"Survival" or " Time-to-event[#] " data

- Examples (events not necessarily 'bad'***)***
  - ***women/couples*** : becoming pregnant;
  - ***fetuses***: being born (gestational age)
  - ***infants***: first sleep through the night, word uttered, walk, tooth, mosquito bite after application of (sham or real) prophylaxis, tooth eruption, caries
  - ***infants***: last breast feeding, diaper (and the 'flip side' thereof *)
  - ***adolescents***: first beer, cigarette, sexual intercourse, driving licence, job, motor vehicle accident; university degree, marriage/cohabitation
  - ***adults***: first gray hair; Ph.D.; divorce; lose job; offspring born; grandchild, cancer diagnosis, menopause, bph, etc....
  - ***new (transient) condition***: (headache, rash, cold,) -> resolution
  - ***(removed??) condition***, e.g. cancer: re-appearance ; death from
  - ***life threatening situation***, eg buried by avalanche: how long survive?
- Play down '***time-to***'; emphasize its reciprocal ***(event rates, hazard function) &*** cumulative incidence
  - at issue is ***exit*** from a state (to another), and the exit **rates**
- Why such data need special techniques
  - not everyone will experience event (no matter how long followed)
  - some haven't been followed for full length of time (enrolled late)
  - some 'lost to view'
  - some die (of unrelated causes) or have the "target" removed
  - [NB "data not symmetrically/normally distributed" not reason ***per se***]
  - [likewise, absence of censored data doesn't mean one can't use survival analysis techniques.. see fruitfly survival data]

[#] **Merriam-Webster** `http://www.m-w.com/cgi-bin/dictionary`
Main Entry: **EVENT**. *Pronunciation*: i-'vent. *Function*: noun. *Etymology*: Middle French or Latin; Middle French, from Latin eventus, from evenire to happen, from e- + venire to come -- *Date*: 1573 1 a archaic : OUTCOME b : the final outcome or determination of a legal action c : a postulated outcome, condition, or eventuality <in the event that I am not there, call the house> 2 a : something that happens : OCCURRENCE b : a noteworthy happening c : a social occasion or activity 3 : any of the contests in a program of sports 4 : the fundamental entity of observed physical reality represented by a point designated by three coordinates of place and one of time in the space-time continuum postulated by the theory of relativity 5 : a subset of the possible outcomes of an experiment

JH would add an 'epi' definition: a ***transition*** from one ***state*** to ***another***.

- Other types of censored data (besides ***right***-censored & ***time***)
  - **left** censored
    - hep c + now, but since when?
    - PSA level post prostatectomy 'undetectable' .. limit of detection
    - thermometer stops at –10C
  - **interval** censored
    - - onset of puberty / caries / when hiv+ : periodic examinations
    - - rounded or grouped measurements (eg age, income)
  - **right** censored
    - measurement off the upper end of instrument scale
    - open-ended category
    - thermometer stops at +40C
- Distinction between censoring and truncation
  - **censoring**
    - **every** (or representative sample of) person(s)/object(s) is observed; have some bounds on the quantity
  - **truncation**
    - **some** person/objects **not** observed / **excluded**, and probability of in/exclusion has to do with the very quantity of interest.. the length of time ... , their size, etc. [length-biased sampling, deliberate exclusions, ..]
      - e.g. 1/2 cross-sectional survey misses those who exit quickly
      - ask in 2004 for list of all the Ph.D. students 'on the books' i.e. active in 1994 and determine in which year (Ph.D. 3, 4, ..) these students got the degree [Alzheimer pts/ Wolfson]; ask in 2004 for list of all patients on the hospital census on randomly selected days in 2002; calculate their average length of stay.
      - e.g. 3/4 sampling design misses objects of short sizes
      - select words by sticking a pin at random on page; measure average length of the words selected. select inter-arrival times of buses, using cross-sectional sampling design
      - e.g. 5/6 measuring instrument misses objects of short sizes
      - e.g. select fish using a given size mesh of net ; lose rapid onset events if counter takes time to reset after previous event .. e.g. cars, radioactive disintegrations etc.
      - e.g. 7/8 exclude pts who die early, before 'an adequate trial of tx; or [for 5-year survival, yes/no], include patients who entered study less than 5 years ago if they already died, but exclude those who entered less than 5 years ago but who have not died.

• [equivalent] Functions: S[t] , hazard h[t] , pdf[t]

T: random variable (duration, time to, time from T_0, etc..)

t: a specific point on T scale (eg 7 days / 5 years post-op)

S[t] (survival function)

$S[t] = Prob[T > t]$ **unconditional**.

can debate whether to use > or ≥ ; by convention in mathematical statistics, we define the complement of the S[t] function, namely 1 – S[t], as
F[t] = Prob[T ≤ t } , so I will use S[t] = Prob[T > t].
In practice, since we measure time in discrete amounts, it is not an issue; survival textbooks are divided on this fine point. F[t] is often called the cdf or cumulative distribution function (maybe that's where the silly term 'cumulative' survival comes from!)

h[t] (hazard function)

$$h[t] = \text{limit, as } \delta t \to 0, \text{ of } \frac{Prob[\, t < T \le t + \delta t \mid T > t \,]}{\delta t} \qquad (1)$$

conditional

Can think of h[t] as a short-term ('instantaneous') rate, in epi sense, with time denominator. To see why, consult page 12, section 1.3 of Collett, or consider the diagram in the next column.

Before taking limit, can see that the conditional probability

$$Prob[\, t < T \le t + \delta t \mid T > t \,]$$

in the top part of expression (1) can be re-written as

$$\frac{Prob[\, t < T \le t + \delta t \,]}{Prob[\, T > t \,]} \qquad (2)$$

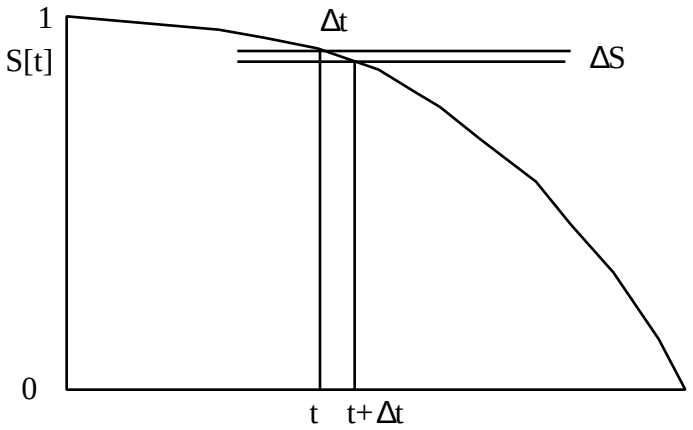
The **Numerator** of (2) is proportional to the **number of deaths** in the interval, just like d_x in a lifetable. ie it is the amount by which S (or lower-case l in lifetable) changes during the interval.

The **Denominator** of (2) is S[t] and, in lifetable terms, is proportional to the number alive at T=t, and so has dimension 'persons'.

Divide the top of (1) by δt to get a quantity proportional to

$$\frac{d}{S[\, t \,] \times \delta t} = \frac{\text{number of deaths}}{\text{Person-time}} \qquad (3)$$

Think of rectangle standing on the base (t, t+ δt) as a person time denominator, and the Δs = d as the 'persons' numerator. As one narrows the δt, the rate hardly changes if the curve is smooth.



In mathematical-statistical terms, we replace d by the product of the probability density function f[t] and the δt, so that the limit, after the δt cancels out, h[t] becomes

$$h[t] = \frac{f[\, t \,]}{S[\, t \,]} = \qquad (4)$$

f[t] is the negative of the derivative of S[t], so can rewrite (4) as

$$h[t] = -\frac{d \log\{ S[\, t \,]\}}{\delta t}, \text{ leading to}$$

$$S[t] = \exp[-\int h[u]\, du \,], \text{ integral from } u=0 \text{ to } u=t \qquad (5)$$

Bottom line.. can reconstruct h[t] from S[t] & vice versa -- or from f[t]

(see alternative derivation Incidence <--> cumulative incidence, survival function Notes by JH in Resources for Lifetables.)

Packages plot the ***negative of the log of the S[t] curve*** against ***t***, since it allows us, when comparing two curves, to judge more easily whether the hazard functions are proportional to each other at all values of t

The integral of h[u] up to t is the area under the hazard function up to t, and is called (not surprisingly) the 'integrated hazard'

• **Summaries** of these (3 equivalent) functions S[t], h[t] and f[t]

- ***median***: the value of t at which S[t] = 1/2 (half-life" or t_{50})

- ***mean***: the area under the (complete) S[t] curve
  (if available)
  equivalent to e_0 in life table

- ***quantile/fractile/percentile***:
  the value of t at which S[t] equals some proportion or %

- ***x-year survival (or cumulative mortality)***:
  the value of S[t] at specified value of t

• "**Cause-specific" Survival; Competing Risks**

treat time of death from another cause (not of interest) as a censored observation (used a lot in cancer statistics)

- ***can give misleading answers if substantial other forces of mortality*** ***(see material on prostate cancer on 626 web page)***

- ***it is possible to have survival curves with 3 categories (alive, dead of target condition, dead of something else)***
  again, see 3-ply curves in Albertsen Hanley et al JAMA Sept 1998

  same would apply to outcomes of starting a Ph.D.. e.g. at 5 years..
  xx% have obtained a Ph.D.
  yy% have decided it is not for them
  zz% are still pursuing it

- ***used (sometimes naively) to calculate 'lifetime probability of developing cancer or other condition (should ask: does the calculation allow for the possibility that one might die of another cause before one could develop the target condition?)***

(Non-Parametric / Semi-Parametric)

Estimation (point&interval) of S[t] , h[t] and pdf[t]

- Lifetable [fixed interval] - Kaplan-Meier [data-determined]
  [Bradford Hill or Armitage] [cf. Armitage]

Comparison of Survival Data/Curves

x-year (e.g. 5-year) survival (or cumulative mortality)

Use SE[$\hat{S}[5]_{\text{index-category}} - \hat{S}[5]_{\text{reference-category}}$]

SE for each determined by formula of
- Greenwood' (Armitage eqn 17.7 p 576)
- Kalbfleisch & Prentice (Armitage eqn 17.8 p 575)
- Peto (Armitage eqn 17.9 p 575)

entire curves

log-rank test [M-H ; one 2x2 table / distinct event-time]
- Armitage section 17.6 p 576)

note that it is a test
can be used to obtain 'relative death rates'
(cf Armitage p 578)]

Wilcoxon (Gehan) test; Peto test
- Kleinbaum Chapter 2

all of these tests have the log-rank format,
but weight the (a - E[a]) differences differently

log rank : gives equal weight to each failure time
Peto : gives more weight to ***early*** failure times

Example: Kaplan-Meier survival curves, log-rank test, and illustration of ***Risksets***

from Statistics at Square One: Survival analysis [http://bmj.bmjjournals.com/collections/statsbk/12.shtml]

"Mclllmurray and Turkie (2) describe a clinical trial of 69 patients for the treatment of Dukes' C colorectal cancer. The data for the two treatments, linoleic acid (tx = 1, n = 25) or control (tx = 0, n = 24) are given in Table 12.1 (3) .. "

	Follow-up Month	1	2	3	6	8	10	12	20	24	30	32	42	44	**Sum**
[a]	tx 1: deaths:	0	0	0	2	0	2	4	0	1	0	1	0	0	10
[b]	tx 1: survived:	25	24	24	21	21	18	13	9	7	5	4	1	1	
		--	--	--	--	--	--	--	--	--	--	--	--	--	
	tx 1: At Risk:	25	24	24	23	21	20	17	9	8	5	5	1	1	
[c]	tx 0: deaths:	0	0	0	4	2	0	2	1	1	1	0	1	0	12
[d]	tx 0: survived:	24	24	24	19	17	17	15	9	7	3	2	0	0	
		--	--	--	--	--	--	--	--	--	--	--	--	--	
	tx 0: At Risk:	24	24	24	23	19	17	17	10	8	4	2	1	0	
		==	==	==	==	==	==	==	==	==	==	==	==	==	
	tx 0&1: deaths:	0	0	0	6	2	2	6	1	2	1	1	1	0	
	tx 0&1: At Risk:	49	48	48	46	40	37	34	19	16	9	7	2	1	
	Riskset # :	.	.	.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	.	
E[a] ...	under H0:	.	.	.	3.0	1.1	1.1	3.0	0.5	1.0	0.6	0.7	0.5	.	11.4
V[a] ...	under H0:	.	.	.	1.3	0.5	0.5	1.3	0.2	0.5	0.2	0.2	0.3	.	5.0

1. Order all the survival times from smallest to largest; identify the distinct death-times; concentrate on those at risk just before each distinct death-time - this is the "**Risk-Set**' (i.e. the 'candidates") for the failure time
Subjects remain. in successive Risk Sets until removed by censoring, or event of interest

2. **Kaplan-Meier curve for each separate group:** Multiply the successive fractions who make it out of (past) each risk set to yield successively lower "estimated fractions still alive". [Skip risk set if no event in that group] eg tx 1: S[6] = (21/23); S[10] = S[6] × (18/20), etc..

3. **Log-Rank Test:** Form 2 x 2 table for the outcome in each risk set, and carry out Mantel-Haenszel test, summing the excesses or deficits (the values of {a – E[a | H_0] }) in the target (usually "a") cell over the tables. Compare the overall deficit/excess with its sampling variation

2 versions of test:- M-H 'focus only on "a"-cell' version, & $(O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2$.

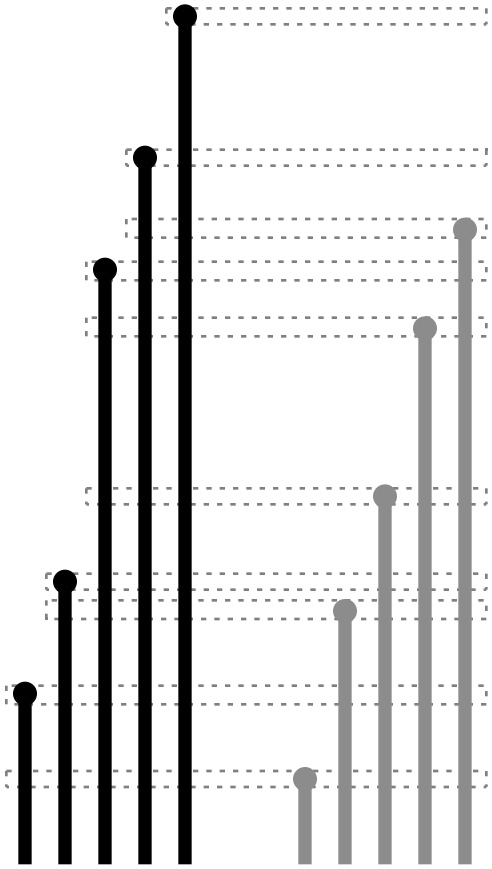
[fruitfly] **FIGURE LEGENDS**

Figure 1. Longevity of $n = 5$ sexually active male fruitflies (gray vertical lines) and $n = 5$ sexually inactive male fruitflies ((black vertical lines, reference group), together with the associated risksets, and Maximum Likelihood estimation of hazard ratio (HR) parameter in the (1-parameter) proportional hazards model which ignores thorax size. Circles denote age at death (longevity, survival time). In order to show all calculations clearly, the survival time axis is not perfectly to scale; the distortion is of no consequence, since the likelihood depends only on the ordering of the deaths. Risksets, one for each distinct event-time, are enclosed by dashed lines. The entries in the corresponding rows are the probabilities, calculated using the HR value in the column, that the death would occur to the subject who did die then, rather than in one of the other candidates in the riskset. The likelihood, for any HR value, is the product of the (column of) probabilities associated with the different risksets. The Maximum (log-)Likelihood occurs at HR = 2.4.

excerpts from draft of a longer article by Jh on proportional hazards model, and Maximum Likelihood estimation of parameters of the model

For now, use diagrams to understand the concept of Risk Sets and use of combining test information from separate strata

Fig 1



	Hazard Ratio (HR) 1	2.	4.	8.
	1/1	1/1	1/1	1/1
	1/2	1/2	1/2	1/2
	1/3	1/2	2/3	4/5
	1/4	1/5	1/7	1/11
	1/5	2/7	4/11	8/19
	1/6	2/9	4/15	8/27
	1/7	1/10	1/16	1/28
	1/8	1/6	1/5	2/9
	1/9	1/13	1/21	1/37
	1/10	2/15	4/25	8/45
Log[Product] , i.e., log Likelihood:	-15.1	-14.43	-14.64	-15.57

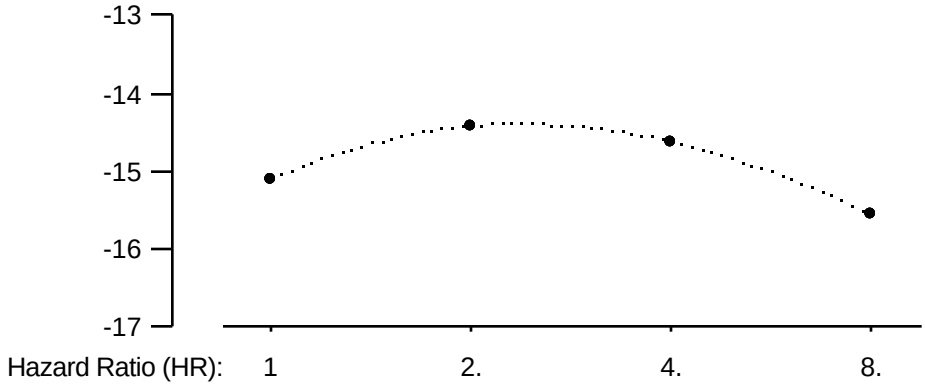
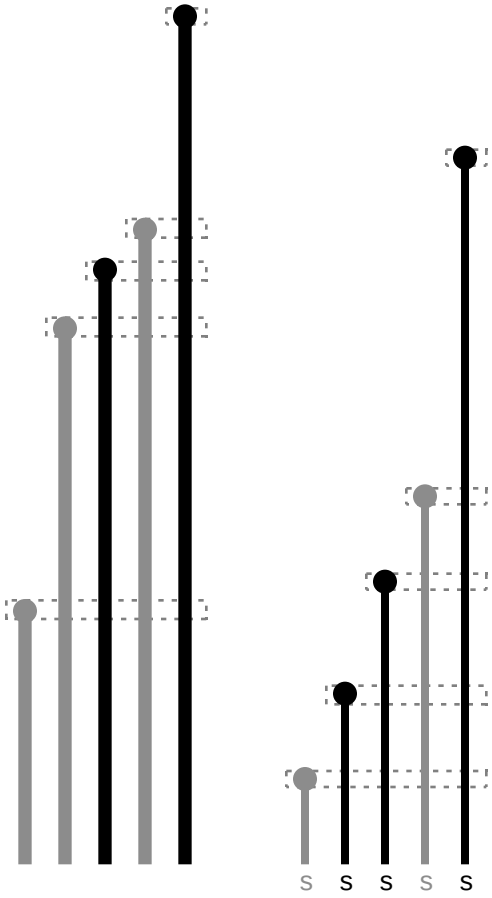


Figure 3. Maximum Likelihood estimation of a 1-parameter proportional hazards model using stratification/matching to eliminate confounding/variation produced by an extraneous variable. Vertical lines represent the longevity of $n = 5$ sexually active fruitflies (shaded line) and $n = 5$ sexually inactive male fruitflies (black, reference group). Three of the latter, and two of the former have shorter than average thorax lengths and are identified by the lowercase letter s and represented by thinner lines, while the remainder, with above average thorax lengths, are represented by thicker lines. Circles denote age at death. Subjects are first segregated (stratified) by thorax size, so that each riskset (enclosed by dashed lines) is homogeneous with respect to this variable. The entries in the corresponding rows are the probabilities, calculated using the HR value in the column, that the death would occur to the subject who did die, rather than in one of the other candidates in the riskset. The likelihood, for any HR value, is the product of the (column of) probabilities associated with the different risksets. The Maximum Likelihood occurs at HR = 2.3. The different log-likelihood scale, compared with Figure 2, stems from the fact that each riskset is smaller, so that the associated probability is larger, and the log-probability is less negative. For this reason, the log-likelihood based on these stratified series cannot be compared with the log-likelihood from the 2-parameter model.

Figure 3



	Hazard Ratio (HR) 1	2.	4.	8.
	1/1	1/1	1/1	1/1
	1/1	1/1	1/1	1/1
	1/2	2/3	4/5	8/9
	1/3	1/4	1/6	1/10
	1/4	1/3	2/5	4/9
	1/2	2/3	4/5	8/9
	1/3	1/4	1/6	1/10
	1/5	1/4	2/7	4/13
	1/4	1/5	1/7	1/11
	1/5	2/7	4/11	8/19
Log[Product] , i.e., log Likelihood:	-9.57	-8.93	-9.16	-10.09