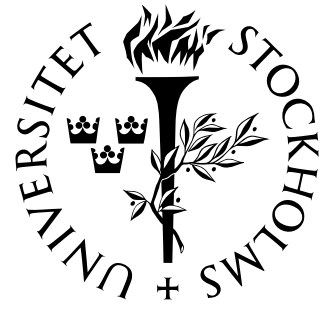


Mathematical Statistics
Stockholm University

Long-term Health Insurance

Marija Miličević

Examensarbete 2003:1

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Long-term Health Insurance

Marija Milicevic

Stockholm, 2002

Supervisor
Arne Sandström
Swedish Insurance Federation

Examiner
Mikael Andersson
Division of Mathematical Statistics
Stockholm University

Abstract

In order to maintain a profitable business of health insurance, the insurance companies must almost continuously adjust the assumptions which form the basis of estimating premiums as well as sickness reserves. These assumptions involve, among others, the probability of getting ill and remaining ill. Estimating this probability is mostly done by parametric methods. However, in recent years, the companies' financial situations suggest that these methods possibly need some adjustment.

This dissertation is done in order to examine whether an adjustment is needed and if so, to indicate the direction of this adjustment. The comparison is made between the parametric G84 and the non-parametric method based on the Nelson-Aalen estimator. Diagrams affirm the notion that the parametric method needs an adjustment as well as the fact that the adjustment will lead to higher premiums for the policyholders.

Contents

Preface

The Swedish Insurance Federation (Sveriges Försäkringsförbund) is the trade association for insurance companies active in Sweden. The Insurance Federation promotes the interests of the member companies and their possibilities to operate in Sweden and internationally by representing the industry primarily to the Government and other authorities. It also provides its members with a range of services such as statistical information among others.

This report is an outcome of my master thesis that was carried out at the Swedish Insurance Federation, under supervision of the Department of Mathematics, Division Mathematical statistics at Stockholm University, Stockholm.

I would like to thank my supervisor Arne Sandström at the Swedish Insurance Federation for his encouragement, understanding and patience, and my professor Mikael Andersson at Stockholm University for his commitment and support. Others I would like to thank are the participating insurance companies and their representatives for making this project possible.

I would also like to thank Hans Ekhult for giving the historical insight on insurance business and the staff at the Swedish Insurance Federation for making my time with them an enjoyable experience.

Marija Milicevic

marijamilicevic@hotmail.com

Stockholm, January 2003

1 Introduction

Non-cancellable, long-term, health insurance has been carried on in Sweden since the beginning of the twentieth century. It is designed to provide the right to a monthly benefit (as compensation for the reduction or loss of income) as long as the loss of working capacity due to sickness or accident is total or amounts to at least 50 per cent. Benefits are payable after the lapse of the waiting period[1] until the expiry of the total insurance period. No benefits are payable before the age of 16 and the upper age limit is usually 65 years. If partial-working capacity is at hand, the benefits payable correspond to the degree of disablement.

In order to maintain profitable business of health insurance, the insurance companies must almost continuously adjust the assumptions which form the basis of estimating premiums as well as sickness reserves. Two of the most important assumptions are the one on the probability of getting ill and the other on the combined probability of getting ill and remaining ill. These probabilities appear in calculations of both premiums and sickness reserves. Estimating the combined probability of getting ill and remaining ill is done mostly by parametric methods whereas the probability of getting ill is calculated from the actual number of ill and healthy policyholders. However, in recent years, the companies' financial situations suggest that these methods possibly need some adjustment.

This dissertation is done in order to examine whether an adjustment is needed and if so, to indicate the direction of this adjustment. Properties as well as quantity of policyholders in each of the companies form the basis of these adjustments. However, the quantity of policyholders is considered rather modest to secure statistically significant results.

Clustering all policyholders into one group and performing analysis would minimise the uncertainty in the results. Four Swedish insurance companies: Folksam, SEB Liv, Länsförsäkringar and Nordea decided to do so and finance an implementation of a program that would produce an estimation of these probabilities by both parametric and non-parametric methods and a comparison between these. Developing the program 'Sjuklighet'[2] and analysing the obtained results was the scope of this dissertation.

[1] The waiting period is usually three months but could be either shorter or longer as well as so-called floating waiting period (R-karens) of no fixed time limit implying right to a benefit as soon as the insured under the national insurance is granted disability pension.
[2] Sickliness.

A policyholder's health condition viewed by the insurance company can be described schematically with a so-called disability model:

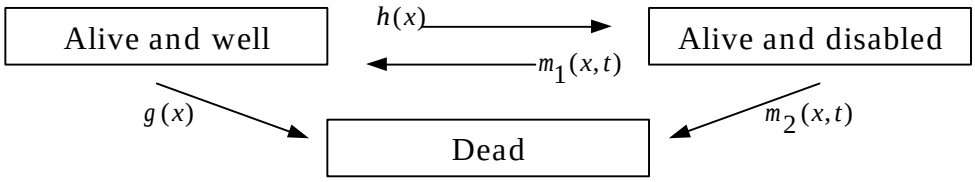


Figure 2

Where:

$h(x)$ - The force of morbidity also called the intensity of disability
$m_1(x,t)$ - The cure rate
$m_2(x,t)$ - The death intensity of disabled
$g(x)$ - The death intensity of healthy[3]

Where:

x -The age at which a person is disabled or dead
t -The time a person has been disabled (and at which he/she terminates from disability to either of the states "Alive and well" or "Dead")

Quantities of interest for the insurance companies are:

$n(x,t)$ – The ***combined*** intensity that a person will fall ill at age x and remain ill t years later. The estimates of $n(x,t)$ will be referred to as ***t-frequencies***.

$l(x,t)$ – The probability that the person who falls ill at the age x will remain ill t years later, i.e. $P(T \geq t|X = x) = 1 - P(T \leq t|X = x) = 1 - F(t|x)$ there $F(t|x)$ is the conditional cumulative distribution function. The estimate of $l(x,t)$ will be referred to as the ***termination function.***

These functions are of vital importance for the insurance business. The importance is illustrated by the following formulas:

The sickness reserves[4] for one "monetary unit":

$$a(x,t,z-x-t) = \int_t^{z-x} \frac{l(x,u)}{l(x,t)} \cdot e^{-d\cdot(u-t)} du$$

[3] Those familiar with the theory of stochastic processes note a certain resemblance with Markov processes. This is in fact a so-called semi-Markov process depending on time of disablement as well as on the duration of illness.

[4] The sickness reserve is the present value of a disability annuity to a person $(x+t)$ years old who, at the age of x, was entitled to a disability annuity expired latest at the age of z.

The single premium[5] for one "monetary unit" per annum, excluding loading:

$$E(k,x,z-x)=\int_0^{z-x-k} e^{-d\cdot t}\cdot\frac{l_{x+s}}{l_x}\cdot h_k(x+s)\cdot l(x+s,k)\cdot\int_k^{z-x-s}\frac{l(x+s,u)}{l(x+s,k)}\cdot e^{-d\cdot u}\,du\,ds$$

Where the *t*-frequency appears as a product:

$$n(x,t)=h(x)\cdot l(x,t)$$

The other terms are:

k	- The waiting period
$\frac{l_{x+s}}{l_x}$	- The probability of a x years old person to live s years later
$d=\ln(1+r)$	- The force of interest
$h_k(\cdot)$	- The intensity of disability that depends on k

1.1 *An internal relationship between* $n(x,t)$ *and* $l(x,t)$

Let $m(x,t)=m_1(x,t)+m_2(x,t)$ be the aggregated intensity[6] of termination from the state "Alive and disabled". Allowing for this assumption is the fact that the insurers are not interested in what caused the termination from the state "Alive and disabled", only that it occurred.

Further on, for arbitrary x:

$$\begin{aligned}m(t)&=\lim_{h\to 0}\frac{P(t<T\le t+h|T>t)}{h}=\lim_{h\to 0}\frac{1}{h}\cdot\frac{P(t<T<t+h,T>t)}{P(T>t)}\\&=\frac{1}{P(T>t)}\lim_{h\to 0}\frac{P(T<t+h)-P(T<t)}{h}=\frac{1}{P(T>t)}\cdot\lim_{h\to 0}\frac{F(t+h)-F(t)}{h}\\&=\frac{1}{P(T>t)}\cdot F'(t)\end{aligned}$$

According to the equation above, $m(t)dt$ can be interpreted as the probability of an instantaneous termination given that policyholder is disabled at time t.

As mentioned earlier:

$$P(T\ge t|X=x)=1-P(T\le t|X=x)=1-F(t|x)$$

$$\frac{\partial}{\partial t}l(x,t)=-\frac{\partial}{\partial t}F(t|x)$$

---

[5] Premiums are estimated as a product of the risk of loss and the amount at risk.

[6] Also called the hazard function.

This implies that for a given x:

$$m(t) = \frac{1}{P(T > t)} \cdot F'(t) = -\frac{l'(t)}{l(t)} = -\frac{d}{dt}(\ln l(t)) \Leftrightarrow l(t) = \exp\left(-\int_0^t m(s)ds\right)$$

Whereas the probability that a person x years of age will be disabled during the small time-interval dx is $h(x)dx$ [7], the combined probability of falling ill and remaining ill is:

$$n(x,t)dx = l(x,t) \cdot h(x)dx$$

2 *t-Frequencies*

The $t-$frequencies are, in the Swedish model, calculated as the ratio between the number of cases of sickness at age x with duration of at least t years, $M(x,t,I)$ and the total number of policyholders of the same age, $N(x,I)$, (both active and disabled):

$$\hat{n}(x,t) = \frac{M(x,t,I)}{N(x,I)} \qquad (I \text{ - The observed time interval})$$

It is understood that the $t-$frequencies are estimated for the values of t at which they can be observed, i.e. for $t \geq k$ where k is the waiting period.

As the time interval I can be a period of say m years, $I = (n_1, n_m)$, $N(x,I)$ can be approximated by:

$$N(x,I) \approx \frac{1}{2} \cdot N_x(n_1) + N_x(n_1 + 1) + \ldots + N_x(n_m - 1) + \frac{1}{2} \cdot N_x(n_m)$$

Where:

$N_x(n)$ - The number of persons at the age x at the time $01-01-n$.

In addition, the calculation of $N(x,I)$ can be simplified by considering the age classes X instead of single ages x. The error introduced by it is presumably small but can not be estimated.
A disadvantage of defining t-frequencies in the manner introduced above is that it depends on the composition of the portfolio. For example, in a portfolio of new policies, the number of actual cases of sickness is zero whereas a decreasing portfolio might, under exceptional conditions, be exclusively composed of disabled persons.

---

[7] This quantity shall not be regarded as a correct estimate of the transition probability but as a little less specific measure of the morbidity.

Thus, special attention should be paid to the proportion between active and disabled persons in upper age groups of older portfolios. When estimating the $t-$frequencies, one might obtain too low values to apply, for instance, to a portfolio of new policies, where the proportion between the active and disabled persons can be expected to be higher than in the older ones.

3 Termination Function $l(x,t)$

In estimating $l(x,t)$ both non-parametric and parametric methods will be used. The non-parametric method is based on the Nelson-Aalen estimator. The mathematical terms appearing in the derivation of the estimator are a multivariate counting process with the corresponding intensity process, martingales and stochastic integrals. These will be addressed in proceeding chapters.

In order to restrict the representation of the theory involved, terms to be clarified subsequently will only be those necessary in derivation of the Nelson-Aalen estimator.

3.1 The non-parametric method

The non-parametric methods are often used to provide a crude estimation of statistical terms. Advantage of the non-parametric methods over the parametric is that the non-parametric methods make direct use of the basic data and need fewer assumptions to be valid.

3.1.1 Time-To-Event Analysis

Time-to-event analysis is used in various fields for analysing data involving the duration between two events, or more generally the times of transition among several states or conditions (see Figure 2). The key characteristic that distinguishes time-to-event analysis from other areas in statistics is that time-to-event data are usually censored.

3.1.1.1 Censoring and truncation

Censoring occurs when incomplete information is available about the termination time of the observed subjects. In order to determine the time to a certain event, defining two time points is necessary: the time at which an original event occurs and the time at which the final event occurs. A subject is said to be *at risk* if the original event has occurred, but the final event has not.

The following figure describes data to be analysed later:

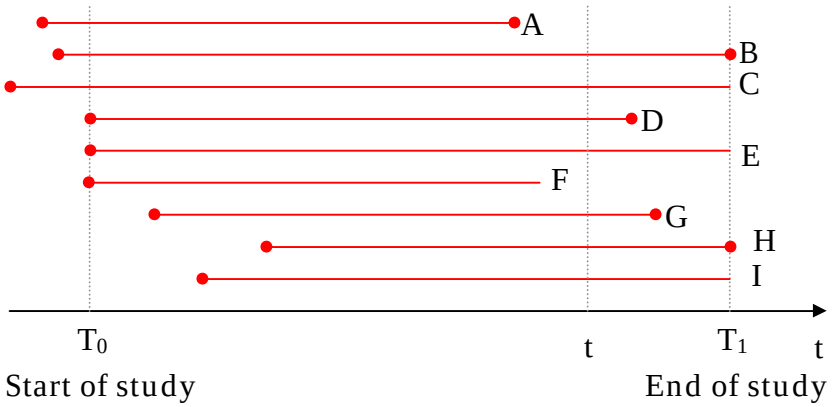


Figure 3

Obviously the time points marked “Start of study” and “End of study”, give rise to censored data. The solid lines represent the *risk period* for each subject whereas the solid points indicate an occurrence of the event of interest; in this case the beginning and the termination[8] of illness. For instance, the entire risk period for subject G falls within the observation period and the times of occurrence are known; hence there is no censoring for this observation. For subject I, the risk period starts during the observation period and the termination event occurs after follow-up is terminated at T_1. The observation of subject I is therefore right censored at T_1. By right censoring, it is meant that the risk period exceeds a certain date. Subject F is another right censored observation where the termination is caused by an event other then the one of interest. While subjects C, E, F and I represent right censoring, A, B and C represent subjects with left truncation. The left truncation means that a risk period of a certain subject is included in the set of subjects who are at risk if a specific condition is satisfied; in this case if a person has been disabled longer than the waiting period, before entering the study. Conditional on knowing the time from the beginning of the risk period to the beginning of the observation period T_0, analysis can be done with the methods developed for right censoring (such as the Nelson-Aalen estimator) with proper adjustment of the risk set. In order to correctly apply these methods some assumption about censoring must be made:

i) Independent censoring:
The risk period is independent of any mechanism that causes an individual’s time to be censored.

ii) Non-prognostic censoring:
Prognosis for an individual to be disabled at time *t* should not be affected by censoring at *t*.[9]

---

[8] In the proceedings no distinction between the causes of termination will be pointed out other than censoring (caused by expiration of either study or insurance).

[9] Independent and non-prognostic censoring models are proven to be special cases of the so-called non-informative censoring model. Its characteristic is that the instantaneous probability of termination in a small interval about y given that the subject was at risk up to y is unchanged by the additional information that the subject was uncensored up to time y.

In the case where the policyholder's duration of illness represents the risk period conditions mentioned above are satisfied.

3.1.2 *Counting Processes and the Multiplicative Intensity Model*

Suppose that there are *n* policyholders suffering from illness. For each of these, the illness duration T_i is observed, which is either the total illness duration or a censored duration[10]. In order to indicate the difference an indicator variable D_i is introduced indicating the true illness duration when $D_i = 1$, otherwise $D_i = 0$. The pair of random variables (T_i, D_i) represents now the information available on each policyholder. Moreover, independence between pairs (T_i, D_i) is assumed.

Consider a process $N_i(t) = I(T_i \le t, D_i = 1)$. It is equal to zero until a policyholder *i* recovers from illness. It changes to one, however, when recovery occurs. This kind of process is called a counting process whose formal definition is as follows:

A counting process $\{N_i(t), 0 \le t < \infty\}$ is a stochastic process that counts the occurrences as time *t* proceeds. It has the following properties:

1) $N_i(0) = 0$
2) $P(N_i(t) < \infty) = 1$
3) The sample paths are right-continuous and piecewise constant with jump of size +1.

A multivariate counting process $N = \{N_1(t), N_2(t), ..., N_n(t)\}$ possesses obviously the same properties as those mentioned above. Moreover, no two component processes are assumed to jump simultaneously, which follows as a consequence of the assumption of the continuity of time *t*.

Imagine now taking a walk along the time axis in Figure 3. At any time *t* (and looking back) you would know whether subject *i* has been observed to terminate from illness (e.g. subject A), been censored (e.g. subject F) or is still suffering from illness and uncensored (e.g. subjects B, C, D, E, G, H and I). This accumulated knowledge about what has happened to, in our case, policyholders up to, but not including *t* is called the *history* or *filtration* of the counting process and is denoted F_{t-}. It is self-evident that $F_s \subseteq F_t$, whenever $s \le t$, that is as time proceeds more and more is known about the population. Let I_{dt} be a small time interval of length *dt* around time *t*. For the first two cases (i.e. subjects A and F) the conditional probability of observing $N_i(t)$ to change from 0 to 1[11] in the interval I_{dt} is 0. For the latter cases (i.e.

[10] Illness duration up to a closing time.
[11] That is observing the true illness duration.

subjects B, C, D, E, G, H and I) this conditional probability is $m_i(t)dt$ [12] because those still suffering from illness and uncensored are at risk of making a transition from being ill to either being well or dead. Define now a function $Y_i(t) = I(T_i \geq t \geq wp)$ where *wp* stands for the waiting period[13]. This function indicates whether subject *i* is at risk of making the transition or not. Denoting the increment of $N_i(t)$ at *t* as $dN_i(t)$ the conditional probability of observing $N_i(t)$ to change from 0 to 1 in the interval I_{dt} can now be written as:

$$P(dN_i(t) = 1 | F_{t-}) = m_i(t) \cdot Y_i(t)dt$$

Define:

$$a_i(t) = m_i(t) \cdot Y_i(t)$$

Then:

$$P(dN_i(t) = 1 | F_{t-}) = a_i(t)dt$$

Where $a_i(t)$ represents the intensity process of the counting process $N_i(t)$.

A multivariate counting process $N(t) = \{N_1(t), N_2(t),..., N_n(t)\}$ has the intensity process of the form $a(t) = \{a_1(t), a_2(t),..., a_n(t)\}$.

Assuming that the population of policyholders is homogeneous, implying $m_1(t) = m_2(t) = ... = m_n(t)$, the formula given above can be rewritten:

$$a_i(t) = m(t) \cdot Y_i(t)$$

Where $m(t)$ denotes the common value.

Aggregating the individual counting processes $N_1(t), N_2(t),..., N_n(t)$ would produce a so-called univariate counting process, which counts the total number of observed terminations in [0, t], i.e.:

$$N(t) = \sum_{i=1}^{n} N_i(t)$$

Its intensity process is given by:

---

[12] See page 3.

[13] See page 1. By defining $Y_i(t)$ in such a manner the issue of left truncation has been taking care of.

$$a(t) = \sum_{i=1}^{n} a_i(t) = \sum_{i=1}^{n} m(t) \cdot Y_i(t) = m(t) \cdot \sum_{i=1}^{n} Y_i(t) = m(t) \cdot Y(t)$$

Where $Y(t) = \sum_{i=1}^{n} Y_i(t)$ counts the total number of, in our case policyholders at risk for transition at an instant just prior to time t.

This approach was introduced by Odd Aalen in 1978 and labelled a multiplicative intensity model for counting processes where the intensity process is given by $a(t) = m(t) \cdot Y(t)$. Properties assumed in this model regarding $m(t)$ are that it is a non-negative deterministic function whereas $Y(t)$ is a non-negative observable stochastic process. This kind of processes is even called *predictable* processes implying that knowing the history of the process up to time t determines its value at t.

3.1.3 *Martingales and Stochastic Integrals*

Let the increment of $N(t)$ at t:

$$dN(t) = N((t + dt)-) - N(t-)$$

This random variable can only take values 0 and 1[14] so taking the conditional expectation gives:

$$E\left[dN(t) \middle| F_{t-}\right] = a(t)dt$$

This equality is valid for any counting process. The intensity process defined in such a manner is characterised by the fact that:

$$M(t) = N(t) - \mathrm{A}(t) \qquad \left[\Leftrightarrow dM(t) = dN(t) - d\mathrm{A}(t) = dN(t) - a(t)dt\right]$$

Where $M(t)$ is a *counting process martingale* and $\mathrm{A}(t)$ is a *cumulative intensity process,* even called *compensator* of the counting process and defined as:

$$\mathrm{A}(t) = \int_0^t a(s)ds\,, \quad t \geq 0$$

One of the defining properties of the counting process martingale is that:

$$E\left[dM(t) \middle| F_{t-}\right] = E\left[dN(t) - d\mathrm{A}(t) \middle| F_{t-}\right] = E\left[dN(t) - a(t)dt \middle| F_{t-}\right] = E\left[dN(t) \middle| F_{t-}\right] - a(t)dt = 0$$

[14] No two component processes are assumed to jump simultaneously.

The last equality is due to the fact mentioned earlier; $a(t) = m(t) \cdot Y(t)$ is a predictable process through the dependence on the predictable process $Y(t)$ and therefore non-random.

A martingale can be considered as being a pure noise process. The systematic part of the counting process is its compensator: a smoothly varying and predictable process, which, subtracted from the counting process, leaves unpredictable zero-mean noise.

The conditional variance of the increment $dM(t)$ is:

$$Var\left(dM(t)|F_{t-}\right) = E\left[(dM(t))^2|F_{t-}\right] - \left(E\left[dM(t)|F_{t-}\right]\right)^2 = E\left[(dM(t))^2|F_{t-}\right]$$

A closer look at the equation above suggests that the conditional variance of increment of M is the increments of the compensator of another process namely M^2. In order to show this let the increment of M^2:

$$d\left(M^2\right)(t) = M\left((t+dt)-\right)^2 - M(t-)^2 = \left(M(t-) + dM(t)\right)^2 - M(t-)^2$$

$$= 2 \cdot M(t-) \cdot dM(t) + \left(dM(t)\right)^2$$

Taking the expectation:

$$E\left[d\left(M^2\right)(t)|F_{t-}\right] = E\left[2 \cdot M(t-) \cdot dM(t) + \left(dM(t)\right)^2|F_{t-}\right] =$$
$$2 \cdot M(t-) \cdot E\left[dM(t)|F_{t-}\right] + E\left[(dM(t))^2|F_{t-}\right] = E\left[(dM(t))^2|F_{t-}\right] = Var\left(dM(t)|F_{t-}\right) = d\langle M\rangle(t)$$

This is known as *M 's predictable variation process* and is denoted by $\langle M\rangle$. Although M is a pure noise process, M^2 has a tendency to increase over time.

Determining the conditional variance:

$$Var\left(dM(t)|F_{t-}\right) = E\left[(dM(t))^2|F_{t-}\right] = E\left[(dN(t) - a(t)dt)^2|F_{t-}\right] =$$
$$E\left[(dN(t))^2|F_{t-}\right] - 2 \cdot a(t)dt \cdot E\left[dN(t)|F_{t-}\right] + (a(t)dt)^2 =$$
$$a(t)dt - 2 \cdot a(t)dt \cdot a(t)dt + (a(t)dt)^2 = a(t)dt - (a(t)dt)^2 = a(t)dt \cdot (1 - a(t)dt) \approx a(t)dt$$

When there are ties in data the approximation in the last step will not hold.

The result:

$$\langle M\rangle(t) = \int_0^t a(s)ds$$

Another concept needed to introduce is the *predictable covariation process* defined by having the increments:

$$d\langle M_i, M_j \rangle(t) = Cov\{dM_i(t), dM_j(t) | F_{t-}\}$$

Two martingales are said to be orthogonal if $\langle M_i, M_j \rangle(t) = 0$.

For martingales $dM_i(t) = dN_i(t) - a_i(t)dt$ and $dM_j(t) = dN_j(t) - a_j(t)dt$ the predictable covariation process is found to be:

$$\begin{aligned} d\langle M_i, M_j \rangle(t) &= E[(dN_i(t) - a_i(t)dt)(dN_j(t) - a_j(t)dt) | F_{t-}] \\ &= E[dN_i(t) \cdot dN_j(t) | F_{t-}] - a_i(t)dt \cdot a_j(t)dt \approx 0 \end{aligned}$$

The approximation in the last step is due to the fact that $N_i(t)$ and $N_j(t)$ never jump simultaneously. Apparently these martingales are orthogonal.

In order to be able to derive the Nelson-Aalen estimator, the definition of stochastic integrals is necessary.

The stochastic integral, $\int_0^t X(s)dY(s)$, i.e. the integration of one stochastic process, $X(s)$, with respect to another, $dY(t)$, is considered here to be a pathwise operation: for a given event $w \in \Omega$ [15], one forms an ordinary Lebesque-Stieltjes integral[16] over a given time interval under the condition that $\int_0^t |X(s)||dY(t)| < \infty$.

Assume now that H(t) is a predictable process[17]. The stochastic integral of such a process with respect to a martingale:

$$W(t) = \int_0^t \mathrm{H}(s)dM(s)$$

This is a martingale itself because:

$$E[\mathrm{H}(t)\,dM\,(t) | F_{t-}] = \mathrm{H}(t) \cdot E[dM\,(t) | F_{t-}] = 0$$

[15] The set of all possible outcomes.
[16] For further information on Lebesque-Stieltjes integral see http://mathworld.wolfram.com/.
[17] See page 9.

The predictable variation process is now easily found to be:

$$Var\left(\mathrm{H}(t)dM(t)\middle|F_{t-}\right)=\mathrm{H}^2(t)\cdot Var\left(dM(t)\middle|F_{t-}\right)=\mathrm{H}^2(t)\cdot d\langle M\rangle(t)$$

So:

$$\langle W\rangle(t)=\int_0^t \mathrm{H}^2(s)d\langle M\rangle(s)$$

3.1.4 *The Nelson-Aalen estimator*

Recall:

$$dN(t)=a(t)dt+dM(t)\approx m(t)\cdot Y(t)dt+"noise"$$

If the whole equation is divided by $Y(t)$:

$$\frac{dN(t)}{Y(t)}=m(t)dt+\frac{dM(t)}{Y(t)}$$

And integrated:

$$\int_0^t \frac{dN(s)}{Y(s)}=\int_0^t m(s)ds+\int_0^t \frac{dM(s)}{Y(s)}$$

In order to deal with the fact that $Y(t)$ could be 0 at times, an indicator variable is introduced:

$$J(t)=I\left(Y(t)>0\right)$$

Taking that into account and defining $\frac{0}{0}=0$ gives:

$$\int_0^t \frac{J(s)}{Y(s)}\cdot dN(s)=\int_0^t J(s)\cdot m(s)ds+\int_0^t \frac{J(s)}{Y(s)}\cdot dM(s)$$

Recall that:

$$l(t)\ =\exp\left(-\int_0^t m(s)ds\right)$$

If the value of $\int_0^t m(s)ds$ is estimated so is the value of $l(t)$.

Rewriting the summation above:

$$\int_0^t J(s)\cdot m(s)ds = \int_0^t \frac{J(s)}{Y(s)}\cdot dN(s) - \int_0^t \frac{J(s)}{Y(s)}\cdot dM(s)$$

The term on the left is essentially the same as the integration that determines the termination function $l(t)$. They are equal in the range where there are observations.

The first term on the right is known as the Nelson-Aalen estimator of $\int_0^t m(s)ds$. It is calculated as a single sum. To see that let $t_1 < t_2 < t_3 < \ldots$ be successive jump-times for $N(t)$ implying that $dN(t)$ is equal one when t equal any of the jumps-times and zero otherwise. Hence $\int_0^t \frac{J(s)}{Y(s)}\cdot dN(s)$ can be rewritten as $\sum_{(i:t_i \le t)} \frac{1}{Y(t_i)}$. This is an increasing and right-continuous step-function.

$\int_0^t \frac{J(s)}{Y(s)}\cdot dM(s)$ is a stochastic integral with respect to a martingale and therefore a martingale itself. As a consequence $\int_0^t \frac{J(s)}{Y(s)}\cdot dN(s)$ is an unbiased estimator of $\int_0^t J(s)\cdot m(s)ds$ meaning that $E\left[\int_0^t \frac{J(s)}{Y(s)}\cdot dN(s)\right] = E\left[\int_0^t J(s)\cdot m(s)ds\right]$.

Denote $\int_0^t \frac{J(s)}{Y(s)}\cdot dM(s) = W(t)$. The predictable variation process of $W(t)$ is easily found using formulas developed for stochastic integrals:

$$\langle W\rangle(t) = \int_0^t \left(\frac{J(s)}{Y(s)}\right)^2 \cdot d\langle M\rangle(s) = \int_0^t \left(\frac{J(s)}{Y(s)}\right)^2 \cdot a(s)ds = \int_0^t \left(\frac{J(s)}{Y(s)}\right)^2 \cdot m(s)\cdot Y(s)ds = \int_0^t \frac{J(s)}{Y(s)}\cdot m(s)ds$$

3.1.4.1 *Some asymptotic results*

Let $Z^{(n)}(t) = \sqrt{n}\cdot W(t) = \sqrt{n}\cdot\left(\int_0^t \frac{J(s)}{Y(s)}\cdot dN(s) - \int_0^t J(s)\cdot m(s)ds\right) = \sqrt{n}\cdot\int_0^t \frac{J(s)}{Y(s)}dM(s)$. For large samples this expression is very close to $\sqrt{n}\cdot\left(\int_0^t \frac{J(s)}{Y(s)}\cdot dN(s) - \int_0^t m(s)ds\right)$. Conditional variance of the increments of $Z^{(n)}(t)$ is:

$$Var\left(dZ^{(n)}(t)|F_{t-}\right)=n\cdot Var\left(dW(t)|F_{t-}\right)=n\cdot Var\left(\frac{dM(t)}{Y(t)}|F_{t-}\right)=n\cdot\frac{d\langle M\rangle(t)}{Y^2(t)}$$
$$\approx n\cdot\frac{a(t)dt}{Y^2(t)}=n\cdot\frac{m(t)\cdot Y(t)dt}{Y^2(t)}=n\cdot\frac{m(t)dt}{Y(t)}=\frac{m(t)dt}{\frac{Y(t)}{n}}$$

According to the law of large numbers, the random variation of $\frac{Y(t)}{n}$ should be small for large n. In other words $\frac{Y(t)}{n}\to y(t)$ when $n\to\infty$ where $y(t)$ is a deterministic function.

So $Var\left(dZ^{(n)}(t)|F_{t-}\right)\approx\frac{m(t)dt}{y(t)}\Leftrightarrow\left\langle Z^{(n)}\right\rangle(t)\approx\int_0^t\frac{m(s)ds}{y(s)}$ for large samples.

Further on, $W_i(t)$ and $W_j(t)$ are orthogonal as a consequence of $M_i(t)$ and $M_j(t)$ being orthogonal. This implies uncorrelated increments for the processes $W_i(t)$, $i=1,2,...,n$ as well as $W_i(t)$ and $W_j(s)$ for any t,s and $i\neq j$.

Another property of large samples is that $Z^{(n)}(t)$ will have many jumps but all of these will be of order $\frac{1}{\sqrt{n}}$.

These characteristics; a deterministic predictable variation process and continuous sample paths are encountered in only one limiting process $Z^{(\infty)}(t)$ and that is the continuous Gaussian martingale. Other characteristics are independent increments[18] and normally distributed finite-dimensional distributions[19]. This basic convergence allows determining the confidence intervals since $Z^{(n)}(t)$ will have an approximate normal distribution with mean 0 and variance:

$$s^2\left(Z^{(\infty)}(t)\right)=\int_0^t\frac{m(s)ds}{y(s)}$$

An estimate of the variance is obtained from:

$$\langle W\rangle(t)=\int_0^t\frac{J(s)}{Y(s)}\cdot m(s)ds$$

[18] For any set of disjunctive intervals $(t_{i-1},t_i),i=1,2,...,k$ the random variables $Z^{(\infty)}(t_i)-Z^{(\infty)}(t_{i-1})$ are independent.

[19] Joint distribution of $\left[Z^{(\infty)}(t_1),...,Z^{(\infty)}(t_k)\right]$ is multivariate normal for any value of k.

Letting $m(t)dt = \frac{dN(t)}{Y(t)}$ gives:

$$\langle \hat{W} \rangle(t) = \int_0^t \frac{J(s)}{Y(s)} \cdot m(s)ds = \int_0^t \frac{J(s)}{Y^2(s)} dN(s)$$

And the difference:

$$\langle \hat{W} \rangle(t) - \langle W \rangle(t) = \int_0^t \frac{J(s)}{Y^2(s)} dN(s) - \int_0^t \frac{J(s)}{Y(s)} \cdot m(s)ds = \int_0^t \frac{J(s)}{Y^2(s)} \cdot (dN(s) - m(s) \cdot Y(s)ds)$$
$$= \int_0^t \frac{J(s)}{Y^2(s)} \cdot (dN(s) - a(s)ds) = \int_0^t \frac{J(s)}{Y^2(s)} dM(s)$$

Evidently difference produces a martingale which implies that $\langle \hat{W} \rangle(t)$ is an unbiased estimator of $\langle W \rangle(t)$. It is calculated as a single sum. To see that let $t_1 < t_2 < t_3 < \ldots$ be successive jump-times for $N(t)$ implying that $dN(t)$ is equal to one when t equals any of the jumps-times and zero otherwise. Hence $\int_0^t \frac{J(s)}{Y^2(s)} dN(s)$ can be rewritten as $\sum_{(i:t_i \le t)} \frac{1}{Y^2(t_i)}$.

3.2 *The parametric method*

The parametric method used to estimate the termination function is called G84. It is based on another parametric method known as G73. Development of these methods, i.e. estimation of coefficients in the equations, is based on long-term experience of health insurance. G84 takes two parameters, x-age at which the policyholder got ill and t-illness duration. Formulas presented next differ somewhat among the insurance companies but the overall structure is (hopefully) preserved.

G84:

$$l_{G84}(x,t) = \begin{cases} l_{G73}(x,t) & \text{if } t \le J(x) \\ l_{G73}(x, J(x)) \cdot \frac{l(t)}{l(J(x))} & \text{if } t > J(x) \end{cases}$$

G73:

$$l_{G73}(x,t) = a(x)e^{-80t} + b(x)e^{-13t} + c(x)e^{-1.5t} + d(x)\left(0.15e^{-0.3t} + 0.85e^{-0.04t}\right)$$

Where:

$$a(x) = 1 - b(x) - c(x) - d(x)$$
$$b(x) = 0.12$$
$$c(x) = 0.006e^{0.04x}$$
$$d(x) = 0.001 + 0.000011e^{0.13x}$$

G73 and consequently G84 distinguish between genders:

Men:

$$J(x) = \begin{cases} 2.5 & \text{if } 0 \le x < 30 \\ 2.5 - 0.07 \cdot (x - 30) & \text{if } 30 \le x < 55 \\ 0.75 & \text{if } x \ge 55 \end{cases}$$

And:

$$l(t) = 0.15e^{-0.3t} + 0.85e^{-0.03t}$$

Women:

$$J(x) = \begin{cases} 2.25 & \text{if } 0 \le x < 30 \\ 2.25 - 0.06 \cdot (x - 30) & \text{if } 30 \le x < 55 \\ 0.75 & \text{if } x \ge 55 \end{cases}$$

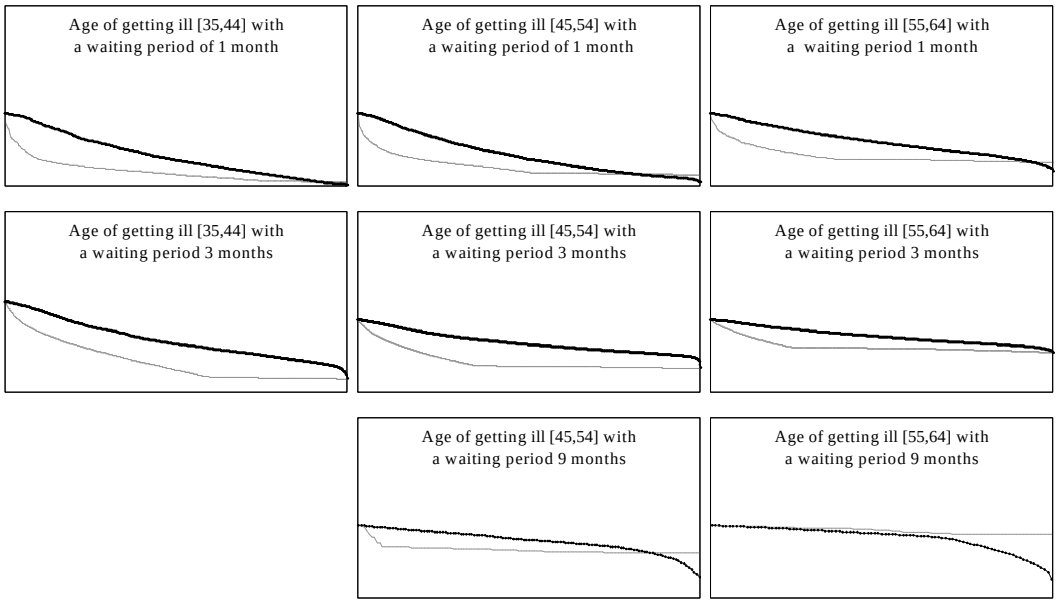
And:

$$l(t) = 0.15e^{-0.3t} + 0.85e^{-0.015t}$$

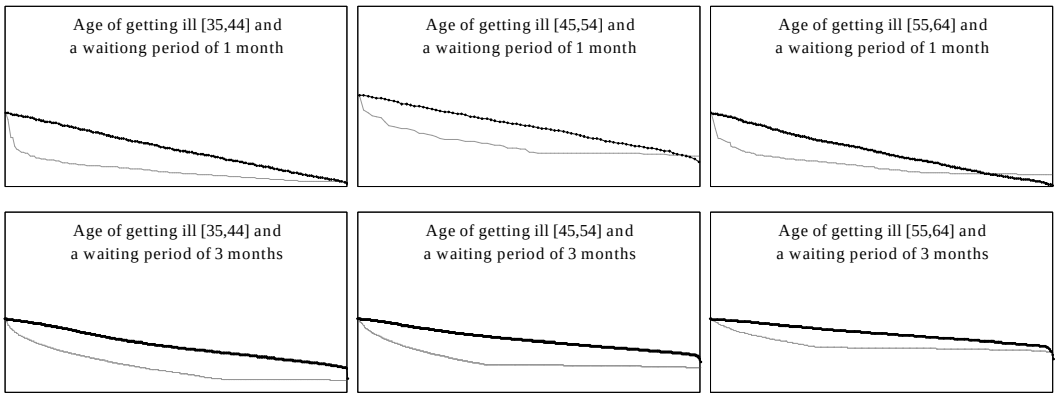
4 The results

As mentioned earlier the purpose of this paper is to compare the non-parametric with the parametric method and if possible draw some conclusions. Unfortunately although highly comprehensible, the participating companies were not enthusiastic about presenting the obtained results. In order to present anything of significance it was agreed that the diagrams presented would not contain scales on axes. Although a constant is added to each of the curves their mutual relationship is preserved. Black lines represent the non-parametric whereas grey the parametric estimation of the termination function.

Termination functions for men:



Termination functions for women:



The termination functions in the diagrams are based on a sample accumulated through the filtration according to gender, age of disablement and a waiting period implying that the curves presented are just functions of one variable namely t.

The termination function is a continuous, monotonically decreasing function. As it represents the probability of getting and remaining ill for a certain period of time it could be neither larger than one nor less then zero.

There are considerable differences between the curves. By definition both start at one but depart shortly afterwards. The largest discrepancy seems to occur for both men and women who had a waiting period of 1 month. The discrepancy for those who had a waiting period of 3 months, both men and women, is slightly less than for those who had a waiting period of 1 month. The peak of the discrepancies for those with a waiting period of 3 months is shifted to the left of the peak for those with a waiting period of 1 month. For men with a waiting period of 9 months the discrepancy almost

vanishes. In all cases the termination function estimated by Nelson-Aalen is greater than G84 implying that calculated premiums for policyholders who form the basis for this analysis are to low as well as the sickness reserves necessary to cover all the claims.

However, caution is well-found in analysing these results. Unfortunately, the number of observations is another issue not to be discussed. It is however appropriate to mention that the number of observations would not (necessarily) imply statistical significance.

5 Conclusion

Examination of the results affirms the notion that the parametric method needs an adjustment. As the non-parametric method gave consistently greater estimates than the parametric method used in most insurance companies, adjustments to be made will certainly lead to higher premiums for the policyholders and larger sickness reserves for the insurance companies.

Modelling the parametric method to fit the observation was outside the scope of this project. However, there are several methods to do that and some are examined in resent dissertations at the Department of Mathematics, Division Mathematical statistics at Stockholm University.

References

New Bases for Non-cancellable Sickness Insurance in Sweden
by Carl-Gösta Dillner
Skand. ActuarTidskr. 1969, Vol.40

New Bases for Long Term Sickness Insurance in Sweden from 1973
by Carl-Gösta Dillner
Skand. Actuarial J. 1974, Vol.4

Censoring Issues in Survival Analysis
by Kwan-Moon Leung, Robert M. Elashoff, and Abdelmonem A. Afifi Annu
Rev. Public Health. 1997, Vol. 18

Survival Analysis: Techniques for Censored and Truncated Data
by John P. Klein, M. L. Moeschberger
Springer Verlag 1997

Statistical Models Based on Counting Processes
by Per Kragh Andersen, Ornulf Borgan, Richard D. Gill, Niels Keiding
Springer Verlag 1995

Counting Process Models for Life History Data: A Review
by Per Kragh Andersen and Ornulf Borgan
Skandinavian Journal of Statistics 1985, Vol. 12

An Empirical Study of t-Frequencies and Termination Functions in Long-Term Sickness Insurance in Sweden
by Arne Sandström
The Research Council of Actuarial Science 1987, Meddelande nr 71