Survival Model PhL 2.8.15

Motivation: I noticed that *my* CHADS2 error bars (from weighted coin flip treatment) were larger than those published by Gage, and he had a footnote that they used an expo survival model, so I studied up a bit on that. I finally figured that they used the second simple model in the Lee book (right here) but still the error bars came out too large. It has to do with how they cut off the data research for each patient, the t_2 time. It is an interesting subject associated with the term Poisson Process.

These notes are very raw, my first time ever looking at this stuff. I really should write this up better. I also have some PDF's on this subject, Lee is an entire book.

First, here is the opening stuff from Lee,

When the survival time T follows the exponential distribution with a parameter λ, the probability density function is defined as

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geqslant 0, \quad \lambda > 0 \\ 0 & t < 0 \end{cases} \tag{6.1.1}$$

The cumulative distribution function is

$$F(t) = 1 - e^{-\lambda t} \qquad t \geqslant 0 \tag{6.1.2}$$

and the survivorship function is then

$$S(t) = e^{-\lambda t} \qquad t \geqslant 0 \tag{6.1.3}$$

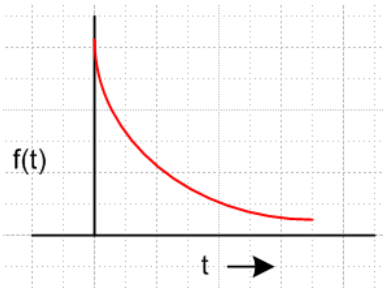
So that, by (2.2.1), the hazard function is

$$h(t) = \lambda \qquad t \geqslant 0 \tag{6.1.4}$$

a constant, independent of t. Figure 6.1 gives the graphical presentation of the three functions.

Suppose the rate of stroke is determined by something like an exponential survival model, where there is some f(t) which determines chance of having a stroke over time. And example would be

$S(t) = e^{-\lambda t}$



You then do a study where you have N patients and you run for T years and you have some subset of the patients n who have strokes, and for each one you have a time t_i when they had their stroke. How would you determine λ from such an experiment?

Well, the wiki page http://en.wikipedia.org/wiki/Exponential_distribution has some info on this subject.

The above model is a constant failure rate, and is memoryless. Here is a good fact:

Confidence intervals [edit]

The 100(1 − α)% confidence interval for the rate parameter of an exponential distribution is given by:[7]

$$\frac{2n}{\widehat{\lambda}\chi^2_{1-\frac{\alpha}{2},2n}} < \frac{1}{\lambda} < \frac{2n}{\widehat{\lambda}\chi^2_{\frac{\alpha}{2},2n}}$$

which is also equal to:

$$\frac{2n\overline{x}}{\chi^2_{1-\frac{\alpha}{2},2n}} < \frac{1}{\lambda} < \frac{2n\overline{x}}{\chi^2_{\frac{\alpha}{2},2n}}$$

where $\chi^2_{p,\nu}$ is the $100(p)$ percentile of the chi squared distribution with ν degrees of freedom, n is the number of observations of inter-arrival times in the sample, and x-bar is the sample average. A simple approximation to the exact interval endpoints can be derived using a normal approximation to the $\chi^2_{p,\nu}$ distribution. This approximation gives the following values for a 95% confidence interval:

$$\lambda_{low} = \widehat{\lambda}\left(1 - \frac{1.96}{\sqrt{n}}\right)$$
$$\lambda_{upp} = \widehat{\lambda}\left(1 + \frac{1.96}{\sqrt{n}}\right)$$

This approximation may be acceptable for samples containing at least 15 to 20 elements.[8]

In the very last, we are just applying our general theory of CI to the parameter λ.

What do the various Lee functions mean? The rate of having a stroke is a constant over time, that is what λ is. We have

$S(t) = e^{-\lambda t}$

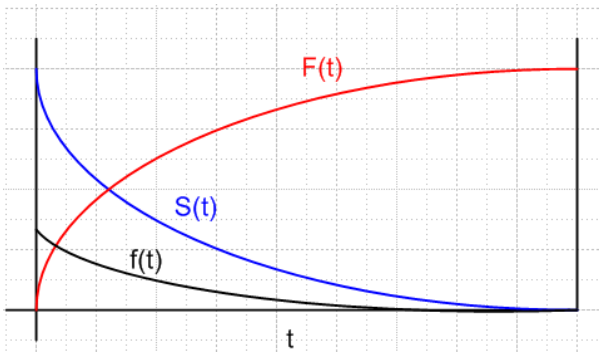
$dS(t) = -\lambda S(t)dt$

$dS(t)/S(t) = -\lambda\, dt$ = constant over time

This says that in any given time interval dt, the same *fraction* of people die and this fraction is λdt.
As time moves on, the actual *number* of people who die gets less, because there are fewer people still around and the fraction is constant.

Here are the three functions of interest defined above: (all are normalized to a population of 1_

$f(t) = \lambda e^{-\lambda t}$	absolute rate of dying at time t	f(t) = dF/dt
$F(t) = 1 - e^{-\lambda t}$	cumulative dead at time t	
$S(t) = e^{-\lambda t}$	cumulative survival at time t.	



Over time period t, here is what happens

probability of patient getting stroke during interval t = S(t)-S(0) = $e^{-\lambda t}$ - 1. ≡ F(t)

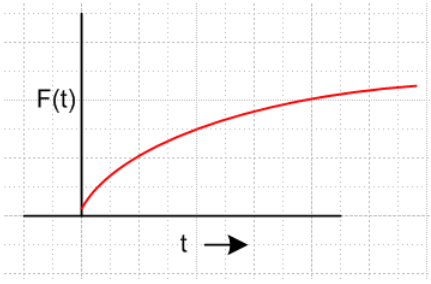
If λ is very small, this probability is very small. OK, if you start with N patients, then

n = N [$e^{-\lambda t}$ - 1] = expected number of strokes

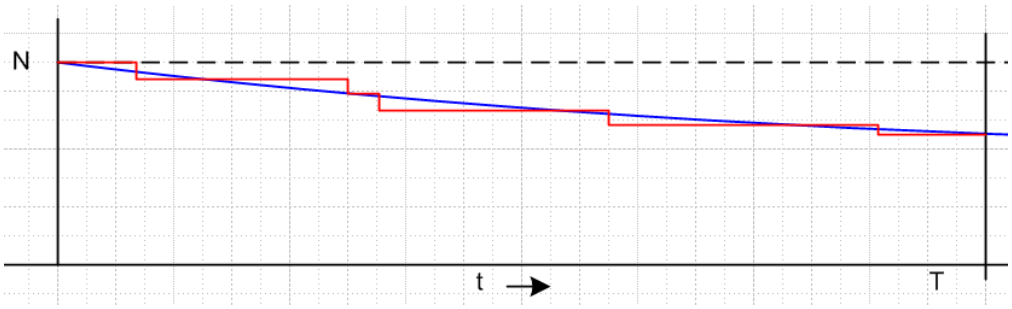
rate over time t = n/N = $e^{-\lambda t}$ - 1

r(t) = $e^{-\lambda t}$ - 1 = F(t)

Now maybe I have something! I am at least relating the stroke rate over period T to λ.
λ is the instantaneous *fractional* rate of stroke (so many percent per day)
r is the rate of stroke over time interval t

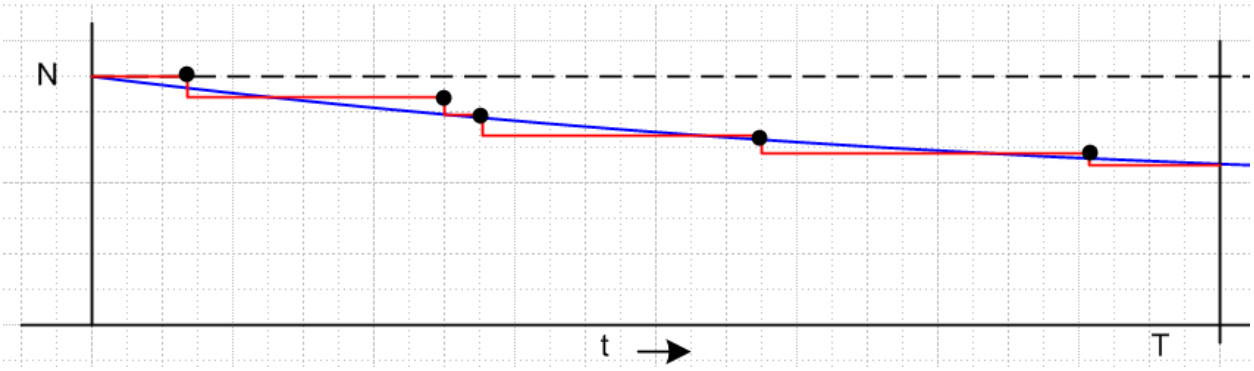


Now something that would be easy to measure for a group of N patients would be the cumulative survival situation. Consider



Here we start with N patients, and each time 1 gets a stroke, our set of surviving patients downticks by 1. The above red graph then captures the information of WHEN that patient had his stroke. Our task would then be to try to fit the red curve with an exponential blue curve of some λ, and in that way we can compute a value for λ. We are moving forward I think!

We could put little dots at the value of N(t) just before each stroke to get this



Then we could use just those dots to try and fit the exponential blue curve. I see where logs would help. We have

$$N_S(t) = N e^{-\lambda t}$$

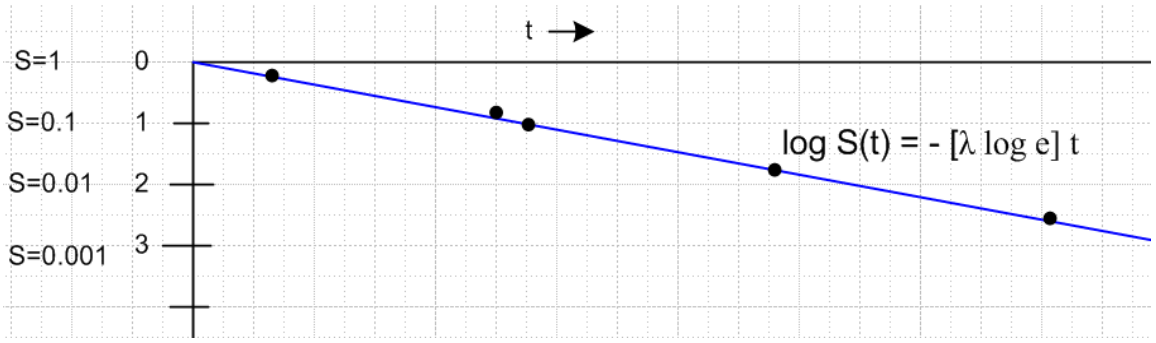
$$\ln N_S(t) = \ln N - \lambda t$$

or more easily

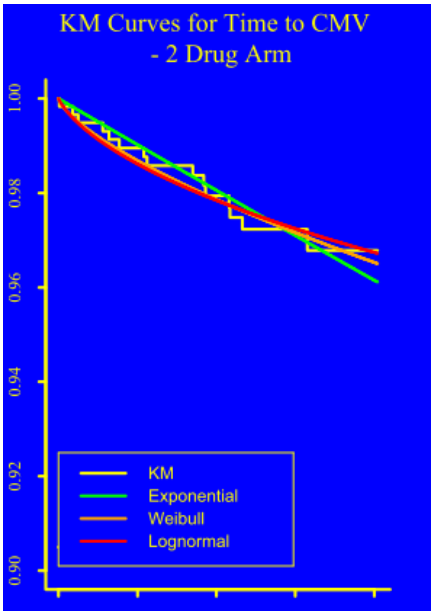
$$S(t) = e^{-\lambda t}$$

$$\log S(t) = -\lambda t \log e$$

Then we could make a linear version of the above plot.

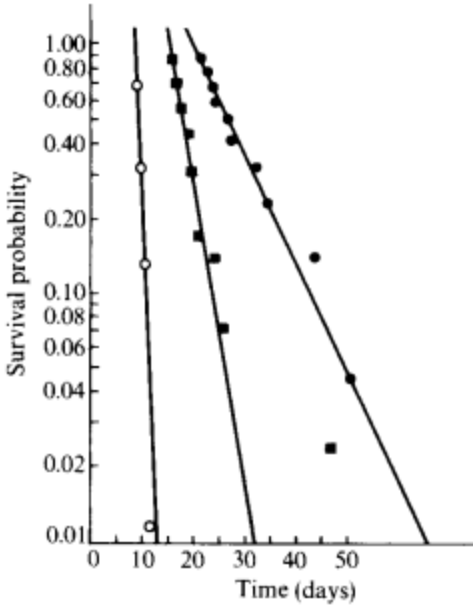


Note added later: I found in a pdf (saved) this graph supporting my interpretation of things above,

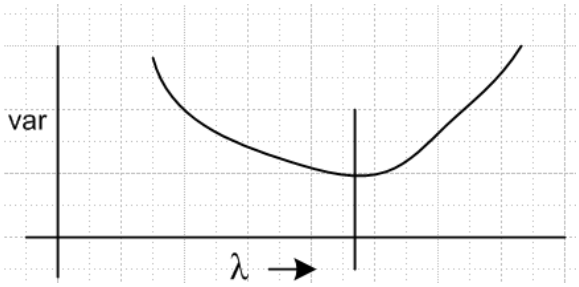


The jaggy line in yellow is the Kaplan-Meier thing, they call KM. For this example, the expo model is not bad, shown in green. The expo is really expo on this linear vertical axis plot, but looks straight.

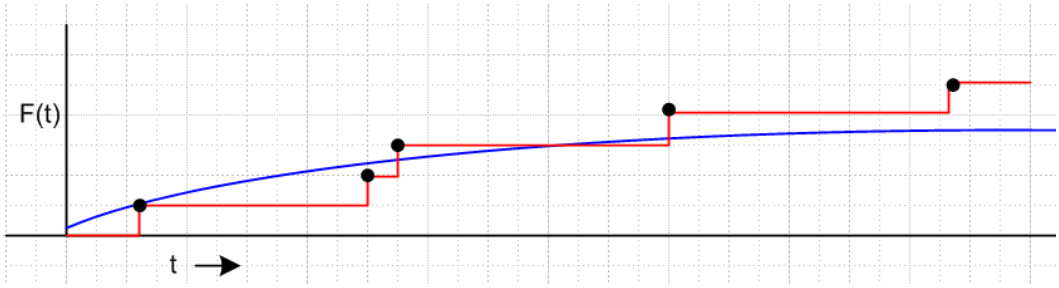
Now first of all, just the fact that the dots all lie on a line supports the expo model. Secondly, we can do a simple fit to obtain λ ! Very simple indeed. I think that is what is being plotted in Example 6.1 from the Lee book. All 25 mice in each test eventually die, but he only has to plot the deaths of the first few to establish that things are linear. The lines don't meet at the origin because this example is also using a second parameter G which is the "guarantee time" where mouse deaths are ignored. Each set of mice is fitted to its own G and λ parameters.



So far so good. Now that can be said about this "fitting to the line" situation? This is an Ng like situation. We just try to minimize the variance as a function of λ. (Ng called this variance J).



Now we could do this all a different way as follows using the cumulative non-survivor function F(t)



We have $F(t) = e^{-\lambda t} - 1$ so then

$$\log F(t) = -(\lambda \log e)\, t$$

and this would result in a similar straight line test and then fit problem.

Let's now jump to Lee Section 7.2 which seems relevant page 176 starts things off, but I am missing what **right-censored** means so push stack to find that out. OK, section 1.2 gives a very clear description. Censored data means that you don't have all the data! If some patients are alive after the trial stops, you don't know when they would have died. They "go off the end" of the study. In our studies we always know when the event occurred, so there is no "left censoring" there is only "right censoring". The other reason for right censoring is a patient does of something else or withdraws from the study. These are all practical things that really happen. So pop stack and back to Section

Section 7.1.1

> Suppose that persons were followed to death or censored in a study. Let t_1, $t_2, \ldots, t_r, t_{r+1}^+, \ldots, t_n^+$ be the survival times observed from the n individuals, with r exact times and $(n-r)$ right-censored times. Assume that the survival times follow a distribution with the density function $f(t, \mathbf{b})$ and survivorship function $S(t, \mathbf{b})$, where $\mathbf{b} = (b_1, \ldots, b_p)$ denotes unknown p parameters $b_1, \ldots, b_p$ in the distribution. As shown in Chapter 6, an exponential distribu-

The idea is to put a + sign after a censored time, fine. Now we have things depending on an Ng vector of parameters! This vector here is called **b**.

Unfortunately, Lee is now going to talk about likelihood functions, and this is a topic that Ng did not address and so I have nothing in my notes. Too bad!

He writes

$$L(\mathbf{b}) = \prod_{i=1}^{r} f(t_i, \mathbf{b}) \prod_{i=r+1}^{n} S(t_i^+, \mathbf{b})$$

where this is a joint probability thing and the first factor group describe people that died at times t_i and the second factor group describes things for those who went off the right end and survived.

represents the joint probability of observing $t_1, t_2, \ldots, t_r, t_{r+1}^+, \ldots, t_n^+$. A similar interpretation applies to continuous survival. $L(\mathbf{b})$ is called the *likelihood function* of $\mathbf{b}$, which can also be interpreted as a measure of the likelihood of observing a specific set of survival times $t_1, t_2, \ldots, t_r, t_{r+1}^+, \ldots, t_n^+$, given a specific set of parameters $\mathbf{b}$. The method of the MLE is to find an estimator of $\mathbf{b}$ that maximizes $L(\mathbf{b})$, or in other words, which is "most likely" to have produced the observed data $t_1, t_2, \ldots, t_r, t_{r+1}^+, \ldots, t_n^+$. Take the logarithm of $L(\mathbf{b})$ and denote it by $l(\mathbf{b})$,

$$l(\mathbf{b}) = \log L(\mathbf{b}) = \sum_{i=1}^{r} \log[f(t_i, \mathbf{b})] + \sum_{i=r+1}^{n} \log[S(t_i^+, \mathbf{b})] \qquad (7.1.1)$$

Then the MLE $\hat{\mathbf{b}}$ of $\mathbf{b}$ is the set of $\hat{b}_1, \ldots, \hat{b}_p$ that maximizes $l(\mathbf{b})$:

$$l(\hat{\mathbf{b}}) = \max_{\text{all b}} (l(\mathbf{b})).$$

It is clear that $\hat{\mathbf{b}}$ is a solution of the following simultaneous equations, which are obtained by taking the derivative of $l(\mathbf{b})$ with respect to each b_j:

$$\frac{\partial l(\mathbf{b})}{\partial b_j} = 0 \qquad j = 1, 2, \ldots, p \qquad (7.1.2)$$

The exact forms of (7.1.2) for the parametric survival distributions discussed in Chapter 6 are given in Sections 7.2 to 7.7. Often, there is no closed solution for the MLE $\hat{\mathbf{b}}$ from (7.1.2). To obtain the MLE $\hat{\mathbf{b}}$, one can use a numerical method. A commonly used numerical method is the Newton–Raphson iterative procedure, which can be summarized as follows.

Interesting that once again I encounter the issue for which I am reading the Buck book! So I get the idea here of maximizing this function, but why is it the right function to maximize? Ng was only interested in minimizing his cost function J($\boldsymbol{\theta}$) to get the best fit for some vector $\boldsymbol{\theta}$.

OK, I think I have it. Suppose your two people die at t_1 and t_2. The probability or "likelihood" of this happening is proportional to $f(t_1)f(t_2)$, a joint probability in the case of independence. Since this combination of the two deaths DID happen, at those two times, it is fairly probable that those two times caused the probability of this happening be at a maximum. As you vary the parameter λ, you may find that some value of λ causes this event which DID happen to in fact do so at a probability maximum. So consider

$$g(\lambda) = f(t_1;\lambda)f(t_2;\lambda) \quad = \lambda e^{-\lambda t1} * \lambda e^{-\lambda t2} = \lambda^2 e^{-\lambda(t1+t2)}. = \lambda^2 e^{-a\lambda}$$

Now $\partial_\lambda g(\lambda) = \lambda^2(-a)e^{-a\lambda} + 2\lambda e^{-a\lambda} = 0 \Rightarrow \quad -a\lambda^2 + 2\lambda = 0 \quad -a\lambda + 2 = 0 \quad \lambda = 2/a$

Then the answer is that $\lambda = 2/(t1+t2)$ which agrees with Lee for the general case of n people dying,

From (7.1.2), the MLE of λ is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} \qquad (7.2.5)$$

This solution for finding λ assumes there was no censoring. This means all n patients in the study are "followed to death". Lee goes on to say

Since the mean μ of the exponential distribution is $1/\lambda$ and a MLE is invariant under an one-to-one transformation, the MLE of μ is

$$\hat{\mu} = \frac{1}{\hat{\lambda}} = \frac{\Sigma_{i=1}^{n} t_i}{n} = \bar{t} \qquad (7.2.6)$$

and we find that the solution $1/\lambda$ is simply the average death time of all the people, very simple. So in the case that you can following everyone to death, if you assume an expo survival function, then this gives a very trivial way to find the solution λ. He then goes on to show that wiki result,

It can be shown $2n\hat{\mu}/\mu$ has an exact chi-square distribution with $2n$ degrees of freedom (Epstein and Sobel, 1953). Since $\lambda = 1/\mu$ and $\hat{\lambda} = 1/\hat{\mu}$, an exact $100(1-\alpha)\%$ confidence interval for λ is

$$\frac{\hat{\lambda}\chi^2_{2n,1-\alpha/2}}{2n} < \lambda < \frac{\hat{\lambda}\chi^2_{2n,\alpha/2}}{2n} \qquad (7.2.7)$$

where $\chi^2_{2n,\alpha}$ is the 100α percentage point of the chi-square distribution with $2n$ degrees of freedom, that is, $P(\chi^2_{2n} > \chi^2_{2n,\alpha}) = \alpha$ (Table B-2). When n is large ($n \geqslant 25$, say), $\hat{\lambda}$ is approximately normally distributed with mean λ and variance λ^2/n. Thus, an approximate $100(1-\alpha)\%$ confidence interval for λ is

$$\hat{\lambda} - \frac{\hat{\lambda}Z_{\alpha/2}}{\sqrt{n}} < \lambda < \hat{\lambda} + \frac{\hat{\lambda}Z_{\alpha/2}}{\sqrt{n}} \qquad (7.2.8)$$

where $Z_{\alpha/2}$ is the $100\alpha/2$ percentage point, $P(Z > Z_{\alpha/2}) = \alpha/2$, of the standard normal distribution (Table B-1).

Notice that (7.2.7) gives an *asymmetric* interval since the two χ^2 objects are not the same. When n > 25 you get the symmetric interval shown. Note that n is the number of people who died and also equals the number of people in this study (they all died). I suspect that if you have a low death rate, the people who did not die don't really matter (they will all have S ≈1), so the analysis above would then apply to our CHADS case where 23 of 512 died.

Lee goes on to do a little example of the above, where 21 subjects all die in the study. Using the method above he gets solution λ = .106 per week. And he finds that

A 95% confidence interval for the relapse rate λ, following (7.2.7), is approximately

$$\frac{(0.106)(24.433)}{42} < \lambda < \frac{(0.106)(59.342)}{42}$$

or (0.062, 0.150). A 95% confidence interval for the mean remission time,

Is this dramatically asymmetric?

24.4/42 = 0.58 .58 * .106 = 0.061

59.3/42 = 1.41 1.41 * .106 = 0.149

Well, you are 42% below on the left and you are 41% high on the right, so not that asymmetrical. But when converted to a range you get

.106 - .062 = 0.044 this is the low side distance

.150-.106 = 0.044 this is the high side distance

so it is still quite symmetrical, nothing like what I see in the CHADS study. But maybe it gets asymmetric if you convert from λ to rate.

Now in fact Lee goes on to do my case where some of the patients survive off the end of the study, and only r of the n people die, so n-r survive. In his example there are 10 people and 5 die and 5 survive off the end. He finds

$$\hat{\lambda} = \frac{r}{\Sigma_{i=1}^{r} t_{(i)} + \Sigma_{i=r+1}^{n} t_{(i)}^{+}} \qquad (7.2.10)$$

So the difference is that for all those that did not die, you have an extra term in the denom which in his example case is just (n-r) t+. Numerically he then gets

$$\lambda = \frac{5}{36 + 50} = 0.058 \text{ per week}$$

and $\hat{\mu} = 1/0.058 = 17.241$ weeks. A 95% confidence interval for λ by (7.2.12) is

$$\frac{(0.058)(3.247)}{(2)(5)} < \lambda < \frac{(0.058)(20.483)}{(2)(5)}$$ or (0.019, 0.119).

So in THIS case we have our 58 sandwiched in the range (19,119) so

left width = 58 - 19 = 39 right width = 119-58 = 61

and things are more seriously asymmetric. But we suspect the reason is that r << 25.

It is now 9:30 PM on 2/8/15. I have learned a lot about the "exponential survival model", and have seen how you might bet an asym range for λ when the number of deaths is < 25 or so. But the sym ranges persist for larger counts. Recall

Table 2. Risk of Stroke in National Registry of Atrial Fibrillation (NRAF) Participants, Stratified by $CHADS_2$ Score*

$CHADS_2$ Score	No. of Patients (n = 1733)	No. of Strokes (n = 94)	NRAF Crude Stroke Rate per 100 Patient-Years	NRAF Adjusted Stroke Rate, (95% CI)†
0	120	2	1.2	1.9 (1.2-3.0)
1	463	17	2.8	2.8 (2.0-3.8)
2	523	23	3.6	4.0 (3.1-5.1)
3	337	25	6.4	5.9 (4.6-7.3)
4	220	19	8.0	8.5 (6.3-11.1)
5	65	6	7.7	12.5 (8.2-17.5)
6	5	2	44.0	18.2 (10.5-27.4)

*$CHADS_2$ score is calculated by adding 1 point for each of the following conditions: recent congestive heart failure, hypertension, age at least 75 years, or diabetes mellitus and adding 2 points for having had a prior stroke or transient ischemic attack. CI indicates confidence interval.
†The adjusted stroke rate is the expected stroke rate per 100 patient-years from the exponential survival model, assuming that aspirin was not taken.

His largest number of deaths is 25, and for that he has 4.6, 5.9, 7.3

left = 5.9-4.6 = 1.3 right = 7.3 - 5.9 = 1.4 so fairly symmetric.

In general the CI ranges are only slightly asymmetric. My puzzle on starting this was merely the question of why they were not perfectly symmetric. Now I know the answer to that question.

Feb 9, 2015

First question: Now is the "rate" in the chads paper related to λ? Some definitions:

rate1: If you have 500 starting a year only trial and 25 have a stroke, rate1 = 25/500 = 5.00%

That is really the rate I have historically been dealing with.

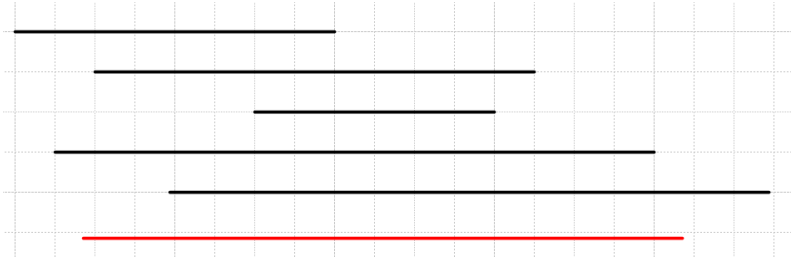
Details about the $CHADS_2$ study

How do you suppose the "trial" was "run" for chads? It was based on Medicare claims data, so retrospective I guess is the word. Yes, just checked that. A historical study = retrospective study. So how

did they do this? Patients are supposed to be AFIB patients. When did each patient "start" in this simulated study? And when did they "end"?

Data was gathered in 7 states by certain qualitative people. The first task was to ID those patients who had AFIB, fine. This was determined at the "**index hospitalization**". At that index event, the properties of each patient were determined, such as high blood pressure. Methods were normalized so all states in the study were done the same way. **MEDPAR** is a Medicare database which tracks stays at hospitals and nursing facilities. They got dates of death from other sources. OK, but not yet clear to me how this data is used. I understand that for a patient, there is some t_1 of an index visit which determines the patient is AFIB and gathers the parameters. But how is t_2 determined, the time of a stroke if there is a stroke? Well, first, after **excluding various people**, they ended up with **1733 AFIB patients** each I guess with a start time at the index visit, all in history of course. OK, the t_2 time is hospitalization for ischemic stroke. Patients were "censored" either if they died of something else, or if they lived beyond 1000 days of the index visit! Only an initial stroke counted for those with several. So OK, there is how they got t_2 times.

So here is what we have: (the red bar shows 1000 days)



Presumably the study did not include any patients whose index t_1 time was less than 1000 days from the time of writing this paper. They probably ruled those out. Then every patient in the study ends either in stroke, other death, or 1000 days duration. That is an interesting way to do it.

They then studied the λ parameter, which in this model is called "the hazard rate". It is the instantaneous fractional rate of stroke. They verified the straight line thing! They uses SAS programs called LIFEREG and LIFETEST. They claim to have obtained CI from the "binomial approximation". Stroke rate was based on "patient years of follow-up data".

Now it says the 1733 patients were "followed up" for a mean of 1.2 years (median 1.0). What exactly does that mean? Is that after time t_2 or in the gap between t_1 and t_2 ? I think it is time in the gap. But if most patients went 1000 days = 3 years with no stroke, why would the mean follow up only be 1 year? Well it says that during this time there was "readmission for ischemic stroke", so perhaps that means the follow up period starts at time t_2 of the original ischemic stroke. This would only serve to determine how many had multiple strokes, or how many died in follow up. OK, I will assume the follow-up starts at t_2 just so they can see what happened to the person who had an ischemic stroke. Nothing is said about hemorrhagic strokes! note: 2121/1733 = 1,24 years / patient

Now they say there were 2121 patient years of follow up during which 94 people were "readmitted" for ischemic stroke. Of these 94, 71 had one set of codes, and the other 24 had TIA type codes, and they counted these all the same. Now consider: 2+17+23+25+19+6+2 = 94, so the "no of strokes" column in their table does add up to 94. So again I am confused by when the follow up period starts. 27% of the 94 stroke people died in 30 days!

They excluded people out of the (65,95) range! They excluded those taking warfarin!

rin. Thus, the study cohort included 1733 patients, aged 65 to 95 years, who had nonrheumatic AF and who were not prescribed warfarin therapy at the time of hospital discharge.

This discharge would be from the index hospital visit. The list off the codes of interest. They had a minimum of 1 year of "follow-up" , so I guess yes, follow-up means period after the t_1 index visit. So the follow-up range is (365,1000) days, that is very clear. This means follow-up claims in Medicare.

event analyses. We censored deaths that were not accompanied by a hospitalization for stroke and then calculated the stroke rate based on patient-years of follow-up data[29] for each risk

29. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia, Pa: Lippincott-Raven Publishers; 1998.

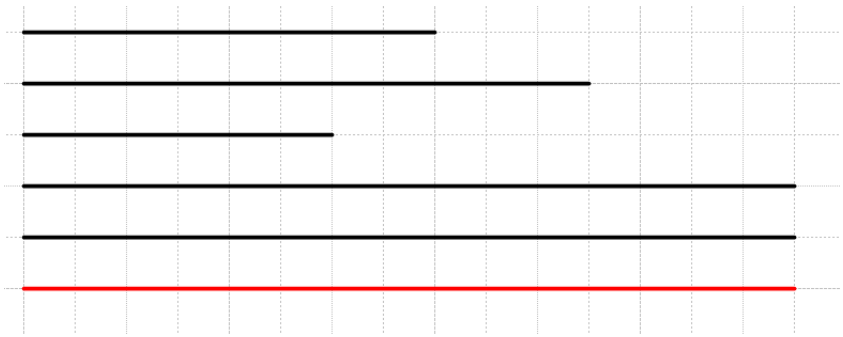
surely a fat book like Lee.

dex hospitalization. For beneficiaries who experienced multiple strokes, we excluded events and patient-days of follow-up that occurred after the initial stroke.

What does this say? Suppose someone had 2 strokes. They then excluded the 2nd stroke (event), and they excluded patient-days of follow-up after the first stroke. So this sounds like these people were just censured out of the whole study.

Now back to that idea that the mean follow-up was 1.2 years. Are they saying that they artificially limited follow-up at random over the group so this was the mean, just to reduce their work required? Maybe this was enough to elicit the results they wanted. Remember that most patients had no stroke at all. Maybe the records did not exist for many of the no-stroke patients. If they never had a stroke admission, there would be no record of anything. So I guess that would not count as follow-up if there was nothing to follow-up.

Suppose I align all the t_1 times to t = 0 to get this situation:



The bottom bar shows 1000 days. The bottom 2 black patients went 1000 days with no stroke, and most of the patients were like this. The top three patients had strokes at the end of the black line segment. They are claiming that the average length of these 3 bars was 1.2 years ? I would expect the average length of these bars to be 500 days since stroke really could happen at any time, and that is 500/365 = 1.37 years, so not too far from 1.2 years. Maybe lots of patients die of other reasons in the first 500 days and are not there for the second 500 days. I suspect that is it.

Now, if you add up the lengths of the top 3 bars, they got 2121 patient-years of follow-up, and during that time there were 94 strokes. So this means there are 94 bars that don't reach 1000 days. But this cannot be right because then sum of bar lengths = 2121 years, and only 94 bars, so then each bar would be 2121/94 = 22.56 years long, and we know that is ridiculous.

Therefore, "follow-up" must include at least the left end of the full-length bars.

So maybe follow-up really was limited in duration by lack of data for all the people. Maybe some people had no records to look at, but that would be in effect follow-up which reveals no stroke. So I now have to assume that "follow-up" includes something more than none of the full length bars, and something less than all of the full length bars. This is very confusing indeed. I need more information about this study to understand what they did. The total length of the full bars is (1733-94)*1000 days which is then (1733-94)*1000/365 = 4490 patient years if none were censured. The total length of the shorter bars had to be less than 94*1000/365 = 257.53 years. So in order to have 2121 years of follow-up, most of that had to be on the full-bars people.

What would they mean by "crude stroke rate per 100 patient years" ? 94/2121 = 0.0443 so in some sense for the entire group you get 4.43 % as stroke rate. If you knew how many of the 2121 were in each bin, you might then come up with their crude result. So I guess "crude" means you just divide these numbers. The final adjusted rates are then somehow based on the expo survival model.

Adjusted Stroke rate for CHADS2 study.

Let's assume they obtained a λ value for each bin of people in the study. How many people then have strokes in a period of 1 year? If nobody got censured during one year, answer would be F(t), which is the total number(as prob) of people who have strokes. So then rate2 = F(1year) where $F(t) = 1 - e^{-\lambda t}$. So I am claiming rate2 = $1 - e^{-\lambda}$. It would follow that $n_{strokes} = N (1 - e^{-\lambda})$, and then

annual stroke rate2 = $n_{strokes}/N_{start} = (1 - e^{-\lambda})$

Now go back to the 94/2121 = 4.43% rate. This says

annual stroke rate1 = $n_{strokes}/N_{patient\ years\ of\ follow\ up}$

How are these different? The rate1 seems incomplete since we have this arbitrariness of the amount of follow up, so maybe that is why it is crude. The rate2 comes from the expo model for which you have determined λ,. so rate 2 I think is the thing you really want.

So maybe I can indirectly deduce the λ for each bin

score 2 bin: annual stroke rate2 = 23/523 = (1 - $e^{-\lambda}$)

Solve to get

$e^{-\lambda}$ = 1 - 23/523

-λ = ln(1-23/523) = -.04497336559

So we find that

λ = .04497 = 4.497%

If you use the small λ approx you get instead

λ ≈ 23/523 = .04397705545

λ ≈ .04398 = 4.398 %

Non-stroke deaths were censured so that might affect things a bit.

Now let's do this another way using (7.2.10) shown above

$$\hat{\lambda} = \frac{r}{\Sigma_{i=1}^{r} t_{(i)} + \Sigma_{i=r+1}^{n} t_{(i)}^{+}} \qquad (7.2.10)$$

In my score 2 bin we have r = 23, n = 523, t+ = 1000 days = 1000/365 = 2.74 years (no strokers). The first sum in the denominator has to be less than 23 * 2.74, the second sum is 500 * 2.74, and is likely to be about half that. So I would say

λ = 23 / [23/2 * 2.74 + 500 * 2.74] =

```
23/((23/2)*2.74 + 500*2.74);
                              .01641087113
```

So using this method I get λ = 1.64 %/year which is a totally different result and this result is not very affected by the exact size of the first sum. The problem here is that they did not really include all those 1000 days for all the healthy people, so this method fails. If their look window was only 1.2 years on average, maybe a better answer would be larger by ratio 2.74/1.2 and that gives 3.747 %.

Now what about CI?

In the Lee model which gives

$$\hat{\lambda} = \frac{r}{\Sigma_{i=1}^{r} t_{(i)} + \Sigma_{i=r+1}^{n} t_{(i)}^{+}} \qquad (7.2.10)$$

where only r people get strokes out of your n people, the error bars are given by :

When n is large, the distribution of $\hat{\lambda}$ is approximately normal with mean λ and variance $\lambda^2/(r-1)$. An approximate $100/(1-\alpha)\%$ confidence interval for λ is then

$$\hat{\lambda} - \frac{\hat{\lambda} Z_{\alpha/2}}{\sqrt{r-1}} < \lambda < \hat{\lambda} + \frac{\hat{\lambda} Z_{\alpha/2}}{\sqrt{r-1}} \qquad (7.2.13)$$

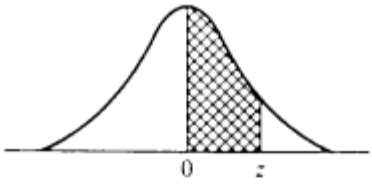
Now in Lee notation, for 95% CI you have α = .05

where $Z_{\alpha/2}$ is the $100\alpha/2$ percentage point, $P(Z > Z_{\alpha/2}) = \alpha/2$, of the standard normal distribution (Table B-1).

So for α = .05, $Z_{.025}$ is the 2.5 % point where 2.5% is in the tail on each side. Now the standard B-1 distribution has σ = 1 and μ = 0 (confirmed on web). So

$Z_{\alpha/2}$ = 1.96

Just by the way, look at his table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: Abridged from Table 1 of *Statistical Tables and Formulas*, by A. Hald, John Wiley & Sons, 1952. Reproduced by permission of John Wiley & Sons.

For 95%, one side must have .4750 of the distribution. This table shows this occurs at z = 1.96.

So OK, here is the error formula

When n is large, the distribution of $\hat{\lambda}$ is approximately normal with mean λ and variance $\lambda^2/(r-1)$. An approximate $100/(1-\alpha)\%$ confidence interval for λ is then

$$\hat{\lambda} - \frac{\hat{\lambda} Z_{\alpha/2}}{\sqrt{r-1}} < \lambda < \hat{\lambda} + \frac{\hat{\lambda} Z_{\alpha/2}}{\sqrt{r-1}} \qquad (7.2.13)$$

where r = number two died, and Z = 1.96. The condition for using the above is that n is large, not that r is large!

Apply this to our chads = 2 group with r = 23

d = $1.96/\sqrt{22}$ = .4179 Maple

1-d = .5821 maple

1+d = 1.4179 maple

Then our range is this

left = .5821 * λ left gap = d*λ right gap = d*λ symmetric

right = 1.4179 * λ

Now in my evaluation above I found for the score 2 bin,

λ ≈ .04398 = 4.398 %

Then the range is

```
L := 4.398;
```

$$L := 4.398$$

```
L*(1-d);
```

$$2.560192769$$

```
L*(1+d);
```

$$6.235807231$$

Now recall in our small λ approx we had,

annual stroke rate2 = (1 - $e^{-\lambda}$) ≈ λ = 4.398%

If we then simplify identify rate2 with λ, we would claim

rate2 = λ = 4.398% CI range = (2.560, 6.236) my computed

The data given by the chads study is this for score 2

$CHADS_2$ Score	No. of Patients (n = 1733)	No. of Strokes (n = 94)	NRAF Crude Stroke Rate per 100 Patient-Years	NRAF Adjusted Stroke Rate, (95% CI)†
0	120	2	1.2	1.9 (1.2-3.0)
1	463	17	2.8	2.8 (2.0-3.8)
2	523	23	3.6	4.0 (3.1-5.1)
3	337	25	6.4	5.9 (4.6-7.3)
4	220	19	8.0	8.5 (6.3-11.1)
5	65	6	7.7	12.5 (8.2-17.5)
6	5	2	44.0	18.2 (10.5-27.4)

So they show

rate = 4.0 CI range = (3.1, 5.1) they show

I am not very close, too bad. My computed range is larger on both sides.

Let's try again a different way. We had

annual stroke rate2 = $n_{strokes}/N_{start} = (1 - e^{-\lambda}) \approx \lambda$

Try

$\lambda \approx$ rate2f = 4.0 % from the table.

Then if r is really 23 we could get

```
L := 4.0;
```

$$L := 4.0$$

```
L*(1-d);
```

$$2.328506384$$

```
L*(1+d);
```

$$5.671493616$$

which says

rate = 4.0 CI range = (2.3, 5.7) my computed

Again my computed range is larger on both sides. So I was really unable to use their "hint" that an expo survival model was involved in order to compute their error bars.

One more time with full accuracy

rate2 = $(1 - e^{-\lambda})$ = 4.000%

$e^{-\lambda} = 1 - .04 = .96$ $\lambda = -\ln(.96) = .04082 = 4.082\ \%$

```
L := 4.0082/100;
```

$L := .04008200000$

```
L_left := L*(1-d);
```

$L_left := .02333279822$

```
L_right := L*(1+d);
```

$L_right := .05683120178$

```
1 -exp(-L_left);
```

.0230626933

```
1 -exp(-L_right);
```

.0552464714

This says

rate = 4.0 CI range = (2.31, 5.52) my computed

so the right side is a little smaller. But still my range is much larger than theirs.

Conclusion: Due to the total lack of clarity in the chads description of how it modeled this study, I cannot really compute anything at all. They fed all their data into some SAS computer program and out came all these results. Perhaps this kind of thing has been done so many times that professional readers would not question their methods.

28. Allison PD. *Survival Analysis Using the SAS System: A Practical Guide*. Cary, NC: SAS Institute Inc; 1995.

29. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia, Pa: Lippincott-Raven Publishers; 1998.

Here from my recently saved PDF:

SAS Commands for Survival Estimation

```
data leuk;
 input t;
cards;
1
1
2
2
3
4
4
5
5
8
8
8
8
11
11
12
12
15
17
22
23
;

proc lifetest data=leuk;
 time t;
run;
```

The list is death times for a set of patients, in monotonic order as Lee does it. Here is the output of the SAS program:

SAS Output for Survival Estimation

The LIFETEST Procedure

Product-Limit Survival Estimates

t	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	21
1.0000	.	.	.	1	20
1.0000	0.9048	0.0952	0.0641	2	19
2.0000	.	.	.	3	18
2.0000	0.8095	0.1905	0.0857	4	17
3.0000	0.7619	0.2381	0.0929	5	16
4.0000	.	.	.	6	15
4.0000	0.6667	0.3333	0.1029	7	14
5.0000	.	.	.	8	13
5.0000	0.5714	0.4286	0.1080	9	12
8.0000	.	.	.	10	11
8.0000	.	.	.	11	10
8.0000	.	.	.	12	9
8.0000	0.3810	0.6190	0.1060	13	8
11.0000	.	.	.	14	7
11.0000	0.2857	0.7143	0.0986	15	6
12.0000	.	.	.	16	5
12.0000	0.1905	0.8095	0.0857	17	4
15.0000	0.1429	0.8571	0.0764	18	3
17.0000	0.0952	0.9048	0.0641	19	2
22.0000	0.0476	0.9524	0.0465	20	1
23.0000	0	1.0000	0	21	0

Then there is another page

SAS Output for Survival Estimation (cont'd)

Summary Statistics for Time Variable t

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	12.0000	8.0000	17.0000
50	8.0000	4.0000	11.0000
25	4.0000	2.0000	8.0000

Mean	Standard Error
8.6667	1.4114

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
21	21	0	0.00

So that is rather interesting. Perhaps LIFETEST assumes everyone in the study died and perhaps it uses a simple expo survival model.

I think I can get unlimited info on the programs and the theory. What is missing is how the chads people really used their data and this theory. The notion of follow-up is totally hazy.

What does chads2vasc have to say on their methods? They also talk about CI using the binomial approximation. They use the ROC curve business. This was a review of past studies, not a hospital records deal. They do give CI ranges, I was wrong about that.

Let's skip to **has-bled** and see what they did. They had 5333 people, study 2003 and 2004. Study described somewhere else. All AFIB people. Did 1-year follow up to assess bleeding. Did use medical records. Their computer program set is SPSS, not SAS.

Data analysis was performed with SPSS, version 16.0 (SPSS Inc., Chicago, IL). The presence of any differences between the groups with and without major bleeding during 1-year follow-up was tested by Fisher exact test for categorical variables and by independent samples *t* test for continuous variables.

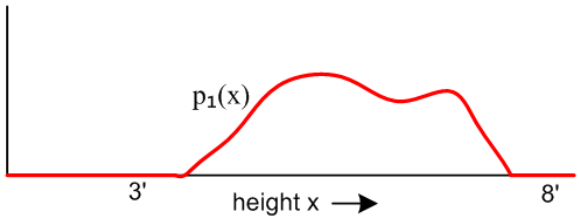
Sounds like they had for each person an exact 1 year long follow-up. I could look up the Fisher test I suppose. They just report out their results.

Why are there no CI estimates? Cannot find a web reason. Ran into Sparctool which has some interesting stuff, gets into all the NOACs.

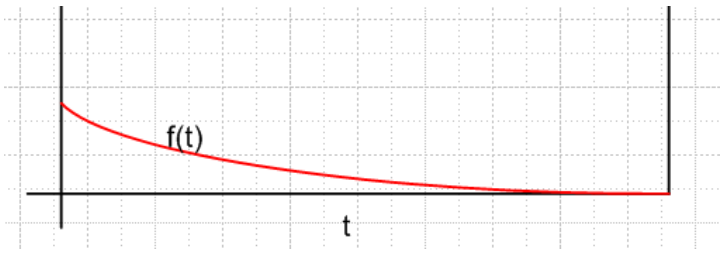
Change subject: there is a connection between the expo model and the Poisson world:

In probability theory and statistics, the **exponential distribution** (a.k.a. **negative exponential distribution**) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution, and it has the key property of being memoryless. In addition to being used for the analysis of Poisson processes, it is found in various other contexts.

Change subject: Can I fit the expo survival model into the universe of p(x) pdf examples like dice, height, stroke?



There is a pdf which is $f(t) = \lambda e^{-\lambda t}$ and it has a shape



For p(x) = f(t), we have $\mu = 1/\lambda$ and $\sigma = 1/\lambda$ (wiki). p(x) is probability of having a stroke at time x.

Unlike Bernoulli, which is a coin-toss type p(x), this one is continuous.

To get the height pdf, you grab some people and measure their heights. The experiment is a person, the outcome is height. For f(g), you grab some people and see who has strokes when.

For rolling one die, there is what I call an underlying p(x) which is p(x) = 1/6 for each face if balanced, and p(x) = something else if weighted.

For measuring one height, there is some "underlying" pdf p(x) which is probably some bell curve and which you try to figure out by measuring many heights.

For measuring stroke rate with the survival model, f(t) is the underlying p(x). But how to you measure it? If everyone had a stroke sooner or later, you could get 1,000,000 people and you just watch them and write down the time when they have their first stroke. You make a histogram with a horizontal axis of 100 years, say, and bin the strokes and you should get the f(t) above.

But if we focus on a time interval during which the rate is roughly constant, such as from 2 to 4 p.m. during work days, the exponential distribution can be used as a good approximate model for **the time until the next phone call arrives**. Similar caveats apply to the following examples which yield approximately exponentially distributed variables:

The time until a radioactive particle decays, or the **time between clicks of a geiger counter**
The time it takes before **your next telephone** call

Wiki has a nice page on the Poisson Distribution. The case above where $f_\lambda(t) = \lambda e^{-\lambda t}$, but the wiki page talks instead about $f_\lambda(k) = \lambda^k e^{-\lambda}/ k!$ which makes no sense to me at all. Elsewhere

- P(*x*; μ): The **Poisson probability** that exactly *x* successes occur in a Poisson experiment, when the mean number of successes is μ.

Well $P(x;\mu) \equiv e^{-\mu} \mu^x / x!$

Why does this have mean μ ? Why is the variance μ? What is the range of x? Well, x is supposed to be an integer, and it ranges from 0,1,2...∞. So we expect that

$1 = ?? = e^{-\mu} \Sigma_{n=0}^{\infty} \mu^n/n! = e^{-\mu} e^{\mu} = 1$, fine.

$\mu = E(n) = \Sigma_{n=0}^{\infty} P_\mu(n) [n] = \Sigma_{n=1}^{\infty} P_\mu(n) [n] = e^{-\mu} \Sigma_{n=1}^{\infty} \mu^n/n! * n$

$= \mu e^{-\mu} \Sigma_{n=1}^{\infty} \mu^{n-1}/(n-1)! = \mu e^{-\mu} \Sigma_{m=0}^{\infty} \mu^m/(m)! = \mu e^{-\mu}e^{\mu} = \mu.$

$E(n^2) = \Sigma_{n=0}^{\infty} P_\mu(n) [n^2] = e^{-\mu} \Sigma_{n=1}^{\infty} \mu^n/n! * n^2 = \mu e^{-\mu} \Sigma_{m=0}^{\infty} \mu^m/(m)! * (m+1)$

$= \mu e^{-\mu} [\mu e^{\mu} + e^{\mu}] = \mu [\mu + 1] = \mu^2 + \mu$

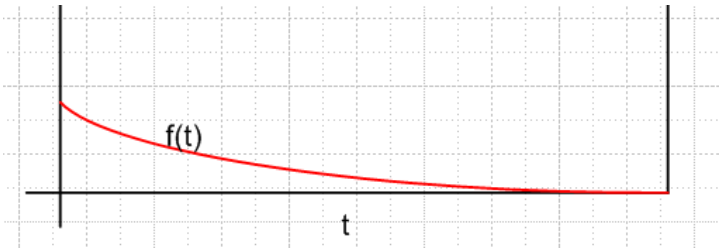
var = $\mu^2+\mu - \mu^2 = \mu$ there it is.

Wolfram shows this Poisson is the limit of the binomial where you set $\mu = Np$ and you keep μ a constant as you let $N \to \infty$. The binomial mean is Np, so it has μ as a mean as well.

The binomial is the probability of n heads versus n in a weighted coin toss repeated N times

So what is the connection between this Poisson distribution and the expo survival model? I am not sure there is any connection. I think it is the Poisson process that is of interest here. Well there is a connection, but I think I will let it lie.

Now go back to $f_\lambda(t) = \lambda e^{-\lambda t}$



as our underlying pdf which has $\mu = 1/\lambda$ and $\sigma = 1/\lambda$. But Lee showed in the simple case that $\mu = \bar{t}$ which is just the average of all the times from your experiment, similar to the height study where $\mu = \bar{h}$. So what does our little formula say,

$\sigma_{mean} = \sigma / \sqrt{N}$ $\sigma = 1/\lambda$ and the mean is $1/\lambda$, so

$\sigma_{mean} = 1/\lambda / \sqrt{N} = \bar{t}/\sqrt{N}$

Then range of mean would be

mean left = $\bar{t} - 1.96\ \bar{t}/\sqrt{N}$ mean right = $\bar{t} + 1.96\ \bar{t}/\sqrt{N}$

and write this as

mean left = $\bar{t}(1 - 1.96 /\sqrt{N})$ mean = $\bar{t}$ mean right = $\bar{t}(1 + 1.96 /\sqrt{N})$

and this agrees with the earlier quote for this simple case

$$\hat{\lambda} - \frac{\hat{\lambda} Z_{\alpha/2}}{\sqrt{n}} < \lambda < \hat{\lambda} + \frac{\hat{\lambda} Z_{\alpha/2}}{\sqrt{n}} \qquad (7.2.8)$$

where you invert the thing.

So if you did this study, you outcome would be $\bar{t}$, the mean time for having a stroke (assuming everyone has a stroke). Death would be better, and then $\bar{t}$ = mean lifetime of a person.

So finally I have applied the "general idea" that $\sigma_{\mathbf{mean}} = \sigma / \sqrt{N}$ to this particular pdf.

The problem again with chads2 is that we cannot just use the simple expo model because not everybody eventually gets a stroke. We could use the modified model Lee discusses where n-r people go off the right end never getting a stroke.

Lets try again on the Lee second model in the chads2 world. We have at score 2 a set of 523 people and of them 23 have strokes. Our mean time to stroke is then $\bar{t}$ = the following expression:

The mean survival time $\mu = 1/\lambda$ can then be estimated by

$$\hat{\mu} = \frac{1}{\hat{\lambda}} = \frac{\Sigma_{i=1}^{r} t_{(i)} + \Sigma_{i=r+1}^{n} t_{(i)}^{+}}{r} \qquad (7.2.11)$$

=

The mean "survival time" here means the mean "time from t = 0 to get a stroke". Now suppose they really did run everybody off the end at 1000 days = 2.74 years. That is just the length of the study as if they all started at once. Then in the numerator, the second sum is 500*2.74 = 1370 years. The first sum we could estimate as 23* 2.74/2 = 31.5 years (assume strokes evenly spread out since rare events). r = 23, so we get

$\bar{t}$ = [1370 + 31.5] / 23 = 60.93 years

I would then argue that this says each year you will have 1/60.93 patients die which is .0164 or 1.64%. But the number chads2 gives is 4%, not 1.6%.

The reason this is off is that this is not the method used by chads2. Remember that the mean "follow up time" was only 1.2 years, and this totally clouds the picture. If the study only ran 1.2 years with the same facts as above, you would get $\bar{t}$ = [600 + 13.8]/23 = 26.68 years, and then rate = 3.7%, closer to their number 4.

I have a feeling that their study had tricky censuring of the patients in a highly variable manner. Why didn't they just "follow up" all 1733 patients for 2.74 years, using medical records? I can see that for 94 patients who had strokes, they would stop the follow-up at the stroke time. But that is only 1/20th of the patients.

Is it possible that a huge fraction of the CHADS2 NRF cohort terminated in a censured manner? If nobody was censured other than the 2.74 years, the study would have 1733*2.74 = 4748.42 patient-years of follow-up. But in practice they only had 2121 such patient-years of follow-up. Only 94 people had a stroke, so they could have removed only 2.74*94 = 258 years at most from the total of patient years. Remember that the NRAF group had average age of 81 !! So perhaps lots of them died? The Gage paper does not say how many of the 1733 patients died less than 1000 days from their index event. If all the people were 81 years old. If you are 80, tables say you get 9 years life expectancy. So how many would die in 2.7 years? Maybe 2.7/9 = 0.3 = 30%, so that is quite a few. That would knock out 4748* .3 = 1424 years / 2 since they would die half way on average, so you only lose 712 patient years that way, and you end up with about 4000 py but they only had 2121 py that got used. Where is the other half??

Idea: maybe you have to have a positive tracking active to really count a patient year. If you simply don't hear anything on the IDC channel, that does not mean they did not have a stroke. You might have

just lost contact and you don't really know anything about that patient, so that patient is censured when it leaves the IDC radar. But how could a Medicare history stop other than death?

Idea: maybe there were limited resources for doing the IDC research, and they could not trace every single person all the way over 1000 days. On average, they only tracked for 1.2 years, and that is around half the 2.74 and maybe that is where the missing half went.

One idea of the censuring deal is that people "contribute" to the study even if their time in the study is right censured, and the reason for such a right censure might be that it cost too much in $ or time to track all the years. Just because a person is in this way censured by the researchers, they still contribute to the study for the time they were active in it. Maybe different states had different resources, and some could track all the way out to 1000 days, but most could not.

So in my group of 523 of whom 23 had strokes, it may be that half the data is lost. So my assumption that 500 of this group made it through 1000 days with no stroke is just wrong. Maybe only 50 of them did that for all I know.

Then basically for every patient of the 1733, there was a t_1 at their index event, and there was some t_2 that was either stroke, death, withdrawal, or "termination" by the research team. Then you have 1733 pairs of the form (t_1, t_2) where all you really know is that $t_2 - t_1 < 1000$ days. Then you have to use the Lee formula,

The mean survival time $\mu = 1/\lambda$ can then be estimated by

$$\hat{\mu} = \frac{1}{\hat{\lambda}} = \frac{\Sigma_{i=1}^{r} t_{(i)} + \Sigma_{i=r+1}^{n} t_{(i)}^{+}}{r} \qquad (7.2.11)$$

where the $t_{(i)}$ are times of stroke, while $t_{(i)}^{+}$ are times of censure. I have no access to this information, so I cannot compute the exact $\bar{t}$ from this formula. But maybe I know this much

$\Sigma_{i=1}^{r} t_i + \Sigma_{i=r+1}^{n} t_i^{+} = 2121$

Then we know that

$\bar{t} = 2121/94 = 22.56$ years $= 1/\lambda$

This would then give an average stroke rate for the entire group,

rate $= 1/22.56 = 0.0443 = 4.43\% = \lambda$

which is in the right ballpark I would guess. I did an estimate of this perhaps. In my 4.7 I get

<rate> = 4.83

which is not too far from the 4.43.

Comment: λ is the fraction getting stroke at any time. At t = 0 all the patients are there. If we find that $\lambda = 4.43\%$, then that is the fraction that are stroking just after the t = 0 start of the race. And that I think is the number of interest.

Now go back to earlier :

d = 1.96/$\sqrt{22}$ = .4179 Maple

1-d = .5821 maple

1+d = 1.4179 maple

This seems correct for the score = 2 bin. But the λ value I now think is 4.0 and is the same as the rate they quote. So then

r_{left} = 4.0*.5821 = 2.328

r_{right} = 4.0 * 1.4179 = 5.67

and then this says

rate = 4.0 % CI = (2.3%, 5.7%) my compute

whereas they give

rate = 4.0 CI = (3.1, 5.1) they show

so as before, my error bars are larger than theirs on both sides, but OK ballpark.

Now I figured out the χ^2 table in B-2 and redid the above using that method and I found

Lfactor = .63
Rfactor = 1.43

and then

r_{left} = 4.0*.63 = 2.52

r_{right} = 4.0 * 1.443 = 5.772

rate = 4.0 CI = (2.5, 5.8) using χ table

so I still don't get much closer to their claim. But I did this wrong, because the

Now it is true that the † thing on chads says "assuming aspirin was not taken". How would that affect things?

rebral ischemia (25%) (TABLE 1). The mean $CHADS_2$ score was 2.1 for the 1204 members of the NRAF cohort who were not prescribed any antithrombotic therapy and 2.3 for the 529 members who were prescribed aspirin.

So quite a few were taking aspirin. Maybe the 523 and 23 include aspirin takers!

aspiring takers = 529/1733 = 0.31 = 31%

Roughly, suppose then we use 69% to knock things down so there were then only 23*.69 = 16 strokes in this group who took no aspirin. But having a smaller r just makes the error bars larger.

When n is large, the distribution of $\hat{\lambda}$ is approximately normal with mean λ and variance $\lambda^2/(r-1)$. An approximate $100/(1-\alpha)\%$ confidence interval for λ is then

$$\hat{\lambda} - \frac{\hat{\lambda} Z_{\alpha/2}}{\sqrt{r-1}} < \lambda < \hat{\lambda} + \frac{\hat{\lambda} Z_{\alpha/2}}{\sqrt{r-1}} \qquad (7.2.13)$$

whereas I would like to have them be smaller.

$$\frac{\hat{\lambda}\chi^2_{2r,1-\alpha/2}}{2r} < \lambda < \frac{\hat{\lambda}\chi^2_{2r,\alpha/2}}{2r} \qquad (7.2.12)$$

As r gets smaller, the χ^2_{2r} get smaller, but the denominator r also gets smaller, so not clear what will happen. Lets try r = 16 and then we have

$\chi^2_{32,.975}$ = 18 $\chi^2_{32,.25}$ = 48 2r = 32

left factor = 18/32 = 0.56 right factor = 48/32 = 1.5

.56*4 = 2.2 1.5*4 = 6.0

rate = 4.0 CI = (2.2, 6.0) using χ table

As I expected, smaller r makes the error band larger, just as it does with the large n simple formula. ,

But in reality, they know exactly how many strokes arose in the aspirin group and I don't know that. But it must be at most 23 so no fiddling with that is going to make smaller error bands.

Conclusions:

(1) the Lee model is not really applicable to what chads2 is doing
(2) the chads2 people have too-small wrong CI numbers
(3) I am misunderstanding something