

---

SISCER Mod 12

# Survival Analysis

## Lecture 1

---

Ying Qing Chen, Ph.D.

Department of Medicine  
Stanford University

Department of Biostatistics  
University of Washington

---

# Before the lectures

## ■ Instructors

□ Ying Qing Chen, Ph.D.

■ yqchen@uw.edu

□ Eric R. Morenz

■ emorenz@uw.edu

## ■ Students

---

---

# About the course

- What is this course about
    - Censored time-to-event analysis
  - Recommended textbooks
    - Kleinbaum, DG (2012) *Survival Analysis: A Self Learning Text, 3<sup>rd</sup> Ed.* Springer.
    - Hosmer, DW & Lemeshow, S (2008) *Applied Survival Analysis, 2<sup>nd</sup> Ed.* Wiley.
  - Instruction hours
    - 8:30am – 10:00am, 10:30am – 12:00noon
  - Student self-evaluation
    - Two Problem Sets
    - Four Quizzes
  - Course calendar
-

---

# Course outlines

- Background
    - Time-to-event and censoring
    - Life tables
  - Parametric Methods
    - Parametric distributions
    - Likelihood functions and maximum likelihood
  - Kaplan-Meier Curves
  - Log-Rank Tests
  - Cox Proportional Hazards Model
  - Additional Topics in Clinical Trials
-

# Example 1: MACS Study

- Multicenter AIDS Cohort Study (MACS) enrolled more than 3,000 men who were at risk for HIV acquisition.
- A subset of  $N = 479$  men were observed to seroconvert.
- |           |
|-----------|
| infection |
|-----------|

 → 

immune decline
----------------

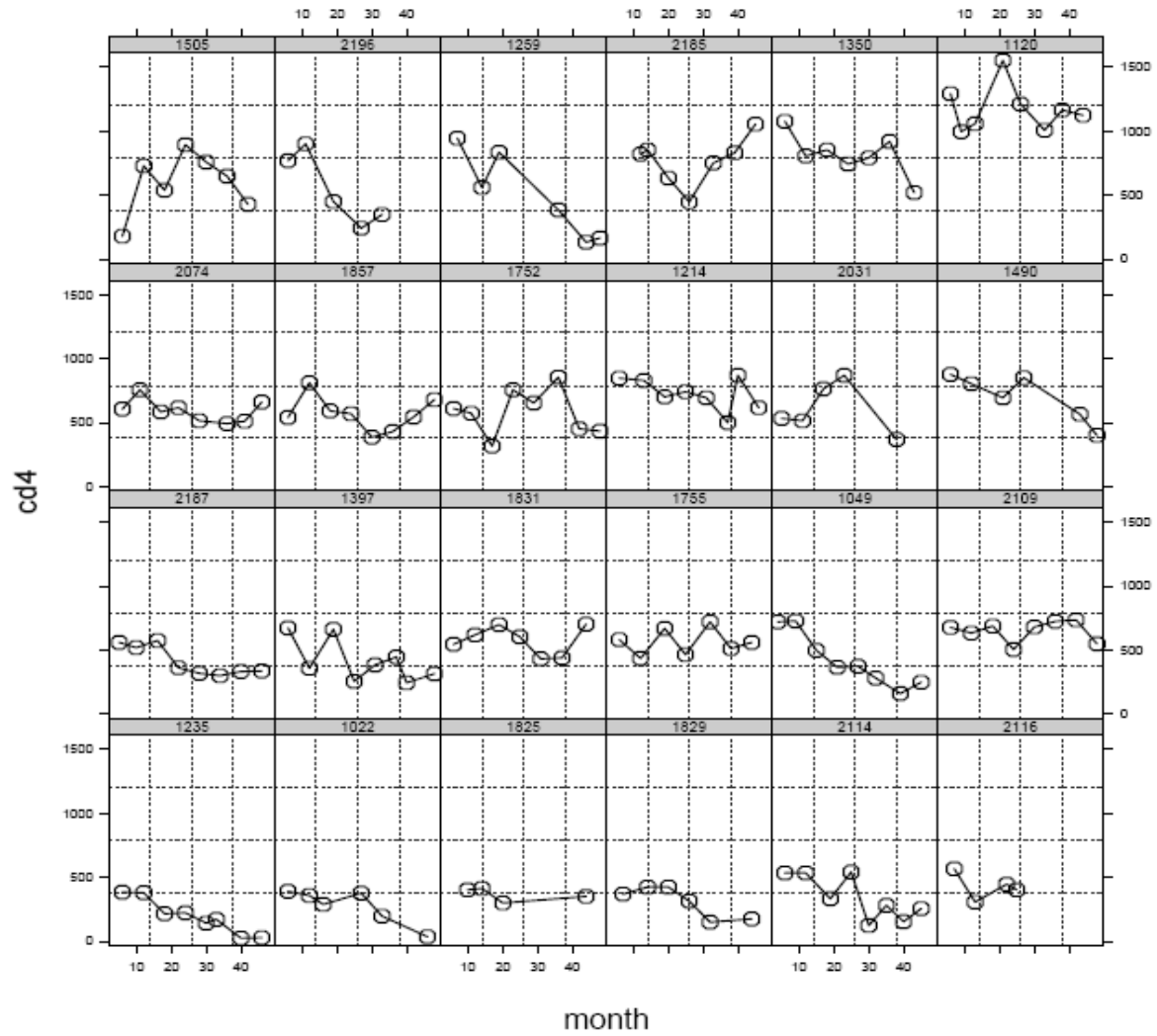
 → 

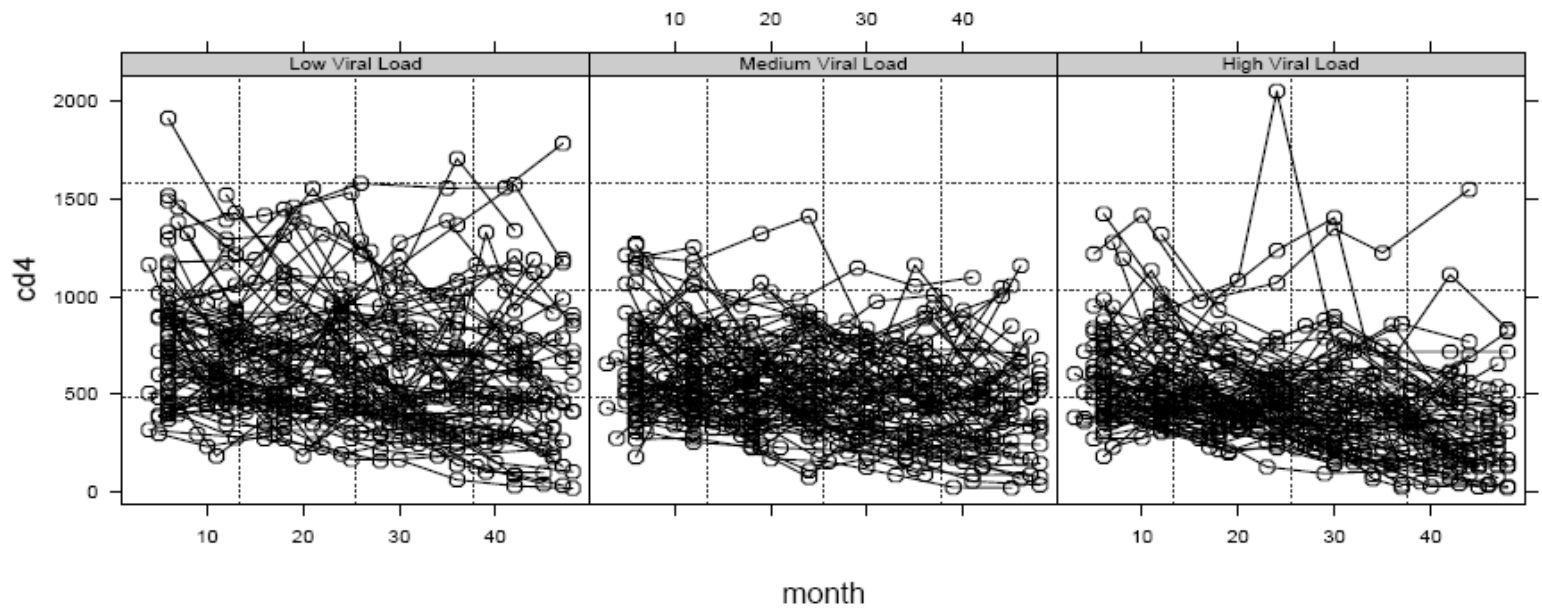
AIDS
------

 → 

Death
-------
- **Q:** Do baseline factors such as CD4 count or viral load predict the subsequent rate at which the immune measures decline?
- **Q:** Do baseline factors such as CD4 count or viral load predict the time until AIDS or death?
- **Q:** Do baseline factors such as CD4 count or viral load and time-varying health measures predict the time until AIDS or death?

Reference: Kaslow et al. (1987) *AJE*, 126: 310-318.





# Example 2: A breast cancer data

## Population:

- Women aged 20-44
- CSS of western Washington, 1983-1992
- Subset of  $n = 253$  with diploid tumor
- Median follow-up as of 5/97: 62 months
- Some delays from diagnosis to enrollment (left truncation)

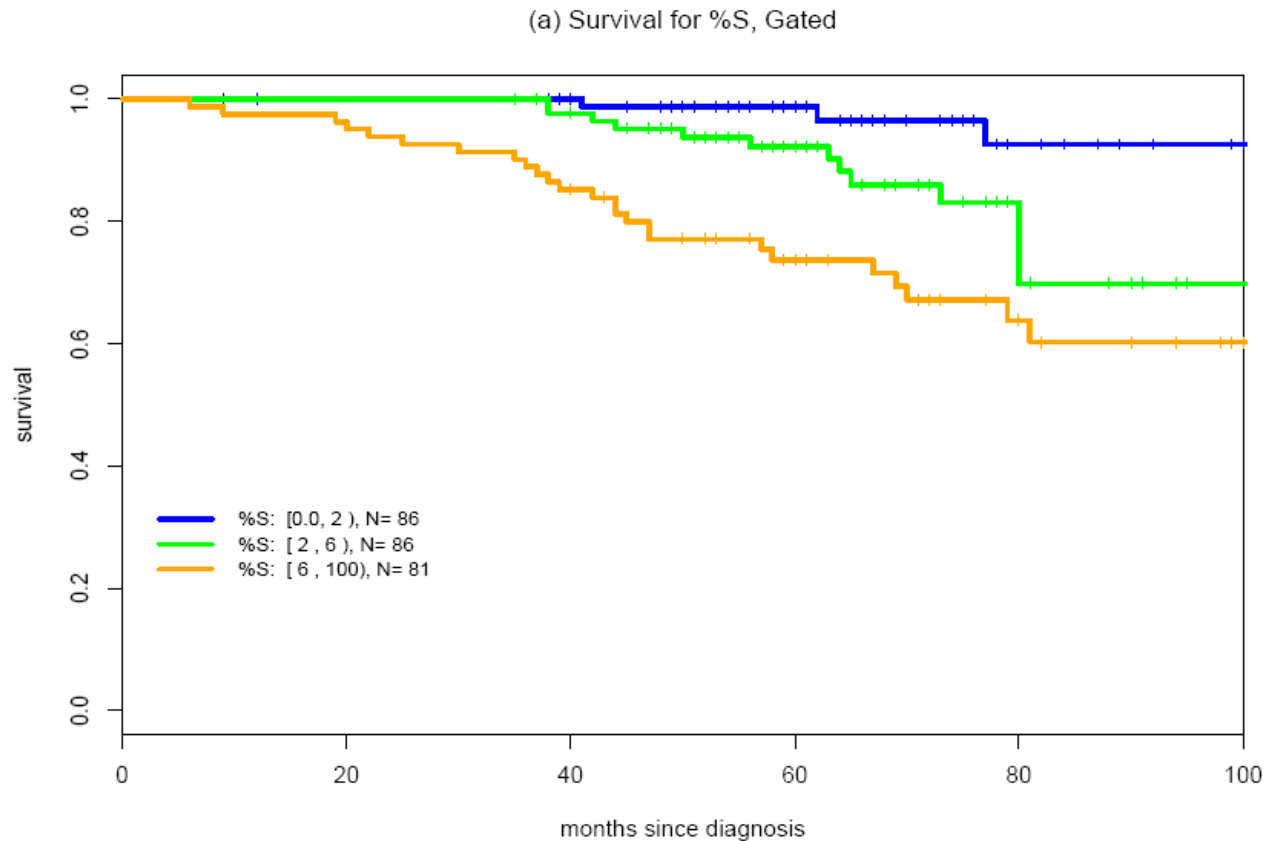
## Question(s):

- Tumor characteristics as predictors?
- Improved cytometry measurements?  
⇒ cytocarotine gating

Reference: Porter et al. (1997) *Nature Medicine*, 3: 222-225.



# How to plot a figure like this?



---

# Example 3: Mayo PBC Data

- Study design
    - Double-blinded randomized trial in primary biliary cirrhosis of the liver (PBC)
    - 1974-1986: 312 subjects randomized
  - Endpoints
    - Disease and vital status: 125 deaths
  - Measurements recorded
    - Clinical
    - Histological
    - Biochemical
    - Serological
  - Scientific questions
    - Compare mortalities between D-penicilamine (DPCA) and control treatment
    - Identify significant prognostic factors in study of natural history
-

## Original Articles

# Prognosis in Primary Biliary Cirrhosis: Model for Decision Making

E. ROLLAND DICKSON, PATRICIA M. GRAMBSCH, THOMAS R. FLEMING,† LLOYD D. FISHER† AND ALICE LANGWORTHY

Division of Gastroenterology and Internal Medicine and Section of Biostatistics, Mayo Clinic and Mayo Foundation, Rochester, Minnesota 55905

The ideal mathematical model for predicting survival for individual patients with primary biliary cirrhosis should be based on a small number of inexpensive, non-invasive measurements that are universally available. Such a model would be useful in medical management by aiding in the selection of patients for and timing of orthotopic liver transplantation. This paper describes the development, testing and use of a mathematical model for predicting survival. The Cox regression method and comprehensive data from 312 Mayo Clinic patients with primary biliary cirrhosis were used to derive a model based on patient's age, total serum bilirubin and serum albumin concentrations, prothrombin time and severity of edema. When cross-validated on an independent set of 106 Mayo Clinic primary biliary cirrhosis patients, the model predicted survival accurately. Our model was found to be comparable in quality to two other primary biliary cirrhosis survival models reported in the literature and to have the advantage of not requiring liver biopsy.

Orthotopic liver transplantation is considered to be potentially life-saving for selected patients with advanced or end-stage primary biliary cirrhosis. The availability of a model to predict survival probability for an individual patient would improve selection of patients for transplantation and the timing of that transplantation. Also, such a model could be used to help to decide which patients are appropriate, medically and ethically, for clinical trials of other treatment modalities. In addition, the model could be used for education and counseling of the patient and the family.

Using the Cox proportional hazards regression procedure (1), Roll et al. at Yale (2) and Christensen et al. in Europe (3) independently developed multivariate survival models. The Yale model used patient's age, serum bilirubin concentration, hepatomegaly and presence of portal fibrosis or cirrhosis to predict survival. The European model used age, bilirubin and albumin concentra-

tions, presence of cirrhosis, presence of cholestasis and whether or not azathioprine was prescribed. However, neither model was developed as a medical management tool, and both models required liver biopsy.

This paper describes a pragmatic model based on inexpensive, noninvasive measurements that are universally and readily available.

### PATIENTS AND METHODS

#### Patient Population

To develop the model, we used natural history data on the 312 primary biliary cirrhosis patients enrolled in either of two double-blind, placebo-controlled, randomized clinical trials at the Mayo Clinic evaluating the use of D-penicillamine for treating primary biliary cirrhosis. To be eligible for these trials, patients had to meet well-established clinical, biochemical, serologic and histologic criteria for primary biliary cirrhosis (4). Patient accrual took place from January, 1974, through May, 1984. One clinical trial (unpublished data) involved patients with histologic Stage 1 or 2 primary biliary cirrhosis; the other involved Stage 3 and 4 patients (4). Both trials found no therapeutic differences between control and D-penicillamine-treated patients. The study protocols required that no patient be taking any antiinflammatory or immunosuppressive medication (other than the study capsule). Therefore, it was deemed appropriate to combine all study participants to determine the natural history of primary biliary cirrhosis.

In addition, we had available 112 patients who were eligible for the trials but declined to participate. None of these patients was taking an immunosuppressive or antiinflammatory medication at the time of trial eligibility. These patients were used for model validation. It is possible that some of the cross-validation patients were exposed to antiinflammatory or immunosuppressive medication during the follow-up period. However, there has been no report of a totally effective regimen for biliary cirrhosis (5). Therefore, it is unlikely that the natural course of their disease was altered by any medication.

#### Data Collection

A comprehensive clinical and laboratory data base was established on each patient. The data were collected prospectively in the trial patients, by using standardized forms, definitions, and study protocols, at entry and at yearly intervals (see Table 1 for the variables measured). For the nontrial patients, the baseline data were collected from patients' records.

At entry, a liver biopsy specimen was obtained, and the

Received June 28, 1988; accepted December 12, 1988.

Supported by Research Grant AM-34238 from the National Institutes of Health.

† Present address: Department of Biostatistics, University of Washington, Seattle, Washington.

Address reprint requests to: E. Rolland Dickson, M.D., Mayo Clinic, 200 First St. SW, Rochester, Minnesota 55905.

# How to understand a table like this?

- Prognostic score derived as linear predictor from Cox model:

$$\lambda(t \mid X_{\text{bili}}, X_{\text{prottime}}, X_{\text{edema}}, X_{\text{alb}}, X_{\text{age}})$$

Covariate	estimate	s.e.	Z
log(bilirubin)	0.928	0.099	9.401
log(prothrombin time)	0.076	0.111	0.678
edema	0.967	0.300	3.221
albumin	-0.961	0.240	-4.001
age	0.036	0.009	4.243

---

# Key words

- Time, time, time
  - Kaplan-Meier
  - Log-rank
  - Cox
  - More?
    - Censoring
    - Truncation
    - Recurrent events
    - Case-cohort, nested case-control
    - Length-bias
-

---

# What's so special about time?

- A logical reflection of repetition of events
  - Ordering and Uni-directional
    - Natural history
    - Causality
  - The ultimate health outcome?
-

---

# How to measure time?

- Measuring time
    - Length of time or time-to-event
      - Time zero/time origin/starting time
      - Event time/endpoint
    - Notation
      - $T$  : a non-negative random variable
    - Issues
      - Staggered entry
      - Length-bias or Truncation
      - Censoring
-

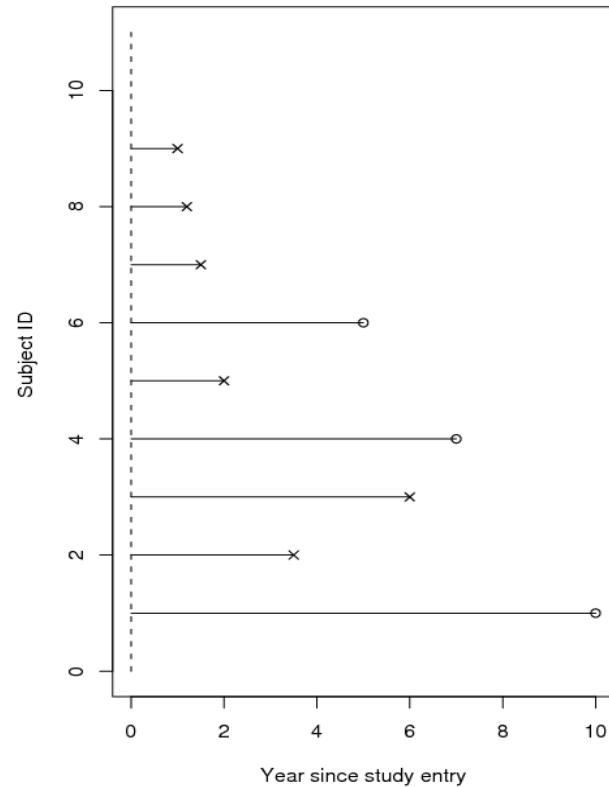
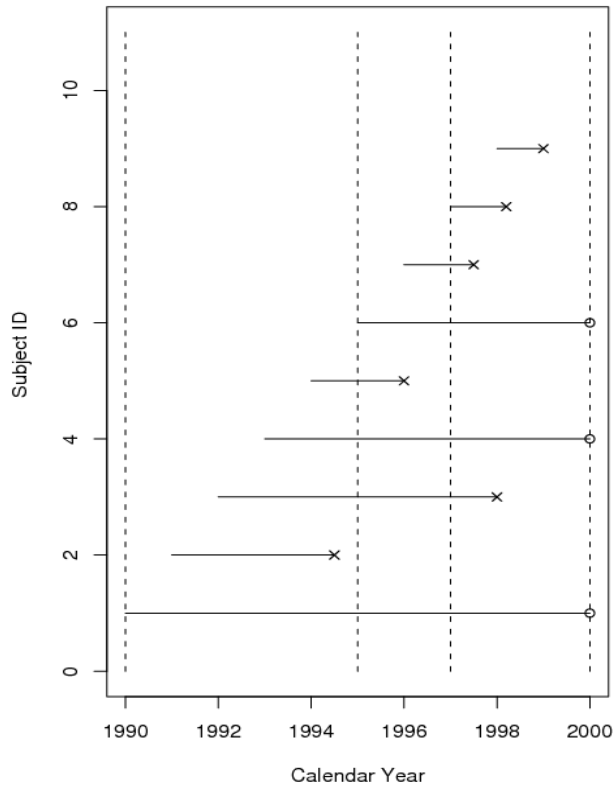
---

# Censoring

- Censoring happens
    - When full length of time is not observable
  - Notation
    - $C$  : censoring time
  - Types of censoring
    - Right and left censoring
    - Interval (doubly) censoring
    - Current status
    - Type I/II and random censoring
    - Examples
      - Lost-to-followup
      - Competing risks
      - Administrative censoring
-



# A typical right-censored dataset



---

# Survival datasets

- How to organize our data?
    - $(X_i, \Delta_i, Z_i), i = 1, 2, \dots, n$
    - $X_i = \min(T_i, C_i), \Delta = I(T_i \leq C_i)$
  - What to do with a survival dataset?
    - Estimate the distribution of time-to-event outcomes
    - Model distributions to identify association
-

---

# How to deal with censoring?

- Naïve approaches

- Ignoring them: simply remove them from our analysis dataset
- Keeping them: assume every censored subject would fail at the end of study

- Better approaches

- Impute censoring: can we impute a failure time for the censored subject, and then analyze dataset based on both imputed and observed failure times?
  - Not impute but treat censoring as if it were “nuisance”
-

# Life-tables

## ■ Data Example

Years	Alive at beginning	Deaths	Censored
0-1	146	27	3
1-2	116	18	10
2-3	88	21	10
3-4	57	9	3
4-5	45	1	3
		76	29
5-6	41	2	11
6-7	28	3	5
7-8	20	1	8
8-9	11	2	1
9-10	8	2	6

## ■ Question

- How do we estimate 5-year survival?

$$S(5) = \Pr(T \geq 5)$$

- Naïve estimates
  - $1 - 76/146 = 47.9\%$
  - $1 - 76/(146 - 29) = 35\%$

# Life-tables

- Better Approaches (1)
  - Assume censoring occurred at the right end of interval

$t$	$n$	$d$	$w$	$q^r = d/n$	$p^r = 1 - q^r$	$\hat{S}^r = \prod p^r$
0-1	146	27	3	0.185	0.815	0.815
1-2	116	18	10	0.155	0.845	0.689
2-3	88	21	10	0.239	0.761	0.524
3-4	57	9	3	0.158	0.842	0.441
4-5	45	1	3	0.022	0.972	0.432

# Life-tables

## ■ Better Approaches (2)

- Assume censoring occurred immediately prior to the left end of interval

$t$	$n$	$d$	$w$	$q^l = d/(n - w)$	$p^l = 1 - q^l$	$\hat{S}^l = \prod p^l$
0-1	146	27	3	0.189	0.811	0.811
1-2	116	18	10	0.170	0.830	0.673
2-3	88	21	10	0.269	0.731	0.492
3-4	57	9	3	0.167	0.833	0.410
4-5	45	1	3	0.024	0.977	0.400

# Life-tables

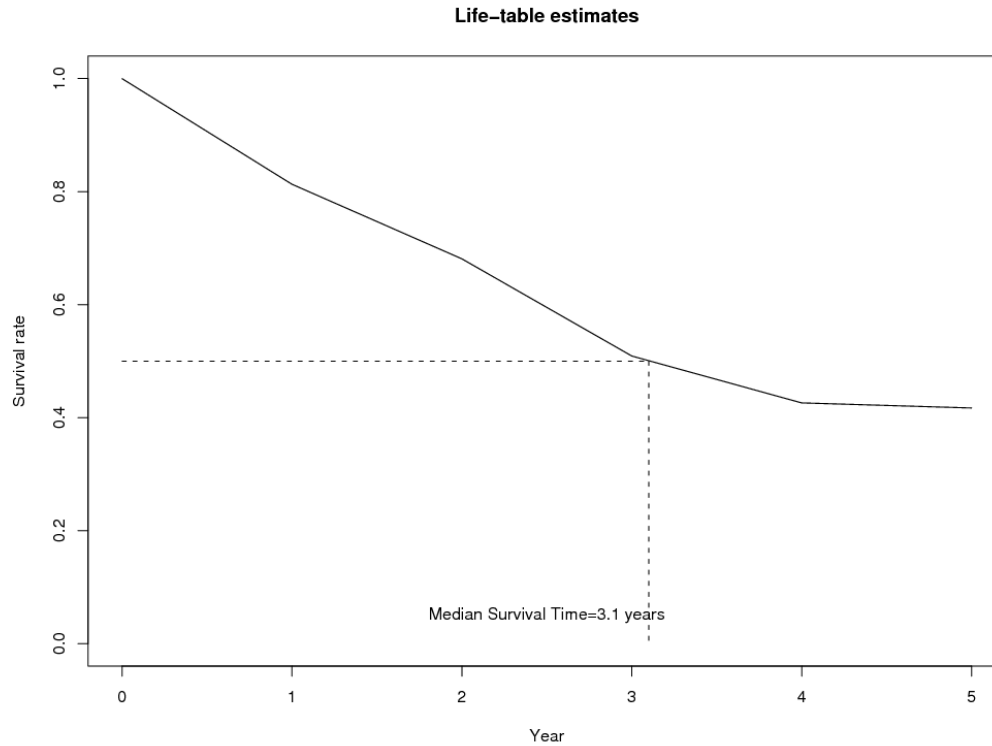
## ■ Better Approaches (3)

- Assume half of censoring occurred immediately prior to the left end of interval and half of censoring occurred at the right end of interval

$t$	$n$	$d$	$w$	$q = d/(n - w/2)$	$p = 1 - q$	$\hat{S} = \prod p$
0-1	146	27	3	0.187	0.813	0.813
1-2	116	18	10	0.162	0.838	0.681
2-3	88	21	10	0.253	0.747	0.509
3-4	57	9	3	0.162	0.838	0.426
4-5	45	1	3	0.023	0.977	0.417

# Life-tables

- Plot of life-table estimates of survival function





---

# Life-tables

- Why better approaches
    - Bias reduction
  - What's next?
    - Error bound
    - Variance calculation
      - Determine how reliable a life-table estimate is
      - This may need heavy theory development
-

# Variance of life-table estimate

1.  $\hat{S}(t) = \prod p \implies \log \hat{S}(t) = \sum \log p$  This is the usual technique to deal with products
2.  $\hat{\text{var}}(p) = pq/(n - w/2)$
3.  $\hat{\text{var}}(\log p) \approx \{(\log p)'_p\}^2 \text{var}(p) = q/\{p(n - w/2)\}$  This is called Delta-method
4.  $\text{var}\{\log \hat{S}(t)\} \approx \sum_{j \leq t} d_j / \{(n_j - w_j/2)(n_j - d_j - w_j/2)\}$
5.  $S(t) = \exp\{\log S(t)\} = \exp\{-\Lambda(t)\}$ , where  $\Lambda(t) = -\log S(t)$
6.  $\text{var}\{\hat{S}(t)\} = ([\exp\{-\Lambda(t)\}]'_\Lambda)^2 \text{var}\{\hat{\Lambda}(t)\}$
7. Standard errors for life-table estimates: Greenwood's formula

$$se\{\hat{S}(t)\} = \hat{S}(t) \left\{ \sum_{j=1}^t \frac{d_j}{(n_j - w_j/2)(n_j - d_j - w_j/2)} \right\}^{1/2}$$

8. 95% confidence intervals for  $S(t)$  is  $\hat{S}(t) \pm 1.96 \times se[\hat{S}(t)]$

# Variance of life-table estimates

Life-table example

$t$	$n$	$d$	$w$	$\hat{S}$	$\text{var}(\log p)$	$se$
0-1	146	27	3	0.813	0.00159	0.0324
1-2	116	18	10	0.681	0.00174	0.0392
2-3	88	21	10	0.509	0.00408	0.0438
3-4	57	9	3	0.426	0.00349	0.0444
4-5	45	1	3	0.417	0.00054	0.0445

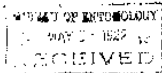
95% confidence interval for  $S(5)$ :

$$0.417 \pm 1.96 \times 0.0445 = (0.331, 0.503)$$

- This life table is not finished
  - See Problem Set 1, Exercise 1
  - STATA: ltable

# Life-tables 100 years ago

DEPARTMENT OF COMMERCE  
BUREAU OF THE CENSUS  
SAM L. ROGERS, Director



## UNITED STATES LIFE TABLES

1890, 1901, 1910, and 1901-1910

Explanatory Text, Mathematical Theory, Computations,  
Graphs, and Original Statistics

ALSO

Tables of United States Life Annuities  
Life Tables of Foreign Countries  
Mortality Tables of Life Insurance Companies

Prepared by JAMES W. CLOVER  
Expert Special Agent of the Bureau of the Census



WASHINGTON  
GOVERNMENT PRINTING OFFICE  
1921

52

### UNITED STATES LIFE TABLES.

TABLE I  
LIFE TABLE FOR BOTH SEXES IN THE  
BASED ON THE ESTIMATED POPULATION JULY 1, 1901 (20,000,000), AND ON THE

Form. The original statistics from which these tables were prepared are: Census of the United States, 1901, New York, 1902; Census of the United States, 1910, New York, 1911; Census of the United States, 1920, New York, 1921.

AGE INTERVAL	Of 100,000 Persons Born Alive		Rate of Mortality		Stationary Population	Stationary Population		Average length of life	Average number of years lived after age
	$l_x$	$d_x$	$q_x$	$Q_x$		At birth	At age $x$		
0-1	100,000	1,000	1.00	1,000	100,000	100,000	100,000	75.00	75.00
1-2	99,000	1,000	1.01	1,010	99,000	99,000	99,000	74.99	74.99
2-3	98,000	1,000	1.02	1,020	98,000	98,000	98,000	74.98	74.98
3-4	97,000	1,000	1.03	1,030	97,000	97,000	97,000	74.97	74.97
4-5	96,000	1,000	1.04	1,040	96,000	96,000	96,000	74.96	74.96
5-6	95,000	1,000	1.05	1,050	95,000	95,000	95,000	74.95	74.95
6-7	94,000	1,000	1.06	1,060	94,000	94,000	94,000	74.94	74.94
7-8	93,000	1,000	1.07	1,070	93,000	93,000	93,000	74.93	74.93
8-9	92,000	1,000	1.08	1,080	92,000	92,000	92,000	74.92	74.92
9-10	91,000	1,000	1.09	1,090	91,000	91,000	91,000	74.91	74.91
10-11	90,000	1,000	1.10	1,100	90,000	90,000	90,000	74.90	74.90
11-12	89,000	1,000	1.11	1,110	89,000	89,000	89,000	74.89	74.89

# Life-tables 100 years later

National Vital Statistics Reports



Volume 57, Number 1

August 5, 2008

## U.S. Decennial Life Tables for 1999–2001, United States Life Tables

by Elizabeth Arias, Ph.D.; Lester R. Curtin, Ph.D.; Rong Wei, Ph.D.; and Robert N. Anderson, Ph.D.

### Abstract

**Objectives**—This report presents period life tables for the United States based on age-specific death rates for the period 1999–2001. These tables are the most recent in a 100-year series of decennial life tables for the United States.

**Methods**—This report presents complete life tables by age, race (white and black), and sex. Also presented are standard errors of the probability of dying and life expectancy. The data used to prepare these life tables are population estimates based on the 2000 decennial census, deaths occurring in the United States to U.S. residents in the 3 years 1999–2001, counts of U.S. resident births in the years 1997–2001, and population and death counts from the Medicare program for years 1999–2001.

**Results**—In 1999–2001, life expectancy at birth was 76.83 years for the total U.S. population, representing an increase of 27.59 years from a life expectancy of 49.24 years in 1900. Between 1900 and 2000, life expectancy increased by 40.06 years for black females (from 35.04 to 75.12), by 35.54 years for black males (from 32.54 to 68.08), by 28.89 years for white females (from 51.08 to 79.97), and by 26.51 years for white males (from 48.23 to 74.74).

### Introduction

The life tables presented in this report are the most recent in a series of decennial life tables for the United States that dates to the beginning of the 20th century. The 1999–2001 life tables are the 11th in the decennial series. The reporting of deaths at the national level began in 1900 with 10 states and the District of Columbia. As the quality of the reporting improved, states were added to the death-registration area. Beginning with the period 1929–1931, the decennial life tables were produced using data for all of the coterminous United States. Alaska and Hawaii were included beginning with the 1959–1961 decennial life tables. Each set of life tables is based on population data from a decennial census and reported deaths of the 3-year period surrounding the census year (the census year is the

middle year in all but the first in the series, in which deaths for 1900–1902 were used because death reports for 1899 were not collected (1)). The decennial life tables differ in one main respect from the life tables prepared and published annually in the Centers for Disease Control and Prevention's National Center for Health Statistics' (NCHS) National Vital Statistics Reports. The annual tables are based on deaths in a single year and, except for census years, on postcensal population estimates rather than on the data from a decennial census.

This report is the first of a series of reports containing life tables for 1999–2001 and other information related to the decennial life table program. Also included in the series is a methodological report that describes in detail the methods employed to estimate the national life tables, a report on national life tables analyzed by major groups of causes of death, and a report containing life tables for individual states and the District of Columbia, including a description of the methods used to estimate individual state life tables.

### Data and Methods

Mortality rates for a specific period may be summarized by the life table method to obtain measures of comparative longevity. There are two types of life tables—the cohort (or generation) life table and the period (or current) life table. The cohort life table provides a longitudinal perspective in that it follows the mortality experience of an actual cohort—for example, all persons born in the year 1945—from the moment of birth through consecutive ages in successive years. On the basis of age-specific death rates observed during consecutive years, the cohort life table reflects the mortality experience of a cohort from birth until no lives remain in the group. To prepare just a single complete cohort life table requires data over many years. Constructing cohort life tables entirely on the basis of observed data for real cohorts is usually not feasible because of data unavailability or incompleteness (2–4).

In contrast, the period life table does not represent the mortality experience of an actual birth cohort. Rather, the period life table

Table 1. Life table for the total population: United States, 1999–2001

Age	Probability of dying between ages $x$ to $x+n$	Number surviving to age $x$	Number dying between ages $x$ to $x+n$	Person-years lived between ages $x$ to $x+n$	Total number of person-years lived above age $x$	Expectation of life at age $x$
$x$ to $x+n$	$q_x$	$l_x$	$d_x$	$L_x$	$T_x$	$e_x$
<b>Days</b>						
0-1	0.00275	100,000	275	274	7,883,447	76.83
1-7	0.00094	99,725	93	1,839	7,883,173	77.04
7-98	0.00095	99,631	94	5,729	7,881,534	77.10
98-99	0.00233	99,537	232	91,794	7,875,935	77.12
<b>Years</b>						
0-1	0.00696	100,000	696	99,436	7,883,447	76.83
1-2	0.00650	99,305	50	99,280	7,584,011	76.37
2-9	0.00033	99,255	33	99,229	7,484,731	75.41
9-4	0.00026	99,222	25	99,210	7,385,492	74.43
4-5	0.00021	99,197	21	99,196	7,286,282	73.45
5-6	0.00019	99,176	19	99,167	7,187,096	72.47
6-7	0.00017	99,158	17	99,149	7,087,925	71.48
7-8	0.00016	99,140	15	99,132	6,988,790	70.49
8-9	0.00015	99,124	15	99,117	6,889,648	69.51
9-10	0.00013	99,110	13	99,103	6,790,531	68.52
10-11	0.00012	99,097	12	99,091	6,691,427	67.53
11-12	0.00012	99,085	12	99,079	6,592,336	66.53
12-13	0.00016	99,073	16	99,065	6,493,257	65.54
13-14	0.00024	99,057	24	99,045	6,394,192	64.55
14-15	0.00036	99,033	36	99,016	6,295,147	63.57
15-16	0.00048	98,998	48	98,974	6,196,131	62.59
16-17	0.00060	98,950	59	98,921	6,097,157	61.62
17-18	0.00070	98,891	69	98,856	5,998,226	60.66
18-19	0.00077	98,822	76	98,783	5,899,380	59.70
19-20	0.00082	98,745	81	98,705	5,800,597	58.74
20-21	0.00088	98,664	87	98,620	5,701,892	57.79
21-22	0.00093	98,577	92	98,531	5,603,272	56.84
22-23	0.00096	98,485	95	98,438	5,504,740	55.89
23-24	0.00097	98,390	95	98,343	5,406,203	54.95
24-25	0.00095	98,295	93	98,248	5,307,960	54.00
25-26	0.00092	98,202	91	98,157	5,209,712	53.05
26-27	0.00091	98,111	89	98,067	5,111,555	52.10
27-28	0.00090	98,022	89	97,978	5,013,488	51.15
28-29	0.00092	97,934	90	97,889	4,915,510	50.19
29-30	0.00096	97,844	94	97,797	4,817,621	49.24
30-31	0.00100	97,750	98	97,701	4,719,825	48.28
31-32	0.00105	97,652	103	97,601	4,622,124	47.33
32-33	0.00111	97,549	109	97,495	4,524,523	46.38
33-34	0.00119	97,441	116	97,382	4,427,028	45.43
34-35	0.00128	97,324	125	97,262	4,329,646	44.49
35-36	0.00138	97,199	134	97,132	4,232,384	43.54
36-37	0.00148	97,065	144	97,005	4,135,252	42.60
37-38	0.00159	96,921	155	96,844	4,038,259	41.67
38-39	0.00172	96,767	167	96,683	3,941,415	40.73
39-40	0.00187	96,600	181	96,520	3,844,732	39.80
40-41	0.00203	96,419	196	96,321	3,748,222	38.87
41-42	0.00220	96,223	212	96,116	3,651,902	37.95
42-43	0.00238	96,010	229	95,996	3,555,785	37.04
43-44	0.00257	95,782	247	95,858	3,459,868	36.12
44-45	0.00278	95,526	266	95,401	3,364,231	35.21
45-46	0.00291	95,259	288	95,124	3,268,890	34.31
46-47	0.00308	94,971	311	94,825	3,173,705	33.41
47-48	0.00333	94,670	335	94,502	3,078,880	32.52
48-49	0.00360	94,335	360	94,155	2,984,378	31.64
49-50	0.00387	94,007	384	93,875	2,890,223	30.78
50-51	0.00427	93,591	410	93,385	2,796,441	29.98
51-52	0.00489	93,180	439	93,061	2,703,055	29.01
52-53	0.00565	92,741	471	92,508	2,610,095	28.14
53-54	0.00648	92,270	508	92,016	2,517,088	27.28
54-55	0.00738	91,762	552	91,486	2,425,573	26.43
55-56	0.00857	91,211	603	90,930	2,334,098	25.59
56-57	0.01024	90,607	661	90,277	2,243,177	24.76
57-58	0.01277	89,947	722	89,586	2,152,900	23.94
58-59	0.01671	89,225	784	88,833	2,063,314	23.12



---

# Modern survival analysis

- Modern survival analysis

- Post-1958

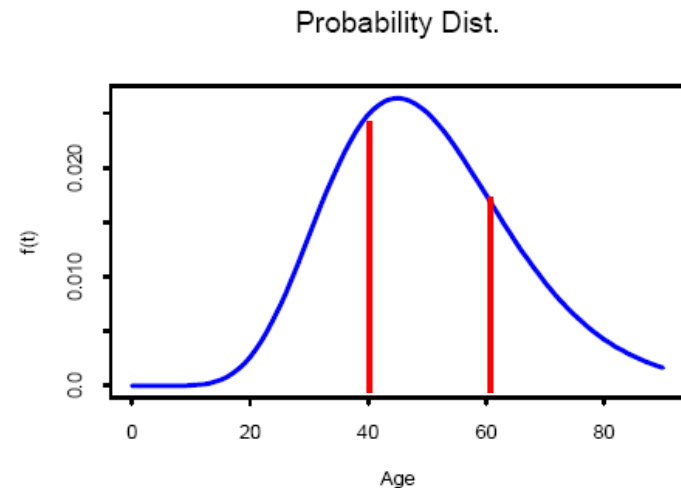
- 1958: Kaplan-Meier estimator (Kaplan & Meier, *Journal of the American Statistical Association*)
      - 1966: Log-rank test (Mantel, *Cancer Chemo. Rep.*)
      - 1972: Cox proportional hazards model (Cox, *Journal of the Royal Statistical Society, Series B*)
-

# Parametric Methods

- Functions to be studied?
  - Histogram: density function

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \Pr\{t \leq T \leq t + \Delta t\}$$

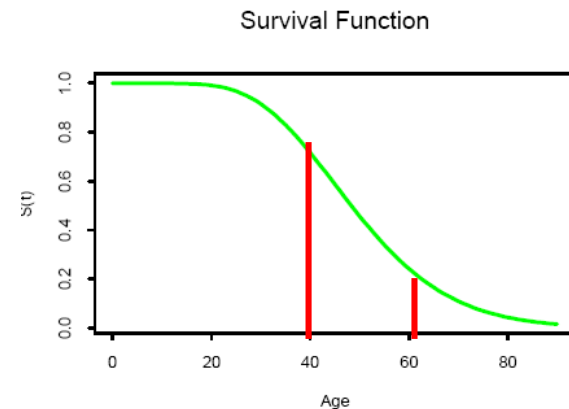
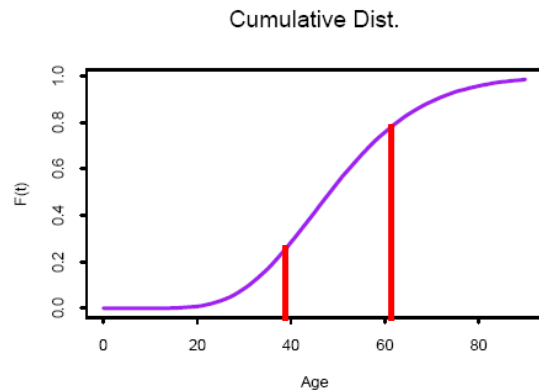
- Area under the curve
  - Probability of an event



## ■ Cumulative distribution (survival) function

$$\begin{aligned} F(t) &= P(T \leq t) \\ &= 1 - P(T > t) = 1 - S(t) \\ &= \int_0^t f(s) ds \end{aligned}$$

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) = 1 - F(t) \\ &= \int_t^{\infty} f(s) ds \end{aligned}$$

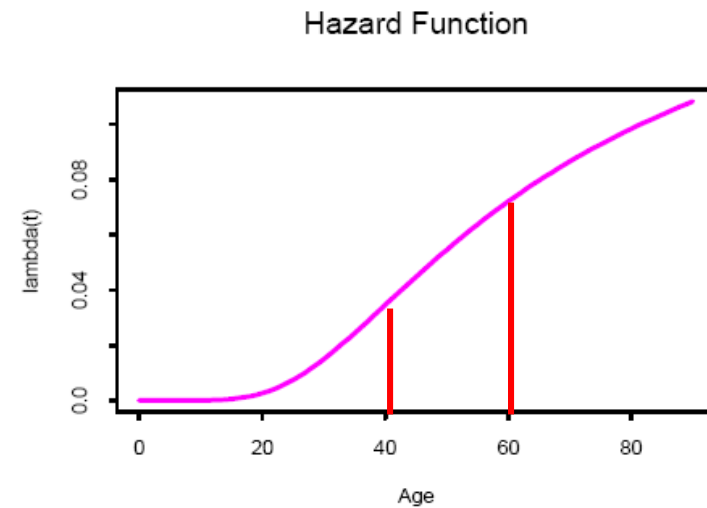




## ■ Hazard function

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t \leq T \leq t + \Delta t \mid T \geq t)$$

$$= f(t)/S(t)$$



## ■ Hazard function

### □ When times are discrete

$$\lambda(t) = P(T = t | T \geq t) = \frac{P(T = t)}{P(T \geq t)}$$

$$S(t) = \frac{P(T \geq x_2)}{P(T \geq x_1)} \cdot \frac{P(T \geq x_3)}{P(T \geq x_2)} \cdot \dots \cdot \frac{P(T \geq x_{j+1})}{P(T \geq x_j)}$$

$$= (1 - \lambda(x_1)) \cdot (1 - \lambda(x_2)) \dots (1 - \lambda(x_j))$$

$$= \prod_{i=1}^j (1 - \lambda(x_i))$$

---

## ■ Hazard function

### □ When times are continuous

$$\lambda(t) = \lim_{\Delta t \rightarrow 0^+} \frac{\Pr(\text{Failure occurring in } [t, t + \Delta t) | T \geq t)}{\Delta t}$$

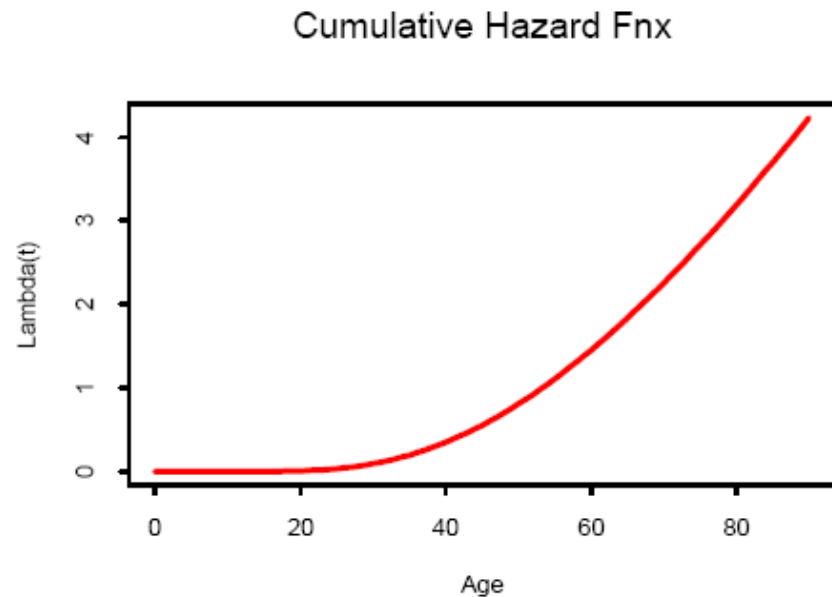
= Instantaneous failure rate at  $t$  given survival up to  $t$

$\lambda(t)\Delta t \approx$  the proportion of individuals experiencing failure in  $[t, t + \Delta t)$  to those surviving up to  $t$

---

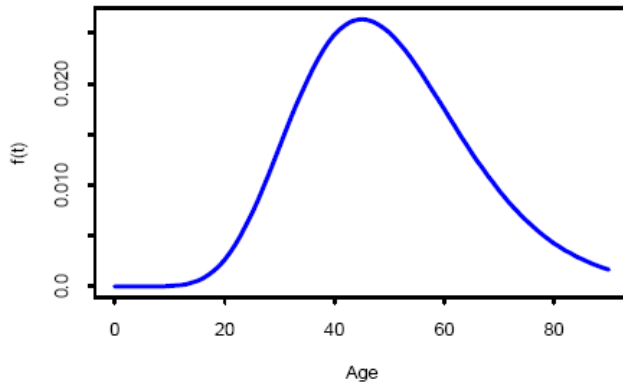
## ■ Cumulative hazard functions

$$\Lambda(t) = \int_0^t \lambda(s) ds$$

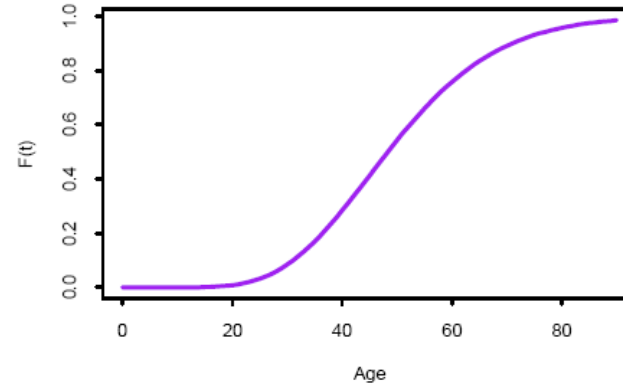


# Summary of functions

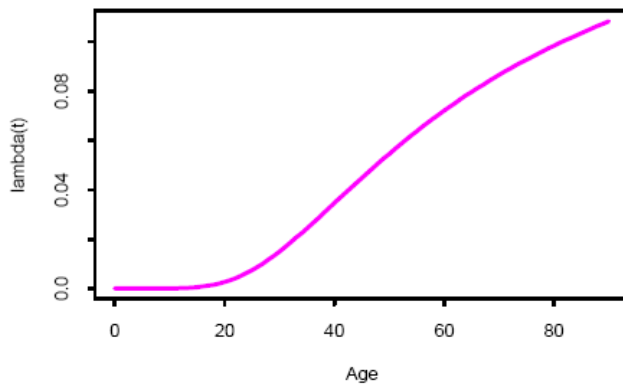
Probability Dist.



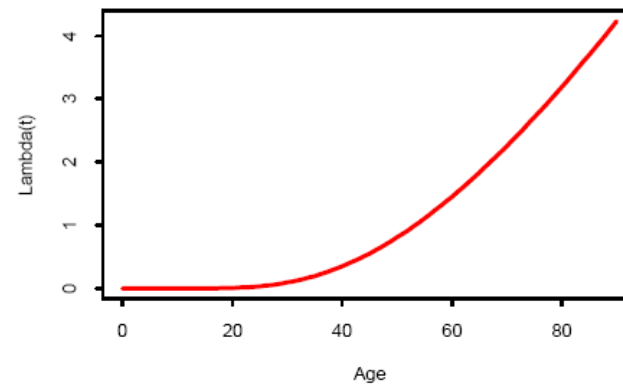
Cumulative Dist.



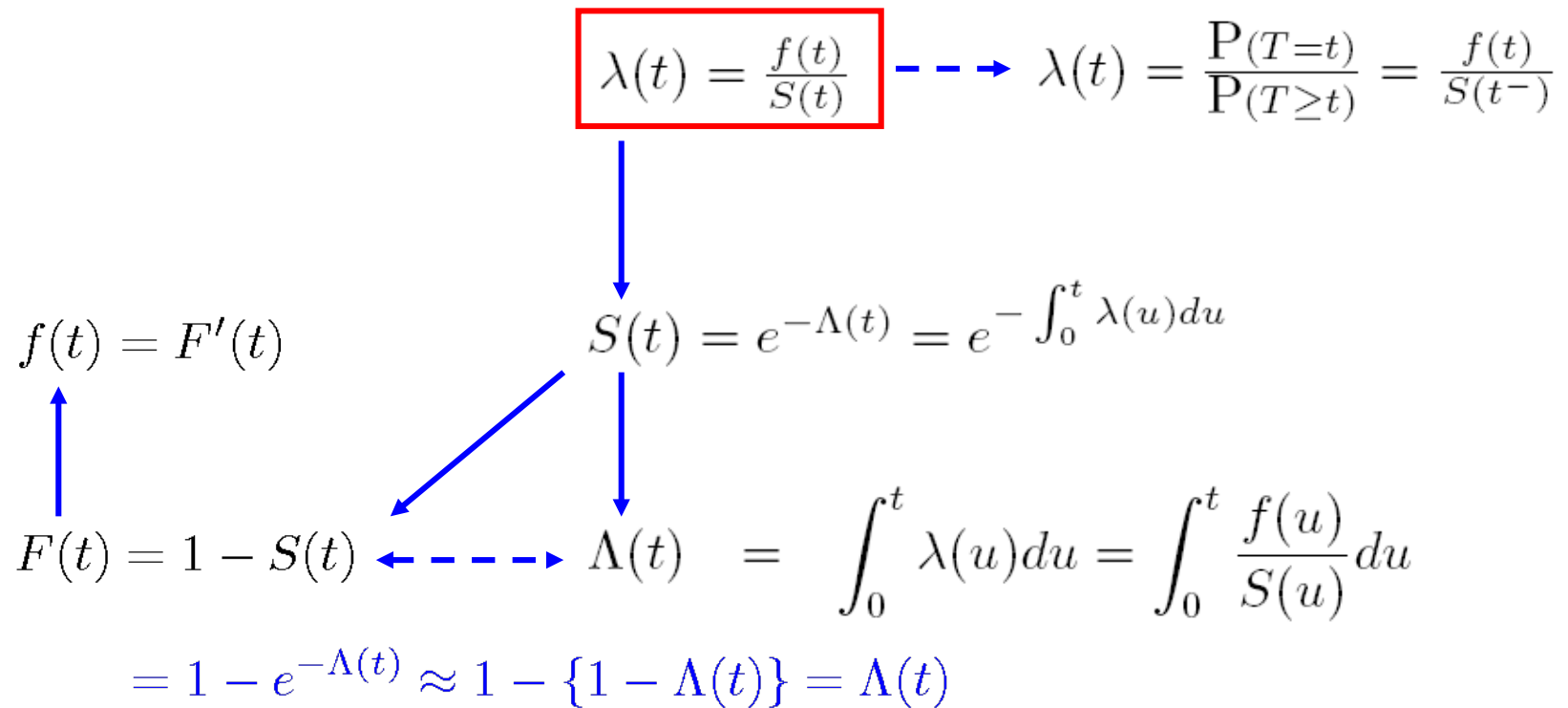
Hazard Function



Cumulative Hazard Fnx



# Relationship of functions



# Summary

- Density function:  $f(t) = \lim_{\Delta t \rightarrow 0} \Pr\{t \leq T \leq T + \Delta t\}$ 
    1.  $\int_0^{\infty} f(u)du = 1$
    2.  $F(t) = \int_0^t f(u)du, S(t) = \int_t^{\infty} f(u)du$
    3.  $\mu(t) = \int_0^t S(u)du, \mu = ET = \int_0^{\infty} S(u)du$
    4. Denominator: all the subjects
  - Hazard function:  $\lambda(t) = \lim_{\Delta t \rightarrow 0} \Pr\{t \leq T \leq T + \Delta t \mid T \geq t\} / \Delta t$ 
    1.  $\lambda(t) = f(t) / S(t)$
    2. Denominator: subjects survival up to  $t$
  - Mean residual life function:  $m(t) = E(T - t \mid T \geq t)$ 
    1. Residual time:  $R_T(t) = (T - t)I(T \geq t) = (T - t)^+ = (T - t) \vee 0$
    2.  $m(t) = \int_t^{\infty} S(u)du / S(t)$
-

# Examples

Example: Weibull distribution. The Weibull distribution with the parameters  $\theta > 0$  and  $\beta > 0$  assumes the parameterized survival function

$$S(t) = e^{-(\theta t)^\beta},$$

for  $t > 0$ . The density function is

$$f(t) = -\frac{dS_{\theta,\beta}(t)}{dt} = \beta\theta(\theta t)^{\beta-1}e^{-(\theta t)^\beta}$$

The hazard function is

$$\lambda(t) = \frac{f(t; \theta, \beta)}{S_{\theta,\beta}(t)} = \beta\theta(\theta t)^{\beta-1}.$$

Note that the hazard function  $\lambda(t)$  is constant if  $\beta = 1$ , increasing in  $t$  if  $\beta > 1$ , and decreasing in  $t$  if  $\beta < 1$ .



---

Example: Gamma distribution. The Gamma distribution with the parameters  $\lambda > 0$  and  $r > 0$  is a continuous distribution with the density function

$$f(t) = \frac{\lambda^r}{\Gamma(r)} t^{r-1} e^{-\lambda t} ,$$

for  $t \geq 0$ , where  $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$ . The survival and hazard functions can be derived from the density function. The mean of the Gamma distribution is  $r/\lambda$  and the variance is  $r/\lambda^2$ .

---

---

Example: Log-normal distribution. A random variable  $T$  is said to have a lognormal distribution with parameters  $-\infty < \mu < \infty$  and  $\sigma > 0$ . The probability density function of  $T$  is

$$f(t) = \frac{1}{\sigma(2\pi)^{1/2}} t^{-1} \exp\{-(\log t - \mu)^2 / 2\sigma^2\} ,$$

for  $t \geq 0$ , from which the survival and hazard functions can be derived.

---

# Parametric methods

- What's common in these distributions?
    - Distributions are completely defined by a fixed number of parameters. They are often called parametric distributions.
    - Parametric distributions are the critical core of the parametric methods.
    - If these distributions are correctly assumed, it means that
      - We may need relatively smaller sample size to estimate these parameters.
      - We may use the estimated parametric models in prediction, including extrapolation (certainly we shall be always cautious about it!)
-

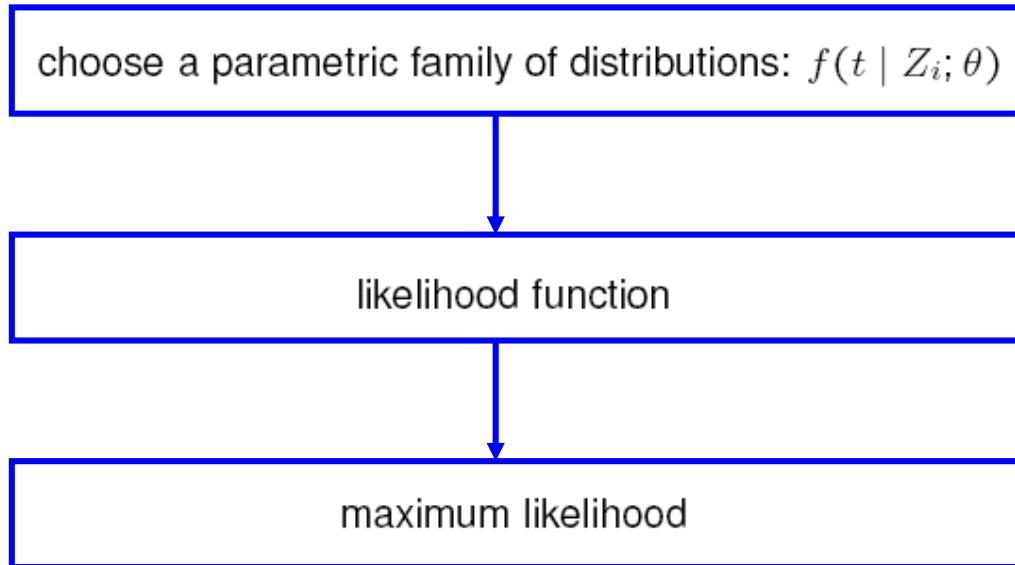
---

# Maximum likelihood estimation (MLE)

## ■ Rationale of MLE

- When we estimate a parameter, the estimate we would obtain should **maximize** the probability of what we actually observe
    - Even the true value of a parameter may **not** necessarily maximize the probability of what we actually observe
  - It is just one of the methods to estimate parameters
-

# Algorithm of MLE



---

# MLE for survival data

- No censoring

if no censoring,  $\{(T_i, Z_i); i = 1, 2, \dots, n\}$

– likelihood function based on observed  $T_i = t_i$  is

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(t_i | Z_i; \theta)$$



## ■ With censoring

if censored, survival data become  $\{(X_i, \Delta_i, Z_i); i = 1, 2, \dots, n\}$

- likelihood contribution is  $\Pr\{X_i = x_i, \Delta_i = \delta_i, Z_i = z_i\}$
- $\Delta_i = 1 \implies X_i = t_i$ , likelihood contribution is  $f_T(t_i | Z_i; \theta)S_C(t_i | Z_i)$
- $\Delta_i = 0 \implies X_i = c_i$ , likelihood contribution is  $f_C(c_i | Z_i)S_T(c_i | Z_i; \theta)$
- likelihood function based on observed data is then

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(x_i | Z_i; \theta)^{\delta_i} S(x_i | Z_i; \theta)^{1-\delta_i}$$

# Key assumptions

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(x_i | Z_i; \theta)^{\delta_i} S(x_i | Z_i; \theta)^{1-\delta_i}$$

What are the key assumptions?

- $Z_i$  do not have information about  $\theta$
- conditional independence: given  $Z_i$ ,  $T_i$  and  $C_i$  are independent
- examples
  - \* clinical trials
  - \*  $Z_i$ : treatment assignment
  - \*  $C_i$ : censoring due to side-effects



# Parameter estimation via MLE

Parameter estimation: maximum likelihood estimation (MLE)

– likelihood function:

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^n f(x_i | Z_i; \theta)^{\delta_i} S(x_i | Z_i; \theta)^{1-\delta_i} \\ &= \prod_{i=1}^n \lambda(x_i | Z_i; \theta)^{\delta_i} S(x_i | Z_i; \theta)\end{aligned}$$

This means we are taking  
“arguments,” to maximize  
it for an estimate

– parameter estimation:  $\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta) = \arg \max_{\theta} l(\theta)$

$$\begin{aligned}l(\theta) &= \log \mathcal{L}(\theta) \\ &= \sum_{i=1}^n [\delta_i \log \lambda(x_i | Z_i; \theta) + \log S(x_i | Z_i; \theta)] \\ &= \sum_{i=1}^n [\delta_i \log \lambda(x_i | Z_i; \theta) - \Lambda(x_i | Z_i; \theta)] \\ &= \sum_{i=1}^n \left[ \delta_i \log \lambda(x_i | Z_i; \theta) - \int_0^{x_i} \lambda(u | Z_i; \theta) du \right]\end{aligned}$$

It is equivalent to  
maximizing the log of  
likelihood function

Solve  $l'(\theta) = 0$  for the MLE  $\hat{\theta}$

# Statistical theory of MLE

- Consistency:

$$\hat{\theta} \rightarrow \theta_0$$

- Normality:

$$\frac{\hat{\theta} - \theta_0}{\sqrt{\text{var}(\hat{\theta})}} \sim \text{Normal}(0, 1)$$

where  $\text{var}(\hat{\theta})$  can be estimated by observed Fisher information

$$- \left\{ \frac{d^2}{d\theta^2} l(\hat{\theta}) \right\}^{-1}$$

**Example.**  $T \sim \exp(\theta)$ . The density function is  $f(t; \theta) = \theta e^{-\theta t} I(t > 0)$ .

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \theta e^{-\theta t_i} \\ \log L(\theta) &= \sum_{i=1}^n [\log \theta - \theta t_i] \\ U(\theta) &= \frac{d}{d\theta} \log L(\theta) = \sum_{i=1}^n \left[ \frac{1}{\theta} - t_i \right] = \frac{n}{\theta} - \sum_{i=1}^n t_i \end{aligned}$$

Thus  $\hat{\theta} = n / \sum_{i=1}^n t_i$  is the mle.

Note that the Fisher information is

$$I(\theta) = \text{E} \left[ -\frac{d^2}{d\theta^2} \log L(\theta) \right] = n/\theta^2. \text{ Thus}$$

$$\hat{\theta} - \theta \stackrel{\text{approx}}{\sim} N \left( 0, \frac{\theta^2}{n} \right) \text{ when } n \text{ is large}$$

or

$$\hat{\theta} \stackrel{\text{approx}}{\sim} N\left(\theta, \frac{\theta^2}{n}\right)$$

Thus  $\text{Prob}\left(\hat{\theta} - 1.96\frac{\theta}{\sqrt{n}} < \theta < \hat{\theta} + 1.96\frac{\theta}{\sqrt{n}}\right) \approx 95\%$ .

An asymptotic 95% confidence interval for  $\theta$  is

$$\left(\hat{\theta} - 1.96\frac{\hat{\theta}}{\sqrt{n}}, \hat{\theta} + 1.96\frac{\hat{\theta}}{\sqrt{n}}\right).$$

- What if there is censoring?
  - Problem Set 1, Exercise 2
  - STATA: `streg`; R/S+: `survreg`