Parabola Volume 24, Issue 1 (1988)

STATISTICAL MODELS FOR SURVIVAL

by

Charles McGilchrist*

Survival of a cockroach is unfortunate, survival of an endangered species is important, survival of a human (probably the least endangered species) is essential. The young take survival to be their right; the old find survival an absorbing interest.

Statistical survival analysis has been of interest to actuaries and statisticians for centuries. The basic theory depends on a knowledge of the concepts of *probability* and *conditional probability*.

Probability and conditional Probability

The use of the word probability is not unique being used by different people to mean different things. Here the *relative frequency* concept of the word is used. Suppose we have a long sequence of trials under identical conditions, each trial not being influenced by results of previous trials and each trial resulting in S=*survival* or D=*death*. The probability of survival is taken to be the limiting relative frequency of S in this long sequence of trials and is denoted by Pr(S). We may, of course, estimate that limiting relative frequency by the relative frequency of S over the finite number of trials that we have.

What sort of trials would be appropriate here? If we are concerned about survival of cockroaches (or blowflies) to insecticide then exposing 10000 random

* Charles is Associate Professor in Statistics at UNSW.

selected insects to a standard dosage would constitute 10000 trials under identical conditions. The proportion of survivors estimates the probability of survival although it is usually *percent kill* that is reported. If we are concerned about survival of sixteen year old males during their seventeenth year then observing the survival experience of 10000 randomly selected sixteen year old males would be given.

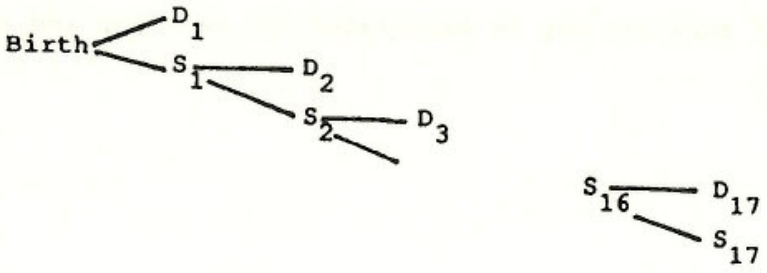
Conditional probability is an extension of this notion. Suppose sixteen year old females are classified as brown/non-brown eyed and we want to refer to the probability of survival of brown eyed females. Each trial may now be classified also as B=*brown eyed*, O=*other female* selected and the sequence of trials is now subdivided into two subsequences based on eye colour.

The limiting relative frequency of survival in the brown-eyed subsequence is called the conditional probability of survival given the female selected is b rown eyed. It is denoted by Pr(S|B) and

$$\begin{aligned}\Pr(S|B) &= \text{limiting relative frequency of SB in the B subsequence}\\ &= \frac{\text{limiting relative frequency of SB in original sequence}}{\text{limiting relative frequency of B in original sequence}}\\ &= \Pr(SB)/\Pr(B).\end{aligned}$$

Crossmultiplying gives $\Pr(SB)=PR(B)\Pr(S|B)$.

Now let us return to survival of sixteen year old males. The probability of survival in the seventeenth year applies only to males who have already survived the first sixteen years. The probability estimated is then really the conditional probability of S_{17} given $S_1S_2..S_{16}$ where S_i is survival of the i^{th} year. It is denoted by $\Pr(S_{17}|S_1S_2 \ldots S_{16})$. If D_i is the event of death in the i^{th}year then an appropriate *tree diagram* is



The total probability of surviving seventeen years is

$$\Pr(S_1 S_2 \ldots S_{17}) = \Pr(S_1 S_2 \ldots S_{16}) \Pr(S_{17} | S_1 S_2 \ldots S_{16})$$
$$= \Pr(S_1) \Pr(S_2 | S_1) \ldots \Pr(S_{17} | S_1 S_2 \ldots S_{16}).$$

The probability $\Pr(S_i | S_1 S_2 \ldots S_{i-1})$ is called the age specific survival rate for the i^{th} year and one minus this probability is the age specific death rate of hazard rate. Hazard rates estimated from recent Australian Bureau of Statistics data for males and females separately give

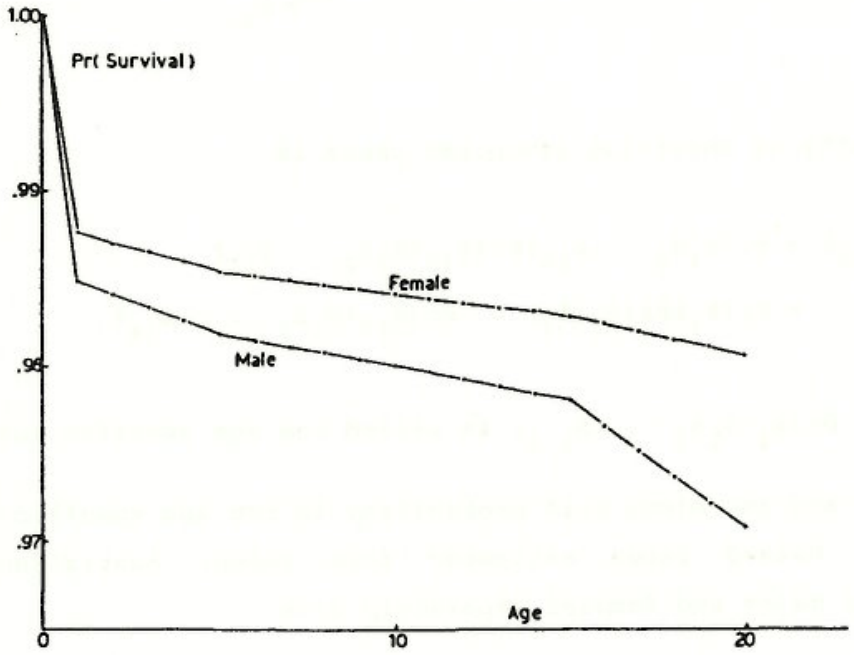
Years	1	2-5	6-10	11-15	16-20
Males	.01515	.00076	.00037	.00040	.00150
Females	.01244	.00054	.00027	.00024	.00047

The probability of male survival to age 17 is estimated at

$$(1-.01515)(1-.00076)^4(1-.00037)^5(1-.00040)^5(1-.00150)^2$$

Note that the age specific death rates or hazard rates for females are lower than for males and indeed remain so for all age groups. The hazard rate for males increases dramatically for the 16-20 age group and this increase is usually attributed to fatality in motor crashes. Notice that the girls are not at such greatly increased hazard possibly because of more careful behaviour.

The probability of survival may be calculated for each age and the results presented graphically as



Such graphs or curves are called survival curves and are a convenient way of summarising a very basic quantity, namely survival. The curves indicate that the probability of surviving to age 21 is estimated as 0.9708 for boys and 0.9806 for girls. Thus the average boy would have lost 3% of contempories by age 21 and the average girl 2%

Hazard Function

So far only annual survival rates have been discussed but in medical studies of survival of cancer patients on a particular treatment regime or coronary patients following an infarct, it is necessary to consider survival not from year to year but rather from one very short time interval to the next. Suppose a person has survived until time t and we wish to consider his survival experience in the next very short time interval denoted by $(t, t+\Delta t)$. If

$h(t)\Delta t$ = Pr[death in $(t,t+\Delta t)$ given survival until time t]

then, for Δt small, $h(t)$ is the instantaneous death rate or hazard rate at time t. Approximately we may think of $h(t)$ as the probability of dying in the next unit of time given survival up to the beginning of that unit of time.

This hazard function $h(t)$ may be linked to another important function called the survivor function defined as

$$S(t) = \text{Pr(surviving to at least time t)}$$

If $S'(t)$ is its derivative and Δt is small then, from the definition of the derivative of a function,

$$S(t) - S(t+\Delta t) = -S'(t)\Delta t$$

is the probability of dying in $(t,t+\Delta t)$. The conditional probability of dying in $(t,t+\Delta t)$ given survival to t is the above probability divided by $S(t)$ so that

$$h(t)\Delta t = -S'(t)\Delta t/S(t) \quad \text{giving} \quad h(t) = -S'(t)/S(t)$$

Integrating from 0 to t and using $S(0) = 1$ gives

$$\int_0^t h(u)\,du = -\ln S(t)$$

and

$$S(t) = \exp - \int_0^t h(u)\,du$$

which is an expression linking the hazard and survivor functions.

Modelling the Hazard Function

While the bottom line is survival, the hazard function or instantaneous death rate is usually more informative about variations in hazard occurring at

different times. Statistical models are usually written in terms of the hazard function although such models clearly imply a survivor function model.

In medical studies interest centres not only on how the hazard function varies over time but also on how it relates to levels of risk factors and to treatment regimes. The hazard function for cancer patients may depend on age, sex of the patient, stage of development of the cancer, white blood cell count as well as the treatment imposed. Analysis of such data was made much easier by the introduction of the *proportional hazards model* by a British statistician D.R. Cox in the Journal of the Royal Statistical Society (series B) as late as 1972. The model is

$$\text{hazard function} = h(t) = \lambda(t) \times g(\text{risk, treatment variables})$$

indicating that there is a basic hazard shape over time denoted by $\lambda(t)$ and this hazard shape is elevated or lowered by a multiplicative factor g which depends on the risk and treatment variables. The model appears to fit well in many studies. If

$$S_o(t) = \exp -\int_c^t \lambda(u)\, du$$

is the survivor function associated with hazard function $\lambda(t)$

then

$$S(t) = \exp -\int_c^t \lambda(u) g(\text{risk, treatment variables})\, du$$

$$= [S_o(t)]^{g(\text{risk, treatment variables})}$$

is the survivor function for the proportional hazards model.

Cox used this model to study the survival experience of leukemia patients and how survival related to treatment by a new drug. Patients were divided randomly into two groups with one group being given no treatment (acting as control group) and the other group being given the new drug. The times of remission in weeks for each leukemia patient are given in the following table.

Group 1 (drug 6-MP)	6*	6	6	6	7	9*	10*	10	11*	13
	16	17*	19*	20*	22	23	25*	32*	32*	34*
	35*									
Group 2 (control)	1	1	2	2	3	4	4	5	5	8
	8	8	8	11	11	12	12	15	17	22
	23									

* indicates patient alive at the end of this period.

The proportional hazards model chosen to fit this data was

$$\text{hazard function} = h(t) = \lambda(t) \times \exp(\beta x)$$

where $x=0$ for a member of the control group and $x=1$ for a member of the treatment group. Note that the function $g(.)$ is chosen as $\exp(\beta x)$. The problem is to decide firstly whether or not the drug is effective and secondly, if it is, to estimate the extent of its beneficial effects.

The main thrust of Cox's work was to give a method of estimating β and, most importantly here, testing to see if β might be zero. If β is zero then the treatment has no effect and there is no point in giving patients useless drugs. It is usual, therefore, to demand that new drugs be shown to be effective in a clinical trial such as described above and this involves the use of an appropriate statistical test procedure. The development of Cox's test and estimation procedures is beyond the scope of this article and indeed requires much more extensive study of theoretical aspects of statistical inference. It suffices to say that Cox's method has been used in hundreds of studies during the past fifteen years and is now basic to the clinical trial literature.