SISCER 2022 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.
    - yqchensu@stanford.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, 3rd Ed.* Springer.
  - Hosmer, DW & Lemeshow, S (2008) *Applied Survival Analysis, 2nd Ed.* Wiley.
- Instruction hours
  - 8:30am – 10:00am, 10:30am – 12:00noon
- Student self-evaluation
  - Two Problem Sets
  - Four Quizzes
- Course calendar

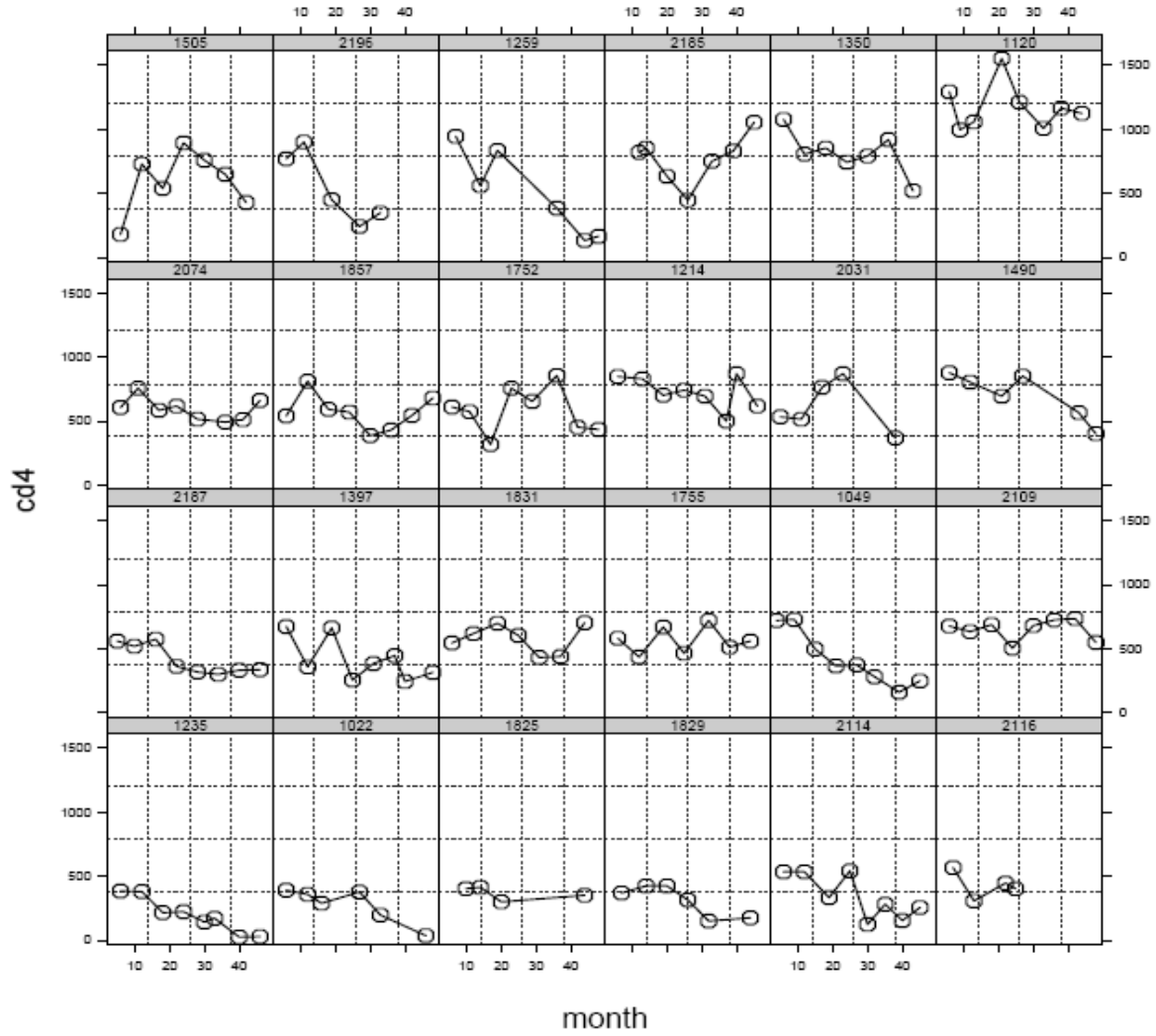
Course outlines

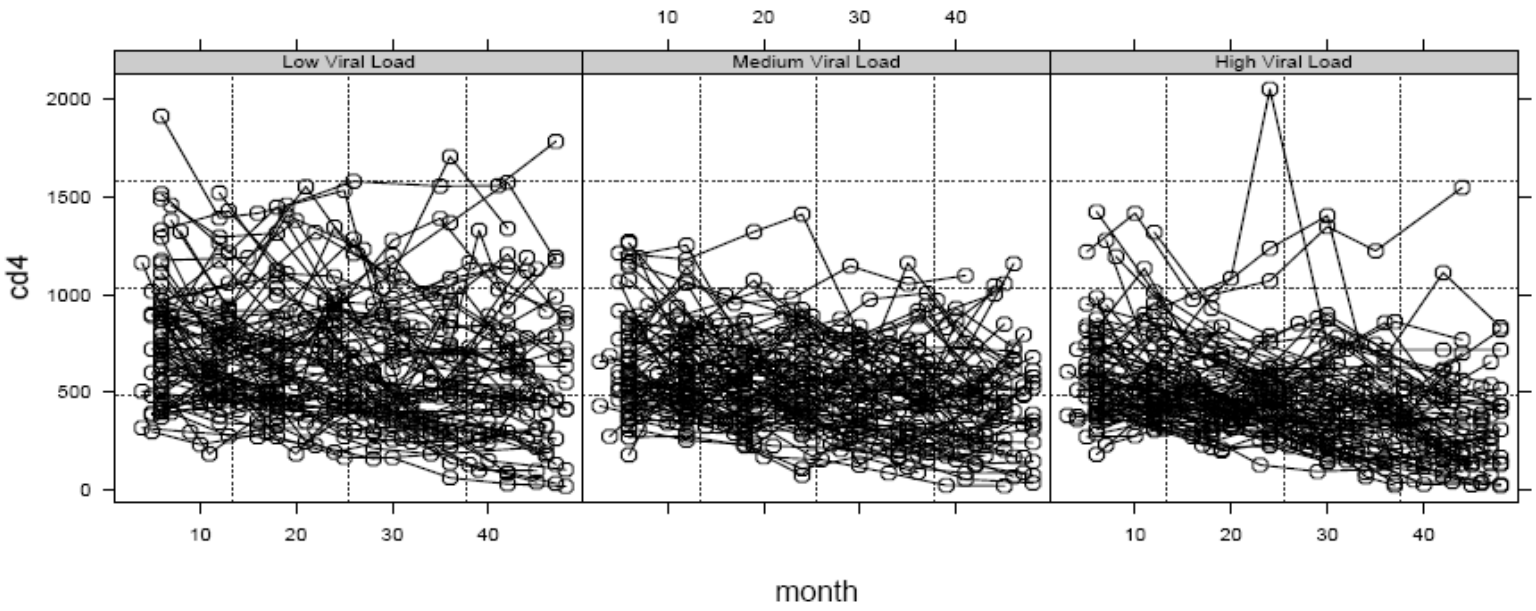
- Background
  - Time-to-event and censoring
  - Life tables
- Parametric Methods
  - Parametric distributions
  - Likelihood functions and maximum likelihood
- Kaplan-Meier Curves
- Log-Rank Tests
- Cox Proportional Hazards Model
- Additional Topics in Clinical Trials

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 $\rightarrow$ immune decline $\rightarrow$ AIDS $\rightarrow$ 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.


1505
2196
1259
2185
1350
1120
2074
1857
1752
1214
2031
1490
2187
1397
1831
1755
1049
2109
1235
1022
1825
1829
2114
2116
cd4
month


Low Viral Load
Medium Viral Load
High Viral Load
cd4
month
0
500
1000
1500
2000
10
20
30
40

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)

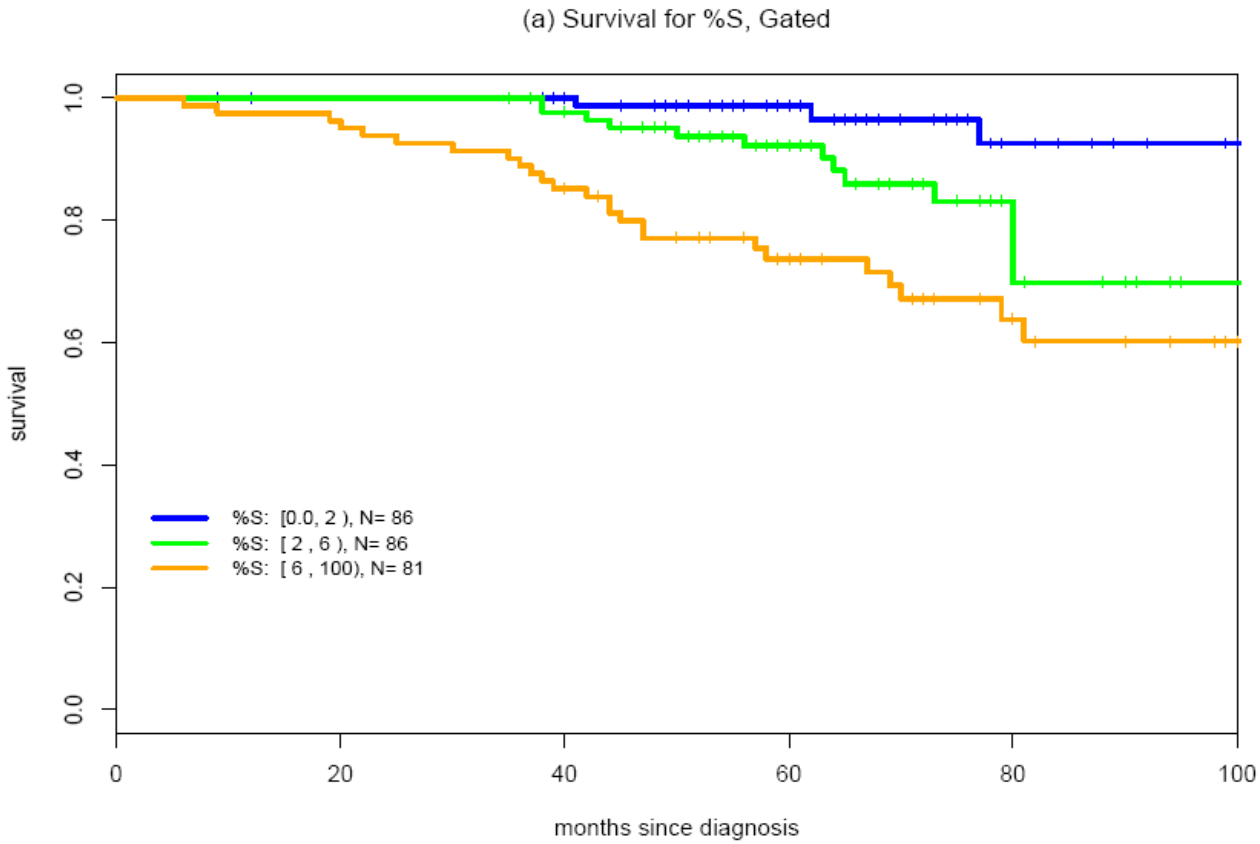
Question(s):

- Tumor characteristics as predictors?
- Improved cytometry measurements?

  $\Rightarrow$ 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 parimary 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

0270-9139/89/1001-0001$02.00/0
HEPATOLOGY
Copyright © 1989 by the American Association for the Study of Liver Diseases
Vol. 10, No. 1, pp. 1–7, 19
Printed in U.S.

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, noninvasive 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 concentrations, 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:

$$\lambda(t \mid X_{\texttt{bili}}, X_{\texttt{protime}}, X_{\texttt{edema}}, X_{\texttt{alb}}, X_{\texttt{age}})$$

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?

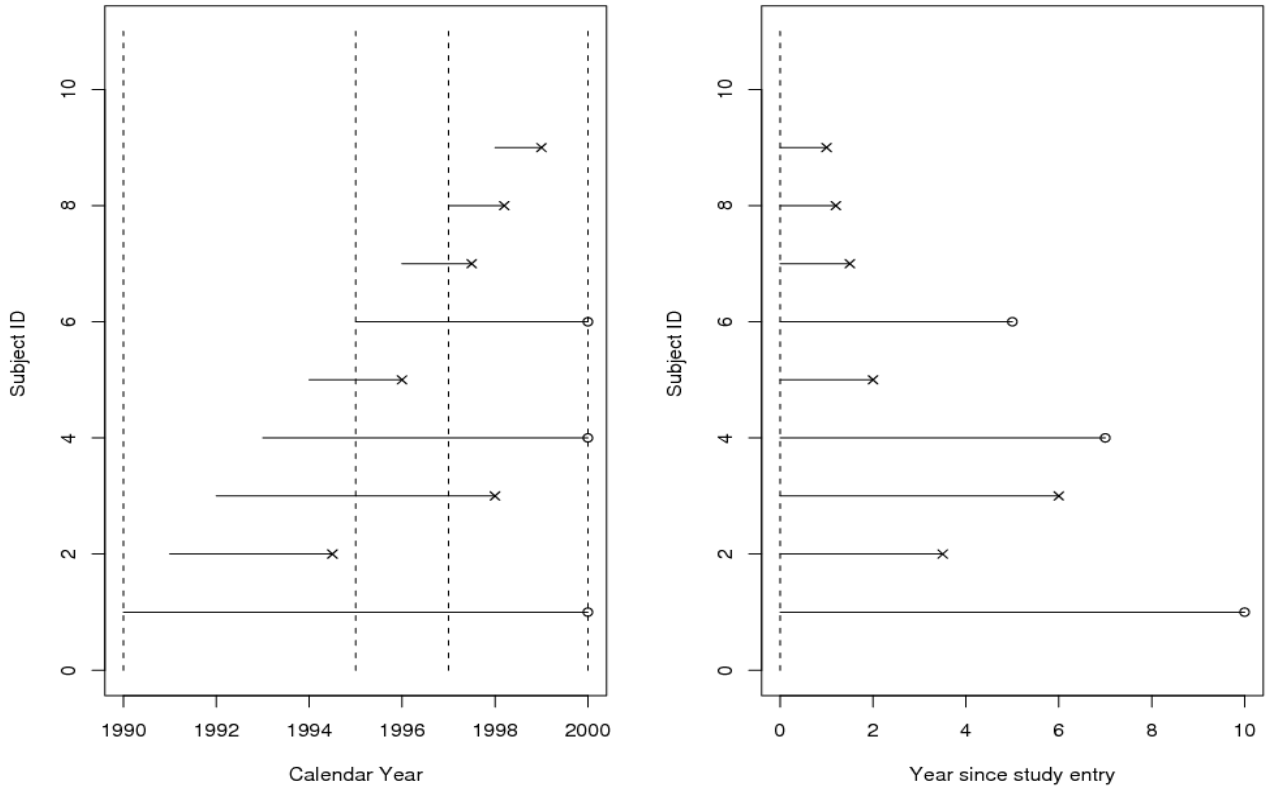
How to measure time?

- 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
  - Examples
    - Lost-to-followup
    - Competing risks
    - Administrative censoring

A typical right-censored dataset



Survival datasets

- How to organize our data?
  - $(X_i, \Delta_i, Z_i), i = 1, 2, .., n$
  - $X_i = \min(T_i, C_i), \Delta = I(T_i \leq C_i)$
- 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

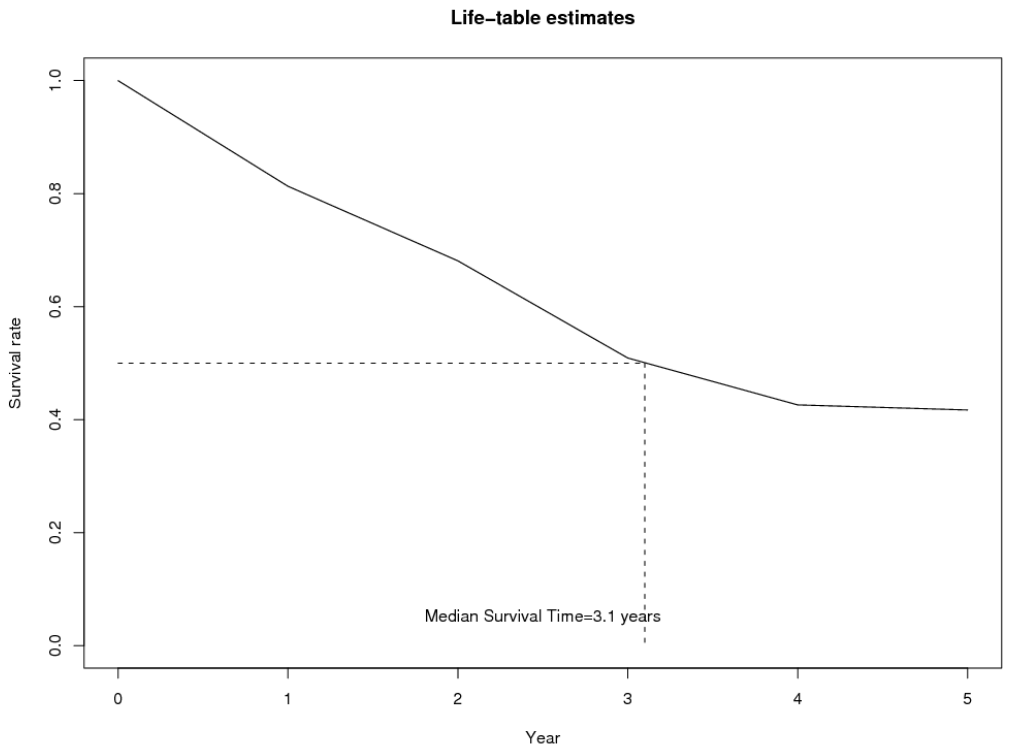
Life-tables

- 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. $\hat{S}(t) = \prod p \Longrightarrow \log \hat{S}(t) = \sum \log p$ ⟶ 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. $\text{var}\{\log \hat{S}(t)\} \approx \sum_{j \le t} d_j/\{(n_j - w_j/2)(n_j - d_j - w_j/2)\}$
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

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

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

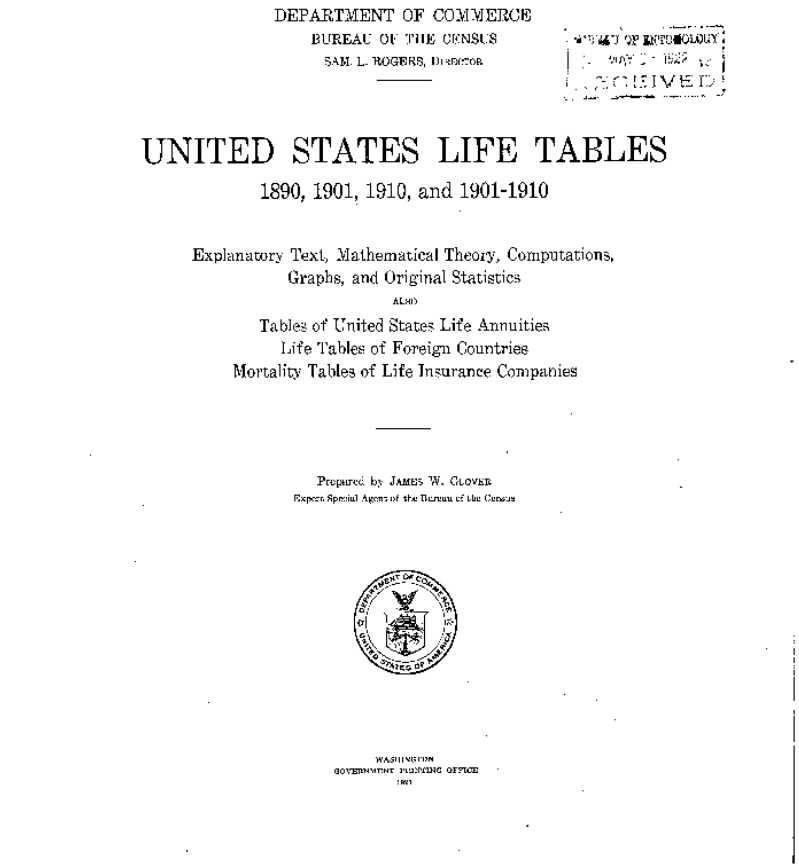
t	n	d	w	$\hat{S}$	$\mathrm{var}(\log p)$	se
0-1	146	27	3	0.813	0.00159	0.0324
1-2	116	18	10	0.681	0.00174	0.0392
2-3	88	21	10	0.509	0.00408	0.0438
3-4	57	9	3	0.426	0.00349	0.0444
4-5	45	1	3	0.417	0.00054	0.0445

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

SOCIETY OF ENTOMOLOGY
MAY 7 1929
RECEIVED

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. GLOVER
Expert Special Agent of the Bureau of the Census

WASHINGTON
GOVERNMENT PRINTING OFFICE
1921

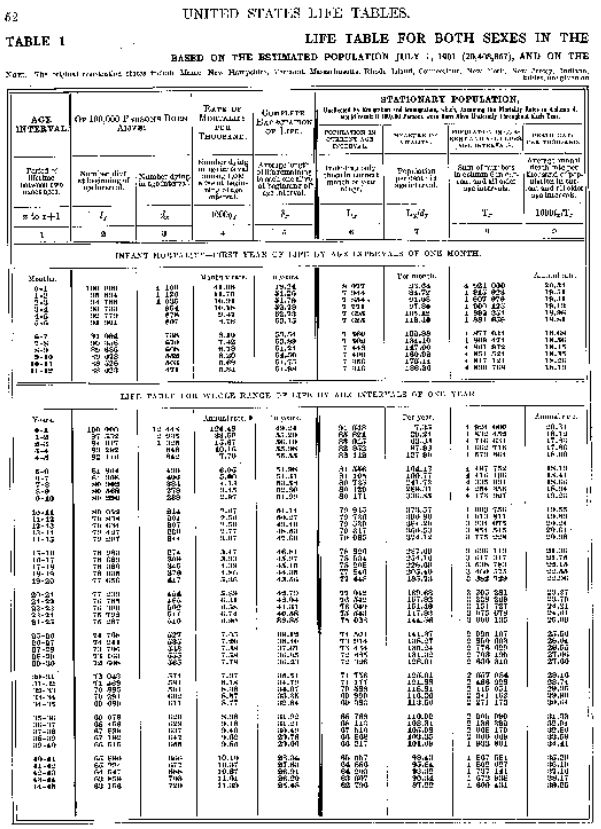


TABLE 1. LIFE TABLE FOR BOTH SEXES IN THE
BASED ON THE ESTIMATED POPULATION JULY 1, 1901 (29,478,867), AND ON THE

Note. The original registration states (which Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Indiana, Ohio, Michigan, and Wisconsin)

AGE INTERVAL.	Of 100,000 Persons Born Alive:		Rate of Mortality at Age Indicated.	Complete Expectation of Life.	Stationary Population. Undertaken in the system of computation, which, among the Mortality Rates in Column 4, supposes it (1) to Persons who have Already Completed Each Year.			
					Population in Current and Decimals.	Number of Deaths.	Population per 1000 of Mean Population.	Sum of Numbers in Column C to C inclusive, and at Close and Under-age.
Period of Lifetime between two exact Ages.	Number alive at beginning of Age Interval.	Number dying in Age Interval.	Number dying in Age Interval out of 1,000 alive at Beginning of Age Interval.	Average length of Life remaining to those who live in Age Interval.	Number living between two exact Ages.	Number dying between two exact Ages.	Population per 1000 of Mean Population.	Sum of Numbers in Column C to C inclusive, and at Close and Under-age.
x to $x+1$	l_x	d_x	$1000q_x$	$\mathring{e}_x$	L_x	L_x/l_0	T_x	$1000e_x/T_x$
1	2	3	4	5	6	7	8	9

INFANT MORTALITY—FIRST YEAR OF LIFE BY AGE INTERVALS OF ONE MONTH.

Months.								
0-1	100,000	1,491	14.91	47.24	8,997	47.54	4,721,000	47.24
1-2	98,509	1,129	11.46	47.65	9,344	89.72	4,712,003	47.83
2-3	97,380	1,052	10.81	47.79	9,244	93.96	4,677,978	48.04
3-4	96,328	1,004	10.42	47.86	9,231	97.89	4,643,734	48.21
4-5	95,324	974	10.22	47.85	9,205	101.82	4,608,503	48.34
5-6	94,350	947	10.04	47.73	9,165	110.10	4,567,298	48.41
6-7	93,403	938	10.04	47.54	9,060	122.58	4,477,611	48.48
7-8	92,465	929	10.05	47.29	8,969	134.15	4,389,471	48.36
8-9	91,536	919	10.04	47.01	8,883	147.90	4,301,302	48.15
9-10	90,617	908	10.02	46.70	8,801	160.93	4,213,521	48.01
10-11	89,709	898	10.01	46.37	8,722	176.14	4,127,511	47.91
11-12	88,811	888	10.00	46.04	8,646	190.93	4,038,709	47.78

LIFE TABLE FOR WHOLE RANGE OF LIFE BY AGE INTERVALS OF ONE YEAR

Years.								
0-1	100,000	12,444	124.44	49.24	91,658	7.35	4,924,600	49.24
1-2	87,556	2,978	34.02	56.04	85,954	30.31	4,832,942	55.19
2-3	84,577	1,328	15.71	56.19	83,923	53.39	4,746,988	56.12
3-4	83,249	948	11.39	55.74	82,753	67.33	4,663,065	56.02
4-5	82,301	842	10.23	55.23	81,788	127.80	4,580,312	55.65
5-6	81,459	698	8.57	54.66	81,110	141.78	4,498,524	55.22
6-7	80,761	636	7.87	54.07	80,443	154.27	4,417,414	54.70
7-8	80,125	599	7.48	53.46	79,825	165.98	4,336,971	54.13
8-9	79,526	570	7.17	52.83	79,241	177.39	4,257,146	53.53
9-10	78,956	548	6.94	52.19	78,682	188.57	4,177,905	52.92
10-11	78,408	533	6.80	51.54	78,142	199.61	4,099,223	52.29
11-12	77,875	524	6.73	50.89	77,613	210.61	4,021,081	51.64
12-13	77,351	524	6.77	50.24	77,089	221.70	3,943,468	50.98
13-14	76,827	534	6.95	49.59	76,560	232.98	3,866,379	50.32
14-15	76,293	554	7.26	48.94	76,016	244.57	3,789,819	49.67
15-16	75,739	584	7.71	48.30	75,447	256.58	3,713,803	49.03
16-17	75,155	624	8.30	47.68	74,843	269.11	3,638,356	48.41
17-18	74,531	674	9.04	47.07	74,194	282.23	3,563,513	47.81
18-19	73,857	734	9.94	46.48	73,490	296.00	3,489,319	47.24
19-20	73,123	803	10.98	45.91	72,722	310.46	3,415,829	46.71
20-21	72,320	881	12.18	45.36	71,880	325.63	3,343,107	46.23
21-22	71,439	966	13.52	44.83	70,956	341.52	3,271,227	45.79
22-23	70,473	1,057	15.00	44.32	69,944	358.12	3,200,271	45.41
23-24	69,416	1,152	16.60	43.83	68,840	375.41	3,130,327	45.09
24-25	68,264	1,249	18.30	43.36	67,639	393.36	3,061,487	44.84
25-26	67,015	1,346	20.08	42.91	66,342	411.93	2,993,848	44.67
26-27	65,669	1,441	21.94	42.48	64,949	431.07	2,927,506	44.58
27-28	64,228	1,532	23.85	42.07	63,462	450.74	2,862,557	44.57
28-29	62,696	1,617	25.79	41.68	61,887	470.89	2,799,095	44.64
29-30	61,079	1,695	27.75	41.30	60,231	491.46	2,737,208	44.81
30-31	59,384	1,764	29.70	40.94	58,502	512.41	2,676,977	45.08
31-32	57,620	1,823	31.64	40.59	56,708	533.71	2,618,475	45.44
32-33	55,797	1,872	33.55	40.25	54,861	555.34	2,561,767	45.91
33-34	53,925	1,910	35.42	39.92	52,970	577.28	2,506,906	46.48
34-35	52,015	1,939	37.28	39.60	51,045	599.54	2,453,936	47.18
35-36	50,076	1,958	39.10	39.28	49,097	622.14	2,402,891	48.00
36-37	48,118	1,969	40.91	38.97	47,133	645.11	2,353,794	48.92
37-38	46,149	1,971	42.71	38.66	45,164	668.47	2,306,661	49.97
38-39	44,178	1,966	44.51	38.35	43,195	692.24	2,261,497	51.19
39-40	42,212	1,954	46.29	38.04	41,235	716.44	2,218,302	52.55
40-41	40,258	1,936	48.09	37.73	39,290	741.09	2,177,067	54.07
41-42	38,322	1,913	49.92	37.41	37,365	766.21	2,137,777	55.78
42-43	36,409	1,885	51.77	37.09	35,466	791.83	2,100,412	57.69
43-44	34,524	1,853	53.67	36.76	33,597	817.96	2,064,946	59.82
44-45	32,671	1,817	55.62	36.43	31,762	844.63	2,031,349	62.19
45-46	30,854	1,778	57.62	36.09	29,965	871.87	1,999,587	64.80
46-47	29,076	1,736	59.70	35.74	28,208	899.72	1,969,622	67.73
47-48	27,340	1,692	61.88	35.38	26,494	928.22	1,941,414	70.99
48-49	25,648	1,647	64.22	35.01	24,825	957.42	1,914,920	74.67
49-50	24,001	1,602	66.74	34.63	23,200	987.37	1,890,095	78.77
50-51	22,399	1,557	69.51	34.24	21,620	1,018.13	1,866,895	83.32
51-52	20,842	1,513	72.59	33.84	20,086	1,049.76	1,845,275	88.36
52-53	19,329	1,470	76.04	33.43	18,594	1,082.33	1,825,189	94.42
53-54	17,859	1,429	79.98	33.01	17,145	1,115.99	1,806,595	101.14
54-55	16,430	1,390	84.59	32.58	15,735	1,150.79	1,789,450	108.93
55-56	15,040	1,353	89.97	32.14	14,363	1,186.79	1,773,715	117.93
56-57	13,687	1,319	96.37	31.69	13,027	1,224.06	1,759,352	128.54
57-58	12,368	1,288	104.14	31.23	11,724	1,262.68	1,746,325	141.19
58-59	11,080	1,260	113.72	30.76	10,450	1,302.72	1,734,601	156.54
59-60	9,820	1,236	125.87	30.28	9,202	1,344.26	1,724,151	175.59
60-61	8,584	1,215	141.57	29.79	7,977	1,387.33	1,714,949	199.75
61-62	7,369	1,197	162.45	29.29	6,770	1,431.99	1,706,972	230.97
62-63	6,172	1,182	191.52	28.78	5,581	1,478.32	1,700,202	271.91
63-64	4,990	1,170	234.47	28.26	4,405	1,526.41	1,694,621	325.98
64-65	3,820	1,161	303.93	27.73	3,239	1,576.36	1,690,216	397.87
65-66	2,659	1,154	433.99	27.19	2,082	1,628.26	1,686,977	494.98
66-67	1,505	1,150	764.12	26.64	930	1,682.21	1,684,895	630.00
67-68	355	355	1000.00	26.08	755	1,738.31	1,683,965	645.36

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 40.08 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

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Center for Health Statistics

Centers for Disease Control and Prevention
National Vital Statistics System

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 $x+n$	$_nq_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x
Days						
0–1	0.00275	100,000	275	274	7,683,447	76.83
1–7	0.00094	99,725	93	1,639	7,683,173	77.04
7–28	0.00095	99,631	94	5,729	7,681,534	77.10
28–365	0.00233	99,537	232	91,794	7,675,805	77.12
Years						
0–1	0.00696	100,000	695	99,436	7,683,447	76.83
1–2	0.00050	99,305	50	99,280	7,584,011	76.37
2–3	0.00033	99,255	33	99,239	7,484,731	75.41
3–4	0.00026	99,222	25	99,210	7,385,492	74.43
4–5	0.00021	99,197	21	99,186	7,286,282	73.45
5–6	0.00019	99,176	18	99,167	7,187,096	72.47
6–7	0.00017	99,158	17	99,149	7,087,929	71.48
7–8	0.00016	99,140	16	99,132	6,988,780	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	67.52
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	35	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	5,998,236	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	5,406,303	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	96,993	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	96,509	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	95,896	3,555,785	37.04
43–44	0.00257	95,782	247	95,658	3,459,889	36.12
44–45	0.00278	95,535	266	95,401	3,364,231	35.21
45–46	0.00301	95,268	288	95,124	3,268,830	34.31
46–47	0.00326	94,981	311	94,825	3,173,705	33.41
47–48	0.00353	94,670	335	94,502	3,078,880	32.52
48–49	0.00380	94,335	360	94,155	2,984,378	31.64
49–50	0.00407	93,975	384	93,783	2,890,223	30.76
50–51	0.00437	93,591	410	93,385	2,796,441	29.88
51–52	0.00469	93,180	439	92,961	2,703,055	29.01
52–53	0.00505	92,741	471	92,506	2,610,095	28.14
53–54	0.00548	92,270	508	92,016	2,517,589	27.28
54–55	0.00598	91,762	552	91,486	2,425,573	26.43
55–56	0.00657	91,211	603	90,909	2,334,086	25.59
56–57	0.00724	90,607	661	90,277	2,243,177	24.76
57–58	0.00797	89,947	722	89,586	2,152,900	23.94
58–59	0.00871	89,225	784	88,833	2,063,314	23.12

Modern survival analysis

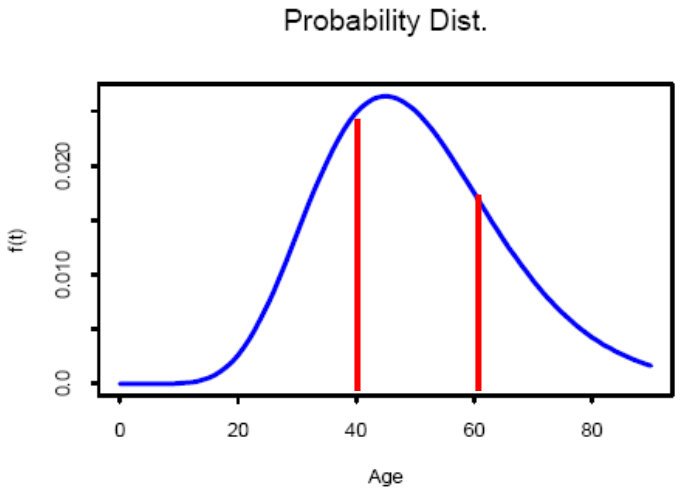
- Modern survival analysis
  - Post-1958
    - 1958: Kaplan-Meier estimator (Kaplan & Meier, *Journal of the American Statistical Association*)
    - 1966: Log-rank test (Mantel, *Cancer Chemo. Rep.*)
    - 1972: Cox proportional hazards model (Cox, *Journal of the Royal Statistical Society, Series B*)

Parametric Methods

- Functions to be studied?
  - Histogram: density function

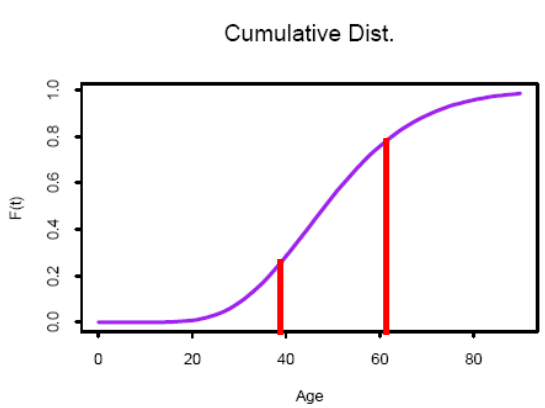
$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\{t \leq T \leq t + \Delta t\}$$

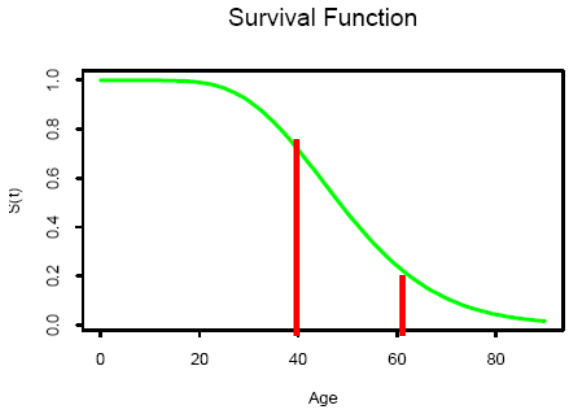
  - Area under the curve
    - Probability of an event



- Cumulative distribution (survival) function

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

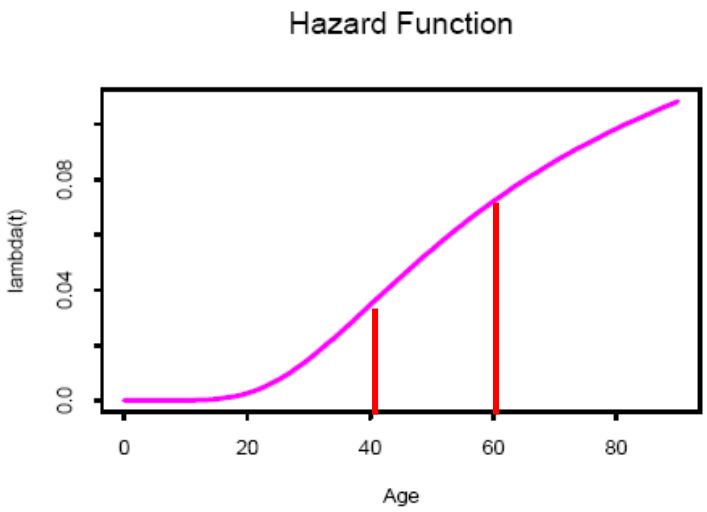




- Hazard function

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

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



- Hazard function
  - When times are discrete

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

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

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

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

- Hazard function
  - When times are continuous

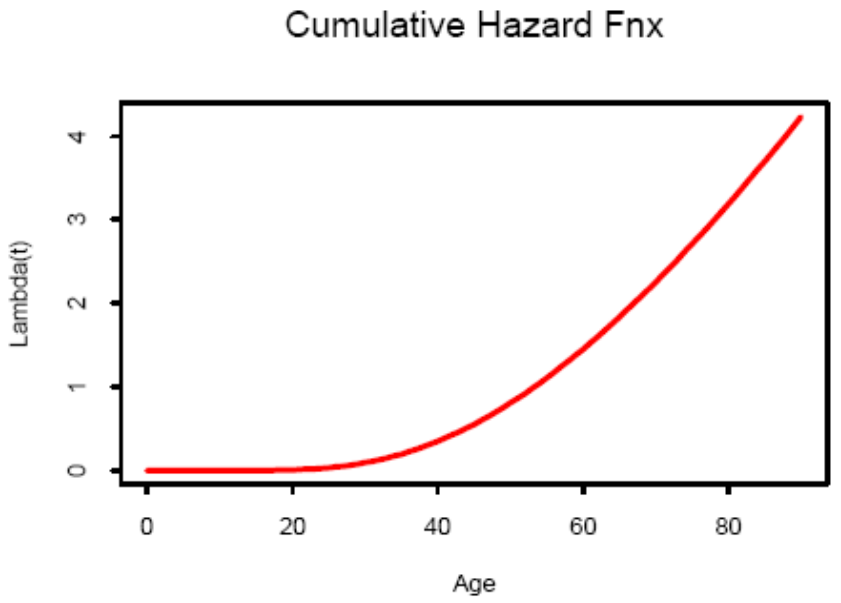
$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\Pr(\text{Failure occurring in } [t, t+\Delta t) | T \geq t)}{\Delta t}$$

$$= \text{Instantaneous failure rate at } t \text{ given survival up to } t$$

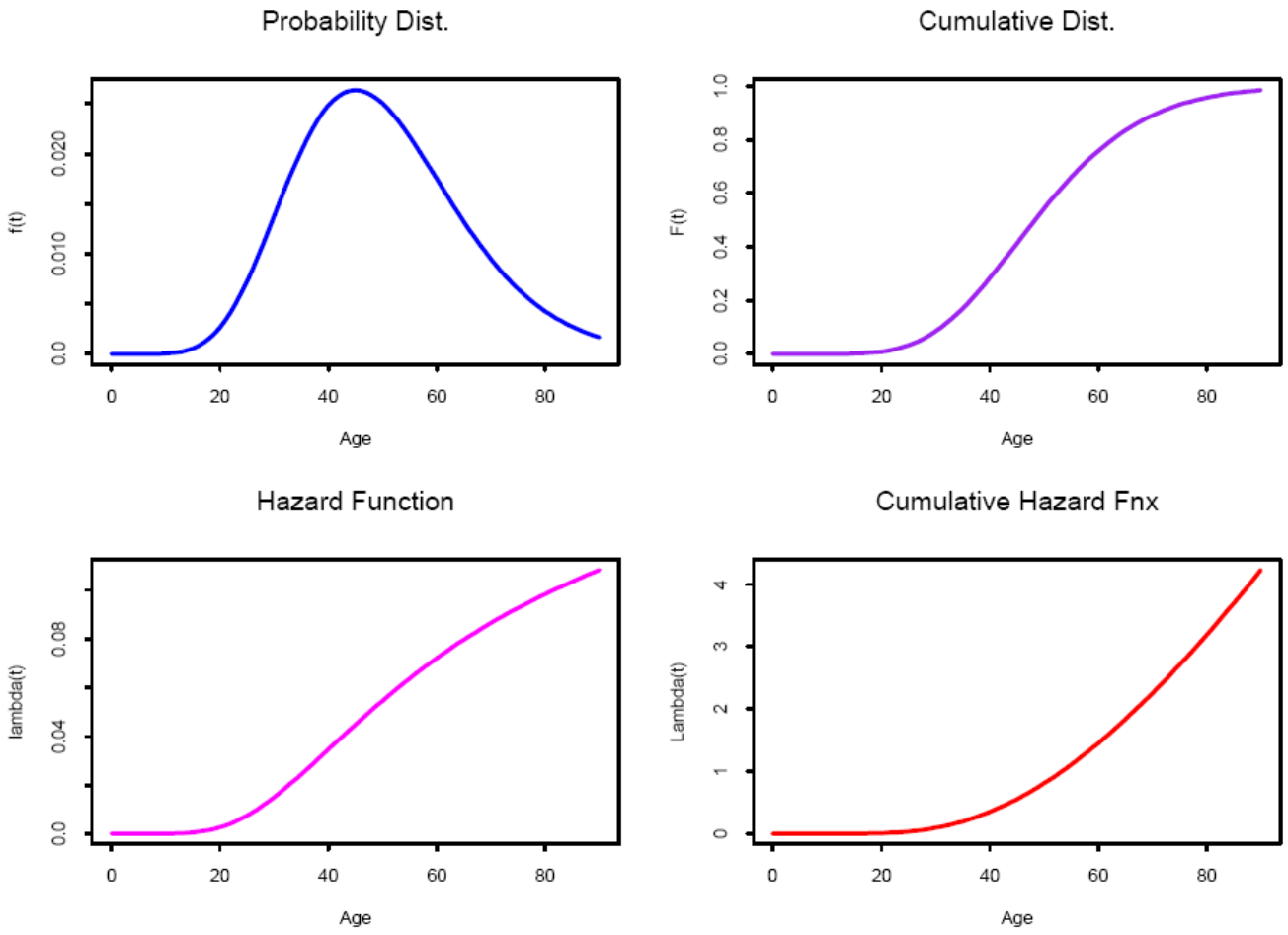
$$\lambda(t)\Delta t \approx \text{the proportion of individuals experiencing failure in } [t, t+\Delta t) \text{ to those surviving up to } t$$

- Cumulative hazard functions

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



Summary of functions



Relationship of functions

$$\lambda(t) = \frac{f(t)}{S(t)} \dashrightarrow \lambda(t) = \frac{\mathrm{P}(T=t)}{\mathrm{P}(T \geq t)} = \frac{f(t)}{S(t^-)}$$

$$f(t) = F'(t) \qquad S(t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(u)du}$$

$$F(t) = 1 - S(t) \dashrightarrow \Lambda(t) = \int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{S(u)}du$$

$$= 1 - e^{-\Lambda(t)} \approx 1 - \{1 - \Lambda(t)\} = \Lambda(t)$$

Summary

- Density function: $f(t) = \lim_{\Delta t \to 0} \Pr\{t \le T \le T + \Delta t\}$
  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: $\lambda(t) = \lim_{\Delta t \to 0} \Pr\{t \le T \le T + \Delta t \mid T \ge t\}/\Delta t$
  1. $\lambda(t) = f(t)/S(t)$
  2. Denominator: subjects survival up to t
- Mean residual life function: $m(t) = E(T - t \mid T \ge t)$
  1. Residual time: $R_T(t) = (T - t)I(T \ge t) = (T - t)^+ = (T - t) \vee 0$
  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/λ and the variance is r/λ^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.

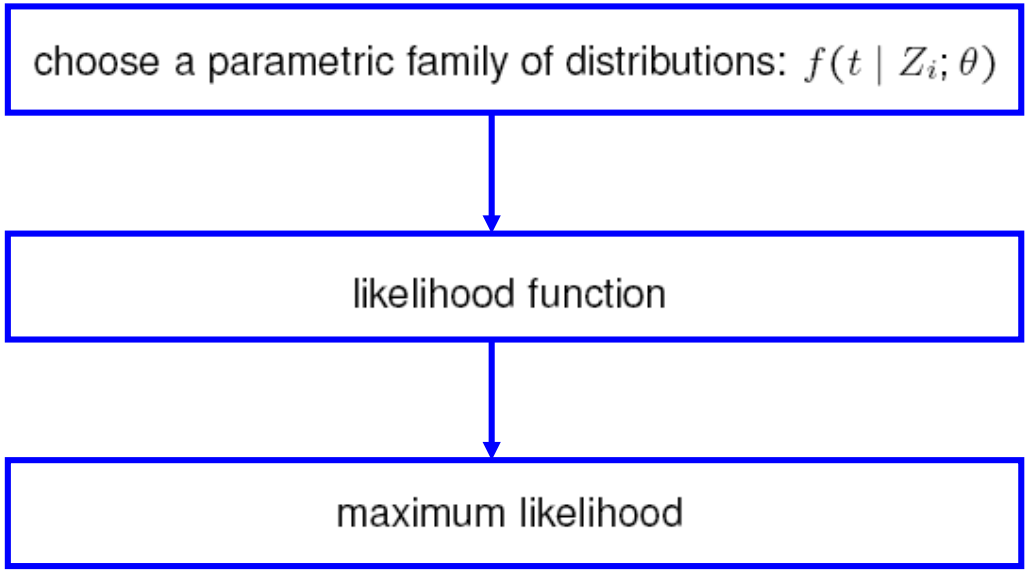
Parametric methods

- What's common in these distributions?
  - Distribution 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, $\{(T_i, Z_i); i = 1, 2, \ldots, n\}$

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

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

- With censoring

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

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

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

Key assumptions

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

What are the key assumptions?

- Z_i do not have information about θ
- 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:

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

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

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

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

It is equivalent to maximizing the log of likelihood function

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

Statistical theory of MLE

- Consistency:

$$\hat{\theta} \to \theta_0$$

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

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

or

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

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 θ 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