < b < 1$.

Proposition 2.5. *Under the uniform assumption,*

$${}_{b-a}q_{x+a} = \frac{(b-a)q_x}{1 - aq_x}.$$

[*] "iff" means "if and only if"

Proof.

$$\begin{aligned}
{}_{b-a}q_{x+a} &= P(T_x < b \mid T_x > a) \qquad \text{[by definition]} \\
&= \frac{P(a < T_x < b)}{1 - P(T_x < a)} \qquad \text{[conditional probability]} \\
&= \frac{P(T_x < b) - P(T_x < a)}{P(T_x > a)} \\
&= \frac{bP(T_x < 1) - aP(T_x < 1)}{1 - aP(T_x < 1)} \qquad \text{[from the uniform assumption]} \\
&= \frac{(b-a)q_x}{1 - aq_x} \qquad \text{[remember that } P(T_x < 1) = q_x\text{]}.
\end{aligned}$$

□

Proposition 2.6. *The FOM $\mu_{x,t}$ of T_x is*

$$\mu_{x,t} = \mu_{x+t} = \frac{q_x}{1 - tq_x}, \quad 0 \le t \le 1.$$

Proof. The first equality follows from Exercise 2.2, while the latter is from the definition of the FOM:

$$\mu_{x,t} = \mu_{x+t} = \frac{\frac{d}{dt}P(T_x < t)}{P(T_x > t)}.$$

But $P(T_x < t) = tq_x$, so $\frac{d}{dt}P(T_x < t) = q_x$, while $P(T_x > t) = 1 - P(T_x < t) = 1 - tq_x$. □

II. The Balducci assumption

This states that $P(T_x < 1 \mid T_x > t) = c_1 + c_2 t$.

Proposition 2.7. *The Balducci assumption holds iff*

$$P(T_x < 1 \mid T_x > t) = (1-t)P(T_x < 1).$$

Proof. Since, for $t = 0$, $P(T_x < 1 \mid T_x > 0) = P(T_x < 1)$, while for $t = 1$, $P(T_x < 1 \mid T_x > 1) = 0$, we have $c_1 + c_2 = 0$, and $c_1 = P(T_x < 1)$. □

Fact 2.9 (Exercise in notation)**.** *The Balducci assumption is equivalent to*

$${}_{1-t}q_{x+t} = (1-t)q_x, \quad 0 \le t \le 1.$$

Proposition 2.8. *Under the Balducci assumption,*

$$P(T_x > t) = \frac{1 - q_x}{1 - (1-t)q_x}.$$

Proof. From Fact 2.7, we have

$$(1-t)q_x = \frac{P(T_x < 1, T_x > t)}{P(T_x > t)} = \frac{P(T_x > t) - P(T_x > 1)}{P(T_x > t)}$$

Solving this for $P(T_x > t)$ we obtain the formula. □

As before, we are interested in finding ${}_{b-a}q_{x+a}$.

Proposition 2.9. *Under the Balducci assumption,*

$$ {}_{b-a}q_{x+a} = \frac{(b-a)q_x}{1-(1-b)q_x}. $$

Proof.

$$ \begin{aligned} {}_{b-a}q_{x+a} &= P(T_x < b \mid T_x > a) \qquad \text{[by definition]} \\ &= \frac{P(a < T_x < b)}{1 - P(T_x < a)} \qquad \text{[conditional probability]} \\ &= \frac{P(T_x > a) - P(T_x > b)}{P(T_x > a)} \\ &= \frac{\frac{1-q_x}{1-(1-a)q_x} - \frac{1-q_x}{1-(1-b)q_x}}{\frac{1-q_x}{1-(1-a)q_x}} \qquad \text{[from Fact 2.8]} \\ &= \frac{(b-a)q_x}{1-(1-b)q_x} \qquad \text{[remember that } P(T_x < 1) = q_x\text{]}. \end{aligned} $$

Finally, we find the hazard rate:

Proposition 2.10. *Under the Balducci assumption, the FOM of T_x is*

$$ \mu_{x,t} = \mu_{x+t} = \frac{q_x}{1-(1-t)q_x}. $$

Proof. The formula is found by applying $\mu_{x,t} = -\frac{d}{dt}P(T_x > t)/P(T_x > t)$ and using Fact 2.8. □

III. Constant FOM assumption

This says that T_x has constant FOM on the interval $[0,1]$, i.e.

$$ \mu_{x,t} = \mu_{x+t} = \mu_x, \quad 0 \le t \le 1. $$

Proposition 2.11. *Under constant FOM assumption,*

$$ P(T_x > t) = e^{-\mu_x t}, \quad 0 \le t \le 1. $$

Proof. This readily follows from the formula of Fact 2.2. □

Fact 2.10. *The FOM assumption is equivalent to*

$$ \mu_{x,t} = \mu_{x+t} = -\log(1-q_x), \quad 0 \le t \le 1, $$

and to

$$ {}_tq_x = P(T_x > t) = (1-q_x)^t. $$

(Exercise.)

Finally, we find ${}_{b-a}q_{x+a}$.

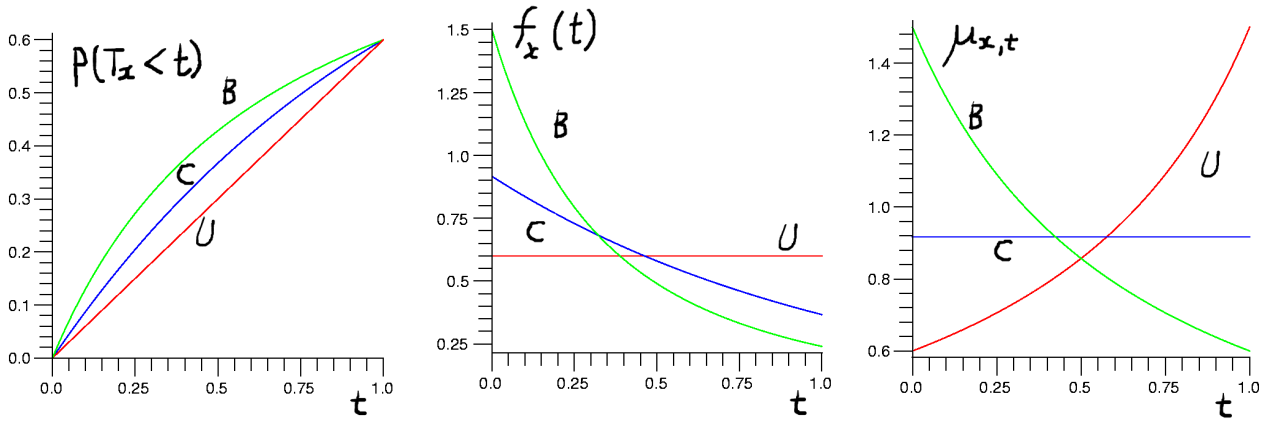
Proposition 2.12. *Under the constant FOM assumption,*

$$ {}_{b-a}q_{x+a} = \frac{(1-q_x)^a - (1-q_x)^b}{(1-q_x)^a}. $$

Proof.

$$ \begin{aligned} {}_{b-a}q_{x+a} &= P(T_x < b \mid T_x > a) \qquad \text{[by definition]} \\ &= \frac{P(a < T_x < b)}{P(T_x > a)} \qquad \text{[conditional probability]} \\ &= \frac{P(T_x > a) - P(T_x > b)}{P(T_x > a)}, \end{aligned} $$

and now use Exercise 2.10. □



The plots above compare the distribution functions $P(T_x < t)$, the distribution functions $f_x(t)$, and the FOMS $\mu_{x,t}$, for $0 \leq t \leq 1$, under the different assumptions. Notice that, despite the apparent similarity of the distribution functions, the FOMs are vastly different. Notice that the FOMs under U and B are symmetric around the vertical line at the point $t = 1/2$. Also, while U results in increasing failure rate, B results in decreasing failure rate.

2.3 Point processes

I shall now devote this small section to explain what a *point process* is.

Definition 2.1. A point process is a discrete random set.

I explain: Discrete means that (i) the set can be enumerated and (ii) the set has no accumulation points. We shall be dealing with point processes which are contained in $\mathbb{R}_+$, the set of non-negative real numbers.

Example 2.1. Consider one random variable T. Consider the set $\{T\}$. This is a point process with one element.

Example 2.2. Let $T_1, \ldots, T_n$ be n random variables (perhaps independent, but not necessarily). The set $\{T_1, \ldots, T_n\}$ is a point process.

Definition 2.2. The *counting process* corresponding to the point process $\{T_1, T_2, \ldots\}$ is defined by

$$N_t = \sum_{n \geq 1} \mathbf{1}(T_n \leq t).$$

I explain: In other words, N_t counts the number of points T_n that fall below t. Note that N_t increases when t increases and has a jump precisely at each of the points T_n. Hence, if we know the function N_t we can recover the points T_n by looking at the jump epochs of N_t. Thus, the set $\{T_1, T_2, \ldots\}$ and the function N_t convey precisely the same information. From now on, I shall be very loose when I speak about a point process. I may be referring to the set $\{T_1, T_2, \ldots\}$ or to the function N_t.

Example 2.3. Let $\tau_1, \tau_2, \ldots$ be i.i.d. exponential random variables, i.e. $P(\tau_i > t) = e^{-\lambda t}\mathbf{1}(t \geq 0)$. Let $T_n = \tau_1 + \cdots + \tau_n$. Then $N_t = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t)$ is a *Poisson process* .

2.4 The force of the force of mortality

The real force of the force of mortality lies in the fact that it can predict what is about to happen in a stochastic discrete-event phenomenon.

Let us redo what we did so far by thinking in terms of point processes. The most trivial point process is, arguably, one that consists of a single point. So let T be a (positive) random variable and consider the singleton $\{T\}$. The counting process of it is

$$N_t = \mathbf{1}(t \geq T), \quad t \geq 0.$$

(We take it to be right-continuous.) Assume that T is absolutely continuous.

Proposition 2.13. *There exists a unique continuous stochastic process* $\{\widetilde{N}_t, t \geq 0\}$ *such that, for each* t, $\widetilde{N}_t$ *is a deterministic function of the collection of random variables* $(N_u, u \leq t)$, *and*

$$E(N_t - N_s \mid N_u, u \leq s) = E(\widetilde{N}_t - \widetilde{N}_s \mid N_u, u \leq s), \quad s < t. \tag{2.1}$$

This process is given by

$$\widetilde{N}_t = \int_0^{t \wedge T} \mu_u du. \tag{2.2}$$

Proof. Obviously, (2.2) is also written as $\widetilde{N}_t = \int_0^t \mu_u \mathbf{1}(u \leq T) du$ and hence it is a deterministic function of $(N_u, u \leq t)$. Note that $(N_u, u \leq s)$ contains the information about the occurrence of T before s. So if $T \leq s$ then $N_s = 1 = N_t$. On the other hand, $\widetilde{N}_s = \widetilde{N}_T = \widetilde{N}_t$. Hence both sides of (2.1) are zero. If $T > s$ then

$$E(N_t - N_s \mid N_u, u \leq s) = P(T \leq t \mid T > s).$$

On the other hand, $\widetilde{N}_t - \widetilde{N}_s = \int_s^t \mu_u \mathbf{1}(T \geq u) du$, so that

$$E(\widetilde{N}_t - \widetilde{N}_s \mid N_u, u \leq s) = \int_s^t \mu_u P(T \geq u \mid T > s) ds.$$

Recall that $P(T \geq u \mid T > s) = \exp - \int_s^u \mu_x dx$ so the last display is further written as

$$\int_s^t \mu_u e^{-\int_s^u \mu_x dx} du = -\int_s^t \frac{d}{du} e^{-\int_s^u \mu_x dx} du = 1 - e^{-\int_s^t \mu_x dx},$$

where the last line follows from the Fundamental Theorem of Calculus.† But then this is $P(T \leq t \mid T > s)$, which verifies 2.1. The proof of uniqueness is omitted. □

Now turn this last fact into a definition:

Definition 2.3 (FOM of a point process)**.** We say that a point process N admits a FOM λ_t if (2.1) holds with $\widetilde{N}_t = \int_0^t \lambda_s ds$.

The reader will excuse me, but, at this point, I shall change terminology and call the FOM of a point process *stochastic intensity*. On the other hand, the process $\widetilde{N}_t$ (the integral of the FOM) is called *compensator*.

Example 2.4. Consider a rate λ Poisson point process. Let N_t be its counting process. Then its stochastic intensity is $\widetilde{N}_t = \lambda t$. Indeed, $E(N_t - N_s \mid N_u, u \leq s) = E(N_t - N_s)$ because a Poisson process has independent increments. Since $N_t - N_s$ is a Poisson random variable with mean $\lambda(t-s)$, the claim follows.

Example 2.5. Consider two independent, positive, absolutely continuous random variables τ_1, τ_2, with FOMs μ_1, μ_2 and let $T_1 = \tau_1, T_s = \tau_1 + \tau_2$. Consider the point process $\{T_1, T_2\}$. Find its stochastic intensity.
Solution: Let me offer you the "quick and dirty way" for doing this. I don't want to be writing stuff like $(N_u, u \leq t)$ so let us denote this by $\mathscr{F}_t$. Basically, we are looking for a $\widetilde{N}_t$ so that, when dt is a small change of time, the corresponding change dN_t of the counting process relates to the change $d\widetilde{N}_t$ via

$$E(dN_t \mid \mathscr{F}_t) = E(d\widetilde{N}_t \mid \mathscr{F}_t) = d\widetilde{N}_t,$$

and where I was allowed to push $d\widetilde{N}_t$ out because of the the continuity. Indeed, the fact that $\widetilde{N}_t$ is continuous allows us to "predict" infinitesimally into the future, something that is not possible for N_t. We shall distinguish three cases: (i) $t < T_1$, (ii) $T_1 \leq t < T_2$, (iii) $T_2 \leq t$. We are careful and observe that each of these events is contained in the information $\mathscr{F}_t$. In case (i) we have, as before,

$$E(dN_t \mid \mathscr{F}_t) = P(T_1 \in dt \mid T_1 > t) = \mu_1(t)dt.$$

In case (ii) we have

$$E(dN_t \mid \mathscr{F}_t) = P(T_1 + \tau_2 \in dt \mid T_1, \tau_2 > t - T_1) = \mu_2(t - T_1)dt.$$

In case (iii) we have

$$E(dN_t \mid \mathscr{F}_t) = 0.$$

So $\widetilde{N}_t$ is a process which is defined through its derivative as:

$$\frac{d}{dt}\widetilde{N}_t = \mu_1(t)\mathbf{1}(t < T_1) + \mu_2(t - T_1)\mathbf{1}(T_1 \leq t < T_2).$$

Check that, upon integrating, we have

$$\widetilde{N}_t = \int_0^{t \wedge T_1} \mu_1(s)ds + \int_{T_1}^{t \wedge T_2} \mu_2(s - T_1)ds, \tag{2.3}$$

(where an integral $\int_a^b$ with $a > b$ is taken to be zero.) Now that we have guessed a formula for $\widetilde{N}_t$, you may rigorously prove that it satisfies (2.1).

†I think that every mathematical discipline has its fundamental theorem. In this course, you are privileged to use the fundamental theorems (i) of Calculus, (ii) of Probability and (iii) of Statistics. Do you know a mathematical discipline whose fundamental theorem is not proved by means of the results of the discipline itself? (Hint: Algebra)

Fact 2.11. *The process $\widetilde{N}$ defined in* (2.3) *is the compensator of the point process N of Example 2.5. (Exercise.)*

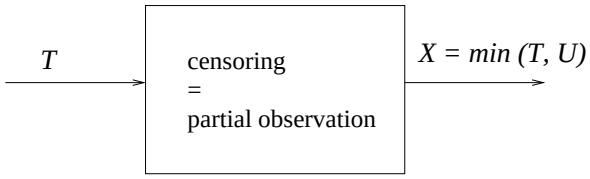
Fact 2.12. *Let T be an absolutely continuous positive random variable with density $f(t)$, distribution function $F(t)$, survival function $S(t) = 1 - F(t)$, and FOM $\mu_t = f(t)/S(t)$. Then the compensator $\widetilde{N}_t$ of the point process $\{T\}$ is given by*

$$\widetilde{N}_t = -\log S(T \wedge t).$$

(Exercise; hint: This same $\widetilde{N}_t$ was derived in equation (2.2)*.)*

Suppose now that T is *censored* by the random variable U. This means that, whereas T represents an actual lifetime, we do not observe T; rather, we observe

$$X := T \wedge U :$$



Consider now

$$N_t := \mathbf{1}(X \le t, T \le U).$$

This is the counting process corresponding to a random set

$$\begin{cases} \{T\} & \text{if } T \le U \\ \emptyset & \text{if } T > U. \end{cases}$$

Fact 2.13. *Let μ_t be the FOM of T and assume that U is independent of T. Then the compensator of N_t is given by*

$$\int_0^t \mathbf{1}(X \ge s)\mu_s ds = -\log S(t \wedge T \wedge U).$$

(Exercise.)

Now let $T_1, T_2, \ldots, T_n$ be independent random variables (positive, absolutely continuous). Suppose that T_i is censored by U_i, and that all random variables are mutually independent. Let

$$X_i = T_i \wedge U_i, \quad n = 1, \ldots, n.$$

Consider the point processes

$$N_t^i = \mathbf{1}(T_i \le t, T_i \le U_i), \quad n = 1, \ldots, n.$$

Their *superposition* , i.e. the counting process corresponding to $\{T_1, \ldots, T_n\}$ is

$$N_t = \sum_{i=1}^{n} N_t^i.$$

To find the compensator $\widetilde{N}_t$ is to find a continuous process such that (2.1) holds. But observe that

$$\begin{aligned} E(N_t - N_s \mid N_u, u \leq s) &= \sum_{i=1}^{n} E(N_t^i - N_s^i \mid N_u, u \leq s) \\ &= \sum_{i=1}^{n} E(N_t^i - N_s^i \mid N_u^i, u \leq s), \end{aligned}$$

where the last equality follows from independence (N^i contains no information about N^j if $j \neq i$). But then we can use Proposition 2.13 to assert that the compensator of N_t^i is given by

$$\widetilde{N_t^i} = \int_0^t \mathbf{1}(X_i \geq s)\mu_s^i ds.$$

Therefore,

$$E(N_t^i - N_s^i \mid N_u^i, u \leq s) = E(\widetilde{N_t^i} - \widetilde{N_s^i} \mid N_u^i, u \leq s).$$

Now assume that the T_i have the same distribution, so that

$$\mu_t^i \equiv \mu_t.$$

In this case, let

$$\widetilde{N}_t = \sum_{i=1}^{n} N_t^i = \int_0^t \sum_{i=1}^{n} \mathbf{1}(X_i \geq s)\mu_s ds.$$

and observe that this is the compensator of N_t. We summarise this important result below.

Fact 2.14. *Let $T_1, \ldots, T_n$ be i.i.d. with common FOM μ_t. Let $U_1, \ldots, U_n$ be independent random variables, and independent of the $T_1, \ldots, T_n$. Define $X_i = T_i \wedge U_i$, $i = 1, \ldots, n$ (censoring). Consider the point process*

$$N_t := \sum_{i=1}^{n} \mathbf{1}(X_i \leq t, T_i \leq U_i).$$

Define the number-at-risk *process*

$$Y_t := \sum_{i=1}^{n} \mathbf{1}(X_i \geq t).$$

Then the compensator of N_t is given by

$$\widetilde{N}_t = \int_0^t Y_u \mu_u du.$$

(Exercise.)

2.5 More force to the force of mortality

You may be wondering, at this point, why we are wasting so much time with new definitions and calculations. But here is why.

First, recall that the compensator $\widetilde{N}_t$ of a point process N_t is a continuous process which satisfies (2.1), which we repeat below:

$$E(N_t - N_s \mid N_u, u \leq s) = E(\widetilde{N}_t - \widetilde{N}_s \mid N_u, u \leq s), \quad s < t.$$

Also, for each t, $\widetilde{N}_t$ it is a function of $(N_u, u \leq t)$. By taking expectations we find that

$$E(N_t - N_s) = E(\widetilde{N}_t - \widetilde{N}_s), \quad s < t.$$

So, in particular,

$$EN_t = E\widetilde{N}_t.$$

This, by itself, is important, because, often $\widetilde{N}_t$ is easier to handle than N_t.

Now suppose that G_t is some continuous (or, simply left-continuous) process, such that, for each t, G_t is a function of $(N_u, u \leq t)$, and consider the quantity $\int_0^t G_u dN_u$ defined by

$$H_t := \int_0^t G_u dN_u = \sum_n G_{T_n}, \tag{2.4}$$

where T_n are the points of N_t. We can then see that H_t has itself a compensator $\widetilde{H}_t$, in the sense that

$$E(H_t - H_s \mid H_u, u \leq s) = E(\widetilde{H}_t - \widetilde{H}_s \mid H_u, u \leq s), \quad s < t. \tag{2.5}$$

Fact 2.15. *The process H_t defined in* (2.4) *has a compensator given by*

$$\widetilde{H}_t = \int_0^t G_u d\widetilde{N}_u,$$

i.e. (2.5) *holds. (Exercise.)*

Example 2.6. People die according to a Poisson process of rate λ. When someone dies, my company has to pay b pounds to the diseased's family. There is also inflation rate α. How much, on the average, money will I pay up to time t?
Answer: The current value of the money I must pay by time t is

$$H_t = \int_0^t be^{-\alpha s} dN_s.$$

To compute EH_t we remember that $EH_t = E\widehat{H}_t$. But

$$\widehat{H}_t = \int_0^t be^{-\alpha s}\lambda ds = b\lambda \frac{1 - e^{-\alpha t}}{\alpha}.$$

This is deterministic, so $E\widehat{H}_t = \widehat{H}_t = b\lambda\frac{1-e^{-\alpha t}}{\alpha}$.

Example 2.7. Compute the variance of N_t (a Poisson process with rate λ), without using the fact that it is Poisson distributed; rather, use a compensator. Answer: We will compute

EN_t^2. Let us write $\Delta N_s = N_s - N_{s-}$ and, using the algebraic identity $(\sum_i a_i)^2 = \sum_i a_i^2 + 2\sum_{i<j} a_i a_j$,

$$N_t^2 = \left(\sum_{s\le t} \Delta N_s\right)^2 = \sum_{s\le t} (\Delta N_s)^2 + 2\sum_{s\le t}\sum_{u<s} \Delta N_u \Delta N_s.$$

But $(\Delta N_s)^2 = \Delta N_s$, because it is 0 or 1. Hence the first term is N_t. Now check that the second term equals $2\int_0^t N_{s-} dN_s$. Thus,

$$N_t^2 = N_t + 2\int_0^t N_{s-} dN_s.$$

The compensator of this is the compensator of N_t (which is λt) plus the compensator of $2\int_0^t N_{s-} dN_s$ (which is $2\int_0^t N_{s-}\lambda ds$):

$$\widehat{N_t^2} = \lambda t + 2\int_0^t N_{s-}\lambda ds.$$

Hence

$$EN_t^2 = \lambda t + 2\int_0^t \lambda s \lambda ds = \lambda t + \lambda^2 t^2.$$

So $\mathrm{var}(N_t) = EN_t^2 - (EN_t)^2 = \lambda t$, as expected, because, after all, N_t is a Poisson random variable with mean λt.

Chapter 3

Lifetime estimation

We first look at nonparametric estimators of a distribution function. Then we have to deal with the fact that, in real life, censored (i.e. incomplete) observations are not discarded for they may provide valuable information. We learn how to construct a *nonparametric estimator*, known as the Nelson estimator, for the cumulative force of mortality and then how to "translate" this into a nonparametric estimator for the survival function; this is the Kaplan-Meier estimator. When observations can be classified into many types, we may use a semi-parametric model introduced by Cox (the proportional hazards model) or its more modern version due to Aalen (the multiplicative intensities model). To understand the estimators we shall use the language, tools and intuition built around the concept of force of mortality and its dynamic version, i.e. the concept of the compensator.

3.1 Introduction

We shall consider the problem of estimation of the distribution of a lifetime variable from real data. As discussed in Section 1.1, data may be incomplete and/or censored.

3.2 Empirical distribution function

The technique of estimating a lifetime based on the method of empirical distribution function (EDF) is one that is typically used in cohort studies. An example of a cohort study is in the estimation of lifetime of a smoker and its comparison vs that of a nonsmoker. The estimation is done separately for the two groups and, in the end, the results are compared. The problem is that such a study may take years to complete.

Suppose that $X_1, \ldots, X_n$ are random variables. Then the EDF is the random function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x), \quad x \in \mathbb{R}.$$

In words: $F_n(x)$ is the fraction of the variables with values not larger than x. Note that the function $x \mapsto F_n(x)$ is a distribution function because it is increasing (right-continuous) and

$F_n(-\infty) = 0$, $F_n(+\infty) = 1$. However, it is random because it is a function of the random variables $X_1, \ldots, X_n$.

Typically, we must assume that the X_i come from the same population. Here we will assume that they are i.i.d. with common (real) distribution $F(x)$. That is, $P(X_i \leq x) = F(x)$. This function is assumed to be unknown.

Note that, due to the *Fundamental Theorem of Statistics* (look at Section B.8), the random function F_n is a consistent estimator of the FUNCTION F. This means that we can uniformly approximate the function F with arbitrarily small probability of error as long as n is large enough (and this is a BIG problem, because n may not be big).

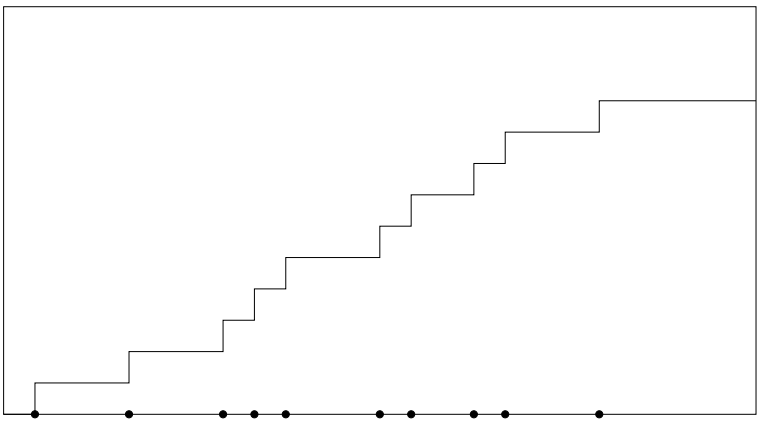


Figure 3.1: *An empirical distribution function for* $n = 10$ *observations. Note that the function jumps at each observation* X_i *and the jump size is* $1/10$

You will recall, from the Fundamental Theorem of Statistics (the Glivenko-Cantelli result of Section B.8), that

$$D_n := \sup_x |F_n(x) - F(x)|$$

converges to 0, as $n \to \infty$, with probability 1.

The *Kolmogorov-Smirnov* test is a *nonparametric test* of the hypothesis that the true distribution is F. Following the Neyman-Pearson formulation, we choose a significance level α (say $\alpha = 5\%$) and try to find a rejection region R (which is a subset of the set of values of our observations), such that the type II error probability is minimised while the type I error probability remains below α. The type I error probability is the probability of rejecting the hypothesis when the true distribution is actually F, while the type II error probability is the probability of accepting the hypothesis when the true distribution is not F. According to the Kolmogorov-Smirnov result, we should try to look for a rejection region of the form

$$R = \{D_n > k\},$$

and try to find k so that

$$P(R) = \alpha.$$

This $P(R)$ is computed under the hypothesis that F is the true distribution. The result also says (which may be surprising at first look!) that $P(R)$ is the same no matter what F is!

So we choose F to be uniform, understand enough probability to compute

$$\alpha = P(\sup_x |F_n(x) - F(x)| > k)$$

(this is done in many books–but also see Section B.3), then tabulate α vs k (or, what is the same, k vs α)* and just compute D_n for the given data. It's beautiful and simple.

3.3 The Nelson estimator of the (cumulative) FOM for censored data

The problem is, again, to estimate an unknown distribution function F but with one extra complication: right-censoring. We will assume that F is the distribution of a positive and absolutely continuous random variable. Let T be such a random variable representing, say, the life of a patient. In this section, we will construct an estimator for the integrated FOM

$$\Lambda(t) := \int_0^t \mu_s ds.$$

Let us think of an experiment involving n such patients with lifetimes $T_1, T_2, \ldots, T_n$, assumed to be i.i.d. random variables with common distribution F.

If all n observations were available, then the problem reduces to that of the previous section. But, in real life, some patients may decide to withdraw from the experiment, before they die. This is censoring.

We model censoring as follows: Let T be the unknown lifetime of a patient and let U be a withdrawal time. We will assume that U is random, independent of T. What we observe, in reality, is the random variable

$$X = \min(T, U).$$

Define

$$\delta = \mathbf{1}(T \leq U).$$

Thus $\delta = 0$ corresponds to censoring.

Now, for the n subjects, define, along with the T_i, the random variables U_i, assumed to be independent and independent of the T_i's. Let $X_i = \min(T_i, U_i)$ and $\delta_i = \mathbf{1}(T_i \leq U_i)$. The X_i's constitute the observations

Consider the point process

$$N_t^i = \mathbf{1}(X_i \leq t, \delta_i = 1) = \mathbf{1}(T_i \leq t, T_i \leq U_i), \quad i = 1, \ldots, n.$$

Each of these point processes is observable, and so is their sum

$$N_t = \sum_{i=1}^n N_t^i.$$

The situation is as described in Proposition 2.14. It was shown there that the compensator of N_t is given by

$$\widetilde{N}_t = \int_0^t Y_u \mu_u du,$$

where Y_t is the number-at-risk process

$$Y_t := \sum_{i=1}^n \mathbf{1}(X_i \geq t).$$

*There ARE tables doing this! But also see Section B.8.1

We now apply Proposition 2.15 with

$$H_t := \int_0^t \mathbf{1}(Y_s > 0)\frac{dN_s}{Y_s} = \sum_{i=1}^{n} \frac{1}{Y_{T_i}}\mathbf{1}(T_i \le t, T_i \le U_i). \tag{3.1}$$

By definition, Y_t is left-continuous, so Proposition 2.15 tells us that H_t has the following compensator:

$$\widetilde{H}_t = \int_0^t \mathbf{1}(Y_s > 0)\frac{d\widetilde{N}_s}{Y_s} = \int_0^t \mathbf{1}(Y_s > 0)\frac{Y_s\mu_s}{Y_s}ds = \int_0^t \mathbf{1}(Y_s > 0)\mu_s ds = \int_0^{t\wedge\tau} \mu_s ds = \Lambda(t\wedge\tau),$$

where τ is the first time t such that $Y_t = 0$. Since $EH_t = E\widetilde{H}_t$, it is reasonable to take H_t as an estimator of $\Lambda(t \wedge \tau)$.

Definition 3.1 (The Nelson estimator)**.** The H_t, defined in (3.1) is the Nelson estimator for the cumulative hazard rate.

The Nelson estimator was introduced in [13]. The time τ represents the earliest time at which all subjects are either dead or have withdrawn; in other words, the number-at-risk is zero after τ. So, obviously, nothing after τ is of relevance. We have also shown that

Proposition 3.1 (Unbiasedness of the Nelson estimator)**.** *Under the true probability measure (i.e. under the distribution F)*

$$EH_t = E\Lambda(t \wedge \tau), \quad \textit{for all } t \ge 0.$$

Out of the times $T_1, \ldots, T_n$, some of them correspond to actual observed deaths ($\delta_i = 1$ or $T_i \le U_i$) and some to censoring ($\delta_i = 0$ or $T_i > U_i$). The point process N_t jumps at the epochs of the actual observed deaths only. Let as then take those times, i.e. those T_i for which $\delta_i = 1$ and let T_1^o be the smallest, T_2^0 be the next one, etc. In this notation, the Nelson estimator (3.1) takes the form

$$H_t = \sum_{k=1}^{N_t} \frac{1}{Y_{T_k^o}}.$$

The 1 in the numerator can be interpreted as the number of deaths occurring and observed at time exactly T_k^o. So, if we let $Y_{T_k^o+}$ be the value of the number-at-risk, just after T_k^O, we have $Y_{T_k^o} - Y_{T_k^o+} = 1$ and so, trivially,

$$H_t = \sum_{k=1}^{N_t} \frac{Y_{T_k^o} - Y_{T_k^o+}}{Y_{T_k^o}}. \tag{3.2}$$

In theory, only one death occurs at each time T_k^o. In practice, because measurements cannot be taken continuously, but every day, or even every week, there may be more than one deaths recorded. That is, in practice, it is (3.2) that is used, rather than (3.1), and the interpretation of the numerator $D_k := Y_{T_k^o} - Y_{T_k^o+}$ is that it represents the number of observed deaths at T_k^o.

But if we look at (3.2) once more, then its form starts making sense. For is we know that $Y_{T_k^o}$ subjects were alive just before T_k^o and D_k died at T_k^o then the FOM at this time can be estimated to be equal to $D_k/Y_{T_k^o}$. Hence the cumulative FOM is the sum of these ratios, exactly as the Nelson estimator prescribes.

Example 3.1. In a clinical experiment, 10 patients were monitored from the time they took a drug until they recovered. The table below shows the time-to-recovery for each patient. Four patients withdrew from the study and so only the times till their withdrawal were recorded (censoring).

patient index	1	2	3	4	5	6	7	8	9	10
observations	5	16	12	9	8	2	6	10	20	14
censoring?	N	Y	N	N	N	N	Y	Y	N	N

The first order of business is to order the observation times, regardless of whether they are censored or not, but we clearly indicated the censored ones by a little $^{\mathsf{c}}$ next to them:

ordered observations	2	5	6^{c}	8	9	10^{c}	12	14^{c}	16^{c}	20

Figure 3.2 shows the observations counting process. Figure 3.4 shows the number-at-risk process and, below it, the Nelson estimator. Note that the sizes of the jumps of the Nelson estimator are equal to the inverse of the number-at-risk process just before the jump time.

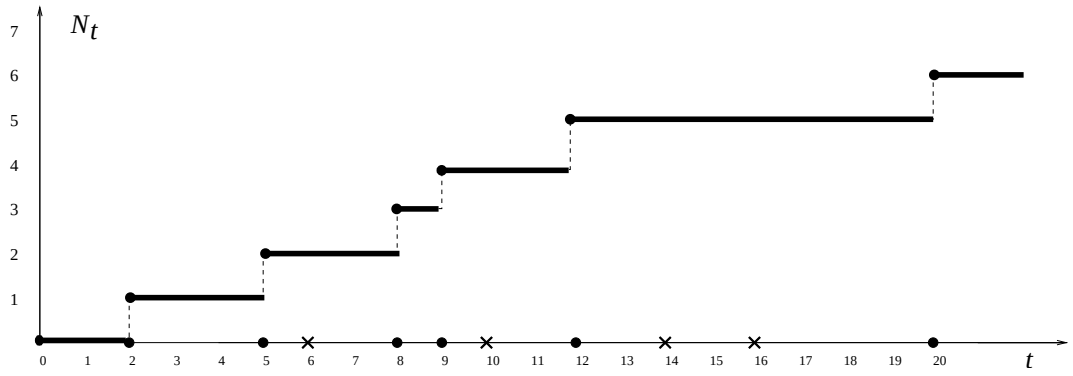


Figure 3.2: *The observations counting process for Example 3.1*

3.4 The Kaplan-Meier estimator for the survival function under censoring

How can we transform the Nelson estimator for $\Lambda(t) = \int_0^t \mu_s ds$ into an estimator for $S(t)$? The answer lies in the relation between $S(t)$ and $\Lambda(t)$. We have:

$$S(t) = 1 - \int_0^t f(s)ds = 1 - \int_0^t S(s)\frac{f(s)}{S(s)}ds = 1 - \int_0^t S(s)\mu_s ds. = 1 - \int_0^t S(s)d\Lambda(s).$$

We can read this as

$$-dS(t) = S(t)d\Lambda(t). \tag{3.3}$$

Hence, by substituting the estimator H_t for $\Lambda(t)$, we hope to obtain an estimator $\widehat{S}(t)$ for $S(t)$. We think of (3.3) as saying that the change in $S(t)$ at time t is equal to its value before the change times the change in $\Lambda(t)$. Now translate this into a statement for an estimator: the change in $\widehat{S}(t)$ at time t is equal to its value before the change times the change in H_t.

In other words, if we agree that $\widehat{S}(t)$ is right-continuous and denote by $\widehat{S}(t-)$ its value just before t, we have

$$\widehat{S}(t-) - \widehat{S}(t) = \widehat{S}(t-)[H_t - H_{t-}].$$

But the change $H_t - H_{t-}$ is either 0 (if t is not a real observed death) or $D_k/Y_{T_k^o}$ if $t = T_k^o$. We write this as

$$\widehat{S}(t-) - \widehat{S}(t) = \widehat{S}(t-)\frac{\Delta N_t}{Y_t},$$

with ΔN_t denoting the change in the observations process (equal to D_k if $t = T_k^o$) and solve for $\widehat{S}(t)$ to obtain

$$\widehat{S}(t) = \widehat{S}(t-)\left[1 - \frac{\Delta N_t}{Y_t}\right].$$

Thus, $\widehat{S}(t)$ is found by *multiplying* its previous value by the factor inside the bracket. Iterating the procedure we obtain

$$\widehat{S}(t) = \prod_{s \le t}\left[1 - \frac{\Delta N_s}{Y_s}\right]. \tag{3.4}$$

Definition 3.2 (Kaplan-Meier estimator for the survival function)**.** This is the estimator given by (3.4).

The Kaplan-Meier estimator was introduced in [10]. It is a biased estimator, but not very much so:

Theorem 3.1. *For all $t \ge 0$, it holds that*

$$0 \le E[\widehat{S}(t) - S(t)] \le F(t)P(X \le t)^n.$$

See Fleming and Harrington [8] for a proof.

Thus the bias $E[\widehat{S}(t) - S(t)]$ is always non-negative and tends to 0 exponentially fast as the number n tends to infinity.

Example 3.2 (continuation of Example 3.1)**.** See the last part of Figure 3.4.

3.5 The Cox and Aalen models

It is often the case in practice that subjects are classified according to their types. For instance, we may wish to study mortality rates of males and females, but it is unreasonable to assume the same distribution for both. We introduce a model, which is neither fully nonparametric nor fully parametric. It is usually referred to as semi-parametric. In it, we assume that each observation is censored, so if T represents the actual lifetime and U a censoring time, we observe

$$X = T \wedge U, \quad \delta = \mathbf{1}(T \le U)$$

but, along with X, we observe a vector of "covariates"

$$Z = (Z_1, \ldots, Z_p).$$

The number p is the number of types and each Z_i is a real number.[†]

[†]More generally, each Z_i can be a random process, i.e. it can vary with time–see below.

For instance, we may want to classify subjects according to SEX $= \{0, 1\}$ (where 0=male, 1=female) and AGE $= \{0, 1, 2\}$ (where 0=young, 1=middle-aged, 2=old). So, here, $p = 2$, and Z_1 takes values in $\{0, 1\}$, while Z_2 takes values in $\{0, 1, 2\}$. The vector $Z = (Z_1, Z_2)$ takes 6 values. A fully nonparametric approach would require to estimate $\lambda(t \mid Z)$ for each of the possible values of Z, but that may be costly.

We assume[‡] the random variable T has a FOM $\lambda(t \mid Z)$ that depends on Z in a multiplicative manner:

$$\lambda(t \mid Z) = \lambda_0(t)g(Z)$$

Usually,

$$g(Z) = \exp \sum_{i=1}^{p} \beta_p Z_p =: \beta' Z,$$

the β_i being unknown parameters, $\beta = (\beta_1, \ldots, \beta_p)$, $Z = (Z_1, \ldots, Z_p)$ and $\beta' Z$ denotes the standard inner product between β and Z. This is the so-called *proportional hazards model*, due to Cox [5].

Note that

$$\lambda(t \mid Z) = \lim_{h \to 0} h^{-1} P(T \leq t + h \mid T > t, Z),$$

knowledge of $\lambda(t \mid Z)$ implies knowledge of the survival function $S(t \mid Z) = P(T > t \mid Z)$, because

$$S(t \mid Z) = \exp - \int_0^t \lambda(u \mid Z) du.$$

Cox also proposed a model for time-varying covariates:

$$\lambda(t \mid Z_t) = P(T \leq t + h \mid T > t, Z_t).$$

The problem with this is that $S(t \mid Z)$ cannot be recovered.

Aalen (1978) had a better suggestion. He argued as follows. Observe that the pair of random variables (X, U) convey precisely the same information as the pair of processes

$$N_t = \mathbf{1}(X \leq t, \delta = 1), \quad Y_t = \mathbf{1}(X \geq t), \quad t \geq 0.$$

Indeed, by observing the evolution of Y_t we can recover X as the first t such that $Y_{t+} = 0$, and then we can recover δ by looking at the value of N_X. If $N_X = 1$ then $\delta = 1$, otherwise $\delta = 0$. Now the compensator of N_t is, as seen before (Proposition 2.14),

$$\widetilde{N}_t = \int_0^t Y_u \mu_u du,$$

where μ_t is the FOM (hazard rate) of T. E.g. in the proportional hazards model,

$$\widetilde{N}_t = \int_0^t Y_u \lambda_0(t) e^{\beta' Z} du.$$

In the time-varying case, this enables us to propose a model as follows:

$$\widetilde{N}_t = \int_0^t Y_u \lambda_0(t) e^{\beta' Z_u} du.$$

This is the *multiplicative intensities* model of Aalen.

Note that, for constant Z, the Cox proportional hazards model coincides with the *Aalen multiplicative intensities model*.

[‡]This IS an assumption that can be refuted, if necessary.

3.5.1 Estimation based on partial likelihood

The problem is to estimate the coefficients $\beta_1, \ldots, \beta_p$ in the proportional hazards model. We assume that there are n i.i.d. random variables $T_1, \ldots, T_n$, representing lifetimes with common, unknown, distribution F. These are independently censored by $U_1, \ldots, U_n$, so that $X_i = T_i \wedge U_i$, $i = 1, \ldots, n$ are the actual observation times. Along with these, we observe the covariates $Z_i = (Z_i^1, \ldots, Z_i^n)$, $i = 1, \ldots, n$. Let $\delta_i = \mathbf{1}(T_i \leq U_i)$.

Example 3.3. Consider the following 15 observations.

index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
observations	5	16	12	9	8	2	6	10	20	14	7	1	18	3	11
sex	M	M	M	F	F	M	M	M	F	F	F	M	M	M	F
censoring?	N	Y	N	N	N	N	Y	Y	N	N	Y	N	N	Y	Y

Thus, e.g., we know that subject $i = 5$ is female (F), that it was not censored, and that we observed $T_5 = 8$. On the other hand, subject $i = 8$ is male (M), there was censoring, so we observed $X_8 = 10$; in other words, we just know that $T_8 = 10$, but not the actual value of T_8. To differentiate between male and female subjects we *arbitrarily* choose $Z = 0$ for male and $Z = 1$ for female, so our proportional hazards model consists of two hazard rates:

$$\lambda(t \mid 0) = \lambda_0(t),$$
$$\lambda(t \mid 1) = \lambda_0(t)e^{\beta}.$$

A fully nonparametric model would require that we estimate two functions $\lambda(t \mid 0)$, $\lambda(t \mid 1)$, without specifying any a priori relationship between them. Thus, we would have to treat the male and female populations separately, and apply the Nelson estimator for each of them. In a real situation this may be "costly" and so we use the model above where we a priory specify that the two unknown functions have a constant (time-independent) but unknown ratio e^{β}. The problem is to estimate this ratio. From the observations we see that just before $t = 18$, there are two subjects alive: subject 13 (M) and subject 9 (F). Recall (Exercise A.2.6) that if we know that the minimum two independent random variables is t, then the chance that the first one is the actual minimum is the ratio of its FOM divided by the sum of the FOMs. Thus, since at time $t = 18$ there is an actual death, the chance that it was actually due to the male subject is

$$\frac{\lambda(t \mid 0)}{\lambda(t \mid 0) + \lambda(t \mid 1)} = \frac{1}{1 + e^{\beta}}.$$

The same logic can be applied at any actual death time, giving an estimate for the probability that the actual death is due to a subject carrying the particular label. Of course, the computation has to be done amongst the subjects that have not expired (died or censored) at that time. Multiplying these terms together gives a quantity that is known as partial likelihood and is only a function of β.

We continue by giving the heuristics for the partial likelihood function. Suppose that (A_i, B_i), $i = 1, \ldots, m$ is a collection of pairs of events. Then the likelihood (probability) of

all of them is

$$\begin{aligned}
&P(A_1B_1A_2B_2\cdots A_mB_m) = \\
&\quad = P(A_1B_1)P(A_2B_2 \mid A_1B_1)P(A_3B_3 \mid A_2B_2A_1B_1)\cdots P(A_mB_m \mid A_{m-1}B_{m-1}\cdots A_1B_1) \\
&\quad = P(A_1 \mid B_1)P(A_2 \mid B_2A_1B_1)P(A_3 \mid B_3A_1B_1A_2B_2)\cdots P(A_m \mid B_{m-1}\cdots A_1B_1) \\
&\qquad \times P(B_1)P(B_2 \mid A_1B_1)P(B_3 \mid A_2B_2A_1B_1)\cdots P(B_m \mid A_{m-1}B_{m-1}\cdots A_1B_1).
\end{aligned}$$

Suppose that the events A_i carry more information about an unknown parameter than the events B_i. It is "reasonable" to ignore that latter terms and use only

$$L := P(A_1 \mid B_1)P(A_2 \mid B_2A_1B_1)P(A_3 \mid B_3A_1B_1A_2B_2)\cdots P(A_m \mid B_{m-1}\cdots A_1B_1)$$

for estimation.

In the case of the proportional hazards, consider the actual deaths, i.e. the points at which N_t jumps. Let $T_0^o = 0$, and, recursively,

$$T_k^o = \inf\{t > T_{k-1}^o : N_t = N_{t-} + 1\}.$$

Each T_k^o corresponds to exactly one T_i. We use the notation (i) for this particular i for which $T_i = T_k^o$. For instance, in Example 3.3 we have that the actual deaths are

$$T_{12} = 1, \quad T_6 = 2, \quad T_1 = 5, \quad T_5 = 8, \quad \ldots$$

So we set

$$T_1^o = T_{12}, \quad T_2^o = T_6, \quad T_3^o = T_1, \quad T_4^o = T_5, \ldots$$

and so we have that the indices (=labels/names of subjects that died) of the ordered actual deaths are

$$(1) = 12, \quad (2) = 6, \quad (3) = 1, \quad (4) = 5, \ldots$$

Define the set-at-risk at time t as

$$R(t) := \{i : \ X_i \geq t\}.$$

Let A_k be the event that the label of the subject that dies at time T_k^o is (k). Let B_k be the event that specifies the observations (necessarily censorings) between T_{k-1}^o and T_k^o. Then the likelihood corresponding to the observed data is $P(A_1B_1A_2B_2\cdots)$. We make the *"reasonable" assumption* that the B_k carry little information about the labels and so the partial likelihood is the product of terms of the form

$$P(A_k \mid B_kA_{k-1}\cdots A_1B_1).$$

From our earlier discussion it should be clear that

$$P(A_k \mid B_kA_{k-1}\cdots A_1B_1) = \frac{\lambda(T_k^o \mid Z_{(k)})}{\sum_{i\in R(T_k^o)} \lambda(T_k^o \mid Z_i)}$$

Hence the partial likelihood is

$$L(\beta) = \prod_{k\geq 1} \frac{\lambda(T_k^o \mid Z_{(k)})}{\sum_{i\in R(T_k^o)} \lambda(T_k^o \mid Z_i)} \tag{3.5}$$

Figure 3.3: *Acheron*

Fact 3.1. *The jolly fellow of Figure 3.3 has something to do with lifetime estimation or with mortality studies. (Exercise: What?)*

We treat the partial likelihood as being a usual density and, applying the fundamentals of non-parametric Statistics, we define the MLE of β as the statistic $\widehat{\beta} = \widehat{\beta}(T_1^o, T_2^o, \ldots; Z_1, \ldots, Z_n)$ that maximised $L(\beta)$:

$$L(\widehat{\beta}) = \max_{\beta} L(\beta).$$

From the general theory of MLE, we expect that $\widehat{\beta}$

1. is strongly *consistent* , i.e. that it converges to the true β as $n \to \infty$,
2. is approximately *unbiased* , i.e. that its expectation under the true β is β,
3. is very *efficient* , i.e. that its variance, under the true β is as small as possible (approximately equal to the lower bound of the Cramér-Rao inequality : $\mathrm{var}_\beta(\widehat{\beta}) \approx 1/I_\beta$, where $I_\beta \approx E_\beta[D_\beta L(\beta)]^2 \approx -E_\beta D_\beta^2 L(\beta))$, when n is large,
4. is *asymptotically normal* , i.e. that the law of $\widehat{\beta}$ is, for large n, approximately $\mathcal{N}(0, 1/I_\beta)$.

3.5.2 Hypothesis testing with partial likelihood

We treat the logarithm of the partial likelihood as being a usual log-likelihood function and apply the standard hypotheses testing methods from Statistics.

z test:

To test the hypothesis that $\beta = \beta^*$ vs its negation, we may use the z-test (see Section B.6). If we believe the hypothesis, then

$$\sqrt{I_{\beta^*}}(\widehat{\beta} - \beta^*)$$

"should" be approximately $\mathcal{N}(0,1)$–distributed. We compute $\sqrt{I_{\beta^*}}(\widehat{\beta} - \beta^*)$ and if it is not typical, we reject the hypothesis.

Score test:

Or we may use the score test (again, see Section B.6). To this end, first compute the score function

$$U(\beta) = \frac{\partial}{\partial \beta} \log L(\beta),$$

where $L(\beta)$ is given by (3.5). Then, if we believe the hypothesis, the quantity

$$U(\beta^*)^2 / I_{\beta^*}$$

"should" be approximately χ^2_1–distributed. But we do not know I_{β^*} because its computation requires taking an expectation under β^*. So we replace it by the observed quantity

$$I(\beta) = \frac{\partial^2}{\partial \beta^2} \log L(\beta),$$

evaluated at $\beta = \beta^*$. Then, believing that the (now computable) quantity

$$U(\beta^*)^2 / I(\beta^*)$$

is approximately χ^2_1–distributed, we compute it and if its value is atypical for the χ^2_1 distribution, we reject the hypothesis.

Likelihood ratio test:

Finally, we may consider the likelihood ratio (see Section B.7)

$$\lambda := \frac{\sup_{\beta \neq \beta^*} L(\beta)}{L(\beta^*)} = \frac{L(\widehat{\beta})}{L(\beta^*)}$$

and use the result that, if β^* is the true value, then $2 \log \lambda$ "should" have a χ^2_1 distribution.

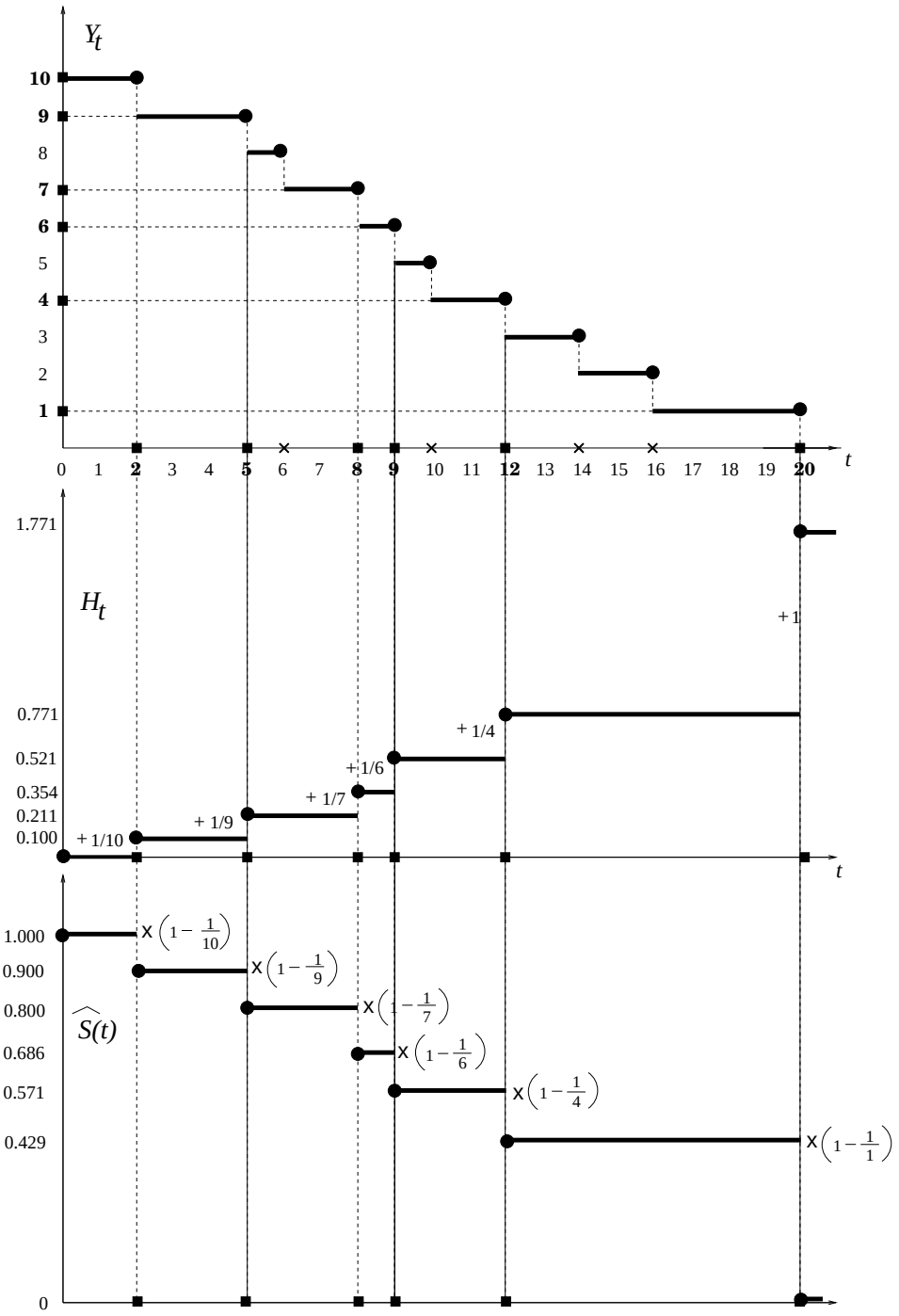


Figure 3.4: *The number-at-risk process Y_t, the Nelson estimator H_t, and the Kaplan-Meier estimator $\widehat{S}(t)$ for Example 3.1*

Chapter 4

Markov chains and estimation

Nonparametrics is desirable, but, in the absence of a lot of data, we have to make do with some more concrete models. Also, a priori experience may render specific models more accurate descriptors of the real situation. The point of the first part is to quickly remind you of the concept of a Markov chain, in continuous time, with discrete (mostly finite) state space, the caveat being that we will allow for time-varying rates (=forces, in actuarial lingo). Of course, tractable models are the good-old time-homogeneous Markov chains, which are reviewed next. We also discuss how to estimate the parameters of a Markovian model, based on long observations from one run of the chain, or from independent trials of the same chain. We only discuss the case of time-homogeneous chains, the more general case being the same, in principle, but analytically more complicated. We shall only deal with Maximum Likelihood Estimators (MLE) whose properties we are also going to review.

4.1 General concepts about (time-varying) chains

Recall that a stochastic process $\{X_t, t \geq 0\}$ with values in some state space S is Markov iff for any t, $(X_s, s > t)$ is independent of $(X_s, s < t)$, conditionally on X_t. This means that joint probabilities can be compute by the chain rule:

$$P(X_{t_i} = x_i, i = 1, 2, \ldots k \mid X_0 = x_0) = \prod_{i=1}^{k} P(X_{t_i} = x_i \mid X_{t_{i-1}} = x_{i-1}),$$

for any $0 = t_0 < t_1 < \cdots t_k$ and any $x_0, x_1, \ldots, x_k \in S$. It follows that the distribution of the process is completely specified by the distribution of the initial random variable X_0 and the conditional distributions

$$P_{s,t}(x, y) = P(X_t = y \mid X_s = x), \quad 0 \leq s < t, \quad x, y \in S.$$

These *transition probabilities* are functions of four variables: two times (s, t) and two states (x, y). It is customary, but also mathematically fruitful, to think of the dependence on the pair of states as a matrix, and thus define the *probability transition matrix* $P_{s,t}$ as a matrix

with entries $P_{s,t}(x,y)$, where x indexes the rows and y the columns. The actuarial notation is

$$ {}_{t-s}p_s^{xy} = P_{s,t}(x,y). $$

Note that the sum of the rows of this matrix is 1. Also note that, owing to the Markov property defined above, the probability transition matrices obey the multiplication rule:

$$ P_{s,t} = P_{s,u}P_{u,t}, \text{ whenever } s < u < t. $$

In Mathematics, one says that the family of matrices $\{P_{s,t}\}$ when $s < t$ run over the positive real numbers form a semigroup which is an algebraic structure with an identity element. The identity here is obviously $P_{t,t}$, for an t. Indeed, $P_{t,t}$ is trivially the identity matrix. Recalling what matrix product means, the multiplication rule reads:

$$ P_{s,u}(x,y) = \sum_z P_{s,u}(x,z)P_{u,t}(z,y). \tag{4.1} $$

We are only going to discuss the case where the cardinality of S is finite. In this case, the function $t \mapsto X_t$ is piecewise constant. We consider only the case where the derivative of $P(X_{t+h} = y \mid X_t = x)$ with respect to h exists. The value of the derivative at $h = 0$ is of particular importance and is known as *transition rate* .

Definition 4.1. The transition rate from state x to a different state y at time t is defined as

$$ q_t(x,y) := \lim_{h \downarrow 0} \frac{1}{h} P(X_{t+h} = y \mid X_t = x). $$

NOTE: The definition itself needs a theorem; namely, that these derivatives exist. But we shall take it for granted that you have learnt this in your undergraduate courses. In actuarial notation:

$$ q_t(x,y) = \mu_t^{xy}. $$

We shall think of the transition rates as a priori "given", as part of the model, and shall derive the transition probabilities from them. Note that we need only know $q_t(x,y)$ for $x \neq y$.

Fact 4.1. *If $x = y$, the derivative of $P(X_{t+h} = x \mid X_t = x)$ equals*

$$ q_t(x,x) := \lim_{h \downarrow 0} \frac{1}{h}[P(X_{t+h} = x \mid X_t = x) - 1] = -\sum_{x \neq y} q_t(x,y). $$

(Exercise.)

Note that $q_t(x,x)$ is a negative number. Again, we find it fruitful to summarise the transition rates information in the matrix Q_t, which is a matrix with entries $q_t(x,y)$. *Note that the sum of the rows of this matrix is zero.*

It is useful to always remember the definitions of the rates by means of the *Landau symbols*:

$$ P(X_{t+h} = y \mid X_t = x) = \begin{cases} q_t(x,y)h + o(h), & \text{if } x \neq y \\ 1 - |q_t(x,x)|h + o(h), & \text{if } x = y \end{cases}, \text{ as } h \to 0. \tag{4.2} $$

Here, $o(h)$ is any function with the property $\lim_{h\to 0} o(h)/h = 0$. Similarly, $o(g(h))$ is any function with the property $\lim_{h\to 0} o(g(h))/g(h) = 0$. Note that a finite sum of $o(h)$ functions is also $o(h)$, whereas $ho(h) = o(h^2)$, and if a function is $o(h^2)$ then it is $o(h)$. We can now translate these into matrix notation. Denote, as usual, by I the identity matrix, i.e. a matrix whose entries in row x and column y is $\mathbf{1}(x = y)$.

Fact 4.2 (Exercise in notation)**.** *Use this notation to see that the definition of transition rates is equivalent to*

$$P_{t,t+h} - I = hQ_t + o(h), \qquad \text{as } h \to 0.$$

Note that, here, $o(h)$ is a matrix with all entries of type $o(h)$.

The matrices $P_{s,t}$ and Q_t are related via

Theorem 4.1.

$$\textit{Kolmogorov's forward equations:} \quad \frac{\partial}{\partial t} P_{s,t} = P_{s,t} Q_t$$
$$\textit{Kolmogorov's backward equations}: \quad \frac{\partial}{\partial s} P_{s,t} = Q_s P_{s,t}.$$

Proof. The adjective forward reminds us to perturb the forward index, i.e. the index t. We have

$$P_{s,t+h} = P_{s,t} P_{t,t+h} = P_{s,t}(I + hQ_t + o(h)) = P_{s,t} + hP_{s,t}Q_t + o(h), \text{ as } h \to 0,$$

(since linear combinations of $o(h)$ is $o(h)$). But this means that

$$\frac{1}{h}(P_{s,t+h} - P_{s,t}) = Q_t + \frac{1}{h} o(h), \qquad \text{as } h \to 0.$$

But the limit of $\frac{1}{h}(P_{s,t+h} - P_{s,t})$ as $h \to 0$ is, by definition of the derivative, $\frac{\partial}{\partial t} P_{s,t}$, and this proves the forward equations. It doesn't require much intelligence at this point to guess that to derive backward equations we have to perturb the backward index, viz. the index s. □

Fact 4.3 (Exercise in notation)**.** *The forward equations, in actuarial notation, read:*

$$\frac{\partial}{\partial t} {}_t p_s^{xy} = \sum_{z \neq y} {}_t p_s^{xz} \mu_{s+t}^{zy} - {}_t p_s^{xy} \mu_{s+t}^{yz}.$$

(Exercise.)

These equations are not, in general, easy to solve. Here people apply numerical methods. Even the simplest, time-varying differential equations are not solvable by quadratures.

4.2 Use of Markov chain models

Consider what is simplest of all: Namely, take $S = \{1, 0\}$ and let $q_t(1, 0) = \mu_t$, a given function of t. Let $q_t(0, 1) = 0$. Interpret 1 as 'alive' and 0 as 'dead'. Thus, μ_t is nothing else but the FOM of the random variable T that is defined as the the first time the process takes the value 1. Our assumption $q_t(0, 1) = 0$ means that there is no chance for resurrection.

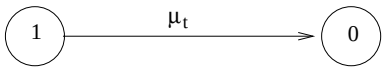


Figure 4.1: *A trivial time-dependent Markov chain*

There is nothing to say here, except that people like to use the diagram of Figure 4.1 for this case.

To make things more interesting and, perhaps, more realistic* introduce a resurrection rate $q_t(0,1) = \lambda_t$ so that the new diagram is as in Figure 4.2 which we may call the punabhava chain. The rate matrix is $Q_t = \begin{pmatrix} -\mu_t & \mu_t \\ \lambda_t & -\lambda_t \end{pmatrix}$ and the forward equations read

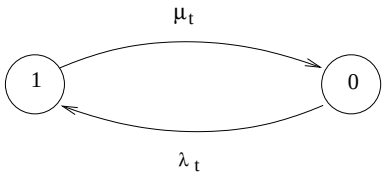


Figure 4.2: *The* PUNABHAVA CHAIN

$$\frac{\partial}{\partial t}\begin{pmatrix} P_{s,t}(1,1) & P_{s,t}(1,0) \\ P_{s,t}(0,1) & P_{s,t}(0,0) \end{pmatrix} = \begin{pmatrix} -\mu_t & \mu_t \\ \lambda_t & -\lambda_t \end{pmatrix}\begin{pmatrix} P_{s,t}(1,1) & P_{s,t}(1,0) \\ P_{s,t}(0,1) & P_{s,t}(0,0) \end{pmatrix}.$$

There appears to be 4 equations, but there are only 2, for the rows of the matrix $P_{s,t}$ add up to 1. A reasonable choice for the functions of the punabhava chain is: $\mu_t = t$ (the later it is the easier it is to die), $\lambda_t = e^{-t}$ (resurrections become rarer and rarer–indeed, there were more of them long time ago and fewer of them today). Denote by X_t the value in $\{1, 0\}$ of the chain at time t. Assume $X_0 = 1$. The following random variables are of interest: the time of the first death $\tau_0 = \inf\{t > 0 : X_t = 0\}$; the time of resurrection measured after the first death: $\tau_1 = \inf\{t > 0 : X_{\tau_0+t} = 1\}$.

Fact 4.4. *Unless μ_t and λ_t are constant, the random variables τ_0, τ_1 are not independent. (Exercise.)*

Now let us assume that $X_s = x$ and define the holding time of the state x by

$$T_s^x := \inf\{t > 0 : X_{s+t} \neq x\}.$$

People are interested in the distribution function of T_s^x:

$${}_tp_s^{\overline{xx}} := P(T_s^x > t \mid X_s = x).$$

Fact 4.5. *The FOM of T_s^x is $\sum_{y\neq x} q_{s+t}(x,y)$, $t \geq 0$, and so (an application of Fact 2.2),*

$${}_tp_s^{\overline{xx}} = \exp - \int_s^{s+t} \sum_{y\neq x} q_u(x,y)du.$$

(Exercise.)

*At least in a Mahayana Buddhist tradition

4.3 The five most trivial chains for survival models

We discuss some very simple models.

Case 4.1 (From LIFE to DEATH without RESURRECTION).
See Figure 4.3. $S = \{\text{ALIVE} = 1, \text{DEAD} = 0\}$. We take $q_t(1,0) = \mu_t$, $q_t(0,1) = 0$. This is nothing else but a model of a single random variable T by means of its FOM μ_t. We have nothing to say, other than that $P(T > t) = \exp - \int_0^t \mu_s ds$.

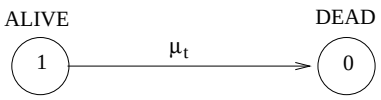


Figure 4.3: *From life to death without resurrection*

Case 4.2 (Oscillating between EMPLOYMENT and UNEMPLOYMENT).
See Figure 4.4. $S = \{\text{EMPLOYED} = 1, \text{UNEMPLOYED} = 0\}$. Rates: $q_t(1,0) = \sigma_t$, $q_t(0,1) = \rho_t$. The model is not solvable, in general.

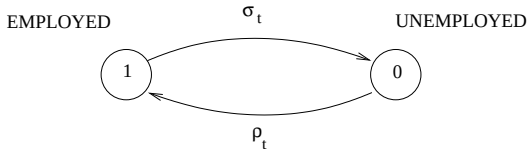


Figure 4.4: *Oscillating between employment and unemployment*

Case 4.3 (From HEALTHY/ILL to DEATH).
See Figure 4.5. $S = \{\text{HEALTHY} = 1, \text{ILL} = 2, \text{DEAD} = 0\}$. we take $q_t(1,2) = \lambda_t$, $q_t(1,0) = \mu_t$, $q_t(2,0) = \nu_t$, and zeros in all other cases. The model is solvable.

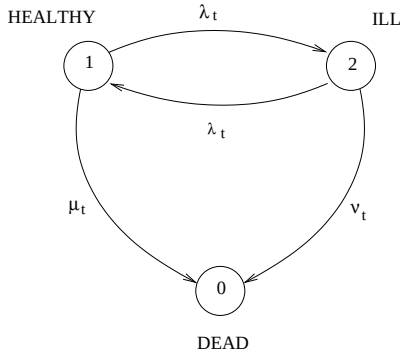


Figure 4.5: *From healthy/ill to death*

Case 4.4 (From LIFE to DEATH for a couple).
See Figure 4.6.
$S = \{\text{BOTH\_ALIVE} = 11, \text{FIRST\_ALIVE\_SECOND DEAD} = 10, \text{FIRST\_DEAD\_SECOND\_ALIVE} = 01, \text{BOTH\_DEAD} = 00\}$. Suppose we take two chains of the first type. Consider first the case where they are independent. Then we have the product chain that can be easily derived if

we let S, T be the two independent lifetimes with FOMs α_t, β_t, respectively:

$$X_t = \begin{cases} 11, & t < S \wedge T \\ 10, & S = S \wedge T \leq t < S \vee T \\ 01, & T = T \wedge S \leq t < S \vee T \\ 00, & t \geq S \vee T. \end{cases}$$

The transition rates are easily found to be: $q_t(11, 10) = q_t(01, 00) = \alpha$, $q_t(11, 01) = q_t(10, 00) = \beta$, and 0 in all other cases. The model is solvable.
The second case is when S, T are not independent. (For instance, consider the case of Romeo and Juliet.) Then we don't have 2 distinct rates, but 4, i.e., $q_t(11, 10) = \alpha_t, q_t(01, 00) = \alpha'_t$, $q_t(11, 01) = \beta_t, q_t(10, 00) = \beta'_t$. The model is solvable. Of course, figuring out the rates from observations may not be a trivial matter.

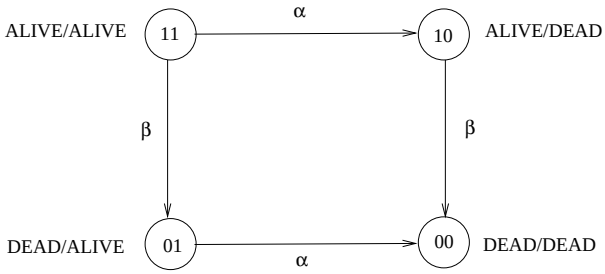


Figure 4.6: *From life to death for a couple*

Case 4.5 (Oscillating between ABLE and ILL till RETIREMENT or DEATH).
See Figure 4.7. $S = \{\text{ABLE} = 1, \text{ILL} = 2, \text{RETIRED} = 3, \text{DEAD} = 0\}$. The only nonzero rates that we consider are: $q_t(1, 2) = \sigma_t, q_t(2, 1) = \rho_t$ (oscillating between ABLE and ILL) and $q_t(1, 3) = \lambda, q_t(1, 0) = \mu_t, q_t(2, 0) = \nu_t$. Wo don't consider the rate $q_t(3, 0)$, simply because we are only interested in the person while he is working or is ill while he is working; when he retires or dies we stop caring. The model is, in general, not solvable. (If the rates σ, ρ are constant then we can solve.)

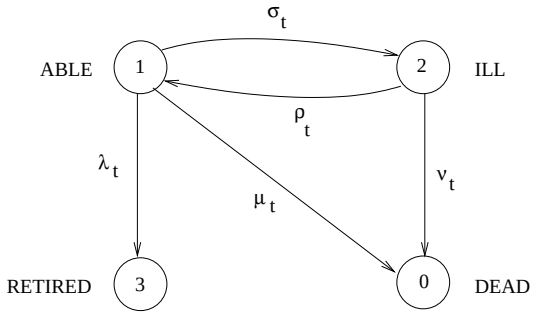


Figure 4.7: *Oscillating between able and ill till retirement or death*

4.4 Homogeneous Markov chains

If we assume that $P_{s,t}$ is a function of $t-s$, then we have the so-called time-homogeneous case. In this case (still assuming S finite) the derivative of $P_{s,t}$ exists, hence we can immediately talk about transition rates. Obviously, the transition rates $q(x,y)$ do not depend on time. And so the rate matrix Q is constant. We are simply going to write P_t for the matrix $P_{s,s+t}$, since it is the same for all s. In this notation, we have

Theorem 4.2.

$$\textit{Kolmogorov's forward equations:} \quad \frac{d}{dt}P_t = P_t Q$$
$$\textit{Kolmogorov's backward equations:} \quad \frac{d}{dt}P_t = Q P_t.$$

These are easily "solvable". Indeed, define, as known from the elementary theory of linear differential equations, the matrix

$$e^{tQ} := \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n,$$

and observe:

Fact 4.6. $\frac{d}{dt}e^{tQ} = Qe^{tQ} = e^{tQ}Q$. *(Exercise.)*

thereby establishing that $P_t = e^{tQ}$ is the solution to the Kolmogorov's equations.

Example 4.1. Consider the punabhava chain with $\mu_t = \mu$, $\lambda_t = \lambda$. Since Q is a 2×2 matrix, any power Q^n with $n \leq 2$ is a combination of I and Q, and so $e^{tQ} = c_1 + c_2 Q$, for appropriate scalars c_1, c_2. Since Q acts on its eigenspaces multiplicatively, this equation becomes scalar one on each of the eigenspaces. The eigenvalues of Q are the numbers s for which $\det(sI - Q) = 0$, i.e. $(s+\mu)(s+\lambda) - \lambda\mu = 0$, or or $s^2 + (\lambda+\mu)s = 0$, whence $s = 0$ or $s = -(\lambda+\mu)$. For each such s we have $e^{ts} = c_1 + c_2 s$. For $s = 0$ we have $c_1 = 1$. For $s = -(\lambda+\mu) =: -\alpha$, we have $e^{-\alpha t} = 1 - c_2\alpha$, whence $c_2 = (1 - e^{-\alpha t})/\alpha$. Assembling things together, we have found that $P_t = e^{tQ} = 1 - Q(1 - e^{-\alpha t})/\alpha = \frac{1}{\alpha}\begin{pmatrix} \frac{\lambda+\mu e^{-\alpha t}}{\lambda+\mu} & \frac{\mu-\mu e^{-\alpha t}}{\lambda+\mu} \\ \frac{\lambda-\lambda e^{-\alpha t}}{\lambda+\mu} & \frac{\mu+\lambda e^{-\alpha t}}{\lambda+\mu} \end{pmatrix}$.

Homogeneous Markov chains (at least in a finite set S) with transition rates $q(x,y)$ can be constructed as follows: First consider a Markov chain $(Y_0, Y_1, \ldots)$ with transition probabilities

$$P(Y_{n+1} = y \mid Y_n = x) = \frac{q(x,y)}{|q(x,x)|}. \tag{4.3}$$

Then, for each $x \in S$, consider a sequence

$$\sigma_x := (\sigma_x(1), \sigma_x(2), \ldots) \tag{4.4}$$

of i.i.d. random variables with common exponential distribution with parameter $|q(x,x)|$. Assume further that the sequences $\{\sigma_x, x \in S\}$ are mutually independent. Define

$$T_0 := 0, \quad T_n := \sigma_{Y_0}(1) + \cdots + \sigma_{Y_{n-1}}(n), \quad n \geq 1,$$
$$X_t := \sum_{n \geq 0} Y_n \mathbf{1}(T_n \leq t < T_{n+1}), \quad t \geq 0. \tag{4.5}$$

Fact 4.7. *The process (X_t) thus constructed is a homogeneous Markov chain with transition rates $q(x, y)$. (Exercise.)*

The backward method We are interested in computing the mean and other distributional aspects of the times

$$\sigma_z := \inf\{t \geq 0 : X_t = z\},$$

or, more generally,

$$\sigma_A := \inf\{t \geq 0 : X_t \in A\}.$$

For, instance, if z is the graveyard state, we are interesting in the first time that the grave is reached. And if z' is the retirement state, and $A = \{z, z'\}$, then σ_A is the first time till retirement or death, whichever comes first.

Notation: $P_x(\cdot), E_x(\cdot)$ mean $P(\cdot \mid X_0 = x), E(\cdot \mid X_0 = x)$, respectively.

Let us find, for instance, the mean of σ_A.

Theorem 4.3 (Backwards equations for mean hitting times)**.**

$$\sum_{y \notin A} q(x, y) E_y \sigma_A = -1, \quad x \notin A.$$

Proof. Let $x \notin A$. Write[†] $E_x\sigma_A = E_x(\sigma_A, \sigma_A \leq \delta) + E_x(\sigma_A, \sigma_A > \delta)$, for any (small) $\delta > 0$. It is easy to see that $E_x(\sigma_A, \sigma_A \leq \delta) = o(\delta)$, as $\delta \downarrow 0$. On the other hand, if $\sigma_A > \delta$ then $\sigma_A = \delta + \sigma'_A$, where $\sigma'_A = \inf\{t \geq 0 : X_{\delta+t} \in A\}$, so

$$E_x(\sigma_A, \sigma_A > \delta) = \sum_y E_x(X_\delta = y, \delta + \sigma'_A) = \sum_y P_\delta(x, y) E_x(\delta + \sigma'_A \mid X_\delta = y).$$

But, by time-homogeneity, $E_x(\sigma'_A \mid X_\delta = y) = E_y\sigma_A$, and so

$$E_x(\sigma_A, \sigma_A > \delta) = \delta + \sum_y P_\delta(x, y) E_y \sigma_A.$$

Since $P_\delta(x, y) = \mathbf{1}(x = y) + \delta q(x, y) + o(\delta)$, we have

$$E_x\sigma_A = o(\delta) + \delta + E_x\sigma_A + \delta \sum_{y \notin A} q(x, y) E_y \sigma_A, \tag{4.6}$$

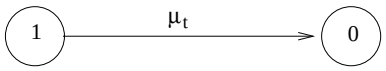
and, upon dividing by δ before sending it to 0, we conclude. □

4.5 Estimation using Markovian models

4.5.1 An utterly trivial case

Let us start with a very simple situation: We observe n i.i.d. lives obeying this model (an assumption, to be sure). Let T_i be the actual death of the i-th life. In other words, $T_1, \ldots, T_n$ are i.i.d. random variables with FOMs μ_t. Suppose that subject i enters the experiment at

[†]The notation $E(X, A)$, where Z is a random variable and A an event, stands for $E(X\mathbf{1}_A)$. Thus, if $A \cap B = \emptyset$, $E(X, A \cup B) = E(X, A) + E(X, B)$.



time a_i, and leaves at time b_i if he has not died by this time. Let us make the simplifying assumption that

$$\mu_t = \mu, \quad a \le t \le b,$$

where $a := \inf_{a \le i \le n} a_i$, $b := \sup_{1 \le i \le n} b_i$. We wish to estimate the one and only uknown parameter of the model, namely, μ.

First of all, we know that $T_i > a_i$. So it is best to consider each of the variables T_i' which, in distribution, equals $T_i - a_i$ given that $T_i > a_i$. What do we observe? We observe the random variables

$$X_i = T_i' \wedge \ell_i$$

$$\delta_i = \mathbf{1}(T_i \le \ell_i),$$

where $\ell_i := b_i - a_i$. Note that $P(T_i' \in dx) = \mu dx$ for $0 < x < \ell_i$, by assumption. Note also that X_i is neither discrete nor continuous; it is a mixture. It is easy to find the joint distribution of (X_i, δ_i). Clearly, X_i has a density on $(0, \ell_i)$ and a discrete mass at ℓ_i. (Recall the notation $\mathfrak{d}_c$ for a mass of size 1 at the point c.) Indeed,

$$P(X_i \in dx, \delta_i = 0) = P(\ell_i \in dx, T_i > \ell_i) = e^{-\mu \ell_i}\, \mathfrak{d}_{\ell_i}(dx) = e^{-\mu x} \mu^0\, \mathfrak{d}_{\ell_i}(dx)$$
$$P(X_i \in dx, \delta_i = 1) = P(T_i \in dx, T_i < \ell_i) = e^{-\mu x} \mu\, dx = e^{-\mu x} \mu^1\, dx.$$

We can compactify this as[‡]

$$P(X_i \in dx, \delta_i = \theta) = e^{-\mu x} \mu^\theta\ [(1-\theta)\, \mathfrak{d}_{\ell_i}(dx) + \theta\, dx], \quad x \ge 0, \quad \theta \in \{0, 1\}.$$

Since we have assumed independence, we can write the joint probability as

$$\begin{aligned} P(X_1 \in dx_1, \delta_1 = \theta_1, \ldots, X_n \in dx_n, \delta_n = \theta_n) &= \prod_{i=1}^n P(X_i \in dx, \delta_i = \theta) \\ &= \prod_{i=1}^n e^{-\mu x_i} \mu^{\theta_i}\ [(1-\theta_i)\, \mathfrak{d}_{\ell_i}(dx_i) + \theta_i\, dx_i] \\ &= e^{-\mu \sum_{i=1}^n x_i} \mu^{\sum_{i=1}^n \theta_i} \prod_{i=1}^n [(1-\theta_i)\, \mathfrak{d}_{\ell_i}(dx_i) + \theta_i\, dx_i]. \end{aligned}$$

Here the variables $x_1, \ldots, x_n$ take values in $[0, \ell_1], \ldots, [0, \ell_n]$, respectively, while the $\theta_1, \ldots, \theta_n$ take values 0 or 1 each. Thus, the likelihood corresponding to the observations $X_1, \ldots, X_n, \delta_1, \ldots, \delta_n$ is

$$L(X, \delta; \mu) = e^{-\mu \sum_{i=1}^n X_i} \mu^{\sum_{i=1}^n \delta_i}.$$

The maximum likelihood estimator $\widehat{\mu}$ is defined by

$$L(X, \delta; \widehat{\mu}) = \max_{\mu \ge 0} L(X, \delta; \mu),$$

and is easily found to be

$$\widehat{\mu} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n X_i}.$$

Note that $\widehat{\mu}$ is a veritable statistic for it is just a function of the observations.

[‡]The variable θ below takes values 1 or 0. It is 1 if $T_i < \ell_i$, i.e. if we observe a death. The choice of the letter θ comes from the greek word $\theta\acute{\alpha}\nu\alpha\tau o\varsigma$ = death.

4.5.2 The general case

At this point, it is useful to recall the construction of homogeneous Markov chains using (4.3) and (4.4) of Section 4.4. When in state x, the chain remains there for an exponentially distributed amount of time (sojourn time) with parameter

$$\mu(x) = |q(x,x)| = \sum_{y \neq x} q(x,y).$$

At the end of this time, the chain jumps to some other state y with probability $q(x,y)/\mu(x)$. An observation consists of a sequence of states $Y_0, Y_1, \ldots$ and the corresponding sojourn times $\sigma_0, \sigma_1, \ldots$. The likelihood of the observation is

$$\begin{aligned} L &= \mu(Y_0) e^{-\mu(Y_0)\sigma_0} \frac{q(Y_0,Y_1)}{\mu(Y_0)} \times \mu(Y_1) e^{-\mu(Y_1)\sigma_1} \frac{q(Y_1,Y_2)}{\mu(Y_1)} \times \cdots \\ &= e^{-\mu(Y_0)\sigma_0} q(Y_0,Y_1) \times e^{-\mu(Y_1)\sigma_1} q(Y_1,Y_2) \times \cdots \\ &= e^{-\sum_x \mu(x) W(x)} \prod_{x \neq y} q(x,y)^{N(x,y)}, \end{aligned} \tag{4.7}$$

where

$$W(x) := \sum_{i \geq 0} \sigma_i \mathbf{1}(Y_i = x), \tag{4.8}$$

$$N(x,y) := \sum_{i \geq 0} \mathbf{1}(Y_i = x, Y_{i+1} = y) \tag{4.9}$$

are the total sojourn time in state x and the total number of jumps from state x to state y, respectively, for the sequence of our observations.

Two things may happen when we observe a Markov chain. Either we stop observing at some predetermined time or the Markov chain reaches an absorbing state and, a fortiori, we must stop the observation. For instance, in the Markov chain of Figure 4.8, there are three absorbing states: a, b, c and two transient ones: $1, 2$. If we start with $Y_0 = 1$, we may be

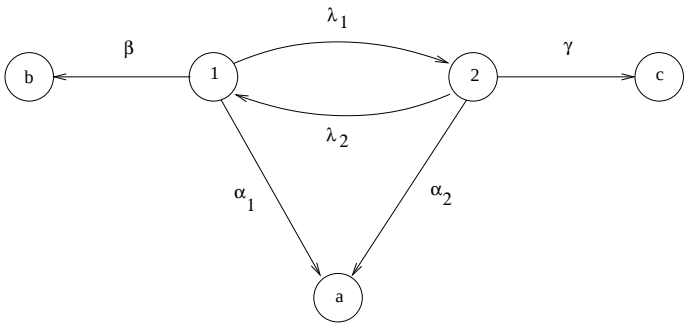


Figure 4.8: *Three absorbing states*

lucky and observe

$$1,\ 2,\ 1,\ 2,\ 1,\ 2,\ 1,\ 2,\ 1,\ 2,\ 1,\ 2,\ \ldots$$

for long time. At some point, we will get bored and stop. Or, we may observe either of the

following trajectories

$$\begin{aligned}
&1\ 2, \ldots, 1,\ a\\
&1\ 2, \ldots, 1,\ 2,\ a\\
&1\ 2, \ldots, 1,\ b\\
&1\ 2, \ldots, 1,\ 2,\ c
\end{aligned}$$

The likelihoods in each case are

$$\begin{aligned}
&e^{-(\lambda_1+\alpha_1+\beta)W(1)}\ \lambda_1^{N(1,2)}\ e^{-(\lambda_2+\alpha_2+\gamma)W(2)}\ \lambda_2^{N(2,1)}\ \alpha_1^{N(1,a)}\\
&e^{-(\lambda_1+\alpha_1+\beta)W(1)}\ \lambda_1^{N(1,2)}\ e^{-(\lambda_2+\alpha_2+\gamma)W(2)}\ \lambda_2^{N(2,1)}\ \alpha_2^{N(2,a)}\\
&e^{-(\lambda_1+\alpha_1+\beta)W(1)}\ \lambda_1^{N(1,2)}\ e^{-(\lambda_2+\alpha_2+\gamma)W(2)}\ \lambda_2^{N(2,1)}\ \beta^{N(1,b)}\\
&e^{-(\lambda_1+\alpha_1+\beta)W(1)}\ \lambda_1^{N(1,2)}\ e^{-(\lambda_2+\alpha_2+\gamma)W(2)}\ \lambda_2^{N(2,1)}\ \gamma^{N(2,c)}.
\end{aligned}$$

The quantities $N(1,a), N(2,a), N(1,b), N(2,c)$ take values 0 or 1 and, amongst them, only one is 1. So we can summarise the likelihood in

$$L = e^{-(\lambda_1+\alpha_1+\beta)W(1)}\ \lambda_1^{N(1,2)}\ e^{-(\lambda_2+\alpha_2+\gamma)W(2)}\ \lambda_2^{N(2,1)}\ \alpha_1^{N(1,a)}\ \alpha_2^{N(2,a)}\ \beta^{N(1,b)}\ \gamma^{N(2,c)},$$

an expression valid for all trajectories.

Clearly, whereas it may be possible to estimate λ_1, λ_2 with the observation of one trajectory alone, this is not the case with the remaining parameters: we need independent trials of the Markov chain. Thus, if we run the Markov chain n times, independently from time to time, a moment of reflection shows that the form of likelihood remains the same as in the last display, provided we interpret $W(x)$ as the total sojourn time in state x over all observations. Same interpretation holds for the quantities $N(x,y)$.

Thus, the general expression (4.7) is valid, in the same manner, for n independent trials. We can, therefore, base our MLE on (4.7). To do this, recall that $\mu(x) = \sum_{y:y\neq x} q(x,y)$ and write the log-likelihood as

$$\log L = \sum_{x,y:\ x\neq y} \big[-W(x)q(x,y) + N(x,y)\log q(x,y)\big].$$

So, for a fixed pair of distinct states x, y,

$$\frac{\partial \log L}{\partial q(x,y)} = -W(x) + \frac{N(x,y)}{q(x,y)}.$$

Setting this equal to zero obtains the MLE estimator

$$\widehat{q}(x,y) = \frac{N(x,y)}{W(x)}. \tag{4.10}$$

N.B. If censoring also occurs, then it can be taken into account, as in Section 4.5.1.

4.6 MLE estimators of the Markov chain rates

There is little to say here, other than what is known for general maximum likelihood estimation: Review the theory of Maximum Likelihood Estimation (summarised in Section B.6).

Consider the problem of estimating a single parameter θ of a Markov chain. Take, for example, θ to be the transition rate $q(\overline{x}, \overline{y})$ between two distinct states $\overline{x}, \overline{y}$. From (4.7) we have

$$\log L = \sum_{x \neq y} [N(x, y) \log q(x, y) - W(x) q(x, y)].$$

The maximum likelihood estimator of $\theta = q(\overline{x}, \overline{y})$ is, as shown in (4.10), given by

$$\widehat{\theta} = \frac{N(\overline{x}, \overline{y})}{W(\overline{x})}.$$

We know that, as the number of observations tends to ∞, $\widehat{\theta}$ converges to the true θ, almost surely. Also, as the number of observations tends to ∞,

$$\operatorname{var} \widehat{\theta} \approx \frac{1}{-E\left(\frac{\partial^2}{\partial \theta^2} \log L\right)},$$

where the expectation is taken under the true θ. This is because, as heuristically argued in Section B.6, $\widehat{\theta}$ asymptotically achieves the Cramér-Rao lower bound, and the denominator in the last display is nothing else but the Fisher information.

Now,

$$-\frac{\partial^2}{\partial \theta^2} \log L = \frac{N(\overline{x}, \overline{y})}{q(\overline{x}, \overline{y})^2}.$$

It is easy to see from (4.8) and (4.9) that

$$EN(\overline{x}, \overline{y}) = q(\overline{x}, \overline{y}) EW(\overline{x}).$$

Hence,

$$\operatorname{var} \widehat{\theta} \approx \frac{q(\overline{x}, \overline{y})}{EW(\overline{x})}.$$

The asymptotic normality statement (B.7) gives that, as the number of observations tends to ∞,

$$\operatorname{var} \widehat{\theta} \text{ is approximately distributed as } \mathcal{N}\left(q(\overline{x}, \overline{y}), \frac{q(\overline{x}, \overline{y})}{EW(\overline{x})}\right).$$

Now let us look at estimators of two parameters simultaneously, such as $(\theta_1, \theta_2) = (q(x_1, y_1), q(x_2, y_2))$. Since the mixed derivatives

$$\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log L = 0,$$

we have that the Fisher information matrix (see Section B.5) is diagonal. Thus, the estimators $\widehat{\theta_1}, \widehat{\theta_2}$ are asymptotically independent and each one asymptotically normal as described above.

Chapter 5

Crude estimation methods

Quite frequently, knowing, or being able to observe, actual times (censoring or deaths) is a luxury. We may only have to rely on counts, i.e. on knowing how many events (of, maybe, a certain type) occurred on some interval of time. This information may be useful, but it may require additional simplifying hypothesis for constructing reasonable maximum likelihood estimators of failure or mortality rates. What we discuss is, in practice, called the Binomial model. We also discuss a classical actuarial formula with very dubious justification and often quite inaccurate.

5.1 The binomial model

This is a very simplistic model, a sort of quick and dirty stochastic model, used when only the number of deaths is known (and not the actual times), within a certain period, a year, say.

Consider N subjects. Suppose you observe D deaths during a year. Make the *ad hoc* assumption that the subjects die independently with the same probability q (which you want to estimate) and so the number of deaths D is given by

$$D = \sum_{i=1}^{N} \eta_i,$$

where $\eta_i = 1$ if the i-th subject dies and 0 otherwise. Since, by assumption, $P(\eta_i = 1) = q$, we have

$$P(D = k) = \binom{N}{k} q^k (1-q)^{N-k}, \quad E(D) = NE(\eta_1) = Nq, \quad \operatorname{var}(D) = N \operatorname{var}(\eta_1) = Nq(1-q).$$

Our standard estimator for q is thus

$$\widehat{q} = \frac{D}{N},$$

so that

$$E(\widehat{q}) = q, \quad \operatorname{var}(\widehat{q}) = \frac{q(1-q)}{N}.$$

To get confidence intervals and other goodies use that

$$\frac{D - Nq}{\sqrt{Nq(1-q)}} \quad \text{is approximately standard normal.}$$

5.2 Maximum likelihood

In this short chapter, we deal with very crude observations: for each subject, we know whether death occurred or not on a specific interval of time. To be specific, let $T_1, \ldots, T_n$ be i.i.d. random variables. Knowing that $T_i > a_i$, for each $i = 1, \ldots, n$, we observe:

$$\mathbf{1}(T_i \in [a_i, b_i]), \quad i = 1, \ldots, n,$$

where $[a_i, b_i]$, $i = 1, \ldots, n$ are fixed intervals. We shall let

$$\delta_i \stackrel{\mathrm{d}}{=} \big[\mathbf{1}(T_i \in [a_i, b_i]) \mid T_i > a_i\big],$$

so that

$$P(\delta_i = 1) = E\delta_i = P(a_i \le T \le b_i \mid T_i > a_i) = {}_{b_i - a_i}q_{a_i}.$$

To ease the notational burden, we let $q_i := {}_{b_i - a_i}q_{a_i}$ and so we have

$$P(\delta_1 = \theta_1, \ldots, \delta_n = \theta_n) = \prod_{i=1}^{n} q_i^{\theta_i}(1 - q_i)^{1-\theta_i}, \quad \theta_1, \ldots, \theta_n \in \{0, 1\},$$

Thus, the likelihood corresponding to the observations is

$$L(\delta_1, \ldots, \delta_n; q_1, \ldots, q_n) = \prod_{i=1}^{n} q_i^{\delta_i}(1 - q_i)^{1-\delta_i}. \tag{5.1}$$

If the q_i are distinct, then trying to obtain MLEs is silly for the maximisation of L with respect to $q_1, \ldots, q_n$ would yield $q_i = \delta_i$ for each i. The problem is meaningful if, for instance, we can group certain q_i's together and consider them identical. For instance, if $a_i = a$, $b_i = b$ for all i, then $q_i = q$ and

$$L(\delta_1, \ldots, \delta_n; q) = q^{\sum_{i=1}^n \delta_i}(1 - q)^{n - \sum_{i=1}^n \delta_i}$$

so that

$$\widehat{q} = \mathrm{Argmax}_{0 \le q \le 1}\, L(\delta_1, \ldots, \delta_n; q) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \tag{5.2}$$

is the MLE of q, with obvious, natural, interpretation: the fraction of observations that where actually deaths.

Note. *We may a priori know that all the T_i have exceeded a certain value x. Then, everything we said above, and everything we'll say below remains unchanged if we replace each T_i by $T_{i,x}$ which (recall!) is defined by $P(T_{i,x} \in \cdot) = P(T_i - x \in \cdot \mid T_i > x)$.*

To deal with the problem in its generality, one resorts to further approximation, referred to as smoothing. Assume that all observation intervals are contained in a common interval of length, say, 1:*

$$[a_i, b_i] \subseteq [0, 1], \quad i = 1, \ldots, n.$$

*There is nothing sacred about this 1; it appears to be a convention among actuaries, so I'll stick to it; just think of 1 as a convenient unit of time, such as one year.

Introduce the parameter $q := P(T < 1)$ and, by making a "reasonable" assumption, think of the probabilities q_i as being certain smooth functions of q:

$$q_i = {}_{b_i-a_i}q_{a_i} \equiv f(a_i, b_i, q).$$

Then the likelihood (5.1) becomes itself a smooth function of q which can be optimised.

For instance, under the *uniform assumption* (See Proposition 2.5),

$$f(a, b, q) = \frac{(b-a)q}{1-aq},$$

under the *Balducci assumption* (See Proposition 2.9),

$$f(a, b, q) = \frac{(b-a)q}{1-(1-b)q},$$

and under the *constant FOM assumption* (See Proposition 2.12),

$$f(a, b, q) = \frac{(1-q)^a - (1-q)^b}{(1-q)^a}$$

Fact 5.1. *Suppose that $a_i = 0$, $b_i = 1$ for all i. Then, under either of the three smoothing assumptions (uniform, Balducci, constant FOM) the MLE of q is as in* (5.2)*. (Exercise.)*

Fact 5.2. *Suppose that $a_i \equiv a$, $b_i \equiv b$ for all i. Then the MLE of q is*

$$\widehat{q} = \begin{cases} \frac{D}{aD+n(b-a)}, & \textit{(uniform)} \\ \frac{D}{(1-b)D+n(b-a)}, & \textit{(Balducci)} \\ 1 - \left(\frac{n}{n-D}\right)^{\frac{1}{a-b}}, & \textit{(constant FOM)} \end{cases},$$

where $D := \sum_{i=1}^n \delta_i$, the number of actual observed deaths. Therefore observe that $\widehat{q}_{\text{unif}} = \widehat{q}_{\text{bald}}$ iff $b = 1 - a$ (why is this reasonable?). (Exercise.)

5.3 The classical actuarial method

This is based † on the Balducci assumption. Just to throw you off, I first give the recipe:
1. Observe each subject between t_i and s_i, where $0 \le t_i < s_i \le 1$. This means that we know that the subject is alive at t_i and withdraws at s_i either due to death due to censoring.
2. Define the quantity

$$\mathsf{E} := \sum_{i=1}^n (s_i - t_i)$$

and call it *(central) exposure*.

3. Count the number of deaths D and, in fact, let δ_i be 1 or 0 depending on whether subject i dies or not.
4. Estimate q by the formula

$$\widehat{q} = \frac{D}{\mathsf{E} + \sum_{i=1}^n (1-s_i)\delta_i}. \tag{5.3}$$

†Actually, it appears not to be based on anything; it is an *ad hoc* method and, I believe, any justification is only *a posteriori*.

5. If you find this formula much too complicated then approximate further by

$$\widehat{q} = \frac{D}{\mathsf{E} + \frac{1}{2}D}. \tag{5.4}$$

6. Introduce a new terminology and call $\mathsf{E} + \frac{1}{2}D$ the *initial exposure*.

Now, here is the a posteriori justification.

First attempt: As I said, the assumption is Balducci. So by taking expectation in $D = \sum_{i=1}^{n} \delta_i$ we find $ED = \sum_{i=1}^{n} q_i$. Then we remember that q_i was my shorthand for ${}_{s_i-t_i}q_{t_i}$ which, under Balducci (Proposition 2.9) ought to be set equal to ${}_{s_i-t_i}q_{t_i} = \frac{(s_i-t_i)q}{1-(1-s_i)q}$. Then try to solve the algebraic equation

$$ED = \sum_{i=1}^{n} \frac{(s_i - t_i)q}{1 - (1 - s_i)q}$$

for q and use this as an estimate provided you replaced ED with the actual number of observed deaths D. This would give an estimate, but this is not the estimate above. So we try again.

Second attempt: Obviously,

$$P(T < 1 \mid T > a) = P(T < b \mid T > a) + P(T > b \mid T > a)P(T < 1 \mid T > b).$$

This is correct. Now use the actuarial notation and do this for each i:

$${}_{1-t_i}q_{t_i} = {}_{s_i-t_i}q_{t_i} + (1 - {}_{s_i-t_i}q_{t_i})\ {}_{1-s_i}q_{s_i}.$$

Nothing wrong with this either. Next replace ${}_{s_i-t_i}q_{t_i}$ by δ_i and write

$${}_{1-t_i}q_{t_i} = \delta_i + (1 - \delta_i)\ {}_{1-s_i}q_{s_i}.$$

This is of course wrong. I know you are tempted to solve for δ_i. But don't. First sum over all i:

$$\sum_{i=1}^{n} {}_{1-t_i}q_{t_i} = \sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} (1 - \delta_i)\ {}_{1-s_i}q_{s_i}.$$

Next use the Balducci assumption: ${}_{1-t_i}q_{t_i} = (1 - t_i)q$, ${}_{1-s_i}q_{s_i} = (1 - s_i)q$:

$$\sum_{i=1}^{n} (1 - t_i)q = D + \sum_{i=1}^{n} (1 - \delta_i)\ (1 - b_i)q,$$

solve for q and lo and behold! You get (5.3)!

And what about (5.4)? Well, if you were convinced that the previous arguments constituted a proof then you will not have any problem in accepting the following argument too: Since s_i occurs somewhere between 0 and 1/2 then, "on the average", it will be 1/2. So replace $1 - s_i$ by 1/2 and voila the proof of (5.4).

Here is a ... little problem. The denominator of (5.3) might be small. For instance, let ε be a small positive number and assume that $s_i = 1 - 2\varepsilon < 1 - \varepsilon = t_i$. Then

$$\widehat{q} = \frac{D}{2\varepsilon n + \varepsilon D}.$$

Since ε could be arbitrarily small, $\widehat{q}$ could be arbitrarily large. Close to infinity... Well, we can't have a probability that is larger than 1, can we?

The problem is not as bad with the approximation of the approximation, i.e. with (5.4), because $D/(\mathsf{E} + 0.5D) \leq D/0.5D = 2$.

5.4 The Poisson model

This is another simplistic model, something to do when you know the number of deaths D on a specific, say, year. Actuaries interpret D/E as a rate of deaths and say "think of D as a random variable which is Poisson-distributed with rate $\mu\mathsf{E}$, where μ is unknown". According to this recipe,

$$P(D=n)=\frac{(\mu\mathsf{E})^n}{n!}e^{-\mu\mathsf{E}},\quad n=0,1,2,\ldots,\quad E(D)=\mu\mathsf{E},\quad \mathrm{var}(D)=\mu\mathsf{E}.$$

Then an estimator for μ is

$$\widehat{\mu}=\frac{D}{\mathsf{E}}.$$

This has

$$E(\widehat{\mu})=\mu,\quad \mathrm{var}(\widehat{\mu})=\frac{\mu}{\mathsf{E}}.$$

These can be used to obtain confidence intervals etc, because

$$\frac{D-\mu\mathsf{E}}{\sqrt{\mu\mathsf{E}}}\quad \text{is approximately standard normal.}$$

Chapter 6

Graduation testing

This chapter has nothing to do with lifetime estimation per se. It is rather a collection of statistical methods for testing how good is a proposed smooth function approximation (or graduation in the actuarial lingo[a]) to crude mortality rates. We discuss goodness of fit and other tests. We will discuss only these methods imposed by the actuarial institution for their exams. Be warned though that some of the tests may be silly or not that good.

[a]It is unclear to me why the terminology "graduation" is used. It appears that it is just tradition and that, nowdays, people whould say "curve fitting" or "smoothing"

6.1 Preliminaries

Consider the problem of fitting a continuous curve to a discrete collection of points in the plane. Specifically, we shall look at pairs of points $(x, \hat{q}_x)$, where x ranges over a finite set, say $x \in \{\ell, \ell + a, \ell + 2a, \ldots, \ell + (n-1)a\}$, representing ages. Here $\hat{q}_x$ is meant to be an estimate for the probability that an individual aged x dies within x and $x + a$ years. These numbers have been obtained by raw data: if we compute the number of deaths D_x among N_x individuals then we set $\hat{q}_x = D_x/N_x$.

People don't like to report pairs of points $(x, \hat{q}_x)$; they'd rather have a formula for a smooth continuous function of x, where x ranges over the real numbers between ℓ and $\ell+(n-1)a$. A procedure for obtaining a smooth function is called "graduation" in the actuarial business. The smooth function is denoted by $\dot{q}_x$.

Here, we shall not learn how to obtain such a function. Rather, we shall describe several tests that test the accuracy of such a function. You will learn how to obtain such a function in the second part of the course, next term. A typical form of such a function used in practice is

$$\dot{q}_x = 1/(1 + e^{-p(x)}),$$

where $p(x)$ is a certain polynomial of degree $k-1$: $p(x) = a_0 + a_1 x + \cdots + a_{k-1}x^{k-1}$, $a_{k-1} \neq 0$. Some terminology: The function $z \mapsto 1/(1 + e^{-z})$ is the inverse of the function

$y \mapsto \log(y/(1-y))$. The latter is called the logistic function:

$$\begin{aligned} \text{logit}(y) &:= \log(y/(1-y)), \quad \text{logit} : (0,1) \to \mathbb{R} \\ \text{logit}^{-1}(z) &:= 1/(1+e^{-z}), \quad \text{logit}^{-1} : \mathbb{R} \to (0,1). \end{aligned}$$

In practice, for each age x, we have a number of individuals N_x. We observe the number D_x of deaths among them and, estimate $\hat{q}_x$ by $\hat{q}_x = D_x/N_x$ and, following *some* method, we compute a smooth function $\dot{q}_x$ that is supposed to approximate $\hat{q}_x$. In doing so, we have to estimate, from the raw data, a number of parameters for the function; e.g., in the logistic equation with a polynomial of degree $k-1$ we need to estimate k coefficients. We then compute, for each x, the standardized D_x, i.e. the quantity

$$Z_x := \frac{D_x - N_x \dot{q}_x}{\sqrt{N_x \dot{q}_x (1-\dot{q}_x)}}.$$

"Theory" tells us that, for each fixed x, this is supposed to be approximately standard normal.

6.2 The χ^2 test

I am going let x range over the set $1, 2, \ldots, n$, where 1 refers to the first age group, 2 the second, and so on, just because I want to have simple subscripts in the variables $Z_1, Z_2, \ldots, Z_n$.

Whereas for each x, Z_x is supposed to be $\mathcal{N}(0,1)$, jointly, the variables $Z_l, Z_2, \ldots, Z_n$ are not independent because of the constraints imposed by the estimation of the parameteres in the function $\dot{q}_x$. In principle, the random vector $Z = (Z_1, Z_2, \ldots, Z_n$ lives in a linear subspace of $\mathbb{R}^n$ which has dimension $n-k$, where k, the number of estimated parameters. This means that the square of the Euclidean norm of Z, namely the quantity

$$||Z||^2 = \sum_{x=1}^{n} Z_x^2,$$

is supposed to have a χ^2_{n-k} distribution (chi-squared with $n-k$ degrees of freedom; degrees of freedom is a name for the dimension of the support of Z).

We can then have the following test for the hypothesis that the model is OK: Compute $||Z||^2$ and if the number is not typical for the χ^2_{n-k} reject the hypothesis. For instance, if we have $n-k=3$ and compute, from data, $||Z||^2 = 11.523$, say, then we should immediately reject the hypothesis that things are OK, because the value 11.523 is not typical for the χ^2_3 distribution. Indeed, a typical value of χ^2_3 is 7.81, because with probability about 95%, a χ^2_3–random variable will fall below 7.81, whereas the chance that a χ^2_3–random variable is above 11 is about 1%.

6.3 The standardised deviations test

Another name for this test is the *likelihood ratio test for the multinomial distribution.* See Section B.7 and, in particular, Theorem B.2. Whereas we argued that the different Z_x are not independent, we shall now make the assumption that they are... Apparently, when n

(number of age groups) is large and the number of parameters k small, then the assumption is not too bad.

The problem then is: Given the $Z_1, \ldots, Z_d$, do the following. Consider a partition of the real line into s disjoint intervals $B_1, \ldots, B_s$, where s is a relatively small number. E.g., $B_1 = (-\infty, -2]$, $B_2 = (-2, -1]$, $B_3 = (-1, 0]$, $B_4 = (0, 1]$, $B_5 = (1, 2]$, $B_6 = (2, \infty)$. Count the number of variables falling in each interval:

$$N(B_i) = \sum_{x=1}^{n} \mathbf{1}(Z_x \in B_i), \quad , i = 1, \ldots, s.$$

If $Z_1, \ldots, Z_d$ were i.i.d. standard normal variables then, clearly, $(N(B_1), \ldots, N(B_s))$ would have a multinomial distribution:

$$P(N(B_1) = n_1, \ldots, N(B_s) = n_s) = \binom{n}{n_1, \ldots, n_s} \theta_1^{*n_1} \cdots \theta_s^{*n_s}.$$

Here

$$\theta_i^* = P(\zeta \in B_i) = \int_{B_i} (2\pi)^{-1/2} e^{-z^2/2} dz$$

is the chance that a standard normal is in B_i. We want to test the hypothesis that this is true, i.e. that the θ_i^* are the correct ones, vs. its negation. By the Likelihood Ratio Test, we have to compute the likelihood ratio

$$\lambda := \frac{\sup_{\theta \neq \theta^*} L(\theta)}{L(\theta^*)} = \frac{\max_\theta L(\theta)}{L(\theta^*)}$$

where $\theta = (\theta_1, \ldots, \theta_s)$ and $L(\theta)$ is the likelihood

$$L(\theta) = \binom{n}{N(B_1), \ldots, N(B_s)} \theta_1^{N(B_1)} \cdots \theta_s^{N(B_s)} \tag{6.1}$$

so that the value of θ that maximises $L(\theta)$ is the MLE statistic. It is not difficult to do this:

Fact 6.1. *Maximising $L(\theta)$ over all $\theta = (\theta_1, \ldots, \theta_s)$ with $\theta_1 + \cdots + \theta_s = 1$ results in the maximiser $\widehat{\theta}_i = N(B_i)/n$, $i = 1, \ldots, s$. (Exercise.)*

Therefore,

$$\lambda = \prod_{i=1}^{s} \frac{\widehat{\theta}_i^{N(B_i)}}{\theta_i^{*N(B_i)}} = \prod_{i=1}^{s} \left(\frac{N(B_i)}{n\theta_i^*} \right)^{N(B_i)}$$

By Theorem B.2, $2 \log \lambda$ should be distributed as χ^2_{s-1}. Now,

$$2 \log \lambda = 2 \sum_{i=1}^{s} N(B_i) \log \left(\frac{N(B_i)}{n\theta_i^*} \right) \approx \sum_{i=1}^{s} \frac{(N(B_i) - n\theta_i^*)^2}{n\theta_i^*}. \tag{6.2}$$

The latter approximation is justified by the Taylor expansion

$$x \log(x/x_0) = (x - x_0) + \frac{(x - x_0)^2}{2x_0} + o((x - x_0)^2),$$

as $x - x_0 \to 0$, and it is justified since, under the hypothesis, $N(B_i)$ should not differ from its mean by too much. The last sum of (6.2) is the *Pearson χ^2 statistic.*

Summarising, we consider the standardised deviations Z_x, split the real line into boxes $B_1, \ldots, B_s$, find the number $N(B_i)$ of the Z_x falling into B_i, for each i, compute the θ_i^* from a normal table and then the Pearson statistic. If the value of it is atypical for the χ^2_{s-1} distribution the we reject the hypothesis; which means that the graduation is not that good.

Note: The test makes no sense when n is not large.

6.4 The sign test

If we believe that the $Z_1, \ldots, Z_n$ are i.i.d. normals then

$$S := \sum_{x=1}^{n} \mathbf{1}(Z_x > 0)$$

should be Binomial$(n, 1/2)$:

$$P(S = s) = \binom{n}{s} 2^{-n}, \quad s = 0, 1, \ldots, n.$$

So we compute S from the data and if S is atypical for the Binomial$(n, 1/2)$ distribution then we are alarmed.

When n is large we can use the standard CLT which says that $(S - E(S))/\sqrt{\text{var}(S)} = (S - (n/2))/\sqrt{n/4}$ is approximately standard normal.

6.5 The change of sign test

This method tests whether changes in sign are typical. Let $\text{sgn}(x)$ be $+1$, -1 or 0 if $x > 0$, $x < 0$ or $x = 0$, respectively. Then consider

$$C = \sum_{x=1}^{n-1} \mathbf{1}(\text{sgn}(Z_x) \neq \text{sgn}(Z_{x+1})).$$

Essentially, we have n i.i.d. random variables $\xi_1, \ldots, \xi_n$ with $P(\xi_1 = 1) = P(\xi_1 = 0) = 1/2$ and then C is distributed as

$$C = \sum_{i=1}^{n-1} |\xi_i - \xi_{i+1}| = \sum_{i=1}^{n-1} \delta_i,$$

where $\delta_i = |\xi_i - \xi_{i+1}|$. We can prove that the δ_i are independent. We do so by induction: First we have $P(\delta_1 = 1, \delta_2 = 1) = P(\xi_1 \neq \xi_2, \xi_2 \neq \xi_3) = P(\xi_1 = 1, \xi_2 = 0, \xi_3 = 1) + P(\xi_1 = 0, \xi_2 = 1, \xi_3 = 0) = 1/8 + 1/8 = 1/4$. Similarly, $P(\delta_1 = 0, \delta_2 = 1) = 1/4$. Next, assume $P(\delta_1 = \varepsilon_1, \ldots, \delta_{n-1} = \varepsilon_{n-1}) = 2^{-(n-1)}$, for all $\varepsilon_1, \ldots, \varepsilon_{n-1} \in \{0, 1\}$ and observe that $P(\delta_1 = \varepsilon_1, \ldots, \delta_n = \varepsilon_n) = P(\delta_1 = \varepsilon_1, \ldots, \delta_{n-1} = \varepsilon_{n-1}) \times P(\delta_{n-1} = \varepsilon_{n-1} \mid \delta_1 = \varepsilon_1, \ldots, \delta_{n-1} = \varepsilon_{n-1}) = 2^{-(n-1)} P(\delta_1 = \varepsilon_1 \mid \delta_{n-1} = \varepsilon_{n-1}) = 2^{-(n-1)} \cdot 2{-}1 = 2^{-n}$. Therefore, C is Binomial$(n - 1, 1/2)$.

So we check whether the value of C is typical for that distribution. Alternatively, if n is large we may check $(C - (n-1)/2)/\sqrt{(n-1)/4}$ against a standard normal.

6.6 The cumulative deviations test

If certain groups of ages are important, it is desirable to test whether the sum of the deviations over this group comes from a normal distribution. Let I be the range of groups deemed important. Then $\sum_{x\in I} Z_x$ is expected to be $\mathcal{N}(0, |I|)$ where $|I|$ is the size of I (e.g., $|I| = n$ if all variables are considered.

6.7 The grouping of signs test

Assume once more that $Z_1, \ldots, Z_n$ are n i.i.d. $\mathbb{N}(0, 1)$ variables, and let

$$\xi_i = \mathbf{1}(Z_i > 0), \quad i = 1, \ldots n.$$

Clearly, $\xi_1, \ldots, \xi_n$ are i.i.d. coin tosses with $P(\xi_i = 1) = P(\xi_i = 0) = 1/2$. We say that $[\xi_\ell, \xi_m]$ is a positive group if $\xi_\ell = \cdots = \xi_m = +1$ and $\xi_{\ell-1} \neq +1$, $\xi_{m+1} \neq +1$. In other words, all variables from index ℓ to index m form a maximal group of positive variables. Let G be the number of positive groups. For example, in

$$+ + + + - - - + - + - - - - - + + + - + - + +$$

there are 6 positive groups.

We define G to be a random variable whose distribution is the number of +groups conditional on the number of positive signs be equal to n_1 (and the number of negative signs equal to $n_2 = n - n_1$). Consider an urn* with a lid and let it contain $n = n_1 + n_2$ items, out of which n_1 are labelled + and n_2 are labelled −. Pick the items at random without replacement and let $\eta_i = \mathbf{1}(i\text{-th item picked has label } +)$. Then, clearly, G is the number of positive groups among $\eta_1, \ldots, \eta_n$. To find the distribution $P(G = t)$ we think as follows: There are $\binom{n}{n_1}$ ways to arrange the $n_1 + n_2$ signs in a row. This is the size of the sample space. So $P(G = t)$ is the size of the set $\{G = t\}$ divided by the size of the sample space. To find the size of the set $\{G = t\}$, i.e. those arrangements that contain exactly t +groups, all we have to do is realise that we have $n_2 + 1$ empty spaces around the n_2 minus signs. E.g., in the example above, we have $n_2 = 11$ minus signs and 12 empty spaces:

$$\square - \square - \square - \square - \square - \square - \square - \square - \square - \square - \square - \square$$

To put $t = 6$ +groups, all we have to do is choose t out of the $n_2 + 1$ empty spaces, e.g., in the example above,

$$\boxed{+} - \square - \square - \boxed{+} - \boxed{+} - \square - \square - \square - \square - \boxed{+} - \boxed{+} - \boxed{+}$$

where a + in a box indicates that this box is replace by a +group. There are $\binom{n_2+1}{t}$ ways to choose t boxes out of the $n_2 + 1$. And now we have to arrange the n_1 +signs in the t +groups. Each +group contains at least one +sign, so there remain $n_1 - t$ +signs to be arranged. But the number of ways to place N identical balls in K boxes is $\binom{N+K-1}{K-1}$. In our

*The concept of an urn is familiar from Statistics. An urn is a container, especially a large round one on a stem, which is used for decorative purposes in a garden, or one which has a lid and is used for holding a dead person's ashes (c.f. `http://dictionary.cambridge.org`). Its use is therefore quite appropriate in mortality studies.

case, $N = n_1 - t$, $K = t$, so we have $\binom{N+K-1}{K-1} = \binom{n_1-1}{t-1}$. Putting things together, we find that the size of the set $\{G = t\}$ is $\binom{n_2+1}{t}\binom{n_1-1}{t-1}$. Therefore,

$$P(G = t) = \frac{\binom{n_2+1}{t}\binom{n_1-1}{t-1}}{\binom{n}{n_1}}, \quad 1 \le t \le \min(n_1, n_2 - 1). \tag{6.3}$$

(Or if we simply let $\binom{N}{M} = 0$ for $M > N$, we need no restriction on t.)

Comparing (6.3) and (B.12) of §B.9 we see that G has a hypergeometric distribution:

$$P(G = t) = H(n, n_2 + 1, n_1, t). \tag{6.4}$$

From this realisation, and formulae (B.13), (B.14) of §B.9, we can find the first two moments of G:

$$EG = \frac{n_1(n_2 + 1)}{n}, \quad \operatorname{var} G = \frac{n_1 n_2 (n_1 - 1)(n_2 + 1)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

(This required setting $K = n$, $R = n_2 + 1$, and $k = n_1$ in (B.13), (B.14).) Note that, when n_1, n_2 are large numbers we have

$$EG \approx \frac{n_1 n_2}{n}, \quad \operatorname{var} G \approx \frac{(n_1 n_2)^2}{n^3}.$$

Finally, we can translate this to the r.v. G and assert that G is approximately $\mathcal{N}(n_1 n_2/n,\ (n_1 n_2)^2/(n_1 + n_2)^3)$. (And we also know the conditions: This is a statement that holds when n is large and when n_i/n remain constant as $n \to \infty$, $i = 1, 2$.)

Finally, one can prove that, for large n_1, n_2,

$$\text{the distribution of } G \text{ is approximately } \mathcal{N}\big(EG, \operatorname{var}(G)\big). \tag{6.5}$$

There is a very neat normal approximation argument going on here, one that requires the observation (i) that G has a hypergeometric distribution and (ii) that the hypergeometric distribution is approximared by a normal. To understand why, read §B.9; look at the result (B.17) stating the normal approximation to a hypergeometric distribution and, using (6.4), translate it to the statement (6.5).

The test then is as follows: Count, from the data, the number n_1 of positive signs, and the number G of positive groups. Let $n_2 = n - n_1$. See if the quantity $(G - EG)/\sqrt{\operatorname{var}(G)}$ is typical for the standard normal distribution.

6.8 The serial correlation test

Define the sample covariance between two random vectors $X = (X_1, \ldots, X_m)$, and $Y = (Y_1, \ldots, Y_m)$ by

$$C_{X,Y} = \frac{1}{m} \sum_{i=1}^{m} (X_i - M_X)(Y_i - M_Y),$$

where M_X, M_Y are the sample means:

$$M_X = \frac{1}{m} \sum_{i=1}^{m} X_i, \quad M_Y = \frac{1}{m} \sum_{i=1}^{m} Y_i.$$

By the Cauchy-Schwarz inequality we have

$$C_{X,Y}^2 \le C_{X,X} C_{Y,Y},$$

where

$$C_{X,X} = \frac{1}{m}\sum_{i=1}^{m}(X_i - M_X)^2, \quad C_{Y,Y} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - M_Y)^2$$

are the sample variances. Hence the quantity

$$\rho_{X,Y} = \frac{C_{X,Y}}{\sqrt{C_{X,X}C_{Y,Y}}} = \frac{\sum_{i=1}^{m}(X_i - M_X)(Y_i - M_Y)}{\sqrt{\sum_{i=1}^{m}(X_i - M_X)^2 \ \sum_{i=1}^{m}(Y_i - M_Y)^2}}$$

is between -1 and $+1$ and is called the sample correlation[†].

The lag-k correlation of $Z_1, Z_2, \ldots, Z_n$ is the correlation between $(Z_1, \ldots, Z_{n-k})$ and $(Z_{k+1}, \ldots, Z_n)$:

$$\rho_n(k) = \frac{\sum_{i=1}^{n-k}(Z_i - M_{n-k})(Z_{k+i} - M'_{n-k})}{\sqrt{\sum_{i=1}^{n-k}(Z_i - M_{n-k})^2 \ \sum_{i=1}^{n-k}(Z_{k+i} - M'_{n-k})^2}},$$

where

$$M_{n-k} = \frac{1}{n-k}\sum_{i=1}^{n-k} Z_i, \quad M'_{n-k} = \frac{1}{n-k}\sum_{i=1}^{n-k} Z_{k+i},$$

The sequence of random variables $\rho_n(k)$, $n = k+1, k+2, \ldots$, has a well-known distributional limit (see e.g. Brockwell and Davis [4]), which can be described by saying that $\sqrt{n-k}\ \rho_n(k)$ is approximately $\mathcal{N}(0,1)$–disributed. This provides a an obvious test.

6.9 An example

In this example, we consider $n = 16$ age groups. Each group is represented by a typical age x. Next to it, we have the number N_x of individuals observed, and the observed deaths D_x. The crude mortality rates are estimated by $q_x = D_x/N_x$. The graduated values are estimated using the equation

$$\dot{q}_x = \text{logit}^{-1}(a_0 + a_1 x + a_2 x^2),$$

where $a_0 = -6.148874\text{m}$ $a_1 = -0.001511$, $a_2 = 0.000697$. The values of $N_x\dot{q}_x$ are summarised in the next column. Finally, the standardised deviations $Z_x := \frac{D_x - N_x\dot{q}_x}{\sqrt{N_x\dot{q}_x(1-\dot{q}_x)}}$ are in the last column.

[†]or sample correlation function

	Age	N_x	D_x	$N_x\dot{q}_x$	Z_x
1.	12	8119	14	18.78	-1.104
2.	17	7750	20	19.68	0.071
3.	22	6525	22	18.83	0.731
4.	27	5998	23	20.37	0.584
5.	32	5586	26	23.10	0.603
6.	37	5245	28	27.35	0.124
7.	42	4659	32	31.70	0.054
8.	47	4222	37	38.77	-0.286
9.	52	3660	44	46.92	-0.429
10.	57	3012	54	55.70	-0.229
11.	62	2500	68	68.81	-0.098
12.	67	2113	87	89.10	-0.226
13.	72	1469	100	97.26	0.287
14.	77	883	95	93.36	0.179
15.	82	418	70	70.87	-0.113
16.	87	181	49	48.40	0.100

We now run various tests to decide on the goodness of the graduation.

The χ^2 test We find $||Z||^2 = \sum_x Z_x^2 = 3.003$. Since we estimated 3 parameters, Z actually lives in a $(n-3) = 13$ –dimensional space. Hence $||Z||^2$ is compared against a χ^2_{13}. We have that the probability, under χ^2_{13} of $(3, \infty)$ is at least 0.99. Hence the observed value is typical and the test shows nothing unsatisfactory.

The standardised deviations test Split the real line intow $s = 4$ boxes

$$B_1 = (-\infty, -1], \quad B_2 = (-1, 0], \quad B_3 = (0, 1], \quad B_4 = (1, \infty).$$

Let θ_i^* be the chance that a standard normal falls in B_i:

$$\theta_1^* = \theta_4^* = 0.159, \quad \theta_2^* = \theta_3^* = 0.5 - 0.159 = 0.341.$$

Let $N(B_i)$ be the number of the Z_x in B_i. We find:

	B_1	B_2	B_3	B_4
$n\theta_i^*$	2.54	5.47	5.47	2.54
$N(B_i)$	1	6	9	0

Next compute Pearson's statistic

$$2 \log \lambda \approx \sum_{i=1}^{4} \frac{N(B_i) - n\theta_i^*)^2}{n\theta_i^*} = 5.87.$$

Since $(5.87, \infty)$ has chance 0.12 under χ^2_3, we have evidence that something goes wrong.

Attempting to define the B_i in different ways, also gives evidence against the goodness of the graduation.

The sign test We count the number S of the Z_x which are positive. We find $S = 9$. We expect that S is Binomial$(16, 1/2)$. The value 9 is typical for this distribution. So the test shows nothing wrong.

The change of sign test We count the number C of sign changes: $C = 5$. We expect that C is Binomial$(15, 1/2)$. Again, $C = 5$ is not atypical.

The cumulative deviations test We compute $\sum_x Z_x = -4.7 \times 10^{-7}$. Clearly, this is quite typical for $\mathcal{N}(0, 16)$.

The grouping of signs test There are $n_1 = 9$ positive and $n_2 = 7$ negative signs. We observe $G = 3$ positive groups. We use (6.3) for the theoretical distribution of G given n_1, n_2. Under this distribution, we find $P(G \leq 3) = 0.16$. Since $EG = 4.5$, $\text{var}(G) \approx 0.984$.

The serial correlation test Compute the lag-1 correlation coefficient, i.e. the correlation coefficient between $(Z_1, \ldots, Z_{n-1})$ and $(Z_2, \ldots, Z_n)$. We finf $\rho_n(1) = 0.461$. So $\sqrt{n}\rho_n(1) = 1.78$. But $\sqrt{n}\rho_n(1)$ is supposed to be $\mathcal{N}(0, 1)$. The value 1.78 is not typical: the standard normal probability of $(1.7, \infty)$ is small, about 0.05. Hence we have evidence against uncorrelatedness.

Why should there be? As I explained initially, the random vector $(Z_1, \ldots, Z_n)$ is supported in an $(n - k)$ –dimensional space. The subsequent tests made the ad hoc assumption, in violation to this, that the components are independent.

Appendix A

DISCUSSION AND EXERCISES

The sections of this appendix follow the order and structure of the preceding chapters. They contain additional points for discussion as well as problems for solution.

A.1 Introduction

The actuarial notation is useful because you will want to talk to actuaries who have been using it for quite some time. Notation is actually a very delicate subject. Good notation can be quite useful in the sense that it helps you think. Imagine that a few centuries ago, people did not use the sign + or the relation =, they simply did not have the symbols.

Here is a symbol I will be using all the time: that of an indicator function. If Ω is a set and $A \subseteq \Omega$, then $\mathbf{1}_A$, often written as $\mathbf{1}(A)$, is the function that assigns the value 1 to every $\omega \in A$, and 0 otherwise. Frequently, A is defined by means of a logical statement, so we tend to interpret $\mathbf{1}_A$ as the truth of A.

EXERCISE A.1.1. Establish that $\mathbf{1}_{A^c} = \mathbf{1}_A$, $\mathbf{1}_{A\cap B} = \mathbf{1}_A\mathbf{1}_B$. Use these rules to show that $\mathbf{1}_{A\cup B} = \mathbf{1}_A+\mathbf{1}_B-\mathbf{1}_{A\cap B}$ and that $\mathbf{1}_{A\cup B\cup C} = \mathbf{1}_A+\mathbf{1}_B+\mathbf{1}_C-\mathbf{1}_{A\cap B}-\mathbf{1}_{B\cap C}-\mathbf{1}_{C\cap A}+\mathbf{1}_{A\cap B\cap C}$.

Now assume that P is a probability on (a class of subsets of) Ω. Then, is A belongs to the domain of P, $E\mathbf{1}_A = P(A)$. Hence

EXERCISE A.1.2. Deduce, from Exercise A.1.1, the inclusion-exclusion formula $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Here is something more subtle, but, as you will probably remember, one of the most useful things in Probability. Suppose that you know that the events A_n, $n = 1, 2, \ldots$, have probabilities $P(A_n)$ so small that they tend to zero so fast that $\sum_{n=1}^{\infty} P(A_n) < \infty$. Show that the the probability of the event "an infinite number of A_n occur" is zero. We approach the problem as follows. Since $\mathbf{1}_{A_n}$ is 1 is A_n occurs, the random variable $Z = \sum_{n=1}^{\infty} \mathbf{1}_{A_n}$ is the total number of events that occur. But $E(Z) = \sum_{n=1}^{\infty} E\mathbf{1}_{A_n} = \sum_{n=1}^{\infty} P(A_n) < \infty$. Since $E(Z) < \infty$, the variable itself must be finite. But Z, being the sum of numbers that are either 0 or 1 is finite iff there are only finitely many 1's. Which means that finitely many of the A_n occur. (This is the *Borel-Cantelli* lemma, one of the most trivial and most useful results in Probability.)

The "dx" notation is useful when dealing with absolutely continuous random variables. For example, suppose that X, Y are independent, exponential rate-1 variables.

Example A.1. Compute the joint density of $X, Z := X + Y$.
The pdf of both X and Y is $f(t) = e^{-t}\mathbf{1}(t > 0)$. To do this, write

$$P(X \in dx, Z \in dz) = P(X \in dx, x + Y \in dz),$$

at which point we remember that X, Y are independent, so the above display equals

$$f(x)dxf(z - x)dz.$$

Thus the joint density of X, Z equals

$$f_{X,Z}(x, z) = e^{-x}\mathbf{1}(x > 0)e^{z-x}\mathbf{1}(z - x > 0) = e^{z}\mathbf{1}(0 < x < z).$$

In the notation q_x, which stands for ${}_1q_x$, and means $P(T_x < 1)$, i.e. $P(T < x + 1 \mid T > 1)$ the constant 1 has no special significance. It means "one unit of time". It could be 1 year, or, 1 month; it depends on the units we choose, and on the granularity of the mortality rate data. Knowing q_x, for x a multiple of the unit 1, is partial information about the random variable.

EXERCISE A.1.3. The random variable T_x was defined distributionally by saying that T_x is distributed as $T - x$ given that $T > x$. Hence T_x takes values in $[0, \infty)$. Another variable that takes values in $[0, \infty)$ is $(T-x)^+ = \max(T-x, 0)$. Are these random variables different? In what way?

Some more notation:

1. $x \vee y := \max(x, y)$
2. $x \wedge y := \min(x, y)$
3. $x^+ = \max(x, 0)$
4. $x^- = \max(-x, 0) = (-x)^+$

EXERCISE A.1.4. If S, T are independent positive random variables, then $(S \wedge T)_x \stackrel{\mathrm{d}}{=} S_x \wedge T_x$

EXERCISE A.1.5. Can you do something as clean and nice for $S \vee T$?

EXERCISE A.1.6. Let $T_1, T_2, \ldots$ be i.i.d. exponential rate $\lambda_1, \lambda_2, \ldots$, respectively. Show that $A_n = \min(T_1, \ldots, T_n)$ is exponential rate $\lambda := \lambda_1 + \cdots + \lambda_n$ ansd that $P(A_n = T_i) = \lambda_i/\lambda$. If the λ_i are all equal, say to 1, show that $B_n = \max(T_1, \ldots, T_n) \stackrel{\mathrm{d}}{=} T_1 + (T_2/2) + (T_3/3) + \cdots + (T_n/n)$. Deduce that $EB_n = 1 + (1/2) + (1/3) + \cdots + (1/n)$, so that, as $n \to \infty$, $EB_n \approx \log n$.

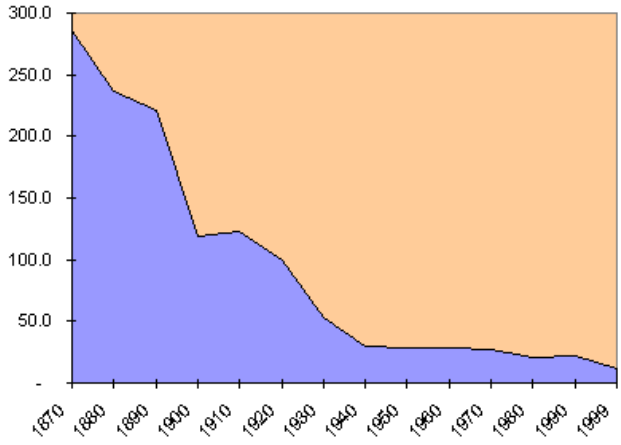
EXERCISE A.1.7. Let S, T be independent exponential rate 1 variables. Compute $E(S \vee T \mid S \wedge T)$. Interpretation: S, T are independent lifetimes. Given that we observed one death, what is the expected time for the other death?

A.2 Force of mortality

As we discussed, FOM's are defined only for absolutely continuous random variables, and we do take the variables to be positive. Since $\mu_t dt$ has the interpretation of a probability of failure within t and $t+dt$, and probability has no units (dimensionless), we measure μ_t in units of TIME^{-1}; which is in agreement with its name "rate". Being a rate, it does not need to be smaller or bigger than 1. It can be anything, as long as it is nonnegative. So, if μ_t is given in units months^{-1} and we want to find $P(T > 267$ days) we can either convert days to months ($267/30 = 8.9$) and do $P(T > 8.9$ months$) = \exp - \int_0^{8.9} \mu_t dt$ or convert the units of μ_t. To convert the units of μ_t, we must convert both the independent and dependent variables. So if μ_t is measured in months^{-1} and t has the month as a unit, then the correct change is $30\mu_{30s}$ in days^{-1} where s now run over days. Indeed, $P(T > 267$ days$) = \exp - \int_0^{267} 30\mu_{30s} ds$.

EXERCISE A.2.1. Recall that if $f(x)$ is the density of a random variable X then the density of X/α is $\alpha f(\alpha x)$.

Mortality rates have been dramatically higher about 100 years ago, even in economically advanced countries



Estimated Chicago infant mortality rates: 1870 to 1999 (deaths under 1 year of age per 1000 live births)

EXERCISE A.2.2. Discuss how you would use the information on the graph above.

EXERCISE A.2.3. Suppose that S, T are independent lifetimes with FOMs μ_t, λ_t. Find the FOM of $S \wedge T$. (Hint: Show it is the sum of the FOM's.) What about the FOM of $S \vee T$? Fix x, y and consider T_x, S_y (independent). Find the FOM of $T_x \wedge S_y$. (Hint: Recall that the FOM of T_x is the FOM of T translated by x.)

EXERCISE A.2.4. Pick your favourite distributions and find their FOM's.

EXERCISE A.2.5. Extensive mortality studies yield the following estimates: $q_0 = 0.500, q_1 = 0.333, q_2 = 0.250, q_3 = 0.200, q_4 = 0.167, q_5 = 0.143, \ldots$, where, as usual, $q_n = P(T < n+1 \mid T > n)$. An actuary decides to model this by letting $q_n = 1/(n+2)$, $n = 0, 1, 2, \ldots$.
(i) Using the Balducci interpolation, whereby $P(T_n > t) = (1-q_n)/(1-(1-t)q_n)$ for $0 < t < 1$, show that the T is modelled as a random variable with $P(T > t) = 1/(t+1)$, $t > 0$.

(ii) Why is this a bad model? (Hint: Consider ET.)
(iii) Under a different interpolation scheme, say $P(T_n > t) =: g_n(t)$, $0 < t < 1$, show that

$$P(T > t) = \frac{\prod_{n=0}^{[t]}(1 - q_n)}{1 - (1 - t + [t])q_{[t]}},$$

where $[t]$ is the integer part of t.
(iv) Would the choice of (i) become more reasonable under a different interpolation scheme?

EXERCISE A.2.6. (i) If S, T are independent with FOMs λ_t, μ_t, respectively, show that

$$P(S < T \mid S \wedge T = t) = \frac{\lambda_t}{\lambda_t + \mu_t}.$$

(ii) Generalise this to an arbitrary number of independent absolutely continuous random variables.

EXERCISE A.2.7. If T has survival function $S(t) = P(T > t)$ then T_x has survival function $S_x(t) = P(T_x > t) = S(x + t)/S(x)$, $t \geq 0$. If T has density f then T_x has density $f_x(t) = f(x + t)/S(x)$, $t \geq 0$.

EXERCISE A.2.8. Prove that $T_{t,x} = T_{t+x}$ for any $t, x \geq 0$.

EXERCISE A.2.9. Show that μ_t is the negative of the logarithmic derivative of $S(t)$, i.e. $\mu_t = -S'(t)/S(t)$.

EXERCISE A.2.10. By making use of the actuarial notation, show that:

$$\begin{aligned}
{}_{s+t}p_x &= {}_tp_x \cdot {}_sp_{x+t} \\
f(t) &= {}_tp_0 \cdot \mu_t \\
f_x(t) &= {}_tp_{x+t} \cdot \mu_{x+t} \\
{}_tp_0 &= \exp - \int_0^t \mu_s ds \\
{}_tp_x &= \exp - \int_x^{x+t} \mu_s ds \\
{}_tq_x &= \int_0^t {}_sp_x \cdot \mu_{x+s} ds
\end{aligned}$$

EXERCISE A.2.11. (i) Suppose that $S(t) = e^{-\lambda t}$, $t \geq 0$, for some positive constant λ. Show that $\mu_t \equiv \lambda$. (ii) Suppose that T is uniform on the interval $[0, 1]$. Show that $\mu_t = 1/(1 - t)$, for $0 \leq t < 1$ (and zero otherwise). (iii) Suppose that $S(t) = t^{-\alpha}$, for $t \geq 1$, for some positive constant α. Show that $\mu_t = \alpha/t$, for $t \geq 1$.

EXERCISE A.2.12. Find a random variable with non-monotonic force of mortality.

EXERCISE A.2.13. Show that, among positive continuous random variables, only the exponential has the memoryless property. Also show that the memoryless property can be written as ${}_tp_s = {}_tp_0$ for all $s, t \geq 0$.

EXERCISE A.2.14. Show Fact 2.4 and thus show that the uniform assumption is equivalent to

$${}_tq_x = tq_x.$$

EXERCISE A.2.15. The Balducci assumption is equivalent to

$$ {}_{1-t}q_{x+t} = (1-t)q_x, \quad 0 \le t \le 1. $$

EXERCISE A.2.16. The FOM assumption is equivalent to

$$ \mu_{x,t} = \mu_{x+t} = -\log(1-q_x), \quad 0 \le t \le 1, $$

and to

$$ {}_tq_x = P(T_x > t) = (1-q_x)^t. $$

EXERCISE A.2.17. Show that the process $\widetilde{N}$ defined in (2.3) is the compensator of the point process N of Example 2.5.

EXERCISE A.2.18. Let T be an absolutely continuous positive random variable with density $f(t)$, distribution function $F(t)$, survival function $S(t) = 1 - F(t)$, and FOM $\mu_t = f(t)/S(t)$. Show that the compensator $\widetilde{N}_t$ of the point process $\{T\}$ is given by

$$ \widetilde{N}_t = -\log S(T \wedge t). $$

Hint: This same $\widetilde{N}_t$ was derived in equation (2.2).

EXERCISE A.2.19. Let μ_t be the FOM of T and assume that U is independent of T. Then the compensator of N_t is given by

$$ \int_0^t \mathbf{1}(X \ge s)\mu_s ds = -\log S(t \wedge T \wedge U). $$

(Exercise.)

EXERCISE A.2.20. Prove this last proposition in detail. Basically, check that the arguments preceding it are correct and fill in the gaps, if necessary.

EXERCISE A.2.21. Let $T_1, \ldots, T_n$ be i.i.d. with common FOM μ_t. Let $U_1, \ldots, U_n$ be independent random variables, and independent of the $T_1, \ldots, T_n$. Define $X_i = T_i \wedge U_i$, $i = 1, \ldots, n$ (censoring). Consider the point process

$$ N_t := \sum_{i=1}^n \mathbf{1}(X_i \le t, T_i \le U_i). $$

Define the *number-at-risk* process

$$ Y_t := \sum_{i=1}^n \mathbf{1}(X_i \ge t). $$

Then the compensator of N_t is given by

$$ \widetilde{N}_t = \int_0^t Y_u \mu_u du. $$

EXERCISE A.2.22. The process H_t defined in (2.4) has a compensator given by

$$ \widetilde{H}_t = \int_0^t G_u d\widetilde{N}_u, $$

i.e. (2.5) holds. (Exercise.)

EXERCISE A.2.23. The jolly fellow of Figure 3.3 has something to do with lifetime estimation or with mortality studies. (Exercise: What?)

A.3 Lifetime estimation

EXERCISE A.3.1. Suppose you have n observations representing actual failure times. For instance, you may have been assigned the job of an inspector in a factory that manufactures electric motors. The motors are run continuously till they fail. You started work in January 2005. During the next 12 months you record the actual day at which a motor fails. But when you started work, because of bureaucracy, nobody could tell you the exact day at which each motor was put into operation. Therefore, the times $T_1, \ldots, T_n$ you record are absolute failure epochs. Your task is to estimate the distribution of the lifetime of a motor. You assume that all lifetimes are i.i.d. following a common unknown distribution F. Prior experience, and some background research, shows that you may assume that $T_i > V_i$ where V_i is a random epoch in the past with known distribution G. Also, the V_i's are i.i.d. and independent of the T_i's. Discuss how you would go about estimating F.

EXERCISE A.3.2 (computer exercise)**.** Pick your favourite distribution F, e.g., uniform on $[0, 1]$. Generate n i.i.d. samples from it and plot the empirical distribution function $\widehat{F}_n$. Do it, e.g., for $n = 100$ and $n = 1000$, using MAPLE. You next want to be 95% confident that the distance between the estimated $\widehat{F}_n$ and the actual F is less than 0.01. How many samples n should you produce?

EXERCISE A.3.3 (computer exercise)**.** Repeat Exercise A.3.2 but do introduce censoring. First introduce censoring by an independent random variable U with mean twice the mean of T, and then with mean 10 times the mean of T. Construct the Kaplan-Meier estimator in each case. Discuss and compare.

EXERCISE A.3.4. The 30 following numbers were produced as i.i.d. samples from a given distribution on $[0, 1]$. Estimate it.
0.864, 0.756, 0.100, 0.841, 0.729, 0.608, 0.688, 0.462, 0.100, 0.784,
0.129, 0.361, 0.676, 0.225, 0.980, 0.756, 0.152, 0.102, 0.722, 0.792,
0.129, 0.462, 0.270, 0.176, 0.921, 0.980, 0.739, 0.864, 0.640, 0.324

A.4 Markov chains and estimation

To better understand the general Markov models, consider the following situation:

$$0 \overset{\alpha_t}{\to} 1 \overset{\beta_t}{\to} 2.$$

We have a stochastic process that starts at $X = 0$, at some point of time it becomes 1 and, later, 2 where it remains forever.

EXERCISE A.4.1 (Nonhomogeneous Markov)**.** Interpret it as a Markovian model with the transition rates shown Let $\tau_0 = \inf\{t \geq 0 : X_t = 1$, $\tau_1 = \inf\{t \geq 0 : X_{\tau_0+t} = 2\}$. Show that τ_0, τ_1 are not independent and derive a formula for their joint density. Find the expectation of $T = \tau_0 + \tau_1$.

Now change the model.

EXERCISE A.4.2 (Semi-Markov)**.** Suppose that τ_0, τ_1 are independent random variables with FOMs α_t, β_t, respectively. Define $X_t = 0, t < \tau_0$, $X_t = 1, \tau_0 \leq t < \tau_0 + \tau_1$, $X_t = 2, t \geq \tau_0 + \tau_1$. Find the expectation of $T = \tau_0 + \tau_1$. Show that $\{X_t, t \geq 0\}$ is not Markov. Let $\sigma_t := \inf\{s \geq 0 : X_{t+s} \neq X_{(t+s)-}\}$. (In words: It takes time σ_t after t till the next jump of the process.) Show that $\{(X_t, \sigma_t), t \geq 0\}$ is Markov.

In a sense, the clock restarts when the jump occurs, in the second model. This is an example of a semi-Markov process. Just a glimpse into these models: Recall the construction of a homogeneous continuous-time Markov chain via the discrete chain (4.3) and the exponential random variables (4.4). Replace the exponential distribution by a general distribution (different for each x), and you get a semi-Markov model.

EXERCISE A.4.3. For the model of Case 4.2:
(i) Write down the Kolmogorov backwards equations.
(ii) Solve them when the rates are constant.
(iii) Find the long-run fraction of time that the person is employed.
(iv) Suppose that the chain starts at state UNEMPLOYED. Find the distribution of the time it takes for this state to be reached for a second time.
Hint: This is the same as the punabhava chain.

EXERCISE A.4.4. For the model of Case 4.3:
(i) Write down and solve the Kolmogorov backwards equations.
(ii) When the rates are constant, and the initial state is HEALTHY, find the average time till DEATH occurs.
(iii) Introduce a rate ρ of resurrection ($q(0, 1) = \rho$). Find the long-run fraction of time the person spends in the DEAD state.
(iv) If we start from the ILL state, how long does it take, on the average to becom HEALTHY again (necessarily through resurrection)?

EXERCISE A.4.5. For the model of Case 4.4, assuming that the initial state is 11:
(i) Assuming α_t, β_t are time-varying, find the expected time till one of the two people passes away.
(ii) Repeat when the rates are constant.
(iii) With constant rates again, find the expected time till both people die.

EXERCISE A.4.6. For the model of Case 4.5:
(i) Assuming constant rates, find the expected time till retirement or death.
(ii) Write and solve the Kolmogorov backwards equations.
(iii) For the time-dependent case, assuming that $\sigma_t \equiv 0$, write and solve the Kolmogorov backwards equations.
(iv) Assuming constant rates, and assuming that the chain never hits the set of states RETIRED, DEATH, find the transition rates from ABLE to ILL. (Practical meaning: If it takes extremely long time till retirement or death occurs, then the model behaves like another Markov chain, in fact, like a punabhava chain; what are its rates?)

EXERCISE A.4.7. Translate the semigroup equation (4.1), $P_{s,u}(x,y) = \sum_z P_{s,u}(x,z)P_{u,t}(z,y)$ in actuarial notation.

EXERCISE A.4.8. If $x = y$, the derivative of $P(X_{t+h} = x \mid X_t = x)$ equals

$$q_t(x,x) := \lim_{h\downarrow 0} \frac{1}{h}[P(X_{t+h} = x \mid X_t = x) - 1] = -\sum_{x\neq y} q_t(x,y).$$

EXERCISE A.4.9. Use Landau's symbols to see that the definition of transition rates is equivalent to

$$P_{t,t+h} - I = hQ_t + o(h), \qquad \text{as } h \to 0.$$

EXERCISE A.4.10. Prove the backward equations.

EXERCISE A.4.11. The forward equations, in actuarial notation, read:

$$\frac{\partial}{\partial t}{}_tp_s^{xy} = \sum_{z\neq y} {}_tp_s^{xz}\mu_{s+t}^{zy} - {}_tp_s^{xy}\mu_{s+t}^{yz}.$$

EXERCISE A.4.12. Write the backward equations in actuarial notation.

EXERCISE A.4.13. Argue that, in the Punabhava chain, unless μ_t and λ_t are constant, the random variables τ_0, τ_1 are not independent.

EXERCISE A.4.14. Show that the FOM of $T_s^x := \inf\{t > 0 : X_{s+t} \neq x\}$ is $\sum_{y\neq x} q_{s+t}(x,y)$, $t \geq 0$, and so show, by applying Fact 2.2, that

$${}_tp_s^{\overline{xx}} := P(T_s^x > t \mid X_s = x) = \exp - \int_s^{s+t} \sum_{y\neq x} q_u(x,y)du.$$

EXERCISE A.4.15. For a time-homogeneous chain, show that $\frac{d}{dt}e^{tQ} = Qe^{tQ} = e^{tQ}Q$.

EXERCISE A.4.16. The process (X_t) constructed in Section 4.4, via (4.3), (4.4), and (4.5), is a homogeneous Markov chain with transition rates $q(x,y)$. Hint: Show that (4.2) hold.

EXERCISE A.4.17. Derive an equation for $P_x(\sigma_A \leq t)$, along the lines of the derivation of (4.6).

EXERCISE A.4.18. Going back to nonhomogeneous chains, discuss why it is necessary to replace the notation P_x, E_x, by $P_{x,t}$, $E_{x,t}$, i.e. $P_{x,t}(A) = P(A \mid X_t = x)$. How could one write backwards equations then?

EXERCISE A.4.19. Consider the chain of Figure 4.4 with constant rates σ, ρ. Suppose we want to simultaneously estimate them. We observe the chain, for a long period of time and find that $N(1,0) = 100$, $N(0,1) = 50$, $W(1) = 400$, $W(2) = 500$, where $N(x,y)$ is the total number of transitions from x to y and $W(x)$ is the total time spent in state x. Compute the MLE $(\widehat{\lambda}, \widehat{\mu})$ of (λ, μ) and find a 99% confidence interval for each of the parameters.

EXERCISE A.4.20. Consider the chain of Figure 4.7 with constant rates. Discuss how to set up an experiment to estimate all the rates. Write down their MLEs. Discuss their asymptotic properties.

A.5 Crude estimation methods

Some of the exercises in Section A.2 are actually useful for this section as well. Do review the relevant material from Section 2.2 on interpolation.

EXERCISE A.5.1. Suppose that $[a_i, b_i]$ of Section 5.2 are either $[a_1, b_1]$ or $[a_2, b_2]$. Formulate and find the MLEs of this 2-parameter problem.

EXERCISE A.5.2. Suppose, in Section 5.2, that $a_i = 0$, $b_i = 1$ for all i. Then show, that under either of the three smoothing assumptions (uniform, Balducci, constant FOM) the MLE of q is as in (5.2).

EXERCISE A.5.3. Suppose, in Section 5.2, that $a_i \equiv a$, $b_i \equiv b$ for all i. Then show that the MLE of q is

$$\widehat{q} = \begin{cases} \frac{D}{aD+n(b-a)}, & \text{(uniform)} \\ \frac{D}{(1-b)D+n(b-a)}, & \text{(Balducci)} \\ 1 - \left(\frac{n}{n-D}\right)^{\frac{1}{a-b}}, & \text{(constant FOM)} \end{cases},$$

where $D := \sum_{i=1}^{n} \delta_i$, the number of actual observed deaths. Therefore observe that $\widehat{q}_{\text{unif}} = \widehat{q}_{\text{bald}}$ iff $b = 1 - a$ (why is this reasonable?).

EXERCISE A.5.4. Within a year, the following were observed: $D = 35$ deaths among $N = 947$ individuals, and $\mathsf{E} = 593$ years. Assuming the Poisson model, find an estimator for the death rate μ as well as a 99% confidence interval.

EXERCISE A.5.5. Here are some data from a mortality inverstigation. We observed 4 individuals during the year 1996. Individual i is observed between a_i and b_i, and $\delta_i = 1$ if he/she dies during this time or 0 otherwise:

i	a_i	b_i	δ_i
1	0.98	0.99	0
2	0.90	0.95	0
3	0.85	0.98	1
4	0.95	0.98	1

Let q be the probability of death by the end of the year. This is an unknown parameter we wish to estimate.
(i) Use the Uniform assumption in the likelihood (5.1) and compute the MLE $\widehat{q}$ of q.
(ii) Estimate the variance of $\widehat{q}$ by using $\text{var}(\widehat{q}) \approx -1/L''(\widehat{q})$ and the standard error by $\sqrt{\text{var}(\widehat{q})}$.
(iii) Apply the classical actuarial formula (5.3) to get an estimate $\widehat{q}_{\text{act}}$ for q.
(iv) How do the two numbers $\widehat{q}$ and $\widehat{q}_{\text{act}}$ compare?

EXERCISE A.5.6. As in the Exercise A.5.5, we have

i	a_i	b_i	δ_i
1	0	1	0
2	0.25	1	0
3	0	1	1
4	0.25	0.75	1

(i) Find MLE of q under Uniform assumption, and an estimate of its standard error.
(ii) Repeat for Balducci.
(iii) Repeat for constant FOM.

EXERCISE A.5.7. Show that the MLE of q, using (5.1) and, either the Uniform or the Balducci assumption, is found by solving a polynomial equation.

A.6 Graduation testing

This chapter is a potpourri of standard statistical tests. Standard numerical exercises can be assigned, and they are of very routine nature: All you have to do is understand Example 6.9.

So, essentially, there is nothing to teach except why and how these tests work.

There are some practical rules you need to know as well. But these you do know from Statistics. For instance, in runnning the Standardised Deviations Test, you may want to chose the boxes so that the probabilities θ_i^* are equal.

EXERCISE A.6.1. Maximise the $L(\theta)$ of (6.1) over all $\theta = (\theta_1, \dots, \theta_s)$ with $\theta_1 + \cdots + \theta_s = 1$ and show that the maximiser is $\widehat{\theta}_i = N(B_i)/n$, $i = 1, \dots, s$.

EXERCISE A.6.2. In n i.i.d. fair coin tosses $\eta_1, \dots \eta_n$ (labelled + or −), let C be the number of sign changes and G the number of +groups. Let C^u be the number of sign changes from − to + and C^d the number of sign changes from + to −. Show that $|C^u - C^d| \le 1$, $G = C^u \vee C^d + \mathbf{1}(\eta_1 = +)$, and so $G \approx C/2$.

EXERCISE A.6.3. Compute the (unconditional) mean and variance of G. Using this show that $EG \approx n/4$, $\mathrm{var}(G) \approx n/16$. Observe that this is compatible with the result of the Exercise A.6.2.

EXERCISE A.6.4. Compute the (unconditional) probability $P(G = t)$.

EXERCISE A.6.5. In $10,000$ fair coin tosses (sides are labelled + or −) find: The probability that there are exactly 5000 +signs, 4000 sign changes and 2000 +groups. Is this event unlikely?

EXERCISE A.6.6. Suppose that if $Z = (Z_1, \dots, Z_n)$ is $\mathcal{N}(0, R)$ where R is a $n \times n$ covariance matrix with rank r. Let $Y = \sum_{k=1}^{n} Z_k^2/\sigma_k^2$, where $\sigma_k^2 = R_{k,k} = \mathrm{var}\, Z_k$. Find (describe an algorithm) r independent $\mathcal{N}(0,1)$ random variables $W_1, \dots, W_r$ so that $Y = W_1^2 + \cdots + W_r^2$.

EXERCISE A.6.7. A graduation has been carried out for $n = 10$ age groups and the following standardised deviations were computed:

$$-0.05,\ -0.34,\ +0.22,\ +1.34,\ -1.31,\ -0.01,\ -0.26,\ +0.24,\ +0.75,\ -0.16,$$
$$+2.05,\ -0.64,\ -1.42,\ -1.45,\ -0.91,\ +0.15,\ -0.31,\ +0.42,\ -0.88,\ -0.04$$

Suppose that 2 parameters were estimated. Carry out verious tests and draw your conclusions.

Appendix B

PROBABILITY & STATISTICS MISCELLANY

B.1 Exponentiality

A positive random variable T is *exponential* if

$$P(T > t + s \mid T > s) = P(T > t),$$

for all s and t. It follows that $P(T > t) = e^{-\lambda t}$, $t > 0$. The positive constant λ is called rate. We have $ET = 1/\lambda$, $\operatorname{var} T = 1/\lambda^2$. We allow λ to take value 0 (then $T = \infty$) or ∞ (then $T = 0$).

A Poisson process on the real line is a point process N with the property that for any disjoint sets $B_1, B_2, \ldots$, the number of points $N(B_1), N(B_2), \ldots$ are independent and $EN(B)$ is finite with mean proportional to the length $|B|$ of B. It follows that $N(B)$ *Poisson* with parameter $\lambda|B|$:

$$P(N(B) = n) = \frac{(\lambda|B|)^n e^{-\lambda|B|}}{n!}, \quad n = 0, 1, 2, \ldots$$

If we let $T_1 < T_2 < \cdots$ be the points of the point process, then $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are i.i.d. exponentials with rate λ. If we know that $N[0, t] = n$, then the n points are actually i.i.d. uniform random variables.

The distribution of T_n is called Gamma(n, λ) and has density

$$\lambda^n \frac{x^{n-1}}{\Gamma(n)} e^{-\lambda x}, \quad x > 0,$$

where $\Gamma(n) = (n-1)!$. The density is a valid density for non-integral values of n provided that we define $\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx$.

B.2 Linearity (= Normality)

A zero-mean random variable X with finite variance is said to be *normal* if for any two numbers a_1, a_2 there exists a number a such that

$$a_1 X_1 + a_2 X_2 \stackrel{\mathrm{d}}{=} aX, \tag{B.1}$$

where X_1, X_2 are i.i.d. copies of X. It follows, by taking variances on both sides, that the Pythagorean theorem

$$a^2 = a_1^2 + a_2^2$$

holds. With $i = \sqrt{-1}$, if we let $\varphi(\theta) = Ee^{i\theta X}$ be the characteristic function of X then we have, directly from (B.1),

$$\varphi(a\theta) = \varphi(a_1\theta)\varphi(a_2\theta),$$

so that, if $\psi(\theta) := \log \varphi(\sqrt{\theta})$,

$$\psi(a_1^2\theta^2 + a_2^2\theta^2) = \psi(a_1^2\theta^2) + \psi(a_2^2\theta^2).$$

By continuity, we see that the only solution to the last equation must be linear: $\psi(x) = cx$. Hence $\varphi(\theta) = e^{c\theta^2}$. But $\varphi''(0) = 2c$ and, on the other hand, $\varphi''(0) = -EX^2$, so $c = -EX^2/2$. Let σ^2 denote EX^2 and so we have

$$\varphi(\theta) = e^{-\sigma^2\theta^2/2}, \quad \theta \in \mathbb{R},$$

Standard Fourier inversion gives the density of X:

$$P(X \in dx) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} dx.$$

When $\sigma^2 = 1$ we have a standard normal random variable.

EXERCISE B.2.1. Suppose that X, Y are two independent, absolutely continuous random variables, such that the joint distribution of (X, Y) does not change under rotations (i.e. if A is a 2×2 rotation matrix then $(X, Y)' = A(X, Y)'$ has the same joint density as (X, Y)). Then show that

$$f(x, y) \propto e^{-(x^2+y^2)}.$$

A finite sequence $X_1, \ldots, X_d$ of zero-mean random variables is said to be normal if for any real numbers $c_1, \ldots, c_d$, the random variable $c_1X_1 + \cdots + c_dX_d$ is normal. This implies that

$$E \exp(i\theta^\intercal X) = \exp(i\theta^\intercal \mu - \frac{1}{2}\theta^\intercal R\theta)\,,$$

where $\theta^\intercal X := \theta_1X_1 + \cdots + \theta_dX_d$, and $\mu = (\mu_1, \ldots, \mu_d)^\intercal = (EX_1, \ldots, EX_d)^\intercal$, $R = E(X - \mu)(X - \mu)^\intercal$, i.e. R is a $d \times d$ matrix with entries $\mathrm{cov}(X_k, X_\ell)$ (the *covariance matrix*). If R is invertible, then X has a density on $\mathbb{R}^d$ given by

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(R)}} \exp\left(-\frac{1}{2}(x - \mu)^\intercal R^{-1}(x - \mu)\right). \qquad \longleftarrow \boxed{\text{The } \mathcal{N}(\mu, R) \text{ density}}$$

More generally, if R has rank r, i.e. r is the dimension of the linear space V spanned by the columns (or rows–the matrix is symmetric) of R then

$$P(X \in V) = 1,$$

and if we choose a coordinate system on V we can express the density of X on V in the above form, provided that we replace μ by its restriction $\mu|_V$ on V, and R by $R|_V$, the restriction of R on V, which is always invertible.

The vector-valued random variable $X = (X_1, \ldots, X_d)$ has $\mathcal{N}(\mu, R)$ distribution if the density is given as above. It is standard normal $\mathcal{N}(0, I)$ if $\mu = 0$, $R = I$, the $d \times d$ identity matrix.

More generally, if X belongs to a subspace V of $\mathbb{R}^d$, we say that it is standard normal if, for all $\theta \in \mathbb{R}^d$,

$$Ee^{i\theta^\intercal X} = e^{-||\theta||^2/2},$$

where $||\theta||^2 = \theta_1^2 + \cdots + \theta_d^2$.

If $(X_1, Y_1, \ldots, Y_m)$ is Normal then conditional expectation $E(X_1 \mid Y_1, \ldots, Y_m)$ is easily computed from the requirement that it is an affine function

$$E(X_1 \mid Y_1, \ldots, Y_m) = c + b_1 Y_1 + \cdots + b_m Y_m,$$

where $c = EX_1$ and the b_i are determined by the orthogonality

$$\mathrm{cov}(X_1 - (c + b_1 Y_1 + \cdots + b_m Y_m),\ Y_i) = 0, \quad i = 1, \ldots, m.$$

It should be noticed that the conditional variance is deterministic and thus easily computed. If $(X, Y) = (X_1, \ldots, X_n; Y_1, \ldots, Y_m)$ is Normal then we let $E(X \mid Y)$ be a vector with entries $E(X_i \mid Y_1, \ldots, Y_m)$ then we can determine it by applying the previous, component-by-component.*

If X is standard normal in V and V has dimension r then we say that $||X||^2$ has a χ_r^2 distribution (standard chi-squared with r degrees of freedom). Clearly, χ_r^2 is the distribution of the sum of the squares of r i.i.d. scalar standard normals with a density that is easily found to be

$$\left(\frac{1}{2}\right)^{r/2} \frac{x^{\frac{r}{2}-1}}{\Gamma(r/2)} e^{-x/2}, \quad x > 0. \qquad \longleftarrow \boxed{\text{The } \chi_r^2 \text{ density}}$$

(This is generalisable to any $r > 0$, not necessarily integer.) Clearly, $\chi_r^2 \equiv \mathrm{Gamma}(r/2, 1/2)$.

When $X_1, \ldots, X_n$ are i.i.d. normals then the sample mean $\overline{X_n} := \frac{1}{n}\sum_{i=1}^n X_i$ and the sample variance $S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X_n})$ are independent with $\overline{X_n}$ being distributed as $\mathcal{N}(\mu, \sigma^2/n)$ and S_n^2 as $\sigma^2 \cdot \chi_{n-1}^2$. (To see that S_n^2 has a *chi-squared* distribution is easy because the vector $Y := (X_1 - \overline{X_n}, \ldots, X_n - \overline{X_n})$ is Normal in $\mathbb{R}^n$. To see that it has $n-1$ degrees of freedom observe that the sum of the components of Y =0, i.e. Y lives in an $(n-1)$–dimensional subspace of $\mathbb{R}^n$. To see that $ES_n^2 = \sigma^2$ is a matter of simple computation. This, incidentally, shows that S_n^2 is an unbiased estimator of σ^2.†) The ratio

$$\frac{\sqrt{n}(\overline{X_n} - \mu)}{S_n^2}$$

has a distribution that does not depend on μ or σ^2 and has density given by

$$\nu^{-1/2} B(1/2, \nu/2)^{-1} (1 + t^2/\nu)^{-(\nu+1)/2}, \quad t > 0, \qquad \longleftarrow \boxed{\text{The } t_\nu \text{ (Student) density}}$$

with $\nu = n - 1$, and is known as the student t_ν density. The density works for any positive value of ν. The function $B(K, L)$ is defined by

$$B(K, L) = \int_0^1 x^{K-1}(1-x)^{L-1} dx = \frac{\Gamma(K)\Gamma(L)}{\Gamma(K+L)},$$

for any $K, L > 0$.

*This is the basis of multivariate linear estimation.

†This is the basis of ANOVA, the Analysis of Variance.

B.3 The Brownian motion and bridge

A collection of random variables $\{X_t\}$ is said to be normal if any finite subcollection is normal. Therefore, a normal collection of random variables[‡] is defined by means of two functions:

$$\mu(t) := EX_t, \quad r(s,t) := EX_sX_t.$$

For instance, when

$$\mu(t) = 0, \quad r(s,t) := \min(s,t), \quad s,t \geq 0,$$

we have the infamous Brownian motion. If $\{X_t, t \geq 0\}$ is a Brownian motion, then

$$B_t := X_t - tX_1, \quad 0 \leq t \leq 1$$

is, by definition, a Brownian bridge. This B, being a linear function of X, is also Gaussian (normal). Hence we just need to compute the two functions mentioned above in order to know its law completely. Clearly,

$$\begin{aligned} EB_t &= 0, \\ EB_sB_t &= E(X_s - sX_1)(X_t - tX_1) = EX_sX_t - sEX_1X_t - tEX_1X_s + stEX_1^2 \\ &= \min(s,t) - st - st + st = \min(s,t) - st \\ &= \min(s,t)[1 - \max(s,t)]. \end{aligned}$$

Via these functions, any event associated with B has, in principle, a well-defined probability. For instance, it can be shown that

$$P(\max_{0\leq t\leq 1} B_t > x) = e^{-2x^2}. \tag{B.2}$$

See, e.g., Karatzas and Shreve [11]. Also,

$$P(\max_{0\leq t\leq 1} |B_t| > x) = 2\sum_{j=1}^{\infty}(-1)^{j+1}e^{-2j^2x^2}. \tag{B.3}$$

See, e.g. Billingsley [2, eq. 11.39].

B.4 The fundamental theorem of Probability

This is the

Theorem B.1 (The Strong "Law" of Large Numbers)**.** *Suppose that $X_1, X_2, \ldots$ are independent random variables with common mean μ. Then*

$$P(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu) = 1.$$

Proof. We will ONLY prove it in a simple, but useful, case: Assume the X_i are i.i.d. and bounded. Let $S_n = X_1 + \cdots + X_n - n\mu = Y_1 + \cdots + Y_n$, where $Y_i = X_i - \mu$. Note that if the

[‡]Also known as a Gaussian process

X_i are bounded then so is S_n. Therefore all its moments exist. In particular, we compute the fourth moment:

$$ES_n^4 = E\Big(\sum_i Y_i^4 + \sum_{i,j}{}' Y_i^2 Y_j^2 + \sum_{i,j}{}' Y_i^3 Y_j + \sum_{i,j,k}{}' Y_i^2 Y_j Y_k + \sum_{i,j,k,\ell}{}' Y_i Y_j Y_k Y_\ell\Big),$$

where a prime over a sum means that it is taken over distinct values of the indices. Since $EY_i = 0$, only the first two sums survive after we take expectation. But then

$$ES_n^4 = nEY_1^4 + (n^2 - n)EY_1^2 EY_2^2 \leq Cn^2,$$

where C is a constant. But then, for any fixed $\varepsilon > 0$,

$$Cn^2 \geq ES_n^4 = ES_n^4 \mathbf{1}(|S_n| > n\varepsilon) \geq n^4\varepsilon^4 E\mathbf{1}(|S_n| > n\varepsilon) = n^4\varepsilon P(|S_n| > n\varepsilon),$$

giving that

$$P(|S_n| > n\varepsilon) \leq \frac{C}{\varepsilon n^2}.$$

The quantity $Z := \sum_{n=1}^{\infty} \mathbf{1}(|S_n| > n\varepsilon)$ is the total number of times n that S_n/n exceeds ε, in absolute value. But

$$EZ = \sum_{n=1}^{\infty} P(|S_n| > n\varepsilon) \leq \frac{C}{\varepsilon}\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty.$$

Since $EZ < \infty$ we have that $P(Z < \infty) = 1$. But this means that, no matter how small ε is, S_n/n will be between $-\varepsilon$ and ε forever after some finite time. But this means precisely that, with probability 1, $S_n/n \to 0$. □

B.5 Unbiased estimators

We here adopt the standard framework of (non-Bayesian) Statistics. There is a family of probability measures $\{P_\theta, \theta \in \Theta\}$, with Θ being an interval of $\mathbb{R}$, and each P_θ has density $p(x,\theta)$. A statistic Y is an *unbiased* estimator[§] of the parameter θ if, under the true probability measure P_θ, its expectation is equal to θ, and this is a requirement that should hold for all θ. This is expressed, of course[¶], as $\int Y(x)p(x,\theta)dx = \theta$ for all θ. The function (random variable) Y, being a statistic, does not depend on θ. We let $X(x) = x$. The random variable $p(X,\theta)$ is called *likelihood*, whereas its (natural) logarithm $\log p(X,\theta)$ is the *log-likelihood*. We shall assume that $p(x,\theta)$ as "smooth" in θ (at least twice differentiable) for all values of θ. We let D_θ denote differentiation with respect to θ and D_θ^2 twice differentiation. Notice that

$$E_\theta D_\theta \log p(X,\theta) = \int D_\theta p(x,\theta)dx = 0,$$

provided we have enough regularity in order to interchange differentiation and integral. Denote by var_θ the variance taken under the measure P_θ. The *Fisher information* is the quantity

$$I_\theta := E_\theta[D_\theta \log p(X,\theta)^2 = \mathrm{var}_\theta\, D_\theta \log p(X,\theta)$$

[§]Unbiasedness is a property that leads us to some nice theory (see below) but one should not consider it as an indispensable property without which estimators are bad; it is, simply, convenient and, sometimes, rational.

[¶]When we write $\int$ without a specification of the range we mean over the whole space.

The *Cramér-Rao inequality* states that for any unbiased estimator

$$\mathrm{var}_\theta(Y) \geq I_\theta^{-1}.$$

This is essentially the *Cauchy-Schwarz inequality*[‖]. To see this, notice that, by unbiasedness,

$$1 = D_\theta E_\theta Y = D_\theta \int Y(x)p(x,\theta)dx = \int Y(x) D_\theta p(x,\theta)dx$$

(provided we have enough regularity in order to interchange differentiation and integral). But then

$$1 = \int Y(x)(-D_\theta \log p(x,\theta))p(x,\theta)dx = \mathrm{cov}_\theta(Y, -D_\theta \log p(X,\theta)$$
$$\leq \mathrm{var}_\theta(Y)\,\mathrm{var}_\theta\, D_\theta \log p(X,\theta) = \mathrm{var}_\theta(Y)\cdot I_\theta.$$

The last inequality is equality iff $D_\theta \log p(X,\theta)$ and Y are linerarly related: $D_\theta \log p(X,\theta) = aY + b$. Taking expectations we find $0 = E_\theta D_\theta \log p(X,\theta) = aE_\theta Y + b = a\theta + b$, so $b = -a\theta$. Taking variances we find $I_\theta = \mathrm{var}_\theta\, D_\theta \log p(X,\theta) = a^2\,\mathrm{var}_\theta\, Y = a^2 I_\theta$, so $a = I_\theta$. Thus, when the lower bound is achieved we have $Y(x) = I_\theta^{-1} D_\theta \log p(X,\theta) + \theta$. (One may check that the expression on the right does not depend on θ; one then knows that the estimator is "best". Incidentally, an unbiased estimator is said to be *efficient* if its variance is close to the Cramér-Rao lower bound.)

EXERCISE B.5.1. Show that, under additional regularity, $I_\theta = -E_\theta D_\theta^2 \log p(X,\theta)$.

We point out that

$$U_\theta(X) := D_\theta \log p(X,\theta)$$

is also known as the *score*.

Note that, in what we said above, we could very well have assumed that X is a random vector in $\mathbb{R}^d$.

EXERCISE B.5.2. If $X = (X_1, \ldots, X_n)$ has i.i.d. components then the *Fisher information* for the vector X is n times the Fisher information of X_1.

To estimate, simultaneously, a parameter $\theta = (\theta_1, \ldots, \theta_s)$ ranging in a set $\Theta \subseteq \mathbb{R}^s$, we consider an unbiased estimator $Y = (Y_1, \ldots, Y_s)$, as an estimator such that $E_\theta Y = \theta$ (where this equality is to be interpreted component-wise: $E_{\theta_i} Y = \theta_i$ for $i = 1, \ldots, s$). We define the score $U_\theta(X)$ as the vector with components $D_{\theta_i} \log p(X,\theta)$ and the information matrix by

$$I_\theta := E_\theta[U_\theta(X) U_\theta(X)^\mathsf{T}].$$

The covariance matrix of Y is

$$\mathrm{cov}_\theta(Y) := E_\theta[(Y-\theta)(Y-\theta)^\mathsf{T}].$$

Note that both matrices are positive-semidefinite[**] Assume that I_θ is also invertible. The analogue of the Cramér-Rao lower bound states that the matrix $\mathrm{cov}_\theta(Y) - I_\theta^{-1}$ is positive semidefinite.[††]

[‖]This states that for any two random variables X, Y with finite variances we have $|\mathrm{cov}(X,Y)| \leq \sqrt{\mathrm{var}\, X}\sqrt{\mathrm{var}\, Y}$, with equality iff $P(aX + bY + c = 0) = 1$ for some constants a, b, c

[**]A symmetric square matrix A is positive semi-definite if $x^\mathsf{T} A x \geq 0$ for all vectors x.

[††]Notice that we can introduce a partial ordering on matrices by saying that A is below B if $B - A$ is positive semi-definite. This relation is reflexive and transitive. So, the Cramér-Rao lower bound here really says that $\mathrm{cov}_\theta(Y)$ is above I_θ in this partial order.

B.6 Maximum likelihood

A Maximum Likelihood Estimator (MLE) for a parameter θ is a statistic Y such that $p(x, Y(x)) = \max_{\theta\in\Theta} p(x,\theta)$. Under "nice" assumptions, MLEs have good properties (consistency, efficiency, ...) and are asymptotically normal. We give the ideas (=heuristics) that lead to this. Suppose that $X_1, \dots, X_n$ are i.i.d. random variables with $P_\theta(X_i \in dt) = f(t,\theta)dt$, so that $P_\theta(X_1 \in dx_1, \dots, X_n \in dx_n) = f(x_1,\theta)\cdots f(x_n,\theta)dx_1\cdots dx_n$. We let $p(x,\theta) = p(x_1,\dots,x_n,\theta) = f(x_1,\theta)\cdots f(x_n,\theta)$ denote the joint density and assume that Y_n is a MLE of θ based on these n random variables: $p(x_1,\dots,x_n,Y_n(x_1,\dots,x_n)) = \max_\theta p(x_1,\dots,x_n,\theta)$. Suppose that θ_0 is the true parameter. We argue that $Y_n(X_1,\dots,X_n) \to \theta_0$ as $n\to\infty$. Consider the log-likelihood evaluated at a parameter θ, not necessarily the true one: $\log p(X_1,\dots,X_n,\theta) = \sum_{i=1}^n \log f(X_i,\theta)$. Thus, $\log p(X_1,\dots,X_n,\theta)$ is a sum of i.i.d. random variables with common mean $E_{\theta_0}\log f(X_1,\theta)$. The strong law of large numbers says that

$$P_{\theta_0}\left(\lim_{n\to\infty}\frac{1}{n}\log p(X_1,\dots,X_n,\theta) = E_{\theta_0}\log f(X_1,\theta)\right) = 1. \tag{B.4}$$

A stronger version of the LLN, and additional assumptions, can guarantee that the limit above is uniform in θ. This in turn implies that $Y_n(X_1,\dots,X_n) = \text{Argmax}\,\frac{1}{n}\log p(X_1,\dots,X_n,\theta)$ converges to $\text{Argmax}\,\frac{1}{n}E_{\theta_0}\log f(X_1,\theta)$. Showing that the later quantity is θ_0 is a consequence of Jensen's inequality:[‡‡]

$$\text{For all } \theta, \quad E_{\theta_0}\log f(X_1,\theta) \le E_{\theta_0}\log f(X_1,\theta_0).$$

We now argue that
\begin{heuristics}

$$E_{\theta_0}Y_n(X_1,\dots,X_n) \approx \theta_0, \tag{B.5}$$

$$\text{var}_{\theta_0}\, Y_n(X_1,\dots,X_n) \approx \frac{1}{nI_{\theta_0}}, \tag{B.6}$$

where $I_{\theta_0} = E_{\theta_0}[D_{\theta_0}\log f(X_1,\theta_0)]^2$, and that

$$\sqrt{nI_{\theta_0}}\,[Y_n(X_1,\dots,X_n) - \theta_0] \quad \text{converges in law to} \quad \mathcal{N}(0,1). \tag{B.7}$$

To do this, write $x = (x_1,\dots,x_n)$ for brevity, consider the function $V(x,\theta) := D_\theta \log p(x,\theta)$ and, assuming enough smoothness, let $V'(x,\theta) = D_\theta V(x,\theta) = D_\theta^2 \log p(x,\theta)$ and Taylor-expand it around a fixed θ:

$$V(x,\eta) = V(x,\theta) + (\eta-\theta)V'(x,\theta) + R(x,\theta,\eta), \tag{B.8}$$

where $R(x,\theta,\eta) = o(|\eta-\theta|)$ as $|\eta-\theta|\to 0$ (a statement which we surely want to hold uniformly in x (assumption!)). Now recall that $\theta = Y_n(x)$ maximises the function $\theta \mapsto \log p(x,\theta)$. Assume that the usual first-order maximality conditions hold. Namely, the first-order derivative with respect to θ vanishes at θ, i.e. $V(x,Y_n(x)) = 0$. Hence, from (B.8), and a Newton-Raphson kind of argument, we can claim that

$$Y_n(X_1,\dots,X_n) - \theta_0 \approx \frac{V(X_1,\dots,X_n,\theta_0)}{-V'(X_1,\dots,X_n,\theta_0)}. \tag{B.9}$$

[‡‡]Jensen's inequality says that if g is a convex function then $Eg(W) \le g(EW)$. In the case at hand, we take $g = -\log$ and $W = f(X_1,\theta)/f(X_1,\theta_0)$ and consider E as E_{θ_0}.

For the denominator we have:

$$-\frac{1}{n}V'(X_1,\ldots,X_n,\theta_0)=\frac{1}{n}\sum_{i=1}^{n}(-D_\theta^2)\log f(X_i,\theta_0),$$

and this converges (Strong Law of Large Numbers), under the true probability P_{θ_0}, to $E_{\theta_0}(-D_\theta^2)\log f(X_1,\theta_0)$, which, by Exercise B.5.1, equals I_{θ_0}. For the numerator we have:

$$\frac{1}{\sqrt{n}}V(X_1,\ldots,X_n,\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_\theta\log f(X_i,\theta_0),$$

where the $D_\theta\log f(X_i,\theta_0)$, $i=1,\ldots,n$, are i.i.d. with mean (under P_{θ_0}) zero and variance I_{θ_0} and so (Central Limit Theorem) the distribution of the scaled numerator converges to a normal $\mathcal{N}(0,I_{\theta_0})$. The different scaling factor in numerator and denominator forces us to multiply (B.9) by $\sqrt{n}$ and, taking into account that a $\mathcal{N}(0,I_{\theta_0})$ random variable divided by I_{θ_0} is distributed as $\mathcal{N}(0,1/I_{\theta_0})$, we see that $\sqrt{n}[Y_n(X_1,\ldots,X_n)-\theta_0]$ converges in law to $\mathcal{N}(0,1/I_{\theta_0})$, which justifies the assertion (B.7).
\end{heuristics}
Incidentally, (B.4) says that the MLE is strongly *consistent* , (B.5) that it is approximately unbiased and (B.6) that it is very efficient, when n is large.

In practice, one may use the above to test the hypothesis that $\theta=\theta^*$ vs its negation.

• The z-test: If we believe that the true parameter is θ^* then $\sqrt{nI_{\theta^*}}(Y_n-\theta^*)$ should be approximately $\mathcal{N}(0,1)$–distributed. So we compute the value of $\sqrt{nI_{\theta^*}}(Y_n-\theta^*)$ from the data and, if it is not typical for the $\mathcal{N}(0,1)$ distribution, we reject the hypothesis; otherwise, we do not reject it.

• The score test: If we believe that the true parameter is θ^* then $V(X_1,\ldots,X_n,\theta^*)^2/nI_{\theta^*}$ should be approximately χ_1^2–distributed (the square of a standard normal). Again, we compute the value of $V(X_1,\ldots,X_n,\theta^*)^2/nI_{\theta^*}$ and, if it is not typical for the χ_1^2, we reject the hypothesis; otherwise, we do not reject it.

B.7 Likelihood ratios

When we test a hypothesis $\theta\in\Theta_0$ vs the alternative $\theta\notin\Theta_0$, a common-sense, but also often mathematically justified, test depends on finding a critical value for the likelihood ratio

$$\lambda(x)=\frac{\sup_{\theta\notin\Theta_0}p(x,\theta)}{\sup_{\theta\in\Theta_0}p(x,\theta)},\tag{B.10}$$

i.e. a value κ such that for a given α (the so-called *size* of the test),

$$\sup_{\theta\in\Theta_0}P_\theta\{x:\lambda(x)>\kappa\}\approx\alpha.\tag{B.11}$$

If we find such a κ, then we reject the hypothesis if the observed $\lambda(x)$ exceeds κ and, otherwise, we do not reject it. This arises from our desire to minimise a *type-II error* while keeping a *type-I error* small. A type-I error is the probability that we reject the hypothesis if it is true. A type-II error is the probability that we accept the hypothesis if is false. Thus, if R is the set of observations x that will tell us when to reject the hypothesis, the type-I error is

$$\varepsilon_{\mathrm{I}}(\theta)=P_\theta(R),\quad\theta\in\Theta_0,$$

while the type-II error is

$$\varepsilon_{\mathrm{II}}(\theta) = P_\theta(R^c), \quad \theta \notin \Theta_0.$$

Incidentally, $1 - \varepsilon_{\mathrm{II}}(\theta)$ is called the *power function* of the test. So we would like to find R so that the worst possible type-II error $\sup_{\theta \notin \Theta_0} P_\theta(R^c)$ is as small as possible, under the constraint that the worst possible type I error $\sup_{\theta \in \Theta_0} P_\theta(R)$ does not exceed α.

For instance, if Θ contains just a two points θ_0, θ_1, with the null hypothesis being $\Theta_0 = \{\theta_0\}$, and we do find a κ such that

$$P_{\theta_0}\{x : \lambda(x) = p(x, \theta_1)/p(x, \theta_0) > \kappa\} = \alpha,$$

then we set

$$R^* := \{x : \lambda(x) = p(x, \theta_1)/p(x, \theta_0) > \kappa\}$$

and if $x \in R^*$ we reject; otherwise we don't. In this case, we can see exactly why R^* is best, i.e. that

Lemma B.1 (Neyman-Pearson)**.** *Among all sets R of x such that $P_{\theta_0}(R) \leq \alpha$, we have $P_{\theta_1}(R^*) \leq P_{\theta_1}(R)$.*

Indeed, from the total probability formula,

$$P_{\theta_1}(R^*) - P_{\theta_1}(R) = P_{\theta_1}(R^* R^c) - P_{\theta_1}(R^{*c} R).$$

But $P_{\theta_1}(A) = P_{\theta_0}(A, \lambda(X))$, for any A. Hence

$$P_{\theta_1}(R^* R^c) \geq \kappa P_{\theta_0}(R^* R^c)$$

because $\lambda \geq \kappa$ on R^*. Since the opposite happens on R^{*c}, i.e. $\lambda < \kappa$ on R^{*c}, we have

$$P_{\theta_1}(R^{*c} R) \leq \kappa P_{\theta_0}(R^{*c} R).$$

Putting the last three displays together, we conclude

$$P_{\theta_1}(R^*) - P_{\theta_1}(R) \geq \kappa(P_{\theta_0}(R^* R^c) - P_{\theta_0}(R^{*c} R)) = \kappa(P_{\theta_0}(R^*) - P_{\theta_0}(R)) = \kappa(\alpha - P_{\theta_0}(R)) \geq 0.$$

In general, solving (B.11) is hard. If only we knew something about the distribution of $\lambda(X)$ we would be in better shape. But the distribution under what probability measure? Observe that (B.11) wants us to know the distribution of $\lambda(X)$ under *all* P_θ for $\theta \in \Theta_0$. So, unless Θ_0 consists of a single point, we still have a problem. However, the GREAT thing is that if $X = (X_1, \ldots, X_n)$ consists of a large number of i.i.d. observations under some P_θ for $\theta \in \Theta_0$ then the distribution of $\lambda(X)$ is, for large n, the same for all $\theta \in \Theta_0$.

Theorem B.2. *Let $\Theta \subseteq \mathbb{R}^s$ and let Θ_0 be a manifold of dimension $s - r$ for some $0 \leq r \leq s - 1$. Then, as $n \to \infty$, the law of $2 \log \lambda(X_1, \ldots, X_n)$ converges to χ^2_r.*

The ideas behind this theorem are simple. Indeed, observe that, because of the maximisations in (B.10), we are dealing with MLEs. so the asymptotic normality result for MLEs should apply. Let θ_0 be the true parameter and suppose $\theta_0 \in \Theta_0$. Then the denominator of (B.10) is maximised at the MLE Y_n for the family $p(x, \theta), \theta \in \Theta_0$. Let also $\widetilde{Y}_n$ be the maximiser of the numerator. Under reasonable assumptions, Y_n is close to $\widetilde{Y}_n$ and so they are both close to θ_0.

TO BE WRITTEN

B.8 The fundamental theorem of Statistics

Consider the empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x), \quad x \in \mathbb{R},$$

corresponding to i.i.d. random variables. From the Fundamental Theorem of Probability we have that

$$P(\lim_{n\to\infty} F_n(x) = F(x)) = 1, \quad \text{for all } x \in \mathbb{R}.$$

Indeed, the $\mathbf{1}(X_i \leq x)$, $i = 1, \ldots, n$, are i.i.d. with common mean

$$E\mathbf{1}(X_i \leq x) = P(X_i \leq x) = F(x).$$

In other words, for each x, the random variable $F_n(x)$ is a consistent estimator of the NUMBER $F(x)$. But the *Fundamental Theorem of Statistics* says something MUCH stronger:

Theorem B.3 (Glivenko-Cantelli)**.**

$$P(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0) = 1.$$

In other words, the function F_n is a consistent estimator of the FUNCTION F.

Proof. (Only for the case where F is continuous.) Suppose that F is a continuous function. It is easy to see that we have to prove that $F_n(x)$ converges to $F(x)$ uniformly for all x on an interval of positive length. Since F is continuous, it is uniformly continuous on any finite interval $[a, b]$. Therefore, for any $\varepsilon > 0$, we can find a $\delta > 0$ such that $|F(x) - F(x')| \leq \varepsilon$ for all $x, x' \in [a, b]$ with $|x - x'| \leq \delta$. Consider then two points x, x' in the interval $[a, b]$ such that $0 < x' - x < \delta$. Since F_n is an increasing function we have that $F_n(x) \leq F_n(x')$. Hence

$$F_n(x) - F(x) \leq F_n(x') - F(x) = F_n(x') - F(x') + F(x') - F(x)$$

and so

$$|F_n(x) - F(x)| \leq |F_n(x') - F(x')| + F(x') - F(x) \leq |F_n(x') - F(x')| + \varepsilon.$$

Hence if we choose an n_0 such that $|F_n(x') - F(x')| \leq \varepsilon$ for all $n \geq n_0$, we can use the same n_0 for all x smaller than x' and at distance at most δ from x' to assert that $|F_n(x) - F(x)| \leq \varepsilon$ for all $n \geq n_0$. But this is exactly uniform convergence over any interval of length δ. Clearly, this can be extended to the interval $[a, b]$ and then to the whole of $\mathbb{R}$. □

B.8.1 Ramification of the Fundamental Theorem of Statistics

We learnt that $F_n(x) - F(x)$ converges to zero, uniformly in x. But we need to know at what rate. In other words, can we multiply this difference, which converges to 0, by a certain function $a(n)$ so that $a(n)[F_n(x) - F(x)]$ has a finite limit (in some sense)? The answer is yes, if we choose $a(n) = \sqrt{n}$, as in the standard Central Limit Theorem.

Theorem B.4. *Assume that F is uniform on $[0, 1]$. The distribution of the random function $\{\sqrt{n}(F_n(x) - F(x)), 0 \leq x \leq 1\}$ converges, as $n \to \infty$, to the distribution of a Brownian bridge $\{B_x, 0 \leq x \leq 1\}$.*

Proof. See Billingsley [2].

To compute the distribution of $D_n := \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ we apply this theorem to get

Corollary B.1. *The random variable $\sqrt{n}D_n$ has a distribution which converges, as $n \to \infty$ to that of the random variable $\max_{0\le x\le 1} |B_x|$.*

The Brownian bridge was introduced in Section B.3. Therefore, we can use the approximation

$$\begin{aligned} P(D_n > k) = P(\sqrt{n}D_n > \sqrt{n}k) &\approx P(\max_{0\le x\le 1} |B_x| > \sqrt{n}k) \\ &\le 2P(\max_{0\le x\le 1} B_x > \sqrt{n}k) = 2e^{-2nk^2}, \end{aligned}$$

where the last equality is from (B.2). See Section 3.2 for an application of this.

If we do not want to use the last crude inequality then we use the exact expression for $P(\max_{0\le x\le 1} |B_x| > \sqrt{n}k)$, from (B.3):

$$\begin{aligned} P(D_n > k) = P(\sqrt{n}D_n > \sqrt{n}k) &\approx P(\max_{0\le x\le 1} |B_x| > \sqrt{n}k) \\ &= 2\sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2nk^2} = 2e^{-2nk^2} - 2e^{-8nk^2} + 2e^{-18nk^2} \pm \cdots \end{aligned}$$

Incidentally, the distribution of $\max_{0\le x\le 1} |B_x|$ is known as the *Kolmogorov distribution*.

B.9 Hypergeometric distribution and approximations

B.9.1 Sampling without replacement

Given an urn with K balls, out of which R are red and $K - R$ blue, pick k balls and ask for the probability $H(K, R, k, r)$ that you picked r red and $k - r$ blue balls. There are $\binom{K}{k}$ ways to pick our sample. The number of possible samples consisting of r red and $k-r$ blue balls is the number of ways to pick r red balls ($= \binom{R}{r}$) times the number of ways to pick $k - r$ blue balls $= \binom{K-R}{k-r}$). Thus

$$H(K, R, k, r) = \binom{K}{k}^{-1} \binom{R}{r} \binom{K-R}{k-r}. \tag{B.12}$$

So let us understand H better. To do this, consider the symbols

$$\xi_1 = \cdots = \xi_R = 1; \quad \xi_{R+1} = \cdots = \xi_K = 0.$$

Let i be a random permutation of the numbers 1 through K. Thus, i maps 1 to i_1, etc., K to i_K, and each $(i_1, \ldots, i_K)$ is a collection of K distinct integers assuming each one of the possible $K!$ values with probability $1/K!$. When we pick the balls from the urn, we pick the random variables $\xi_{i_1}, \xi_{i_2}, \ldots$. We pick k of them, i.e., we pick $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_k}$. The number of 1's amongst them is $X = \xi_{i_1} + \xi_{i_2} + \cdots + \xi_{i_k}$. Hence

$$H(K, R, k, r) = P(X = r).$$

From this we have

$$\sum_r rH(K,R,k,r) = EX = E(\xi_{i_1} + \xi_{i_2} + \cdots + \xi_{i_k}) = kE\xi_{i_1} = kR/K. \tag{B.13}$$

Here we used the fact that ξ_{i_1}, $\xi_{i_2}, \ldots$ all have the same distribution, and that $E\xi_{i_1} = P(\xi_{i_1} = 1) = R/K$. We can continue this and compute a few moments of X. For instance,

$$\begin{aligned}\sum_r r^2 H(K,R,k,r) = EX^2 &= E(\xi_{i_1} + \xi_{i_2} + \cdots + \xi_{i_k})^2 \\ &= \sum_{p,q} E\xi_{i_p}\xi_{i_q} = \sum_{p=q} E\xi_{i_p}\xi_{i_q} + \sum_{p\neq q} E\xi_{i_p}\xi_{i_q} \\ &= kE\xi_{i_1}^2 + (k^2-k)E\xi_{i_1}\xi_{i_2} \\ &= kP(\xi_{i_1}=1) + (k^2-k)P(\xi_{i_1}=1, \xi_{i_1}=2) \\ &= k\frac{R}{K} + (k^2-k)\frac{R}{K}\frac{R-1}{K-1}.\end{aligned}$$

The variance of X is

$$\operatorname{var} X = EX^2 - (EX)^2 = \frac{R}{K}\ \frac{K-R}{K}\ \frac{k(K-k)}{K-1}. \tag{B.14}$$

B.9.2 Normal approximation

Consider the situation corresponding to an urn with K balls, R of which are red. Pick k balls at random without replacement. Let $H(K,R,k,r)$ denote the probability $P(X=r)$ that our sample contains exactly r red balls. (X denotes the number of red balls in our sample.) This probability is given by formula (B.12) which is hard to apply. (For instance, think of $K = 10,000$, $R = 5,500$, $k = 1000$, $r = 900$.)

We will approximate this formula by a normal distribution, when K is large. To start with, we assume that

$$R = [sK], \quad k = [tK], \tag{B.15}$$

where $0 \le s \le t \le 1$, and $[a]$ denotes the integer part of the real number a. WARNING: We will pretend sK, tK are integers and omit the integer part.

We consider a different experiment: that of tossing i.i.d. coins with probability of heads equal to s. Let η_i be the result of the i-th toss. We let $\eta_i = 1$ if the i-th toss is heads, or $\eta_i = 0$ if tails. Thus,

$$P(\eta_i = 1) = s, \quad P(\eta_i = 0) = 1 - s, \quad i = 1, 2, \ldots$$

Let

$$S_m = \eta_1 + \cdots + \eta_m.$$

We first show that, conditionally on the event

$$A_K := \{S_K = R\},$$

the variables $(\eta_1, \ldots, \eta_K)$ behave like the variables $(\xi_{i_1}, \ldots, \xi_{i_K})$ of the previous section. This is easily seen by computing their joint probabilities. For instance, $P(\eta_1 = 1|A_K) = P(\eta_2 = 1|A_K) = \cdots = P(\eta_K|A_K)$, and, since $\sum_{i=1}^K P(\eta_1 = i|A_K) = \sum_{i=1}^K E(\eta_i|A_K) =$

$E(S_K|A_K) = R$, we have $P(\eta_1 = 1|A_K) = R/K$. We can compute $P(\eta_1 = 1, \eta_2 = 1|A_K)$ by a similar method. Observe that $E(S_K^2|A_K) = R^2$. But $E(S_K^2|A_K) = E((\sum_{i=1}^K \eta_i)^2|A_K)$ $= KE(\eta_1^2|A_K) + (K^2 - K)E(\eta_1\eta_2|A_K) = KP(\eta_1 = 1|A_K) + (K^2 - K)P(\eta_1 = \eta_2 = 1|A_K)$ $= R + (K^2 - K)P(\eta_1 = \eta_2 = 1|A_K)$. Thus, $R^2 = R + (K^2 - K)P(\eta_1 = \eta_2 = 1|A_K)$, which gives $P(\eta_1 = \eta_2 = 1|A_K) = R(R-1)/K(K-1)$, as expected. The other joint probabilities are computed in the same fashion.

Hence formulae (B.13) and (B.14) apply. Note that the variable denoted as X there has the same law as S_k given that A_K occurs. Therefore, (B.15), we have:

$$\begin{aligned} E(S_k|A_K) &= kR/K \\ \operatorname{var}(S_k|A_K) &= \frac{R}{K}\,\frac{K-R}{K}\,\frac{k(K-k)}{K-1}. \end{aligned}$$

Now take into account (B.15) to write these as

$$\begin{aligned} E(S_{tK}|A_K) &= tsK \\ \operatorname{var}(S_{tK}|A_K) &= s(1-s)t(1-t)\frac{K^2}{K-1}. \end{aligned}$$

The standard Central Limit Theorem (CLT) asserts that the distribution of $(S_m - ES_m)/\sqrt{m}$ converges to a normal distribution with mean 0 and variance equal to the variance of η_1, i.e.,

$$\operatorname{var}\eta_1 = E\eta_1^2 - (E\eta_1)^2 = s - s^2 = s(1-s).$$

We denote such a normal distribution by $\mathcal{N}(0, s(1-s))$. We let Z_s, Z_s' i.i.d. random variables with distribution $\mathcal{N}(0, s(1-s))$. We denote the aforementioned convergence by

$$(S_m - ES_m)/\sqrt{m} \Rightarrow Z_s$$

We apply the CLT as follows.

$$\begin{aligned} (S_{tK} - ES_{tK})/\sqrt{K} &\Rightarrow \sqrt{t}Z_s \\ (S_K - S_{tK} - E(S_K - S_{tK}))/\sqrt{K} &\Rightarrow \sqrt{1-t}Z_s' \end{aligned}$$

The reason that the second holds is that $S_K - S_{tK}$ has the same distribution as $S_{(1-t)K}$. What is important to realise is that, since S_{tK} and $S_K - S_{tK}$ are independent, we can ALSO assert that the joint distribution of $(S_{tK} - ES_{tK})/\sqrt{K}$ and $(S_K - S_{tK} - E(S_K - S_{tK}))/\sqrt{K}$ converges to the joint distribution of $\sqrt{t}Z_s$ and $\sqrt{1-t}Z_s'$. We write this as

$$\frac{1}{\sqrt{K}}\big(S_{tK} - tsK,\ S_K - S_{tK} - (1-t)sK)\big) \Rightarrow \big(\sqrt{t}Z_s,\ \sqrt{1-t}Z_s'\big).$$

Hence the first component, together with the sum of the two components will converge:

$$\frac{1}{\sqrt{K}}\big(S_{tK} - tsK,\ S_K - sK\big) \Rightarrow \big(\sqrt{t}Z_s,\ \sqrt{t}Z_s + \sqrt{1-t}Z_s'\big).$$

This implies that the conditional distribution of $S_{tK} - tsK$ given that $S_K - sK = 0$ (i.e., given that A_K occurs) converges to the conditional distribution of $\sqrt{t}Z_s$ given that $\sqrt{t}Z_s + \sqrt{1-t}Z_s'$:

$$\frac{1}{\sqrt{K}}\big(S_{tK} - tsK \mid A_K\big) \Rightarrow \big(\sqrt{t}Z_s \mid \sqrt{t}Z_s + \sqrt{1-t}Z_s' = 0\big). \tag{B.16}$$

It remains to read this formula. All we have to do is to compute the latter conditional distribution. But this is conditioning between two jointly normal random variables.

We do the computation separately. Let W_1, W_2 be jointly normal, both with 0 mean, and let $r_{i,j} = EW_iW_j$, $i, j = 1, 2$. It is known that the distribution of W_1 given W_2 is again normal with mean $E(W_1|W_2)$ and variance $\mathrm{var}(W_1|W_2)$. To compute the mean, we recall that $E(W_1|W_2)$ is a linear function of W_2, i.e. $E(W_1|W_2) = \alpha W_2$, where the constant α is computed from the requirement that the error $W_1 - \alpha W_2$ be independent of W_2:

$$0 = E(W_2(W_1 - \alpha W_2)) = r_{12} - \alpha r_{22}, \quad \alpha = r_{12}/r_{22}.$$

But then

$$\begin{aligned}\mathrm{var}(W_1|W_2) &= \mathrm{var}(W_1 - \alpha W_2|W_2) = \mathrm{var}(W_1 - \alpha W_2) = EW_1^2 - \alpha^2 EW_2^2 \\ &= r_{11} - (r_{12}/r_{22})^2 r_{22} = (r_{11}r_{22} - r_{12}^2)/r_{22}.\end{aligned}$$

In our case, we have $W_1 = \sqrt{t}Z_s$, $W_2 = \sqrt{t}Z_s + \sqrt{1-t}Z'_s$. So we have

$$\begin{aligned}r_{11} &= EW_1^2 = ts(1-s) \\ r_{22} &= EW_2^2 = ts(1-s) + (1-t)s(1-s) = s(1-s) \\ r_{12} &= EW_1W_2 = ts(1-s).\end{aligned}$$

and so

$$\begin{aligned}E(W_1|W_2) &= tW_2 \\ \mathrm{var}(W_1|W_2) &= t(1-t)s(1-s).\end{aligned}$$

Taking these into account, we write (B.16) as

$$\frac{1}{\sqrt{K}}\big(S_{tK} - tsK \mid A_K\big) \Rightarrow \mathcal{N}(0, s(1-s)t(1-t)).$$

For all practical purposes, we read the latter as

The distribution of $(S_{tK}|A_K)$ is $\mathcal{N}(tsK, s(1-s)t(1-t)K)$ when K is large.

Remembering that $(S_{tK}|A_K)$ has a hypergeometric distribution, we can now assert that the hypergeometric distribution converges to a normal, i.e.,

$$\sum_{r \le x} H(K, sK, tK, r) \approx \int_{-\infty}^{x} \varphi((y + st)/\sqrt{s(1-s)t(1-t)})dy, \tag{B.17}$$

with $\varphi(y) = 2\pi^{-1/2}\exp(-y^2/s)$.

Bibliography

[1] Aalen, O.O. (1978) Nonparametric inference for a family of counting processes. *Ann. Stat.* **6**, 701-726.

[2] Billingsley, P. (1968) *Convergence of Probability Measures.* Wiley

[3] Błaszczyszyn, B. and Rolski, T. (2004) *Podstawy Matematyki Ubezpieczeń na Życie* (Introduction to Life Insurance Mathematics). WNT, Warszawa

[4] Brockwell, P.J. and Davis, R.A. (1998) *Time Series: Theory and Methods.* Springer-Verlag.

[5] Cox, D.R. (1972) Regression models and life tables. *J. Royal Stat. Soc. B* **34**, 187-220.

[6] Elandt-Johnson, R.C. and Johnson, N.L. (1980) *Survival Models and Data Analysis.* Wiley

[7] Euler, L. (1767) Sur les rentes viageres. *Memoires de l'Academiei de Sciences de Berlin* **16**, 165-175. Translated as: Concerning Life Annuities (2005): `http://arxiv.org/pdf/math.HO/0502421`

[8] Fleming, T.R. and Harrington, D.P. (2005) *Counting Processes and Survival Analysis.* Wiley

[9] Gerber, H.U. and Cox, S.H. (2004) *Life Insurance Mathematics.* Springer

[10] Kaplan, E.L. and Meier, P. (1958) Nonparametric estimator from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457-481.

[11] Karatzas, I. and Shreve, S.E. (2004) *Brownian Motion and Stochastic Calculus.* Springer

[12] Liptser, R.S. and Shriyaev A.N. (2001) *Statistics of Random Processes, I: General Theory, II: Applications.* Springer

[13] Nelson, W. (1969) Hazard plotting for incomplete failure data. *J. Quart. Technol.* **1**, 27-52.

[14] Williams, D. (2001) *Weighing the Odds: A Course in Probability and Statistics.* Cambridge

[15] NIST/SEMATECH e-Handbook of Statistical Methods. `www.itl.nist.gov/div898/handbook/index.htm`

[16] StatSoft, Inc., Electronic Statistics Handbook. (2005) `www.statsoft.com/textbook/stathome.html`

Index