Actuarial Statistics - II

Introduction

In this section, we study claim models, survival function and Life Tables. The individual claim model and the sum of the claims of many insured individuals are discussed. With the assumption that the claims of individuals are assumed to be independent, we obtain the probability distributions of these factors. This will be useful in assessing claims that involve one or many components. Survival function is a useful function in the study of lifetime of any individual or any objects that will expire in due course of time. Using survival distribution various other functions are obtained. They are, Force of Mortality, probability of an individual lives until a specified age, probability of death between two age points, etc. Life Table is a table which shows, for each age, what the probability is that a person of that age will die before completion of that age. Various other components of this table are discussed. This table is very useful in the study of actuarial premium calculation.

Models for individual claims and their sums

The term general insurance essentially applies to an insurance risk that is not a life insurance or health insurance risk, and so the term covers familiar forms of personal insurance such as motor vehicle insurance, home and contents insurance, and travel insurance. Let us focus on how a motor vehicle insurance policy typically operates from an insurer's point of view. Under such a policy, the insured party pays an amount of money (the premium) to the insurer at the start of the period of insurance cover, which we assume to be one year. The insured party will make a claim under the insurance policy each time the insured party has an accident during the year that results in damage to the motor vehicle, and hence requires repair costs. There are two sources of uncertainty for the insurer: how many claims will the insured party make, and, if claims are made, what will be the amounts of those claims? Thus, if the insurer were to build a probabilistic model to represent its claims outgo under the policy, the model would require a component that modelled the number of claims and another that modelled the amounts of those claims. This is a general framework that applies to modelling claims outgo under any general insurance policy, not just motor vehicle insurance, and we will describe these. However let us begin with assumption that there will be only one claim if any in the given period. That is, during the period there could be one claim of amount b with probability q and there is no claim with probability 1-q. The claim random variable, X, has a probability function given by

$$f_X(x) = \begin{cases} 1-q & for\ x = 0 \\ q & for\ x = b \\ 0 & elsewhere, \end{cases}$$

The expected value of the claim is E[X]= $b\,q$ and E[X^2]= $b^2\,q$ hence, Var[X]= $b^2\,q(1\text{-}q)$.
The random variable X can also be written as X = I b where I is a Bernoulli random variable which takes 0 and 1 as values is can take. We also refer it as an indicator function, because it indicates the occurrence, I=1, or nonoccurrence, I =0 of given event or claim.
We seek more general models in which the amount of claim is also a random variable and several claims can occur in a period. Health, automobile, and other property and liability coverages provide immediate examples. Here we postulate $X = I\,B$, where X is the claim number random variable for the period, B gives the total claim amount incurred during the period. and I is the indicator for the event that at least one claim has occurred.

Sometimes the claim random variable could be mixed random variable. Also claim amount need not be fixed, it could vary continuously.

Consider an example of amount of claim is a random variable. If maximum claim is Rs 2,00,000 which occurs with probability 0.1 and no claim occurs with probability 0.5. Claim amount is positive but less than 2,00,000 with probability 0.4 and the distribution of the claim has the following conditional distribution function

$$F_1(x) = 1 - \left(1 - \frac{x}{200000}\right)^2 \quad for \quad 0 < x < 2{,}00{,}000.$$

Here the claim distribution can be written as

$F(x)$ = Pr($X \leq x$) = Pr(Claim $\leq x$)

= Pr(Claim $\leq x$ | No claim) P(No claim)

+ Pr(Claim $\leq x$ | positive claim but less than 2,00,000) Pr (Positive claim but less than 2,00,000) + Pr(Claim $\leq x$ | Full claim) P(Full claim).

Hence the claim distribution function is

$$F(x) = \begin{cases} 0 & for \quad x < 0 \\ 0.5 + 0.4\left[1 - \left(1 - \frac{x}{200000}\right)^2\right] & for \quad 0 \leq x < 2{,}00{,}000 \\ 1 & for \quad x \geq 2{,}00{,}000. \end{cases}$$

Here we observe that P(X=0) = P(No claim) = $F(0) = 0.5$,
and P(X=2,00,000)=P(Full claim)= $F(2{,}00{,}000)$-$F(2{,}00{,}000-)$=1-0.9 = 0.1,
P(Claim amount is at most 50000) = $F(50000) = 0.5+0.4[1-(50000/200000)^2] = 0.875$,
P(Claim amount is at least 100000) = 1-$F(100000)$ = 1-{0.5+0.4[1-1/4]}=0.2.

Expected claim amount is

$$E[X] = 0 \times P[X = 0] + 0.4 \times \int_0^{200000} x \, \frac{2}{200000}\left(1 - \frac{x}{200000}\right) dx$$

$$+200000 \times P[X = 200000]$$

$$E[X] = 0.4 \times 66{,}666.67 + 200000 \times 0.1$$

= 46,666.67

Sums of independent random variables

In the individual risk model, claims of an insuring organization re modelled as the sum of the claims of many insured individuals. The claims of individuals are assumed to be independent

in most applications. The probability distribution of sum of random variables can be obtained by the method of convolution.

Let X and Y are two individual claims with distribution function $F_X(.)$ and $F_Y(.)$ respectively and are independently distributed. Then the distribution function or cumulative distribution function (cdf) of X+Y is

$$F_{X+Y}(s) = \int_{-\infty}^{\infty} F_Y(s-x)\, dF_X(x) = F_X * F_Y(s)$$

is called the convolution of the cdfs of $F_X(.)$ and $F_Y(.)$.
Another approach is to use moment generating functions whenever they exist.

Suppose that X follows Uniform(0,1) and Y follows Uniform(0,2) and are independent. Then the distribution function of cdf of $X+Y$:

$$F_{X+Y}(s) = \frac{1}{4}\, s^2\, I_{[0,1)}(s) + \frac{1}{4}\,(2s-1) I_{[1,2)}(s) + \left[1 - \frac{1}{4}\,(3-s)^2\right] I_{[2,3)}(s).$$

The sum can be extended to any number of fixed random variables.

For example, if X, Y and Z are independent exponential random variables with parameters 1, 2, and 3 respectively. Further, they are independent. Then the probability density function of X+Y+Z is

$$f(s) = 3\, e^{-s} - 6\, e^{-2s} + 3\, e^{-3s} \quad for\ \ x > 0.$$

If X and Y are discrete random variables with probability functions or pmfs $p_X(.)$ and $p_Y(.)$ respectively. We find for the cdf of $X+Y$ and the corresponding density through convolution is

$$p_{X+Y}(s) = \sum_{x=0}^{s} p_Y(s-x) p_X(x)$$

Let X follow Poisson(λ) and Y follows Poisson(μ) be independent random variables.
For s = 0,1,2, . . ., p(s) = P($X+Y$ =s) is

$$p_{X+Y}(s) = \sum_{x=0}^{s} p_Y(s-x)\ p_X(x) = \frac{e^{-\mu-\lambda}}{s!} \sum_{x=0}^{s} \binom{s}{x} \mu^{s-x}\, \lambda^x$$

$$= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^s}{s!}$$

When number of individuals are more, that is when n is large we can use normal approximation for sum of the random variable on the basis of central limit theorem.
The usual statement of the theorem is for a sequence of independent and identically distributed random variables, $X_1, X_2, \ldots..X_n$, $E[X_i]= \mu$ and $Var[X_i]= \sigma^2$.

With $\bar{X} = \frac{X_1+X_1+\cdots+X_n}{n}$, $\sqrt{n}\,(\bar{X}-\mu)/\sigma$ is standard normal variate. That is it has mean 0 and variance 1.

Survival Function

Many insurance policies provide a benefit on the death of the policyholder. When an insurance company issues such a policy, the policyholder's date of death is unknown, so the insurer does not know exactly when the death benefit will be payable. The problems associated with life insurance involves the variability in the claim made. In other types of insurance the amount of the claim is also a random variable. The central difficulty in issuing life insurance is that of determining the length of the future life of the insured. In order to estimate the time at which a death benefit is payable, the insurer needs a model of human mortality, from which probabilities of death at particular ages can be calculated. Let X denote the random variable which represents the future lifetime of a newborn. Assume that the distribution function of X is absolutely continuous. The survival function of X, denoted by $s(x)$ is defined by the formula $s(x) = P[X > x] = P[X \geq x]$ where the last equality follows from the continuity assumption. The assumption that $s(0) = 1$ will always be made. The survival function in Reliability Theory is defined as the reliability function.

Let us say, we are interested in studying the life of persons who have attained certain age, say x. For convenience let (x) denote a life aged x. The death of (x) can occur at any greater than x, and we denote the future lifetime of a life aged x by $T(x)$. That is, life time of a person after the current age of x is $T(x)$ and $x + T(x)$ represent age-at-death random variable for (x).

In the past there has been some interest in modelling survival functions in an analytic way. The simplest model is that due to Abraham DeMoivre. He assumed that $s(x) = 1 - x/\omega$ for $0 < x < \omega$ where ω is the limiting age by which all have died. The DeMoivre law is simply the assertion that X has the uniform distribution on the interval $(0, \omega)$.

Life insurance is usually issued on a person who has already attained a certain age x. In many insurance problems we are interested in the probability of survival rather than death. The survival function for (x) is $P[T(x) > t]$. According to the standard notation, set ${}_tp_x = P[T(x) > t]$. This is the survival function for (x), that is, probability that (x) will attain age $x+t$. Let ${}_tq_x = P[T(x) \leq t]$, the distribution function of life from age x. This is the probability that (x) will die within t years. When $t = 1$ the prefix is omitted and one just writes p_x and q_x respectively. That is the probability that a person with age x will survive more than one year is p_x and the probability that a person with age x will die within one year is q_x.

${}_tp_x = P[T(x) > t] = P[X > x + t \,|X > x] = s(x+t)/s(x)$.

This gives another relation $P[X > x + t] = s(x+t) = s(x)\,{}_tp_x$.

Similarly, ${}_tq_x = P[T(x) \leq t] = P[X \leq x + t \,|X > x] = 1 - s(x+t)/s(x)$. Hence the density function of $T(x)$ is given by $f_{T(x)}(t) = -s'(x + t)/s(x)$

Further, ${}_{t+u}p_x = P[T(x) > t+u] = s(x+t+u)/s(x)$

Further, $_{t+u}p_x = P[T(x) > t+u] = \frac{s(x+t+u)}{s(x)}$

$= \frac{s(x+t+u)}{s(x+t)}\frac{s(x+t)}{s(x)} = {}_up_{x+t}\ {}_tp_x = {}_tp_x\ {}_up_{x+t}$

Note that since a life of currently aged x scurvies zero years is sure, we get, ${}_0p_x=1$.
Further, all lives eventually die we get $\lim_{t\to\infty} {}_tp_x = 0$.
Suppose p_x denotes ${}_1p_x$, we can write for integer $t \geq 1$, ${}_t p_x = p_x\, p_{x+1} \ldots\ldots p_{x+t-1}$.

There is one more special symbol. Set ${}_{t|u}q_x = P[t < T(x) \leq t+u] = P[T(x) \leq t+u] - P[T(x) \leq t]$ which represents the probability that (x) survives at least t and no more than $t+u$ years. Again, if $u = 1$ one writes ${}_t|q_x$. The relations ${}_t|_uq_x = {}_{t+u}q_x - {}_tq_x = {}_tp_x - {}_{t+u}p_x$ follow immediately from the definition.

The **curtate future lifetime** of (x), denoted by $K(x)$, is defined by the relation $K(x) = [T(x)]$. Here $[t]$ is the greatest integer function. Note that $K(x)$ is a discrete random variable with density $P[K(x) = k] = P[k \leq T(x) < k+1]$. The curtate lifetime, $K(x)$, represents the number of complete future years lived by (x). That is, $P[K(x) = k]$ is the probability of (x) living exactly more than k year but dies before $(k+1)^{st}$ year after age x.

Force of mortality:

We denote the force of mortality for those attaining age x by μ(x) and define it as

$$\mu(x) = \lim_{dx\to 0^+} \frac{1}{dx}\Pr[T(0) \leq x + dx \,|\, T(0) > x].$$

Using the definition of T_x, we see that above expression is equivalent as

$$\mu(x) = \lim_{dx\to 0^+} \frac{1}{dx}\Pr[T(x) \leq dx].$$

In terms of survival function, `

$$\mu(x) = \lim_{dx\to 0^+} \frac{1}{dx}\left[1 - \frac{s(x+dx)}{s(x)}\right] = -\frac{s'(x)}{s(x)}.$$

Hence, $\mu(x) = \frac{f_X(x)}{1-F_X(x)}$

The force of mortality represents the death rate per unit age per unit survivor for those attaining age x.

Intuitively the force of mortality is the instantaneous 'probability' that someone exactly age x dies at age x. (In component reliability theory this function is often referred to as the *hazard rate*.) Integrating both sides of this equality gives the useful relation

$$s(x) = exp\{-\int_0^x \mu(t)\, dt\}.$$

In reliability theory, the study of the survival probabilities of manufactured parts and systems, $\mu(x)$ is called the failure rate function or hazard rate function.

If the force of mortality is constant the life random variable X has an exponential distribution. This is directly in accord with the "memoryless" property of exponential random variables. This memoryless property also has the interpretation that a used article is as good as a new one. For human lives (and most manufactured components) this is a fairly poor assumption,

at least over the long term. The force of mortality usually is increasing, although this is not always so. There are many distributions for certain parameters exhibit increasing force of mortality. However, in actuarial science we estimate this function based on the population pertaining to the area of study.

Using the relation between survival function and force of mortality given above, we observe that $-\mu(x) = d \log s(x)$. Integrating this expression from x to $x + n$, we have

$$-\int_x^{x+n} \mu(y)dy = \log\left[\frac{s(x+n)}{s(x)}\right] = \log\ {}_np_x$$

Hence, $${}_np_x = exp\left(-\int_x^{x+n} \mu(y)dy\right)$$

This is equivalent to $${}_np_x = exp\left(-\int_0^n \mu(x+t)\ dt\right).$$

In particular, to convert to the notation we defined earlier, we can also write as

$${}_xp_0 = exp\left(-\int_0^x \mu(t)dt\right) = s(x)$$

In addition, $F_X(x) = 1 - s(x) = 1 - exp\left(-\int_0^x \mu(t)dt\right)$ is the probability distribution function of X. The density function of X can be obtained by differentiating $F(x)$. That is

$$F'_X(x) = f_X(x) = exp\left(-\int_0^x \mu(t)dt\right)\ \mu(x) =\ {}_np_0\ \mu(x).$$

Let $F_{T(x)}(t)$ and $f_{T(x)}(t)$ denote, respectively, the probability distribution function and probability density function of $T(x)$, the future lifetime of (x). Also note that $F_{T(x)}(t) = {}_tq_x$; therefore, $f_{T(x)}(t) = \frac{d}{dt}\ {}_tq_x = \frac{d}{dt}\left(1 - \frac{s(x+t)}{s(x)}\right)$

$$= \frac{s(x+t)}{s(x)}\left[-\frac{s'(x+t)}{s(x+t)}\right] = {}_tp_x\,\mu(x+t) \quad t \geq 0.$$

Thus ${}_tp_x\,\mu(x+t)\,dt$ is the probability that (x) dies between t and $t + dt$, and

$$\int_0^\infty {}_tp_x\,\mu(x+t)dt\ = 1$$

where the upper limit on the integral is written as positive infinity to indicate there is no upper bound.

We also get $\frac{d}{dt}\ (1 -\ {}_tp_x\ ) = -\frac{d}{dt}\ {}_tp_x =\ {}_tp_x\,\mu(x+t)$

This equivalent form is useful in several developments in actuarial science.

Life Tables

In practice the survival distribution is estimated by compiling mortality data in the form of a life table. Given a survival model, with survival probabilities ${}_tp_x$, we can construct the **life table** for the model from some initial age x_0 to a maximum age ω. We define a function $\{l_x\}$ for $x_0 \leq x \leq \omega$ as follows. Let l_{x_0} be an arbitrary positive number (called the **radix** of the table) and, for $0 \leq t \leq \omega - x_0$, define $l_{x_0+t} = l_{x_0}\ {}_tp_{x_0}$.

From this definition we see that for $x_0 \leq x \leq x + t \leq \omega$,

$$l_{x+t} = l_{x_0\ x+t-x_0} p_{x_0} = l_{x_0\ \ x-x_0} p_{x_0\ \ t} p_x$$
$$= l_{x\ \ t} p_x ,$$

so that ${}_t p_x = l_{x+t} / l_x$.

For any $x \geq x_0$, we can interpret l_{x+t} as the expected number of survivors to age $x + t$ out of l_x independent individuals aged x. This interpretation is more natural if l_x is an integer, and follows because the number of survivors to age $x + t$ is a random variable with a binomial distribution with parameters l_x and ${}_tp_x$. That is, suppose we have l_x independent lives aged x, and each life has a probability ${}_tp_x$ of surviving to age $x + t$. Then the number of survivors to age $x + t$ is a binomial random variable, B_t , say, with parameters lx and ${}_tp_x$. The expected value of the number of survivors is then $E[B_t] = l_x\ {}_tp_x = l_{x+t}$.
We always use the table in the form l_y/l_x which is why the radix of the table is arbitrary – it would make no difference to the survival model if all the l_x values were multiplied by 100, for example. We can use the l_x function to calculate survival probabilities. We can also calculate mortality probabilities. For example,

$$q_{30} = 1 - \frac{l_{31}}{l_{30}} = \frac{l_{30} - l_{31}}{l_{30}}$$
$$and \quad {}_{15|30}q_{40} = {}_{15}p_{40}\ {}_{30}q_{55} = \frac{l_{55}}{l_{40}} \left(1 - \frac{l_{85}}{l_{55}}\right) = \frac{l_{55} - l_{85}}{l_{40}} .$$

In principle, a life table is defined for all x from the initial age, x_0, to the limiting age, ω. In practice, it is very common for a life table to be presented, and in some cases even defined, at integer ages only. In this form, the life table is a useful way of summarizing a lifetime distribution since, with a single column of numbers, it allows us to calculate probabilities of surviving or dying over integer numbers of years starting from an integer age.
It is usual for a life table, tabulated at integer ages, to show the values of d_x, where $d_x = l_x - l_{x+1}$, in addition to l_x, as these are used to compute q_x, we have

$$d_x = l_x \left(1 - \frac{l_{x+1}}{l_x}\right) = l_x\ (1 - p_x) = l_x q_x .$$

We can also arrive at this relationship if we interpret d_x as the expected number of deaths in the year of age x to $x + 1$ out of l_x lives aged exactly x, so that, using the binomial distribution again $d_x = l_x q_x$.
Further, it is easy to see that ${}_nd_x = lx\ lx{+}n$.

Imagine that at time 0 there are l_0 newborns. Here l_0 is called the radix of the life table and is usually taken to be some large number such as 10,000,000. These newborns are observed and l_x is the number of the original newborns who are still alive at age x. Similarly ${}_nd_x$ denotes the number of the group of newborns alive at age x who die before reaching age $x + n$. As usual, when $n = 1$ it is suppressed in the notation.
It is easy to see that ${}_nd_x = l_x\ l_{x+n}$.

Summary of notations used in the Life Tables
${}_tq_x = 1 - {}_tp_x = P[T(x) \leq t] = P[X \leq x + t \mid X > x] = 1 - s(x+t)/s(x)$, ${}_1q_x = q_x$
l_x = number of persons aged x living.
${}_nd_x$ = denotes the number of the group of newborns alive at age x who die before reaching age $x + n$.
d_x = number of persons dying between ages x; x +1, that is $d_x = l_x - l_{x+1} = l_x\ q_x$.

In the cohort life-table model, imagine a number l_0 of individuals born simultaneously and followed until death, further based on the data on d_x we determine l_x and probabilities q_x. This is the deterministic procedure.
In the probabilistic model, given the number l_0 we can determine the probabilities q_x and l_x using survival function $s(x)$.

Practical 2: Computation of various components of life tables

Given l_0 = 100000 and column d_x only, complete the table.

The table is completed using the formula $l_{x+1} = l_x - d_x$ *and* $q_x = d_x / l_x$.

Age x	l_x	d_x	q_x
0	100000	2629	0.0263
1	97371	141	0.0014
2	97230	107	0.0011
3	97123	63	0.0006
4	97060	63	0.0006
5	96997	69	0.0007
6	96928	69	0.0007
7	96859	52	0.0005
8	96807	54	0.0006
9	96753	51	0.0005
10	96702	33	0.0003
11	96669	40	0.0004
12	96629	47	0.0005
13	96582	61	0.0006
14	96521	86	0.0009
15	96435	105	0.0011
16	96330	83	0.0009
17	96247	125	0.0013
18	96122	133	0.0014
19	95989	149	0.0016
20	95840	154	0.0016
21	95686	138	0.0014
22	95548	163	0.0017
23	95385	168	0.0018
24	95217	166	0.0017
25	95051	151	0.0016
26	94900	149	0.0016

Age x	l_x	d_x	q_x
40	92315	295	0.0032
41	92020	332	0.00361
42	91688	408	0.00445
43	91280	414	0.00454
44	90866	464	0.00511
45	90402	532	0.00588
46	89870	587	0.00653
47	89283	680	0.00762
48	88603	702	0.00792
49	87901	782	0.0089
50	87119	841	0.00965
51	86278	885	0.01026
52	85393	974	0.01141
53	84419	1082	0.01282
54	83337	1088	0.01306
55	82249	1213	0.01475
56	81036	1344	0.01659
57	79692	1423	0.01786
58	78269	1476	0.01886
59	76793	1572	0.02047
60	75221	1696	0.02255
61	73525	1784	0.02426
62	71741	1933	0.02694
63	69808	2022	0.02897
64	67786	2186	0.03225
65	65600	2261	0.03447
66	63339	2371	0.03743

27	94751	166	0.0018
28	94585	157	0.0017
29	94428	133	0.0014
30	94295	160	0.0017
31	94135	149	0.0016
32	93986	152	0.0016
33	93834	160	0.0017
34	93674	199	0.0021
35	93475	187	0.002
36	93288	212	0.0023
37	93076	228	0.0024
38	92848	272	0.0029
39	92576	261	0.0028

67	60968	2426	0.03979
68	58542	2356	0.04024
69	56186	2702	0.04809
70	53484	2548	0.04764
71	50936	2677	0.05256
72	48259	2811	0.05825
73	45448	2763	0.06079
74	42685	2710	0.06349
75	39975	2848	0.07124
76	37127	2832	0.07628
77	34295	2835	0.08267
78	31460	2803	0.0891

Given the number l_0=100000, using probabilistic method we determine entire Table. First determine q_x=1-s(x+1)/s(x) for all x. Then using the formula $d_x = q_x\, l_x$, and $l_{x+1} = l_x - d_x$.