STAT 100 Chance and Data Analysis November 4, 2002

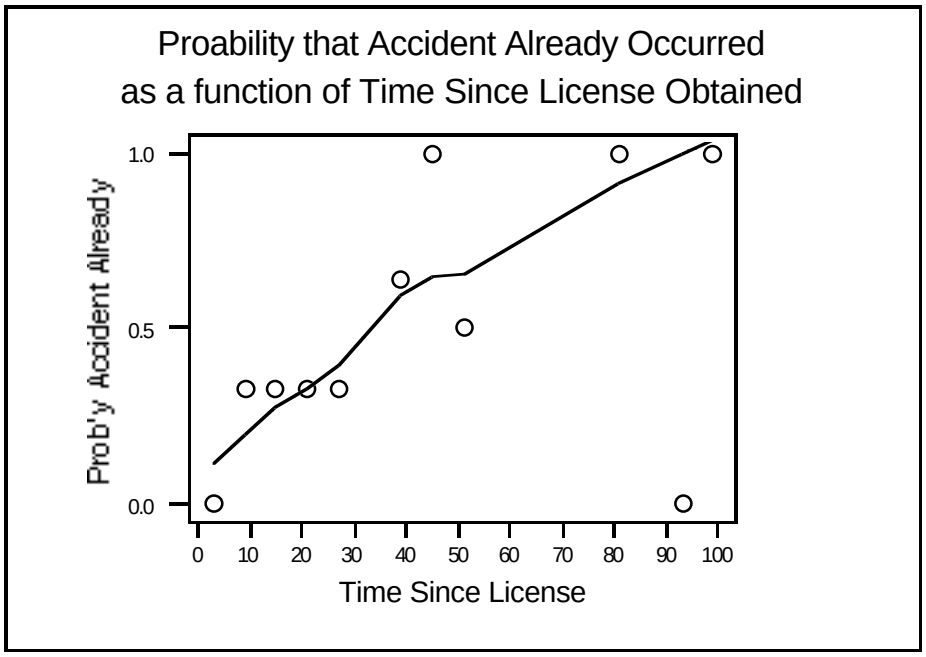
Today: Lifetime & Survival data analysis
Correction to Assignment 6 Problem #2.

License Date	Accident?	Today	Obs. Time(Months)
Oct-02	0	Nov-01	1
Mar-02	0	Nov-01	8
Mar-02	1	Nov-01	8
Feb-02	0	Nov-01	9
1-Jan	0	Nov-01	10
	1		11
Dec-01	0	Nov-01	11
Jul-01	0	Nov-01	16
Jul-01	0	Nov-01	16
Jul-01	1	Nov-01	16
1-May	0	Nov-01	18
Apr-01	0	Nov-01	19
Apr-01	0	Nov-01	19
Jan-01	1	Nov-01	22
Dec-00	1	Nov-01	23
Dec-00	0	Nov-01	23
Nov-00	1	Nov-01	24
Oct-00	0	Nov-01	25
Sep-00	0	Nov-01	26
Sep-00	1	Nov-01	26
Aug-00	0	Nov-01	27
Jul-00	0	Nov-01	28
Nov-99	0	Nov-01	36
Nov-99	0	Nov-01	36
Sep-99	0	Nov-01	38
Aug-99	1	Nov-01	39
Aug-99	1	Nov-01	39
Aug-99	1	Nov-01	39
Aug-99	1	Nov-01	39
Jul-99	0	Nov-01	40
Jun-99	1	Nov-01	41
Jun-99	1	Nov-01	41
Jun-99	1	Nov-01	41
Dec-98	1	Nov-01	47
Nov-98	0	Nov-01	48
Sep-98	1	Nov-01	50
Mar-96	1	Nov-01	80
Jan-95	0	Nov-01	94
Jun-93	1	Nov-01	113

As in the simulation from last day, I will estimate probabilities using 6 month accumulations.
For example, for those observed at least 6-11 months, of the 6 observed, 2 reported having already had an accident. So the estimated probability that a students AFST (accident-free survival time) is 9 months or less is 2/6=0.33. Summarizing the data in this way produces the following table:

No. Accidents	No. Observed	Est Prob	Mid-interval
0	1	0.00	3
2	6	0.33	9
1	3	0.33	15
2	6	0.33	21
2	6	0.33	27
7	11	0.64	39
1	1	1.00	45
1	2	0.50	51
1	1	1.00	81
0	1	0.00	93
1	1	1.00	111

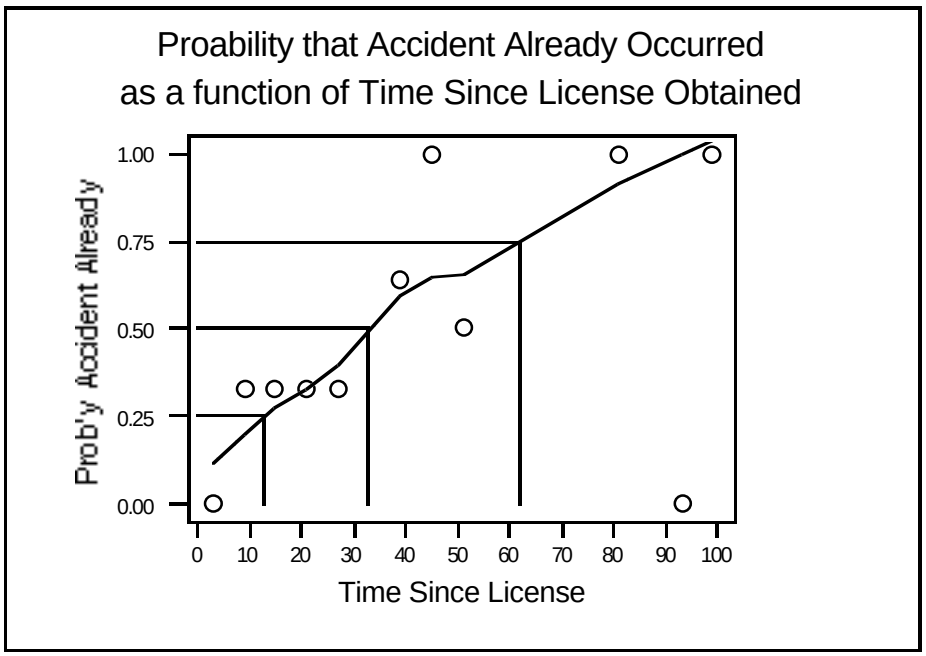
And if we plot the third column (vertical axis) against the fourth column (hor. axis):



The fitted curve is a “lowess” smoothing of the points. (All you need to know about lowess is that it is an automatic smoothing method.) Note that some points are estimated with more data than others and this is not taken into account. The points on the right side of the graph are based on proportions with a denominator of 1 or 2, and so are expected

to be quite variable (only a few respondents have had their license for 60 or more months.). It is possible to weight the points but we avoid this complication here.

What does the graph say? It says that we can estimate the probability that an AFST is less than any particular value. For example, from the smoothed relationship we can estimate that the quartiles of the AFST distribution are Q1=12, Q2=33, and Q3=62. See following graph:



Note that we have not used any special probability distribution in the estimation above. What does the data tell us about the prospects of students in this class?

For those who have already had an accident, you can compare your experience with the class by seeing what percentile your AFST falls at. For example, if your AFST was 50 months, then this would be estimated to be be estimated to be longer than about 70% of the students. (50 months is approximately the 70^{th} percentile.)

If you have not had an accident yet, and it has been 50 months since you got your license, your AFST is definitely greater than 50 months. But what is your chance next month of having an accident, and realizing your AFST in 51 months? Well if the smooth curve can be approximated by a straight line over the range 30-70 months, the slope appears to be about (.85-.45)/(70-30) = .01 so that for each month increase over 50, the probability that the SFST is less that increases by about .01 – this means that the probability that AFST is less than 51 is .01 larger than the probability that AFST is less than 50. So this means

that for a student that has not yet had an accident, and is already at AFST=50, the chance of an accident in month 51 is .01 or 1%.

One of our basic assumptions in this is that all students see the same risk of an accident, and this is not really true. But to an outside observer who has not more detailed information to go on (like an insurance company), this assumption is still useful since it will estimate the accident rates for a group of people (a group similar to the class).

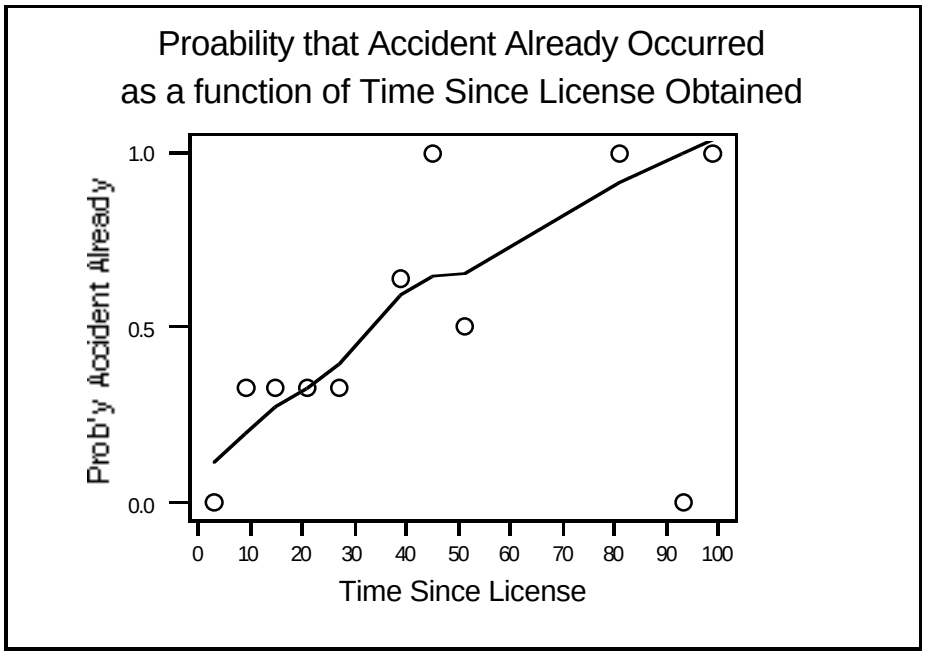
So far we have not used any special probability law, like the Geometric distribution that was introduced last time. We will now make use of this model for the accident data, but first let's review what the Geometric distribution is:

The general context may be described as the one of tossing a coin, although we should include the possibility of the coin being biased. Suppose the chance of a head is p (a number between 0 and 1, but not necessarily 0.5). Then if we toss the coin until the first head appears, the number of tosses it takes may be 1,2,3, ...and the relative frequency of all these possibilties is given by the probabilty distribution. For a fair coin, the probabilities are, .5, .25, .125, .0625, ...for the number of tosses to be 1,2,3,4,... respectively. (The kth probability is just 1/2 raised to the kth power). But for a biased coin, with probability of a head of p, the probabilities of the various possible numbers of tosses to get the first head will be p, (1-p)p, $(1-p)^2p$, $(1-p)^3p$, etc... For example, if p were only 0.1 (very hard to get a Head), the probabilities are 0.1, 0.09, 0.081, 0.0729, ... The sequence p, (1-p)p, $(1-p)^2p$, $(1-p)^3p$, ... applied to the possible outcomes 1,2,3,4,.... is the Geometric distribution.

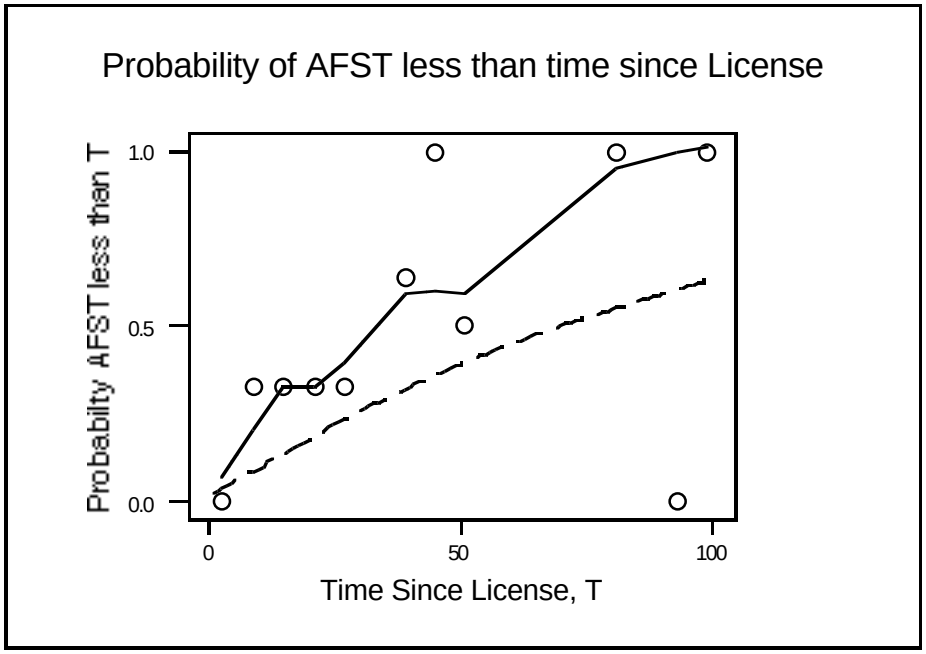
Now in our context of AFST, the p is the probability of an accident in one month, and the AFST is the number of months until the first accident. We know that p will be quite small, and our rough estimate from the previous analysis (without the Geometric model) was that p =.01. In this case the sequence of probabilities would be .01, .0099, .0098, .0097, ... With more significant digits we have

1	1	0.0100000
2	2	0.0099000
3	3	0.0098010
4	4	0.0097030
5	5	0.0096060

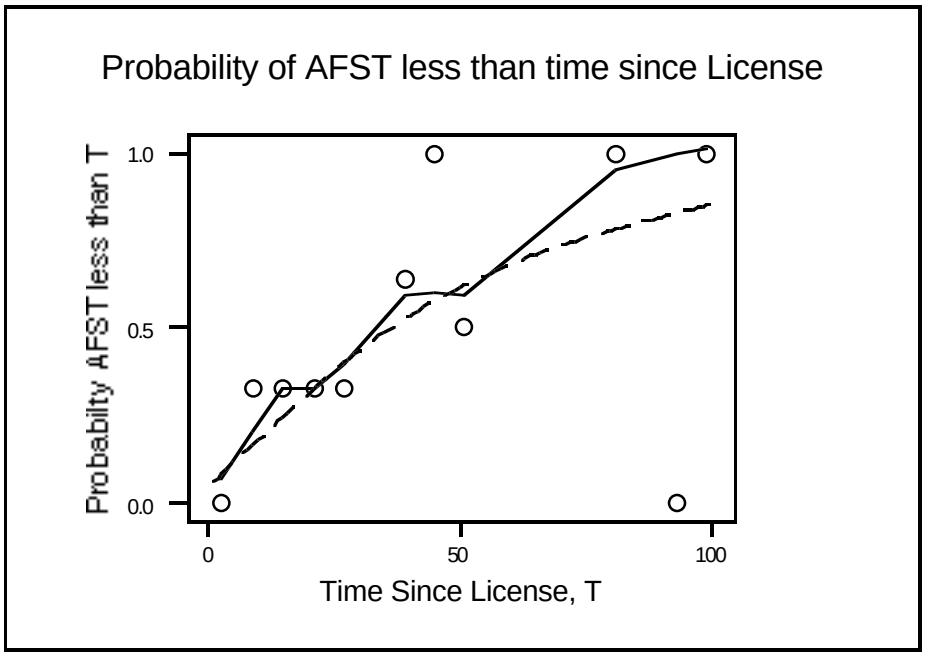
and so on. These are computed using thr formula with terms like $(1-p)^3p$. The question in our current context is, does this Geometric distribution fit the data? In order to compare this with the distribtion shown in the graph below



we need to compute comparable probabilities from the formula. We know how that the probabilities for 1,2,3,4 are .01, .0099, .0098, .0097 but what we really need are the probabilities that the AFST is less than or equal to 4 (and 1, and 2, and 3, and 5, ...). But we just add these to get P(AFST ≤4)= 0.01 + .0099 + .0098 + .0097 = .0396 and similarly P(AFST≤k) for k=1,2,3,4,5,..... Using MINITAB this is easily done and we get



We can see that the fit is not good, but the shape seems right. We seem to have too small a value of p in our geomtric model. Let's try again with p=0.02.



This is pretty good, and we could fiddle a bit more but we need to remember that the points at months >50 are not well estimated, so perhaps p=.02 is good enough.

Which method is right? In this case, probably the Geometic method is more accurate, since we used a straight line fit (from 30 to 70, remember) to guess the p=.01, whereas we are now using all the data to get p=.02. (Note that the straight line assumption for the probability graph in our first method is certainly erroneous, since probabilities cannot be greater than 1.0, and time to license can be greater than 100.)

So how do you compute your probability of a accident in the next month, assuming you have not had one yet since getting your license? It turns out that, using the Geometric distribution, the probability is p = .02. This is the case for any T, the AFST so far. The assumption underlying the Geometric model is that this "hazard rate", p, stays constant over varying T. Is this realistic?

Well, the data obtained without the assumption seems to fit the data with the assumption. Also, when I asked in class if you thought the hazard rate would increase or decrease as T increased, there was a difference of opinion. Some thought the hazard would decrease with increasing T since the driver would get better with experience. Others thought that, if the driver had no accident for long time T, they might become more careless and have a higher hazard. So perhaps as an approximation, constant hazard is a reasonable assumption. It is this assumptions that leads to the Geometric model.

Correction to Assignment 6 Problem #2

2. In an application of the randomized response technique, a sensitive question might be "Do you currently pay any income tax in Canada?" Suppose that in a class of 500 students, each student rolls a die to determine if they will answer Q1 or Q2 below:

Q1. Did you get a 1 or a 2 face-up on the die?
Q2. Do you currently pay any income tax in Canada?

If the student gets a 1 or a 2 on the die, they answer question 1. If they get a 3,4,or 5, they answer Q2. If they get a 6, they roll again. The instructor asks for the Yes answers to the selected questions and counts 400 "Yes" responses.

a) Estimate the proportion of the 500 students that would have answered Yes to Q2.
b) If a oridinary survey of the class had selected a random sample of 300 students and obtained the proportion of Yes responses by guaranteeing anonymity in some other way (so you can assume truthful answers),what would be the SD of the estimated percentage responding Yes?
c) Do you think this SD in b) is larger, smaller, or about the same as the SD of the estimate in part a)? Say why.