STAT 100 Chance and Data Analysis November 1, 2002

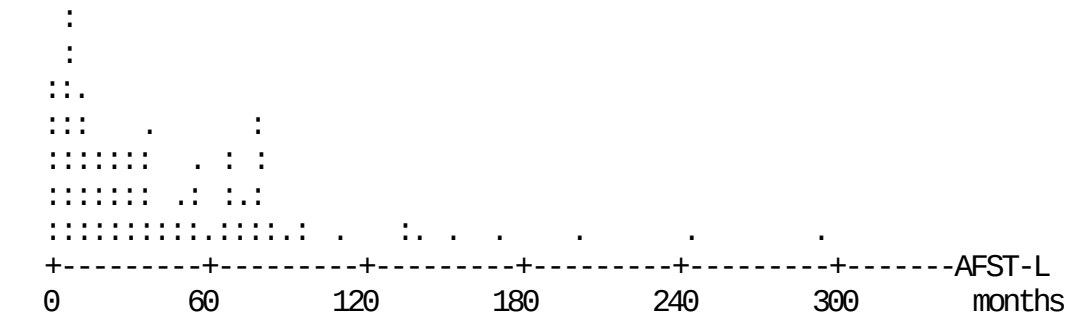
Next topic: Survival Analysis

(More follow-up from recent survey next week)

Think about when you got your driver's license, and how long after that you had your first auto accident (include minor scrapes, and "not my fault", in this). Call this time your accident-free survival time (AFST). What would the distribution of AFST look like?

Here is a suggestion – the time unit below is in months. (For me it was less about 8 months - KLW).

First consider what it would look like if we had data on 25 years of driving:



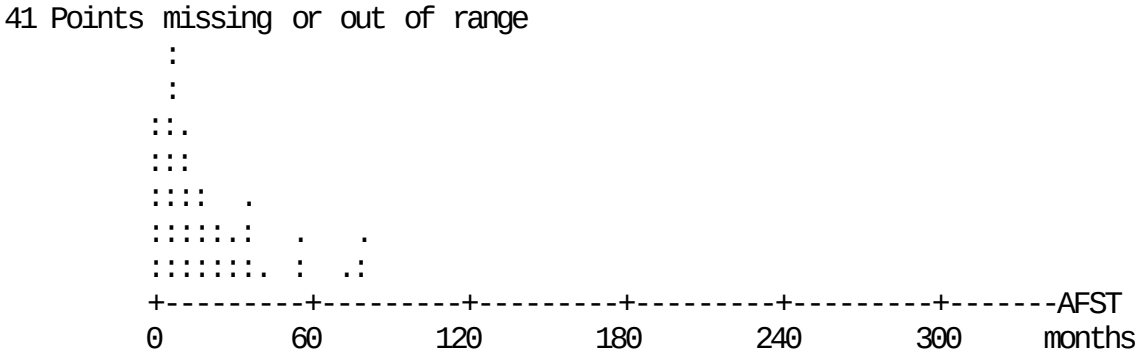
AFST-L means that these are the AFST over a lifetime of driving.

But now suppose we collect the data from the current class, where the observation time (from license date to present) may well be less than the AFST. We might have (partial display of 100 student responses)

Data Display

OBS	AFST	
7	2	7 months since obtained license, accident 2 months later
64	39	
23	15	
51	8	
134	71	
.	.	
.	.	
57	>57	57 months since obtained license, more than 57 months to first accident
80	>80	
133	>133	
38	>38	
30	>30	

Now our data will include many “no accident yet” and the ones that have had an accident would have AFSTs that look like this:



Of 100 students with driving licenses, 59 have had an accident of some kind, and 41 have not yet had an accident.

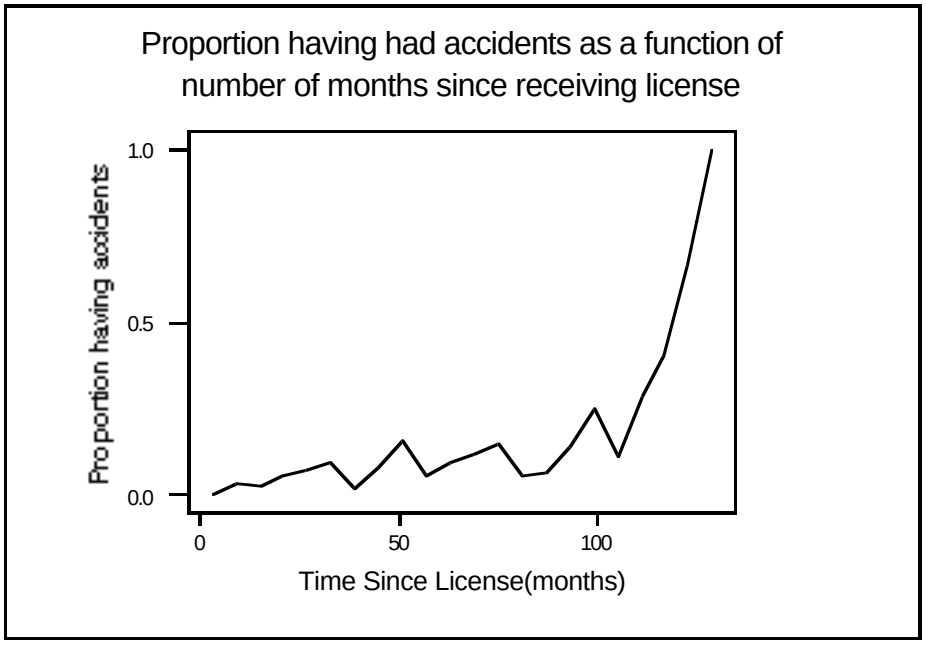
Q: Does this mean that hardly any students survive accident-free for more than about 80 months?

Not really – most students have not been “at risk” for more than 80 months, so those that would remain accident-free for longer than 80 months do not appear on the AFST dotplot above.

How does one use this kind of data to compute the risk of an accident as a function of time?

The method is this. Find all the people who have been observed for a given time period (like 6 months) and compute the proportion that have had an accident in that time. This is the 6 month “hazard” rate. Do the same thing for 12 months, etc.

With the data used here, 100 people observed for 0-6 months, no accidents occurred for any of these people for this duration, so the accident rate for 0-6 months is estimated as 0/100 =0.00. There were two people only observed for 6 months or less so that left 98 people at risk of an accident in the 6-12 month period. there were 3 accidents in this duration so the hazard rate during this period was 3/98 = .031. there were 8 people who were observed for 6-12 months only and so when we look at those observed for 12-18 months, there are only 100-2-8=90 left. In this 12-18 month duration there were 2 more accidents, so the hazard rate in this period is 2/90 = .022. Continuing in this way we get the following graph:



The apparent surge during the last few months is just a result of the small denominator for the estimate, and should be ignored. The general trend is for the proportion of accidents to increase fairly steadily (and linearly) as the observation period grows. In fact, it looks as though the hazard rate (number of accidents divided by number of people at risk and observed) increase about 0.2 per 120 months or .033 per 6 months. That means the accident rate is about 3.3% per 6 month period.

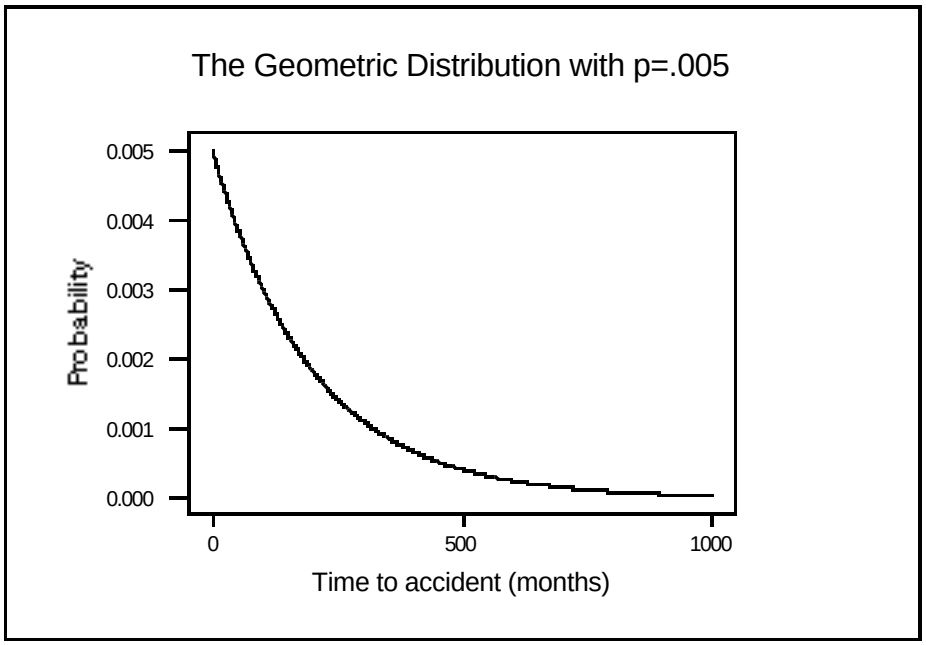
This is a useful finding (although based in these notes on fictitious data) – an insurance company uses this kind of information. Note that we have estimated this even though a large percentage of the group of 100 people studied have still not had an accident, and all respondents have been driving for various lengths of time.

That is the kind of data analysis that is used on survival data. Now lets consider how to model this kind of data.

We need a model for the probabilities that something will last a certain length of time. One simple and very useful model is called the Geometric Distribution. Any driver will experince a 3.3% risk of an accident in 6 months, or say about 1/2 % risk per month (to keep the calculations simple). The probability that the driver had an accident in the first month is .005. If the driver does not have an accident (an event with probability .995) than there is a probability of .005 of an accident in the second month. So 1/2 % of the drivers will have an accident in the first month and 99.5% of 1/2 % will have an accident

in the second month (but not the first). Similarly the probability that a driver waits until moth 3 for a accident is .995 x .995 x .005. The general formula is
P(first accident in month n) = $.995^n$ .005

The graph of the frequency distribution looks like this:

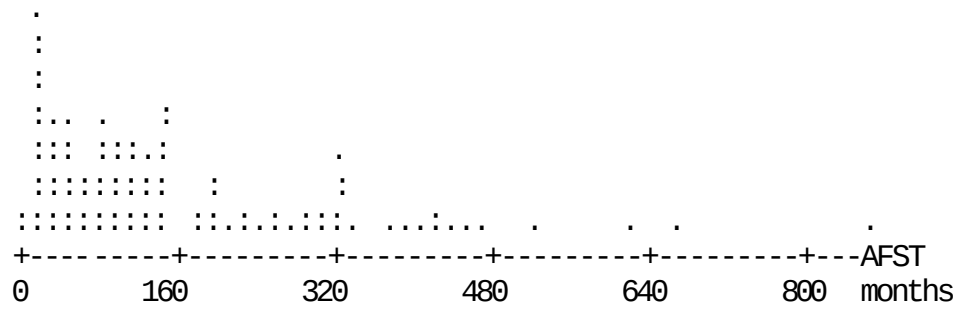


These probabilities apply to the future prospects of a single individual, but we can also interpret it in a population sense – if we started with 1000 individuals, how many would have had their accident by the 25^{th} month? We simply compute $.995^n$ x .005 for n = 0,1,2,…24 and add up the probabilities and apply them to the 1000 individuals. This sum comes to 0.118 so the average number having the accident by month 25 would be 118 out of the 1000 that started.

Now, let us consider this application with a bit of contextual knowledge. Is it reasonable to assume that drivers have the same risk of the first accident no matter how long they have been driving? Another reaonable theory might be that, the longer the time that the driver is accident free, the LESS the risk of an accident in the next month. Maybe drivers are able to do more safe things automatically. Or, yet another theory might be that as drivers experience accident free driving, they become careless and have MORE risk of an accident in the next month. These situations are referred to as decreasing or increasing hazard rate, respectively. The Geometric Distribution for AFST always has a constant hazard rate and so this model is not as general as we would like it to be. More complex models are beyond the scope of this course. But note how useful it is to have the

"wrong" model for this situation, the Geometric. It gives us something to compare the reality to. We will attempt to estimate the real hazard function (that is, the probability of an accident in the next month) for drivers with various amounts of driving experience. (We won't separate males and females, but this simplification is sure to cause some inaccuracies.) We can look for an increasing or decreasing hazard rate over the duration of exposure (= time since licensing).

Another observation – Note that the data we see here is severely "censored" by the length of time a driver has had a license. If drivers really had a .005 rate of accident per month, and we could observe all drivers for a lifetime, the AFST distribution would look like this:



The method we have used gets around the censoring in the data to recover the desired information about hazard rate (i.e. risk of having an accident in the next month).