Chapter 1: Introduction of Survival Analysis

Chien-Fu Jeff Lin, MD. PhD.

cflin@mail.ntpu.edu.tw

http://web.ntpu.edu.tw/~cflin

Department of Statistics
National Taipei University, Taipei, Taiwan

Spring 2007

Outline I

Example: A Simple Question

- 100 patients were admitted to hospital on Sep 7, 2000,
- 99 patients were discharged on Sep 11, 2000,
- 1 patient died on Sep 12, 2000,
- What's the death rate on Sep 12, 2000 ?
- $\frac{100}{?}$
- $\frac{1}{1}$?

The answer really depends on how you collect your data!

Example: A Simple Question

- 100 patients were admitted to hospital on Sep 7, 2000,
- 99 patients were discharged on Sep 11, 2000,
- 1 patient died on Sep 12, 2000,
- What's the death rate on Sep 12, 2000 ?
- $\frac{100}{?}$
- $\frac{1}{1}$?

The answer really depends on how you collect your data!

Example: A Simple Question

- 100 patients were admitted to hospital on Sep 7, 2000,
- 99 patients were discharged on Sep 11, 2000,
- 1 patient died on Sep 12, 2000,
- What's the death rate on Sep 12, 2000 ?
- $\frac{100}{?}$
- $\frac{1}{1}$?

The answer really depends on how you collect your data!

Example: A Simple Question

- 100 patients were admitted to hospital on Sep 7, 2000,
- 99 patients were discharged on Sep 11, 2000,
- 1 patient died on Sep 12, 2000,
- What's the death rate on Sep 12, 2000 ?
- $\frac{100}{?}$
- $\frac{1}{1}$?

The answer really depends on how you collect your data!

Example: A Simple Question

- 100 patients were admitted to hospital on Sep 7, 2000,
- 99 patients were discharged on Sep 11, 2000,
- 1 patient died on Sep 12, 2000,
- What's the death rate on Sep 12, 2000 ?
- $\frac{100}{?}$
- $\frac{1}{1}$?

The answer really depends on how you collect your data!

Example: A Simple Question

- 100 patients were admitted to hospital on Sep 7, 2000,
- 99 patients were discharged on Sep 11, 2000,
- 1 patient died on Sep 12, 2000,
- What's the death rate on Sep 12, 2000 ?
- $\frac{100}{?}$
- $\frac{1}{1}$?

The answer really depends on how you collect your data!

Example: A Simple Question

- 100 patients were admitted to hospital on Sep 7, 2000,
- 99 patients were discharged on Sep 11, 2000,
- 1 patient died on Sep 12, 2000,
- What's the death rate on Sep 12, 2000 ?
- $\frac{100}{?}$
- $\frac{1}{1}$?

The answer really depends on how you collect your data!

Example: A Simple Question

- 100 patients were admitted to hospital on Sep 7, 2000,
- 99 patients were discharged on Sep 11, 2000,
- 1 patient died on Sep 12, 2000,
- What's the death rate on Sep 12, 2000 ?
- $\frac{100}{?}$
- $\frac{1}{1}$?

The answer really depends on how you collect your data!

Outline

Background I

1. Survival analysis are used to analyze the length of time between a starting event (entry into follow-up) and an outcome event (such as death).
2. Such data are often censored; that is, not all subjects who enter the study are followed long enough to observe the time of the outcome event.
3. In addition, such data often have highly skewed distributions.
4. Special statistical methods are required in order to analyze such data.

Background II

5. More precisely, survival analysis is the study of the distribution of life times, and is a loosely defined statistical term that encompasses variety of statistical technique for analyzing positive-valued random variables.
6. Typically, the value of the random variable is the times from an initiating event time point to some terminal event time point, i.e. from time of birth (start of treatment) to death(relapse).

Survival Analysis in Various Researches I

1. Births to Deaths (Medicine, Epidemiology, Demography)
2. Diagnosis of Disease(s) (Cancer, Stroke) to Death (Medicine, Epidemiology)
3. Hospitalizations and Length of Stay (Medicine, heath management)
4. Repair and Replacement in Mechanical System (Engineering, Reliability)
5. Marriages to Divorces, Re-marriages to Re-divorces (Sociology, Event History Analysis)
6. Residence Changes, (Politics, Sociology)
7. Immigration and Migration (Politics, Sociology)

Survival Analysis in Various Researches II

8. Commit Crimes, Arrests, Probation, and Convictions (Law and Criminal Justice)
9. Job Changes, Loss, and Promotions (Business, Sociology)
10. Business Closing, Bankruptcies (Business, Tobit Model)
11. Customer Loyalty (Business, Data Mining)

Outline

Example: Complete Observations of One Year Study I

1. A 1 year study of 50 animals
2. All survived for 1 year and 20 developed skin cancer.
3. The incidence proportion developing skin cancer during the first year is estimated to be $\hat{p} \times 100 = 40$ percent with $s.e.(\hat{p}) = \sqrt{\hat{p}\hat{q}/n} \times 100 = \pm 6.9\%$.

Example: Complete Observations of One Year Study II

4. An approximate confidence interval can be computed based on the normal distribution.
5. An exact confidence interval can be computed using the exact binomial distribution.
6. The estimation methods would be different if half the animals died cancer-free during the year.

Example: Complete Observations of Life Time Study I

1. A life-time study of 50 animals
2. 20 developed skin cancer
3. Estimate the life-time cancer incidence proportion.
4. This is almost the same as the example above, for a 1 year study.
5. Simple proportions, as used here, are only appropriate if all subjects are followed for an equal interval of time.
6. In this case the interval of time is defined as a lifetime.
7. An "equal follow-up interval" is usually defined by a fixed time interval, such as 1 year, but can be measured in any units of time, such as lifetimes or generations.

Example: Complete Observations of Life Time Study II

8. Note that the average time to cancer and the median time to cancer are not meaningful for the entire population, since not all animals developed cancer during follow-up.
9. The average time to cancer among those developing cancer during a one year follow-up is less interpretable.
10. The average time to cancer among those developing cancer during a follow-up that varied between 1 and 5 years is very difficult to interpret.

Complete Observations I

Consider the complete observations (all deaths observed) of an 20 children with acute leukemia survival study in the following Table 1:

Table 1: Time (months) to death data: complete observations

3.1	5.6	7.1	9.6	6.4	34.3	18.5	51.2	14.1	17.3
5.2	7.8	46.3	25.0	8.8	29.1	23.7	33.9	4.7	43.9

Complete Observations: R I

```
> # R commands
> time.comp<-c(3.1, 5.6, 7.1, 9.6, 6.4, 34.3, 18.5,
               51.2, 14.1, 17.3, 5.2, 7.8, 46.3, 25.0,
               8.8, 29.1, 23.7, 33.9, 4.7, 43.9)
> summary(time.comp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.100   6.925  15.700  19.780  30.300  51.200
> sd(time.comp)
[1] 15.35344
```

Complete Observations: R II

```
> # R commands
> stem(time.comp)

  The decimal point is 1 digit(s) to the right of the
  0 | 35566789
  1 | 0479
  2 | 459
  3 | 44
  4 | 46
  5 | 1

> plot.density(density(time.comp))
```

Complete Observations: R I

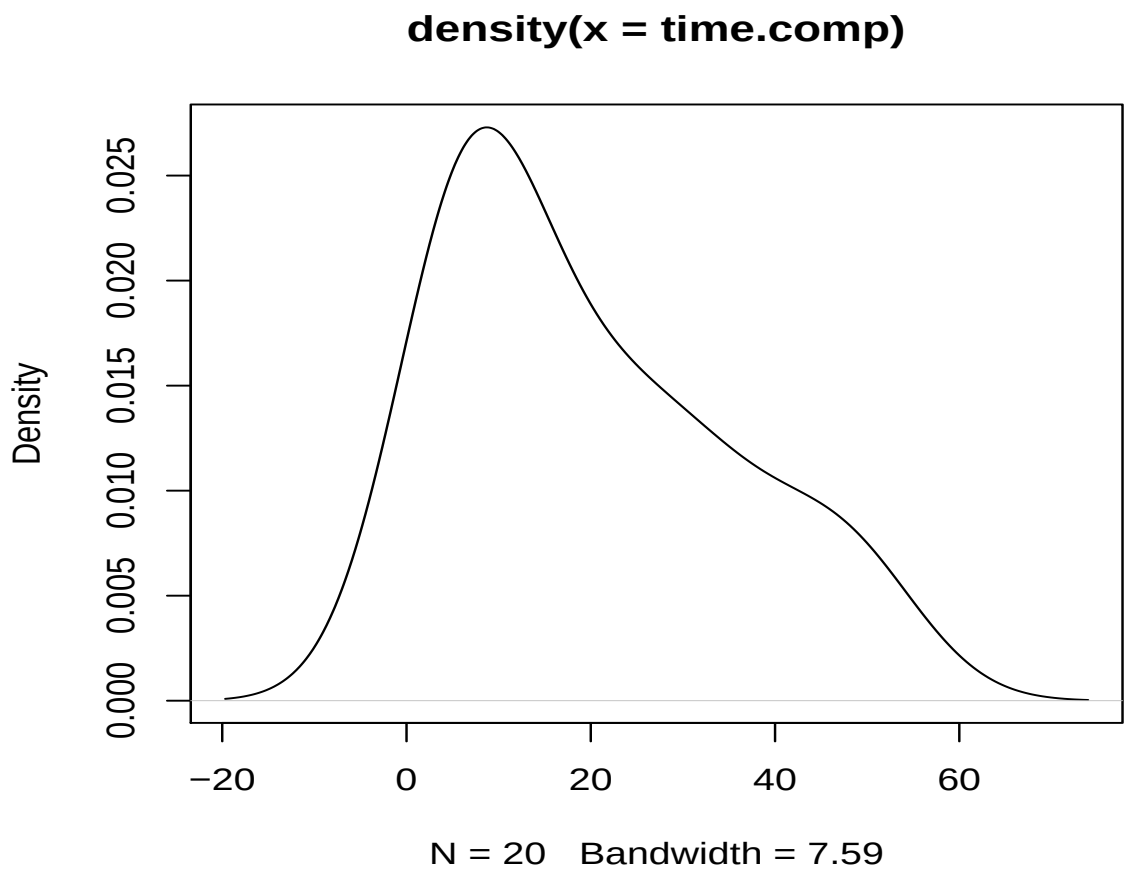

Figure 1: Density Plot: Complete Observations of Survival Time

Complete Observations I

1. The average time from entry to death is 19.7 months (S.D.$=$ 15.3).
2. The time range from 3.1 to 51.2 months.
3. Half of the subjects were dead below 15 months. The distribution of time is not normal distributed and is highly skew.

Outline

Censored Data I

1. Censored data arise from losses to follow-up and from varying follow-up intervals.
2. Censored data make it more difficult to compute interpretable summaries.
3. How would you compute the 5 year death fraction based on the following outcomes from a 5 year study of 50 subjects in the following Table 2?

Censored Data II

Table 2: Time (months) to death data with censored observations

Number	Observed outcome
10	Drop-out alive before 5 years
5	die during study
35	alive at 5 years

Censored Data III

4. What do we know about the 10 subjects who were lost to follow-up would have died within 5 years, had they didn't dropped out.
5. If all study subjects are followed for a fixed equal period of time, then an event proportion (risk) can be estimated for that interval with no ambiguity.
6. When subjects are followed for differing lengths of follow-up, it is often more appropriate to estimate an event rate.

Outline

Survival Studies

In clinical (medical) studies, most subjects enter the study over a period of time. Others will become lost to contact during the follow-up period so that their information is not available at the end of the study. The length of time of the study or follow-up will therefore not be the same for each subject.

Calendar Time of Follow-up Data I

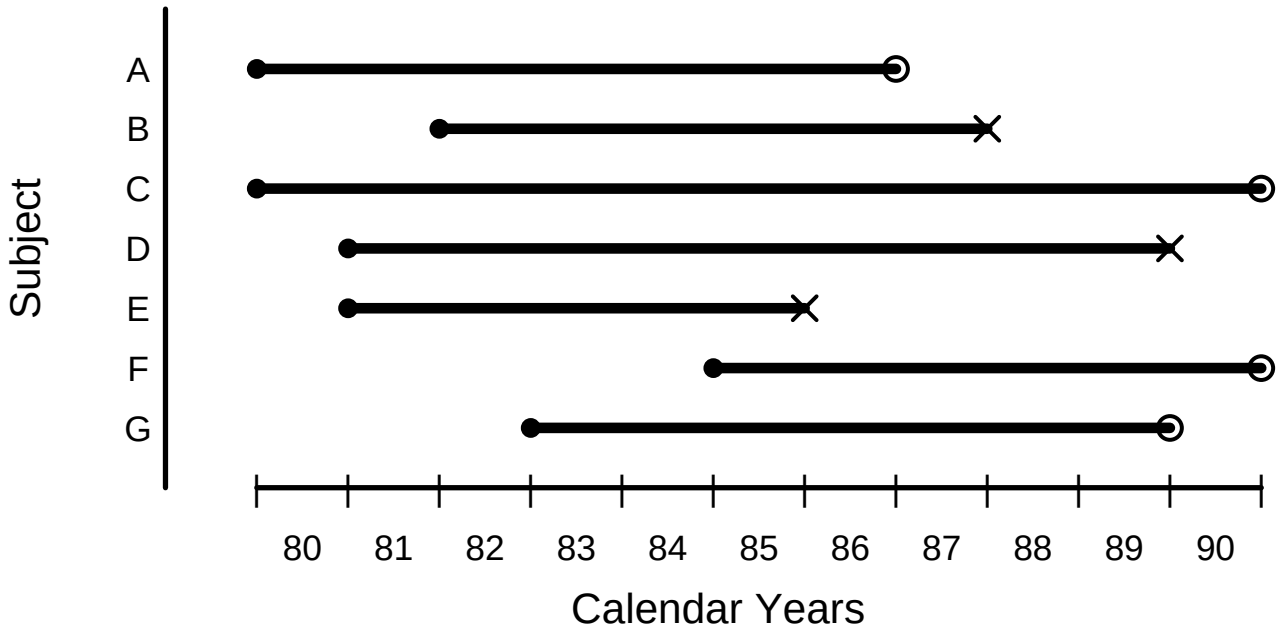

Figure 2: Calendar Time: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990

Example: Calender Time: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990 I

1. Figure 2 shows the long-term follow-up results of 7 subjects with lung cancer from January 1, 1980 to December 31, 1990.
2. $\times$ denote death and $\bigcirc$ denote alive at the last visit.
3. Researchers often assume that all subjects enter the study on the same date, (i.e., day zero) and calculate the length of time from day zero the last follow-up visit.

Example: Calender Time: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990 II

4. Subject B, D, E are observed dead during the study time
5. Others were alive at last visit.
6. These data are plotted on the calendar (actual) time scale in Figure 2.
7. Then, researchers investigate the death fraction (or survival proportion) over a certain period of time.

Patient Time of Follow-up Data I

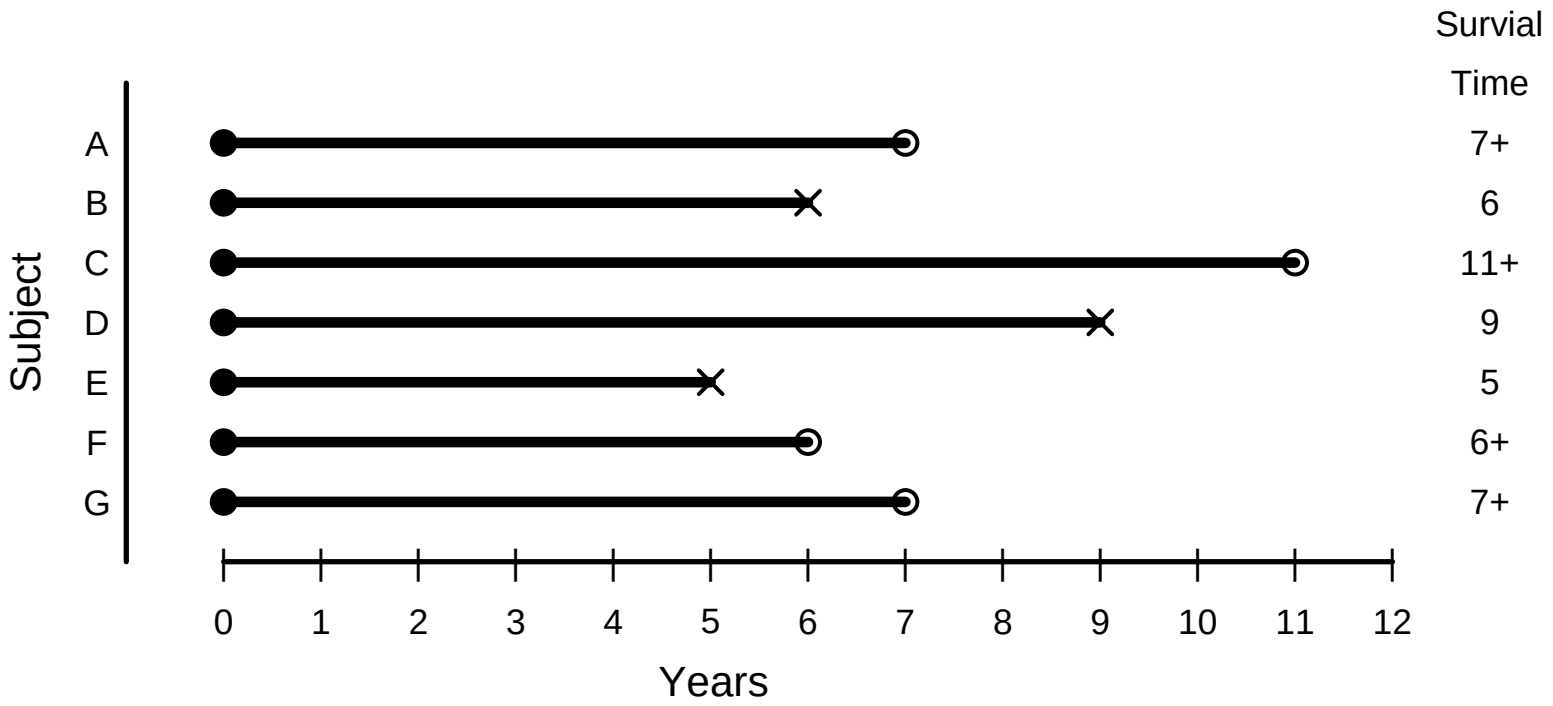

Figure 3: Patient Time: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990: II

Example: Patient (Study) Time: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990 I

1. Figure 3 shows the long-term follow-up results of 7 subjects with lung cancer from January 1, 1980 to December 31, 1990.
2. $\times$ denote death and $\bigcirc$ denote alive at the last visit.
3. Researchers calculate the length of time from the date of diagnosis, day zero, to the date of last visit, death date or alive at the last visit date.
4. When the subject is alive at the last visit, researches put a $+$ (or $\star$) sign beside the length of time, i.e., $7+$ ($7\star$).

Example: Patient (Study) Time: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990 II

5. As far as each subject is concerned, the period of time that a subject spends in the study, measured from that subject's time origin, is often referred to as "patient time" as in Figure 3.
6. Note that each subject's time has been plotted as if he or she were enrolled at exactly the same date and were followed until his or her respective end point.

Example: Patient (Study) Time: Follow-up 7 Subjects with Lung Cancer from 1980 to 1990 III

7. The period time, patient (study) time, from the time origin to the death, $\times$ of a subject is then the survival time (or failure time), and is recorded for the individuals B, D, E.
8. The survival time so the remaining individuals are righted-censored $\bigcirc$.

Survival Time (Failure Time) I

1. In practice, the actual data recorded will be the date on which each individual enters the study, and the date on which each individual dies or was last known to be alive.
2. The survival time (or failure Time) in days, weeks, months, or years, whichever is the most appropriate, can the be calculated.

Outline

Remission Duration from a Clinical Trial for Acute Leukemia I

1. Freich et al. (1963) and Gehan (1965) report the results of a randomized controlled clinical trial of 6–mercaptopurine (6–MP) versus placebo in 42 children.
2. Patients were followed until their leukemia return (relapse).
3. The data are shown in Table 3;
4. The time unit is month, censor denotes the last follow-up visit is dead as 1 and alive as 0.
5. The investigators were interested in the differences of survival rate between two treatments.

Table 3: A clinical trial of 6-mercaptopurine (6-MP) versus placebo.

Placebo				6-MP			
Time	Censor	Time	Censor	Time	Censor	Time	Censor
1	1	5	1	10	1	20	0
22	1	4	1	7	1	19	0
3	1	15	1	32	0	6	1
12	1	8	1	23	1	17	0
8	1	23	1	22	1	35	0
17	1	5	1	6	1	6	1
2	1	11	1	16	1	13	1
11	1	4	1	34	0	9	0
8	1	1	1	32	0	6	0
12	1	8	1	25	0	10	0
2	1			11	0		

Outline

Failure Times and White Blood Cell Count (WBC) I

1. Table 4 shows, for two groups of leukemia patients, failure time (time to death) in weeks and white blood cell count (WBC), Feigl and Zelen, (1965).
2. The presence of continuous explanatory variable, WBC, AG, lies in that the division into groups is based on an (uncontrolled) measurement for each individual rather than on a randomized allocation.
3. The investigators were interested in the effect of WBC on the survival time.

Table 4: Failure Times and White Blood Cell Count, WBC (Feigl and Zelen, 1965)

AG+			AG−		
Patient Number	WBC Count	Failure Time (Weeks)	Patient Number	WBC Count	Failure Time (Weeks)
1	2300	65	18	4400	56
2	750	156	19	3000	65
3	4300	100	20	4000	17
4	2600	134	21	1500	7
5	6000	16	22	9000	16
6	10500	108	23	5300	22
7	10000	121	24	10000	3
8	17000	4	25	19000	4
9	5400	39	26	27000	2
10	7000	143	27	28000	3
11	9400	56	28	31000	8
12	32000	26	29	26000	4
13	35000	22	30	21000	3
14	100000	1	31	79000	30
15	100000	1	32	100000	4
16	52000	5	33	100000	43

Outline

Surgical result of stage I lung cancer at VGH from 1970 to 1995 I

1. A retrospective study assessed the surgical results of stage I lung cancer at Taipei Veteran General Hospital.
2. Table 5 show the partial results of 7 subjects in the study.
3. The variable OPDATE is the date of operation.
4. The variable FUDATE is the date of the date of death or the date of last follow-up visit.

Surgical result of stage I lung cancer at VGH from 1970 to 1995 II

5. The variable FUMO (FUDAY) is the duration of survival time measured in month (DAY) from the date of operation to the date of death or the date of last follow-up visit.
6. The STATUS is 1 if the subject died, 2 if the subject died of other causes, and 0 if the subject is alive at the last visit.
7. The investigator are interested in the overall and cancer-specific survival results.
8. In practice, the actual data recorded will be the date on which each individual enters the study, and the date on which each individual dies or was last known to be alive.

Table 5: The surgical results of stage I lung cancer at VGH from 1970 to 1995

STATUS	OPDATE	FUDATE	FUMO	FUDAY
1	1980/08/02	1985/05/11	58.10	1743
1	1980/09/01	1993/04/15	153.63	4609
1	1980/09/03	1980/11/14	2.40	72
1	1980/09/15	1981/04/05	6.73	202
0	1980/10/14	1995/03/04	175.13	5254
1	1980/10/14	1989/02/08	101.30	3039
2	1980/12/11	1983/09/22	33.83	1015

Outline

VGH Lung Cancer Follow-up Survival Study: 1970–1995 I

1. Three hundred and twenty-one patients who underwent complete surgical resection for stage I non-small cell lung cancer were reviewed retrospectively by Wu et al. (2003).
2. The aim of this study carried out at the Taipei Veteran general Hospital was to examine the association between the number of totally removed lymph nodes during thoracotomy and other explanatory variables or covariates and the survival time of patients.
3. Part of the data are in the file survVGHlungca95class.xls and Table 6 shows the descriptions of the variables.

Table 6: The variables in the VGH Lung Cancer Follow-up Survival Study Data: 1970–1995

Variable	Description.
age	Age in years.
sex	Gender: 1 = Male; 2 = female.
cough	Cogh at diagnosis; 0 = no; 1 = yes.
smoke	Smoking amount measured as pack-years.
hemopty	Hemoptysis at diagnosis; 0 = no; 1 = yes.
chestpain	Chest pain at diagnosis; 0 = no; 1 = yes.
wtloss	Body weight loss at diagnosis; 0 = no; 1 = yes.
surgeonid	Sugeron indicator, 1, 2, 3, 4, 6, 9.
opdate	Date of operation
cell	Cell type: 1 = squamous cell; 2 = others.
tumorsize	Tumor size measured in cm.
lnd	Total number of the removed lymphoid nodes.
metadate	Date of metastasis.
fudate	Date of death or date of last follow-up visit.
status	Censor status at last visit:
	0 = alive; 1 = death of cancer; 2 = death of other causes.

Outline

Survival Times of Stanford Heart Transplant Programme I

1. We consider the survival of patients accepted into the Stanford heart transplant programme (Example 8.9 and Table 8.1 in Cox and Oakes, 1984).
2. Table 7 showed the descriptions of the variables and the data are in file survCoxHeart.xls.
3. Table 8 show the results of the first 5 subjects.
4. The data covered the period from the start of the programme in 1967, to February 1980.

Table 7: Survival Times of Stanford Heart Transplant Programme

Variable	Description.
wait	Waiting time (in days) before transplantation.
transplant	Transplantation indicator: 1 = yes; 0 = no.
survtime	Survival time (in days) post-transplantation
status	Final status, 1 = dead, 0 = alive.

Table 8: Partial Results of five Subjects Stanford Heart Transplant Programme

wait	transplant	survtime	status
49	0		1
5	0		1
0	1	15	1
35	1	3	1
17	0		1

Variables: Survival Times of Stanford Heart Transplant Programme I

1. By this time a total of 249 patients had been accepted for transplantation, but only 184 patients had actually received transplants.
2. Of the remaining 65 patients all but eight died awaiting transplantation.
3. By contrast, 65 of the transplanted patients were still alive at the closing date.

Variables: Survival Times of Stanford Heart Transplant Programme II

4. The design of the programme is crucial to a proper understanding of the data.
5. Patients are accepted into the programme when judged by the physicians to be suitable, candidates for transplantation.
6. When a donor heart becomes available, medical judgement is used to select the candidate who should receive it.

Variables: Survival Times of Stanford Heart Transplant Programme III

7. For transplant patients both the waiting, from the date of acceptance to the date of transplant, and post-transplant survival time are observed, together with an indicator of the final vital status.
8. For non-transplant patients, the survival time from the date of acceptance and the vital status indicator are available.

Outline

Characteristic of Survival Time I

1. It is often useful to summarize the survival experience of a study group.
2. The summary is especially useful if the study group is representative of a larger population.
3. Then the survival experience of the study group is an estimate of the survival experience of the wider population.
4. If the study group is a random sample from the target population, then probabilistic measures of the accuracy of the estimate can be computed.

Characteristic of Survival Time I

1. Censored Data
   1. Censored Data is those observations whose times to event we do not get to observe completely.
   2. Censored Data prevent the use of standard methods of statistical summarization and inference.
   3. In particular, right censored data are reported as lower bounds for the actual unobserved event times.
   4. If the dependent variable is the time of the event, what do you do with these cases?

Characteristic of Survival Time II

2. Not Normally Distributed
   1. Survival times often have a non-normal distribution in the population that is very different from a Gaussian (Normal) distribution used in most statistical inferences.
   2. Many standard approximate statistical methods are not accurate for such data.

Characteristic of Survival Time III

3. Whole Distribution of Survival Times
   1. Many standard statistical methods are instead oriented towards inference for the mean survival time, μ and standard deviation σ.
   2. The extremes of the distribution of times to event (extreme quantiles) are often of interest in survival analysis.
   3. For example, many people hope that they will live to the 95^{th} percentile, rather than the 50^{th} percentile.
   4. The rate of occurrence of events per unit time is often of interest in survival analysis.

Characteristic of Survival Time IV

4. Time-Dependent Covariates
   1. Some explanatory variables (i.e., blood pressure value) change over time during the study.
   2. How do you handle these time-dependent covariates variables in a regression analysis?

Outline

Data Collection I

1. Survival analysis methods were developed to address several issues that arise in data collection due to the one-directional flow of time.
2. Survival analysis methods are oriented towards analyzing the distribution of the (random) length of time to an event.
3. Equivalently, we examine the rate of occurrence of the event per unit time.
4. The time to event depends not only upon the definition of the event of interest, but also upon the definition of the time origin for each subject. The following Table 9 summarize the data collection about survival analysis.

Data Collection I

Table 9: Some important times and dates in data collection

Name	Whose Time Axis?	Description
Study Start	Data Collector	Calendar date of data collection start
Event Time Origin	Subject	Time 0 for measuring the time to event for a subject
Entry into Follow-Up	Subject	Time when follow-up starts
Observation Times	Data collector	Calendar times when subject is observed
Loss to Follow-Up	Subject	Time of drop out from study
End of Study	Data Collector	Calendar date of data collection end
		Follow-up usually stops for all subjects at this time
Time of Event	Subject	Only partially known for some subjects (censoring)

Right Censored Data I

1. Limited time resources during a study can prevent observation of those times to event that are long.
2. This leads limitation to right censored data.
3. Standard designs for data collection lead to different potential lengths of follow-up for different subjects.
4. If the time to event for a subject is longer than the potential follow-up time, then the time to event for that subject is right-censored.
5. Other aspects of the enrollment of subjects, dropout from the study, planned study termination, and frequency of observation during follow-up can lead to various types of censoring.

Truncated Data I

1. In some studies, subjects are not enrolled at time 0.
2. The sample is then limited to subjects whose time to event is at least as long as the time until their enrollment.
3. This type of sample is a truncated sample.
4. Statistical inference must account for the fact that the observed data are conditional upon being in the sample.

Interval Censored Data I

1. In some studies, the observation times are episodic (not continuous in time).
2. This leads to interval censored data.

Summary of Data Collection I

Table 10: What's different about survival analysis?

Concept of Interest	Approach
Symmetry	Doesn't happen. Too bad.
Normal Distribution	Usually not relevant to the population.
	Fundamental for sampling distributions.
Typical Value	Percentiles (not the average).
Distribution	All but the right-most part of the distribution for right censored data.
Distribution	All but left-most part for truncated data.
Cumulative Distribution	Survival Curve $= 1-$ Distribution Function.
Density-type function	Hazard = Density / Survival Curve.
Regression Parameter	Hazard ratios.

Outline

Survival/Proportion/Poisson Models I

1. The first thing is to avoid the Type 0 error: not recognizing the type of problem you have.
2. Investigators often want to study the occurrence of an event.
3. And data involve follow-up of subjects and recording the occurrence of an event.

Survival/Proportion/Poisson Models II

4. Survival analysis yields estimates of survival curves (fraction without an event) and event rates (per unit time).
5. Logistic regression yields estimates of the fraction with (or without) the event.
6. Poisson analysis yields estimates of event rates (per unit time). Which should be used?

Guidelines for Survival/Proportion/Poisson Models I

Guidelines (which should be interpreted for each case):

1. Use logistic if the period of follow-up is the same for every subject and has a well-defined time origin and only the first event is of interest. This often arises when the event occurs quickly after the time origin (death within 30 days of a heart attack).

Guidelines for Survival/Proportion/Poisson Models II

2. Use Poisson if the period of follow-up differs among subjects and the event rate is constant during follow-up (so the time origin doesn't matter). Comparison to published death rates is often based on a Poisson analysis.

Guidelines for Survival/Proportion/Poisson Models III

3. Use survival if the period of follow-up differs among subjects and each subject has a well-defined time origin and the event rate varies with time during follow-up. Loss to follow-up or censoring is a clue that this is occurring. Long follow-ups also suggest survival analysis, since the event rate is likely to vary over a long time period.

Often, a combination of analyses is appropriate.

Summary: Survival Data I

1. Survival time: positive random variable from initial event to outcome event
2. Censored data
3. Not normally distributed
4. Time-dependent covariates