1 Probability models

1.1 Observation, experiments and models

Stochastic Models[1]

Normal vs *Bernoulli* and *Poisson*: We need to distinguish between *individual* observations, governed by Bernoulli and Poisson (or if quantitative rather than all-or-none or a count, Normal) and *statistics* formed by aggregation of individual observations. If a large enough number of individual observations are used to form a statistic, its (sampling) distribution can be described by a Gaussian (Normal) probability model. So, ultimately, this probability model is just as relevant.

1.1.1 Epidemiologic [subject-matter] models [JH]

We need to also make a distinction between the quantity(quantities) that is(are) of substantive interest or concern, the data from which this(these) is(are) estimated, the *statistical* models used to get to the the quantity(quantities) and the relationships of interest.

For example, of medical, public health or personal interest/concern might be the

- level of use of cell phones while driving
- average and range [across people] of reductions in cholesterol with regular use of a cholesterol-lowering medication
- amount of time taken by health care personnel to decipher the handwriting of other health care personnel
- (average) number of times people have to phone to reach a 'live' person
- reduction in one's risk of dying of a specific cancer if one is regularly screened for it.
- appropriate-size tracheostomy tube for an obese patient, based on easily easily obtained anthropometric measurements
- length of central venous catheter that can be safely inserted into a child as a function of the child's height etc.
- rate of automobile accidents as a function of drivers' blood levels of alcohol and other drugs, numbers of persons in the car, cell-phone and other activities, weather, road conditions, etc.
- Psychological Stress, Negative Life Events, Perceived Stress, Negative Affect Smoking, Alcohol Consumption and Susceptibility to the Common Cold
- The force of mortality s a function of age, sex and calendar time.
- Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction
- Are seat belt restraints as effective in school age children as in adults?
- Levels of folic acid to add to flour, so that most people have sufficiently high blood levels.
- Early diet in children born preterm and their IQ at age eight.
- Prevalence of Down's syndrome in relation to parity and maternal age.

Of broader interest/concern might be

- the wind chill factor as a function of temperature and wind speed
- how many fewer Florida votes Al Gore got in 2000 because of a badly laid-out ballot
- a formula for deriving one's "ideal" weight from one's height
- yearly costs under different cell-phone plans
- yearly maintenance costs for different makes and models of cars
- car or life insurance premiums as a function of ...
- cost per foot^2 of commercial or business rental space as a function of ...
- Rapid Changes in Flowering Time in British Plants
- How much money the City of New York should revover from Brink's for the losses the City incurred by the criminal activities of two Brink's employees (they collected the money form the parking meters, but kept some of it!).

[1] 'Stochastic' `http://www.allwords.com/word-stochastic.html` French: stochastique(fr) German: stochastisch(de) Spanish: estocstico(es) Etymology: From Ancient Greek (polytonic,), from (polytonic,) "aim at a target, guess", from (polytonic,) "an aim, a guess". Parzen, in his text on Stochastic Processes .. page 7 says: <<The word is of Greek origin; see Hagstroem (1940) for a study of the history of the word. In seventeenth century English, the word "stochastic" had the meaning "to conjecture, to aim at a mark." It is not clear how it acquired the meaning it has today of "pertaining to chance." Many writers use the expression "chance process" or "random process" as synonyms for "stochastic process."<<

1.1.2 From behaviour of statistical 'atoms' to statistical 'molecules'

1 condition' or 1 circumstance' or 'setting' [also known as "1-sample problems"]

The smallest statistical element or unit (?atom): its quantity of interest might have a Y distribution that under sampling, could be represented by a discrete random variable with '2-point' support (Bernoulli), 3-point support, $k-$point support, etc. or interval support (Normal, gamma, beta, log-normal, ...)

The *aggregate* or summary of the values associated with these elements is often a sum or a count: with e.g., a Binomial, Negative Binomial, gamma distribution. Or teh summary might be more complex – it could be some re-arrangement of the data on the individuals (e.g., the way the tumbler longevity data were summarized). This brings in the notion of "sufficient statistics".

More complex: t, F, ...

2 or conditions' or 1 circumstances' or 'settings', indexed by possible values of 'X' variable(s). Think of the 'X' variable(s) as 'covariate patterns' or 'profiles.'

unknown conditions or circumstances Sometimes we don't have any measurable (or measured) 'X' variable(s) to explain the differences in Y from say family to family or person to person.There instead of the usual multiple regression approach, we use the concept of a hierarchical or random-effects or latent class or mixture model.

1.2 Binary data

It is worth recalling from the first semester, the concepts of states and events (transitions from one state to another).

Cohort studies with fixed follow-up time

Recall: *cohort* is another name for a closed population, with membership (entry) defined by some event, such as birth, losing one's virginity, obtaining one's first driver's permit, attaining age 21, graduating from university, entering the 'ever-married' state, undergoing a certain medical intervention, enrolling in a follow-up study, etc. Then the *event of interest* is the *exit* (transition) from a/the state that prevailed at entry. So *death* is the transition from the *living* state to the *dead* state, receiving a *diagnosis* of cancer changes one's state from 'no history of cancer since entry' to 'have a history of cancer', being convicted of a traffic offense changes one's state from 'clean record' to 'have a history of traffic offenses.' We can also envision more complex situations, with a transition from 'never had a stroke,' to 'have had 1 stroke,' to 'have had 2 strokes,' ... or 'haven't yet had a cold this winter,' to 'have had 1 cold,' to 'have had 2 colds,' etc.

Censoring: to be distinguished from *truncation*. Truncation implies some observations are missed by the data-gathering process, i.e., that the observed distribution is a systematic distortion of the true distribution. Note that we can have censoring of any quantitative variable, not just one that measures the duration until the event of interest. For example, the limits on say a thermometer or a weight scale or a chemical assay may mean that it cannot record/detect values below or above these limits. Also, the example in C&H implicitly refers to *right* censoring: one can have *left* censoring, as with lower limits of detection in a chemical assay, or *interval* censoring, as – in repeated cross-sectional examinations – with the date of sero-conversion to HIV.

Incidence studies: the word *new* means a change of state since entry.

"*Failure*": It is a pity that C&H didn't go one step more and use the even more generic term "*event*". That way, they would not have to think of graduating with a PhD (i.e., *getting out of – exiting from – here*) as "*failure*" and still being here" as "*survival*." This simpler and more general terminology would mean that we would not have to struggle to find a suitable label of the 'y' axis of the $1 - F(t)$, usually called $S(t)$, function. One could simply say *"proportion still in initial state,"* and substitute the term for the initial state, i.e., proportion still in PhD program, proportion event-free, etc.

N or n? D or d? JH would have preferred lower case, at least for the denominator. In *sampling* textbooks, N usually denotes the *population* size, and n the *sample* size. In the style manual used in *social sciences*, n is the sample size in each stratum, whereas N is the overall sample size. Thus, for example, a study might report on a sample of $N = 76$ subjects, composed of $n = 40$ females and $n = 36$ males.

Cohort studies with variable follow-up time: If every subject entered a study at least 5 years ago, then, in principle, one should be able to determine D and $N - D$, and the 5-year survival proportion. However, *losses to follow-up* before 5 years, and before the event of interest, lead to observations that are typically regarded as censored at the time of the loss. Another phenomenon that leads to censored observations is *staggered entry*, as in the JUPITER trial. Unfortunately, some losses to follow-up may be examples of *informative*' censoring.

Cross-sectional prevalence data

Recall again that prevalence refers to a *state*. Examples would include the

proportion (of a certain age group, say) who wear glasses for reading, or have undetected high blood pressure, or have high-speed internet at home, or have a family history of a certain disease, or a certain 'gene' or blood-type.

From a purely *statistical* perspective, the analysis of *prevalence* proportions of the form D/N and *incidence* proportions of the form D/N takes the same form: the underlying statistical 'atoms' are N Bernoulli random variables.

1.3 The binary probability model

JH presumes they use this heading as a shorthand for 'the probability model for binary responses' (or 'binary outcomes' or binary random variables)

... to "*predict* the outcome" : JH takes this word *predict* in its broader meaning. If we are giving a patient the probability that he will have a certain *future* event *say within the next 5 years*, we can talk about predicting the outcome: we are speaking of *pro*gnosis; but what if we are giving a woman the probability that the suspicious finding on a mammogram does in fact represent an existing breast cancer, we are speaking of the *present*, of whether a phenomenon already *exists*, and we use a prevalence proportion as an estimate of the *dia*gnostic probability. Note that prevalence and incidence refer to aggregates.

The risk parameter

Risk typically refers to the *future*, and can be used when speaking to or about one person; we don't have a comparable specialized term for *the probability that a state exists* when speaking to or about one person, and would therefore just use the generic term probability.

The odds parameter

The sex-ratio is often expressed as an odds, i.e., as a ratio of males to females. If the proportion of males is 0.51, then the male:female ratio is 51:49 or (51/49):1, i.e., approximately 1.04:1. This example is a good reason why C&H should have used a more generic pair of terms than failure and survival (or success and failure).

In betting on horse races (at least where JH comes from), odds of 3:1 are the odds *against* the horse winning; i.e., the probability of winning is 1/4. When a horse is a heavy favourite so that the probability of winning was 75%, the "bookies" would give the odds as "3:1 *on*."

Rare events

One of the tricks to make events *rare* will be to slice the time period into small slices or windows.

Death, the first of the two only sure events (taxes is the other) is also rare - in the short term!

Also, it would be more correct to speak of a *rare events*, since disease is often used to describe a process, rather than a transition. And since most transitions are rapid, the probability of a transition (an event) occurring within a given short sub-interval will usually be small.

If the state of interest being addressed with cross-sectional data is uncommon (or rare), then yes, the prevalence odds and the prevalence proportion will be very close to each other.

Supplementary Exercise 1.1. If one rounds probabilities or risks or prevalences (π's), or their corresponding odds, $\Omega = \pi/(1-\pi)$, to 1 decimal place, at what value of π will the rounded values of π and Ω be differerent? Also, why use lowercase π for proportion, and uppercase Ω for odds?

1.4 Parameter Estimation

Should you be surprised if the estimate were π were other than D/N? Consult Google or Wikipedia on "the rule of succession," and on Laplace's estimate of the probability that the sun will rise tomorrow, given that it has unfailingly risen ($D = 0$) for the past 6000 years, i.e., $N \approx 365 \times 6000$.

Supplementary Exercise 1.2. One has 2 independent observations from the model

$$E[y|x] = \beta \times x.$$

The y's might represent the total numbers of typographical errors on x randomly sample pages of a large document, and the data might be $y = 2$ errors in total in a sample of $x = 1$ page, and $y = 8$ errors in total in a separate sample of $x = 2$ pages. The β in the model represents the mean number of errors per page of the document. Or the y's might represent the total weight of x randomly sample pages of a document, and the data might be $y = 2$ units of weight in total for a sample of $x = 1$ page, and $y = 8$ units for a separate sample of $x = 2$ pages. The β in the model represents the mean weight per page of the document. We gave this 'estimation of β' problem to several statisticians and epidemiologists, and to several grade 6 students, and they gave us a variety of estimates, such as $\hat{\beta} = 3.6$/page, 3.33/page, and 3.45!

How can this be?

1.5 Is the model true?

I wonder if they were aware of the quote, attributed to the statistician George Box that goes something like this

> "all models are wrong; but some are more useful than others"

`http://en.wikiquote.org/wiki/George_E._P._Box`

2 Conditional probability models

2.1 Conditional probability

JH is suprised at how few textbooks used trees to explain conditional probabilities. Probability trees make it easy to see the direction in which one is preceeding, or looking, where simply algebraic symbols can not, and make it easier to distinguish 'forward' from 'reverse' probabilities.

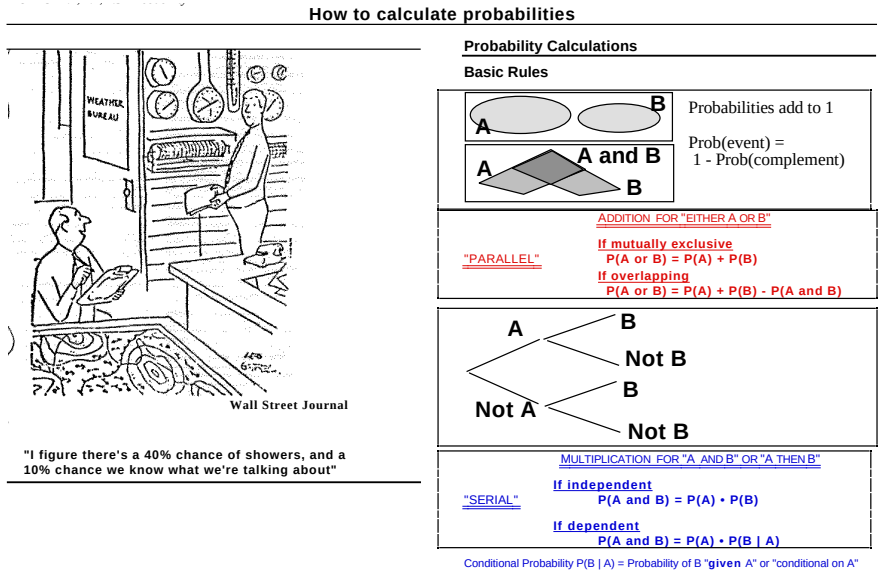

Figure 1: From JH's notes for EPIB607, introductory biostatistics for epidemiology

Trees show that the probability of a particular sequence is always a fraction of a fraction of a fraction .. , and that if we start with the full probability of 1 at the single entry point on the extreme left, then we need at the right hand side to account for all of this (i.e., the 'total') probability.

Statistical dependence and independence

JH likes to say that with independence, one doesn't have to look over one's shoulder to the previous event to know which probability to multiple by.. The illustrated example on the gender composition of 2 independent births, and of a sample of 2 persons sampled (without replacement) from a pool of 5 males and 5 females, show this distinction: in the first example, when one comes to the second component in the probability product, $Pr(y_2 = male)$ is the same

whether one has got to there via the 'upper' path, or the 'lower' one. know

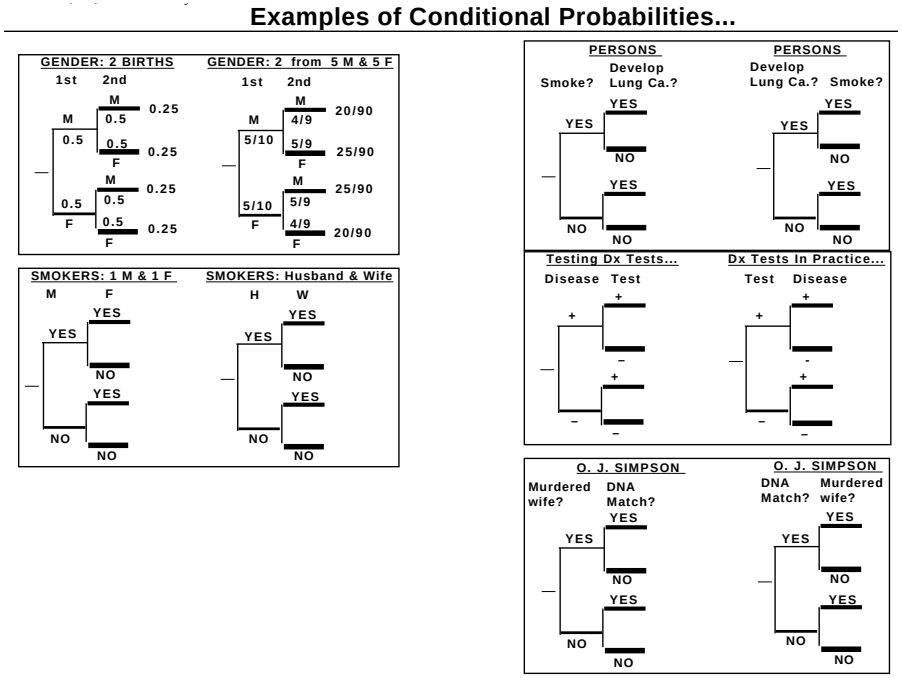

Figure 2: JH examples of independence/dependence, and 'forward'/'reverse' probabilities

2.2 Changing the conditioning: Bayes' rule

The right hand half of JH Figure 2 shows 3 examples of 'forward' (on left) and 'reverse' probabilities.

These same distinctions between 'forward' and 'reverse' probabilities is at the heart of the frequentist p-values (probabilities) versus Bayesian posterior probabilities. To state it simply,

$$Probability[data|Hypothesis] \neq Probability[Hypothesis|data]$$

or, if you prefer something that rhymes,

$$Probability[data|theta] \neq Probability[theta|data].$$

Two striking – and frightening – examples of misunderstandings about them are given on the next page.

U.S. National Academy of Sciences under fire over plans for new study of DNA statistics: Confusion leads to retrial in UK.[2]

[...] He also argued that one of the prosecution's expert witnesses, as well as the judge, had confused two different sorts of probability.

One is the probability that DNA from an individual selected at random from the population would match that of the semen taken from the rape victim, a calculation generally based solely on the frequency of different alleles in the population. The other is the separate probability that *a match between a suspect's DNA and that taken from the scene of a crime could have arisen simply by chance* – **in other words that the suspect is innocent despite the apparent match.**[3] This probability depends on the other factors that led to the suspect being identified as such in the first place.

During the trial, a forensic scientist gave the first probability in reply to a question about the second. Mansfield convinced the appeals court that the error was repeated by the judge in his summing up, and that this slip – widely recognized as a danger in any trial requiring the explanation of statistical arguments to a lay jury – justified a retrial. In their judgement, the three appeal judges, headed by the Lord Chief Justice, Lord Farquharson, explicitly stated that their decision "should not be taken to indicate that DNA profiling is an unsafe source of evidence."

Nevertheless, with DNA techniques being increasingly used in court cases, some forensic scientists are worried that flaws in the presentation of their statistical significance could, as in the Deen case, undermine what might otherwise be a convincing demonstration of a suspect's guilt.

Some now argue, for example, that quantified statistical probabilities should be replaced, wherever possible, by a more descriptive presentation of the conclusions of their analysis. "The whole issue of statistics and DNA profiling has got rather out of hand," says one. Others, however, say that the Deen case has been important in revealing the dangers inherent in the '**prosecutor's fallacy**'. They argue that this suggests the need for more sophisticated calculation and careful presentation of statistical probabilities. "The way that the prosecution's case has been presented in trials involving DNA-based identification has often been very unsatisfactory," says David Balding, lecturer in probability and statistics at Queen Mary and Westfield College in London. "Warnings about the prosecutor's fallacy should be made much more explicit. After this decision, people are going to have to be more careful."

[2] NATURE p 101-102 Jan 13, 1994.

[3] italics by JH. The wording of the italicized phrase is imprecise; the text in bold wording is much better .. if you read "despite" as "given that" or "conditional on the fact of"t

"The prosecutor's fallacy": Who's the DNA fingerprinting pointing at? [4]

Pringle describes the successful appeal of a rape case where the primary evidence was DNA fingerprinting. In this case the statistician Peter Donnelly opened a new area of debate. He remarked that

> forensic evidence answers the question "What is the probability that the defendant's DNA profile matches that of the crime sample, assuming that the defendant is innocent?"
>
> while the jury must try to answer the question "What is the probability that the defendant is innocent, assuming that the DNA profiles of the defendant and the crime sample match?" [5]

Apparently, Donnelly suggested to the Lord Chief Justice and his fellow judges that they imagine themselves playing a game of poker with the Archbishop of Canterbury. If the Archbishop were to deal himself a royal flush on the first hand, one might suspect him of cheating. Assuming that he is an honest card player (and shuffled eleven times) the chance of this happening is about 1 in 70,000.

But if the judges were asked whether the Archbishop were honest, given that he had just dealt a royal flush, they would be likely to place the chance a bit higher than 1 in 70,000 *.

The error in mixing up these two probabilities is called the "the prosecutor's fallacy," and it is suggested that newspapers regularly make this error.

Apparently, Donnelly's testimony convinced the three judges that the case before them involved an example of this and they ordered a retrial

[* Comment by JH: This is a very nice example of the advantages of Bayesian over Frequentist inference .. it lets one take one's prior knowledge (the fact that he is the Archbishop) into account.

The book 'Statistical Inference" by Michael W. Oakes is an excellent introduction to this topic and the limitations of frequentist inference.]

[4] New Scientist item by David Pringle, 1994.01.29, 51-52; cited in Vol 3.02 Chance News

[5] (JH) Donnelly's words make the contrast of the two types of probability much "crisper." The fuzziness of the wording on the previous story is sadly typical of the way statistical concepts often become muddied as they are passed on.

2.3 Examples

2.3.1 Example from genetics

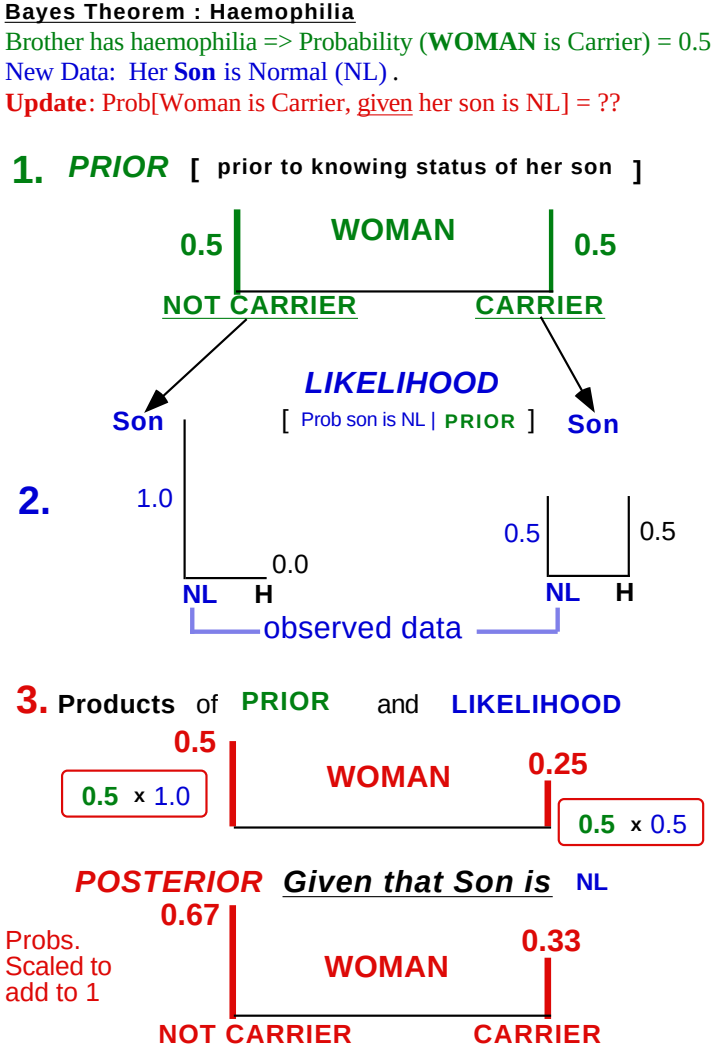

Figure 3: a simpler (but now outdated) example – nowadays there are direct tests for being a carrier: so one doesn't have to wait to have a son to alter the probabilities

2.3.2 Twins: Excerpt from an article by Bradley Efron

Modern science and the Bayesian-Frequentist controversy

> Heres a real-life example I used to illustrate Bayesian virtues to the physicists. A physicist friend of mine and her husband found out, thanks to the miracle of sonograms, that they were going to have twin boys. One day at breakfast in the student union she suddenly asked me what was the probability that the twins would be identical rather than fraternal. This seemed like a tough question, especially at breakfast. Stalling for time, I asked if the doctor had given her any more information. "Yes", she said, "he told me that the proportion of identical twins was one third". This is the population proportion of course, and my friend wanted to know the probability that her twins would be identical.
>
> Bayes would have lived in vain if I didn't answer my friend using Bayes' rule. According to the doctor the prior odds ratio of identical to nonidentical is one-third to two-thirds, or one half. Because identical twins are always the same sex but fraternal twins are random, the likelihood ratio for seeing "both boys" in the sonogram is a factor of two in favor of Identical. Bayes' rule says to multiply the prior odds by the likelihood ratio to get the current odds: in this case 1/2 times 2 equals 1; in other words, equal odds on identical or nonidentical given the sonogram results. So I told my friend that her odds were 50-50 (wishing the answer had come out something else, like 63-37, to make me seem more clever.) Incidentally, the twins are a couple of years old now, and "couldnt be more non-identical" according to their mom.

Supplementary Exercise 2.1. Depict Efron's calculations using a probability tree.

Supplementary Exercise 2.2 Use a probability tree to determine the best strategy in the Monty Hall problem

```
( http://en.wikipedia.org/wiki/Monty_Hall_problem )
```

Supplementary Exercise 2.3 A man has exactly two children: you meet the *older* one and see that it's a boy. A woman has exactly two children; you meet *one* of them [don't know if its the younger/older] and see is a boy. What is the probability of the man's younger child being a boy, and what is the probability of the woman's "other" child being a boy?

3 Likelihood

"*We need a way of choosing a value of the parameter(s) of the model*" (1st paragraph): It is clear from the later text that they do not mean to give the impression that one is only interest in a single value or point-estimate. For any method to be worthwhile, it needs to be able to provides some measure of uncertainty, i.e. an interval or range of parameter values.

"In simple statistical analyses, these stages of model building and estimation may seem to be absent, the analysis just being an intuitively sensible way of summarizing the data." Part of the reason is that (as an example) a sample mean may simply seem like a natural quantity to calculate, and it does not seem to require an explicit statistical model. The mean can also be seen as the least squares estimate, in the sense that the sum of the squared deviations of the sample values from any other value than the sample mean would be larger than the sum of the squared deviations about the mean itself, i.e., the sample mean is a least squares estimate. But that purely arithmetic procedure still does not require any assumptions about the true value of the parameter value μ, or about the shape of the distribution of the possible values on both sides of μ. For the grade 6 exercise about the mean number of errors per page, it seemed to make sense to divide the total number of errors by the total number of pages; but what if the task was to estimate the mean weight of the pages? We discussed in class at least two different statistical models – that would lead to different estimates.

"In modern statistics the concept which is central to the process of parameter estimation is likelihood." Older and less sophisticated methods include the method of moments, and the method of minimum chi-square for count data. These estimators are not always efficient, and their sampling distributions are often mathematically intractable. For some types of data, the method of weighted least squares is a reasonable approach, and we will also see that iteratively-reweighed least squares is a way to obtain ML estimates without formally calculating likelihoods.

Likelihood is central not just to obtain frequentist-type estimators per se, but also to allow Bayesian analyses to combine prior beliefs about parameter values to be updated with the data at hand, and arrive at what one's post-data beliefs should be.

Likelihood provides a very flexible approach to combining data, provided one has a probability model for them. As a simple example, consider the challenge of estimating the mean μ from several independent observations for a $N(\mu, \sigma)$ process, but where each observation is recorded to a different degree of numerical 'rounding' or 'binning.' For example, imagine that because of

the differences with which the data were recorded, the $n = 4$ observations are $y_1 \in [4, 6),\ y_2 \in [3, 4),\ y_3 \in [5, \infty),\ y_4 \in [-\infty, 3.6)$. Even if we were told the true value of σ, the least squares method cannot handle this uni-parameter estimation task.

"*The main idea is simply that parameter values which make the data more probable are better supported than values which make the data less probable.*" Before going on to *their first example*, with a parameter than in principle could take any values in the unit interval, consider a *simpler example* where there are just two vlaues of π. We have sample of candies from one of two sources: American, where the expected distribution of colours is 30%:70% and the other Canadian where it is 50%:50%. In our sample of $n = 5$, the observed distribution is 2:3. Do the data provide more support for the one source than the other?

3.1 Likelihood in the binary model

Notice the level of detail at which the observed data are reported in Figure 3.1: not just the numbers of each (4 and 6) but the actual *sequence* in which they were observed. The Likelihood function uses the probability of the observed data. Even if we did not know the sequence, the probability of observing 4 and 6 would be ${}^{10}C_4 = 210$ times larger; however since we assume there is no order effect, i.e., that π is constant over trials, the actual sequence does not contain any information about π, and we would not include this multiplier in the Likelihood. In any case, we think of the likelihood as a function of π rather than of the observed numbers of each of the two types.: these data are considered fixed, and π is varied.. contrast this with the tail area in a frequentist p-values, which includes other non-observed values more extreme than that observed. Likelihood and Bayesian methods do not do this.

"$\pi = 0.5$ *is more likely than* $\pi = 0.1$" Please realize that this statement by itself could be taken to mean that we should put more money on the 0.5 than the 0.1. It does not mean this. in the candy source example, knowing where the candies were purchased, or what they tasked like, would be additional information that might in and of itself make one source more likely than the other. The point here is not to use terms that imply a prior or posterior probability distribution on π. The likelihood function is based just on the data, and in real life any extra prior information about π would be combined with the information provided by the data. It would have been better if the authors had simply said "*the data provide more support for* "$\pi = 0.5$ *than* $\pi = 0.1$." Indeed, I don't think "$\pi = 0.5$ *is more likely than* $\pi = 0.1$" is standard terminology. The terminology "0.4 is the ML estimate of π" is simpler and less ambiguous.

History: there is some dispute as to who first used the principle of ML for the choice of parameter value. The name of Gauss is often mentioned. The *seldom mentioned* 1912 paper by Fisher, while still a student, is a nice clean example, and shows how Likelihood (he did not use the word likelihood in the paper) is flexible and allows for the different bins sizes with which observations might be recorded, etc. It is worth reading that original paper, but don't spend too much time on section 5, where he deals with the ML estimation of the parameters μ and σ of a Normal distribution: the ML estimate of σ^2 involves a divisor of n rather than $n-1$, and embarrassment for Fisher, who was from early on, insisted on the correct degrees of freedom when assessing variation. His 1912 paper can be found in the digital archives in Adelaide, Australia (he spent his last years there) but JH has put a copy in the Resources folder.

The *usual* reference is to papers by Fisher in the early 1920's, where he worked of many of the properties of ML estimators.

One interesting feature of the 1912 paper is that Fisher never defined the likelihood as a *product* of probabilities; instead he defined the log-likelihood as a *sum* of log-probabilities. This is very much in keeping with his summation of *information* over observations. Indeed, there is a lot in his writings about choosing the most *informative* configurations at which to observe the experimental or study units.

3.2 Supported range

The choice of critical value is much less standardized or conventional than say the one for a significance test, or confidence level, or a highest posterior density.

Fig 3.4 (based on 20/50) vs. Fig 3.3 (based on 4/10): the authors don't say it explicitly, but the sharpness of the likelihood function is measured formally by the second derivative at the point where it is a maximum.

3.3 The log likelihood

The (log-)likelihood is invariant to alternative monotonic transformations of the parameter, so one often chooses a parameter scale on which the function is more symmetric.

3.4 Censoring in follow-up studies

See applications below. These will be more relevant after we consider all of the fitting options, and the benefits/felxibility of a Likelihood approach.

3.5 Other fitting methods

We mentioned earlier that the method of least squares does not make an explicit assumption about the distribution of the deviations from or even that the observed data are a sample from a larger universe. Another older method, that does not make explicit assumptions about the variations about the postulated means, is the method of minimum chi-square. It was used for fitting simpler models for dose response data involving count data. This minimum chi-square criterion does not lead to simple methods of estimation, or to estimators with easily derived sampling distributions. Nevertheless, it is one of the thee methods (the others are ML – which requires a fully specified model for the variations, and LS, that does not) used in the java applet `http://www.biostat.mcgill.ca/hanley/MaxLik3D.swf`. The applet allows you to fit a linear model to the above-described 2-point data, and to monitor how the log-likelihood, the sum of squared deviations, and the chi-square goodness of fit statistics vary as a function of the entertained values of β.

The applet shows that the LS method which measures lack of fit on the same scale that the y's are measured on (cf the two red lines). The min-X^2 method – applied to y's that represent counts or frequencies, is similar, in that the "loss function" is $\sum(y-\hat{y})/\hat{y}^2$. The criterion for the ML fitting of a Poisson model is very different, in that it is measured on the probability or log-probability scale, a scale that is shown in blue, and projecting out from the $x-y$ plane.

Under some Normal models with homoscedastic variation, the LS and ML methods give the same estimates for the parameter(s) that make up the mean. If $y|x \sim Normal(\mu_x, \sigma^2)$, then Lik $= \prod(1/\sigma)\exp[-\{(y_i-\beta x_i)^2/2\sigma^2\}]$. This is maximized when the exponentiated quantity is minimized. The minimization is the same one involved in the LS estimation.

Supplementary Exercise 3.1. Grouped Normal data (from Fisher's paper[6]).Three hundred observed measurement errors (ϵ's) from a $N(0,\sigma)$ distribution are grouped (binned) in nine classes, positive and negative values being thrown together as shown in the following table:-

Bin	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	All
Frequency (f)	114	84	53	24	14	6	3	1	1	300

Estimate σ^2 ...

1. as $(1/300)\sum f \times \epsilon_{mid}^2$. Note that we estimate it using a divisor of n rather than $n-1$, since we do not have to estimate μ : the errors are deviations from *known* values, so $\mu = 0$ (structurally).
2. Using Sheppard's correction for the grouping, i.e, by subtracting $w^2/12$, where w is the width of each bin, in this case 1. Incidentally, can you figure out why Sheppard subtracts this amount? Shouldn't grouping *add* rather than subtract noise?
3. Using the method of Minimum χ^2.
4. Using the method of Maximum Likelihood.

Supplementary Exercise 3.2 [2012 only]. Frequency data, the subject of Galton's 1894 correspondence with the Homing News and Pigeon Fanciers' Journal.[7]

Significance magazine (`http://www.significancemagazine.org/`) has special Galton coverage in 2011, the 100th anniversary of his death – Galton was born in 1822, the same year, he noted himself, as the geneticist Gregor Mendel. In the article "Sir Francis Galton and the homing pigeon", Fanshawe writes...

> "The results for the 3,207 "old birds" are shown in the table. The table shows the proportion of birds in each category. Galton suggests summarising the figures by their mean and "variability", which he estimates as 976 and 124 yards per minute respectively. It is not clear which quantity Galton calls the "variability" – his figure appears too small to be a standard deviation.
>
> The second row of figures are Galton's, and arise from the proportions that would be expected by approximating the original data by a Normal distribution. The fit appears extremely good."

Using these frequencies and bin-boundaries[8] from the journal article, and the Normal distribution assumed by the journal and by Galton,

Bin	-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14+	All
Freq	22	43	164	284	598	645	683	396	132	120	120	3207

[6] On the Mathematical Foundations of Theoretical Statistics, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 222 (1922), pp. 309-368

[7] Material (3p of journal, Fanshawe's article, and R code) avialable under Resources.

[8] 5-6 means 500-600 yards per minute, etc.

estimate μ and σ, and, where possible, using SE($\hat{\mu}$) and SE($\hat{\sigma}$), [9] form symmetric (frequentist) confidence intervals for μ and σ,

1. by concentrating the frequencies at the midpoints, and at suitably chosen values for the two open-ended categories
2. via the method of Minimum χ^2, and
3. via the method of Maximum Likelihood. Then
4. determine whether Fanshawe is correct: i.e., is the "124 yards" measure of "variability" indeed too small to be a standard deviation (SD)?
5. Galton rarely used the SD.[10] Instead he – as Gosset often did – used the Probable Error (PE), i.e., 1/2 the IQR.[11]

   In a Gaussian distribution, how much smaller/larger is the PE than the SD?

   Does this factor explain how Galton arrived at the 124 yards per minute?

The sample size is so large here that the symmetric (z-based) CI for σ is quite accurate. By what if the sample size were quite small? In this case you could use the tails of the (non-symmetric) distribution of the distribution of s^2 to derive an asymmetric first-principles frequentist confidence interval for σ^2, and by transformation, for σ.[12]

Suppose that for each dart thrown, one calculates the squared distance from the center, ie $d_i^2 = e_{1,i}^2 + e_{2,i}^2$. Show that $(1/n)\sum_i d_i^2$ is an unbiased estimator of $2\sigma^2$. What sampling statistical distribution does each d_i^2 follow? What is a common name for the distribution of the square root of this random variable?

[9] Since $s^2 \sim (1/\nu) \times \sigma^2 \times ChiSq(d.f. = \nu)$, then $\text{Var}[s^2] = (1/\nu^2) \times \sigma^4 \times 2\nu$. By Delta method,

$$\text{Var}[s] \approx \text{Var}[s^2] \times \left\{\frac{ds}{ds^2}\right\}^2 = \underbrace{(1/\nu^2) \times \sigma^4 \times 2\nu} \times \underbrace{(1/4) \times \{1/\sigma^2\}^{-1}} = (1/\nu^2) \times \sigma^2,$$

so $\text{SE}[s] \approx (1/\nu^2)^{-1/2} \times \sigma$.

[10] Karl Pearson was the one who promoted the SD.

[11] Thus, it is equally probable (50:50) for an observation to be more/less than this amount from the middle (truth).

[12] Hint: (taking some semantic liberties) a first-principles 100(1-α)% frequentist CI, (L, U) for θ is the pair of *statistics* (L, U), such that $\text{Prob}(\hat{\theta} \geq \hat{\theta}_{\text{observed}} \mid \theta = L) = \alpha/2$ and $\text{Prob}(\hat{\theta} \leq \hat{\theta}_{\text{observed}} \mid \theta = U) = \alpha/2$.

3.6 Other Applications: exercises

3.6.1 2 datapoints and a model

One has 2 independent observations from the (no-intercept) model

$$E[y|x] = \mu_{y|x} = \beta \times x.$$

The y's might represent the total numbers of typographical errors on x randomly sampled pages of a large document, and the data might be $y = 2$ errors in total in a sample of $x = 1$ page, and $y = 8$ errors in total in a separate sample of $x = 2$ pages. The β in the model represents the mean number of errors per page of the document. Or the y's might represent the total weight of x randomly sample pages of a document, and the data might be $y = 2$ units of weight in total for a sample of $x = 1$ page, and $y = 8$ units for a separate sample of $x = 2$ pages. The β in the model represents the mean weight per page of the document.

We gave this 'estimation of β' problem $\{ (x, y) = (1, 2) \ \& \ (2, 8)\}$ to several statisticians and epidemiologists, and to several grade 6 students, and they gave us a variety of estimates, such as $\hat{\beta} = 3.6$/page, 3.33/page, and 3.45!

Supplementary Exercise 3.3

How can this be? The differences have to do with (i) what model they (implicitly or explicitly) used for the variation of each $y \mid x$ around the mean $\mu_{y|x}$ and (ii) the method of fitting.

1. From 1st principles derive both the LS and (if possible the) ML estimators of β when
   (a) $y \mid x \sim$???$(\mu_{y|x})$
   (b) $y \mid x \sim Poisson(\mu_{y|x})$
   (c) $y \mid x \sim N(\mu_{y|x}, \sigma)$ [assume σ is known]
   (d) $y \mid x \sim N(\mu_{y|x}, \sigma^2 = x \times \sigma_0^2)$ [assume σ_0^2 is known]
2. Where possible, match the estimators with the various numerical estimates above.
3. One of the numerical estimates came from another fitting method, namely the (now seldom-used) method of Minimum Chi-square, which seeks the value of β that minimizes $\sum \frac{(O-E)^2}{E} = \sum \frac{(y-\beta x)^2}{\beta x}$ in this example. Verify that the one remaining estimate of unknown origin is in fact obtained using this estimator.

See the (Flash) applet on `http://www.biostat.mcgill.ca/hanley/software/`

One of the messages of this exercise is that for one to use a likelihood approach, one must have a fully-specified probability model so that one can write the probability of each observed observation.

And, with different distributions of the y's around the mean $\mu_{y|x} = E(y|x) = \beta \times x$, the probabilities (and thus the overall likelihood, and its maximum, would be different.

3.6.2 Application: Estimation of parameters of gamma distribution fitted to tumbler mortality data [interval-censored and right-censored data].

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 and for Statistical Models. 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. Since inspections were only once a week, the lengths of service of the items that *did* fail are also censored, but *within* [in most instances] a 1-week interval. This type of censoring is called '*interval-censoring*'.

Supplementary Exercise 3.4

Using the data in Table 1 for the article [contained in the various versions of the R code in the same link] , determine the MLEs of the two parameters of the gamma distribution, and compare them with those obtained by the original authors [they use a slightly approx. ML method]. Do so in two ways (they should give the same likelihood function, and thus the same MLEs):

1. using an *unconditional* approach, based on 549 contributions – one per tumbler, with each tumbler considered in isolation from the other 548 – so that each failure (unconditional) contributes one term and each (ULTIMATELY) censored observation (also unconditional) contributes another. [of course, there are 'multiplicities'; thus, instead of a sum of 549 log-likelihhods, you can use the multiplicities (and multiplication of a 1-item log-likelihood by the multiplicity) to reduce the computation].

2. using the binomial structure created by the authors: a row that has n exposed tumblers *that* week (and that only considers whether the tumbler that began that week survived *that* week) makes n Bernoulli-based log-likelihoods, (or 1 Binomial-based log-likelihood) for that *week*.

This exercise shows that there is more than 1 way to set up the likelihood.

3.6.3 Application: Estimation of parameters of a parametric distribution fitted to avalanche mortality data [all observations are censored – either left-censored or right-censored. Such data are often referred to as "current-status" data].

One example of status-quo data is data from a cross-sectional survey of menarche status in girls, or the prevalence of decayed-missing-or-filled (DMF) teeth (or say permanent dentition) in dental public health, or HIV prevalence in the general population or in specific sub-populations, such as partners of persons who contracted HIV though blood donations.

Another is the data from the Avalanche Survival Chances by Falk et al. in the journal *Nature* in 1994. The article and the data are available under Resources.

The authors fitted a non-parametric model. We will discuss in class which parametric models (or mixtures of different parametric models) might make sense. But, just to get some practice with this type of data, we will start with a very simply one, even if we know a priori it is too simplistic.

Supplementary Exercise 3.5

Using the raw data, and (for now) the simplistic parametric model we agreed on in class, determine the MLEs of the two parameters of this gamma distribution, and compare the fit with the fit of the smooth and non-parametric curves shown in the authors' article.

3.6.4 Application: Distribution of Observations in a Dilution Series.

(Again, Text from Fisher's 1922 paper). An important type of discontinuous distribution occurs in the application of the dilution method to the estimation of the number of micro-organisms in a sample of water or of soil. The method here presented was originally developed in connection with Mr. Cutler's extensive counts of soil protozoa carried out in the protozoological laboratory at Rothamsted, and although the method is of very wide application, this particular investigation affords an admirable example of the statistical principles involved.

In principle the method consists in making a series of dilutions of the soil sample, and determining the presence or absence of each type of protozoa in a cubic centimetre of the dilution, after incubation in a nutrient medium. The series in use proceeds by powers of 2, so that the frequency of protozoa in each dilution is one-half that in the last. The frequency at any stage of the process may then be represented by

$$m = \frac{n}{2^x},$$

when x indicates the number of dilutions. Under conditions of random sampling, the chance of any plate receiving 0, 1, 2, 3 protozoa of a given species is given by the Poisson series

$$e^{-m}\left(1, m, \frac{m^2}{2!}, \frac{m^3}{3!}, \dots\right),$$

and in consequence the proportion of sterile plates is

$$p = e^{-m},$$

and of fertile plates

$$q = 1 - e^{-m}.$$

In general we may consider a dilution series with dilution factor a so that

$$\log p = -\frac{n}{a^x},$$

and assume that s plates are poured from each dilution. The object of the method being to estimate the number n from a record of the sterile and fertile plates, we have

$$L = S_1(\log p) + S_2(\log q),$$

when S_1 stands for summation over the sterile plates, and S_2 for summation over those which are fertile.

Supplementary Exercise 3.6 Estimate n from the following dilution series data:

Dilution:	0.25	0.5	1	2	4	8	16	32	64	128
Number of plates:	5	5	5	5	5	5	5	5	5	5
Number of fertile plates:	5	5	5	5	4	3	2	2	0	0

3.6.5 Application: Pooled testing:- old and new uses

The following excerpts are from a 1976 article "Group testing with a new goal, estimation", in Biometrika, 62, 1, p. 181 by authors Sobel and Elashoff. They begin by referring to the Dorfman, whose article, in the Annals of Mathematical Statistics, 1943, first used the ideas of group testing, with a binomial model, to reduce the number of medical tests necessary to find all members of a group of size N that have the syphilis antigen. They continued...

> Another aspect of the group-testing problem arises when one is interested not in the *classification of all the individuals* but in the *estimation of the frequency* of a disease, or of some property, when group-testing methods can be used. Given a random sample of size N, say, from a binomial population, the best estimate of the prevalence rate p, in the sense of minimizing the mean square error, will be obtained by testing each unit separately. However, if N is large and the tests are costly, then a different criterion, that includes testing costs, may indicate that group-testing designs should be used. We might expect benefits from group testing to increase as p decreases.
>
> [....] Example: Rodents are collected from the harbour of a large city, and, after being killed, dissected, etc., their liver is to be carefully examined under a microscope for the presence or absence of a specific type of bacterium. The goal of the study is to estimate the proportion p of rodents that carry this bacterium using an economical experimental design. In this application the cost of obtaining the animals is negligible compared to the cost of testing, i.e. the microscopic search. It was proposed that an economical design to estimate p should be possible by combining in a single sample a small portion of the liver from each of several test animals and then carrying out a microscopic search on a homogeneous mixture of these liver portions. The problem is to find the best number, say A, of liver portions to combine and how to estimate the prevalence rate p from such a design. In addition, if this bacterial type is present in some particular tests, then the pathologists want to know whether they should carry out another test on a subset of these same animals or go on to test a new group of A animals.
>
> [...] Thompson (1962) estimated the proportion of insect vectors capable of transmitting asteryellows virus in a natural population of the six-spotted leafhopper, an aphid. Instead of putting one insect with a previously unexposed aster test plant, he puts several insects with one test plant, for eoonomic reasons, and waits to see if the

plant develops the symptoms of this virus. If it does, then at least one of these insects carried the virus; otherwise it is assumed that none carried it. The statistical problem is to choose an optimal number A of insects to be put with one test plant.

Contemporary uses: (can also Google *Minipool testing*)

The following text is an excerpt form Canadian Blood Services : Customer Letter #2005-18, 2005-05-17, entitled "Planned Measures to Protect the Blood Supply from West Nile Virus (WNV) - 2005 Season."

Dear Colleague:

West Nile season is approaching once again and this letter is to inform you about enhanced measures Canadian Blood Services has put in place to further protect the safety of the blood supply during the 2005 season.

For the summer of 2005, Canadian Blood Services will again use single-unit testing (SUT) to enhance the sensitivity of the West Nile Virus nucleic acid test. Minipool testing (6 samples/pool) is used throughout the year.

- In the summer of 2005, a 'trigger' will be used to initiate SUT. SUT will be initiated in a health region when a presumptive positive blood donor is detected using minipool testing, OR the prevalence of recent confirmed human cases in the preceding two weeks exceeds 1/1,000 population in rural areas, or 1/2,500 in urban areas.
- SUT will cease in a health region when there have been no positive donors for two weeks or the occurrence of WNV cases in the population falls below the aforementioned population triggers.

Supplementary Exercise 3.7 Suppose that in order to estimate the prevalence (π)of a characteristic in a population, one tests N randomly sampled objects by pooling them into n_b batches of size k (so that $N = n_b \times k$) and determining, for each batch, i.e. collectively, if at least one of its members is positive. Suppose that n_{b+} batches are found to be positive. Develop estimators of π using the method of moments, and using minimum χ^2 and Maximum Likelihood criteria.

3.6.6 Application: Measuring one's accuracy at darts

In 2011, Tibshirani (junior!) et al.[13] published a very instructive essay. In addition to its innovative use of a personalized heatmap to show the optimal strategy for throwing darts, it provides an engaging example for teaching several statistical concepts and techniques, such as fast Fourier transforms, the EM algorithm, Monte Carlo integration, importance sampling, and the Metropolis Hastings algorithm. It is a delightful blend of the applied and the theoretical, the algebraic and the graphical.

It also continues the tradition of statisticians' fascination with the imagery of marksmen (Turner, 2010). In her chapter on metaphor and reality of target practice, Klein (1997) writes of 'men reasoning on the likes of target practice' and describes how this imagery has pervaded the thinking and work of natural philosophers and statisticians. Klein shows a frequency curve, by Yule, for 1,000 shots from an artillery gun in American target practice. Pearson used it in his 1894 lectures on evolution; he decomposed the frequency curve into two chance distributions centered slightly to the right and left of the target, gave reasons why this might occur, and used it to illustrate the interplay between random variation and natural selection. He also used it in his 1900 paper in one of the illustrations of his test of goodness of fit. Incidentally, Klein also reminds us of the origin of the term '*stochastic.*' In Liddell and Scott (1920) we find the following entries:

$\sigma\tau o\chi o\varsigma$	an aim, shot. a guess, conjecture.
$\sigma\tau o\chi a\sigma\mu\alpha$	a missile aimed at a mark; an arrow, javelin.
$\sigma\tau o\chi a\sigma\tau\iota\kappa o\varsigma$	able to hit: able to guess, shrewd, sagacious.

Since the optimal aiming spot in darts – and thus the heatmap provided by the online applet – depends strongly on one's accuracy, much of the Tibshirani et al. article is devoted to the challenge of estimating the (co)variance parameter(s) that describes this accuracy. All of the estimators rely on the data generated by throwing n darts, aiming each time at the centre of the board, i.e., the double-bulls-eye, and recording the result for each throw.

[13] Tibshirani, R.J., Price, A, and Taylor, J. A statistician plays darts. J. R. Statist. Soc. A (2011) 174, Part 1, 213-226. [See also the follow-up letter from S. Sadhukhan, Z Liu, and J Hanley, along with the references • Klein, J.L. (1997). Statistical Visions in Time: A History of Time Series Analysis 1662- 1938. pp. 3-11. Cambridge. Cambridge University Press. • Liddell, H.G. and Scott R. (1920). A Lexicon, abridged from Liddell and Scott's Greek-English Lexicon. p. 653. London. Oxford at the Clarendon Press. • Tibshirani, R.J., Price, A, and Taylor, J. A statistician plays darts. J. R. Statist. Soc. A (2011) 174, Part 1, 213-226. • Turner, E.L. and Hanley, J.A. (2010) Cultural imagery and statistical models of the force of mortality: Addison, Gompertz and Pearson. J. R. Statist. Soc. A, 173, Part 3, 483-499.

The authors noted that they would lose considerable information by not measuring the actual *locations* where the darts land but considered this to be too time-consuming and error-prone. Instead, they chose the individual *scores* produced by the throws (the 44 possible scores are 0:22, 24:28, 30, 32:34, 36, 38:40, 42, 45, 48, 50, 51, 54, 57, 60). Based on $n = 100$ throws by authors 1 and 2, assuming the simplest variance model (equal, uncorrelated vertical and horizontal Gaussian errors), their standard deviations were estimated to be $\hat{\sigma} = 64.6$ and 26.9 respectively (the applet gives $\hat{\sigma}$ to 2 decimal places)

Our follow-up letter provides a measure of the statistical precision of these accuracy estimates (for example, we calculate that the 95% limits to accompany the reported point estimate 64.6 derived from 100 scores are approximately 56 and 75). More importantly, we show that more precise estimates of σ can often be achieved with the same number of throws (or the same precision with fewer throws) if one uses a simpler yet more informative version of the result from each throw.

Here, as in the letter, we focus on the simplest variance model, where horizontal and vertical errors, e_x and e_y, are Gaussian, centered on (0,0), independent of each other and of the same amplitude, i.e., $\sigma_{e_x} = \sigma_{e_y} = \sigma; \rho_{e_x,e_y} = 0$.

We first consider the most mathematically tractable, but least practical, method of estimating σ, namely to measure the exact (x, y) locations where the n darts land. We then consider the almost as mathematically tractable, but much more practical – and almost as statistically efficient – method of estimating σ, namely to merely record in which 'ring' each dart lands. We leave to later the the authors' more complex – but sometimes less efficient – method based on actual 0-60 scoring system used in darts games.

Denote by $e_{c,i}$ the error in the c-th co-ordinate (1='x', 2='y') of the i-th dart.

Supplementary Exercise 3.8

1. Show that $(1/2n)\sum_c\{\sum_i e_{c,i}^2\}$ is an unbiased estimator of σ^2 and that it is the method-of-moments, the LS, and the ML estimator.

   What sampling statistical distribution does this estimator follow?

   Use the two separate $\alpha/2$ tails of this (slightly non-symmetric) distribution to derive an asymmetric first-principles frequentist confidence interval for σ^2.[14]

   Suppose that for each dart thrown, one calculates the squared distance from the center, ie $d_i^2 = e_{1,i}^2 + e_{2,i}^2$. Show that $(1/n)\sum_i d_i^2$ is an unbiased estimator of $2\sigma^2$. What sampling statistical distribution does each d_i^2 follow? What is a common name for the distribution of the square root of this random variable?

2. Suppose we simply divide the dartboard into 7 '*rings*' [15] and record which one the dart lands in: 1. the double-bulls-eye; 2. the single-bulls-eye; the ones formed by the: 3. single-bulls-eye and inner triple; 4. inner and outer triple; 5. outer triple and inner double; and 6. inner and outer double, wires respectively; and 7. beyond the outer double wire (i.e., the throw misses the board). In other words, we divide the dartboard into just 7 regions. Suppose that the distribution of the results of $n = 100$ throws is as follows:

ring:	1	2	3	4	5	6	7	all
frequency:	0	6	77	5	12	0	0	100

   Calculate (and plot) the logLik(σ^2) function and find the MLE of σ^2.

3.7 Bayesian approach to parameter estimation

Given that the Bayesian approach is a very important and conceptually different way of making inference about the parameters of a model, and even though they mentioned Bayes rule in Chapter 2, it is surprising that Clayton and Hills do not make a statement about the Bayesian approach until Chapter 10; and even then, they do not give it much space. Maybe it's because they wanted the reader to become quite comfortable with Likelihood (which provides the Bridge between the prior and posterior distributions) before doing so.

[14] Hint: (taking some semantic liberties) a first-principles 100(1-α)% frequentist CI, (L, U) for θ is the pair of *statistics* (L, U), such that $\text{Prob}(\hat{\theta} \geq \hat{\theta}_{\text{observed}} \mid \theta = L) = \alpha/2$ and $\text{Prob}(\hat{\theta} \leq \hat{\theta}_{\text{observed}} \mid \theta = U) = \alpha/2$.

[15] In fact, the innermost region is a circle, the next 5 are rings, and the outermost one is all of the remaining area.

.

4 Consecutive follow-up intervals

4.1 A sequence of binary models

The lifetable as a sequence of Bernoulli models: Efron (1977) was one of the early authors to point out that the likelihood contribution of a subject, followed for t units of time, is equivalent to the likelihood for a sequence of a large number, $n = t/\Delta$, of Bernoulli trials, with time-dependent probabilities of failure. For a trial that corresponds to the small interval $(t, t+\Delta)$, the failure probability can be well approximated by $p = h(t)\Delta$, where $h(t)$ is called the hazard function (see later). The sequence ends with the n^{th} trial, at the time of the event of interest or when follow-up was otherwise terminated. In a subsequent article Efron (1988) focused on discretizations of the t-axis and on using logistic regression to fit various smooth-in-t hazard and survival functions in the one-sample situation, where the usual non-parametric alternative is the Kaplan-Meier estimator of survival rate.

The probabilities of surviving one, two, and three years without failing are called the *cumulative survival probabilities* for the cohort: JH continues to argue that the word cumulative is misleading. The complement of the (unconditional) survival probability is the *cumulative incidence*. It is an increasing function. Would we call a declining fraction, obtained as a product of more and more fractions, a *cumulative* fraction?

4.2 Estimating the conditional probabilities of failure

The subjects who contribute to the estimation of the conditional probabilities do not have to have been followed from the beginning. One can splice together estimates based on separate samples. This is what is done to create current lifetables. And in any case, when (a subset of) those who "survive" a specific time band are used again in the next band, the estimates are treated as independent of each other – just as if they were from different persons. In current lifetables, they are different persons!

Table 17.1 in p. 570 of the Survival Analysis chapter (17) of the 4th edition of Statistical Methods in Medical Research by Armitage, Berry & Matthews, illustrates the difference between 'current' (aka 'period') and 'cohort' lifetables.

The entire 'current' lifetable is calculated, as a product of conditional probabilities, using the *observed* age-specific mortality rates in England and Wales in 1930-1932. In this sense it is fictitious, since those who computed the table

in the 1930's didn't know for sure that the world would even exist in 2010, when those remaining from the fictional 1000 who started out at age 0 would reach their 80th birthday. And even if they did, they could not have anticipated exactly what force of mortality these 80-year olds would face in 2010, even though they might have foreseen that mortality rates would improve over time. The force of mortality these 80-year olds would face in 2010 is a good deal lower than the force of mortality the 80-year olds actually faces in 1930-32. For example, the death rate in the male 75-79 age category in Denmark was 9.4/100MY in 1930-34 and 4.2/100MY in 2000-04.

570 Survival analysis

Table 17.1 Current and cohort abridged life-tables for men in England and Wales born around 1931.

	Current life-tables 1930–32			Cohort life-table, 1931 cohort
Age (years) x	Probability of death between age x and $x+1$ q_x	Life-table survivors l_x	Expectation of life $\mathring{e}_x$	Life-table survivors l_x
0	0·0719	1000	58·7	1000
1	0·0153	928·1	62·2	927·8
5	0·0034	900·7	60·1	903·6
10	0·0015	890·2	55·8	894·8
20	0·0032	872·4	46·8	884·2
30	0·0034	844·2	38·2	874·1
40	0·0056	809·4	29·6	861·8
50	0·0113	747·9	21·6	829·7
60	0·0242	636·2	14·4	—
70	0·0604	433·6	8·6	—
80	0·1450	162·0	4·7	—

"The cohort life-table describes the actual survival experience of a group, a 'cohort' of individuals born at about the same time. Those born in 1900, for instance, are subject during their first year to the mortality under 1 year of age prevailing in 1900-1; if they survive to 10 years of age they are subject to the mortality at that age in 1910-11; and so on. Cohort life-tables summarize the mortality at different ages at the times when the cohort would have been at these ages. The right-hand side of Table 17.1 summarizes the l_x column from the cohort life-table for men in England and Wales born in the 5 years centred around 1931. As would be expected. the values of l_1 in the two life-tables are very similar, being dependent on infant mortality in about the same calendar years At higher ages the values of l are greater for the cohort table because this is based on mortality rates at the higher ages whIch were experienced since 1932."

4.3 A cohort life table

These [survival] plots are useful for studying whether the probability of failure is changing with follow-up time, and for calculating survival probabilities for different periods of time. In fact, it is not that easy to check if the probability of failure is changing from survival curves. The probability of failure the authors write of is a conditional, i.e. time-specific, probability, and so the hazard function, which uses as a denominator the numbers of persons at risk at that time, makes it easier to monitor this probability.

4.4 The use of exact times of failure and censoring

"[...] *choosing the bands so short that each failure occupies a band by itself.*" This is the same assumption that allows us to derive the Poisson distribution as a limiting case of the Binomial distribution, and the link between the Poisson distribution and the exponential distribution of inter-event times.

"*This method of estimating the cumulative survival probabilities is called the Kaplan-Meier method*" It is also called the product-limit method, since it is derived by slicing time into smaller and smaller bands, and not having to be materially concerned about where within the band an observation becomes censored. In the JUPITER trial example JH is using in the EPIB-634 course, the follow-up ranges from just over a year to almost 5 years, or approximately 400 to 1600 days. The 200+ events in the placebo arm, and the 100+ in the treatment arm, are distributed over these 1600 days. If we use one day as the width of each band, and estimate $S(1000)$, the 1000-day "event-free survival" then this estimate is a product of 1000 estimated conditional probabilities, many of them estimated at unity. *So the changes in the product take place only at the days in which there were events.* See also the COMPARE trial.

The persons at risk just before the event on a particular day (including the person(s) who did suffer the event that day) are called the *riskset.* They are the *candidates* for the event.

4.4.1 $\widehat{S(t)}_{KM}$ is a Maximum-Likelihood estimator of $S(t)$

As is rigorously justified in their 1958 paper, the Kaplan-Meier estimator is a non-parametric ML estimator within the class of all possible $S(t)$ functions.

Supplementary Exercise 4.1 Take a small survival dataset with just 3 observations, 1 censored and 2 not, such as the 3 values 5, 7+ and 10. Show that

$\widehat{S(t)}_{KM}$	Interval	Point (t)	Prob. Mass at Point
1	$t < 5$		
		$t = 5$	1/3
2/3	$5 \le t < 10$		
		$t = 10$	2/3
0	$t \ge 10$.		

maximizes the Likelihood, ie the probability of the observed data as a function of $S(t)$, i.e., that no other $\widehat{S(t)}$ can yiled a larger likelihood.

4.4.2 $\widehat{S(t)}_{KM}$ as a '<u>self-consistent</u>' and as a <u>Distribute</u> <u>mass</u> <u>to</u> <u>the right</u>' estimator of $S(t)$

The K-M estimator, based on n observations $T_1, \ldots, T_n$, some censored, some not, can also be seen as obeying the self-consistent estimating equation:

$$S(t) = \frac{1}{n}\left\{ \sum_{all} I[T_i > t] + \sum_{censored\ <\ t} \frac{S(t)}{S(T_i)} \right\}$$

Observations known to exceed t [even if censored after t]are counted as survivors (1's) while observations for which we don't know if they will exceed t are counted as fractions or probabilities: those which are already close to reaching t are given higher chances of eventually exceeding it, those which are further to the left of t are given lower chances of doing so: the chance of eventually exceeding t, given that one has already reached a value $T < t$, is $S(t)/S(T)$.

The K-M estimator can also be seen as a **distribute to the right** procedure: Initially, each of the n observations is given a mass of $1/n$. Then, the mass given to the leftmost censored observation is redistributed (equally) to all observations to the right of it, and that leftmost observation is removed. The process is repeated until all censored observations are removed, and all of their mass has been redistributed.[16]. The procedure will remind some of the EM algorithm.

Supplementary Exercise 4.2 Take a simple survival dataset with just 5 observations, 2 censored and 3 not, such as the 5 values 2, 5+, 6, 7+ and 9. Derive the K-M estimate of $S(t)$. Illustrate the 'self-consistency' of the KM estimator, and that the 'distribution to the right' procedure produces the KM estimate.

4.4.3 The Nelson-Aalen estimator of $S(t)$

Just as with K-M, divide the entire interval $[0, t]$ into J *narrow* event-containing sub-intervals; ignore the 'non-event-containing' sub-intervals. Sub-interval j is defined by distinct event-time t_j, with n_j at risk just before the event(s) [death(s)] in that interval. (there can be more than 1 event at the same t_j, particularly if time is measured coarsely). The (step-)function $n(t)$ is the number at risk at each time point in $(0, t)$. 'Riskset'$_j$ = the n_j'candidates' for the event(s) at t_j. Suppose s_j survive event-containing sub-interval j, and that the remaining $d_j = n_j - s_j$ do not [the letter d is used here because in many applications, the 'transition' ('event') is from the initial state of 'alive' to the destination state of '$d$ead', but transitions may be desirable or underirable.]

The Nelson-Aalen Estimator uses the same general formula that links the $S(t)$ and $ID(t)$ or $\lambda(t)$ functions:

$$\widehat{S_{NA}(t)} = \exp\left\{ -\int_0^t ID(u)du \right\} = \exp\left\{ -\int_0^t \lambda(u)du \right\} = \exp\left\{ -\sum \frac{d_j}{n_j} \right\}$$

Think of a fitted ID function $ID(t)$ with $\widehat{ID(t)} = 0$ in the non-event-containing sub-intervals of $(0, t)$ and $\widehat{ID(t)} = d/PT = d/(n \times \delta t)$ in each event-containing interval of width δt; thus $\widehat{ID(t)} = d_j/(n_j \times \delta t)$ in event-containing interval j.

Supplementary Exercise 4.3 (a) Using the $\widehat{ID(t)}$ function just described, evaluate the integral of $\int_0^t \widehat{ID(u)}du$ and use it to obtain the Nelson-Aalen estimator of $S(t)$. (b) Derive the conditions under which the K-M estimator $\prod \frac{s_j}{n_j} = \prod\{1 - \frac{d_j}{n_j}\}$ gives a result that is very close to that of the Nelson-Aalen estimator. (c) Assuming $d_j \sim Poisson(n_j \times \delta t)$, derive an expression for $Var[\widehat{S(t)_{NA}}]$.

[16] Google "Efron distribute to the right Kaplan Meier"

4.5 Examples of the Kaplan-Meier method

Example 1 Cf. JUPITER data on the website for course `EPI634`.

The `R` code calls the "canned" routines, but also derives the K-M-based cumulative incidence curves 'from scratch.'

Example 2 Figure 2 below is from the article: "Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial" (Lancet 2007; 369: 643-656). If interested, and if you don't have direct access to the Lancet site, the full article is also available under "resources for rates" in course `EPI634`. There you will also find a companion article for a similar randomized trial, with similar estimates of benefit, carried out in Uganda, and published back to back with the one from Kenya.

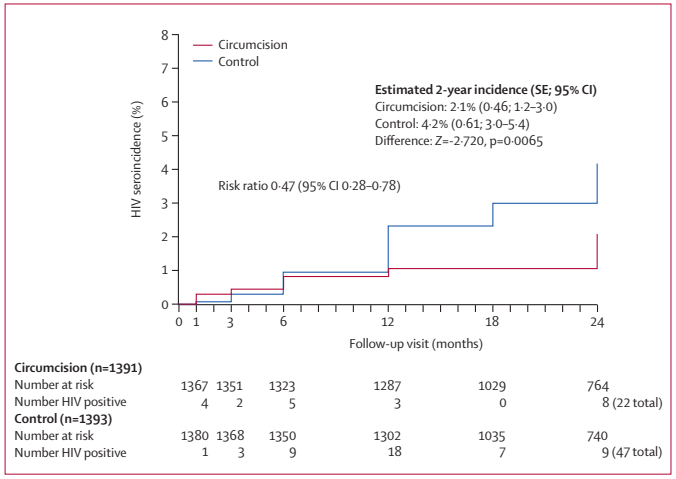

Figure 2: **Cumulative HIV seroincidence across follow-up visits by treatment**
Time to HIV-positive status is taken as the first visit when a positive HIV test result is noted. Time is credited as the follow-up visit month. Participants without HIV-positive status are censored at the last regular follow-up visit completed where HIV testing was done, credited specifically as months 1, 3, 6, 12, 18, and 24.

Supplementary Exercise 4.4 Replicate the statistics reported in the insert beginning with the text "Estimated 2-year incidence" in the top right portion of the above Figure 2.

Example 3 The items below are from "Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial," Lancet 2007; 369: 657-666.

	Intervention group	Control group	Incidence rate ratio (95% CI)	p value
0–6 months follow-up interval				
Number of participants	2263	2319		
Incident events	14	19		
Person-years	1172·1	1206·7		
Incidence per 100 person-years	1·19	1·58	0·76 (0·35–1·60)	0·439
6–12 months follow-up interval				
Number of participants	2235	2229		
Incident events	5	14		
Person-years	1190·7	1176·3		
Incidence per 100 person-years	0·42	1·19	0·35 (0·10–1·04)	0·0389
12–24 months follow-up interval				
Number of participants	964	980		
Incident events	3	12		
Person-years	989·7	1008·7		
Incidence per 100 person-years	0·30	1·19	0·25 (0·05–0·94)	0·0233
Total 0–24 months follow-up				
Cumulative number of participants	2387	2430		
Cumulative incident events	22	45		
Cumulative person-years	3352·4	3391·8		
Cumulative incidence per 100 person-years	0·66	1·33	0·49 (0·28–0·84)	0·0057

Table 3: **HIV incidence by study group and follow-up interval, and cumulative HIV incidence over 2 years**

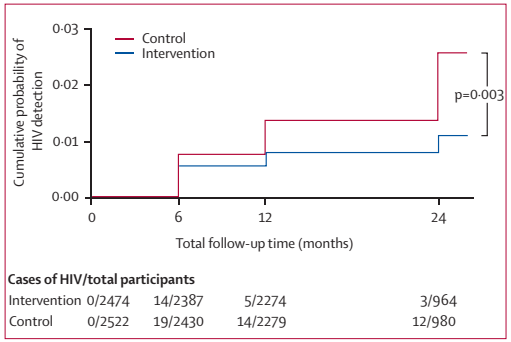

Figure 2: **Kaplan-Meier cumulative probabilities of HIV detection by study group**
Actual visits grouped by the three scheduled visits at 6 months, 12 months, and 24 months after enrolment. The cumulative probabilities of HIV infection were 1·1% in the intervention group and 2·6% in the control group over 24 months.

Supplementary Exercise 4.5 Comment on the appropriateness of (i) the term "*Cumulative incidence per 100 person-years*" in the last row of Table 3 (ii) using a single incidence (hazard) rate ratio of 0.49 for the full 2 years, and in the abstract, reporting that the estimated efficacy of intervention was 51%.

5 Rates

5.1 The probability rate (hazard rate)

JH is not sure why the authors used the term *probability rate*, when the term *hazard rate*[17], or short-term incidence density, or even just *rate*, or *instantaneous rate*, would have done. The only virtue JH sees for this term is that – unlike the term hazard rate – it is somewhat explanatory: the term does indeed convey, and help you remember, the idea that it is the *probability per unit time*. JH has seen many people struggle to remember and accurately reproduce the definition of the hazard rate. The one item that is not conveyed directly by any of these terms is the *conditional* nature of the probability: it has as its denominator those people, or that person time experience lived by those, who reached the "t" that marks the beginning of the small (infinitesimal) interval.

Another way to think of it is as the limit, as the width of the time band is shrunk to zero, of the incidence density (ID).

Since every realistic and epidemiologically interesting time interval has a non-zero width, and since in any case we usually use the hazard rate as a smooth function of time, the idea of it as an instantaneous rate is merely a mathematical nicety. Indeed, we would immediately multiply this rate into some amount of person time PT (which we can depict as a rectangle with height P persons and width T time units) to get an expected number of events, or for the individual, the conditional probability.[18] The point is that if we were to reverse the process from the expected number of events in a certain PT, the ratio of no. of events to PT would remain the same as we shrunk the width of this time slice, and the corresponding number of events. If it did not, it would imply that the intensity is changing quickly over time, and that a single average intensity (or the corresponding conditional probability) is misleading.

In fact, the force of human mortality is – after a certain age – a monotonically increasing function of attained age (note the conditioning on attained age) but practically speaking, the values of the hazard function at age 32.564 and at 32.565 (or indeed over the age range 32 to 33) are similar enough that we can quite closely approximate this monotonically increasing hazard function (force of mortality) in this age band as a constant, and over a larger age range as piecewise constant within each 1-year age band. If we were concerned with the shape of the hazard function after an attained age or 104, we might want to make the time bands narrower. And at age 32, we might want to make them a bit wider than 1 year: see the value of the q function in the 1-year Canadian lifetables, where $q$is the conditional failure probability for age bands 1 year wide (h=1 in the terminology of section 5.3)

"The probability rate refers to an individual subject. This is counterintuitive to many epidemiologists."

This is also counterintuitive to JH, who doesn't understand where these authors are coming from on this. An incidence density is certainly not about an individual person. How are we to think of a failure rate of 8 ruptures per 10000-pipe-kilometer-years of operating pipeline of a water distribution system?

The authors however do well to ask us to distinguish between the definition of the *parameter*, and an *estimate* (or estimator) of the value of this parameter in a particular context (e.g. the rupture rate when the temperature is in the vicinity of -20C.

Mathematically, then, here are a few definitions of what they call the probability rate, or simply the instantaneous rate, at time t. Since it is a parameter, we will, as they do, give it the Greek letter lambda, λ. With P the number of persons at risk at t, or more realistically, the average number of persons at risk over the entire interval $(t, t+\delta t)$,

$$\lambda(t) = \lim_{\delta t \to 0} \frac{\text{Expected no. of events}}{P \times \delta t}$$

One can re-write this as

$$\lambda(t) = \lim_{\delta t \to 0} \frac{\text{Expected no. of events}}{P} \div \delta t$$

so that the Expected no. of events/$Person$ is a probability. This probability, when divided by δt becomes the (conditional) failure probability *per unit time* that the authors use as their definition.

One will also see in survival analysis textbooks the definition of $\lambda(t)$ or $h(t)$ as

$$h(t) = \lambda(t) = f(t)/S(t),$$

[17] The Website `jeff560.tripod.com/h.html` "Earliest Known Uses of Some of the Words of Mathematics" tells us: HAZARD RATE came into use in statistics in the 1960s as a general term for what is called the force of mortality in demography and the intensity function in extreme value theory. David (2001) finds "hazard rate" in R. E. Barlow; A. W. Marshall & F. Proschan "Properties of Probability Distributions with Monotone Hazard Rate," Annals of Mathematical Statistics, 34, (1963), 375-389. A JSTOR search found "death-hazard rate" in D. J. Davis "An Analysis of Some Failure Data," Journal of the American Statistical Association, 47, (1952), 113-150.

[18] Freedman, in his nice article, Survival Analysis: A Primer" in the American Statistician in May 2008 (see resources for survival for course `EPI634`) puts it nicely: "The intuition behind the formula is that $h(t)dt$ represents the conditional probability of failing in the interval $(t, t+dt)$, given survival until time t."

where $S(t)$ is the 'survival' function, i.e., $1 - F(t)$, and $f(t)$ the probability density function, of the 'time to event' random variable. This is no different from the definition above, since we can write it as

$$h(t) = \lambda(t) = \frac{f(t)\delta t}{S(t)} \div \delta t.$$

S(t) is the proportion of persons who are at risk (event-free) at time t, and $f(t)\delta t$ is the (unconditional) fraction of events that occur within the interval $(t, t+\delta t)$, so $\frac{f(t)\delta t}{S(t)}$ is itself a (conditional) fraction of a fraction.

Moreover, we can rewrite the definition as

$$h(t)dt = \lambda(t)dt = \frac{-dS(t)}{S(t)}$$

and integrate both sides over the interval $(0, T)$ to get

$$\int_o^T h(t)dt = \int_o^T \lambda(t)dt = \int_o^T \frac{-dS(t)}{S(t)} = -\log S(T).$$

Then, exponentiating both sides, we get the fundamental relationship between the incidence density function (alias hazard function ($h(t)$, or the maybe more familiar term 'failure rate function', $\lambda(t)$) and the complement of cumulative incidence (CI), namely

$$1 - CI_{0\to T} = S(T) = \exp\left[-\int_o^T h(t)dt\right] = \exp\left[-\int_o^T \lambda(t)dt\right].$$

Notice also the (welcomed) use throughout the book of λ as an event *rate*, and not – as some books use it – as the expected *number* of events, i.e. as the mean parameter of a Poisson distribution. JH has tried to be consistent in using the Greek letter μ for the expected number of events, since after all it is the mean or expected value of the random variable, and since it is important to keep the distinction between the numerator and denominator of an event rate parameter.

5.2 Estimating *the* probability rate parameter

Notice the use of the word *the*, i.e., that the parameter value is assumed constant in the follow-up period of interest.

5.3 The likelihood for a rate parameter

You might find it strange that the authors don't go directly to the representation of the observed rate as an observed Poisson numerator divided by a known PT denominator. I think they did this to emphasize the idea of subdividing the PT into person-clicks.

It is interesting that in 1907 Gosset (of Student-t fame) derived the Poisson distribution 'from scratch' using this same conceptual subdivision of a plate (or field in a microscope) into a large number of small squares, small enough that only one yeast cell would fit in it (C&H in section 4.4 write of time bands so narrow that "each failure occupies a band by itself").[19] If the mean number of cells per plate was μ and the area of the plate was A, or $N = A/a$ small squares of area a each, then the probability π that a small square contains a square is $\pi = \mu/N$. The probability that the total area A will contain y yeast cells is then

$$Pr(y \text{ occupied cells}) = {}^{N}C_y \; \pi^y (1-\pi)^{N-y}.$$

Gosset used Stirling's approximation, and the definition of $e^x = \exp[x]$ as a limit, to go from this binomial probability to the Poisson probability $\exp[-\mu]\ \mu^y/y!$

If we worked with μ directly, then (ignoring the factorial, which doesn't involve this parameter), the likelihood based on an observed count of D is

$$\exp[-\mu]\ \mu^D.$$

Substituting $\mu = \lambda Y$, where Y is C&H's notation for amount of person-Years (what we call the denominator) gives

$$\exp[-\lambda Y]\ (\lambda Y)^D,$$

or, ignoring items that do not involve λ, as

$$\exp[-\lambda Y]\ (\lambda)^D,$$

so that the log-likelihood is indeed

$$-\lambda Y + D \log\,(\lambda),$$

[19] JH has put this very readable 1907 article "*On the Error of Counting with a Haemacytometer*" under the resources for rates in course `EPI634`

5.3.1 Example: Likelihood for parameter of exponentially distributed random variable, with interval censoring.

The Uganda and Kenya 'circumcision in the prevention of HIV' studies are examples of interval-censored (as well as the usual right-censored) data, since one cannot know exactly when a person became HIV+, only that it occurred in the interval between the last negative test and the first positive one.

Before setting up the likelihood for such data, let us consider a simple statistical model for the data, and let us focus for now on the placebo group. *We will assume that the sero-conversion rate λ is constant over the 2 years,* i.e., that $\lambda(t) = \lambda$ over that interval. Up until now, we treated the number of events in the 'aggregated-across-subjects' person time as a Poisson random variable. *Another way to look at this is to consider the inter-event times, (or the time-to-event times) and their distribution.* We know from BIOS601 that if the event rate is λ, and there is always one unit at risk, then the inter-event times have an exponential distribution with mean $1/\lambda$. Thus, we can say that the 'time-to-event' for each subject is a realization of an exponential random variable with mean or expected value $1/\lambda$. If we call this r.v. 'T', then

$$T \sim \exp(\mu_T = 1/\lambda),$$

$$S_T(t) = \exp[-\lambda t],$$

$$F_T(t) = 1 - S_T(t) = 1 - \exp[-\lambda t],$$

$$f_T(t) = F'_T(t) = \lambda \exp[-\lambda t] = (1/\mu_T)\exp[-(1/\mu_T)t].$$

In the control group in the Uganda trial, 2319 initially HIV- men were tested at the 6-month, or 0.5year follow-up, and 19 of them were found to be HIV+, and the remaining 2300 were found to be HIV-.

The likelihood, based just on this first follow-up test is therefore the probability (as a function of the seroconversion rate λ) of observing this pattern of results. First we write it as a product of 2319 probabilities:

$$Likelihood = \prod_{i=1}^{i=2319} Pr[obs'd\ outcome\ for\ subject\ i] = \prod_{i=1}^{i=19} Pr_i \prod_{i=20}^{i=2319} Pr_i$$

With T denoting the r.v. 'time to HIV+', each Pr_i in the second product is of the form $Pr[T > 0.5 \mid \lambda] = \exp[-0.5\lambda]$, while each Pr_i in the first product is of the form $Pr[T < 0.5 \mid \lambda] = 1 - \exp[-0.5\lambda]$. The likelihood based on this first test can thus be simplified to

$$L_{1st\ test} = \exp[-2300 \times 0.5\lambda] \ \times \ (1 - \exp[-0.5\lambda])^{19}$$

Some 2229 of those HIV- at 6-months were tested at the 12-month, or 1year follow-up, and 14 of them were found to be HIV+, and the remaining 2215 were found to be HIV-. Thus the likelihood based on this second test can thus be simplified to

$$L_{2nd\ test} = \exp[-2215 \times 0.5\lambda] \ \times \ (1 - \exp[-0.5\lambda])^{14}$$

Notice that with this exponential distribution, the fact that these 2229 had got throught the first interval HIV-free has nothing to do with their (now conditional) probabilities for the next 6 months. Technically, we call this the "memoryless" property of the exponential distribution.[20] Thus, $Pr[T > t \mid T > t_{given} = Pr[T > t - t_{given}]$, and so, whereas we would normally have to use the *conditional* probability $\{F(1.0) - F(0.6)\}/S(0.5)$, here we can use the unconditional probability of escaping infection for 6 months. In effect, we can 'reset the clock to zero at T=0.5,' and imagine it was just like back at $T = 0$.

Some 980 of those HIV- at 12-months were tested at the 24-month, or 2year follow-up, and 12 of them were found to be HIV+, and the remaining 968 were found to be HIV-. The likelihood based on this third test can thus be simplified to

$$L_{3rd\ test} = \exp[-968 \times 1.0\lambda] \ \times \ (1 - \exp[-1.0\lambda])^{12}$$

Thus the likelihood based on all three tests is

$$L_{all\ 3\ tests} = L_{1st\ test} \ \times L_{2nd\ test} \ \times \ L_{3rd\ test}$$

ie

$$L = \exp[-(2300 \times 0.5 + 2215 \times 0.5 + 968 \times 1.0)\lambda]$$
$$\times$$
$$(1 - \exp[-0.5\lambda])^{12} \ \times \ (1 - \exp[-0.5\lambda])^{14} \ \times \ (1 - \exp[-1.0\lambda])^{12}$$

Supplementary Exercise 5.1. (i) Maximize L with respect to λ. (ii) What would happen to L, and to the ease of estimation, if subjects were tested more frequently, e.g. every month, every week, every day?

[20] In industrial life-testing, this property is referred to as the 'used is the same as new' property. In failure time distributions where the failure is a function of age or duration of use (e.g. a computer or hard disk), the hazard is — maybe after a certain run-in period – an increasing function of its age or accumulated hours of work, and so the testers say 'older is worse (less 'reliable') than newer;' initially, before those units doomed to early failure have been weeded out, it may be that 'newer is worse than older.' Sadly, most human hazards, other than being struck by a meteor, are from internal sources to do with our own bodies, and so while the hazard function or force of mortality decreases until about age 8 – see Canada lifetables – it is monotonically increasing thereafter.

5.4 Cum. survival probability as fn. of rate parameter

We saw this in BIOS601 as $S(T) = \exp[-\int_0^T h(t)dt]$, or cumulative incidence as $CI_{0\to T} = 1 - S(T) = 1 - \exp[-\int_0^T h(t)dt]$.

We also came up with a 'heuristic' ("a usually speculative formulation serving as a guide in the investigation or solution of a problem") whereby the integral $\int_0^T h(t)dt$ can be seen as the expected number of events, μ, if there was always one unit (person) at risk for the period 0 to T. Thus if an event (failure) occurred at any point in this interval, the failed unit is immediately replaced by another of the same profile: e.g., if $h(t)$ referred to computers, we would replace a computer that failed at time t_1 by another of the same age, and if this failed before T, at time t_2 say, we would in turn replace it by another of age t_2, and so on until we got to T. So by the end, we would have observed the 1-unit system for a total of T units of time, and we might have observed $0, 1, 2, \ldots$ failures (and had to make this many replacements), in order to have the system in continuous operation for this duration. The expected number of failures in that period would be the integral of (the area under) the $h(t)$ curve. We saw in first term that the Poisson distribution has the 'closed under addition' property; in this application, we can think of the total number of events in $(0, T)$ as (the limit of) a sum of more and more Poisson random variables, representing the numbers of events in smaller and smaller intervals $(t, t + dt)$, with expected numbers of events $h(t)dt$. In the limit, this sum of small expectations is nothing more than the overall expected number of events,

$$\mu = \int_0^T h(t)dt$$

The observed sum is thus the realization of a single Poisson random variable with mean μ, and so the probability that the initial unit will 'survive' the entire interval is just the probability that there will be no event in the entire period, i.e.,

$$S(T) = Pr(Poisson.RV[\mu] = 0) = \exp[-\mu] = \exp[-\text{integral of } h(t)].$$

The other concept that is reinforced by this heuristic, and the computer example, is that the computer-days are interchangeable. Imagine we had a large bank of computers all of the same vintage: we could imagine having a different one of these computers be the one that ran the system (was 'on duty') for the day, and we could even draw lots for which computer is the one on duty at any time. Assuming that the 'on duty' computer didn't age any faster than the ones that were 'off duty' that day, we can now see that the probability that a *specific* computer would fail before time T is the same as the probability that a *sequence of computer-days – or computer-hours, or computer-minutes* (each one contributed by a possibly different computer) would contain at least one failure. This *interchangeability* of (impersonal, indistinguishable, unnamed) units of the same age, i.e., with the same $h(t)$, is central to the concept of 'person-clicks' that C&H use.. it is not the particular person that matters to the contribution, but the person's *profile* – his/her $h(t)$ value.

If the rate is a constant over the period $(0, T)$, so that the integral is $\mu = \lambda \times T = \lambda T$, then we get the simple expression for the (cumulative) survival probability given at the top of page 46, namely $S(T) = exp[-\lambda T]$.

This section also discusses the simple approximation to $exp[-\mu]$ when μ is small, namely $1 - \mu$. In this situation, the cumulative risk (in fact, the word *cumulative* is redundant!) can thus be approximated by

$$\text{Risk} = \text{Cumulative Incidence } \approx 1 - \mu = 1 - \lambda T \quad [\mu \text{ small}].$$

Whether or not the integral μ is small, if λ is constant over $(0, T)$, then – apart from random variations –

$$\log\{S(t)\} = \log\{\exp[-\lambda t]\} = -\lambda t,$$

so that

$$\text{the plot of } -\log\{S(t)\} \textit{ versus } t \text{ should be linear in } t, \text{ with slope } \lambda.$$

5.5 Rates that vary with time

JH's comments in section 5.4 discussed both piecewise-linear (and in the limit a) general smooth form(s) for $h(t)$ or $\lambda(t)$, and so there is little to add for this section, other than to make one remark about their use of the term "*cumulative failure rate*." JH finds this term too close to "cumulative incidence", which is a proportion. C%H's "cumulative failure rate" is in fact the integral we discussed above, and so has as its dimension or units the expected number of events in the period $(0, T)$ if one unit were always operating, i.e., 'at risk.' He would prefer that you use the more common term "*integrated hazard*" often denoted by an upper case letter,

$$H(T) = \int_0^T h(t)dt \quad or \quad \Lambda(T) = \int_0^T \lambda(t)dt.$$

C&H tell us that "it follows that the relationship

$$\log[S(t)] = -Cum.\ failure\ rate \ \{\ \log[S(t)] = -H(t) \ \ in\ our\ notation\ \}$$

still holds when the rate varies from one band to the next... and will be used to calculate $S(t)$." We have already used the exponentiated version of this to calculate $S(t)$. But this relationship in the log scale is also used to check whether an assumed form or model for $h(t)$ fits with the observed data: it is more difficult to judge fit on the S scale, where $S(t)$ is likely to be quite curvilinear, than on the H scale, where $H(t)$ may have a simpler form, such as piecewise linear.

Supplementary Exercise 5.2. For the Uganda HIV data, assume a different λ for each of the 3 intervals, and estimate each one separately. Do the data provide evidence against this assumption? Answer by maximizing L under the larger (3 possibly different λs) and smaller (all three λs are the same) models, and computing the likelihood ratio.

5.6 Rates varying continuously in time: Kaplan-Meier (K-M) and Nelson-Aalen (N-A) estimators

"*The assumption that the rate parameter is constant over broad bands of time, but changes abruptly from one band to the next, is widely used, but an alternative model, useful when exact times of failure and censoring are known,* ***is to allow the rate parameter to vary from click to click****. In Chapter 4 this kind of model led to the Kaplan-Meier estimate of the survival curve; when using rates it leads to the estimate known as the Aalen-Nelson estimate.*"

This is a very nice way of putting it. First, it says that the Kaplan-Meier curve is a limiting case of a probability-based lifetable, with the time bands made narrower and narrower. In the limit (and the Kaplan-Meier table is sometimes referred to as the 'product-limit' table) one need only be concerned with products of continuation probabilities from the event-containing intervals. It also explains why the Kaplan-Meier curve is called 'non-parametric': by making the bands narrower and narrower, the curve follows the data exactly.

The Kaplan-Meier estimate can be seen as a product of *empirical* continuation *probabilities*, each one governed by the *binomial* model. We formally acknowledge this when we use Greenwood's formula for the SE of $\widehat{S(t)}$.

The Nelson-Aalen estimate can be seen as a product of model-based continuation *probabilities*, with each estimated probability calculated from the theoretical relation between the (in this case shortterm incidence or) hazard rate and cumulative incidence, viz. $S_{t\to t+dt} = 1 - CI_{t\to t+dt} = \exp[-\int_t^{t+dt} h(u)du$

If an interval $t, t+dt)$ involves n persons at risk, and d events (deaths), then the person time is ndt and so the estimate of the incidence is $\frac{d}{n\times dt}$. each one governed by the *binomial* model. If d is zero, then the estimate of the incidence is zero. Thus, the empirical hazard function is a square-wave function,

$$\widehat{h(t)} = \begin{cases} 0 & \text{if } (t, t+dt) \text{ contains } d = 0 \text{ events,} \\ \frac{d}{n\times dt} & \text{if } (t, t+dt) \text{ contains } d > 0 \text{ events.} \end{cases}$$

Thus,

$$\widehat{h(t)dt} = \begin{cases} 0 & \text{if } (t, t+dt) \text{ contains } d = 0 \text{ events,} \\ \frac{d}{n} & \text{if } (t, t+dt) \text{ contains } d > 0 \text{ events.} \end{cases}$$

Thus

$$\int_0^T \widehat{h(t)}dt = \sum \frac{d}{n},$$

with the summation over those event-containing narrow bands where $t < T$. The persons at risk in these event-containing bands are called *risksets*.

The `EPIB634` site has R code that divides the JUPITER follow-up time into 1-year, then 1-month, then 1-week, then 1-day bands. The resulting h(t) function becomes more and more erratic, but in doing so – just like the K-M curve – it conforms exactly to the data.

Just as the K-M curve is based on a product of *binomial*-based probability estimates, the N-A curve can be seen as an integral (the limit of a sum) of *Poison*-based rate (hazard) estimates: provided that each n is large, the 'd' that forms the numerator of the empirical elemental area can be seen as a realization of a Poisson random variable. Its estimated variance can therefore be estimated as d, and the variance of $\frac{d}{n}$ as $\frac{d}{n^2}$. Thus,

$$\widehat{Var}\left[\int_0^T \widehat{h(t)}dt\right] = \sum \frac{d}{n^2}.$$

For the numerators in this variance expression, some textbooks use binomial-based variances of $n \times \frac{d}{n} \times \frac{n-d}{n}$ instead of the Poisson-based variances of d. If each $n-d$ is large, as it is in the JUPITER study, then the difference between the two formulations is miniscule.

Most software packages plot the N-A curve as a step-function, just as they do the K-M curve. The conf. intervals are first calculated for the estimated integral, and then for $\widehat{S(t)}$.

Supplementary Exercise 5.3. Calculate the Nelson-Aalen and Kaplan-Meier curves, and the SE's, for the placebo arms of the Uganda and Kenya circumcision trials, and the JUPITER trial.

6 Time

6.1 When do we start the clock?

Examples JH has dealt with include the analysis of longevity of

- The Titanic survivors, where the two time scales are (i) age (years elapsed since birth) and (ii) 'survivor-time', the years elapsed since the April 15, 1912 sinking;
- Oscar nominees, where the two time scales are (i) age and (ii) nominee-time', the years elapsed since first being nominated for an Oscar;
- Nobel Prize nominees, where the two time scales are (i) age and (ii) 'nominee-time', the years elapsed since first being nominated for a Nobel Prize;
- Jazz musicians, where the two time scales are (i) age and (ii) performer-time', the years elapsed since first becoming a jazz musician;
- Popes versus artists;
- Baseball Hall of Famers versus players who were nominated by not inducted;
- Rock Stars who become famous early versus later (or not at all).

  For more details on these examples, see `bios601/Epidemiology2/`

For more on the choice of time scale, Google "Multiple time scales in survival analysis." or find the articles that cite the 1979 Applied Statistics article by Farewell and Cox "A note on multiple time scales in life testing."

There is also the interesting article *The two-way proportional hazards mode*l by Efron in J. R. Statist. Soc. B (2002) 64, Part 4, pp. 899-909, applied to "patient histories in a study of heart transplant recipients treated at the Stanford Medical Center between 1980 and 1996; some 110 of the patients suffered a *serious bacterial infection*, their infection times ranging from a few *days after transplantation* to nearly 9 years, these being the observed lifetimes that would usually be featured in a proportional hazards analysis of the infection process. In this case, however, the investigators' *main interest centred on calendar date*: was the *incidence rate* of bacterial infections *declining over the course of the study?* Incidence is itself a hazard rate, in the simplest situation the number of new cases per eligible subject per unit time, and it is natural to answer the question with a hazard rate analysis."

6.2 Age-specific rates

"*To ignore this variation [of incidence and mortality rates with age] runs the risk that comparisons between groups will be seriously distorted, or confounded, by differences in age structure.*"

It's good to have a few handy real examples of *age-confounding* that are easily understood by non-statisticians. Two immediately come to mind (i) the overall death rate is higher in Canada than Ethiopia (ii) the higher death rate among non-smokers in a 20-year follow-up study of smokers and non-smokers [Does Smoking Improve Survival? `www.whfreeman.com/statistics/ips/eesee4/eesees4.htm`; this is also described in chapter 1 of Rothman 2002, with finer age-categories]

"*For longer studies it will be necessary to take account of changing age during the study, and to treat age properly - as a time scale. This scale is then divided into bands and a separate estimate of the rate is made within each age band as described in Chapter 5. In this latter analysis, a subject can pass through several age bands during the course of the study.*"

Not only can a subject pass through several age bands but she can also change from one 'exposure' category to another – as in the Oscars exercise.

6.3 The expected number of failures

"*One reason for subdividing the total follow-up experience of a cohort into age bands is to determine whether the observed number of failures is more or less than we might have expected. Since mortality and incidence rates usually increase quite sharply with age, the distribution of person years observation between age bands is an extremely important determinant of the number of events we would expect to observe.*"

It is not clear what is the basis for the "expectation" i.e., whether it is a 'what if' comparison against external rates, or an internal one against the rates in a comparison group constructed and followed by the investigators. One can think of the 'expected number' of 16.77 cases in exercise 6.3 as the number one would expect in a scaled-down version of England and Wales (E&W), scaled down to the same sample size (974 women) followed for the same cell-specific numbers of person years as those shown in Table 6.4. In other words, it as as thought one had

974 treated by HRT	974 from E&W, same age & follow-up, untreated
15 cases	16.7 cases

Of course, the fact that the 16.7 is based on observed rates in the whole of

E&W means that it is not subject to the same degree of random variation as is the number of cases in the actual cohort. With this solid a basis for it, the expected number is usually taken to be a constant, so only one standard error (SE) is involved in the 15 vs. 16.7 comparison – the one associated with the 15.

"The expected number of cases, as calculated above, is not quite the same as the expected number in the usual statistical sense. The latter cannot depend upon the outcome of the study, but the former does."

C&H are saying that the numbers of Woman-years in the second column of Table 6.4 are random variables: they would not have been known ahead of time. For some 15 women – the 15 being a random variable – the follow-up was terminated by the event of interest. Likewise, any terminations for other reasons might also be unpredictable ahead of time. However, if these are not related to the person's probability of a future event, they don't have a great influence on the sampling behaviour of the estimators of interest.

6.4 Lexis diagrams

`en.wikipedia.org/wiki/` Wilhelm Lexis (1837-1914) was an eminent German statistician, economist, and social scientist and a founder of the interdisciplinary study of insurance.

The "Lexis diagram", in which lifelines are displayed as 45-degree lines on a grid with age on the vertical axis and calendar year on the horizontal axis, is very helpful in epidemiology, and in survival analysis with 2 time scales.

The `Epi` package for `R` has several functions that make it easy to convert the data of the type shown in Table 6.2 into the person-year segments shown Figure 6.3. Previously, this was a very laborious computing process.

Once we have the tabulated person years and cases in each Lexis rectangle (the cells don't have to be square), we can calculate the expected number of cases if a specified set of external rates applied, or make internal rectangle-by-rectangle comparisons, and thus a summary of these comparisons. We can also use them to fit (Poisson) regression models for rates.

Here is the `R` code, and some of its output, for the data in C&H Table 6.2.

```
library(Epi)

id = c(1,2,3,4);
yr.birth = c(1904,1924,1914,1920);
yr.entry = c(1943,1948,1945,1948);
yr.exit  = c(1952,1955,1961,1956);
fail = c(0, 1, 0, 0) );

ds=data.frame(id, yr.birth, yr.entry, yr.exit, fail); ds

  id yr.birth yr.entry yr.exit fail
1  1     1904     1943    1952    0
2  2     1924     1948    1955    1
3  3     1914     1945    1961    0
4  4     1920     1948    1956    0


# Define as Lexis object with timescales calendar time and age

Lexis <- Lexis( entry = list( calendar.year = yr.entry ),
                exit  = list( calendar.year = yr.exit, age = yr.exit - yr.birth ),
         exit.status  = fail,
                data  = ds )
Lexis

  calendar.year age lex.dur lex.Cst lex.Xst lex.id id yr.birth yr.entry yr.exit fail

1          1943  39       9       0       0      1  1     1904     1943    1952    0
2          1948  24       7       0       1      2  2     1924     1948    1955    1
3          1945  31      16       0       0      3  3     1914     1945    1961    0
4          1948  28       8       0       0      4  4     1920     1948    1956    0

# Default plot of follow-up

plot(Lexis)

# With a grid and deaths as endpoints

plot(Lexis, grid=0:5*5, col="black" )
points(Lexis, pch=c(NA,16)[Lexis$lex.Xst+1] )

# With a lot of bells and whistles:  [ *** SEE PLOT NEXT PAGE *** ]

plot(Lexis, grid=0:20*5, col="black", xaxs="i", yaxs="i",
      xlim=c(1940,1965), ylim=c(20,50), lwd=3, las=1 )
points(Lexis, pch=c(NA,16)[Lexis$lex.Xst+1], col="red", cex=1.5 )

# Split time along two time-axes

L2 = splitLexis(Lexis,breaks=seq(1940,1965,5),
   time.scale="calendar.year")
L2 = splitLexis(L2,   breaks=seq(20,50,5), time.scale="age" )
str( L2 )
```

```
L2
```

	lex.id	calendar.year	age	lex.dur	lex.Cst	lex.Xst	id	yr.birth	yr.entry	yr.exit	fail
1	1	1943	39	1	0	0	1	1904	1943	1952	0
2	1	1944	40	1	0	0	1	1904	1943	1952	0
3	1	1945	41	4	0	0	1	1904	1943	1952	0
4	1	1949	45	1	0	0	1	1904	1943	1952	0
5	1	1950	46	2	0	0	1	1904	1943	1952	0
6	2	1948	24	1	0	0	2	1924	1948	1955	1
7	2	1949	25	1	0	0	2	1924	1948	1955	1
8	2	1950	26	4	0	0	2	1924	1948	1955	1
9	2	1954	30	1	0	1	2	1924	1948	1955	1
10	3	1945	31	4	0	0	3	1914	1945	1961	0
11	3	1949	35	1	0	0	3	1914	1945	1961	0
12	3	1950	36	4	0	0	3	1914	1945	1961	0
13	3	1954	40	1	0	0	3	1914	1945	1961	0
14	3	1955	41	4	0	0	3	1914	1945	1961	0
15	3	1959	45	1	0	0	3	1914	1945	1961	0
16	3	1960	46	1	0	0	3	1914	1945	1961	0
17	4	1948	28	2	0	0	4	1920	1948	1956	0
18	4	1950	30	5	0	0	4	1920	1948	1956	0
19	4	1955	35	1	0	0	4	1920	1948	1956	0

```
# Tabulate the cases and the person-years

summary( L2 )

tapply( status(L2,"exit")==1, list( timeBand(L2,"age","left"),
          timeBand(L2,"calendar.year","left") ), sum )
```

	1940	1945	1950	1955	1960
20	NA	0	NA	NA	NA
25	NA	0	0	NA	NA
30	NA	0	1	NA	NA
35	0	0	0	0	NA
40	0	0	0	0	NA
45	NA	0	0	0	0

```
tapply( dur(L2),  list( timeBand(L2,"age","left"),
          timeBand(L2,"calendar.year","left") ), sum )
```

	1940	1945	1950	1955	1960
20	NA	1	NA	NA	NA
25	NA	3	4	NA	NA
30	NA	4	6	NA	NA
35	1	1	4	1	NA
40	1	4	1	4	NA
45	NA	1	2	1	1

```
> summary( L2 )

Transitions:
     To
From  0 1 Records:  Events:  Risk time:
   0 18 1        19         1           40

Rates:
     To
From 0    1 Total
   0 0 0.02  0.02
```

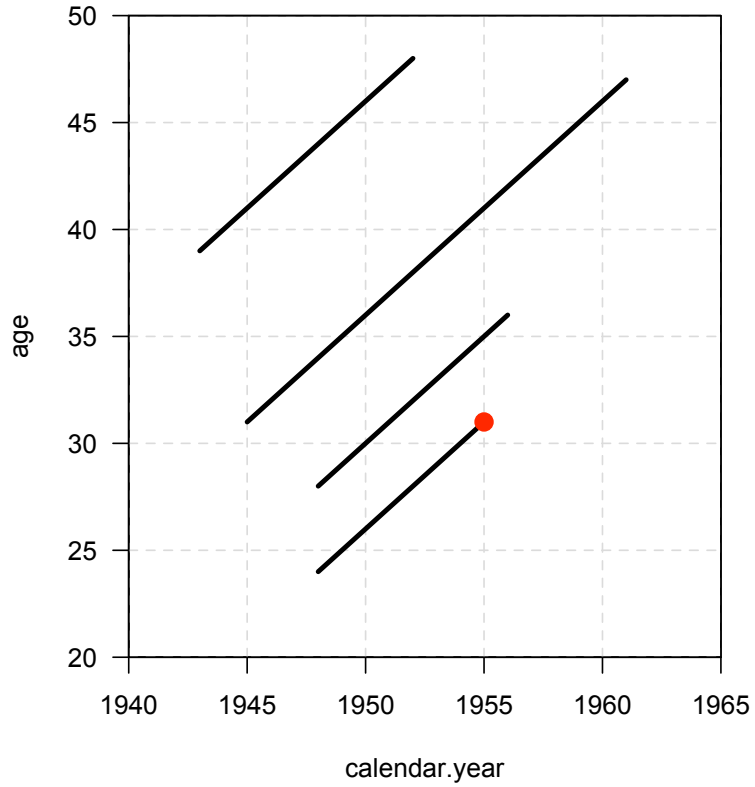

Figure 4: Lexis Diagram, from `Epi` package in `R`

Supplementary Exercise 6.1. Death rates in those who survived the sinking of the Titanic vs. in the sex-and age-matched US general population, together with some other investigations

Under 'For Person-Years Analyses' in Resources for 'Fitting Models to Grouped Data [B & D vol II, ch4]' in the BIOS602 website you will find (a) the Titanic longevity data set (b) USA death rates (within 5 x 5 rectangles, called '*quinquinquennia*') from the Berkeley Mortality Database.[21] You will also find some R code that uses the Epi package to create – for each passenger – the durations in and exit status from each quinquinquennium, then aggregates these over all the persons traversing each quinquinquennium, etc.

1. Convert each survivor's record into the experience in the (age, period) quinquinquennia traversed, i.e the number of years spent in the rectangle, and the status (e.g., $d = 0$ if alive, 1 if dead) at the end of these years. Rather than program the calculations from scratch, two possibilities are `http://epi.klinikum.uni-muenster.de/pamcomp/pamcomp.html` – which some people used last year – and the `R` 'Epi' package `http://staff.pubhealth.ku.dk/~bxc/Epi/` The key functions in the latter are `Lexis` (and associated plotting functions) and `splitLexis`, which, when applied twice, calculates the time spent, and exit status from each quinquinquennium. The 'bogus example' in the documentation of the `splitLexis` function illustrates these, while the example on the notes for C&H chapter 6 shows the application to the 4-person cohort used in that chapter.

2. How much higher/lower is the *set* of age-specific death rates for male Titanic survivors than that for the general US population? for female survivors? Answer in two ways: first, calculate sex-specific observed/expected ratios, where the numerator is the total number of deaths observed in the sex-specific cohort, and the denominator is the sum of the expected numbers of deaths in these cells, using the USA age-sex-period death rates; second, calculate sex-specific Mantel-Haenszel summary incidence ratios (Rothman terminology) or incidence density ratios (Miettinen terminology) or mortality rate ratios (everyone's terminology), using age and period as 'strata.'[22] Assume that each of the USA death rates is based on a denominator of one million person years.[23] Assume that the death rates after 1995 are the same as those in 1990-95.

3. 'On average,' [24], for the age-span 40-90 in the period 1990-1995, how much higher are the USA age-specific male death rates in males than females? Answer by plotting the log of the male:female death rate ratio vs age, (or the two separate sets of log-death-rates on the same graph), and taking some 'typical' value for the ratio. Are you comfortable giving a single ratio? i.e., is the mortality-rate-ratio (M:F) reasonably constant over that age-span?

4. The previous question refers to cross-sectional rates, i.e., those in a specified *period*.[25] On average, over the age-span 40-90 in the 1900 *birth-cohort*, how much higher are the USA age-specific death rates in males than females? Answer by plotting the log of the male:female death rate ratio vs age, (or the two separate sets of log-death-rates on the same graph), and taking some 'typical' value for the ratio. Are you comfortable giving a single ratio? i.e., is the mortality-rate-ratio (M:F) *reasonably* constant over that age-span?

5. For the age-span 40-90, in a single number describe how much age-and specific death rates have fallen over the 20th century (the changes may be more subtle that this, so your answer will necessarily be a simplification).

6. For the Titanic survivors, was there a gradient in mortality rates across the 3 passenger classes?

Supplementary Exercise 6.2. Mortality of performers while in the 'still hoping to win' vs in the 'already a winner' state

1. Divide the performer-years into those spent as Oscar nominees and as Oscar winners and then subdivide these into quinquinquennia.

2. Compare the death rates in the performer-years spent as nominees versus those spent as winners. Do so using both 'adjusted' expected numbers and purely-internal comparisons.

[21].] This site, `http://www.demog.berkeley.edu/~bmd/index.html`, contains historical lifetable and death rate data for the USA and other countries.

[22] As is illustrated in equation 8-5 in Rothman 2002, the formula is

$$\frac{\sum_{strata}(no.\ of\ cases,\ index\ category) \times (py,\ ref.\ category)/(py\ in\ stratum)}{\sum_{strata}(no.\ of\ cases,\ ref.\ category) \times (py,\ index\ category)/(py\ in\ stratum)}$$

[23] If the ratio of the amount of experience in the ref. category to that in the index category goes to infinity, the M-H summary ratio converges to $\sum_{strata} O / \sum_{strata} E = O/E$.

[24] Even if the average is not representative.

[25] Cross-sectional rates are what are used to make 'current' or 'period' lifetables, by far the more common type of lifetable.