# Survival analysis

Paul W. Dickman
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
paul.dickman@mep.ki.se

Summer School on Modern Methods in Biostatistics and Epidemiology
Cison di Valmarino, Treviso, Italy
20–25 June, 2005

http://www.bioepi.org/



## Overview

- Central concepts is survival analysis: censoring, truncation, survivor function, hazard function.
- Estimating survival (or cumulative incidence) using the actuarial and Kaplan-Meier methods.
- Testing for differences in survival between groups using the log-rank test.
- Estimating rates and modelling them using Poisson regression.
- Parametric survival models (very limited coverage).
- Cox proportional hazards model.
- Diagnostics for the Cox model (time permitting).
- Comparison of the Cox and Poisson regression models.



## Teaching format

- Generally lectures in the morning followed by exercises in the afternoon with a greater emphasis on lectures during the early part of the course.
- I have constructed exercises and provided solutions to most exercises. I will suggest appropriate exercises for each afternoon but you are welcome to diverge from the suggested exercises if you wish.
- Course participants have a wide range of backgrounds and diverse interests. It is hoped that the exercise questions will provide time for you to study or ask questions about topics of special interest.
- The lecture notes contain topics which I will only cover briefly or not at all. Feel free to ask about these topics in the exercise sessions if you're interested.



## Key concepts

- Special methods (i.e., survival analysis) are required when the outcome of interest has a time dimension.
- The outcome can be presented as a survival proportion or an event rate. The two measures are mathematically related.
- Epidemiological cohort studies can (and should) be analysed in the framework of survival analysis. 'Time' may be a confounder or effect modifier.
- Cox regression and Poisson regression are very similar.
- Reinforce key concepts in statistical modelling of epidemiological data
  - Studying confounding and effect modification in a modelling framework
  - Reparameterising a statistical model to estimate interaction effects



## Textbook

- I haven't selected a textbook since, irrespective of which one I choose, less than half of you will like it – not a good situation given the cost of textbooks.
- Many textbooks on survival analysis are available and I suggest you look through a few until you find one you like.
- Very few books are targeted at epidemiologists (e.g., you won't find Poisson regression mentioned in many books).
- The definitive text for epidemiologists is 'Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies' by Breslow and Day [1] although it is rather advanced.
- Many general biostatistics textbooks (e.g., Rosner [2]) contain a chapter on survival analysis.



- 'An Introduction to Survival Analysis Using Stata' [3] is recommended for Stata users. Many parts, however, assume a solid grasp of mathematical statistics.
- The SAS 'books by user' [4, 5] are recommended for SAS users.



## Exam

- Held on the final day.
- Approximately 30 minutes.
- You will be asked to interpret some Stata output (Cox regression) and asked questions such as
  1. Interpret the estimated hazard ratio for the effect of exposure, including a comment on statistical significance.
  2. Is there evidence of confounding?
  3. Is there evidence of effect modification?



## Computing

- We will be using Stata version 9.
- All examples in the lecture notes and all exercises were prepared using Stata version 8. I have confirmed that the code works in version 9 although I expect we will discover new features during the course.
- Please install bioepi from within Stata, typing at the command line:

```
. net install bioepi, from(http://bioepi.altervista.org/stata)
. help bioepi
```

- All Stata datasets can be accessed directly from the web

```
. use http://www.bioepi.org/teaching/sa/colon.dta
```

## For SAS users

- I am actually more familiar with SAS and have prepared additional notes (available on the web) on how the methods described in the course can be implemented using SAS.
- All data sets used in the course are available on the web in SAS format along with SAS code for completing some of the exercises.

  `http://www.meb.ki.se/biostat/courses/2005/biostat_III/sas/`
- Although I'm more familiar with SAS, I much prefer Stata for analysis of epidemiologic data.
- I also wrote some notes on Cox regression in SAS version 9.

  `http://www.pauldickman.com/teaching/sas/phreg/`



## Analysis of Time-to-Event Data

- Time-to-event data are generated when the response measurement of interest is the time from a well-defined origin of measurement to occurrence of an event of interest.
- Three basic requirements define time-to-event measurements

  a. agreed scale of measurement for time (e.g. time since diagnosis, attained age)
  b. unambiguous origin for the measurement of 'time'
  c. precise definition of 'response,' or occurrence of the event of interest
- Time-to-event analysis is also known as failure time analysis (primarily in engineering), lifetime analysis, and survival analysis.



## Examples of time-to-event measurements

- Time from diagnosis of cancer to death due to the cancer
- Time from diagnosis of cancer to death due to any causes
- Time from diagnosis of localised cancer to metastases
- Time from randomisation to death in a cancer clinical trial
- Time from randomisation to recurrence in a cancer clinical trial
- Time from remission to relapse of leukemia
- Time to re-offending after being released from jail
- Time between two attempts to donate a unit of blood for transfusion purposes
- Time from HIV infection to AIDS
- Time to the first goal (or next goal) in a hockey game
- Time from exposure to cancer incidence in an epidemiological cohort study

- Epidemiological cohort studies are time-to-event studies and are analysed in the framework of survival analysis.
- Examples of time-to-event data can be found in almost every discipline.



- In some studies, the event of interest (e.g. death) is bound to occur if we are able to follow-up each individual for a sufficient length of time.
- However, whether or not the event of interest is inevitable has no consequence for the design, analysis, or interpretation of the study.
- In some studies the time-to-event (or survival probability) is of primary interest whereas in epidemiological cohort studies we may be primarily interested in comparing the event rates between the exposed and unexposed.
- The basic statistical methodology is similar for randomised and observational studies, although some methods are more appropriate for some designs than others (e.g. need to control for confounding in observational studies).
- The characteristic of time-to-event data that renders standard statistical methods inappropriate is *censoring* — unobserved values of the response measurement of interest.



## Censoring

- Refers to the situation where individuals cannot be observed for the full time to event.
- In studying the survival of cancer patients, for example, patients enter the study at the time of diagnosis (or the time of treatment in randomised trials) and are followed up until the event of interest is observed. Censoring may occur in one of the following forms:
  - Termination of the study before the event occurs;
  - Death due to a cause not considered to be the event of interest (in cause-specific survival analyses); and
  - Loss to follow-up, for example, if the patient emigrates.
- We say that the survival time is censored.
- These are examples of right censoring, which is the most common form of censoring in medical studies.



- With right censoring, we know that the event has not occurred during follow-up, but we are unable to follow-up the patient further. We know only that the true survival time of the patient is greater than a given value.
- More details of censoring and a related issue, truncation, are given on slide 73.



## A sample of 35 patients diagnosed with colon carcinoma in Finland during 1985–94; followed-up until the end of 1995

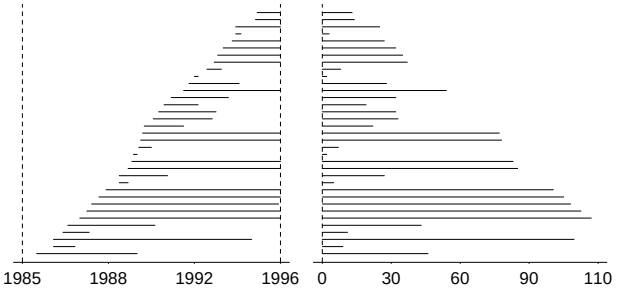



Figure 1: Real time (left) and time from entry in months (right)



| ID | Sex | Age at dx | Clinical stage | dx date mmyy | Surv. time mm | Surv. time yy | Status |
|---|---|---|---|---|---|---|---|
| 1 | male | 72 | Localised | 2.89 | 2 | 0 | Dead - other |
| 2 | female | 82 | Distant | 12.91 | 2 | 0 | Dead - cancer |
| 3 | male | 73 | Distant | 11.93 | 3 | 0 | Dead - cancer |
| 4 | male | 63 | Distant | 6.88 | 5 | 0 | Dead - cancer |
| 5 | male | 67 | Localised | 5.89 | 7 | 0 | Dead - cancer |
| 6 | male | 74 | Regional | 7.92 | 8 | 0 | Dead - cancer |
| 7 | female | 56 | Distant | 1.86 | 9 | 0 | Dead - cancer |
| 8 | female | 52 | Distant | 5.86 | 11 | 0 | Dead - cancer |
| 9 | male | 64 | Localised | 11.94 | 13 | 1 | Alive |
| 10 | female | 70 | Localised | 10.94 | 14 | 1 | Alive |
| 11 | female | 83 | Localised | 7.90 | 19 | 1 | Dead - other |
| 12 | male | 64 | Distant | 8.89 | 22 | 1 | Dead - cancer |
| 13 | female | 79 | Localised | 11.93 | 25 | 2 | Alive |
| 14 | female | 70 | Distant | 6.88 | 27 | 2 | Dead - cancer |
| 15 | male | 70 | Regional | 9.93 | 27 | 2 | Alive |
| 16 | female | 68 | Distant | 9.91 | 28 | 2 | Dead - cancer |
| 17 | male | 58 | Localised | 11.90 | 32 | 2 | Dead - cancer |
| 18 | male | 54 | Distant | 4.90 | 32 | 2 | Dead - cancer |
| 19 | female | 86 | Localised | 4.93 | 32 | 2 | Alive |
| 20 | male | 31 | Localised | 1.90 | 33 | 2 | Dead - cancer |
| 21 | female | 75 | Localised | 1.93 | 35 | 2 | Alive |
| 22 | female | 85 | Localised | 11.92 | 37 | 3 | Alive |
| 23 | female | 68 | Distant | 7.86 | 43 | 3 | Dead - cancer |
| 24 | male | 54 | Regional | 6.85 | 46 | 3 | Dead - cancer |
| 25 | male | 80 | Localised | 6.91 | 54 | 4 | Alive |
| 26 | female | 52 | Localised | 7.89 | 77 | 6 | Alive |
| 27 | male | 52 | Localised | 6.89 | 78 | 6 | Alive |
| 28 | male | 65 | Localised | 1.89 | 83 | 6 | Alive |
| 29 | male | 60 | Localised | 11.88 | 85 | 7 | Alive |
| 30 | female | 71 | Localised | 11.87 | 97 | 8 | Alive |
| 31 | male | 58 | Localised | 8.87 | 100 | 8 | Alive |
| 32 | female | 80 | Localised | 5.87 | 102 | 8 | Dead - cancer |
| 33 | male | 66 | Localised | 1.86 | 103 | 8 | Dead - other |
| 34 | male | 67 | Localised | 3.87 | 105 | 8 | Alive |
| 35 | female | 56 | Distant | 12.86 | 108 | 9 | Alive |

## Sample data sets

- The following data sets will be used during the course:

  **colon** colon carcinoma diagnosed in Finland during 1975–1994 with follow-up to 31 December 1995.
  **melanoma** skin melanoma diagnosed in Finland during 1975–1994 with follow-up to 31 December 1995.
  **colon_sample** a random sample of 35 patients from the colon data.
  **diet** data from a pilot study evaluating the use of a weighed diet over 7 days in epidemiological studies. The primary hypothesis is the relation between dietary energy intake and incidence of coronary heart disease (CHD).

- The 3 cancer data sets have been kindly provided by the Finnish Cancer Registry.

- The diet data are analysed extensively by David Clayton and Michael Hills in their textbook [6]. These data are also used in examples in the Stata manual (for example, `stplit`, `strate`, and `stptime`).



## Variables in the colon carcinoma data set

```
. use http://www.bioepi.org/teaching/sa/colon
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. describe
 obs:  15,564   vars:     11
-------------------------------------------------------------------------------
 1. sex       byte   %8.0g    sex        Sex
 2. age       float  %9.0g               Age at diagnosis
 3. stage     byte   %9.0g    stage      Clinical stage at diagnosis
 4. mmdx      float  %9.0g               Month of diagnosis
 5. yydx      float  %9.0g               Year of diagnosis
 6. surv_mm   float  %9.0g               Survival time in completed
                                          months
 7. surv_yy   float  %9.0g               Survival time in completed years
 8. status    byte   %17.0g   status     Vital status at last date of
                                          contact
 9. subsite   byte   %22.0g   colonsub   Anatomical subsite of tumour
10. year8594  float  %15.0g   year8594   Year of diagnosis 1985-94
11. agegrp    float  %9.0g    agegrp     Age in 4 categories
-------------------------------------------------------------------------------
```



## Variables in the skin melanoma data set

```
. use melanoma, clear
(Skin melanoma, all stages, Finland 1975-94, follow-up to 1995)
. describe
  obs: 7,775    vars: 12
-------------------------------------------------------------------------------
variable    type   format  label      variable label
-------------------------------------------------------------------------------
sex         byte   %8.0g   sex        Sex
age         float  %9.0g              Age at diagnosis
stage       byte   %9.0g   stage      Clinical stage at diagnosis
mmdx        float  %9.0g              Month of diagnosis
yydx        float  %9.0g              Year of diagnosis
surv_mm     float  %9.0g              Survival time in completed months
surv_yy     float  %9.0g              Survival time in completed years
status      byte   %17.0g  status     Vital status at last contact
subsite     byte   %16.0g  melsub     Anatomical subsite of tumour
year8594    float  %15.0g  year8594   Year of diagnosis 1985-94
agegrp      float  %9.0g   agegrp     Age in 4 categories
osr_fail    float  %9.0g              Indicator for death due to any cause
-------------------------------------------------------------------------------
```



## Variables in the diet data set

```
. describe
  obs: 337       vars: 12
-------------------------------------------------------------------------------
              storage  display   value
variable      type   format    label      variable label
-------------------------------------------------------------------------------
id            int    %9.0g                Subject identity number
chd           byte   %9.0g                Failure: 1=chd, 0 otherwise
y             float  %9.0g                Time in study (years)
hieng         float  %12.0g    hieng     * Indicator for high energy
energy        float  %9.0g                Total energy (kcals per day)
job           byte   %9.0g     job       * Occupation
month         byte   %8.0g                Month of survey
height        float  %9.0g                Height (cm)
weight        float  %9.0g                Weight (kg)
doe           int    %dDmCY               Date of entry
dox           int    %dDmCY               Date of exit
dob           int    %dDmCY               Date of birth
                                         * indicated variables have notes
-------------------------------------------------------------------------------
```



## Skin melanoma 1985–94

Table 1: Codes for vital status with corresponding frequency counts 1985–94

| Code and description | Male | Female |
|---|---|---|
| 0 Alive | 1554 | 1786 |
| 1 Dead: melanoma was the cause | 543 | 376 |
| 2 Dead: other cause of death | 238 | 247 |
| 4 Lost to follow-up | 0 | 0 |
| Total | 2335 | 2409 |

Table 2: Codes for subsite with corresponding frequency counts 1985–94

| Code and description (ICD-7) | Male | Female |
|---|---|---|
| 1 Head and neck (190.0–190.4) | 332 | 375 |
| 2 Trunk (190.5) | 1276 | 678 |
| 3 Limbs (190.6–190.7) | 561 | 1231 |
| 4 Multiple and NOS | 166 | 125 |
| Total | 2335 | 2409 |

Note that the sample data sets also include patients diagnosed 1975–1984.



## Colon carcinoma 1985–94

Table 3: Codes for vital status with corresponding frequency counts 1985–94

| Code and description | Male | Female |
|---|---|---|
| 0 Alive | 1476 | 2081 |
| 1 Dead: due to colon carcinoma | 1806 | 2618 |
| 2 Dead: other cause of death | 519 | 586 |
| 4 Lost to follow-up | 1 | 0 |
| Total | 3802 | 5285 |

Table 4: Codes for subsite with corresponding frequency counts 1985–94

| Code and description (ICD-7) | Male | Female |
|---|---|---|
| 1 Coecum and ascending (153.0) | 1289 | 1951 |
| 2 Transverse (153.1) | 707 | 901 |
| 3 Descending and sigmoid (153.2,153.3) | 1566 | 2095 |
| 4 Other and NOS | 240 | 338 |
| Total | 3802 | 5285 |

Note that the sample data sets also include patients diagnosed 1975–1984.



## Terminology

- In the strictest sense, a *ratio* is the result of dividing one quantity by another. In the sciences, however, it is mostly used in a more specific sense, that is, when the numerator and the denominator are two separate and distinct quantities [7].

- A *proportion* is a type of ratio in which the numerator is included in the denominator, e.g. the incidence proportion (aka cumulative incidence).

- A *rate* is a measure of change in one quantity per unit of another quantity. In epidemiology, rates typically have units events per unit time.

- The 'survival rate' of a group of patients over a specified time period is therefore not strictly a rate, but a proportion.

- We will be estimating both proportion (e.g., survival proportions) and rates (e.g., mortality rates) and should recognise that these are conceptually different.



## The survivor function

- The survivor function, $S(t)$, gives the probability of surviving until at least time $t$.

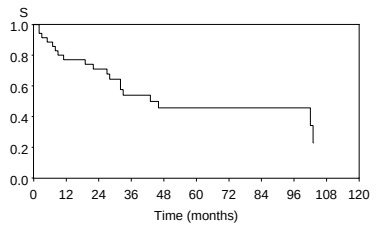



Figure 2: Estimates of $S(t)$ for the 35 patients diagnosed with colon carcinoma. All deaths are considered events ($S(t)$ is called the observed survivor function).

- $S(t)$ is a nonincreasing function with a value 1 at the time origin and a value 0 as $t$ approaches infinity.
- Note that $S(t)$ is a function (the survivor function) which depends on $t$ and should not be referred to as the survival rate.
- The survivor function evaluated at a specific value of $t$ is often referred to as the 'survival rate', for example, the '5-year survival rate'.
- We prefer to use the term 'survival proportion', for example, the '5-year survival proportion'.
- For example, the 5-year survival proportion for the data presented in Figure 2 is 45%.
- Nonparametric methods for estimating $S(t)$ (described later) generally involve estimating the survival proportion at discrete values of $t$ and then interpolating these to obtain an estimate of $S(t)$.



## Interpreting $S(t)$ and comparing estimates of $S(t)$ between groups

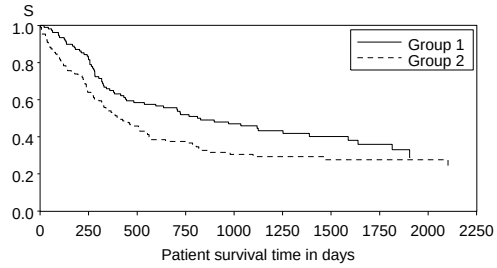



Figure 3: Estimated survivor function (S) for two groups of patients



- It is clear that the individuals in group 1 experience superior survival compared to individuals in group 2.
- It is, however, difficult to determine the essence of the failure pattern, and even more difficult to compare it between groups, simply by studying plots of the survivor function.
- The rate of decline of the survivor function is a measure of the risk of experiencing the event at time $t$ (the instantaneous mortality rate at time $t$).
- In survival analysis, this is called the hazard function, $\lambda(t)$[1].
- The hazard function is described in more detail on slide 161.
- Patients in group 1 have superior survival for the interval up to 850 days following diagnosis but then have worse survival than group 2 after 850 days.

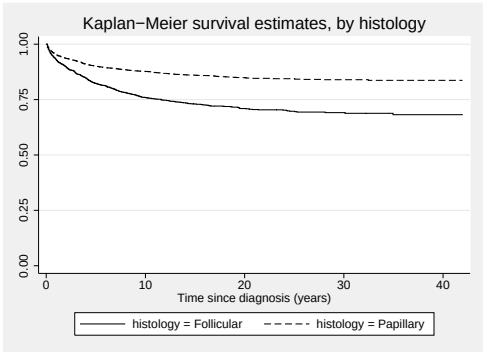
[1] strictly, the hazard is the derivative of the negative logarithm of the survivor function, $\lambda(t) = d/dt H(t) = d/dt - \ln[S(t)]$



- This is an example of non-proportional hazards, a concept we will return to later.
- The survival experience of a cohort can be expressed in terms of the survival proportion or the hazard rate.
- In epidemiological cohort studies where incidence is the outcome, we often present the cumulative incidence, given by $1 - S(t)$, rather than $S(t)$.
- We may then model the hazard function (the incidence rate) and estimate the incidence rate ratio (hazard ratio) for the exposed compared to the unexposed.
- Often it is the hazard ratio, rather than the survivor function, which is of primary interest.

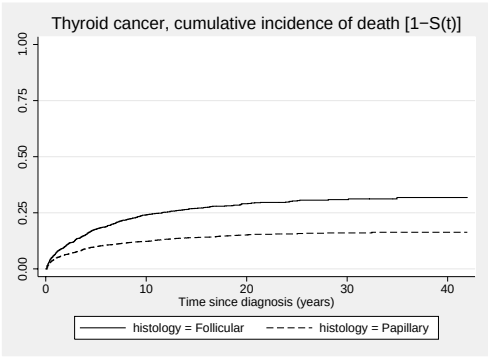


## Survival of patients with differentiated thyroid cancer






```
sts graph, by(histology) failure
```



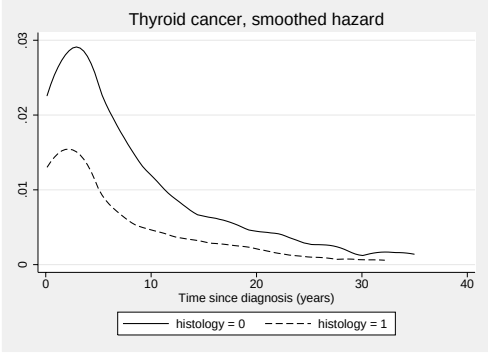



## An introduction to the hazard function, $\lambda(t)$

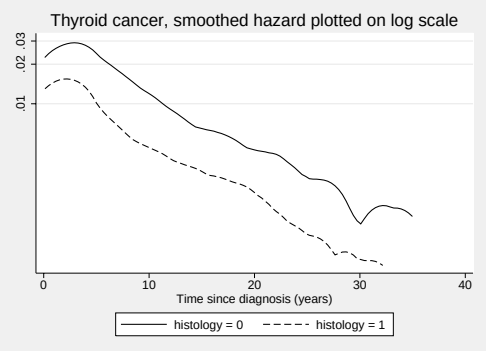
- The term 'hazard rate' is the generic term used in survival analysis to describe the 'event rate'. If, for example, the event of interest is disease incidence then the hazard represents the incidence rate.
- The hazard function, $\lambda(t)$, is the instantaneous event rate at time $t$, conditional on survival up to time $t$. The units are events per unit time.
- In contrast to the survivor function, which describes the probability of *not* failing before time $t$, the hazard function focuses on the failure rate at time $t$ among those individuals who are alive at time $t$.
- That is, a lower value for $\lambda(t)$ implies a higher value for $S(t)$ and vice-versa.
- Note that the hazard is a rate, not a probability, so $\lambda(t)$ can take on any value between zero and infinity, as opposed to $S(t)$ which is restricted to the interval $[0, 1]$. More details on slide 161.



```
sts graph, by(histology) hazard
```



```
sts graph, by(histology) hazard yscale(log)
```





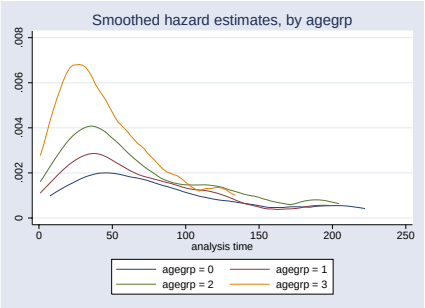




Figure 4: Localised skin melanoma. Plot of the estimated hazard function for each age group.



## A subset of the 35 patients diagnosed with colon carcinoma

| ID | Sex | Age at dx | Clinical stage | dx date mmyy | Surv. time mm | Surv. time yy | Status |
|---|---|---|---|---|---|---|---|
| 1 | male | 72 | Localised | 2.89 | 2 | 0 | Dead - other |
| 2 | female | 82 | Distant | 12.91 | 2 | 0 | Dead - cancer |
| 3 | male | 73 | Distant | 11.93 | 3 | 0 | Dead - cancer |
| 4 | male | 63 | Distant | 6.88 | 5 | 0 | Dead - cancer |
| 5 | male | 67 | Localised | 5.89 | 7 | 0 | Dead - cancer |
| 6 | male | 74 | Regional | 7.92 | 8 | 0 | Dead - cancer |
| 7 | female | 56 | Distant | 1.86 | 9 | 0 | Dead - cancer |
| 8 | female | 52 | Distant | 5.86 | 11 | 0 | Dead - cancer |
| 9 | male | 64 | Localised | 11.94 | 13 | 1 | Alive |
| 10 | female | 70 | Localised | 10.94 | 14 | 1 | Alive |
| 11 | female | 83 | Localised | 7.90 | 19 | 1 | Dead - other |
| 12 | male | 64 | Distant | 8.89 | 22 | 1 | Dead - cancer |
| 13 | female | 79 | Localised | 11.93 | 25 | 2 | Alive |
| 14 | female | 70 | Distant | 6.88 | 27 | 2 | Dead - cancer |
| 15 | male | 70 | Regional | 9.93 | 27 | 2 | Alive |
| 16 | female | 68 | Distant | 9.91 | 28 | 2 | Dead - cancer |



## Estimating the survivor function, $S(t)$

- Consider the sample data for the 35 colon cancer patients introduced on slide 15 and reproduced (in part) on the previous slide.
- We may be interested in estimating $S(t)$ where the event of interest is death due to any cause.
- An estimate of $S(t)$ could be obtained by simply calculating the proportion of individuals still alive at selected values of $t$, such as completed years.
- Eight of the 35 patients died during the first year of follow-up so the estimate for $S(1)$ is $\hat{S}(1) = (35-8)/35 = 27/35 = 0.771$.
- We encounter problems when attempting to estimate $S(2)$. Ten patients died within two years of follow-up, but 2 patients (patients 9 and 10) could not be followed-up for a full 2 years.



- We could exclude these two patients from the analysis and let $\hat{S}(2) = (33-10)/33$, but this will underestimate the true survival proportion since it ignores the fact that each of these two patients were at risk of death for between one and two years but did not die while under observation.
- If we instead use $\hat{S}(2) = (35-10)/35$ then we will overestimate the true survival proportion, since we are assuming that each of these two patients survived for a full two years.
- Two common (and similar) methods for estimating $S(t)$ in the presence of censoring are the lifetable (actuarial) method and the Kaplan-Meier (product-limit) method.



## Life table method for estimating $S(t)$

- Also known as the actuarial method. The approach is to divide the period of observation into a series of time intervals and estimate the conditional (interval-specific) survival proportion for each interval.
- The cumulative survivor function, $S(t)$, at the end of a specified interval is then given by the product of the interval-specific survival proportions for all intervals up to and including the specified interval.
- In the absence of censoring, the interval-specific survival proportion is $p = (l-d)/l$, where $d$ is the number of events (deaths) observed during the interval and $l$ is the number of patients alive at the start of the interval.
- In the presence of censoring, it is assumed that censoring occurs uniformly throughout the interval such that each individual with a censored survival time is at risk for, on average, half of the interval. This assumption is known as the actuarial assumption.



- The effective number of patients at risk during the interval is given by $l' = l - \frac{1}{2}w$ where $l$ is the number of patients alive at the start of the interval and $w$ is the number of censorings during the interval.
- The estimated interval-specific survival proportion is then given by $p = (l'-d)/l'$.
- For the first interval, $l = l' = 35$ and $p = (35-8)/35 = 0.771$. The estimated 1-year survival proportion is therefore $\hat{S}(1) = 0.771$.
- For the second interval, $l' = 27 - \frac{1}{2} \times 2 = 26$ and $p = (26-2)/26 = 0.923$.
- The estimated 2-year survival proportion is then $\hat{S}(2) = 0.771 \times 0.923 = 0.71209$.
- The cumulative survival estimated is estimated as the product of conditional survival proportions, where the estimate of each conditional survival proportion is based upon only those individuals under follow-up.



- That is, the individuals who are censored are assumed to have the same prognosis as those individuals who could be followed up.
- This requires the assumption that censoring is *non-informative*.
- That is, we make the assumption that, conditional on the values of any explanatory variables, censoring is unrelated to prognosis (the probable course and outcome of the disease).
- Informative censoring is discussed further on slide 68.
- In the first exercise you will construct (by hand) a life table on these same data but with death due to cancer as the outcome.

Table 5: Life table with annual interval for the 35 patients.

| time | $l$ | $d$ | $w$ | $l'$ | $p$ | $S(t)$ |
|---|---|---|---|---|---|---|
| [0-1) | 35 | 8 | 0 | 35.0 | 0.77143 | 0.77143 |
| [1-2) | 27 | 2 | 2 | 26.0 | 0.92308 | 0.71209 |
| [2-3) | 23 | 5 | 4 | 21.0 | 0.76190 | 0.54254 |
| [3-4) | 14 | 2 | 1 | 13.5 | 0.85185 | 0.46217 |
| [4-5) | 11 | 0 | 1 | 10.5 | 1.00000 | 0.46217 |
| [5-6) | 10 | 0 | 0 | 10.0 | 1.00000 | 0.46217 |
| [6-7) | 10 | 0 | 3 | 8.5 | 1.00000 | 0.46217 |
| [7-8) | 7 | 0 | 1 | 6.5 | 1.00000 | 0.46217 |
| [8-9) | 6 | 2 | 3 | 4.5 | 0.55556 | 0.25676 |
| [9-10) | 1 | 0 | 1 | 0.5 | 1.00000 | 0.25676 |

- $l$ is the number alive at the start of the interval
- $d$ is the number of events (deaths) during the interval
- $w$ is the number of censorings (withdrawals) during the interval
- $l'$ is the effective number at risk for the interval
- $p$ is the interval-specific survival proportion
- $S(t)$ is the estimated cumulative survivor function at the end of the interval



## Survival analysis using Stata

- In order to analyse survival data it is necessary to specify (at a minimum) a variable representing survival time and a variable specifying whether or not the event of interest was observed (called the failure variable).
- Instead of specifying a variable representing survival time we can specify the entry and exit dates (this is necessary if subjects enter the study at different times).
- In many statistical software programs (such as SAS), these variables must be specified every time a new analysis is performed.
- In Stata, these variables are specified once using the `stset` command and then used for all subsequent survival analysis (`st`) commands (until the next `stset` command).



- For example

```
. use melanoma
. stset surv_mm, failure(status==1)
```

- The above code shows how we would `stset` the skin melanoma data in order to analyse cause-specific survival with survival time in completed months (`surv_mm`) as the time variable.
- Of the four possible values of `status`, we have specified that only code 1 indicates an event (death due to melanoma).
- If we wanted to analyse observed survival (where all deaths are considered to be events) we could use the following command

```
. stset surv_mm, failure(status==1,2)
```



- Some of the Stata survival analysis (`st`) commands relevant to this course are given below. Further details can be found in the manuals or online help.

```
stset       Declare data to be survival-time data
stsplit     Split time-span records
stdes       Describe survival-time data
stsum       Summarize survival-time data
sts         Generate, graph, list, and test the survivor
              and cumulative hazard functions
strate      Tabulate failure rate
stptime     Calculate person-time at risk and failure rates
stcox       Estimate Cox proportional hazards model
stphtest    Test of Cox proportional hazards assumption
stphplot    Graphical assessment of the Cox
              proportional hazards assumption
stcoxkm     Graphical assessment of the Cox
              proportional hazards assumption
streg       Estimate parametric survival models
```



- Once the data have been `stset` we can use any of these commands without having to specify the survival time or failure time variables.
- For example, to plot Kaplan-Meier estimates of the cause-specific survivor function by sex and then fit a Cox proportional hazards model with sex and calendar period as covariates

```
. sts graph, by(sex)
. stcox sex year8594
```

- The `ltable` command (life table estimation) is not an `st` command – the data do not have to be `stset` to use this command but we need to specify the survival time and failure variable every time we use the command.



## Life table estimates in Stata – the ltable command

```
. **- Create a failure time variable to use with the ltable command -**
. generate fail=0
. replace fail=1 if status==1 | status==2
. ltable surv_yy fail
```

| Interval | Beg. Total | Deaths | Lost | Survival | Std. Error | [95% Conf. | Int.] |
|---|---|---|---|---|---|---|---|
| 0 1 | 35 | 8 | 0 | 0.7714 | 0.0710 | 0.5946 | 0.8785 |
| 1 2 | 27 | 2 | 2 | 0.7121 | 0.0769 | 0.5307 | 0.8336 |
| 2 3 | 23 | 5 | 4 | 0.5425 | 0.0884 | 0.3567 | 0.6958 |
| 3 4 | 14 | 2 | 1 | 0.4622 | 0.0918 | 0.2786 | 0.6274 |
| 4 5 | 11 | 0 | 1 | 0.4622 | 0.0918 | 0.2786 | 0.6274 |
| 6 7 | 10 | 0 | 3 | 0.4622 | 0.0918 | 0.2786 | 0.6274 |
| 7 8 | 7 | 0 | 1 | 0.4622 | 0.0918 | 0.2786 | 0.6274 |
| 8 9 | 6 | 2 | 3 | 0.2568 | 0.1197 | 0.0698 | 0.4994 |
| 9 10 | 1 | 0 | 1 | 0.2568 | 0.1197 | 0.0698 | 0.4994 |



## Life table for 6274 patients with localised colon carcinoma

```
. use http://www.bioepi.org/teaching/sa/colon
. gen csr_fail=(status==1)
. ltable surv_yy csr_fail if stage==1
```

| Interval | Beg. Total | Deaths | Lost | Survival | Std. Error | [95% Conf. | Int.] |
|---|---|---|---|---|---|---|---|
| 0 1 | 6274 | 550 | 202 | 0.9109 | 0.0036 | 0.9035 | 0.9177 |
| 1 2 | 5522 | 392 | 519 | 0.8430 | 0.0047 | 0.8336 | 0.8520 |
| 2 3 | 4611 | 253 | 460 | 0.7944 | 0.0053 | 0.7837 | 0.8046 |
| 3 4 | 3898 | 175 | 449 | 0.7565 | 0.0058 | 0.7449 | 0.7677 |
| 4 5 | 3274 | 123 | 363 | 0.7264 | 0.0062 | 0.7141 | 0.7383 |
| 5 6 | 2788 | 89 | 364 | 0.7016 | 0.0065 | 0.6887 | 0.7141 |
| 6 7 | 2335 | 53 | 326 | 0.6845 | 0.0067 | 0.6711 | 0.6975 |
| 7 8 | 1956 | 23 | 284 | 0.6758 | 0.0069 | 0.6621 | 0.6891 |
| 8 9 | 1649 | 16 | 250 | 0.6687 | 0.0071 | 0.6547 | 0.6823 |
| 9 10 | 1383 | 13 | 216 | 0.6619 | 0.0072 | 0.6475 | 0.6759 |



## The actuarial assumption

- The survival proportion, despite commonly being called the survival rate, is a proportion.
- For example, the interval-specific survival proportion has the form

$$\frac{\text{number surviving the interval}}{\text{number alive at the start of the interval}}$$

- When using life table methods, instead of the 'number alive at the start of the interval' we subtract $\frac{1}{2}w$ to obtain the 'effective number at risk'.
- This is similar to what we might do if we were estimating person-time at risk and invites the question of why we don't also subtract $\frac{1}{2}d$ since those who die are not at risk for the entire interval.

- If we were estimating a rate (events/person-time) then we would do this.
- We are not, however, estimating a rate.
- We are estimating a proportion and applying a correction (to both the numerator and the denominator) to account for censoring.
- It is possible to estimate the survivor function $S(t)$ by first estimating the cumulative hazard function and then transforming the estimate to get $S(t)$.



## The Kaplan-Meier method for estimating $S(t)$

- Also known as the product-limit method but is more commonly known as the Kaplan-Meier method, after the two researchers who first published the method in English in 1958 [8].
- The method was published much earlier (1912) in German [8, 9].
- In essence, the Kaplan-Meier method is the life table method where the interval size is decreased towards zero so that the number of intervals tends to infinity. Each life table interval is of infinitesimal length, just enough for one event or time increment.
- In practice, survival time is measured on a discrete scale (e.g. minutes, hours, days, months, or years) so the interval length is limited by the accuracy to which survival time is measured.



| ID | Sex | Age at dx | Clinical stage | dx date mmyy | Surv. time mm | yy | Status |
|---|---|---|---|---|---|---|---|
| 1 | male | 72 | Localised | 2.89 | 2 | 0 | Dead - other |
| 2 | female | 82 | Distant | 12.91 | 2 | 0 | Dead - cancer |
| 3 | male | 73 | Distant | 11.93 | 3 | 0 | Dead - cancer |
| 4 | male | 63 | Distant | 6.88 | 5 | 0 | Dead - cancer |
| 5 | male | 67 | Localised | 5.89 | 7 | 0 | Dead - cancer |
| 6 | male | 74 | Regional | 7.92 | 8 | 0 | Dead - cancer |
| 7 | female | 56 | Distant | 1.86 | 9 | 0 | Dead - cancer |
| 8 | female | 52 | Distant | 5.86 | 11 | 0 | Dead - cancer |
| 9 | male | 64 | Localised | 11.94 | 13 | 1 | Alive |
| 10 | female | 70 | Localised | 10.94 | 14 | 1 | Alive |
| 11 | female | 83 | Localised | 7.90 | 19 | 1 | Dead - other |
| 12 | male | 64 | Distant | 8.89 | 22 | 1 | Dead - cancer |
| 13 | female | 79 | Localised | 11.93 | 25 | 2 | Alive |
| 14 | female | 70 | Distant | 6.88 | 27 | 2 | Dead - cancer |
| 15 | male | 70 | Regional | 9.93 | 27 | 2 | Alive |
| 16 | female | 68 | Distant | 9.91 | 28 | 2 | Dead - cancer |



- The Kaplan-Meier method was developed for applications where survival time is measured on a continuous scale.
- We should therefore use as accurate a time scale as possible. That is, don't base the estimate on time in days if time in minutes is also known.
- In practice, only those intervals containing an event contribute to the estimate, so we can ignore all other intervals.
- To obtain Kaplan-Meier estimates of survival, the patient survival times are first ranked in increasing order.
- The times where events (deaths) occur are denoted by $t_i$, where $t_1 < t_2 < t_3 < \ldots$.
- The number of deaths occurring at $t_i$ is denoted by $d_i$.
- If both censoring(s) and death(s) occur at the same time, then the censoring(s) are assumed to occur immediately after the death time.



- That is, individuals with survival times censored at $t_i$ are assumed to be at risk at $t_i$.
- The Kaplan-Meier estimate of the cumulative survivor function at time $t$ is given by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \le t}(1 - \frac{d_i}{l_i}) & \text{if } t \ge t_1 \end{cases} \quad (1)$$

- A plot of the Kaplan-Meier estimate of the survivor function (slide 56) takes the form of a step function, in which the survival probabilities decrease at each death time and are constant between adjacent deaths times.
- Censorings do not affect the estimate of $S(t)$, but contribute in Equation 1 by decreasing $l_i$, the number of persons at risk, at the next death time.
- If the largest observed survival time (which we will call $t_z$) is a censored survival time, then $\hat{S}(t)$ is undefined for $t > t_z$, otherwise $\hat{S}(t) = 0$ for $t > t_z$.



- The standard error of the estimate can be obtained using Greenwood's method [10] (slide 78).
- An example is shown on slide 54.
- At $t = 2$ months we observed 2 deaths among the 35 patients at risk, so $p_1 = 1 - 2/35 = 0.9428$.
- At $t = 3$ months we observed 1 death among the 33 patients at risk, so $p_2 = 1 - 1/33 = 0.9697$.
- Subsequently, $\hat{S}(t) = 0.9429 \times 0.9697 = 0.9143$ for $3 \le t < 5$.



## K-M estimates for the sample data (up to 25 months)

| $t$ | at risk | observed deaths | $p_i$ | $S(t)$ | SE |
|---|---|---|---|---|---|
| 0 | 35 | 0 | 1.0000 | 1.0000 | – |
| 2 | | | | | |
| 2 | 35 | 2 | 0.9429 | 0.9429 | 0.0392 |
| 3 | 33 | 1 | 0.9697 | 0.9143 | 0.0473 |
| 5 | 32 | 1 | 0.9688 | 0.8857 | 0.0538 |
| 7 | 31 | 1 | 0.9677 | 0.8571 | 0.0591 |
| 8 | 30 | 1 | 0.9667 | 0.8286 | 0.0637 |
| 9 | 29 | 1 | 0.9655 | 0.8000 | 0.0676 |
| 11 | 28 | 1 | 0.9643 | 0.7714 | 0.0710 |
| 13+ | 27 | 0 | | | |
| 14+ | 26 | 0 | | | |
| 19 | 25 | 1 | 0.9600 | 0.7406 | 0.0745 |
| 22 | 24 | 1 | 0.9583 | 0.7097 | 0.0776 |
| 25+ | 23 | 0 | | | |
| | | ... | | | |



## Kaplan-Meier estimates in Stata

```
. stset surv_mm, failure(status==1,2)
. sts list
         Beg.          Net     Survivor     Std.
Time    Total   Fail   Lost     Function    Error     [95% CI]
-------------------------------------------------------------
   2       35      2      0       0.9429   0.0392   0.7903   0.9854
   3       33      1      0       0.9143   0.0473   0.7573   0.9715
   5       32      1      0       0.8857   0.0538   0.7236   0.9555
   7       31      1      0       0.8571   0.0591   0.6903   0.9379
   8       30      1      0       0.8286   0.0637   0.6577   0.9191
   9       29      1      0       0.8000   0.0676   0.6258   0.8992
  11       28      1      0       0.7714   0.0710   0.5946   0.8785
  13       27      0      1       0.7714   0.0710   0.5946   0.8785
  14       26      0      1       0.7714   0.0710   0.5946   0.8785
  19       25      1      0       0.7406   0.0745   0.5603   0.8558
  22       24      1      0       0.7097   0.0776   0.5271   0.8323
  25       23      0      1       0.7097   0.0776   0.5271   0.8323
-------------------------------------------------------------
```

## Graphical presentation of $S(t)$

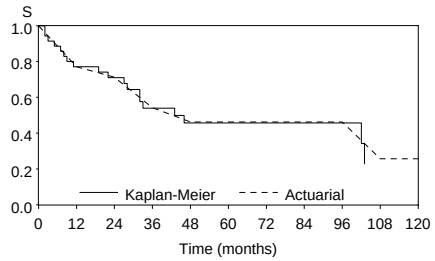



Figure 5: Estimates of the cumulative observed survivor function using the actuarial method (annual intervals) and the Kaplan-Meier method (based on survival time in months).

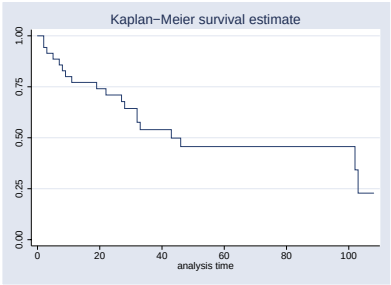


- The Kaplan-Meier estimates are graphed as a step function, with vertical drops at each death time.
- The actuarial method provides estimates of the survivor function at the end of each interval, and no estimate of the survivor function is made between these points.
- It is customary to interpolate the actuarial estimates by 'joining the dots', which corresponds to an approximately even distribution of deaths within each interval.



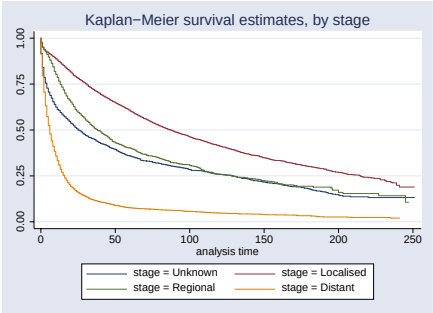
## Plotting Kaplan-Meier estimates of $S(t)$ using Stata

```
. use http://www.bioepi.org/teaching/sa/colon_sample
. stset surv_mm, failure(status==1,2)
. sts graph
```





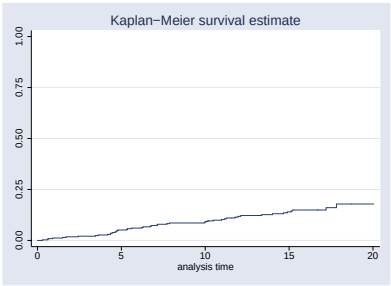

```
. use http://www.bioepi.org/teaching/sa/colon
. stset surv_mm, failure(status==1,2)
. sts graph, by(stage)
```






## Plotting estimates of the incidence proportion $(1 - S(t))$

```
. use http://www.bioepi.org/teaching/sa/diet
. stset dox, fail(chd) origin(doe) scale(365.25)
. sts graph, failure
```






## Ties in survival data

- If two individuals have the same survival time (time to event or time to censoring), we say that the survival times are 'tied'.
- Many of the standard methods for survival analysis, such as the Kaplan-Meier method and the Cox proportional hazards model, assume that survival time is measured on a continuous scale and that ties are therefore rare.
- In population-based survival analysis, however, ties are common.
- For example, among the 9087 Finns diagnosed with colon carcinoma during 1985–1994, 490 died during the first month of follow-up and 542 during the second month of follow-up (although there were no censorings during these months since every individual had a potential follow-up time of at least 12 months).



## Comparison of the Kaplan-Meier and actuarial methods

- Estimates of the survivor function can be presented in either tabular or graphical form.
- For tabular presentations, we rarely require estimates of the survivor function for interval lengths shorter than one year so the actuarial method suffices.
- We will therefore focus on a comparison of the two methods where the aim is to present estimates of $S(t)$ graphically.
- Since both the Kaplan-Meier and actuarial methods aim to estimate the same quantity, the survivor function, we would expect the estimates to be similar, if not identical.
- The estimates in Figure 5 are similar, although not identical. If the Kaplan-Meier estimates are made using survival time in years (rather than months) as the basis of calculation we see even greater disparity between the two methods (Figure 6).



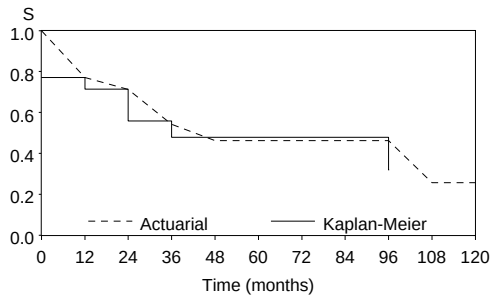



Figure 6: Estimates of the cumulative observed survivor function using the actuarial method (annual intervals) and the Kaplan-Meier method (based on survival time in years).

- One reason for this difference is the way in which ties are handled.
- If both censoring(s) and death(s) occur at the same time, the Kaplan-Meier method assumes that all of the individuals with censored survival times were at risk at the time of the death(s), whereas the actuarial method assumes that half of these individuals were at risk at the time of the death(s).
- There is little reason to assume that all deaths should precede all concurrent withdrawals when time is measured on a discrete scale (e.g. in days or longer units).
- There are no censorings during the first 12 months in our example so the two estimates of $S(12)$ are identical in Figure 6.
- The Kaplan-Meier estimator was developed for applications where survival time is measured on a continuous scale, where ties are rare, although the Kaplan-Meier estimator is discrete in nature.



- The actuarial method takes account of ties through application of the actuarial assumption, whereas the Kaplan-Meier method overestimates the survivor function (to a very small degree) in the presence of ties by assuming that all deaths precede concurrent withdrawals.
- Another consequence of the Kaplan-Meier method being developed for continuous time and the actuarial method for grouped data is in the way in which the estimates are interpreted.
- The actuarial method provides estimates of the survivor function at the end of each interval, and no estimate of the survivor function is made between the interval endpoints.
- Actuarial estimates of $S(t)$ for values of $t$ between the interval endpoints are obtained by interpolation (as in Figures 5 and 6), which corresponds to assuming an approximately even distribution of deaths within each interval.



- This assumption may not always be valid, especially during the first year of follow-up where mortality is often highest during the first few months, and the first month in particular.
- The Kaplan-Meier method, on the other hand, provides an estimate of $S(t)$ for all values of $t$, although the estimate of $S(t)$ is constant between event times.
- The interpretation of the Kaplan-Meier estimate of $S(t)$ presented in Figure 6 is that 22% of the patients die immediately following diagnosis but no deaths occur for another 12 months, indicating that the Kaplan-Meier method is clearly inappropriate for heavily grouped data.
- For data which is not heavily grouped, such as cancer registry data when survival time is estimated to the nearest month, the Kaplan-Meier method could be considered superior to the actuarial method with annual intervals.



- However, the actuarial method based on monthly intervals is an even better alternative, since it accounts for the ties in the data through use of the actuarial assumption.
- The actuarial method with annual intervals has been popular in population-based survival analysis since it requires fewer arithmetic calculations than would be required for the Kaplan-Meier method with survival time measured in months.
- With the advent of computers, however, this is no longer a significant advantage and a method based on survival time measured in months is preferable for producing graphical estimates of $S(t)$.
- Either the Kaplan-Meier method or the actuarial method will suffice, although the actuarial method could be considered technically superior if the data contain ties (although the differences are insignificant in practice).



## Informative right-censoring

- To make it possible for statistical analysis we make the crucial assumption that, conditional on the values of any explanatory variables, censoring is unrelated to prognosis (the probable course and outcome of the disease).
- The statistical methods used for survival analysis assume that the prognosis for an individual censored at time $t$ will be no different from those individuals who were alive at time $t$ and were under follow-up past time $t$.
- One way to think of this is that, conditional on the values of any explanatory variables, the individuals censored at time $t$ should be a random sample of the individuals at risk at time $t$.
- This is known as noninformative censoring. Under this assumption, there is no need to distinguish between the different reasons for right-censoring described on slide 12.



- When withdrawal from follow-up is associated with prognosis, this is known as informative censoring and standard methods of analysis will result in biased estimates.
- Common methods for controlling for informative censoring are to stratify or condition on those explanatory factors on which censoring depends.
- Censoring due to termination of the study, or accidental death, are usually uninformative, but careful consideration must be given to other forms of censoring.
- Determining whether or not censoring is informative is not a statistical issue — it must be made based on subject matter knowledge.



## Example of informative censoring — dairy cows

- Consider a cohort study of dairy cows where the aim is to study whether the type of feed and type of housing (indoor/outdoor) are associated with incidence of a disease.
- If a cow is slaughtered without the disease being diagnosed, is it appropriate to consider the survival time as censored in the analysis?
- No, not if the disease, or a precursor to the disease, affects milk production and cows with low milk yield are sacrificed.
- If this were the case then the sacrificed cows would be more likely to be diagnosed with the disease than cows who were not sacrificed.



## Example of informative censoring — colon cancer in IBD patients

- In a historical cohort study, 19,500 individuals with inflammatory bowel disease (IBD) were identified in the Swedish hospital inpatient separations register and IBD registers maintained in Uppsala and Stockholm.
- We were interested in risk factors for cancer of the colon; the cohort was followed up using the Swedish cancer register.
- Some patients had their colon surgically removed (colectomy) without being diagnosed with colon cancer, so were not at risk for colon cancer.
- These were the patients with the most extensive type of IBD, and it is known that risk of colon cancer is proportional to the extent of the IBD.
- Therefore, censoring due to colectomy is informative.

## Censoring due to death from competing risks is often informative

- When estimating cause-specific survival, deaths due to causes other than the cancer of interest are considered censored.
- Deaths due to other causes are more likely to occur among elderly patients and these patients usually have a higher risk of dying due to cancer.
- As such, censoring due to death from competing risks is informative.
- The solution is to stratify by age. The survival of censored individuals is then estimated by patients of a similar age.



## Censoring and truncation

- With right censoring, the most common form of censoring in medical studies, we know that the event has not occurred during follow-up, but we are unable to follow-up the patient further. We know only that the true survival time of the patient is greater than a given value.
- Less common is left-censoring, where we know the event has occurred prior to the time of observation but we don't know exactly when.
- Interval censoring occurs when we know that the event has occurred between two time points but don't know the exact date (e.g. HIV infection between two test dates, or cancer between two screens).
- Standard methods for survival analysis assume that all censored data are right censored and we will assume that this is the case.
- Special methods are required for analysing left censored and interval censored data, which will not be covered in this course.



- Censoring, in general, refers to the situation where we can identify the individuals in our study but we do not have precise information on the event time for all individuals (we know only that it is in some interval).
- A second feature of survival studies, often confused with censoring, is *truncation*.
- Truncation refers to the situation where certain subjects are screened such that the investigator is not aware of their existence.
- Left truncated data occurs when we only observe the individual if they are event free after a certain follow-up time. For example, late entry to the study.
- Consider a study of exposures in utero on IQ measured at entrance to military service.
- Right truncated data occurs when only individuals who experience the event of interest are included in the study.



- Special methods of analysis are required for analysing truncated data, such as use of a conditional likelihood or a method which uses a selective risk set (see Klein & Moeschberger (1997) [11]).
- Nevertheless, there are examples in the literature where right truncated data have been analysed using standard methods (see Altman and Bland (1998) [12] for some examples).
- Such an approach, for example, led to the dubious finding that left handed people die, on average, seven years younger than right handed people [13].
- The investigators studied all individuals who died during a certain time-interval (i.e. the data were right truncated).
- Those individuals in the cohort who died, for example, during their 70s, belong to a birth cohort where left handedness was less prevalent. Those who died, for example, during their 20s, would be much more likely to be left handed.



## Estimating AIDS incubation time: An example of right truncated data

- Knowledge of the time between HIV infection and development of AIDS (called the incubation period) is important in AIDS research.
- The median incubation time in developed countries is currently 12-14 years, and is increasing as treatment becomes more effective.
- The first reliable estimates of incubation time were obtained in the early 1980's by studying individuals who developed AIDS from blood transfusions (before prospective donors were screened for HIV).
- Only individuals who experienced the event could be studied. That is, the data were right truncated.
- Not all blood recipients were exposed to HIV, and not everyone who was exposed had developed AIDS at the time of the analysis.



- Nevertheless, by studying those individuals who developed AIDS as a result of HIV exposure at transfusion, using appropriate statistical methods, it was possible to estimate incubation time.



## Estimating the standard error of $S(t)$

- The most widely used method for estimating the standard error of the estimated survival proportion is the method described by Greenwood (1926) [14, 10].
- Appropriate for both the actuarial and Kaplan-Meier methods.
- Appropriate for both observed and cause-specific survival.
- Known as Greenwood's method or Greenwood's formula. The formula,

$$\mathrm{SE}({}_1p_i) = {}_1p_i \left[ \sum_{j=1}^{i} \frac{d_j}{l'_j(l'_j - d_j)} \right]^{\frac{1}{2}}, \qquad (2)$$

  is slightly laborious for hand calculation, but readily available in many computer programs.



- This is the default method for the software used in this course.
- Non-integer values for $l'_i$, e.g. $l'_i = 20.5$, do not cause any problems in practical use.
- For a single interval, Equation 2 reduces to

$$\mathrm{SE}(p_i) = p_i \left\{ \frac{d_i}{l'_i(l'_i - d_i)} \right\}^{\frac{1}{2}} = \sqrt{p_i(1-p_i)/l'_i},$$

  which is the familiar binomial formula for the standard error of the observed interval-specific survival proportion based on $l'_i$ trials.
- It can also be shown for the general case that Equation 2 reduces to the binomial standard error in the absence of censoring.

## Confidence intervals for estimated survival proportions

- Confidence intervals can be calculated for any estimated survival proportion in order to provide a measure of uncertainty associated with the point estimate.
- A 95% confidence interval (CI) is an interval, i.e. a range of values, such that under repeated sampling, the true survival proportion will be contained in the interval 95% of the time.
- The CI is often called an interval estimate for the true survival proportion, while the estimated survival proportion is called the point estimate.
- In nationwide population-based cancer registries, such as Finland, we assume that every person with a diagnosis of cancer in the population has been registered.
- The idea of sampling from a population which has a 'true mean' survival time is therefore not obvious.



- We actually assume that among the entire population of Finland, the survival time for every individual, given a diagnosis of cancer, follows some theoretical distribution with a 'true mean' survival time, and the people actually diagnosed make up our sample.
- Estimated confidence intervals provide an indication of the level of statistical uncertainty in the estimated survival proportions. They do not represent the range of possible prognoses for an individual patient.
- A confidence interval for the true survival proportion can be obtained by assuming that the estimated survival proportion is normally distributed around the true value with estimated variance given by the square of the standard error.
- A two-sided $100(1-\alpha)\%$ confidence interval ranges from $p - z_{\alpha/2}\mathsf{SE}(p)$ to $p + z_{\alpha/2}\mathsf{SE}(p)$, where $p$ is the estimated survival proportion (which can be an interval-specific or cumulative observed, cause-specific, or relative survival), $\mathsf{SE}(p)$ the associated standard error, and $z_{\alpha/2}$ the upper $\alpha/2$ percentage point of the standard normal distribution.



- For a 95% confidence interval, $z_{\alpha/2} = 1.96$, and for a 99% confidence interval, $z_{\alpha/2} = 2.58$.
- The standard error of the observed and cause-specific survival proportion can be obtained using Greenwood's method (slide 78).
- As a rule of thumb, the normal approximation for a single interval $i$ is usually appropriate when both $l_i' p_i$ and $l_i'(1-p_i)$ are greater than or equal to 5 [15].
- Confidence intervals obtained in this way are symmetric about the point estimate and can sometimes contain implausible values for the survival proportion, i.e., values less than zero or greater than one.



## Constructing confidence intervals on the log-hazard scale

- One method of obtaining confidence intervals for the observed survival proportion in the range [0,1] is to transform the estimate to a value in the range $[-\infty, \infty]$, obtain a confidence interval for the transformed value, and then back-transform the confidence interval to [0,1].
- One such transformation is the complementary log-log transformation, $\log[-\log(p)]$, which is equivalent to constructing the confidence intervals on the log-hazard scale.
- To estimate confidence intervals the relative survival ratio using this method, we first transform the estimated cumulative observed survival rate (OSR).
- We will write this transformation as $g(\mathsf{OSR}) = \log[-\log(\mathsf{OSR})]$, where $g$ is the complementary log-log transformation.
- We also require an estimate of the variance of the OSR on the log hazard scale.



- Using a Taylor series approximation, the variance of a function, $g$, of a random variable, $X$, can be approximated by

$$\mathsf{var}\{g(X)\} \approx \left\{\frac{\mathsf{d}g(X)}{\mathsf{d}X}\right\}^2 \mathsf{var}(X)$$

- If we denote the cumulative observed survival proportion by $X$ then, noting that

$$\frac{\mathsf{d}\log[f(X)]}{\mathsf{d}X} = \frac{1}{f(X)}\frac{\mathsf{d}f(X)}{\mathsf{d}X},$$

we have

$$\mathsf{var}\{g(X)\} = \mathsf{var}\{\log[-\log(X)]\} \approx \frac{1}{[X\log(X)]^2}\mathsf{var}(X).$$

- An estimated 95% confidence interval on the log hazard scale is therefore given by $g(\mathsf{OSR}) \pm 1.96\sqrt{\mathsf{var}\{g(OSR)\}}$, which is then back-transformed to give a 95% confidence interval for the OSR.



## Expectation of life

- The expectation of life is a single measure summarising the survival of the patients.
- It is calculated as the area under the survivor function.
- Extrapolation techniques can be used if the survivor function does not reach zero while the patients are under follow-up.
- By comparing the expectation of life of the patients to the expectation of life of a comparable group from the general population, it is possible to estimate the 'proportion of expected life lost'.



## Median survival time

- The median survival time is another measure used to summarise the survival experience of the patients.
- The median survival time is the time beyond which 50% of the individuals in the population are expected to survive.
- It is estimated from the life table as the time at which the cumulative observed survival proportion falls below 0.5.
- The median is estimated by extrapolation if the cumulative observed survival proportion does not sink below 0.5 during the period the patients are under follow-up.



## Comparison of survival between groups

- Comparing survival at a fixed time point (e.g. five years) wastes available information.
- It is invalid to compare the proportion surviving at a given time, based on the comparison of two binomial proportions, where the time point for comparison is chosen after viewing the estimated survivor functions (e.g. testing for a difference at the point where the Kaplan-Meier curves show the largest difference).
- Various tests are available (parametric and non-parametric) for testing equality of survival curves. The most common is the log rank test, which is non-parametric.
- Start by tabulating the number at risk in each group and the total number of events (deaths) at every time point when one of more deaths occur.

- Under the null hypothesis that the two survival curves are the same, the expected number of deaths in each group will be proportional to the number at risk in each group.
- For example (see slide 89), at $t = 2$ months we observed 2 deaths (one male and one female). Conditional on 2 deaths being observed, we would expect $2 \times 19/35 = 1.086$ deaths among the 19 males at risk and $2 \times 16/35 = 0.914$ deaths among the 16 females at risk.
- Now calculate the totals of the observed and expected number of deaths for each group, calling them $O_1$, $O_2$, $E_1$, and $E_2$, and calculate the following test statistic

$$\theta = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}. \quad (3)$$

- Under the null hypothesis, $\theta$ will approximately follow a $\chi^2$ distribution with 1 degree of freedom. That is, if $\theta$ is greater than 3.84 then we reject the null hypothesis and conclude that there is a statistically significant difference between the two survival curves.



## Log rank test for comparing survival of males and females

| event | males | | | females | | |
|---|---|---|---|---|---|---|
| time | at risk | obs | exp | at risk | obs | exp |
| 2 | 19 | 1 | 1.086 | 16 | 1 | 0.914 |
| 3 | 18 | 1 | 0.545 | 15 | 0 | 0.455 |
| 5 | 17 | 1 | 0.531 | 15 | 0 | 0.469 |
| 7 | 16 | 1 | 0.516 | 15 | 0 | 0.484 |
| 8 | 15 | 1 | 0.500 | 15 | 0 | 0.500 |
| 9 | 14 | 0 | 0.483 | 15 | 1 | 0.517 |
| 11 | 14 | 0 | 0.500 | 14 | 1 | 0.500 |
| 19 | 13 | 0 | 0.520 | 12 | 1 | 0.480 |
| 22 | 13 | 1 | 0.542 | 11 | 0 | 0.458 |
| 27 | 12 | 0 | 0.545 | 10 | 1 | 0.455 |
| 28 | 11 | 0 | 0.550 | 9 | 1 | 0.450 |
| 32 | 11 | 2 | 1.158 | 8 | 0 | 0.842 |
| 33 | 9 | 1 | 0.563 | 7 | 0 | 0.438 |
| 43 | 8 | 0 | 0.615 | 5 | 1 | 0.385 |
| 46 | 8 | 1 | 0.667 | 4 | 0 | 0.333 |
| 102 | 2 | 0 | 0.500 | 2 | 1 | 0.500 |
| 103 | 2 | 1 | 0.667 | 1 | 0 | 0.333 |

Totals: $O_1 = 11$, $E_1 = 10.488$, $O_2 = 8$, $E_2 = 8.512$



- The test statistic is $\theta = (O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2 = 0.056$, implying no evidence of a difference in survival between males and females.
- For $k$ groups, the log rank test statistic is

$$\theta = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

  which has an approximate $\chi^2_{k-1}$ distribution under the null hypothesis.
- The log rank test is designed to be sensitive to departures from the null hypothesis in which the two hazards (instantaneous death rates) are proportional over time. It is very insensitive to situations in which the hazard functions cross.
- The log rank test puts equal weight on every failure (irrespective of the number at risk at the time of the failure).



- An alternative test, the generalised Wilcoxon test, is constructed by weighted the contribution of each failure time by the total number of individuals at risk and is consequently more sensitive to differences early in the follow-up period (when the number at risk is larger).
- The Wilcoxon test is more powerful than the log rank test if the proportional hazards assumption does not hold.
- It is difficult to apply the log rank test while simultaneously controlling for potential confounding variables (a regression approach is preferable).
- In a randomised clinical trial, however, potential confounders are controlled for in the randomisation, so we can use the log rank test to compare survival curves for the different treatment groups.
- The log rank test provides nothing more than a test of statistical significance for the difference between the survival curves, it tells us nothing about the size of the difference. A regression approach allows us to both determine statistical significance and to estimate the size of the effect.



## Log rank test in Stata

```
. use colon_sample
. stset surv_mm, failure(status==1,2)
. sts test sex
Log-rank test for equality of survivor functions
------------------------------------------------
       |  Events
sex    |  observed       expected
-------+------------------------
Male   |        11          10.49
Female |         8           8.51
-------+------------------------
Total  |        19          19.00
             chi2(1) =       0.06
             Pr>chi2 =     0.8113
```



## A slight change of focus

- I'm now moving to the diet example, which is an epidemiological cohort study.
- Epidemiologists place a greater emphasis on modelling rates and lesser emphasis on estimating and comparing proportions and this will be reflected in my presentation.
- This is due to tradition; the methods presented in this course are equally appropriate for, for example, cancer survival analysis as they are for analysing epidemiological cohort studies.



## Measures of effect

- For a failure response, where we estimate $\lambda$ (the rate), possible measures of effect are $\lambda_1 - \lambda_0$ (rate difference) $\lambda_1/\lambda_0$ (rate ratio)
- For a metric response, where we estimate $\mu$ (the mean), possible measures of effect are $\mu_1 - \mu_0$ (difference in means) $\mu_1/\mu_0$ (ratio of means)
- The scale should be chosen so that the exposure has roughly the same effect regardless of level, so that we can combine information about effects from different studies.
- For a metric response, we generally assume that differences in means are constant over levels of other factors.
- For a failure response, ratios are more likely to be constant than the differences, but interpretation also matters.



## Additive vs multiplicative models

- Consider the following rates (of some event) per 1000 person-years for 'exposed' and 'unexposed' individuals.

| agegrp | unexposed | exposed |
|---|---|---|
| 0 | 19.60 | 14.92 |
| 5 | 1.82 | 1.37 |
| 15 | 2.58 | 2.29 |
| 25 | 3.03 | 2.82 |
| 35 | 4.91 | 3.91 |
| 45 | 10.37 | 7.85 |
| 55 | 22.66 | 17.18 |
| 65 | 54.84 | 43.04 |
| 75 | 147.03 | 131.37 |

- How might we measure the effect of exposure?

## Effect of exposure

| agegrp | difference | ratio |
|---|---|---|
| 0 | -4.68 | 0.76 |
| 5 | -0.45 | 0.75 |
| 15 | -0.29 | 0.89 |
| 25 | -0.21 | 0.93 |
| 35 | -1.00 | 0.80 |
| 45 | -2.52 | 0.76 |
| 55 | -5.49 | 0.76 |
| 65 | -11.79 | 0.78 |
| 75 | -15.67 | 0.89 |
| combined | -0.72 | 0.81 |



- In this example, the rate ratio would be the preferred measure for describing the effect of exposure since it is consistent across age groups.
- That is, we would choose a multiplicative model. The effect of exposure is to multiply the rate by $0.81$.
- Here we have based on choice of model on goodness-of-fit to the observed data.
- Knowledge of (or assumptions about) the underlying biologic mechanism may also influence our choice of model.



## Choice of time scale

- There are several time scales along which rates might vary. These differ from one another only in the choice of *time origin*, the point at which time is zero.
- What is the time origin for each of the following questions?
  - What is the time?
  - How old are you?
  - For how long have you been staying at castel brando?
- In which units did you specify time? Could different units have been used?
- Time progresses in the same manner but, in answering these questions, we have applied a different time origin and used different units.



## Common time scales in epidemiology

| Origin | Time scale |
|---|---|
| Birth | Age |
| Any fixed date | Calendar time |
| First exposure | Time exposed |
| Entry into study | Time in study |
| Disease onset | Time since onset |
| Diagnosis | Time since diagnosis |
| Start of treatment | Time on treatment |

- In many of the methods used in survival analysis, effects are adjusted for the underlying time scale. Choice of time scale therefore has important implications.
- On many time scales, subjects do not enter follow-up at the time origin, $t = 0$.
- To deal with these issues `stset` has two additional options, one to specify the origin of time, the other to specify the time of entry to the study.



## The diet data

- For the diet data
  - time of entry = `doe`
  - time of exit = `dox`
  - event indicator = `chd`

```
. stset time, fail() enter() origin() scale()
```

sets the `st` variables and the time scale on which all following analyses are to be carried out. To set the time scale as time since entry:

```
. stset dox, fail(chd) entry(doe) origin(doe) scale(365.25)
```

- Each individual enters the study (becomes 'at risk') at the date specified by `doe`.
- The date of entry is also the time origin (time zero).



- By specifying `scale(365.25)` we are scaling the time unit from days to years.
- To use attained age as the time scale we specify

```
. stset dox, fail(chd) entry(doe) origin(dob) scale(365.25)
```

- Individuals enter the study at `doe` (as before) but the time origin is now the date of birth.
- To use calendar time as the time scale we specify a fixed date as the time origin. For example

```
. stset dox, fail(chd) entry(doe) origin(d(1/1/1900))
```



## Estimating CHD rates according to energy intake

- We first `stset` the data using time since entry as the timescale.

```
. stset dox, fail(chd) origin(doe) scale(365.25) id(id)
     failure event:  chd != 0 & chd < .
obs. time interval:  (dox[_n-1], dox]
 exit on or before:  failure
    t for analysis:  (time-origin)/365.25
            origin:  time doe
------------------------------------------------------------
      337  total obs.
        0  exclusions
------------------------------------------------------------
      337  obs. remaining, representing
      337  subjects
       46  failures in single failure-per-subject data
 4603.669  total analysis time at risk, at risk from t = 0
```



- The `stptime` command tabulates the number of events and person time-at risk and calculates event rates.

```
. stptime, by(hieng) per(1000)
         failure _d:  chd
   analysis time _t:  (dox-origin)/365.25
             origin:  time doe
hieng | person-time  failures     rate   [95% Conf. Interval]
------+-----------------------------------------------------
  low |   2059.4305     28   13.595992   9.387478    19.69123
 high |   2544.2382     18   7.0748093   4.457431     11.2291
------+-----------------------------------------------------
total |   4603.6687     46   9.9920309   7.484296    13.34002
```

- Note that person-time is in years but the rates are per 1000 years.

- The `strate` command performs similar calculations.

```
. strate hieng, per(1000)
        failure _d:  chd
  analysis time _t:  (dox-origin)/365.25
            origin:  time doe
Estimated rates (per 1000) and lower/upper bounds of 95% CI
(337 records included in the analysis)
  +-------------------------------------------------+
  | hieng    D        Y      Rate    Lower     Upper |
  |-------------------------------------------------|
  |   low   28   2.0594   13.5960   9.3875   19.6912 |
  |  high   18   2.5442    7.0748   4.4574   11.2291 |
  +-------------------------------------------------+
```



- The incidence rate ratio (IRR) for individuals with a high compared to low energy intake is $7.1/13.6 = 0.52$.
- That is, without controlling for any possible confounding factors, we estimate that individuals with a high energy intake have a CHD risk that is approximately half that of individuals with a low energy intake.
- This is sometimes called a 'crude estimate'; it is not adjusted for potential confounders.
- Is this a true effect? What important confounder might we need to consider?



## A model for the incidence rate

- When working with rates, we believe that effects are most likely to be multiplicative.
- That is, we believe that the rate in the high energy group ($\lambda_1$) is likely to be a multiple of the rate in the low energy group ($\lambda_1$). The multiplication factor is the incidence rate ratio, $\theta$.

$$\lambda_1 = \lambda_0\theta, \text{ for example, } 7.1 = 13.6 \times 0.52$$

- If the explanatory variable $X$ is equal to 1 for individuals with a high energy intake and 0 for individuals with a low energy intake then we can write

$$\lambda(X) = \lambda_0 \times \theta^X$$

- That is, $\lambda = \lambda_0$ when $X = 0$ and $\lambda = \lambda_0\theta$ when $X = 1$.



- In practice, it is more convenient to work on a logarithmic scale.

$$\begin{aligned} \lambda &= \lambda_0 \times \theta^X \\ \ln(\lambda) &= \ln(\lambda_0 \times \theta^X) \\ &= \ln(\lambda_0) + \ln(\theta^X) \\ &= \ln(\lambda_0) + \ln(\theta)X \\ \ln(\lambda) &= \beta_0 + \beta_1 X \end{aligned}$$

where $\beta_0 = \ln(\lambda_0)$ and $\beta_1 = \ln(\theta)$ is the log IRR.

- $\ln(\lambda) = \beta_0 + \beta_1 X$ is a Poisson regression model with one binary explanatory variable, $X$.
- Exercise: What are the estimates of $\beta_0$ and $\beta_1$?



## Three regression models commonly applied in epidemiology

- Linear regression

$$\mu = \beta_0 + \beta_1 X$$

- Logistic regression

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

- Poisson regression

$$\ln(\lambda) = \beta_0 + \beta_1 X$$

- In each case $\beta_1$ is the effect per unit of $X$, measured as a change in the mean (linear regression); the change in the log odds (logistic regression); the change in the log rate (Poisson regression).



## The effect of high energy, using Poisson regression

| hieng | X | D | Y | Rate per 1000 |
|---|---|---|---|---|
| low | 0 | 28 | 2059.4 | 13.60 |
| high | 1 | 18 | 2544.2 | 7.07 |

$$\begin{aligned} \ln(\lambda) &= \beta_0 + \beta_1 X \\ \ln(28/2059.4) &= \beta_0 = -4.3 \\ \ln(18/2544.2) &= \beta_0 + \beta_1 \\ \ln\left(\frac{18/2544.2}{28/2059.4}\right) &= \beta_1 \\ -0.6532 &= \beta_1 \\ 0.52 &= \exp(\beta_1) \end{aligned}$$



## Poisson regression in Stata

```
. poisson chd hieng, expos(y)
  chd |      Coef.   [95% Conf. Interval]
------+------------------------------------
hieng |  -.6532341  -1.245357   -.0611114
_cons |   -4.29798  -4.668379   -3.927582
on a log scale, and
. poisson chd hieng, expos(y) irr
    chd |       IRR   [95% Conf. Interval]
--------+------------------------------------
  hieng |  .5203602   .2878382    .9407184
on a ratio scale.
```

- The model is estimated using the method of maximum likelihood.
- Confidence intervals are constructed by assuming the estimated regression parameters are normally distributed.



- That is, confidence intervals are constructed on the log scale, as is standard for ratio measures.
- We see that the confidence limits for the IRR are simply the exponentiated limits of the log IRR.
- As such, the CI for the IRR is not symmetric around the point estimate.
- Can also use the `streg` (which fits the model in the framework of parametric survival models) or `glm` (generalised linear model) commands.

## Categorical exposures with more than two levels

- The variable hieng has two levels
- The variable eng3, created below, has 3 levels.

```
. egen eng3=cut(energy),at(1500,2500,3000,4500)
```

```
     eng3 |    Rate
----------+--------
     1500 |   16.90
     2500 |   10.91
     3000 |    4.88
------------------
```



- Effect(s) of eng3

```
Levels  Effect    95% Confidence Interval
2/1     0.6452    [ 0.339 , 1.229 ]
3/1     0.2886    [ 0.124 , 0.674 ]
```

- To include eng3 in regression commands we need to use indicator variables

| eng3 | $X1$ | $X2$ | $X3$ |
|---|---|---|---|
| 1500 | 1 | 0 | 0 |
| 2500 | 0 | 1 | 0 |
| 3000 | 0 | 0 | 1 |

for the 3 levels.



```
. tabulate eng3, generate(X)
. poisson chd X2 X3, e(y) irr
chd |     IRR  [95% Conf. Interval]
----+------------------------------
 X2 |   .6452  .3388815    1.228561
 X3 |   .2886  .1235342    .6744495
```

- The variable that indicates the category with the lowest energy intake is omitted, meaning this is the reference category.
- In terms of the parameters

$$\begin{aligned} \ln(\lambda) &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 \\ &= \beta_0 \quad \text{(level 1)} \\ &= \beta_0 + \beta_2 \quad \text{(level 2)} \\ &= \beta_0 + \beta_3 \quad \text{(level 3)} \end{aligned}$$



## Automatic generation of indicators

```
. xi: poisson chd i.eng3, e(y) irr
-----------------------------------------------
         chd |        IRR  [95% Conf. Interval]
-------------+---------------------------------
 _Ieng3_2500 |   .6452416  .3388815    1.228561
 _Ieng3_3000 |   .2886479  .1235342    .6744495
```

- xi stands for e(x)pand (i)ndicators
- The baseline is, by default, the first level, but this can be changed to (say) the third level (3000–) with

```
. char eng3[omit] 3000
```

To re-set the default, use `char eng3[omit]`.



## Metric exposure variables

- The effect of energy on failure.

```
. poisson chd energy , e(y) irr
-----------------------------------------
   chd |        IRR  [95% Conf. Interval]
-------+---------------------------------
energy |     .99885  .9981367    .9995637
```

- For each 1 unit increase in energy intake, the CHD rate is reduced by 0.1%. The units of energy are kcals per day.

```
. summarize energy
Variable |  Obs     Mean    Std. Dev.     Min      Max
---------+-----------------------------------------
  energy |  337    2828.9    441.8      1748.4   4395.8
```



```
. gen energy100=energy/100
. poisson chd energy100, e(y) irr
-----------------------------------------
      chd |        IRR  [95% Conf. Interval]
----------+---------------------------------
energy100 |   .8913034  .8298593    .9572968
```

- The estimated IRR is $0.99885^{100} = 0.8913$. That is, for each 100 unit increase in energy intake, we estimate that the CHD rate is reduced by 11%.



## The model of constant effect over strata

- If the true effect of exposure varies across strata there is said to be 'effect modification' — the effect of exposure cannot then be represented by a single number.
- For example, the effect of high energy may depend on age - we say that the effect is modified by age (or that there is an interaction between energy intake and age).
- But if the estimates differ only randomly, we can consider a model in which the true effect is constant. This allows us to combine the information from different strata to yield a single estimate of exposure effect.
- We shall call this the estimate of effect 'controlled for' the stratifying variable(s).
- Statistical tests for the presence of effect modification are available (although there are no statistical tests for confounding).



## Effect of high energy controlled for job

```
Crude       Effect   95% CI
            0.5204   [ 0.288 , 0.941 ]
Effects     Level
by job      of job  Effect   95% CI
            1       0.4103   [ 0.124 , 1.362 ]  driver
            2       0.6551   [ 0.227 , 1.888 ]  conductor
            3       0.5177   [ 0.212 , 1.267 ]  bank worker
```

- These estimates are sufficiently close to be combined into a single estimate.

```
Effect of hieng controlled for job
Level        Effect    95% CI
2 vs 1       0.5248   [ 0.290 , 0.949 ]
```

## Effects of job controlled for hieng

```
Crude effects
levels 2/1  1.3720
levels 3/1  0.8766
Effects by energy
levels 2/1                levels 3/1
hieng        Effect       hieng         Effect
-------------------       -------------------
low          1.1369       low           0.8134
high         1.8153       high          1.0265
-------------------       -------------------
controlled   1.3584       controlled    0.8843
for hieng                 for hieng
```



## Using Poisson regression

```
. xi: poisson chd i.hieng i.job, e(y) irr
------------------------------------
         chd |        IRR   Std. Err.
-------------+----------------------
   _Ihieng_1 |   .5247666   .1585834
     _Ijob_2 |   1.358442   .5344426
     _Ijob_3 |   .8843023   .3229207
```

- The Stata Poisson regression command makes no distinction between the exposure variable and the control variable.
- The first number reported is the effect of `hieng` controlled for `job`, and the next two are the effects of `job` controlled for `hieng`.



## Models and parameters

- In the Poisson regression model we estimated 4 parameters (the intercept is not reported when we use the `irr` option). One parameter (the intercept) is a log rate and the other three are incidence rate ratios.
- The model is $\ln(\lambda) = \beta_0 + \beta_1 \text{hieng} + \beta_2 \text{cond} + \beta_3 \text{bank}$
- $\exp(\beta_0)$ is the predicted rate (not rate ratio) for an individual with all covariates at the reference level (i.e., a driver with a low energy intake).
- The estimated incidence rate ratios are

$$\begin{aligned} \theta &= \exp(\beta_1) \text{ (comparing high to low energy)} \\ \phi_c &= \exp(\beta_2) \text{ (comparing conductors to drivers)} \\ \phi_b &= \exp(\beta_2) \text{ (comparing bank to drivers)} \end{aligned}$$



## Parameters estimates with and without the `irr` option

```
. xi: poisson chd i.hieng i.job, e(y)
------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Ihieng_1 |  -.6448017   .3021979    -2.13   0.033    -1.237099   -.0525046
     _Ijob_2 |   .3063385   .3934232     0.78   0.436    -.4647568    1.077434
     _Ijob_3 |  -.1229563     .36517    -0.34   0.736    -.8386764    .5927639
       _cons |  -4.324988   .3118157   -13.87   0.000    -4.936136   -3.713841
------------------------------------------------------------------------------
. xi: poisson chd i.hieng i.job, e(y) irr
------------------------------------------------------------------------------
         chd |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Ihieng_1 |   .5247666   .1585834    -2.13   0.033      .290225     .9488499
     _Ijob_2 |   1.358442   .5344426     0.78   0.436     .6282879     2.937133
     _Ijob_3 |   .8843023   .3229207    -0.34   0.736     .4322823     1.808981
------------------------------------------------------------------------------
```

- The estimated rate for each combination of explanatory variables, as a



function of the baseline rate $\lambda$ and the three incidence rate ratios, is as follows.

| job | hieng=0 | hieng=1 |
|---|---|---|
| driv | $\lambda$ | $\lambda\theta$ |
| cond | $\lambda\phi_c$ | $\lambda\theta\phi_c$ |
| bank | $\lambda\phi_b$ | $\lambda\theta\phi_b$ |

- The estimated incidence rate for drivers with a low energy intake is $\exp(-4.325) = 0.0132$ events/person-year.
- The estimated incidence rate for conductors with a high energy intake is $0.0132 \times 1.358 \times 0.525 = 0.0094$ events/person-year.



## Effect modification

- Does `job` modify the effect of `hieng`?

```
job       Effect of hieng
driver       0.41
conductor    0.66
bank         0.52
```

- The figures represent the incidence rate ratios (comparing high to low energy intake) within the designated job category.
- If the effect of high energy is not modified by job then we would expect these to be similar.



- To compare these effects use ratios:

```
job       Effect
driver       0.41  0.41/0.41 = 1
conductor    0.66  0.66/0.41 = 1.60
bank         0.52  0.52/0.41 = 1.26
```

- The numbers 1.60 measures how much the effect of `hieng` differs between conductors and drivers, while 1.26 measures how much the effect of `hieng` differs between bank workers and drivers.
- They are called the interactions between `hieng` and `job`.



## Interactions are symmetric

- Does `hieng` modify the effects of `job`?

```
Level 2 vs level 1 of job        Level 3 vs level 1 of job
hieng  Effect Ratio              hieng  Effect Ratio
low    1.14   1.14/1.14 = 1      low    0.81   0.81/0.81 = 1
high   1.82   1.82/1.14 = 1.60   high   1.03   1.03/0.81 = 1.26
```

- The interactions between `hieng` and `job` are the same as those between `job` and `hieng`.

## Using Poisson regression

```
. xi: poisson chd i.hieng*i.job, e(y) irr
------------------------------------
         chd |        IRR   Std. Err.
-------------+----------------------
   _Ihieng_1 |   .4102648   .2512349
     _Ijob_2 |   1.136857   .5684285
     _Ijob_3 |    .813427   .3712769
_IhieXjob_~2 |   1.596755   1.303745
_IhieXjob_~3 |   1.261973   .9638479
```

- 0.41 is the effect of `hieng` when `job` is at its first level.
- 1.14 and 0.81 are the effects of `job` when `hieng` is at its first level.
- 1.60 and 1.26 are the interactions between `hieng` and `job`.



## Testing for interaction

```
. xi: poisson chd i.hieng*i.job, e(y) irr
------------------------------------
         chd |        IRR   Std. Err.
-------------+----------------------
   _Ihieng_1 |   .4102648   .2512349
     _Ijob_2 |   1.136857   .5684285
     _Ijob_3 |    .813427   .3712769
_IhieXjob_~2 |   1.596755   1.303745
_IhieXjob_~3 |   1.261973   .9638479
. testparm _IhieXjob*
  chi2(  2) =    0.33
Prob > chi2 =    0.8475
```

- No evidence of a statistically significant interaction. Could also use a likelihood ratio test.



## Models and parameters

| job | hieng=0 | hieng=1 |
|---|---|---|
| driv | $\lambda$ | $\lambda\theta$ |
| cond | $\lambda\phi_c$ | $\lambda\theta\phi_c\xi_{c1}$ |
| bank | $\lambda\phi_b$ | $\lambda\theta\phi_b\xi_{b1}$ |

- $\xi_{c1}$ and $\xi_{b1}$ are the interaction parameters.
- They measure deviations from the hypothesis of common effect of `hieng` in all `job` categories.



## How to make Stata produce stratified effects

- The trick is to define 0/1 variables as follows:

| job | hieng=0 | hieng=1 |
|---|---|---|
| driv | $\lambda_d$ | $\lambda_d\theta_d$ |
| cond | $\lambda_c$ | $\lambda_c\theta_c$ |
| bank | $\lambda_b$ | $\lambda_b\theta_b$ |

```
. gen ld = ( job == 1 )
. gen lc = ( job == 2 )
. gen lb = ( job == 3 )
. gen thd = ( job == 1 ) * ( hieng == 1 )
. gen thc = ( job == 2 ) * ( hieng == 1 )
. gen thb = ( job == 3 ) * ( hieng == 1 )
```



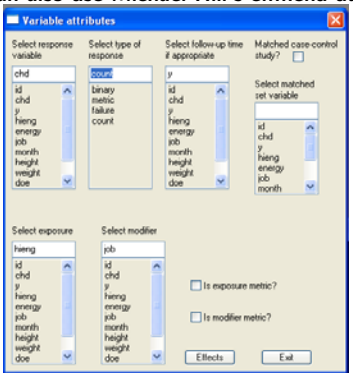
## Stata output

```
. poisson chd ld lc lb thc thd thb,
          e(y) irr nocons
chd |        IRR   [95% Conf. Interval]
----+-----------------------------------
 ld |   .0144648   .0072338     .028924
 lc |   .0164445   .0082238    .0328825
 lb |   .0117661   .0066821    .0207183
thd |   .4102648   .1235412    1.362438
thc |   .6550924   .2273009    1.888008
thb |   .5177431    .211639    1.266581
```

- The first three parameters are the base-line rates in the three job strata, and the next three parameters are the stratum-specific effects of `hieng`.



## Can also use Michael Hill's effmenu utility

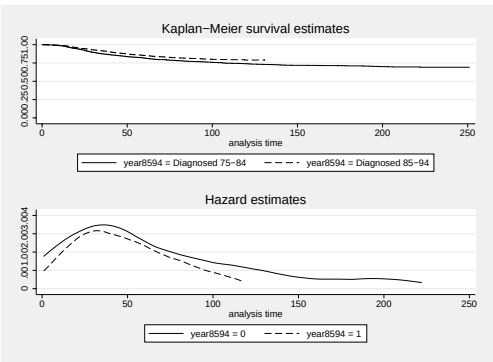





```
. effmenu1
Response variable is chd type failure
Follow-up time variable is y
Exposure variable is hieng  which is categorical
Modifier variable is job which is categorical
Effects measured as rate ratios
Effects of hieng levels 2/1
Level of job       Effect    95% Confidence Interval
driver             0.4103      [ 0.124 , 1.362 ]
conductor          0.6551      [ 0.227 , 1.888 ]
bank               0.5177      [ 0.212 , 1.267 ]
Overall test for effect modification
chi2(  2)     =   0.331  (P-value = 0.847)
```



## Localised melanoma, survival as a function of period



```
. use melanoma, clear
. stset surv_mm if stage==1, failure(status==1)
. strate year8594, per(1000)

        failure _d:  status == 1
  analysis time _t:  surv_mm

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(5318 records included in the analysis)
```

| year8594 | D | Y | Rate | Lower | Upper |
|---|---|---|---|---|---|
| Diagnosed 75-84 | 572 | 270.8810 | 2.1116 | 1.9455 | 2.2920 |
| Diagnosed 85-94 | 441 | 189.9790 | 2.3213 | 2.1145 | 2.5484 |

- The graphs suggest that patients diagnosed in the recent period have lower mortality (better survival) but the estimated rates suggest otherwise.

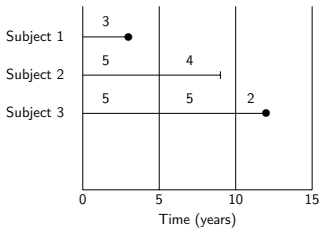


## Time as a confounder

- When the rate changes with time then time may confound the effect of exposure.
- We will, for the moment, assume that the rates are constant within broad time bands but can change from band to band.
- This approach (categorising a metric variable and assuming the effect is constant within each category) is standard in epidemiology.
- We often categorise metric variables — the only difference here is that the variable is 'time'.



- Consider a group of subjects with rates $\lambda_1$ during band 1, $\lambda_2$ during band 2, etc.




- What are the estimated failure rates for each of the bands?



## Splitting the records by follow-up time

- A convenient way to fit these models using a computer is to replace the single record for this subject by three new records, one for each band of observation.
- The new subject–band records can be treated as independent records.

| subject | timeband | follow-up | failure |
|---|---|---|---|
| 1 | 0-5 | 3 | 1 |
| 2 | 0-5 | 5 | 0 |
| 2 | 5-10 | 4 | 0 |
| 3 | 0-5 | 5 | 0 |
| 3 | 5-10 | 5 | 0 |
| 3 | 10-15 | 2 | 1 |

- The rate for timeband 0-5 is then $1/(3+5+5)$, and so on for other time bands.
- This method can be used whether rates are varying simply as a function of time or in response to some time–varying exposure.



## System variables created by stset

| | |
|---|---|
| _t0 | time at entry |
| _t | time at exit |
| _d | failure indicator |
| _st | inclusion indicator |

- For example, to stset the data with time since entry as the time scale.

```
. use http://www.bioepi.org/teaching/sa/diet
. stset dox, id(id) fail(chd) origin(doe) entry(doe) sc(365.25)
. list id  _t0 _t _d _st doe dox in 1/5, clean
 id   _t0         _t   _d   _st         doe         dox
127     0   16.791239    0    1   16Feb1960   01Dec1976
200     0   19.958932    0    1   16Dec1956   01Dec1976
198     0   19.958932    0    1   16Dec1956   01Dec1976
222     0   15.394935    0    1   16Feb1957   10Jul1972
305     0   1.4948665    1    1   16Jan1960   15Jul1961
```



## Splitting on 'time in study' (time since entry)

```
. use http://www.bioepi.org/teaching/sa/diet
. stset dox, id(id) failure(chd) origin(doe) ent(doe) sc(365.25)
. list id  _t0 _t _d _st if id==78, clean
      id   _t0          _t   _d   _st
 28.  78     0   5.6180698    1    1
. stsplit timeband, at(0(2)20) trim
(0 + 4 obs. trimmed due to lower and upper bounds)
(2122 observations (episodes) created)
. list id  _t0 _t _d _st if id==78, clean
       id   _t0          _t   _d   _st
 189.  78     0           2    0    1
 190.  78     2           4    0    1
 191.  78     4   5.6180698    1    1
```



## Rates for different time bands

```
. strate timeband, per(1000)
```

| timeband | D | Y | Rate | Lower | Upper |
|---|---|---|---|---|---|
| 0 | 6 | 0.6658 | 9.01205 | 4.04876 | 20.05973 |
| 2 | 3 | 0.6499 | 4.61589 | 1.48872 | 14.31189 |
| 4 | 11 | 0.6187 | 17.77860 | 9.84579 | 32.10291 |
| 6 | 8 | 0.5947 | 13.45180 | 6.72721 | 26.89835 |
| 8 | 1 | 0.5670 | 1.76370 | 0.24844 | 12.52060 |
| 10 | 8 | 0.4919 | 16.26292 | 8.13305 | 32.51949 |
| 12 | 2 | 0.4148 | 4.82158 | 1.20586 | 19.27877 |
| 14 | 5 | 0.3619 | 13.81571 | 5.75048 | 33.19266 |
| 16 | 2 | 0.1778 | 11.24988 | 2.81357 | 44.98197 |
| 18 | 0 | 0.0610 | 0.00000 | . | . |



- Poisson regression can also be performed using the `streg` command. This is preferable when the data have been 'stsplit'.

```
. xi: streg hieng, dist(exp)
Exponential regression -- log relative-hazard form
No. of subjects =          337         Number of obs   =      2455
No. of failures =           46
Time at risk    =  4603.504449
                                        LR chi2(1)      =      4.82
Log likelihood  =   -175.00017          Prob > chi2     =    0.0282
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5203748   .1572099    -2.16   0.031     .2878463    .9407449
------------------------------------------------------------------------------
```

- The effect of hieng controlled for timeband is found with:

```
. xi: streg hieng i.timeband, dist(exp)
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5192307   .1568778    -2.17    .2871988    .9387245
_Itimeband_2 |    .513561   .3631324    -0.94    .1284453    2.053363
_Itimeband_4 |   1.994116   1.012076     1.36    .7374608    5.392148
_Itimeband_6 |   1.509825   .8154229     0.76    .5238557    4.351527
_Itimeband_8 |   .1977629   .2136155    -1.50    .0238074    1.642776
_Itimeban~10 |   1.808471   .9766678     1.10     .627507       5.212
_Itimeban~12 |   .5339343   .4359425    -0.77    .1077719    2.645272
_Itimeban~14 |   1.536125   .9301557     0.71    .4688202    5.033231
_Itimeban~16 |   1.261781   1.030245     0.28    .2546695    6.251594
_Itimeban~18 |   1.29e-06   .0015521    -0.01           0           .
------------------------------------------------------------------------------
```

- There is no reason to believe that time-on-study would be a confounder for these data. This would, however, be of interest in the cancer examples.



## Splitting the follow–up on the age scale

- Attained age is a possible confounder for the diet study. Attained age is more interesting as a potential confounder than age at entry.

```
. stset dox, fail(chd) entry(doe) origin(dob) sc(365.25) id(id)
. list id _t0 _t _d _st if id==163
 id         _t0          _t   _d   _st
163    47.55373   60.922656    1     1
. stsplit ageband, at(30,40,50,60,70) trim
. list id ageband _t0 _t _d _st if id==163
id   ageband       _t0          _t   _d   _st
163       40  47.55373          50    0     1
163       50        50          60    0     1
163       60        60   60.922656    1     1
```

- We see that, as expected, the CHD incidence rate depends on attained age.

```
. strate ageband, per(1000)
```



```
ageband    _D         _Y     _Rate
     30     0     0.0963     0.0000
     40     6     0.9070     6.6152
     50    18     2.1070     8.5428
     60    22     1.4933    14.7325
```



## The effect of hieng controlled for attained age

```
. xi: streg hieng i.ageband, dist(exp)
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |   .5370318   .1625245   .2967542    .9718588
_Iageband_40 |    4864030   1.36e+10          0           .
_Iageband_50 |    5963317   1.66e+10          0           .
_Iageband_60 |   1.03e+07   2.86e+10          0           .
------------------------------------------------------------------------------
```

- Poor choice of baseline for ageband!



- Let's use a different reference category.

```
. char ageband[omit] 40
. xi: streg hieng i.ageband, dist(exp)
------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|
-------------+----------------------------------------
       hieng |   .5370318   .1625245    -2.05   0.040
_Iageband_30 |   2.06e-07   .0005734    -0.01   0.996
_Iageband_50 |   1.226003   .5787785     0.43   0.666
_Iageband_60 |   2.109768    .973276     1.62   0.106
------------------------------------------------------
```

- Is there evidence that the effect of hieng is confounded by attained age?



## Assessing goodness-of-fit of Poisson regression models

- Since Poisson regression is a generalised linear model, methods of assessing goodness-of-fit of GLMs can be applied (many of which you've seen previously with logistic regression).
- For a GLM fitted to non-sparse data, model goodness-of-fit can be assessed using the deviance or the Pearson chi-square statistic, both of which have an approximate $\chi^2$ distribution under the assumption that the model fits, with degrees of freedom equal to the number of observations minus the number of parameters estimated in the model (including the intercept) [16].
- The deviance is the difference in twice the log likelihood between the fitted model and what is called the saturated model.
- The saturated model is the model which contains one parameter for every observation, such that the fitted values equal the observed values.



- For data which is cross-classified by $k$ categorical variables, as cancer registry data usually are, the saturated model contains all 2-way, 3-way, up to $k$-way interactions.
- As such, if a model is fitted containing all main effects, the deviance is essentially a test for interaction (where interaction is equivalent to non-proportional excess hazards).
- The asymptotic $\chi^2$ assumption for the deviance and the Pearson chi-square statistic is only valid for 'non-sparse' data.



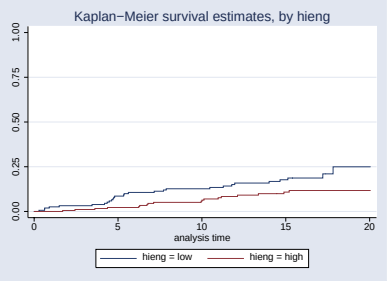
- A rule-of-thumb for chi-square based statistics of agreement between observed and fitted values is that both the expected number of successes and the expected number of failures must be 5 or more in at least 80% of the cells and at least 1 in each cell.
- In practice, individual-level data should be grouped.
- The exact distributions of the deviance and the Pearson chi-square statistic are not known, and there is no agreement in the literature regarding which is the best measure of goodness-of-fit.
- However, the two statistics should be similar for a model that provides a good fit to the data, and a large discrepancy between the two statistics is generally indicative of sparse data.
- When data are sparse, we typically see a deviance less that the degrees of freedom and a Pearson chi-square much greater than the degrees of freedom.

- Values of the deviance and Pearson chi-square significantly greater than the associated degrees of freedom can be due to a number of factors, including
  1. an incorrectly specified functional form (an additive rather than a multiplicative model may be appropriate);
  2. overdispersion; or
  3. the absence of important explanatory variables (or interactions) from the model.
- In most cases, lack-of-fit is due to missing explanatory variables (or interactions) from the model.
- Model goodness-of-fit can also be assessed using plots of residuals and influence statistics.



## Plotting estimates of the incidence proportion by hieng

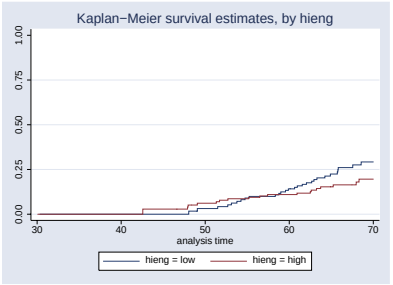
```
. stset dox, fail(chd) origin(doe) scale(365.25)
. sts graph, failure by(hieng)
```






## Incidence proportion by hieng with age as the timescale

```
. stset dox, fail(chd) entry(doe) origin(dob) scale(365.25)
. sts graph, failure by(hieng) noorigin
```






## Log-rank test

```
. sts test hieng
Log-rank test for equality of survivor functions
      |   Events         Events
hieng |  observed       expected
------+-------------------------
low   |        28          21.09
high  |        18          24.91
------+-------------------------
Total |        46          46.00
           chi2(1) =       4.20
           Pr>chi2 =     0.0403
```



- This is a test of equality of CHD mortality rates between individuals with a high and low energy intake, adjusted for attained age, and assuming that the mortality rates in the two groups are proportional with respect to attained age.
- That is, it is a very similar test to that performed in the framework of Poisson regression on slide 148.
- The P-value for the test of the effect of `hieng` in the Poisson regression model was 0.040 (very similar to the P-value above).
- A slight difference is that attained age was categorised in the Poisson regression model and the rate assumed to be constant within each category.



## Statistical models

- Multiple regression models are important in that they allow simultaneous estimation and testing of the effect of many prognostic factors on survival.
- The aim of statistical modelling is to derive a mathematical representation of the relationship between an observed response variable and a number of explanatory variables, together with a measure of the uncertainty of any such relationship.
- The uses of a statistical model can be classified into the following three areas:
  1. Descriptive: To describe any structure in the data and quantify the effect of explanatory variables, and to study the pattern of any such associations;
  2. Hypothesis testing: To statistically test whether an observed response variable is associated with one or more explanatory variables; and
  3. Prediction: For example, predicting excess mortality for a future time period, or predicting the way in which the outcome may change if certain explanatory variables changed in value.



- Note that a statistical model is never true, but may be useful.
- When making inference based on the model we assume that the model is true.
- If the model is badly misspecified then inference will be erroneous.
- It is therefore important to consider the validity of any assumptions (e.g. proportional hazards) underlying the model and to check for evidence of lack-of-fit.



## A mathematical framework for survival analysis

- It is assumed that the survival times of the source population are described by a probability distribution (which may be parametric or nonparametric).
- Parametric distributions, such as the normal or binomial distributions, can be expressed as a mathematical function of the parameters.
- For example, the binomial probability distribution is a function of the parameters $n$ and $p$.

$$\Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \text{ for } r = 0, 1, 2, 3, \ldots n.$$

- Parametric distributions used to describe survival times include the exponential and Weibull (not the normal distribution).
- A nonparametric distribution is one which cannot be expressed as a mathematical function.

- The probability density function is denoted by $f(t)$ and the corresponding cumulative distribution function $F(t)$.
- Quantities which follow a probability distribution are called random variables (as opposed to fixed constants).
- The actual survival time for an individual is denoted by $t$ and is assumed to be a realisation of the random variable $T$.
- The survivor function, $S(t)$, specifies the probability that the random variable $T$ exceeds the specified time $t$.

$$S(t) = \Pr(T > t) = 1 - F(t). \tag{5}$$



## The hazard function, $\lambda(t)$

- The hazard function, $\lambda(t)$, is simply the name used in survival analysis for what epidemiologists call the incidence rate (when the event of interest is disease incidence).
- The hazard function is formally defined by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{6}$$

- The hazard function is sometimes called the hazard rate or hazard and can be interpreted as the instantaneous event rate at time $t$, conditional on survival up to time $t$.
- From Equation 6, one can see that $\lambda(t)\Delta t$ may be viewed as the 'approximate' probability of an individual who is alive at time $t$ experiencing the event in the next instant, $\Delta t$.



- The units are events per unit time.
- In contrast to the survivor function, which describes the probability of *not* failing before time $t$, the hazard function focuses on the failure rate at time $t$ among those individuals who are alive at time $t$.
- That is, a lower value for $\lambda(t)$ implies a higher value for $S(t)$ and vice-versa.
- Note that the hazard is a rate, not a probability, so $\lambda(t)$ can take on any value between zero and infinity, as opposed to $S(t)$ which is restricted to the interval $[0, 1]$.



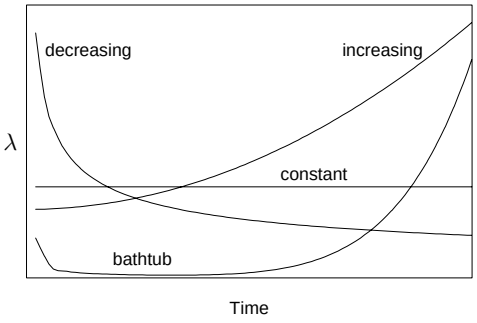
- Each of the functions $f(t)$, $F(t)$, $S(t)$, and $\lambda(t)$ can be obtained mathematically from any one of the other functions.
- Consequently assuming a distribution for $f(t)$ implies assuming a shape for $\lambda(t)$.
- One relationship of particular importance is

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \lambda(s)\,\mathrm{d}s\right] \\ &= \exp(-\Lambda(t)), \end{aligned} \tag{7}$$

where $\Lambda(t)$ is called the cumulative hazard (or integrated hazard) at time $t$.



## Common forms for the hazard function






- A bathtub-shaped hazard is appropriate in most human populations followed from birth, where the hazard rate decreases to almost zero after an initial period of infant mortality, and then starts to increase again later in life.
- A decreasing hazard function is appropriate following the diagnosis of most types of cancer, where mortality due to the cancer is highest immediately following diagnosis, and then decreases with time as patients are cured of the cancer.
- A constant hazard function is often used for modelling the lifetime of electronic components, but is also appropriate following the diagnosis of some types of cancer, most notably cancers of the breast and prostate, where the level of excess mortality due to the cancer is relatively constant over time and persists even 15-20 years after diagnosis.
- A constant hazard function implies that survival times can be described by an exponential distribution (which has one parameter, the hazard $\lambda$). This distribution is 'memoryless' in that the expected survival time for any individual is independent of how long the individual has survived so far.



- The average time to winning a prize for a regular lotto player, for example, can be described by an exponential distribution.
- An exponential distribution has also been used to model the time between goals in hockey [17, 18].
- The survivor function has the same basic shape (a nonincreasing function from 1 to 0) for all types of data and the hazard function is often a more informative means of studying differences between patient groups.



## Parametric models

- If we assume that survival times follow an exponential distribution, we could model the hazard as a function of one or more covariates.
- We could then obtain an estimate of the hazard ratio for the treatment group compared to the control group while adjusting for other explanatory variables.
- The disadvantage of this method is that assuming an exponential distribution for survival times implies the assumption of a constant hazard function over time, which may not be appropriate.
- The Weibull distribution, which has two parameters, is a more flexible distribution in which the hazard can be either monotonic increasing, decreasing, or constant.
- The Weibull and the Gompertz distributions have proved to be applicable in several types of medical survival studies.

- If a parametric distribution is appropriate, such models will result in more efficient estimates (narrower confidence limits) of the parameters of interest.
- Most common statistical procedures are parametric, for example, t-tests, ANOVA, and linear regression all assume normal distributions.
- Inference based on the above procedures is, however, quite robust to violations of the distributional assumptions. For example, application of a standard t-test will generally lead to the correct conclusion even if the two samples are not drawn from populations with normal distributions.
- This is not necessarily the case when assuming a parametric distribution for survival time. The assumption of an inappropriate distribution can result in erroneous conclusions.
- That is, when using parametric survival models, special attention must be paid to testing the appropriateness of the model.



## The Cox proportional hazards model

- The most commonly applied model in medical time-to-event studies is the Cox proportional hazards model [19].
- The Cox proportional hazards model does not make any assumption about the shape of the underlying hazards, but makes the assumption that the hazards for patient subgroups are proportional over follow-up time.
- We are usually more interested in studying how survival varies as a function of explanatory variables rather than the shape of the underlying hazard function.
- In most statistical models in epidemiology (e.g. linear regression, logistic regression, Poisson regression) the outcome variable (or a transformation of the outcome variable) is equated to the 'linear predictor', $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.
- $X_1, \ldots, X_k$ are explanatory variables and $\beta_0, \ldots, \beta_k$ are regression coefficients (parameters) to be estimated.



- The $X$s can be continuous (age, blood pressure, etc.) or if we have categorical predictor variables we can create a series of indicator variables ($X$s with values 1 or 0) to represent each category.
- We are interested in modelling the hazard function, $\lambda(t; \boldsymbol{X})$, for an individual with covariate vector $\boldsymbol{X}$, where $\boldsymbol{X}$ represents $X_1, \ldots, X_k$.
- The hazard function should be non-negative for all $t > 0$; thus, using

$$\lambda(t; \boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

may be inappropriate since we cannot guarantee that the linear predictor is always non-negative for all choices of $X_1, \ldots, X_k$ and $\beta_0, \ldots, \beta_k$.



- However, $\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$ is always positive so another option would be

$$\log \lambda(t; \boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- In this formulation, both the left and right hand side of the equation can assume any value, positive or negative.
- This formulation is identical to the Poisson regression model. That is,

$$\log \frac{\text{no. events}}{\text{person-time}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- The one flaw in this potential model is that $\lambda(t; \boldsymbol{X})$ is a function of $t$, whereas the right hand side will have a constant value once the values of the $\beta$s and $X$s are known.
- This does not cause any mathematical problems, although experience has shown that a constant hazard rate is unrealistic in most practical situations.



- The remedy is to replace $\beta_0$, the 'intercept' in the linear predictor, by an arbitrary function of time — say $\log \lambda_0(t)$; thus, the resulting model equation is

$$\log \lambda(t; \boldsymbol{X}) = \log \lambda_0(t) + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- The arbitrary function, $\lambda_0(t)$, is evidently equal to the hazard rate, $\lambda(t; \boldsymbol{X})$, when the value of $\boldsymbol{X}$ is zero, i.e., when $X_1 = \cdots = X_k = 0$.
- The model is often written as

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{X}\beta).$$

- It is not important that an individual having all values of the explanatory variables equal to zero be realistic; rather, $\lambda_0(t)$ represents a reference point that depends on time, just as $\beta_0$ denotes an arbitrary reference point in other types of regression models.



- This regression model for the hazard rate was first introduced by Cox [19], and is frequently referred to as the Cox regression model, the Cox proportional hazards model, or simply the Cox model.
- Estimates of $\beta_1, \ldots, \beta_k$ are obtained using the method of maximum partial likelihood (slide 201).
- As in all other regression models, if a particular regression coefficient, say $\beta_j$, is zero, then the corresponding explanatory variable, $X_j$, is not associated with the hazard rate of the response of interest; in that case, we may wish to omit $X_j$ from any final model for the observed data.
- As with logistic regression and Poisson regression, the statistical significance of explanatory variables is assessed using Wald tests or, preferably, likelihood ratio tests.
- The Wald test is an approximation to the likelihood ratio test. The likelihood is approximated by a quadratic function, an approximation which is generally quite good when the model fits.



- In most situations, the test statistics will be similar.
- Differences between these three test statistics are indicative of possible problems with the fit of the model.
- The assumption of proportional hazards is a strong assumption, and should be tested (see slide 204).
- Because of the inter-relationship between the hazard function, $\lambda(t)$, and the survivor function, $S(t)$, (Equation 7, slide 163) we can show that the PH regression model is equivalent to specifying that

$$S(t; \boldsymbol{X}) = \{S_0(t)\}^{\exp(\beta_1 X_1 + \cdots + \beta_k X_k)}, \tag{8}$$

where $S(t; \boldsymbol{X})$ denotes the survivor function for a subject with explanatory variables $\boldsymbol{X}$, and $S_0(t)$ is the corresponding survivor function for an individual with all covariate values equal to zero.



- Most software packages, will provide estimates of $S(t)$ based on the fitted proportional hazards model for any specified values of explanatory variables.
- For example, the Stata `stcurve` can be used after `stcox` to plot the cumulative hazard, survival, and hazard functions at the mean value of the covariates or at values specified by the at() options.

## Interpreting the Estimated Regression Coefficients

- Recall that the basic PH regression model specifies

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\beta_1 X_1 + \cdots + \beta_k X_k) \; ;$$

equivalently,

$$\log \lambda(t; \boldsymbol{X}) = \log \lambda_0(t) + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- Note the similarity to the basic equation for multiple linear regression, i.e.,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- In ordinary regression we derive estimates of all the regression coefficients, i.e., $\beta_1, \ldots, \beta_k$ and $\beta_0$.



- In PH regression, the baseline hazard component, $\lambda_0(t)$, vanishes from the partial likelihood; we only obtain estimates of the regression coefficients associated with the explanatory variates $X_1, \ldots, X_k$.

- Consider the simplest possible setup, one involving only a single binary variable, $X$; then the PH regression model is

$$\log \lambda(t; X) = \log \lambda_0(t) + \beta X \; ,$$

or equivalently,

$$\begin{aligned} \beta X &= \log \lambda(t; X) - \log \lambda_0(t) \\ &= \log \{\lambda(t; X)/\lambda_0(t)\} \; . \end{aligned}$$

- Since $\lambda_0(t)$ corresponds to the value $X = 0$,

$$\beta = \log \{\lambda(t; X = 1)/\lambda_0(t)\} \; .$$



- That is, $\beta$ is the logarithm of the ratio of the hazard rate for subjects belonging to the group denoted by $X = 1$ to the hazard function for subjects belonging to the group indicated by $X = 0$.

- The parameter $\beta$ is a log relative risk and $\exp(\beta)$ is a relative risk of response; PH regression is sometimes called "relative risk regression".

- If we conclude that the data provide reasonable evidence to contradict the hypothesis that $X$ is unrelated to response, $\exp(\hat{\beta})$ is a point estimate of the rate at which response occurs in the group denoted by $X = 1$ relative to the rate at which response occurs at the same time in the group denoted by $X = 0$.

- A confidence interval for $\beta$, given by $\hat{\beta} \pm 1.96\text{SE}$, represents a range of plausible values for the log relative risk associated with the corresponding explanatory variable.



- Corresponding confidence intervals for the relative risk associated with the same covariate are obtained by transforming the confidence interval for $\beta$, i.e.,

$$(\beta_\ell, \beta_u) \Rightarrow \left(e^{\beta_\ell}, e^{\beta_u}\right) .$$

- When more than one covariate is involved, the principle is the same; $\exp(\hat{\beta}_j)$ is the estimated relative risk of failure for subjects that differ only with respect to the covariate $X_j$.

- If $X_j$ is binary, $\exp(\hat{\beta}_j)$ estimates the increased/reduced risk of response for subjects corresponding to $X_j = 1$ versus those denoted by $X_j = 0$.

- When $X_j$ is a numerical measurement then $\exp(\hat{\beta}_j)$ represents the estimated change in relative risk associated with a unit change in $X_j$.

- Since the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are obtained simultaneously, these estimated relative risks adjust for the effect of all the remaining covariates included in the fitted model.



## Example: Localised colon carcinoma 1975–1994

- We fitted a proportional hazards model to study the effect of sex, age (in 4 categories), and calendar period (2 categories) on cause-specific mortality (only deaths due to colon cancer were considered events).

- We'll begin by restricting the data to localised cases only.

```
.  use http://www.bioepi.org/teaching/sa/colon, clear
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. keep if stage==1
(9290 observations deleted)
```



- We `stset` the data where only deaths due to colon cancer (status=1) are considered 'failures'.

```
. stset surv_mm, failure(status==1)
     failure event:  status == 1
obs. time interval:  (0, surv_mm]
 exit on or before:  failure
------------------------------------------------------------------------------
     6274  total obs.
        0  exclusions
------------------------------------------------------------------------------
     6274  obs. remaining, representing
     1734  failures in single record/single failure data
 424049.7  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =       251
```

- Now we estimate the Cox model.



```
. xi: stcox sex i.agegrp year8594
i.agegrp          _Iagegrp_0-3     (naturally coded; _Iagegrp_0 omitted)
         failure _d:  status == 1
   analysis time _t:  surv_mm
Cox regression -- Breslow method for ties
No. of subjects =         6274           Number of obs   =      6274
No. of failures =         1734
Time at risk    =    424049.72
                                         LR chi2(5)      =    197.23
Log likelihood  =   -14348.889           Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9151101   .0451776    -1.80   0.072     .8307126    1.008082
_Iagegrp_1 |   .9491689   .1314101    -0.38   0.706      .723597     1.24506
_Iagegrp_2 |   1.338501   .1682956     2.32   0.020     1.046148    1.712553
_Iagegrp_3 |    2.24848   .2834768     6.43   0.000     1.756199    2.878751
  year8594 |   .7548672   .0372669    -5.70   0.000     .6852479    .8315596
------------------------------------------------------------------------------
```

- The output commences with a description of the outcome and censoring



variable and a summary of the number of subjects and number of failures.

- The default method for handling ties (the Breslow method) is used.

- The test statistic `LR chi2(5) = 197.23` is not especially informative. The interpretation is that the 5 parameters in the model (as a group) are statistically significantly associated with the outcome ($P < 0.00005$).

- The variable sex is coded as 1 for males and 2 for females. Since each parameter represents the effect of a one unit increase in the corresponding variable, the estimated hazard ratio for sex represents the ratio of the hazards for females compared to males.

- That is, the estimated hazard ratio is 0.92 indicating that females have an estimated 8% lower colon cancer mortality than males. There is some evidence that the difference is statistically significant ($P = 0.07$).

- The model assumes that the estimated hazard ratio of 0.92 is the same at

each and every point during follow-up and for all combinations of the other covariates.

- That is, the hazard ratio is the same for females diagnosed in 1975–1984 aged 0–44 (compared to males diagnosed in 1975–1984 aged 0–44) as it is for females diagnosed in 1985–1994 aged 75+ (compared to males diagnosed in 1985–1994 aged 75+).
- The indicator variable year8594 has the value 1 for patients diagnosed during 1985–1994 and 0 for patients diagnosed during 1975–1984.
- The estimated hazard ratio is 0.75. We estimate that, after controlling for age and sex, patients diagnosed 1985–1994 have a 25% lower mortality than patients diagnosed during 1975–1984. The difference is statistically significant ($P < 0.0005$).
- We chose to group age at diagnosis into four categories; 0–44, 45–59, 60–74, and 75+ years.



- It is estimated that individuals aged 75+ at diagnosis experience 2.25 times higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant ($P < 0.0005$).
- Similarly, individuals aged 60–74 at diagnosis have an estimated 34% higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant ($P < 0.02$).



- These significance tests test the pairwise differences and tell us little about the overall association between age and survival – we need to perform a general test.

```
. test  _Iagegrp_1 _Iagegrp_2 _Iagegrp_3
 ( 1)  _Iagegrp_1 = 0
 ( 2)  _Iagegrp_2 = 0
 ( 3)  _Iagegrp_3 = 0
           chi2(  3) =  174.13
         Prob > chi2 =    0.0000
```

- This is a Wald test of the null hypothesis that all age parameters are equal to zero, i.e. that age is not associated with the outcome.
- We see that there is strong evidence against the null hypothesis, i.e. we conclude that age is significantly associated with survival time.
- To perform a likelihood ratio test we fit the reduced model (the model without age) and see that the log likelihood is $-14436.387$.



```
. xi: stcox sex year8594
No. of subjects =         6274          Number of obs   =      6274
No. of failures =         1734
Time at risk    =    424049.72
                                        LR chi2(2)      =     22.23
Log likelihood  =   -14436.387          Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9978866   .0487896    -0.04   0.965     .9066997    1.098244
    year8594 |     .79287   .0390053    -4.72   0.000     .7199909    .8731261
------------------------------------------------------------------------------
```

- The log likelihood for the model containing age is $-14348.889$; for the model excluding age it is $-14436.387$.
- The likelihood ratio test statistic for the association of age with survival is calculated as $2 \times (-14348.889 - (-14436.387)) = 175.0$, which is compared to a $\chi^2$ distribution with 3 degrees of freedom (P=0.0001).



- We see that the Wald test statistic ($174.1$) is very similar in value to the likelihood ratio test statistic ($175.0$).



- You can also get Stata to calculate the likelihood ratio test statistic for you (you have to explicitly fit both models and save the estimates for the first).

```
xi: stcox sex i.agegrp year8594
est store A
xi: stcox sex year8594
lrtest A
```

- The output of the final command is as follows

```
. lrtest A
likelihood-ratio test           LR chi2(3)  =    175.00
(Assumption: . nested in A)     Prob > chi2 =    0.0000
```

- We might choose to model age as a numeric variable.



```
. stcox sex age year8594
No. of subjects =         6274          Number of obs   =      6274
No. of failures =         1734
Time at risk    =    424049.72
                                        LR chi2(3)      =    248.82
Log likelihood  =    -14323.09          Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9029577   .0445694    -2.07   0.039     .8196956    .9946773
         age |   1.033981   .0024182    14.29   0.000     1.029253    1.038732
    year8594 |   .7488609   .0369492    -5.86   0.000     .6798334    .8248973
------------------------------------------------------------------------------
```

- For each (and every) one year increase in age at diagnosis, we estimate that mortality is 3.4% higher.
- For a 10-year increase in age at diagnosis the estimated hazard ratio is $1.033981^{10} = 1.396$.



## An introduction to likelihood inference

- The aim of statistical inference is to estimate population parameters of interest from observed data. For example,
  - Estimate the hazard ratio for exposed/unexposed in a study of cancer patient survival. The parameter of interest is the log hazard ratio $\beta$ in the population from which the sample is drawn. This parameter is estimated using the sample of patients observed.
  - Similar with logistic regression (the parameter of interest is the log odds ratio $\beta$).
  - Estimate the recombination fraction, $\theta$, in parametric linkage analysis from the observed pedigrees (marker genotypes and phenotypes).
- A simple example: Imagine we are interested in estimating the proportion, $p$, that a toss of a coin will result in heads.
- We toss the coin 10 times and observe 4 heads.

- We wish to estimate the parameter of interest, $p$, from the observed data (the 10 tosses of the coin). Issues of interest are
  - What is the most likely value for $p$?
  - What is a range of likely values for $p$?
  - Is $p = 0.5$ a plausible value (c.f., linkage analysis)?
- The likelihood approach is to calculate the probability of observing the observed data, given the probability model, for all possible values of the parameter(s) of interest and choosing the values of the parameter(s) that make the data most likely.
- That is, for what value of $p$ is the probability of tossing 4/10 heads most likely?
- We will calculate the probability of observing 4 heads in 10 tosses for a range of possible values of $p$.



- If the true value is $p = 0$, what is the probability of observing 4 heads in 10 tosses?
- That one was easy (the probability is zero), but what if $p = 0.1$?
- If $p = 0.1$ then the number of observed heads can theoretically be any integer between 0 and 10 and the probability of each is described by the binomial distribution.
- Recall that if $X$ is a random variable described by a binomial distribution with parameters $n$ and $p$ then the probability distribution of $X$ is given by

$$\Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \text{ for } r = 0, 1, 2, 3, \ldots n.$$

- $\Pr(X = r)$ is the probability of obtaining $r$ 'successes' (e.g., toss heads) in a sample of size $n$ where the true proportion is $p$.



- For $p = 0.1$ and $n = 10$ the probability of observing each of the possible outcomes is as follows.

| $r$ | Prob($r$ heads) |
|---|---|
| 0 | 0.35 |
| 1 | 0.39 |
| 2 | 0.19 |
| 3 | 0.06 |
| 4 | 0.01 |
| 5 | 0.00 |
| 6 | 0.00 |
| 7 | 0.00 |
| 8 | 0.00 |
| 9 | 0.00 |
| 10 | 0.00 |
| $\Sigma$ | 1.00 |



## Binomial distribution with $n = 10$ for various values of $p$

| | Assumed value of $p$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $r$ | 0.00 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 |
| 0 | 1.00 | 0.35 | 0.11 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1 | 0.00 | 0.39 | 0.27 | 0.12 | 0.04 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2 | 0.00 | 0.19 | 0.30 | 0.23 | 0.12 | 0.04 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3 | 0.00 | 0.06 | 0.20 | 0.27 | 0.21 | 0.12 | 0.04 | 0.01 | 0.00 | 0.00 | 0.00 |
| 4 | 0.00 | 0.01 | 0.09 | 0.20 | 0.25 | 0.21 | 0.11 | 0.04 | 0.01 | 0.00 | 0.00 |
| 5 | 0.00 | 0.00 | 0.03 | 0.10 | 0.20 | 0.25 | 0.20 | 0.10 | 0.03 | 0.00 | 0.00 |
| 6 | 0.00 | 0.00 | 0.01 | 0.04 | 0.11 | 0.21 | 0.25 | 0.20 | 0.09 | 0.01 | 0.00 |
| 7 | 0.00 | 0.00 | 0.00 | 0.01 | 0.04 | 0.12 | 0.21 | 0.27 | 0.20 | 0.06 | 0.00 |
| 8 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.04 | 0.12 | 0.23 | 0.30 | 0.19 | 0.00 |
| 9 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.04 | 0.12 | 0.27 | 0.39 | 0.00 |
| 10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.11 | 0.35 | 1.00 |
| $\Sigma$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |



## 'Likelihood' for a range of values of $p$

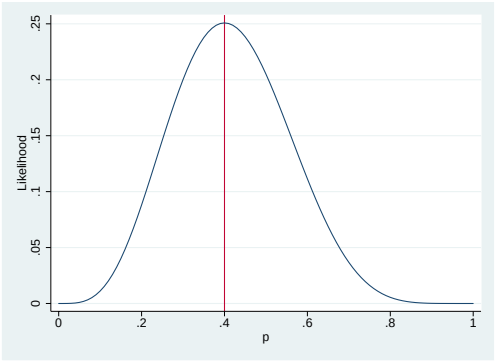
| $p$ | Prob($r = 4$) |
|---|---|
| 0.00 | 0.00 |
| 0.10 | 0.01 |
| 0.20 | 0.09 |
| 0.30 | 0.20 |
| 0.40 | 0.25 |
| 0.50 | 0.21 |
| 0.60 | 0.11 |
| 0.70 | 0.04 |
| 0.80 | 0.01 |
| 0.90 | 0.00 |
| 1.00 | 0.00 |

- This is the likelihood function[2]. The value of $p$ for which the likelihood is greatest is $p = 0.4$. This is called the maximum likelihood estimate.

[2] In practice, we would exclude the constant $\frac{10!}{4!(10-4)!}$ from the likelihood function.



## Plot of the binomial likelihood






## What are other likely values for $p$

- We can see that $p = 0.5$ is also quite likely. The probability of the data is 0.21 when $p = 0.5$ compared to a probability of 0.25 when $p = 0.4$ (the MLE).
- We can test whether $p = 0.5$ is a likely value by studying the ratio of the likelihoods.

$$L(0.5)/L(0.4) = 0.21/0.25 = 0.8176$$

- A result in mathematical statistics tells us that, if the true value of $p$ was 0.5, then minus twice the log likelihood ratio will have a chi square distribution with 1 degree of freedom.

$$-2\ln[L(0.5)/L(0.4)] = -2[l(0.5) - l(0.4)] = 0.40279$$

where $l$ is the log likelihood (the natural logarithm of the likelihood).



```
. di chi2tail(1,0.403)
.52554398
```

- We see that, if the true value of $p$ was 0.5, then we would observe a test statistic at least as large as that we observed 53% of the time. That is, we cannot reject the hypothesis that the true value of $p$ is 0.5.

## Mathematically

- We wish to find the value of $p$ that maximises the likelihood function

$$L(p) = \frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}, \text{ for } r = 0, 1, 2, 3, \ldots n.$$

- It is generally easier to maximise the log likelihood (the maximum will occur at the same value). Ignoring the constant,

$$l(p) = \ln[L(p)] = r\ln(p) + (n-r)\ln(1-p).$$

- The derivative of $l(p)$ wrt $p$ is $l'(p) = r/p - (n-r)/(1-p)$.
- The maximum value of $l(p)$ will occur when $l'(p) = 0$ which $\hat{p} = r/n$.



## Likelihood calculations for the Cox model

- Estimation is based on the concept of *risk sets*.
- The risk set at each failure time is the collection of subjects who were at risk of failing at that time.
- In theory, only one individual can fail at each failure time and we can calculate the conditional probability of failure for the subject who actually failed.
- The likelihood function is the product of these conditional probabilities of failure.
- Imagine 5 individuals at risk at time $t$ of which one fails.
- These individuals have hazards $\lambda_1, \lambda_2, \ldots, \lambda_5$ which may be different since the individuals have different covariate values.



- Conditional on one of the five failing, the probability it is number 2 is

$$\frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}$$

- Since $\lambda(t) = \lambda_0(t)\exp(x\beta)$ we can write this as

$$\frac{\lambda_0(t)\exp(x_2\beta)}{\lambda_0(t)\exp(x_1\beta) + \lambda_0(t)\exp(x_2\beta) + \ldots + \lambda_0(t)\exp(x_5\beta)}$$

- The baseline hazard, $\lambda_0(t)$, cancels and we have

$$\frac{\exp(x_2\beta)}{\sum_{i\in R}\exp(x_i\beta)}$$

  where $R$ represents the risk set.

- The likelihood function is the product of these conditional probabilities.



- If we have $k$ distinct failure times then

$$L(\beta) = \prod_{j=1}^{k}\left[\frac{\exp(x_j\beta)}{\sum_{i\in R_j}\exp(x_i\beta)}\right] \quad (9)$$

- Note that these calculations do not depend on the underlying failure times; only the ordering of failure times is important.
- Although this is not a likelihood in the strict sense, it is a partial likelihood, it can for all intents and purposes be treated as a likelihood.
- In practice we often observe multiple failures at the same time (ties) and need to use an approximation to equation 9.



## Assessing the appropriateness of the proportional hazards assumption

- The proportional hazards assumption is a strong assumption and its appropriateness should always be assessed.
- The model assumes that the *ratio* of the hazard functions for any two patient subgroups (i.e. two groups with different values of the explanatory variable $\boldsymbol{X}$) is constant over follow-up time.
- Note that it is the hazard ratio which is assumed to be constant. The hazard can vary freely with time.
- When comparing an aggressive therapy vs a conservative therapy, for example, it is not unusual that the patients receiving the aggressive therapy do worse earlier, but then have a lower hazard (i.e. better survival) than those receiving the conservative therapy.



- In this situation, the ratio of the hazard functions will not be constant over time, as is assumed by the PH model.
- Figure 3 (slide 25) shows an example of non-proportional hazards, although this may not be obvious to the untrained eye; it is difficult to assess the PH assumption by looking at the estimates of the survivor function.
- If the hazard functions cross, it is possible that the effect of treatment will not be statistically significant despite the presence of a clinically interesting effect.
- As such, it is important to plot survival curves before fitting the model and to assess the appropriateness of the proportional hazards assumption of the proportional hazards assumption after the model has been fitted.
- Note that the hazard functions do not have to cross for the PH assumption to be violated. For example, a hazard ratio of 4 which gradually decreases with time to a value of 1.5 is an example of non-proportional hazards.



- Hess (1995) [20] reviews methods for assessing the appropriateness of the proportional hazards assumption.
- Therneau & Grambsch [21] give a more up-to-date review and include code for implementing the various methods in SAS and S-PLUS.
- Following is a list of commonly used methods for assessing the appropriateness of the proportional hazards assumption (in increasing order of utility):

  1. Plotting the cumulative survivor functions and checking that they do not cross. This method is not recommended, since the survivor functions do not have to cross for the hazards to be non-proportional (e.g. Figure 3).
  2. Plotting the log cumulative hazard functions over time and checking for parallelism.
  3. Including time-by-covariate interaction terms in the model and testing statistical significance. For example, a statistically significant time-by-exposure term would indicate a trend in the hazard ratio with time.
  4. Plotting Schoenfeld's residuals against time to identify patterns.



- The first two methods do not allow for the effect of other covariates, whereas the second two methods do.
- Including a time-by-covariate interaction in the model has the advantage that we obtain an estimate of the hazard ratio as a function of time.

## Plots of the log cumulative hazard function

- Recall from Equation 8 (slide 174) that

$$S(t;\boldsymbol{X}) = \{S_0(t)\}^{\exp(\beta_1 X_1 + \cdots + \beta_k X_k)}.$$

- Consider the situation where we have only a single binary variable, $X$, then

$$S(t;X=1) = \{S(t;X=0)\}^r,$$

where $r = \exp(\beta)$ is the hazard ratio.

- Taking natural logarithms of both sides gives

$$\log S(t;X=1) = r\log\{S(t;X=0)\}.$$

- Taking natural logarithms of the negatives of both sides gives

$$\log[-\log S(t;X=1)] = \log r + \log[-\log\{S(t;X=0)\}].$$



- Consequently, if the proportional hazards model is appropriate, plots of $\log[-\log S(t)]$ vs $t$ for each group will be parallel, with the constant difference between them equal to $\log r$, which is the coefficient $\beta$.
- From Equation 7 (slide 163), we see that $-\log S(t)$ is equivalent to the cumulative hazard function, $\Delta(t)$.
- Consequently, plots of $\log[-\log S(t)]$ are often called log cumulative hazard plots.
- Figure 7 was constructed using the following command.

```
stphplot, by(year8594)
```

- The estimated regression coefficient for calendar period is $\ln(0.755) = -0.28$, so we would expect a constant difference of approximately 0.28 between the curves.



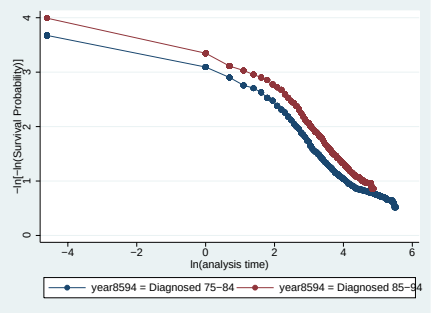



Figure 7: Log cumulative hazard plot by calendar period for the localised colon carcinoma data



- This appears to be the case, except possibly for higher values of $\log(t)$.
- The proportional hazards assumption for calendar period appears to be appropriate.
- Note that the lines do not have to be straight, it is only necessary for there to be a constant difference between the lines.
- Plotting $\log(t)$ (as opposed to $t$) on the $x$ axis results in straighter lines and it is therefore easier to study whether the difference is constant.
- Note that Figure 7 is based on estimates made using the Kaplan-Meier method which, unlike the estimates from the Cox model, are not adjusted for age and sex.
- It is, however, possible to construct adjusted plots in Stata.



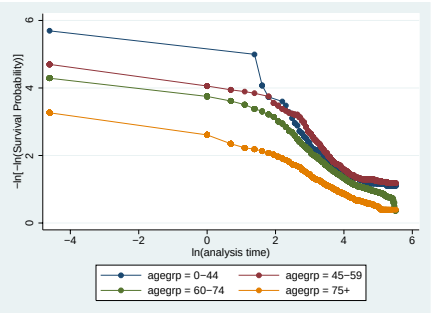



Figure 8: Log cumulative hazard plot by age for the localised colon carcinoma data, showing some evidence of non-proportional hazards



## Using time-varying covariates to assess the PH assumption

- If the effect of an exposure is modified by time then this can be modelled using what is often called a time-varying covariate.
- This is nothing more than an interaction between the exposure and the effect modifier, except the situation is slightly complicated when the effect modifier is time.
- In population-based cancer survival analysis, the values of explanatory variables are generally known at the start of follow-up and do not change over time.
- In other types of studies, however, the values of explanatory variables can change during follow-up. For example, blood pressure, occupational exposure to carcinogens.
- We may therefore wish to take account of this in the model by incorporating what are known as time-varying covariates into the Cox model.



- Another application is in observational studies where an intervention may occur at any point in the follow-up. At the time of the intervention, the explanatory variable associated with the intervention changes value from 0 (false) to 1 (true).
- Using a time-varying covariate for an explanatory variable implies that we have removed the assumption that the hazard ratio for that variable is constant with time.
- We can make use of time-varying covariates to test whether the hazard ratio for a fixed covariate is constant over time.
- Consider again a proportional hazards model with one single binary variable, $X_1$, which takes the value 1 if an exposure is present and 0 if it is absent

$$\lambda(t;\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1).$$

- The hazard ratio for exposed to unexposed is given by $\exp(\beta_1)$.



- We now construct a second variable, $X_2 = X_1 t$ and include this in the model, in addition to $X_1$. The variable $X_2$ takes the value $t$ if the exposure is present and 0 if it is absent

$$\lambda(t;\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_1 t).$$

- Based on this model, the hazard ratio for exposed to unexposed is given by $\exp(\beta_1 + \beta_2 t)$.
- An estimate for $\beta_2$ significantly different from 0 indicates that the hazard ratio is non-constant over time. $\beta_2 > 0$ indicates that the hazard ratio increases with time and $\beta_2 < 0$ indicates it decreases with time.
- This is not a general test of the proportional hazards assumption. It tests against the alternative that the hazard ratio changes monotonically with time.

- Another alternative might be that the hazard ratio is constant for an initial time period, say $t = 2$ years, but takes on a different (constant) value for the remainder of follow-up.
- To test against this alternative, we construct a variable $X_2$ which takes the value 1 if the exposure is present and $t > 2$ years, and 0 otherwise.
- In the resulting model containing the variables $X_1$ and $X_2$, the hazard ratio for exposed to unexposed for the period $t \leq 1$ year is given by $\exp(\beta_1)$ and for $t > 2$ years it is given by $\exp(\beta_1 + \beta_2)$.
- An estimate for $\beta_2$ significantly different from 0 indicates that the hazard ratio is different between the two time periods.
- We will now extend the model for the colon carcinoma data by including a term which allows different hazard ratios for calendar period before and after 2 years (24 months).



- This can be conveniently done using the `tvc()` and `texp()` options to `stcox`. The `nohr` option requests that estimates of the regression parameters, rather than the estimated hazard ratios, are reported.

```
. xi: stcox sex i.agegrp year8594, tvc(year8594) texp(_t >= 24) nohr
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rh           |
         sex |  -.0893258   .0493693    -1.81   0.070    -.1860879    .0074362
_Iagegrp_1 |  -.0519054   .1384465    -0.37   0.708    -.3232555    .2194448
_Iagegrp_2 |   .2903707   .1257308     2.31   0.021     .0439429    .5367986
_Iagegrp_3 |   .8110017   .1260577     6.43   0.000     .5639331     1.05807
  year8594 |  -.4206795   .0653073    -6.44   0.000    -.5486794   -.2926797
-------------+----------------------------------------------------------------
t            |
  year8594 |   .3212311   .0988321     3.25   0.001     .1275237    .5149385
------------------------------------------------------------------------------
```



- Now present the estimates as hazard ratios.

```
. xi: stcox sex i.agegrp year8594, tvc(year8594) texp(_t >= 24)
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rh           |
         sex |   .9145475   .0451506    -1.81   0.070     .8302006    1.007464
  _Iagegrp_1 |   .9494187   .1314437    -0.37   0.708     .7237889    1.245385
  _Iagegrp_2 |   1.336923   .1680924     2.31   0.021     1.044923    1.710522
  _Iagegrp_3 |   2.250161   .2836501     6.43   0.000     1.757572    2.880806
    year8594 |   .6566005   .0428808    -6.44   0.000     .5777122     .7462612
-------------+----------------------------------------------------------------
t            |
    year8594 |   1.378824   .1362721     3.25   0.001     1.136012    1.673536
------------------------------------------------------------------------------
Note: Second equation contains variables that continuously vary with respect
      to time; variables are interacted with current values of _t >= 24.
```



- The coefficient for the additional time-varying covariate represents the additional hazard experienced by patients diagnosed in 1985–94 during the period beyond 24 months after diagnosis.
- The time varying covariate was statistically significant in the model ($P = 0.001$).
- That is, the PH assumption was not appropriate for calendar period.
- The estimated hazard ratio, based on the above model, for patients diagnosed 1985–94 compared to 1975–84 is $\exp(-0.4207) = 0.657$ for the period up to 2 years of follow-up and $\exp(-0.4207 + 0.3212) = 0.905$ for the period after 2 years of follow-up.
- The estimated hazard ratio for the period after two years of follow-up can be obtained by multiplying the two hazard ratios, $0.657 \times 1.379 = 0.905$.



- The cutoff at 24 months was chosen arbitrarily. For the first 6 months of follow-up the estimated hazard ratio was 0.724, for the first year it was 0.676, and for the first two years it was 0.657.
- Choosing the cutpoint after inspection of the data will invalidate statistical inference (i.e. reported P-values will be too low).
- We have described two possible alternatives to proportional hazards. In practice, it is possible to fit any model of the form

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_1 f(t)),$$

  where $f(t)$ is a function of time.
- A similar approach to testing the PH assumption is to partition the time axis and fit separate models for different time periods.
- If the PH assumption is appropriate, the parameter estimates will be similar for each model.



- The problem is in how to partition the time axis, i.e., the choice of cutpoints.
- Possibilities are to choose cutpoints that lead to a similar number of events in each interval, or that lead to a similar number of observation times (events and censorings) in each interval.



## Tests of the PH assumption based on Schoenfeld residuals

- If the PH assumption holds then the Schoenfeld residuals (a diagnostic specific to the Cox model) should be independent of time.
- A test of the PH assumption can be made by modelling the Schoenfeld residuals as a function of time and testing the hypothesis of a zero slope.
- `stphtest` can be used after `stcox` to test the proportional hazard assumptions based on Schoenfeld residuals.
- Use of this command requires that you previously specified `stcox`'s `schoenfeld()` option (if the global test is desired) and/or `stcox`'s `scaledsch()` option (if the detailed test is desired).



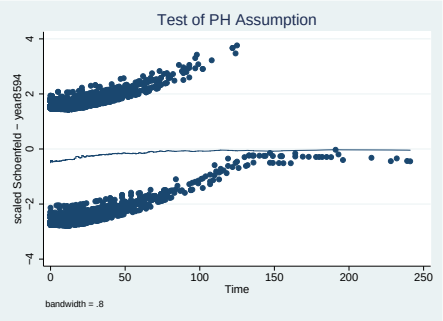
```
. xi: stcox sex i.agegrp year8594, schoenf(sc*) scaledsch(ssc*)
  [output suppressed]
. stphtest, detail
Test of proportional hazards assumption
------------------------------------------------
             |    rho       chi2    df  Prob>chi2
-------------+----------------------------------
sex          |   0.00840     0.12     1    0.7262
_Iagegrp_1   |   0.01367     0.32     1    0.5694
_Iagegrp_2   |   0.04178     3.05     1    0.0809
_Iagegrp_3   |  -0.00177     0.01     1    0.9412
year8594     |   0.07231     9.29     1    0.0023
-------------+----------------------------------
global test  |              27.72     5    0.0000
------------------------------------------------
```

- The tests suggest that there is evidence that the hazards are nonproportional by calendar period (and possibly age).
- Use `estat phtest, detail` in Stata version 9.

- Rather than just fitting a straight line to the residuals and testing the hypothesis of zero slope (as is done by `stphtest`) we can study a plot of the residuals along with a smoother to assist us in determining how the mean residual varies as a function of time.
- The smooth illustrates how the log hazard ratio varies as a function of time. We see, for example, that the effect of period is larger during the initial years of follow-up.

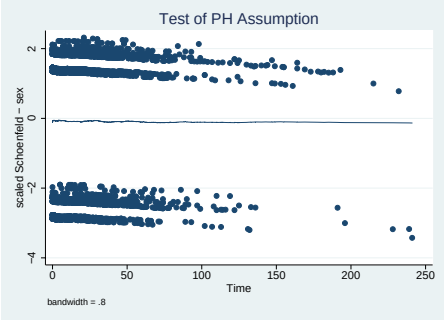


```
. estat phtest, plot(year8594)
```






```
. estat phtest, plot(sex)
```






## A model including stage

```
. use http://www.bioepi.org/teaching/sa/colon, clear
. drop if stage == 0 /* remove unknown stage */
. stset surv_mm, failure(status==1)
. xi: stcox sex i.agegrp i.stage year8594, sch(sc*) scaledsch(ssc*)
```

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| sex | .9559269 | .0232954 | -1.85 | 0.064 | .911342 | 1.002693 |
| _Iagegrp_1 | 1.087061 | .0693794 | 1.31 | 0.191 | .9592411 | 1.231913 |
| _Iagegrp_2 | 1.308011 | .0767528 | 4.58 | 0.000 | 1.165907 | 1.467436 |
| _Iagegrp_3 | 1.835699 | .1089947 | 10.23 | 0.000 | 1.634035 | 2.062252 |
| _Istage_2 | 2.300746 | .0945407 | 20.28 | 0.000 | 2.122715 | 2.493708 |
| _Istage_3 | 8.072185 | .2375035 | 70.98 | 0.000 | 7.619854 | 8.551367 |
| year8594 | .8601408 | .0206306 | -6.28 | 0.000 | .8206413 | .9015415 |



```
. stphtest, detail
      Test of proportional hazards assumption
      Time:  Time
      ----------------------------------------------------------------
                  |       rho           chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      sex         |    -0.00182          0.02        1          0.8773
      _Iagegrp_1  |    -0.00121          0.01        1          0.9180
      _Iagegrp_2  |     0.02014          2.92        1          0.0875
      _Iagegrp_3  |    -0.00743          0.39        1          0.5300
      _Istage_2   |    -0.04084         11.89        1          0.0006
      _Istage_3   |    -0.15971        168.36        1          0.0000
      year8594    |     0.02513          4.58        1          0.0323
      ------------+---------------------------------------------------
      global test |                    210.44        7          0.0000
      ----------------------------------------------------------------
```

- There is evidence that the hazards are heavily non-proportional by stage.
- A plot of the empirical hazards (slide 230) suggests that individuals diagnosed

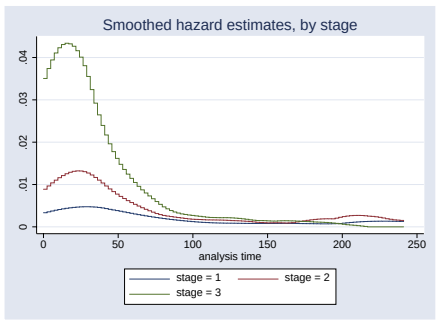


with distant metastases have proportionally much higher mortality early in the follow-up but once they have survived several years their mortality is not that much higher than the other age groups.

- The plots of the fitted hazards (slide 231) show the effect of the assumption of proportional hazards.

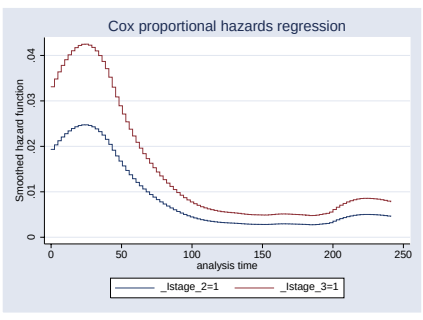


```
/* empirical hazards by stage */
. sts graph, hazard by(stage)
```

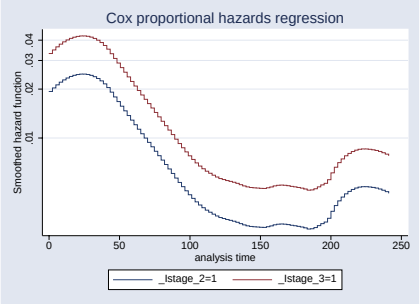





```
/* fitted hazards by stage */
. stcurve, hazard at1(_Istage_2=1) at2(_Istage_3=1)
```

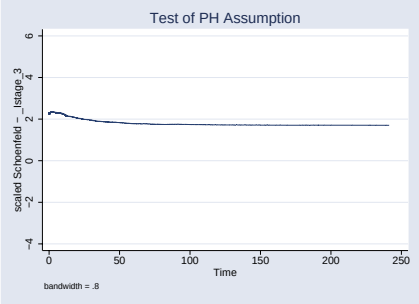


```
/* fitted hazards by stage (on log scale) */
. stcurve, hazard at1(_Istage_2=1) at2(_Istage_3=1) yscale(log)
```






```
/* Can also plot a smooth of the scaled Schoenfeld residuals */
. stphtest, plot(_Istage_3) s(i)
```






## The stratified Cox model

- The Cox model assumes that the baseline hazard (e.g., instantaneous mortality rate in the reference group) is an arbitrary function of time.
- The hazard functions for each of the other groups are assumed to be proportional to the baseline.
- It is possible to relax this assumption to allow separate baseline hazards for each level of, for example, age at diagnosis.
- This is known as a stratified proportional hazards model and is a useful method for modelling data where non-proportional hazards are suspected for a factor that is not of primary interest.
- A model stratified on agegrp is analogous to including an agegrp*time interaction in a Poisson regression model.
- Use the `strata()` option in Stata to specify up to 5 strata variables.



```
. use http://www.bioepi.org/teaching/sa/colon, clear
. keep if stage==1
. stset surv_mm, failure(status==1)
. xi: stcox sex year8594, strata(agegrp)

Stratified Cox regr. -- Breslow method for ties

No. of subjects =         6274     Number of obs   =      6274
No. of failures =         1734
Time at risk    =    424049.72
                                   LR chi2(2)      =     34.92
Log likelihood  =   -12452.776     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t | Haz. Ratio  Std. Err.     z  P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
     sex |   .9151112  .0451873  -1.80 0.072  .8306965    1.008104
year8594 |   .7553578   .037324  -5.68 0.000  .6856348    .8321711
------------------------------------------------------------------------------
                                                  Stratified by agegrp
```



## Time-varying exposures vs time-varying effect of exposure

- We have seen how 'time-varying covariates' can be used in order to allow the effect of exposure to depend on time.
- We may also encounter the situation where the exposure varies with time (effect of the exposure may or may not depend on time), for example, CD4 count, blood pressure, or cumulative exposure to cigarettes or HRT.
- In exercise 13 we will analyse some data collected to study a possible effect of *marital bereavement* (loss of husband or wife) on all–cause mortality in the elderly (see Clayton & Hills, §32.2). Bereavement is a time-varying exposure – all subjects enter as not bereaved but may become bereaved at some point during follow–up.
- A distinction is made between internal variables (which relate to an individual and can only be measured while a patient is alive) and external variables (which do not necessarily require the survival of the patient for their existence).



- Care should be taken when modelling time-dependent covariates, particularly with internal variables [22, 23].



## Analysing the diet data using Cox regression

- Use attained age as the timescale.

```
. use http://www.bioepi.org/teaching/sa/diet
. stset dox, fail(chd) entry(doe) origin(dob) scale(365.25)
. stcox hieng

        failure _d:  chd
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
  enter on or after:  time doe
No. of subjects =          337          Number of obs   =       337
No. of failures =           46
Time at risk    =   4603.66872
                                         LR chi2(1)      =      4.20
Log likelihood  =   -234.78217           Prob > chi2     =    0.0405
```



```
------------------------------------------------------------------------------
   _t | Haz. Ratio   Std. Err.      z    P>|z|  [95% Conf. Interval]
------+-----------------------------------------------------------------------
hieng |   .5426351   .1643032   -2.02   0.043  .2997606    .9822933
------------------------------------------------------------------------------
```

- This is a test of equality of CHD mortality rates between individuals with a high and low energy intake, adjusted for attained age, and assuming proportional hazards with respect to attained age.
- That is, it is a very similar test to that performed in the framework of Poisson regression on slide 148 (and to the log-rank test on slide 155).
- The effect estimate and P-value for the test of the effect of `hieng` are very similar in the Cox and Poisson regression models.
- A slight difference is that attained age was categorised in the Poisson regression model and the rate assumed to be constant within each category.

## Comparison of Cox regression to Poisson regression for the analysis of cohort studies

- The methods are very similar; the basic formulation of both models is

$$\log(\text{rate}) = \boldsymbol{X}\beta.$$

- In both cases, the $\beta$ parameters are interpreted as log rate ratios.
- Both models are multiplicative (i.e. both assume proportional hazards).
- That is, if the RR for males/females is 3 and the RR for smokers to non-smokers is 4, then the RR for male smokers to female non-smokers is 12 (in a model with no interaction terms).
- In Poisson regression, follow-up time is classified into bands and a separate rate parameter is estimated for each band, thereby allowing for the possibility that the rate is changing with time.



- It is assumed that the rate is constant within each band, so if the rate is changing rapidly with time we may have to choose very narrow bands.
- In Cox regression, we essentially choose bands of infinitesimal width; each band is so narrow that it includes only a single event.
- Unlike in Poisson regression, we do not estimate the baseline rates within each time band; instead, we estimate the relative rates for the different levels of the covariates.
- Cox regression is more efficient in this respect if we have a small study (few events).
- When using Poisson regression we must classify events and person-time by subsets of 'time' and require a sufficient number of events to be able to estimate the rate in each stratum.
- The fact that we have to estimate additional parameters for bands of time can be considered a disadvantage (particularly for small studies).



- However, the fact that we obtain estimates for these parameters can be considered an advantage.
- Time-by-covariate interactions are, in practice, easier to model in the framework of Poisson regression.
- In Poisson regression we are not forced to choose a single scale for 'time'.



## Sampling from the risk set: the nested case-control design

- When fitting the Cox model we essentially compare, at every failure time, the characteristics of the individual who failed to the characteristics of all individuals who did not fail (equation 9).
- We could think of this as a case-control study matched on time; at each failure time we have one case and several hundred (or more) controls.
- We could instead select, for example, 5 controls per case with little loss of efficiency.
- Our controls are selected from the risk set; a single individual may be a control at multiple time points and a control may later become a case.
- This is a nested case-control design; a case-control study nested within a cohort.



- This design has become popular because it allows for statistically efficient analysis of data from a cohort with substantial savings in cost and time.
- We may wish, for example, to extract information from medical records for the patients diagnosed with colon carcinoma in order to study additional explanatory variables.
- This would be an ideal setting for a nested case-control design; we extract information for all individuals who died but only a sample of those who did not.
- Another ideal application is where we establish a population-based cohort and take blood samples with the aim of studying the association between genotype and disease risk.
- We store the blood samples and only after following up the cohort do we analyse the samples for the cases (individuals who developed the disease) along with a sample of controls.



- Generating a nested case-control study is very easy in Stata. First, however, we'll repeat the full cohort analysis of the localised colon carcinoma data.

```
. use http://www.bioepi.org/teaching/sa/colon, clear
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. keep if stage==1
(9290 observations deleted)
. gen id=_n
. stset surv_mm, failure(status==1) id(id)
```



```
. xi: stcox sex i.agegrp year8594
i.agegrp          _Iagegrp_0-3        (naturally coded; _Iagegrp_0 omitted)

Cox regression -- Breslow method for ties

No. of subjects =         6274         Number of obs   =      6274
No. of failures =         1734
Time at risk    =    424049.72
                                       LR chi2(5)      =    197.23
Log likelihood  =   -14348.889         Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9151101   .0451776    -1.80   0.072     .8307126    1.008082
_Iagegrp_1 |   .9491689   .1314101    -0.38   0.706      .723597     1.24506
_Iagegrp_2 |   1.338501   .1682956     2.32   0.020     1.046148    1.712553
_Iagegrp_3 |    2.24848   .2834768     6.43   0.000     1.756199    2.878751
  year8594 |   .7548672   .0372669    -5.70   0.000     .6852479    .8315596
------------------------------------------------------------------------------
```



- We will generate a nested case-control study with one control per case.

```
. sttocc, n(1)

          failure _d:  status == 1
    analysis time _t:  surv_mm
                  id:  id
       matching for:  sex

There were 148 tied times involving failure(s)
  - failures assumed to precede censorings,
  - tied failure times split at random

There are 1734 cases
Sampling 1 controls for each case
```

- We could have easily matched on, for example, sex.

```
. sttocc, match(sex) n(1)
```

- The resulting nested case-control study is analysed using conditional logistic regression (theoretically very similar to Cox regression).

```
. xi: clogit _case sex i.agegrp year8594, group(_set) or

Conditional logistic regression     Number of obs   =       3468
                                    LR chi2(5)      =      90.61
                                    Prob > chi2     =     0.0000
Log likelihood = -1156.6132         Pseudo R2       =     0.0377
------------------------------------------------------------------------------
       _case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    .906497   .0648059    -1.37   0.170      .787977    1.042843
  _Iagegrp_1 |   1.010284   .1816638     0.06   0.955     .7102069    1.437149
  _Iagegrp_2 |   1.365991   .2293711     1.86   0.063      .982919    1.898356
  _Iagegrp_3 |   2.298342   .3916342     4.88   0.000     1.645778    3.209654
    year8594 |    .760862   .0541835    -3.84   0.000     .6617425     .8748281
------------------------------------------------------------------------------
```

- Estimates are similar to the full cohort but standard errors are slightly higher.



## Introduction

- Our interest is typically in net mortality (mortality associated with a diagnosis of cancer).
- We have used cause-specific mortality to estimate net mortality — only those deaths which can be attributed to the cancer in question are considered to be events.

$$\text{cause-specific mortality} = \frac{\text{number of deaths due to cancer}}{\text{person time at risk}}$$



## Potential disadvantages of cause-specific survival

- Using cause-specific mortality requires that reliably coded information on cause of death is available.
- Even when cause of death information is available to the cancer registry via death certificates, it is often vague and difficult to determine whether or not cancer is the primary cause of death.
- How do we classify, for example, deaths due to treatment complications?
- Consider a man diagnosed with prostate cancer and treated with estrogen who dies following a myocardial infarction. Do we classify this death as 'due entirely to prostate cancer' or 'due entirely to other causes'?
- Welch *et al.* [24] studied deaths among surgically treated cancer patients that occurred within one month of diagnosis. They found that 41% of deaths were not attributed to the coded cancer.



## Relative survival

- Can instead estimate excess mortality: the difference between observed (all-cause) and expected mortality.

$$\begin{array}{ccccc} \text{excess} & = & \text{observed} & - & \text{expected} \\ \text{mortality} & & \text{mortality} & & \text{mortality} \end{array}$$

- Relative survival is the survival analog of excess mortality — the relative survival ratio is defined as the observed survival in the patient group divided by the expected survival of a comparable group from the general population.
- It is usual to estimate the expected survival proportion from nationwide (or statewide) population life tables stratified by age, sex, calendar time, and, where applicable, race [25].
- Although these tables include the effect of deaths due to the cancer being studied, Ederer *et al.* [14] showed that this does not, in practice, affect the estimated survival proportions.



- A major advantage of relative survival (excess mortality) is that information on cause of death is not required, thereby circumventing problems with the inaccuracy [26] or nonavailability of death certificates.
- We obtain a measure of the excess mortality experienced by patients diagnosed with cancer, irrespective of whether the excess mortality is directly or indirectly attributable to the cancer.
- Deaths due to treatment complications or suicide are examples of deaths which may be considered indirectly attributable to cancer.



## Cervical cancer diagnosed in New Zealand 1994 – 2001
## Life table estimates of patient survival

Women diagnosed 1994 - 2001 with follow-up to the end of 2002

| | | | | | Interval-specific | | | Interval-specific | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Effective number | observed | Cumulative observed | Cumulative expected | relative | Cumulative relative |
| I | N | D | W | at risk | survival | survival | survival | survival | survival |
| 1 | 1559 | 209 | 0 | 1559.0 | 0.86594 | 0.86594 | 0.98996 | 0.87472 | 0.87472 |
| 2 | 1350 | 125 | 177 | 1261.5 | 0.90091 | 0.78014 | 0.98192 | 0.90829 | 0.79450 |
| 3 | 1048 | 58 | 172 | 962.0 | 0.93971 | 0.73310 | 0.97362 | 0.94772 | 0.75296 |
| 4 | 818 | 32 | 155 | 740.5 | 0.95679 | 0.70142 | 0.96574 | 0.96459 | 0.72630 |
| 5 | 631 | 23 | 148 | 557.0 | 0.95871 | 0.67246 | 0.95766 | 0.96679 | 0.70218 |
| 6 | 460 | 10 | 130 | 395.0 | 0.97468 | 0.65543 | 0.94972 | 0.98284 | 0.69013 |
| 7 | 320 | 5 | 129 | 255.5 | 0.98043 | 0.64261 | 0.94198 | 0.98848 | 0.68219 |
| 8 | 186 | 3 | 134 | 119.0 | 0.97479 | 0.62641 | 0.93312 | 0.98405 | 0.67130 |
| 9 | 49 | 1 | 48 | 25.0 | 0.96000 | 0.60135 | 0.91869 | 0.97508 | 0.65457 |



## Issues with relative survival

- The central issue in estimating relative survival is defining a 'comparable group from the general population' and estimating expected survival.
- If not all of the excess mortality is due to the cancer then the relative survival ratio will underestimate net survival (overestimate excess mortality).
- For example, patients diagnosed with smoking-related cancers will experience excess mortality, compared to the general population, due to both the cancer and other smoking related conditions.
- Should the patients be a selected group from the general population, for example, with respect to social class, the national population might not be an appropriate comparison group.



## Statistical cure

- The life table is a useful tool for describing the survival experience of the patients over a long follow-up period.
- In particular, an interval-specific relative survival ratio equal to one indicates that, during the specified interval, mortality in the patient group was equivalent to that of the general population.
- The attainment and maintenance of an interval-specific RSR of one indicates that there is no excess mortality due to cancer and the patients are assumed to be 'statistically cured'.
- An individual is considered to be medically cured if he or she no longer displays symptoms of the disease.
- Statistical cure applies at a group, rather than individual, level.

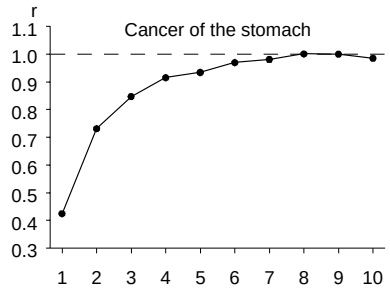



Figure 9: Plots of the annual (interval-specific) relative survival ratios (r) for males and females diagnosed with cancer of the stomach in Finland 1985–1994 and followed up to the end of 1995.



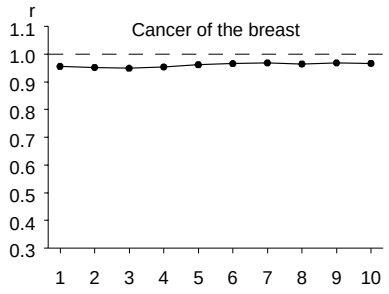



Figure 10: Plots of the annual (interval-specific) relative survival ratios (r) for females diagnosed with cancer of the breast in Finland 1985–1994 and followed up to the end of 1995.



- Plots of the interval-specific RSR are also useful for assessing the quality of follow-up.
- If the interval-specific RSR levels out at a value greater than 1, this generally indicates that some deaths have been missed in the follow-up process.
- An interval-specific relative survival ratio of unity is generally not achieved for smoking-related cancers, such as cancer of the lung and kidney.
- Compared to the general population, these patients are subject to excess mortality due to the cancer in addition to excess mortality due to other conditions caused by smoking, such as cardiovascular disease.



## Estimating relative survival using a period approach

- In 1996 Hermann Brenner suggested estimating cancer patient survival using a period, rather than cohort, approach [27].
- Time at risk is left truncated at the start of the period window and right censored at the end.
- This suggestion was initially met with scepticism although studies based on historical data [28] have shown that
  - period analysis provides very good predictions of the prognosis of newly diagnosed patients; and
  - highlights temporal trends in patient survival sooner than cohort methods.



## Cervical cancer diagnosed in New Zealand 1994 – 2001
## Life table estimates of patient survival

Women diagnosed 1994 - 2001 with follow-up to the end of 2002

| I | N | D | W | Effective number at risk | Interval-specific observed survival | Cumulative observed survival | Cumulative expected survival | Interval-specific relative survival | Cumulative relative survival |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1559 | 209 | 0 | 1559.0 | 0.86594 | 0.86594 | 0.98996 | 0.87472 | 0.87472 |
| 2 | 1350 | 125 | 177 | 1261.5 | 0.90091 | 0.78014 | 0.98192 | 0.90829 | 0.79450 |
| 3 | 1048 | 58 | 172 | 962.0 | 0.93971 | 0.73310 | 0.97362 | 0.94772 | 0.75296 |
| 4 | 818 | 32 | 155 | 740.5 | 0.95679 | 0.70142 | 0.96574 | 0.96459 | 0.72630 |
| 5 | 631 | 23 | 148 | 557.0 | 0.95871 | 0.67246 | 0.95766 | 0.96679 | 0.70218 |
| 6 | 460 | 10 | 130 | 395.0 | 0.97468 | 0.65543 | 0.94972 | 0.98284 | 0.69013 |
| 7 | 320 | 5 | 129 | 255.5 | 0.98043 | 0.64261 | 0.94198 | 0.98848 | 0.68219 |
| 8 | 186 | 3 | 134 | 119.0 | 0.97479 | 0.62641 | 0.93312 | 0.98405 | 0.67130 |
| 9 | 49 | 1 | 48 | 25.0 | 0.96000 | 0.60135 | 0.91869 | 0.97508 | 0.65457 |



## Modelling excess mortality (relative survival)

- The hazard at time since diagnosis $t$ for persons diagnosed with cancer is modelled as the sum of the known baseline hazard, $\lambda^*(t)$, and the excess hazard due to a diagnosis of cancer, $\nu(t)$ [29, 30, 31, 32, 33].

$$\lambda(t) = \lambda^*(t) + \nu(t)$$

- It is common to assume that the excess hazards are piecewise constant and proportional. Provides estimates of relative excess risk.
- The model can be easily estimated in the framework of generalised linear models using standard statistical software (e.g., SAS, Stata, R) [29].
- Non-proportional excess hazards are common but can be incorporated by introducing follow-up time by covariate interaction terms.



## Modelling excess mortality using Poisson regression

- The model can be written as

$$\ln(\mu_j - d_j^*) = \ln(y_j) + \mathbf{x}\beta, \tag{10}$$

  where $\mu_j = E(d_j)$, $d_j^*$ the expected number of deaths, and $y_j$ person-time.
- This implies a generalised linear model with outcome $d_j$, Poisson error structure, link $\ln(\mu_j - d_j^*)$, and offset $\ln(y_j)$.
- Such models have previously been described by Breslow and Day (1987) [1, pp. 173–176] and Berry (1983) [32].
- The usual regression diagnostics (residuals, influence statistics) and method for assessing model fit for generalised linear models can be utilised.



## Estimating and modelling relative survival in Stata

- First create a new directory and download the strs package and ancillary files.

```
. mkdir c:\strs
. net install http://www.pauldickman.com/rsmodel/stata_colon/strs, all
```

- The ado files will be installed in the ado directory and some sample data files (including the colon cancer data we have been working with during the course) will be copied to the current working directory.

```
use colon, clear
keep if stage==1
gen id=_n
stset surv_mm, fail(status==1 2) id(id) scale(12)
```

- We now produce life table estimates of observed and relative for each combination of sex, period, and agegroup.

```
strs using popmort, br(0(1)10) mergeby(_year sex _age) by(sex year8594 agegrp)
```

```
-> sex = Male, year8594 = Diagnosed 75-84, agegrp = 0-44
+-------------------------------------------------------------------------------------+
|start end    n   d   w       p   p_star       r      cp   cp_e2   cr_e2 lo_cr_e2 hi_cr_e2 |
|-------------------------------------------------------------------------------------|
|    0   1   75   4   0  0.9467   0.9970  0.9495  0.9467  0.9970  0.9495   0.8667   0.9826 |
|    1   2   71   8   0  0.8873   0.9968  0.8902  0.8400  0.9938  0.8452   0.7401   0.9114 |
|    2   3   63   1   1  0.9840   0.9965  0.9875  0.8266  0.9903  0.8346   0.7272   0.9041 |
|    3   4   61   3   0  0.9508   0.9963  0.9544  0.7859  0.9866  0.7966   0.6837   0.8747 |
|    4   5   58   3   0  0.9483   0.9960  0.9521  0.7453  0.9827  0.7584   0.6412   0.8439 |
|-------------------------------------------------------------------------------------|
|    5   6   55   2   0  0.9636   0.9956  0.9679  0.7182  0.9784  0.7340   0.6144   0.8241 |
|    6   7   53   0   0  1.0000   0.9953  1.0047  0.7182  0.9738  0.7375   0.6173   0.8280 |
|    7   8   53   0   0  1.0000   0.9949  1.0051  0.7182  0.9688  0.7413   0.6205   0.8322 |
|    8   9   53   1   0  0.9811   0.9945  0.9865  0.7046  0.9635  0.7313   0.6090   0.8246 |
|    9  10   52   2   0  0.9615   0.9942  0.9672  0.6775  0.9579  0.7073   0.5830   0.8048 |
+-------------------------------------------------------------------------------------+
```

- strs can estimate expected survival using 3 different methods (Ederer I, Ederer II, Hakulinen) and supports both cohort and period estimation. See the help file.
- SAS code is also available, see `http:\\www.pauldickman.com\rsmodel\` for both the SAS and Stata code.



## Contents of grouped.dta

| | |
|---|---|
| start | Start of interval |
| end | End of interval |
| n | Alive at start |
| d | Deaths during the interval |
| d_star | Expected number of deaths |
| ns | Number of survivors |
| w | Withdrawals during the interval |
| n_prime | Effective number at risk |
| y | Person-time at risk |
| p | Interval-specific observed survival |
| se_p | Standard error of P |
| lo_p | Lower 95% CI for P |
| hi_p | Upper 95% CI for P |
| p_star | Interval-specific expected survival |
| r | Interval-specific relative survival |
| se_r | Standard error of R |
| lo_r | Lower 95% CI for R |



| | |
|---|---|
| hi_r | Upper 95% CI for R |
| cp | Cumulative observed survival |
| se_cp | Standard error of CP |
| lo_cp | Lower 95% CI for CP |
| hi_cp | Upper 95% CI for CP |
| cp_e2 | Cumulative expected survival(Ederer II) |
| cr_e2 | Cumulative relative survival(Ederer II) |
| lo_cr_e2 | Lower 95% CI for CR (Ederer II) |
| hi_cr_e2 | Upper 95% CI for CR (Ederer II) |
| sex | Sex |
| year8594 | Indicator for year of dx 1985-94 |
| agegrp | Age in 4 categories |



## Modelling relative survival (excess mortality) using Stata

- First need to use strs to produce life table estimates stratified by all of the desired explanatory variables.

```
strs using popmort, br(0(1)10) mergeby(_year sex _age) by(sex year8594 agegrp)
```

- We can then fit a Poisson regression model to the first 5 years of follow-up

```
use grouped if end < 6, clear
xi: glm d i.end i.sex i.year8594 i.agegrp, fam(pois)
                        link(rs d_star) lnoffset(y) eform
```



## Output from the glm command

```
Generalized linear models                   No. of obs      =        80
Optimization     : ML                       Residual df     =        70
                                            Scale parameter =         1
Deviance         =  131.4342128             (1/df) Deviance =  1.877632
Pearson          =  130.1530694             (1/df) Pearson  =   1.85933

Variance function: V(u) = u                 [Poisson]
Link function    : g(u) = log(u-d*)         [Relative survival]

                                            AIC             =   6.39959
Log likelihood   = -245.9836017             BIC             = -175.3077
```



```
------------------------------------------------------------------------------
             |                 OIM
           d |      ExpB   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iend_2 |   .7984084   .0730515    -2.46   0.014     .6673339    .955228
     _Iend_3 |   .6230213   .0671961    -4.39   0.000     .5043086   .7696785
     _Iend_4 |   .4969433   .0645561    -5.38   0.000     .3852391   .6410374
     _Iend_5 |   .4334347    .065147    -5.56   0.000      .322838   .5819191
     _Isex_2 |   .9564493   .0729823    -0.58   0.560     .8235891   1.110742
_Iyear8594_1 |   .7308044   .0539291    -4.25   0.000     .6323935   .8445296
  _Iagegrp_1 |   .8642841   .1353083    -0.93   0.352      .635911   1.174672
  _Iagegrp_2 |   1.071568   .1534869     0.48   0.629     .8092774   1.418869
  _Iagegrp_3 |   1.436319   .2146593     2.42   0.015     1.071613   1.925147
           y | (exposure)
------------------------------------------------------------------------------
```

- Exercise: Compare the estimated hazard ratios to those obtained from Poisson regression fitted to cause-specific survival. Are they similar? Would you expect them to be similar?



## Selection bias in observational studies

- The aim of a study is usually to derive from an available subset of patients, statements about their patterns of survival which will be generalisable to a wider body of patients.
- Selection bias can occur when patients treated at a given clinic are not representative of a general class of patients. For example, if seriously ill patients are transferred to a specialist clinic then neither patients treated at the 'general' clinic or the specialist clinic will be representative of 'all patients'.
- Selection bias also occurs when treatment is assigned based on characteristics of the patients, thereby precluding comparisons between treatment groups.
- Patients treated aggressively are generally healthier than patients treated conservatively. For example:

(i) Radical prostatectomy vs 'watchful waiting' (expectant therapy) for men diagnosed with localised prostate cancer.



(ii) Bone marrow transplant and high dose chemotherapy vs conventional therapies for women diagnosed with advanced breast cancer.

- High-dose chemotherapy accompanied by transplant involves harvesting bone marrow or stem cells from the patient prior to chemotherapy.
- The patient then receives high-dose chemotherapy, which adversely affects bone marrow. After chemotherapy, the stem cells and bone marrow are replaced with the hope that the drugs have killed the cancer cells and the bone marrow will regenerate before the patient dies of infection.
- The procedure was started in 1979 and by the late 1980s the results looked very promising. Patients with advanced breast cancer who were given transplants had remission rates of 50 to 60 percent compared with the 10 to 15 percent remission rates achieved by conventional means.
- Subsequent examination of the data showed that the women receiving the transplant treatment were carefully selected to be younger than 60 and in general good health.

- Recently completed randomised clinical trials found no difference in survival for women who were randomly assigned to have transplants and those who were assigned to conventional therapy.
- In general, comparison of survival according to treatment should be avoided in observational studies.
- It is possible to adjust for factors which make a patient subgroup atypical (e.g. disease characteristics, presence of comorbid conditions, age, etc.) but there is no substitute for a randomised experimental trial for evaluating different treatments.



## Another problem which arises when comparing treatments

- A common question is whether a combination treatment (e.g. surgery followed by radiation therapy) is preferable to the single treatment (e.g. surgery alone).
- Survival time is usually measured from date of diagnosis, date of first hospital admission, or date of first treatment.
- In order to receive the combination treatment, one must survive a sufficient period after surgery in order to receive the radiation therapy.
- Those who die during, or immediately after, surgery are included in the 'surgery only' group.
- A naive analysis would show that the group receiving combination therapy experience superior survival.



## Screening and lead time bias in cancer survival analysis

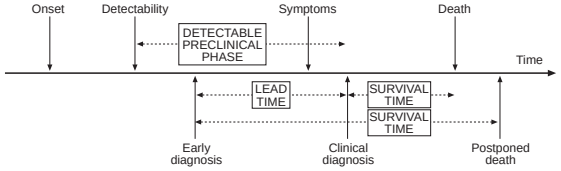



Figure 11: Natural history of chronic illnesses.

- The survival time for cancer patients is measured from the date of diagnosis to the date of death (Figure 11).
- As such, any factor which affects either the definition of the illness, the date of diagnosis, or the date of death will also affect the survival time.



- It is possible, for example, to increase patient survival time by bringing forward the date of diagnosis without altering the date of death.
- The implementation of a mass-screening program leads to cancers being detected earlier than they would have been without screening.
- This difference in the time of diagnosis is called the lead time (Figure 11) and can bias the comparison of survival between patient groups, the so-called lead time bias [34].
- In practice, bringing forward the date of diagnosis should also delay the date of death by increasing the effectiveness of standard therapies.
- The implementation of a mass-screening program is not the only way to bring about early diagnosis of cancer.
- Increased contact with the health system for any reason may lead to early clinical diagnosis.



- Public education programs, such as those aimed at educating the public of the warning signs for melanoma, and encouraging skin self-examinations and consultation with a physician when these warning signs are observed, will lead to earlier diagnosis of melanoma.



## Publishing the results of survival studies

- Discussion of the articles by Altman et al. [35, 36].



## References

[1] Breslow NE, Day NE. *Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies*. IARC Scientific Publications No. 82. Lyon: IARC, 1987.

[2] Rosner BA. *Fundamentals of Biostatistics*. Duxbury, 5th edn., 2000.

[3] Cleves MA, Gould WW, Gutierrez RG. *An Introduction to Survival Analysis Using Stata*. Stata Press, 2004.

[4] Cantor A. *SAS Survival Analysis Techniques for Medical Research*. BBU Press, 2nd edn., 2003.

[5] Allison PD. *Survival Analysis Using the SAS System: A Practical Guide*. Cary, NC: SAS Institute Inc., 1996.

[6] Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford: Oxford University Press, 1993.

[7] Elandt-Johnson RC. Definition of rates: Some remarks on their use and misuse. *American Journal of Epidemiology* 1975;**102**:267–271.

[8] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958;**53**:457–481.

[9] Böhmer PE. Theorie der unabhängigen Wahrscheinlichkeiten. *Rapports Memoires et Proces verbaux de Septieme Congres International dActuaires Amsterdam* 1912;**2**:327–343.



[10] Greenwood M. *The Errors of Sampling of the Survivorship Table*, vol. 33 of *Reports on Public Health and Medical Subjects*. London: Her Majesty's Stationery Office, 1926.

[11] Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, 1997.

[12] Altman DG, Bland JM. Time to event (survival) data. *British Medical Journal* 1998; **317**:468–469.

[13] Halpern DF, Coren S. Handedness and life span. *New England Journal of Medicine* 1991; **324**:998.

[14] Ederer F, Axtell LM, Cutler SJ. The relative survival rate: A statistical methodology. *National Cancer Institute Monograph* 1961;**6**:101–121.

[15] Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.

[16] McCullagh P, Nelder JA. *Generalized Linear Models*. London: Chapman and Hall, 2nd edn., 1989.

[17] Lock RH, Danehy TJ. Using a Poisson model to rate teams and predict scores in ice hockey. In: *ASA Proceedings of the Section on Statistics in Sports*. American Statistical Association (Alexandria, VA), 1997, 1997; 25–30.

[18] Danehy TJ, Lock RH. Chodr – Using statistics to predict college hockey. *Stats The Magazine for Students of Statistics* 1995;**13**:10–14.

[19] Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* 1972;**34**:187–220.

[20] Hess KR. Graphical methods for assessing vioations of the proportional hazards assumption in Cox regression. *Statistics in Medicine* 1995;**14**:1707–1723.

[21] Therneau TM, Grambsch PM. *Modelling Survival Data: Extending the Cox Model*. Springer: New York, 2000.

[22] Fisher LD, Lin DY. Time-dependent covariates in the cox proportional-hazards regression model. *Annu Rev Public Health* 1999;**20**:145–57.

[23] Wolfe RA, Strawderman RL. Logical and statistical fallacies in the use of cox regression models. *Am J Kidney Dis* 1996;**27**:124–9.

[24] Welch HG, Black WC. Are deaths within 1 month of cancer-directed surgery attributed to cancer? *J Natl Cancer Inst* 2002;**94**:1066–70.

[25] Berkson J, Gage RP. Calculation of survival rates for cancer. *Proceedings of Staff Meetings of the Mayo Clinic* 1950;**25**:270–286.

[26] Percy CL, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health* 1981;**71**:242–250.

[27] Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer* 1996;**78**:2004–2010.

[28] Brenner H, Gefeller O, Hakulinen T. Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation, computational realisation and applications. *European Journal of Cancer* 2004;**40**:326–35.



[29] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Stat Med* 2004;**23**:51–64.

[30] Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine* 1990;**9**:529–538.

[31] Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987;**36**:309–317.

[32] Berry G. The analysis of mortality by the subject-years method. *Biometrics* 1983; **39**:173–184.

[33] Pocock S, Gore S, Kerr G. Long term survival analysis: the curability of breast cancer. *Stat Med* 1982;**1**:93–104.

[34] Day NE. The assessment of lead time and length bias in the evaluation of screening programmes. *Maturitas* 1985;**7**:51–58.

[35] Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer* 2004;**91**:4–8.

[36] Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. *British Journal of Cancer* 1995;**72**:511–518.