1. This item is from one of the editions of the magazine The (Scouting) Leader, published by Scouts Canada, in 1991.

   Is Scouting Safe?

   Over the past year, leaders have been showing a growing commitment to provide each member a safe and enjoyable Scouting experience. In support of efforts in the field, we conducted a study to establish baseline data on scouting accident and injury trends so that we can make informed decisions about activity precautions or the need for higher safety standards. This column highlights the findings.

   The first question we asked ourselves was, "Is Scouting a safe program for members?" Statistics Canada, Health Division, told us that 11 out of every 1,000 males aged 5-19 are hospitalized for at least one night a year. When we compared similar information taken from Scouting accident forms, we found our members are hospitalized at a rate of only one per thousand a year. Given that we run active programs and heavily use the outdoors, Scouting falls far below the average rate for daily living risk to males in this age group...

   (a) Which is the *single biggest flaw* in the analysis of the scouting injuries. List two others that might on their own might be major – but not nearly as large as the distortion produced by the *big one* ! In previous years, most students didn't recognize the flawed person-time calculations. If you don't either, come back to (a) once you have answered (b).

   (b) If Scouting members spend 2 hours per week in activities during the course of a normal 30-week Scouting year (September-June), and 2 weeks (24/7) in summer camps, how many Scouting-activity (SA) hours are contributed by 1000 scouts over the course of 1 year, including the 2 full weeks in the summer? Denote this number by "# SA-hrs generated by 1000 scouts in 1 year." From this, and the general population hospitalization rate derived in (c) below, calculate an incidence density *ratio*, i.e.,

   $$\frac{1 \text{ hospitalization } \div \{\#\text{ SA-hrs generated by 1000 scouts in 1 year}\}}{1.263 \text{ events per } 10^6 \text{ child-hours}}$$

   You can also call it an (incidence) *rate ratio.*

   Comment on the difference between this and the rate ratio implied by the magazine article – and (if necessary) revise your answer to (a).

   (c) It is not clear how the Statistics Canada data were collected; nor is it clear how the 1.1% was arrived at. Although it probably wasn't arrived at this way, we will *assume, for this exercise* that the 1.1% was a *1-year cumulative incidence.* We can then use the equation linking incidence rates and cumulative incidence, namely

   $$CI_{jan1-dec31} = 1 - \exp\left[- \int_0^{365} ID(t)\delta t \right] = 0.011,$$

   to back-calculate the *average* daily (or hourly) incidence density $\overline{ID(t)}$. We can express $\overline{ID(t)}$ as the number of hospitalizations per child-day or per child-hour.

   Exercise: Show that, in order to have $0.011 = 1 - \exp\{-\int ID(t)\}$, the average ID over the year must be 0.01106095 events per child-year, or $0.01106095/(365 \times 24)$ events per child-hour, i.e., 1.263 events per 10^6 child-hours. [1]

2. The reported percutaneous Injury (PI) rate for obstetrics/gynecology (OB/GYN) residents (Table 1 of Ayas et al, "Extended work duration & the risk of self-reported percutaneous injuries in interns" – available, if interested, in the resources for Surveys) was 0.0975 injuries/Intern-Month.

   (a) Using this incidence rate, calculate the probability that an average-risk ob/gyn resident would have no (or the complement, at least one), percutaneous Injury by the end of (a) 1 month (b) 12 months of experience? i.e. what is the probability of 'surviving' these amounts of experience without a PI? The complement is often referred to as 'cumulative incidence' or 'risk'. *Note that the integral in the formula for the relationship between incidence density (or event rate) and cumulative incidence is the expected number of events if one always had 1 resident on duty over the interval in question, even if that meant replacing one who was injured. In this case, since the ID is assumed to be constant over time, the integral, ie., the expected number, has a simple form.*

   (b) What would the 6- and 12-month 'injury-free-survival' be if the incidence density varied linearly from: (i) 0.070 at $t = 0$ to 0.013 at $t = 12$ (ii) 0.013 to 0.007 (iii) 0.0007 to 0.0013?

   What approximation suggests itself for (iii)? Devise a rule-of-thumb for when the true value & the approx'n. agree to 1 decimal place.

[1] This is *very* close to the ID of $1.1 \div (100 \times 365 \times 24) = 1.256$ events per 10^6 child-hours we would have obtained if we treated the 1.1 *not as a proportion or percentage*, but as an *incidence density* of 1.1 events per 100 child-years.]

3. The following questions (relating to discrete-in-time acts, similar to repeated 'Russian roulette') are based on part of a letter to Editor of The Lancet, May 21, 1994:

> Mastro estimated the probability of HIV-1 transmission, per sexual contact, from female prostitutes to male military prostitutes in Northern Thailand. His conservative estimate of the transmission probability, based on all men, was 0.031 (95% CI 0.025 - 0.040). In a subgroup of men not reporting a history of other sexually transmitted diseases (STDs) his estimate was 0.012 (0.006 - 0.025). He attributes this unexpectedly high value to the possible presence of STDs in female prostitutes (which may have enhanced HIV transmission) and/or high levels of infectivity among the prostitutes who are likely to be at an early stage of HIV infection.
>
> Mastro apparently overlooks these explanations and assumes that the probability of transmission of HIV between regular partners would be the same as that in prostitute-client contacts. He then used this probability of 0.031 to calculate that over 90% of initially uninfected regular partners of seropositive persons would acquire infection over 1 year. This is inconsistent with data from prospective studies in developing countries suggesting seroconversion rates among HIV-discordant partners of about 10% per year. If it is assumed that couples on average have two sexual contacts per week, then on the basis of simple probability calculations, this gives an average transmission probability per sexual contact of about 0.001 (over 30 times smaller than the conservative estimate of Mastro)

   (a) Use the Binomial distribution with $\pi = 0.031$ to arrive at an estimate of "over 90%" [second sentence of paragraph 2]; assume two sexual contacts per week on average, or 104 in a year.

   (b) Assuming again an average of two contacts per week, do the *reverse* binomial calculation [*from* the 10% 1-year seroconversion risk] that produces an estimate of "about 0.001" [last sentence of para. 2].

   (c) The 'per-act' transmission probability for human papillomavirus (H*P*V) is thought to be much higher than it is for H*I*V [cf Pubmed for PhD student Ann Burchell]. Assuming a frequency of sexual intercourse of 2 /week with an infected partner, what would the 3-month (13 week) cumulative incidence (seroconversion risk) be if (a) the per-act transmission probability was 10%? (b) this per-act transmission probability could be halved by condom use?

   (d) Using the constant 0.031 per-act probability above, plot the proportion of initially uninfected regular partners of seropositive persons that remain infection-free at various times over 1 year. To do the calculations, you might use the handy "cumulative product" function[2] `cumprod` in `R`:

```
prob.escape.infection.during.act=rep(0.969,104); prob.escape.infection.during.act
prob.remain.uninfected = cumprod(prob.escape.infection.during.act)
plot((1:104)/2,prob.remain.uninfected,
      ylab="Proportion Uninfected",xlab="Week",ylim=c(0,1.08))
points((1:104)/2,1-prob.remain.uninfected)
text(104/2,1.005-prob.remain.uninfected[104],"Cumulative Incidence",adj=c(1,0))
text(1/2,1.05,
  "S(week) = Proportion still in Initial State, i.e.
Proportion still uninfected, at indicated week",adj=c(0,0))
segments(0,1,104/2,1)
```

   Repeat with a constant 0.001 per-act probability, and comment on how the shape of the 'survival' curve, and of its complement, the cumulative incidence curve, compare with those based on a 0.031 per-act probability.

   You might be interested to adapt the "Russian Roulette" `R` code under Resources to see how cumulative incidence is determined by the per-act probabilities and the number of acts. No matter whether the 'per-act' probabilities are constant over time, or change over time, the proportion still in the initial state is the product of the conditional act-wise probabilities of remaining in the initial state, i.e., 'a fraction of a fraction of a fraction ... '

4. Refer to the 2002-2002 (current) Lifetable for Canadian females.

   (a) From the *conditional* probabilities, $p_{0\to1}$ = prop'n who reach their 1st birthday; $p_{1\to2}$ = prop'n of those reaching their 1st who reach their 2nd, etc, calculate the prop'ns $l_{25}, l_{50}, \ldots$ of a (fictitious) 'cohort' still *l*iving on their 25th, 50th, ... *[the set of q values is provided under Resources]*

   (b) Technically, each q is a proportion, or a 1-year CI. Convert each q into its equivalent incidence density, by reversing the $ID \to CI$ relationship. Then plot $\log[ID]$ versus age. How well does it follow Gompertz' law?

5. Breslow and Day (Vol I, pp 52-53) used the 1968-1972 incidence[3] rates in table 2.3 to arrive at, in Table 2.4[4], the cumulative incidence (risk) for 4 cancers over various age spans.

[2] The cum. products of the probabilities $P_1, P_2, P_3, \ldots$ is P_1, $P_1 \times P_2$, $P_1 \times P_2 \times P_3, \ldots$

[3] In epidemiology, incidence is typically used in relation to receiving a diagnosis of a illness; although deaths from an illness do, strictly speaking, pertain to a type of incidence, epidemiologists use the term mortality rates, rather than incidence rates, in this context.

[4] Note that what we (and survival analysis texts) call the 'integrated hazard' (Λ), Breslow & Day call a 'cumulated rate'.

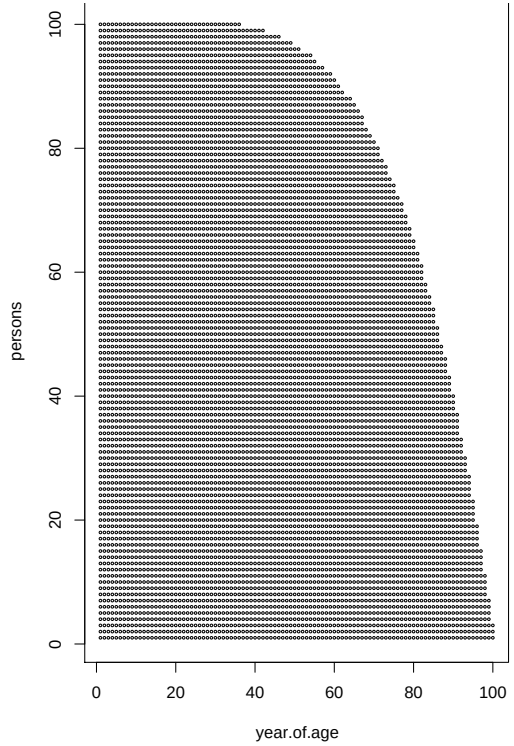
For these same age spans, and also the age span 0-85 (close to "lifetime"), calculate the risk of dying from (a) lung cancer (b) breast cancer (c) any cause. Use as input to these risk calculations the cancer-specific and all-cause mortality rates[5] observed for Québec females in 2002.[6] For (a) and (b) ignore competing mortality, i.e., assume that one cannot die of another cause first.

6. The notes give an expression for the expected value of the $Y_{l.b.}$ be the random variable representing the value of Y in a unit selected by Length-Biased Sampling. Derive this expression.

7. Let T be a positive r.v. denoting the longevity of a randomly selected product/item/person (e.g. ink cartridge, battery, computer, iPod, or human). Denote the associated cumulative distribution function by $F_T(t)$, the survival function by $S_T(t) = 1 - F_T(t)$, the probability density function by $f_T(t)$, and the expectation $\int_0^\infty t f_T(t)dt$ by μ_T. Show that

$$\mu_T = \int_0^\infty S_T(t)dt.$$

Heuristically: the mean longevity of 82.21 years in Fig. 1 is the total number of person-years (8221 P-Y) ÷ the number of persons (100). The 8221 P-Y can be seen as the sum of the lengths of the horizontal lines (i.e., first sum the years for the same person, and then sum over persons) or the sum of the lengths of the vertical lines (i.e., first sum the persons for the same year of age, and then sum over years).



8221 years lived by 100 persons.

8. Consider the 'potential years of life lost' (PYLL) by a woman who dies of cervical cancer at age 45, i.e., the additional years she could have expected to live had she been protected against this cancer. Because the equation in (b) may not be valid beyond 85, use the life-span 45-85 for both (a) and (b).

   Calculate (i) the (conditional) probability that a woman who reaches her 45th birthday will be alive on her 85th birthday, and (ii) the expected (mean) number of additional years that women who reach 45 will live over the next 40 years. Determine (i) and (ii) in 2 ways, based on the...

   (a) 2000-2002 (Current) Complete Life-table, Canadian Females

   (b) hazard rate / mortality rate / incidence density function[7], $h(age)$, fitted to the observed age-specific all-cause mortality rates for

[5] Calculate mortality rates as no. of deaths ÷ (mid-year population size × 1 year). {Had numbers of deaths for say each of the years 2001-2003 been available to you, you could have aggregated them to arrive at a more stable rate, namely the combined number of deaths ÷ (2002 population size × 3 years), or ÷ (sum of the person years in the 3 years in question).}

[6] Data provided in a Excel file, and a .csv file that can be read into R or SAS, located under the heading 'Datasets and programs' at the bottom of the Resources webpage for epidemiology/concepts/measures.

[7] Gompertz (1779-1865) observed in 1825 that the force (intensity, I) of mortality at age a had the form $I_0\beta^a$ over a wide age-span i.e., age-specific death rates were log-linear-in-age (Gompertz 'Law of Mortality'). Random variables whose hazard functions follow this form are said to follow the Gompertz Distribution.

Québec women aged 45-85 in 2002:

$$\log h(age) = -6.7 + 0.10 \times (age - 45).$$ see footnote[8]

9. Calculate the not-for-profit[9] 1-year life-insurance premium for

   (a) a Canadian woman aged 50, in "average" health, based on the 2000-2002 (Current) Complete Life-table for Canadian females;

   (b) a Québec woman aged 50, in "average" health, based on the fitted hazard function given above.

10. The English actuary T R Edmonds (some of his writings are available under Resources for Epidemiology) claimed that the force of mortality function across the entire age range could be simplified to just 3 (or 4 or 5) constants – see pages 170-179 of Woods for an easier to read version. Examine the modern day force of mortality function to see if the same parsimony applies today: use the year-by-year q values from the current Canadian life table, females, 2002-2002 [under Lifetables... Examples in Resources for Epidemiology] and plot the older and modern $\log(q)$ functions on the same graph. *Note*: Woods defined q_x as if it were unconditional, whereas q_x is in fact *conditional* on reaching the previous year – the q's are not unconditional probabilities that add to 1 (the entries in the d_x column in a modern lifetable are unconditional, and their sum is either 1, or 100,000 or whatever is the radix of the table.) The product of the complements of the q's (i.e., the product of the p's) gives the survival function.

11. Premature Death in Jazz Musicians: Fact or Fiction? *letter from retired professor to American J Public Health 1991 June; 81(6): 804-805*

> "Jazz musicians tend to be more liable than other professions to die early deaths from drink, drugs, women, or overwork.[10]
>
> "The career of the ODJB (Original Dixieland Jazz Band) was both as fantastic and as typical as any that jazz has had to offer. Its story features... the petty jealousies, alcoholism, premature deaths, and all the rest."[2] "Catlett's career was a singularly queer one, even for jazz, whose history is filled with the wreckage of poverty, sudden obscurity, and premature death."[3]

Statistical study of 86 jazz musicians listed in a university syllabus refutes these tenets,[4] the second and third of which were made by two of America's most respected critics, and all of which foster the commonly held view that jazz players die prematurely. Dates of birth, and of death when it had occurred, were tabulated, and longevity matched with that expected in the United States by year of birth, race, and sex.[5–7] One musician who had not reached the age of his life expectancy was excluded from the list; the musicians were born in the US.

Birth years ranged from 1862 to 1938; 16 births occurred before 1900, 23 between 1900 and 1909, 19 between 1910 and 1919, 22 between 1920 and 1929, and five between 1930 and 1939. Comparison with national values showed that 70 (82%) of the musicians exceeded their life expectancy; four-fifths of the Black men, three fourths of the White men, and all the women lived longer than expected as shown in this frequency distribution.

	Male			Female		
	Total	n	%	Total	n	%
White	19	14	74	-	-	-
Black	59	49	83	7	7	100

Jazz was born in the "sporting houses" of New Orleans and nurtured in the speakeasies and night clubs of Chicago, Kansas City, and New York. Its association with vice and crime in its early days has led to the assumption that to play jazz is to court depravity and death. Although the size and sex distribution of the sample limits the inferences to be drawn, the data suggest that jazz musicians do not die young. Most of the 85 musicians in this study have survived the potential hazards of irregular hours of work and meals, the ready temptation of drugs and alcohol, and the perils of racial prejudice, and to have overcome "the problem of the artist who is creative within a socially and racially discriminatory world."[8]

Questions:

(a) Give two reasons why the author's comparison group gave the jazz

[8] Integral of $h(age)$ has closed form; or, could use num'l integration, e.g. `integrate` in `R`

[9] Such that in a large number of such insured persons, the premiums collected would just balance the total amount of the death benefits (each one valued at $\$10,000$) paid out.

[10] [1] Lindsay M: Teach Yourself Jazz. London: English Universities Press, 1958. [2] Schuller G: Early Jazz. Its Roots and Musical Development. New York: Oxford University Press, 1968: 176. [3] Balliett W: The Sound of Surprise. New York: Da Capo Press, 1918: 144. [4] Norton P, Schumacher HJ: Topics in American Culture: Jazz Styles. Ann Arbor, MI: University of Michigan Extension Service, 1978; xiv-xvi. [5] Chilton J: Who's Who of Jazz. Philadelphia: Chilton Book Co. 1972. [6] Feather L: The Encyclopedia of Jazz. New Ed. Bonanza Books, 1960. [7] US Department of Commerce: Historical Statistics of the United States. Colonial Times to 1957. Washington, DC: Govt Print ing Olfice, 1461; 24-25. [8] Berendt J: The Jazz Book. New York: Lawrence Hill, 1915: 256.

musicians an unfair longevity advantage. [11] [12]

(b) "Comparison with national values showed that 70 (82%) of the musicians exceeded their life expectancy". In the Canadian lifetable used in the previous exercise, what percentage of 100000 newborns would be expected to exceed the life expectancy at birth? Comment![13]

(c) What is the shape of the distribution of the ages-at-death in that lifetable? Does this explain how, as in Lake Wobegon, `http://en.wikipedia.org/wiki/Lake_Wobegon_effect`, more than 50% can genuinely be "above average"?[14]

12. Run the java applet for the Bridge of Life for Sweden, Female, 1751-1851 (cohort); England, Male, 1871-1880 (current); and Canada, Female, 2000-2002 (current) until you consider the hazard functions have stabilized. What do you think is the factor that governs the stability of the hazard and log-hazard function at a given age? *The applets can be accessed using the Bridge of Life link at the bottom left of JH's home page.*

13. The important but seldom-visited article "Tumbler Mortality" by Brown and Flood in JASA in 1947 shows the "survival" of tumblers (Free Online Dictionary: a. A drinking glass, originally with a rounded bottom. b. A flat-bottomed glass having no handle, foot, or stem.) in a cafeteria. The article is available under Resources for Epidemiology. Note that whereas the authors used the word truncation for the observations on tumblers that were still in service at the end of the test, we would use the word 'right-censored' today.

(a) Using the data in Table 1, try to re-create the week-by progress of the "service-test" (it ran for 78 weeks) by creating an array of 60 rows and 78 columns. Let row i represent the i-th original tumbler and its replacement and its replacement, and so on. and column j the j-th week. The entries should represent the number of weeks the currently-being tested tumbler has been under test, so that a row that starts with the integers

$$1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1, 2, 3, 1, \ldots$$

tells us that the first tumbler failed during its 5-th week of service, its replacement during its 11-th, its replacement during its 3-rd, etc. Do so 'by hand' or using (random) durations generated from the fitted distribution.

(b) How long does it seem to take before the data in a *column* represent the steady state age-distribution of the tumblers in service? And what shape would that age-distribution have?

Hint: to understand, use a very simple example, with only 2 possible lengths of service, and make it reasonably realistic, switch contexts: imagine 2/3rds of hospital admissions for a specific procedure involve a hospital stay of 2 days, and 1/3rd involve 5 days, so the average stay is 3 days. Say the hospital unit has a limited no. of beds, and that they are always full, so that on the day a bed becomes free, there is a new patient admitted to it. Now imagine, at steady state, a cross-sectional survey. It will preferentially capture the longer stays, and indeed the ratio of the number of long stays to short stays will not be $1:2$ but $1 \times 5 \ : \ 2 \times 2$ or $5:4$. On average, 1 of the 5 long stays will be captured on the first day of the stay, 1 on the 2nd, *dots* , 1 on the 5th. Likewise, on average, 1 of the 4 short stays will be captured on the first day of the stay, and on the 2nd. So the 9 'ages' will be 1, 2, 3, 4, 5, 1, 1, 2, and 2, and so the distribution will be

'age':	1 day	2 days	3 days	4 days	5days
frequency:	3	3	1	1	1

Now, how does the shape and location of this distribution of the (cross-sectional) ages relate to the shape and location of the distribution of the stays (or longevity)?

It is left as an exercise to show that it is proportional to the survival function $S(t) = 1 - F(t)$ of the distribution of the lengths of stay. i.e., if we denote by the pdf $g(t)$ of the cross-sectional $(x.s.)$ ages , and by $f(t)$ the pdf whose cdf is $F(t)$ and survival function is $S(t)$, then

$$g_{x.s.}(t) \propto S(t).$$

[11] Hint: one has to do with a *vital* –pun intended– requirement for becoming a famous jazz musician; the other with the difference between current and truly cohort lifetables, and the trends in age-specific US mortality rates over the last century. Drawing lifelines on a Lexis diagram for years 1862-1991, ages 0-100, may help illustrate the 2 issues.

[12] If interested in a fairer comparison, you could look at the Methods in the 'Elvis to Eminem' study (bottom of Resources for risk/cumulative incidence, lifetables, webpage).

[13] JH suspects that the author, if asked, would have argued: 'under the null hypothesis, the expected percentage that exceeded the life expectancy at birth should be 50%.'

[14] The editor of Amer. Journal of Roentgenology missed this point in JH's 1994 article, as did the British newspaper *The Independent* when it wrote "The usually wonderful Jeremy Paxman, introducing a Newsnight discussion last Friday on the teaching of reading skills, expressed dismay that 'a third of our primary schoolchildren have below-average reading ability'. Had he paid more attention in his 'rithmetic lessons, perhaps Paxman would have realised that half our schoolchildren are below average in everything. As, indeed. are half our Newsnight presenters."

Another way to gain some intuition for this is to realize that the age-distribution of stays that end each calendar day must be represented by $f(t)$: after all, that's what f is. But, these endings (they are called 'separations' in the hospital business) are produced by applying the age-specific 'ending-rates' to the age-specific numbers of candidates for separation, ie by multiplying the 'hazard function' by the g function, i.e.,

$$f(t) \propto h(t) \times g_{x.s.}(t).$$

But the very definition of the hazard function is

$$h(t) = \frac{f(t)}{S(t)},$$

and in our case, $g_{x.s.} \propto S(t)$, so we have

$$g_{x.s.}(t) \propto \frac{f(t)}{h(t)} = S(t).$$

Indeed,

$$g_{x.s.}(t) = \frac{S(t)}{\int (S(u)du}.$$

See the website for some `R` code that shows the steady state age-distibution of tumblers on test.

(c) The authors fit a parametric form to the longevity distribution. This was before the classic 1958 paper by Kaplan and Meier, who showed how to calculate a non-parametric estimate of the survival function. Can you think of how to do this? [Hint: see question 4(a)]

14. See the fascinating article "Evidence for Cardiomyocyte Renewal in Humans" by Bergmann (senior author Frisén) in *Science* in 2009 on the re-generation of heart muscle (under Resources for Epidemiology).

    In Fig 4 the authors fit (a) the turnover rate and (b) its inverse as straight line functions of age. When you substitute these functions into the fundamental formula linking incidence and cumulative incidence or survival, what fractions of cardiomyocytes remaining from birth are predicted for ages 25, 50 and 75?