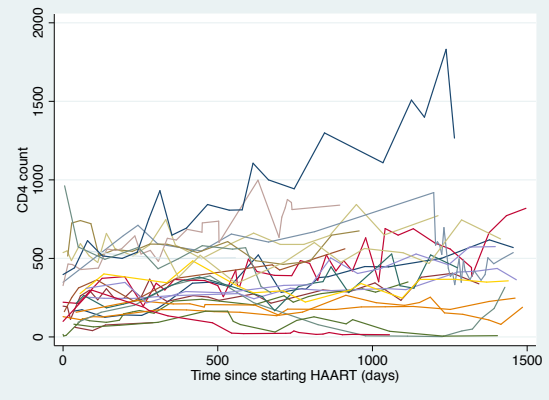
Longitudinal Data
Fall 2015



Survival Analysis

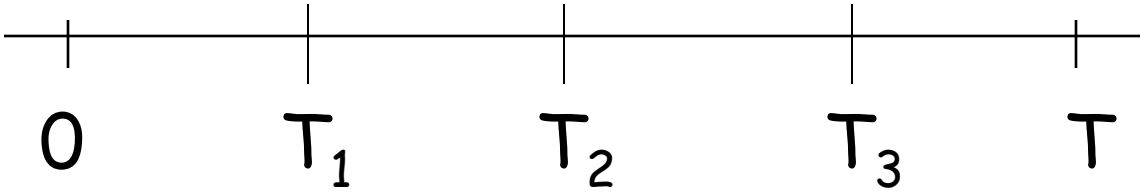
Instructors

Nick Jewell (jewell@berkeley.edu)

GSI

Robin Mejia (mejia@nasw.org)

Measures of Disease Incidence



Consider incidence rate in sub-intervals

Study rates over separate sub-intervals

Make sub-intervals shorter and shorter

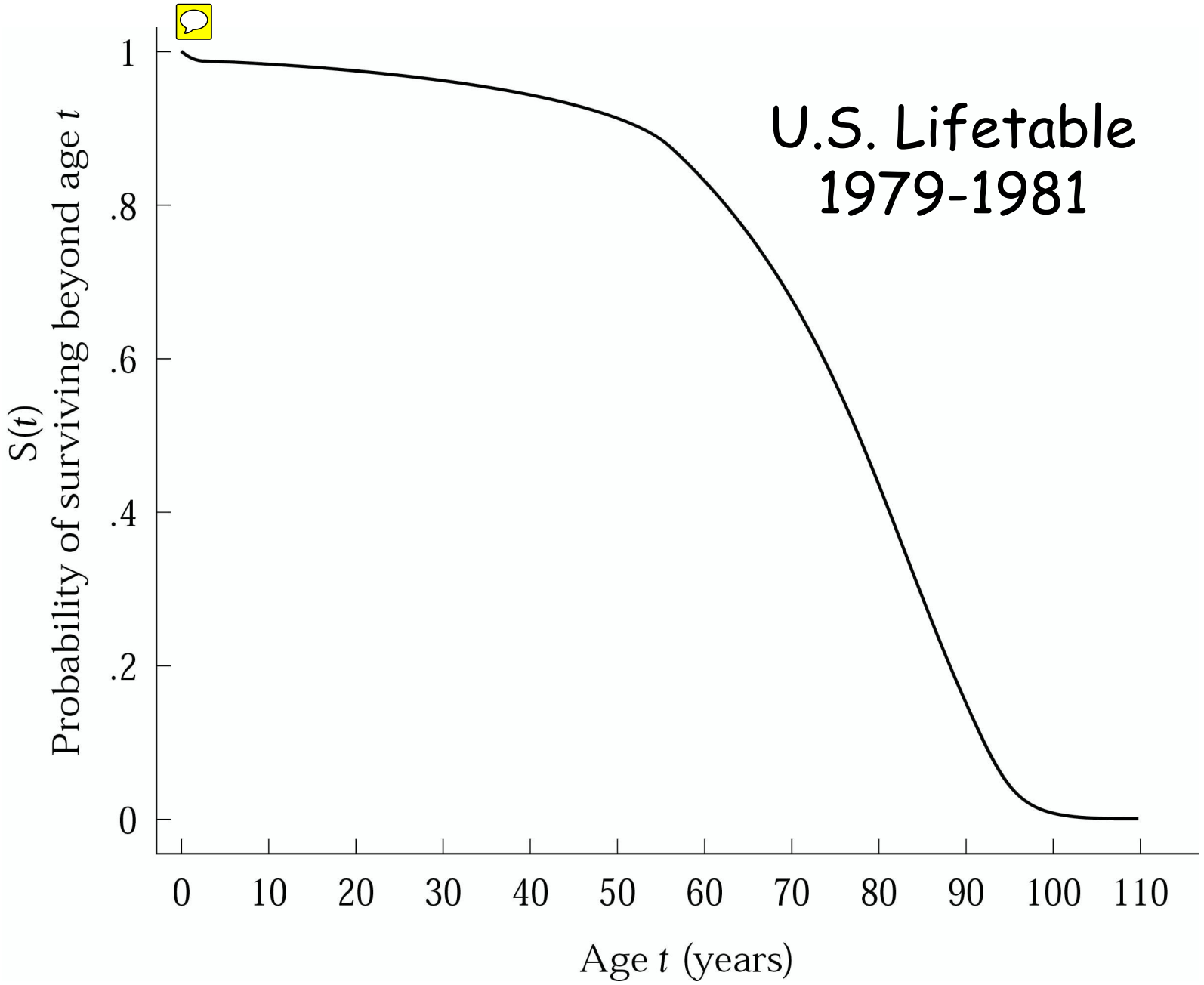
Hazard function *h(t)*

The Hazard Function

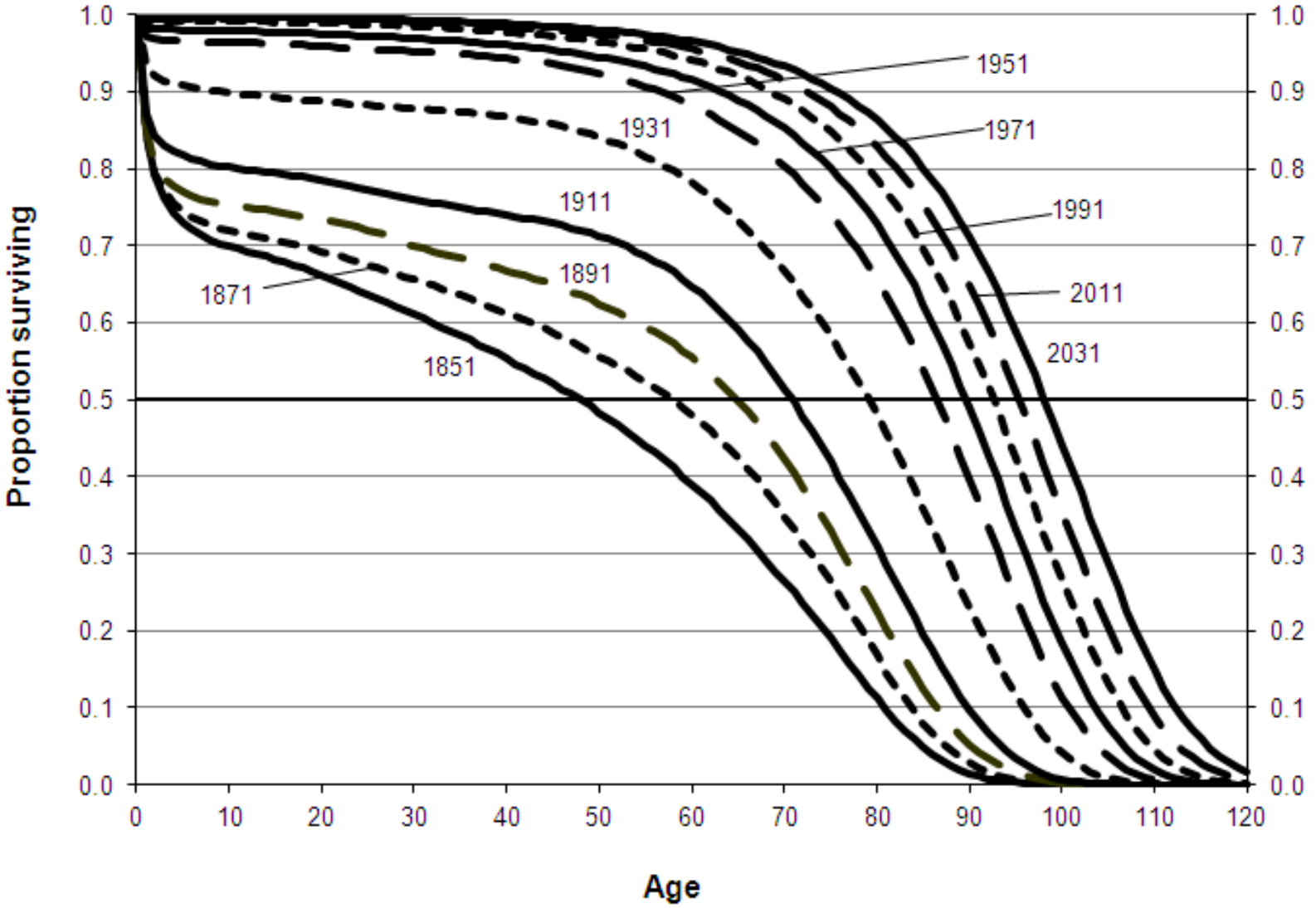
- Instantaneous incidence rate
- Related to Incidence Proportion *I(t)* and survival function *S(t) = 1-P(t)*

$$h(t) = \frac{\frac{dP(t)}{dt}}{(1-P(t))} = -\frac{\frac{dS(t)}{dt}}{S(t)}$$

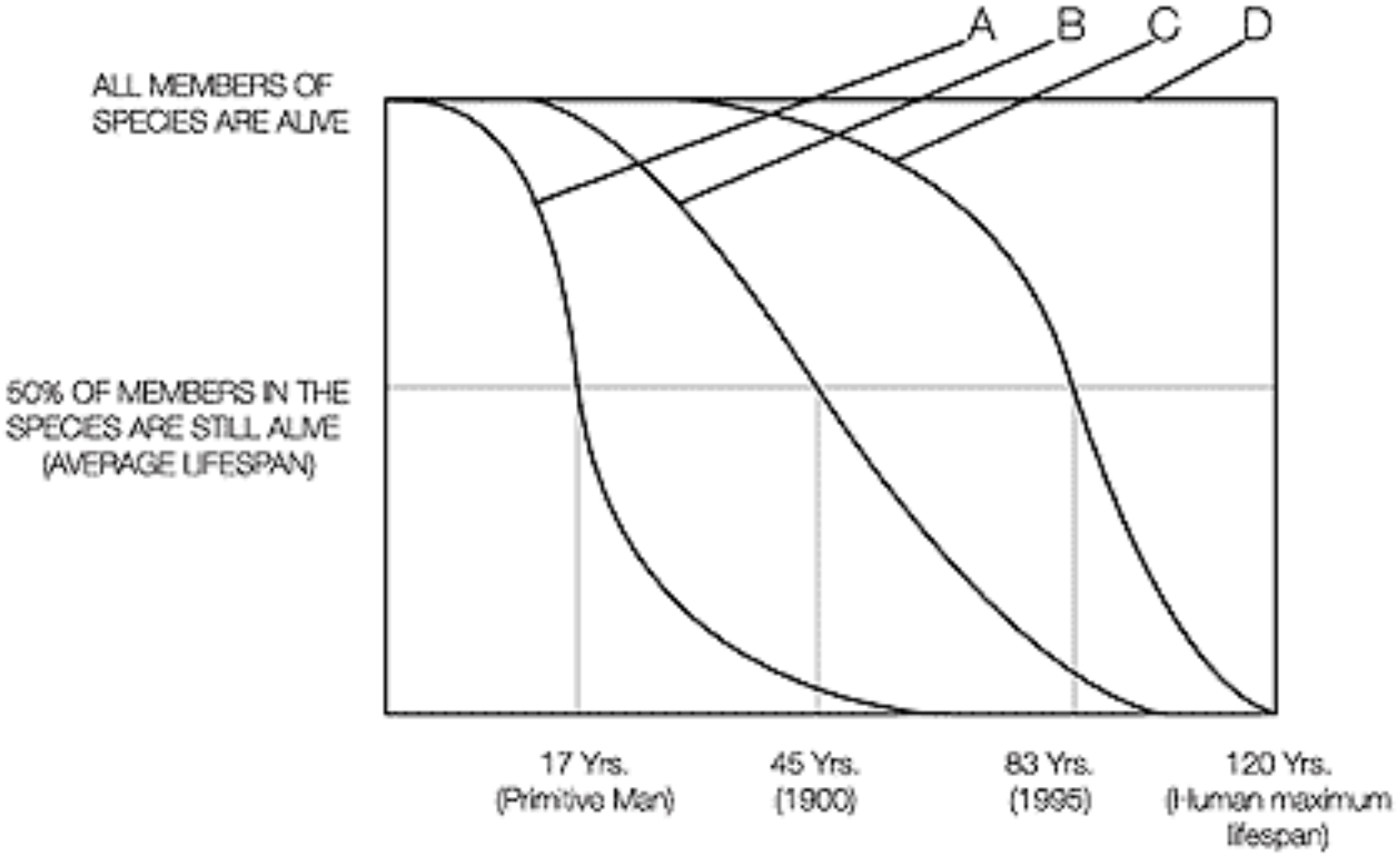
Note: *dP(t)/dt* is the probability density of incidence or death times

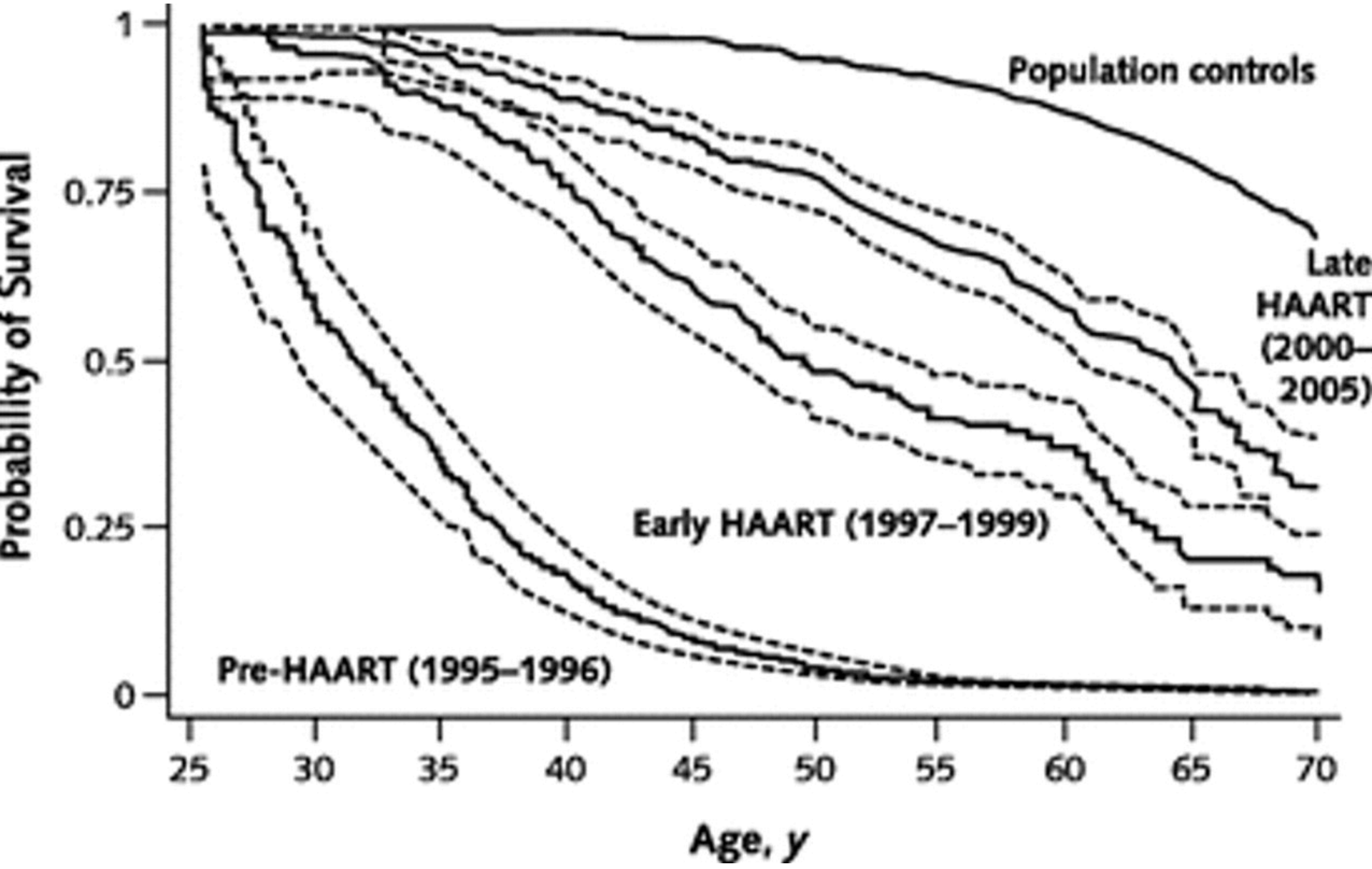

U.S. Lifetable
1979-1981
S(t)
Probability of surviving beyond age t
1
.8
.6
.4
.2
0
0
10
20
30
40
50
60
70
80
90
100
110
Age t (years)

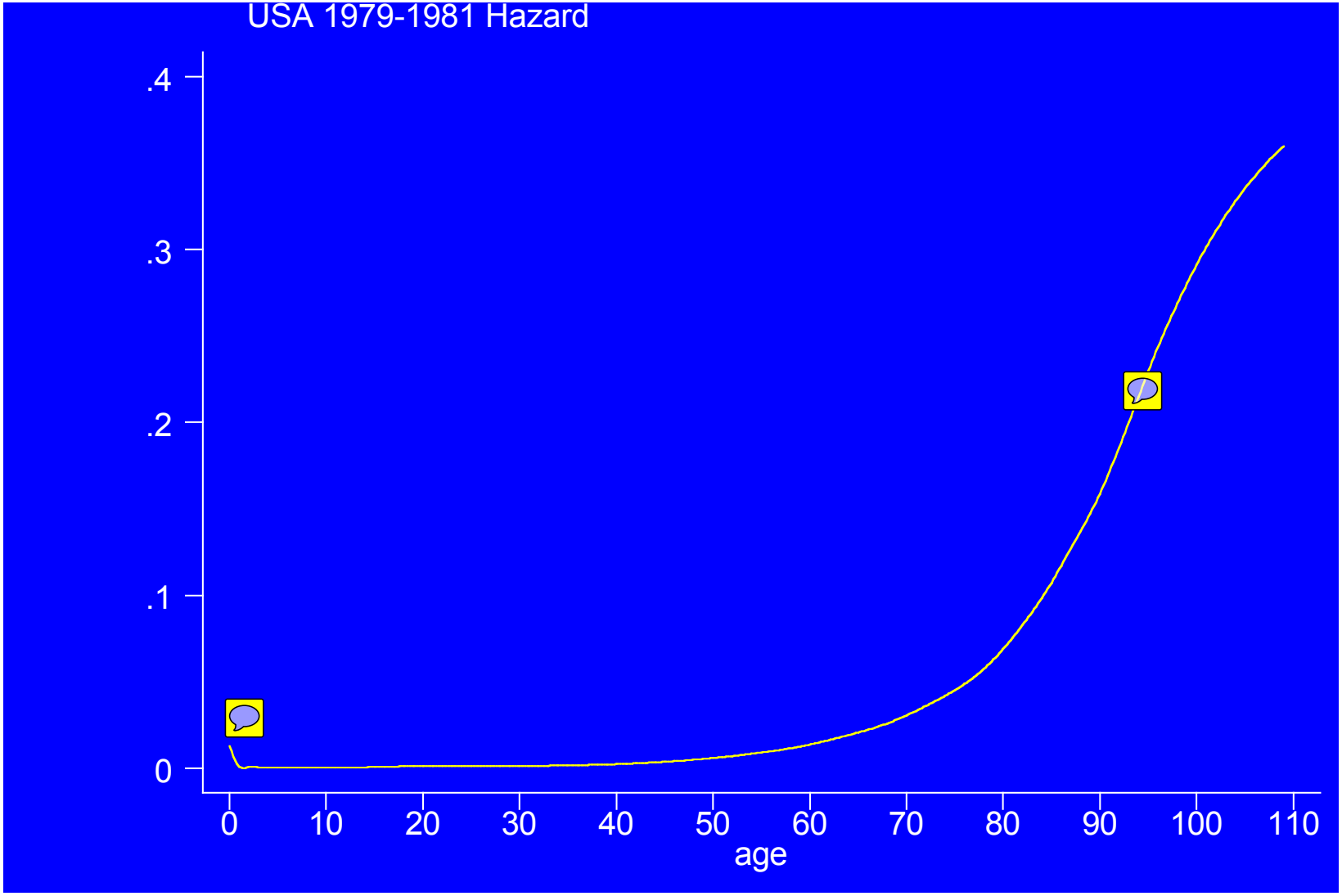
Proportion of persons surviving (on a cohort basis) to successive ages, according to mortality rates experienced or projected, persons born 1851-2031, England and Wales

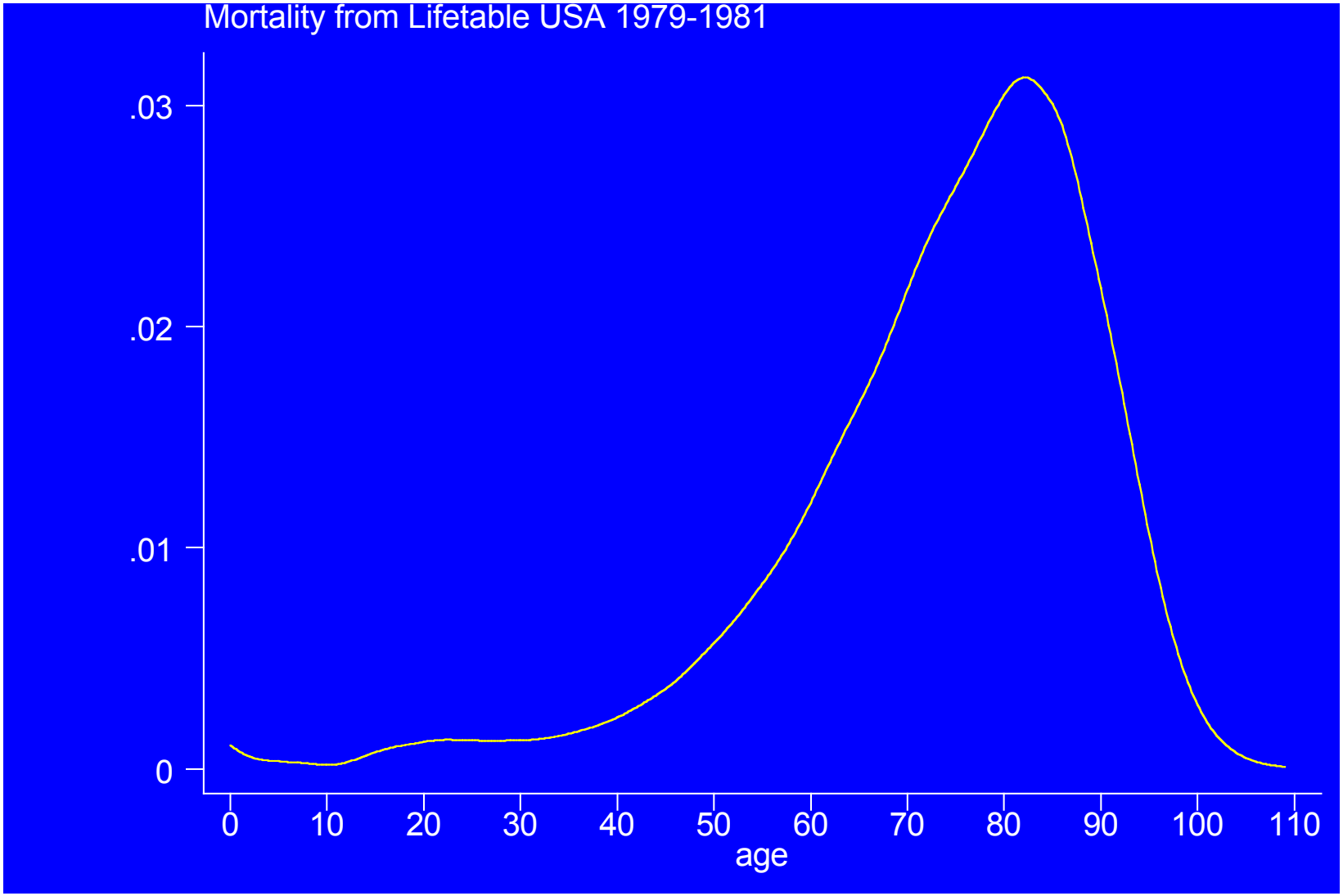


LIFESPAN/SURVIVAL CURVE




Probability of Survival
1
0.75
0.5
0.25
0
Population controls
Late HAART (2000–2005)
Early HAART (1997–1999)
Pre-HAART (1995–1996)
25
30
35
40
45
50
55
60
65
70
Age, y


USA 1979-1981 Hazard
.4
.3
.2
.1
0
0
10
20
30
40
50
60
70
80
90
100
110
age



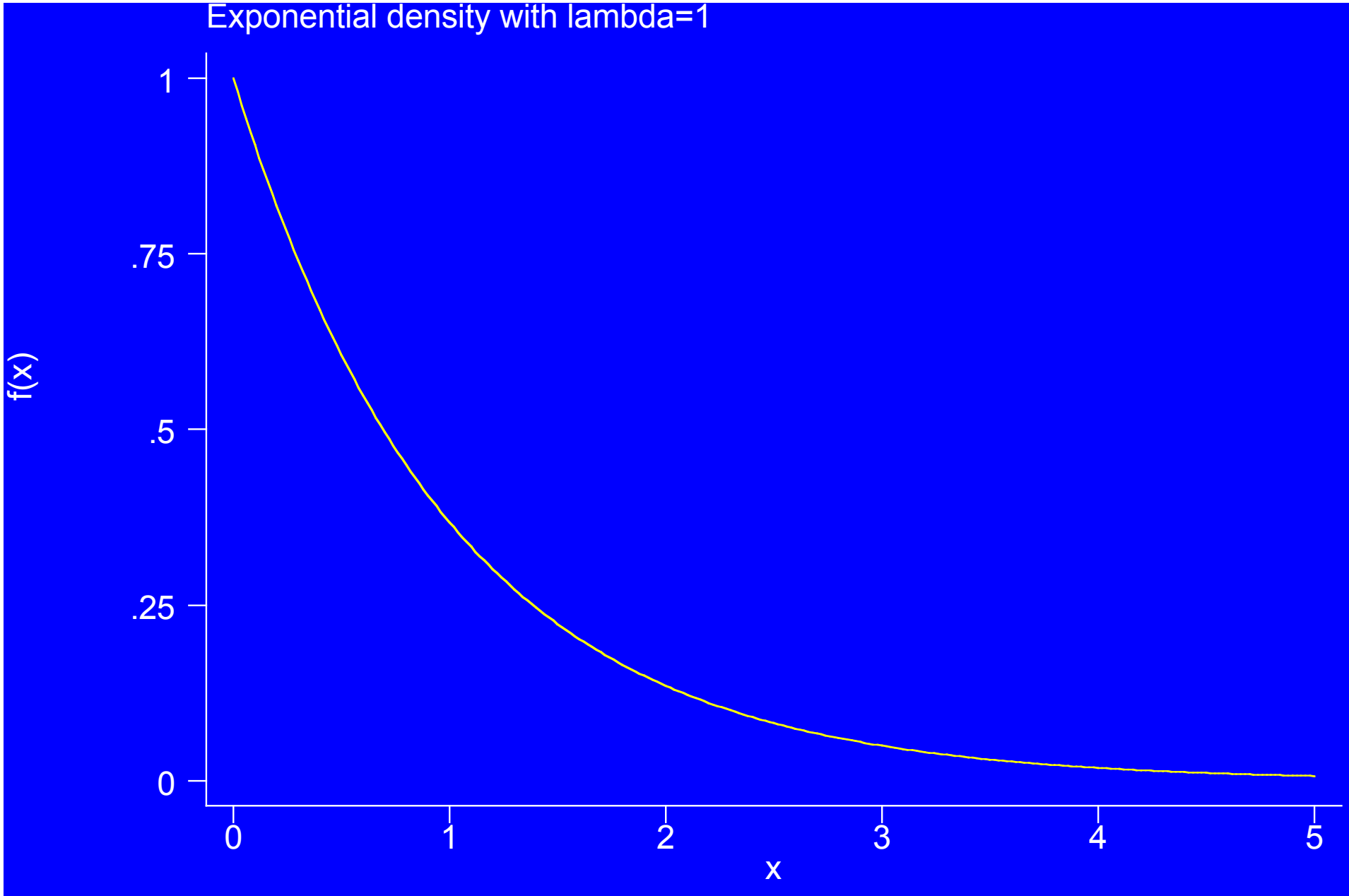
$$f(t) = \frac{dP(t)}{dt}$$

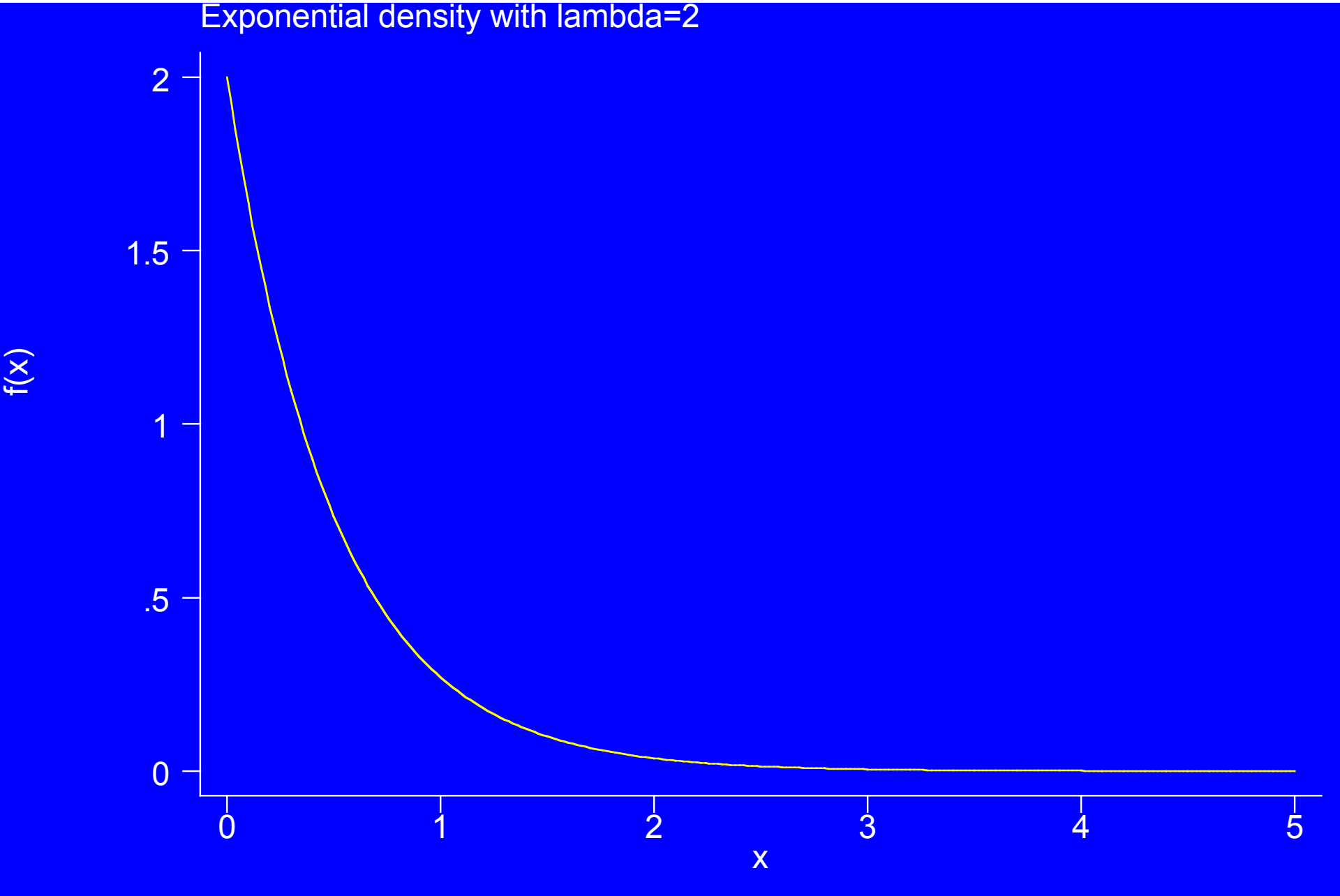
Exponential distribution

$$f(x) = \lambda e^{-\lambda x}$$

$$S(x) = e^{-\lambda x}$$

$$h(x) = \frac{f(x)}{S(x)} = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}} = \lambda$$


Exponential density with lambda=1
f(x)
1
.75
.5
.25
0
0
1
2
3
4
5
x

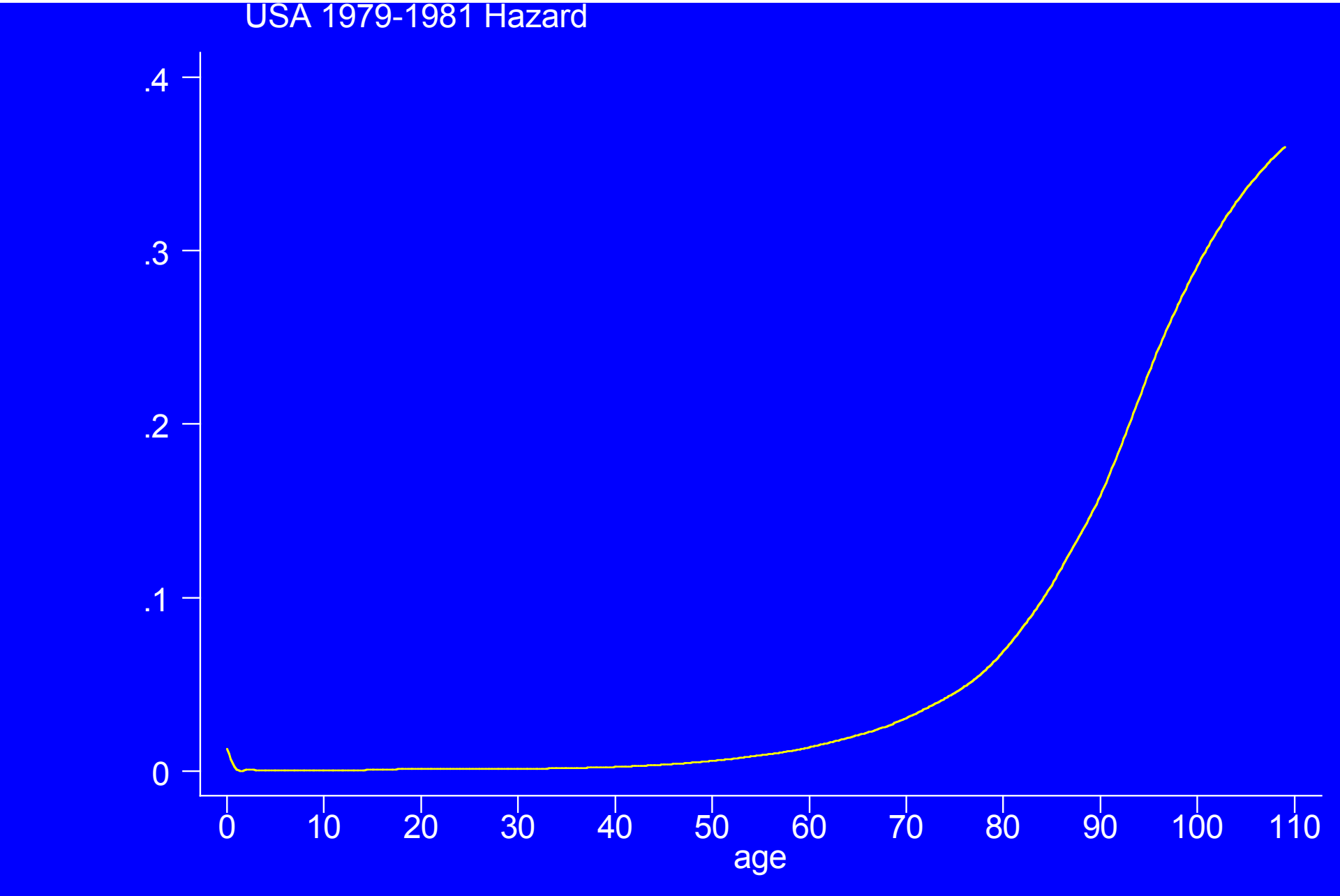

Exponential density with lambda=2
f(x)
2
1.5
1
.5
0
0
1
2
3
4
5
x

A constant hazard is used in electronics, and it is appropriate for some cancers such as lung and breast where the hazard, over and above 'natural' causes, is constant and may persist for 15-20 years.

A constant hazard is 'memoryless' in that the expected survival is independent of how long the individual has survived.

- A U shaped hazard is common. e.g. humans: very high mortality in 1st year, decreases till teen years, then increases.

- A decreasing hazard is appropriate for some cancers where hazard is high after diagnosis, then decreases to cure.


USA 1979-1981 Hazard
.4
.3
.2
.1
0
0
10
20
30
40
50
60
70
80
90
100
110
age

There are various parametric forms for hazard functions—such as the exponential, Weibull, Gompertz, Pareto, etc…, but we will focus on non-parametric, or semi-parametric inference.

Estimation of Survival (or Hazard) Function

- Suppose we have follow-up data on a sample of (independent) individuals that describes the *time* at which they became an incidence case (or died)
- How do we use the data to estimate *S(t)* or *h(t)*

Kaplan-Meier or Product-Limit Estimator

Simple Example

Interval from AIDS to death Hemoph. (Age < 41)

Patient	Survival (months)
1	2
2	3
3	6
4	6
5	7
6	10
7	15
8	15
9	16
10	27
11	30
12	32

Product Limit Method

Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1					
2	3	1					
3	6	2					
4	7	1					
5	10	1					
6	15	2					
7	16	1					
8	27	1					
9	30	1					
10	32	1					

Product Limit Method

Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12				
2	3	1	11				
3	6	2	10				
4	7	1	8				
5	10	1	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	1/12=.083			
2	3	1	11				
3	6	2	10				
4	7	1	8				
5	10	1	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917		
2	3	1	11				
3	6	2	10				
4	7	1	8				
5	10	1	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917		
2	3	1	11	1/11=.091	0.909		
3	6	2	10				
4	7	1	8				
5	10	1	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917		
2	3	1	11	0.091	0.909		
3	6	2	10	2/10=.2	0.8		
4	7	1	8				
5	10	1	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917		
2	3	1	11	0.091	0.909		
3	6	2	10	0.2	0.8		
4	7	1	8	1/8=.125	0.875		
5	10	1	7	1/7=.143	0.857		
6	15	2	6	2/6=.333	0.667		
7	16	1	4	1/4=.25	0.75		
8	27	1	3	1/3=.333	0.667		
9	30	1	2	1/2=.5	0.5		
10	32	1	1	1/1=1	1		

Product Limit Method

Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917	1x.917=.917	
2	3	1	11	0.091	0.909		
3	6	2	10	0.2	0.8		
4	7	1	8	0.125	0.875		
5	10	1	7	0.143	0.857		
6	15	2	6	0.333	0.667		
7	16	1	4	0.25	0.75		
8	27	1	3	0.333	0.667		
9	30	1	2	0.5	0.5		
10	32	1	1	1	1		

Product Limit Method

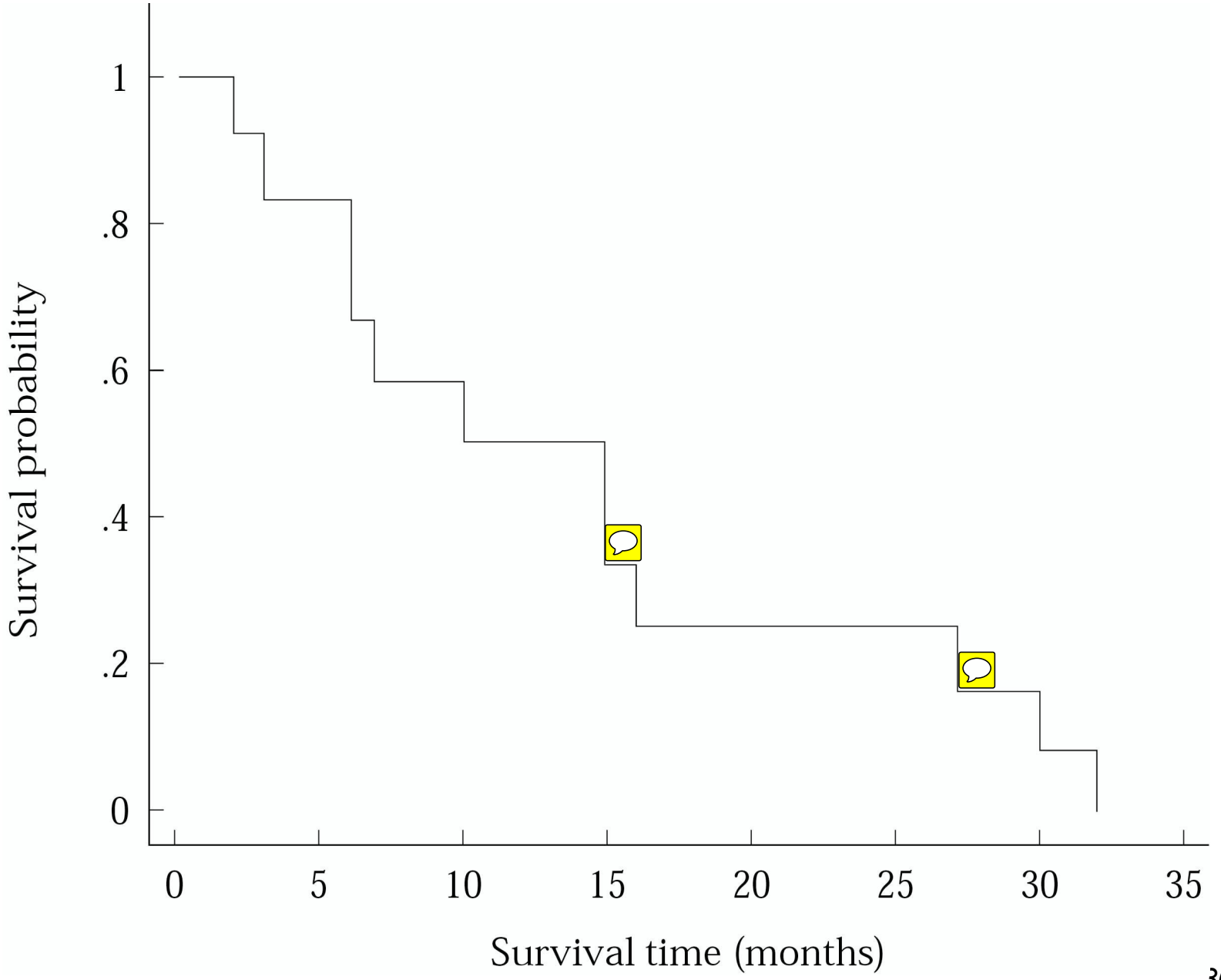
Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917	0.917	
2	3	1	11	0.091	0.909	.917x.909=.833	
3	6	2	10	0.2	0.8		
4	7	1	8	0.125	0.875		
5	10	1	7	0.143	0.857		
6	15	2	6	0.333	0.667		
7	16	1	4	0.25	0.75		
8	27	1	3	0.333	0.667		
9	30	1	2	0.5	0.5		
10	32	1	1	1	1		

Product Limit Method

Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917	0.917	
2	3	1	11	0.091	0.909	0.833	
3	6	2	10	0.2	0.8	.833x.8=.667	
4	7	1	8	0.125	0.875		
5	10	1	7	0.143	0.857		
6	15	2	6	0.333	0.667		
7	16	1	4	0.25	0.75		
8	27	1	3	0.333	0.667		
9	30	1	2	0.5	0.5		
10	32	1	1	1	1		

Product Limit Method

Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917	0.917	
2	3	1	11	0.091	0.909	0.833	
3	6	2	10	0.2	0.8	0.667	
4	7	1	8	0.125	0.875	.67x.875=.583	
5	10	1	7	0.143	0.857	.583x.857=.5	
6	15	2	6	0.333	0.667	.5x.667=.333	
7	16	1	4	0.25	0.75	.333x.75=.25	
8	27	1	3	0.333	0.667	.25x.667=.333	
9	30	1	2	0.5	0.5	.333x.5=.167	
10	32	1	1	1	0	.167x0=0	


Survival probability
1
.8
.6
.4
.2
0
0
5
10
15
20
25
30
35
Survival time (months)

Variation in Follow-up Periods -- Censoring

Suppose some of the patients are still not "dead'' at the time the analysis is done --

(right-) censored observations

In example, suppose individuals who "failed" at times 3 and 10 months actually dropped out at that point (lost to follow-up)

Simple Example

Interval from AIDS to death Hemoph. (Age < 41)

Patient	Survival (months)
1	2
2	3+
3	6
4	6
5	7
6	10+
7	15
8	15
9	16
10	27
11	30
12	32

What about left-censoring?

What is truncation and is it the same?

Think of examples of right truncation and left truncation

Can such effects impact "normal" longitudinal data analysis?

What about left-censoring?

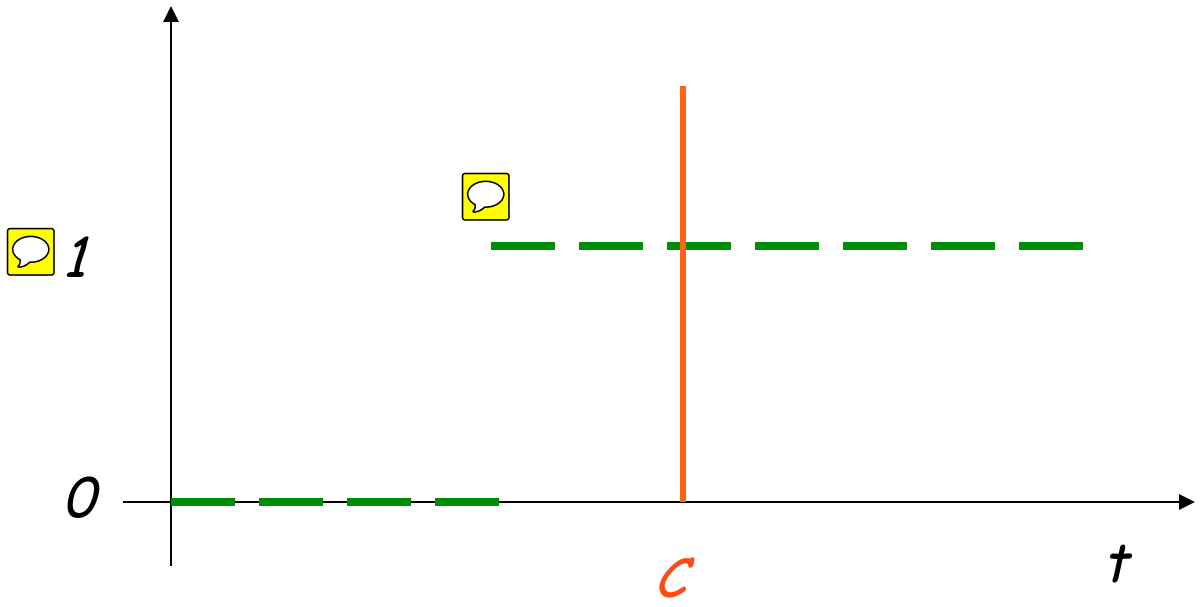
How about interval censoring?

What is truncation and is it the same?

Think of examples of right truncation and left truncation

Can such effects impact "normal" longitudinal data analysis?

Current Status Data



Instead of observing entire history (say up to T), we only observe "status" (in our case, dead or not, say) at a single point in time, C

Extreme form of Interval Censoring Inference about longitudinal effects from cross-sectional observations

Product Limit Method

Event Number	Time of Death or censoring (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1					
2	3+	0					
3	6	2					
4	7	1					
5	10+	0					
6	15	2					
7	16	1					
8	27	1					
9	30	1					
10	32	1					

Product Limit Method

Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12				
2	3+	0	11				
3	6	2	10				
4	7	1	9				
5	10	0	8				
6	15	2	8				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	1/12=.083			
2	3+	0	11				
3	6	2	10				
4	7	1	8				
5	10+	0	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917		
2	3+	0	11				
3	6	2	10				
4	7	1	8				
5	10+	0	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917		
2	3+	0	11	0/11=0	1		
3	6	2	10				
4	7	1	8				
5	10+	0	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917		
2	3+	0	11	0	1		
3	6	2	10	2/10=.2	0.8		
4	7	1	8				
5	10+	0	7				
6	15	2	6				
7	16	1	4				
8	27	1	3				
9	30	1	2				
10	32	1	1				

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917		
2	3+	0	11	0	1		
3	6	2	10	0.2	0.8		
4	7	1	8	1/8=.125	0.875		
5	10+	0	7	0/7=0	1		
6	15	2	6	2/6=.333	0.667		
7	16	1	4	1/4=.25	0.75		
8	27	1	3	1/3=.333	0.667		
9	30	1	2	1/2=.5	0.5		
10	32	1	1	1/1=1	1		

Product Limit Method

Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917	1x.917=.917	
2	3+	0	11	0	1		
3	6	2	10	0.2	0.8		
4	7	1	8	0.125	0.875		
5	10+	0	7	0	1		
6	15	2	6	0.333	0.667		
7	16	1	4	0.25	0.75		
8	27	1	3	0.333	0.667		
9	30	1	2	0.5	0.5		
10	32	1	1	1	1		

Product Limit Method

Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917	0.917	
2	3+	0	11	0	1	.917x1=.917	
3	6	2	10	0.2	0.8		
4	7	1	8	0.125	0.875		
5	10+	0	7	0	1		
6	15	2	6	0.333	0.667		
7	16	1	4	0.25	0.75		
8	27	1	3	0.333	0.667		
9	30	1	2	0.5	0.5		
10	32	1	1	1	1		

Product Limit Method

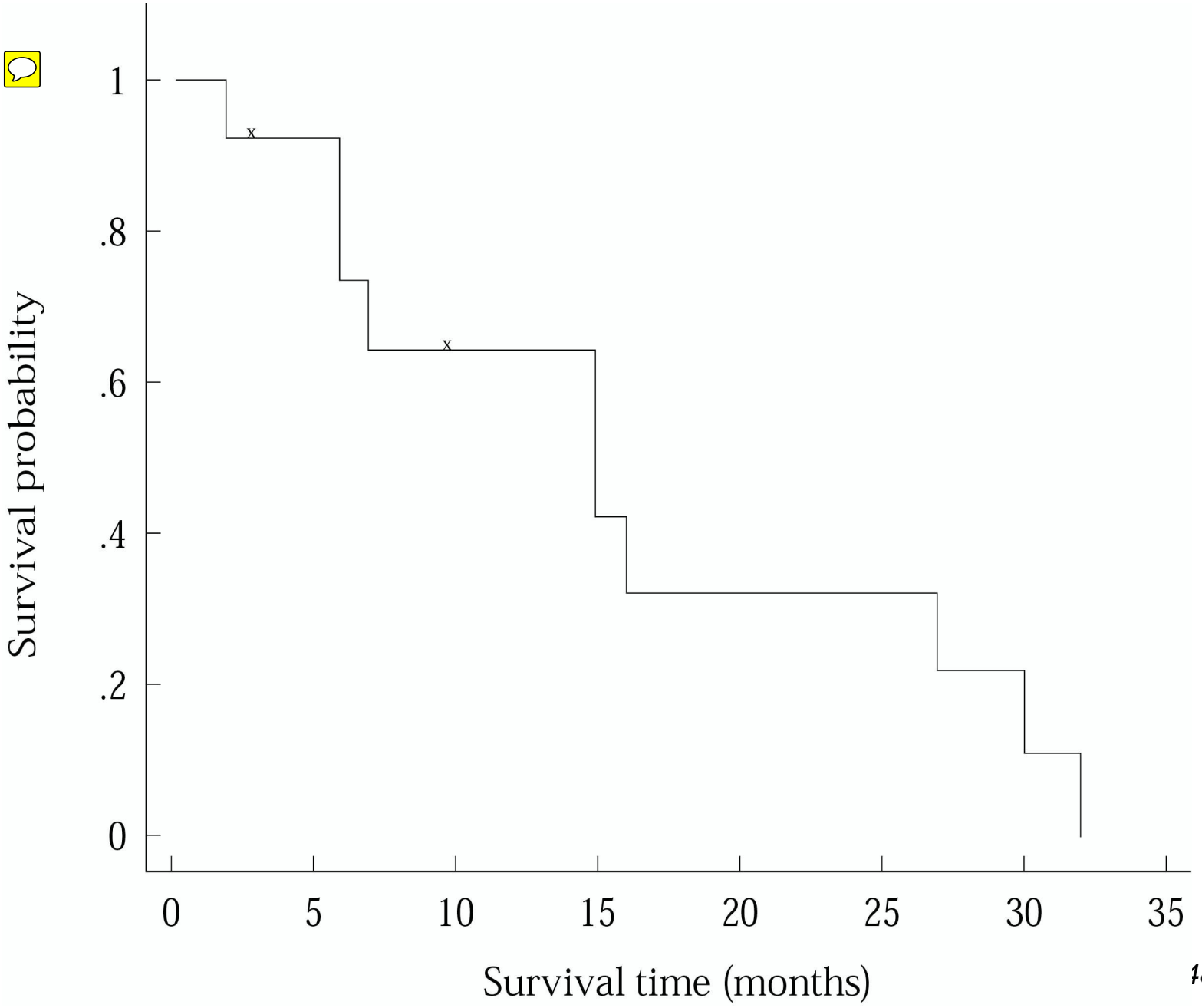
Event Number	Time of Death (t)	# who fail at time *t*	# at risk at time *t*	*q*	*p=1-q*	*S(t+)*	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917	0.917	
2	3+	0	11	0	1	0.917	
3	6	2	10	0.2	0.8	.917x.8=.733	
4	7	1	8	0.125	0.875		
5	10+	0	7	0	1		
6	15	2	6	0.333	0.667		
7	16	1	4	0.25	0.75		
8	27	1	3	0.333	0.667		
9	30	1	2	0.5	0.5		
10	32	1	1	1	1		

Product Limit Method

Event Number	Time of Death (t)	# who fail at time t	# at risk at time t	q	$p=1-q$	$S(t+)$	
0	0	0	12	0	1	1	
1	2	1	12	0.083	0.917	0.917	
2	3+	0	11	0	1	0.917	
3	6	2	10	0.2	0.8	0.733	
4	7	1	8	0.125	0.875	.733x.875=.642	
5	10+	0	7	0	1	.642x1=.642	
6	15	2	6	0.333	0.667	.642x.667=.428	
7	16	1	4	0.25	0.75	.428x.75=.321	
8	27	1	3	0.333	0.667	.321x.667=.214	
9	30	1	2	0.5	0.5	.214x.5=.107	
10	32	1	1	1	0	.107x0=0	

Note that this makes sense if we can treat the censored observations the same as the others.

i.e. we assume independent censoring


Survival probability
1
.8
.6
.4
.2
0
0
5
10
15
20
25
30
35
Survival time (months)

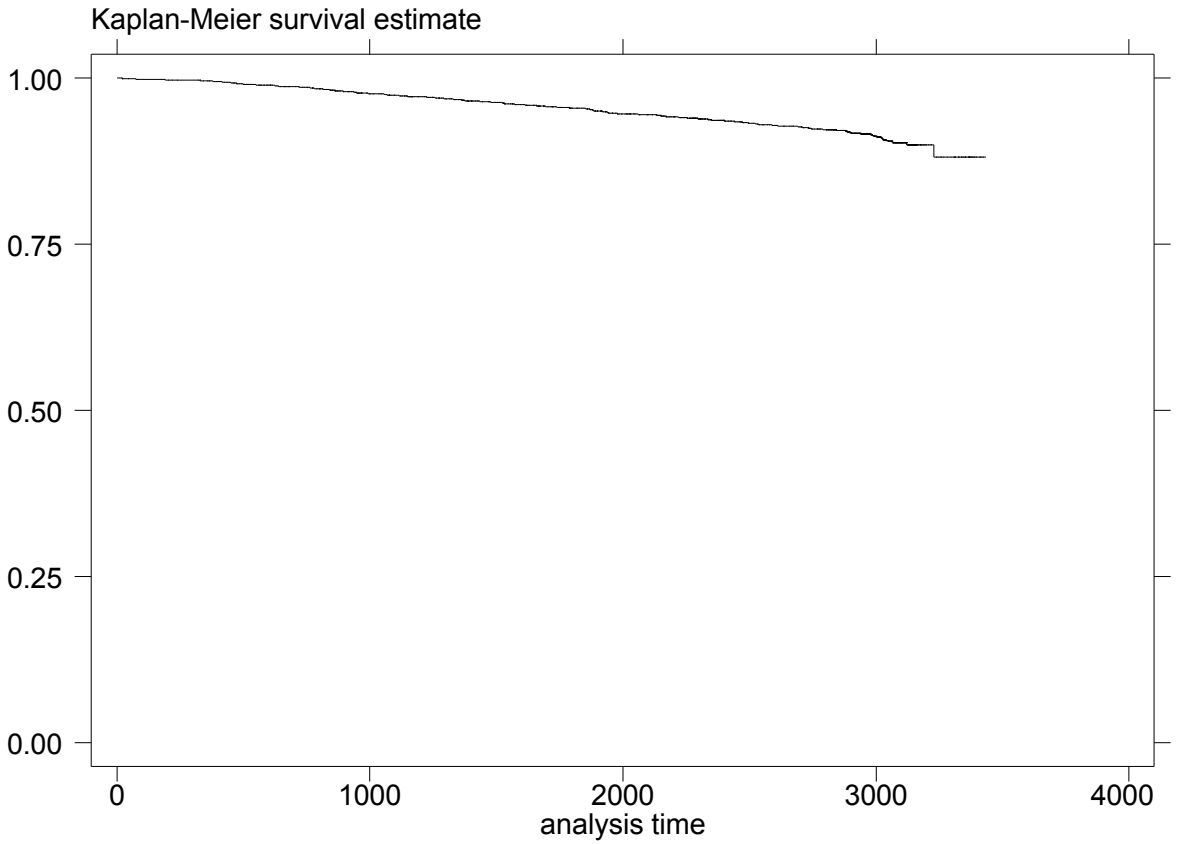
Western Collaborative Group Study

- Collected follow-up data on 3,154 employed men from 10 Californian companies (1960-61)
- Aged 39-59 years old at baseline
- Looked for onset of CHD for about 9 years
- Risk factors measured: smoking, blood pressure, cholesterol,weight, behavior type
- 257 CHD “events”

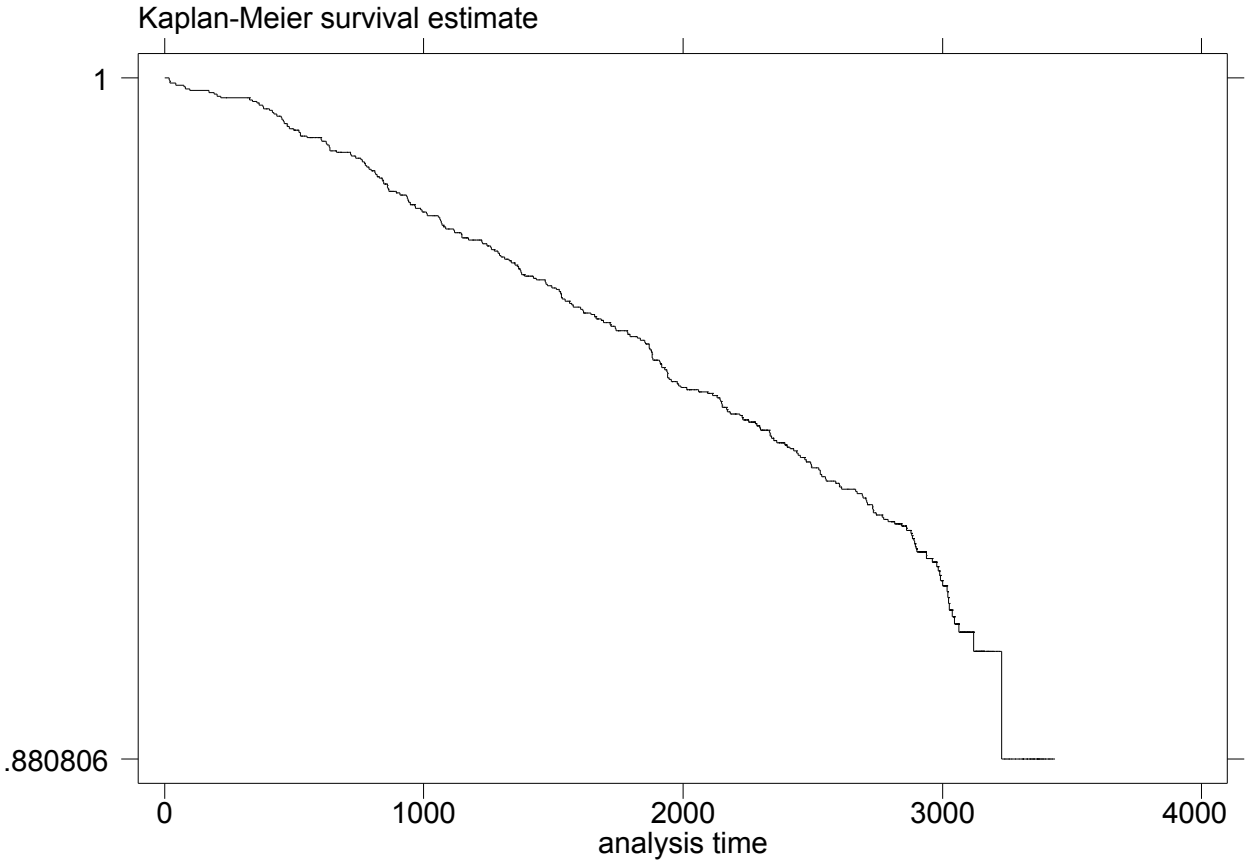
id	age0	height0	weight0	chol0	behpat0	ncigs0	chd69	time169
2001	49	73	150	225	2	25	0	1664
2002	42	70	160	177	2	20	0	3071
2003	42	69	160	181	3	0	0	3071
2004	41	68	152	132	4	20	0	3064
2005	59	70	150	255	3	20	1	1885
2006	44	72	204	182	4	0	0	3102
2007	44	72	164	155	4	0	0	3074
2008	40	71	150	140	2	0	0	3071
2009	43	72	190	149	3	25	0	3064
2010	42	70	175	325	2	0	0	1032
2011	53	69	167	223	2	25	0	3091
2013	41	67	156	271	2	20	0	3081
2014	50	72	173	238	1	50	1	1528
2017	43	72	180	189	3	30	0	3072

```
. stset  time169, failure( chd69)

     failure event:  chd69 ~= 0 & chd69 ~= .
obs. time interval:  (0, time169]
 exit on or before:  failure

------------------------------------------------------------------------------
     3154  total obs.
        0  exclusions
------------------------------------------------------------------------------
     3154  obs. remaining, representing
      257  failures in single record/single failure data
  8464892  total analysis time at risk, at risk from t =         0
                        earliest observed entry t =         0
                             last observed exit t =      3430
```

```
sts graph
        failure _d:  chd69
  analysis time _t:
time169
```

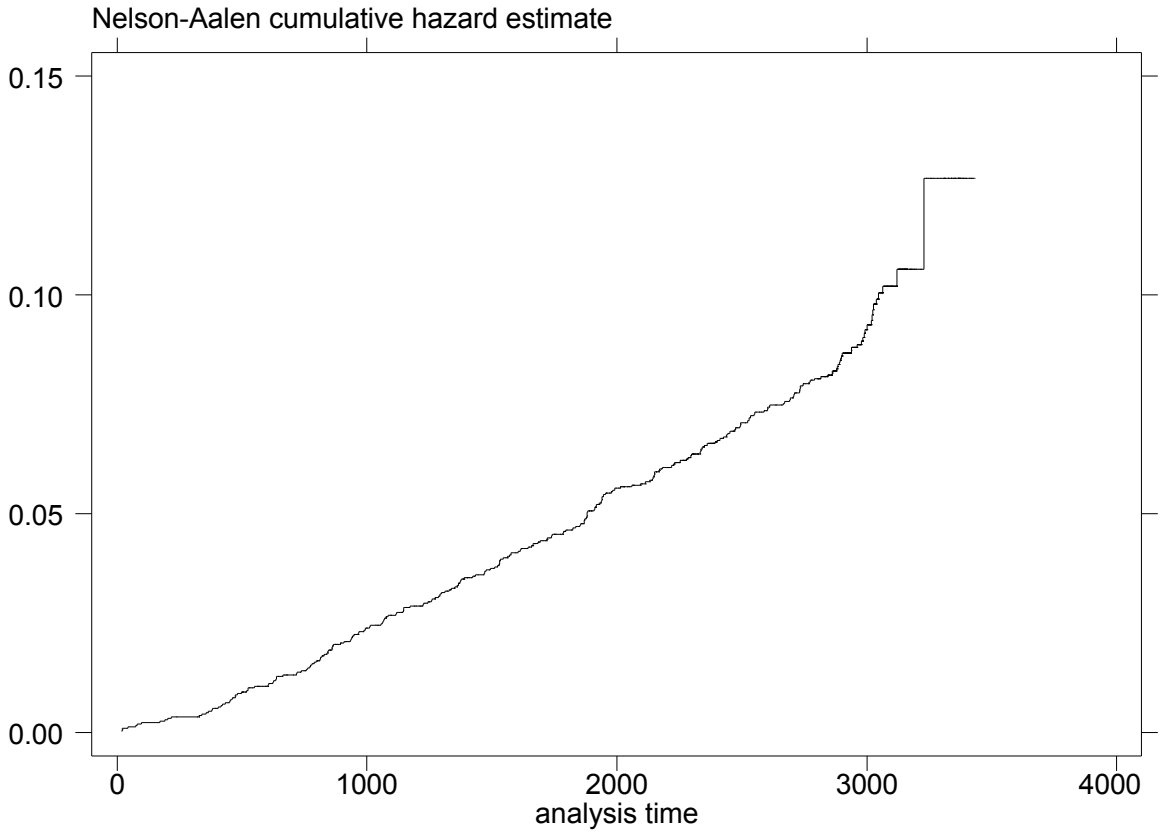


```
. sts graph, yas
        failure _d:  chd69
  analysis time _t:  time169
```



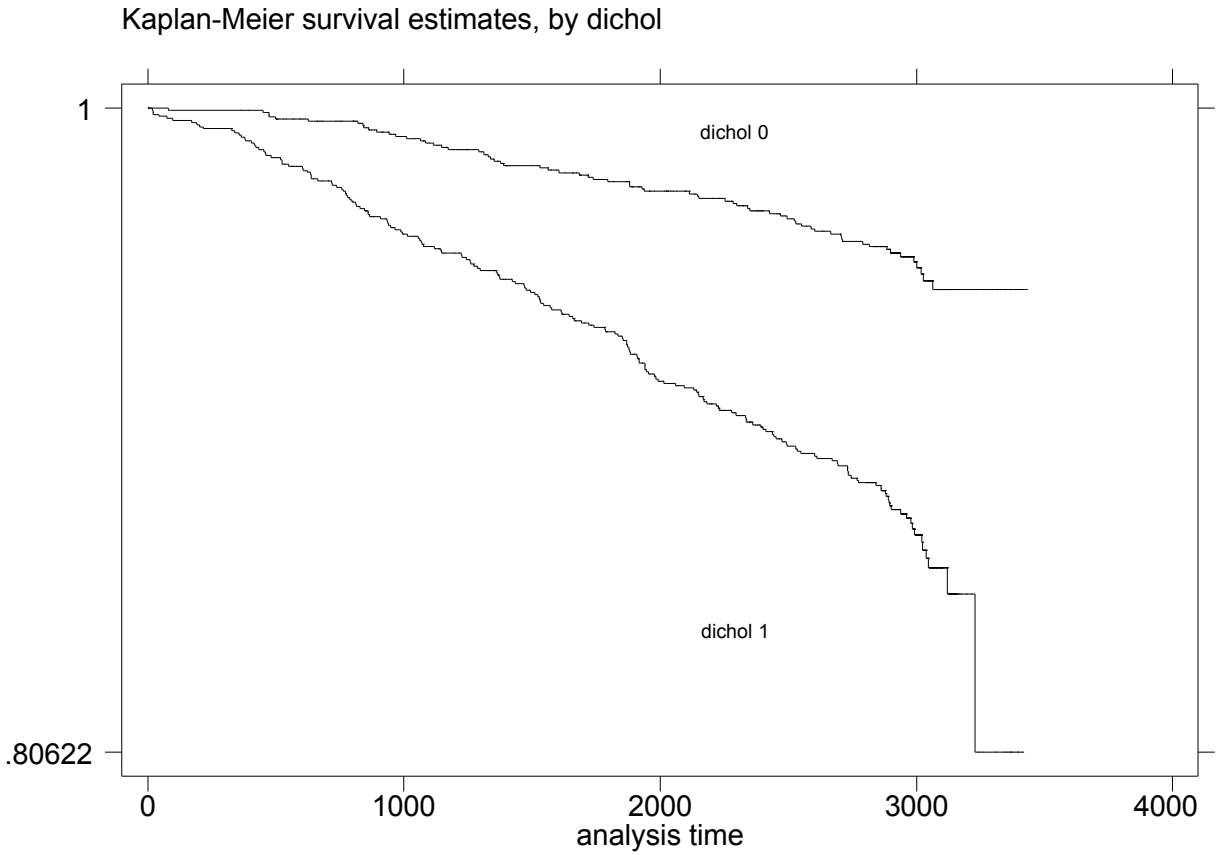
Cumulative Hazard Function

sts graph, na



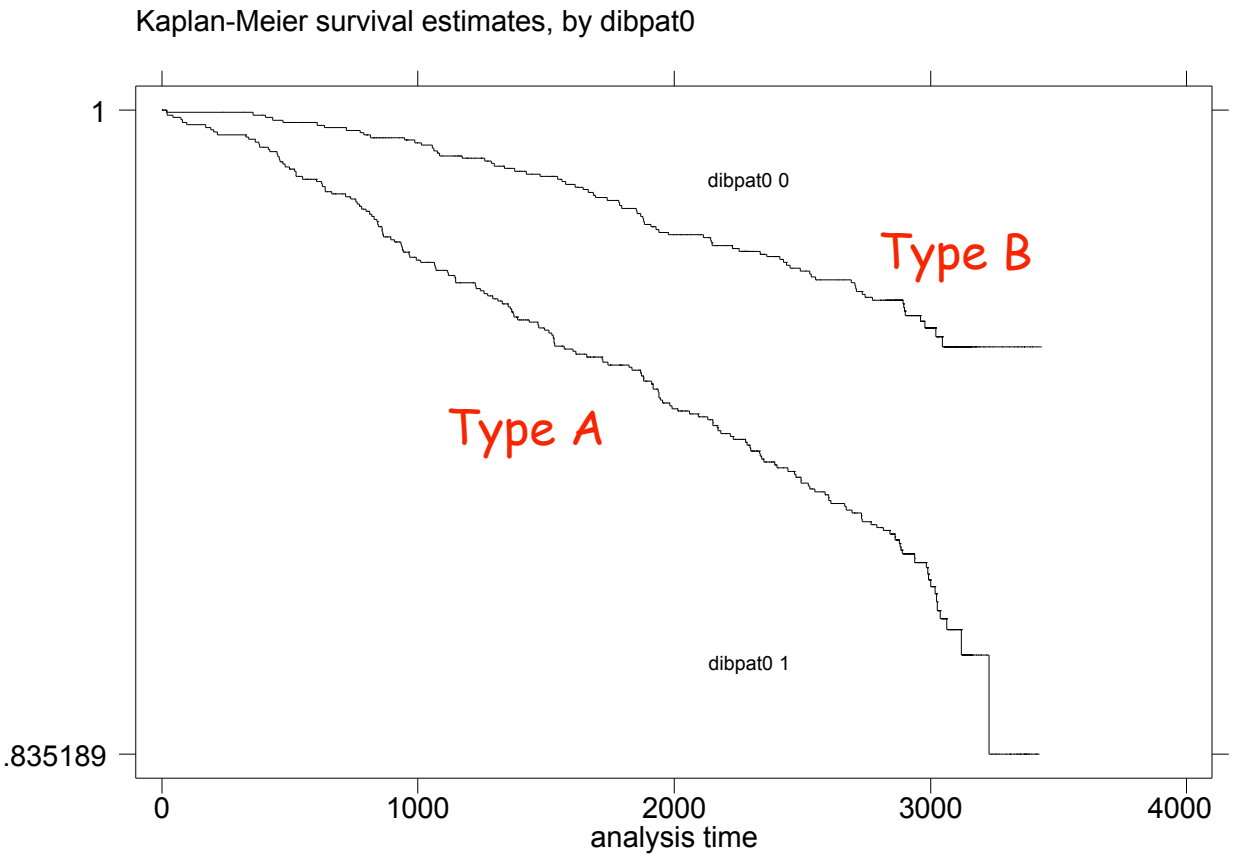
Comparing Groups

sts graph, by(dichol) yas



Dichol =0 if cholesterol < 223 mg/dl

sts graph, by(dibpat0) yas



Regression Model for Hazard Functions

Measure of Association in two groups:

Relative Hazard: $RH(t) = \frac{h_1(t)}{h_0(t)}$

Hard to summarize: assume $RH(t) = \frac{h_1(t)}{h_0(t)} = K$

Proportional Hazards Assumption

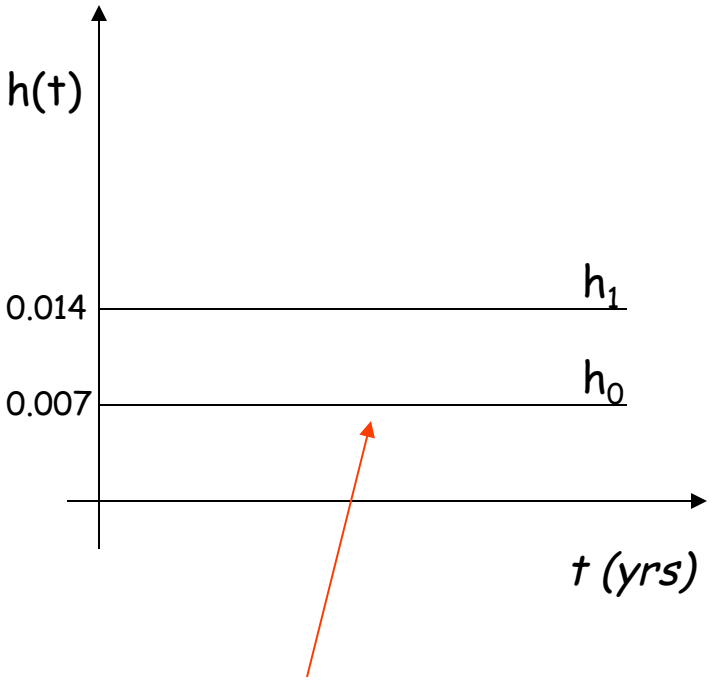
Rare Disease

If $P_0(T)$ is small, then it is easy to see that, under proportional hazards

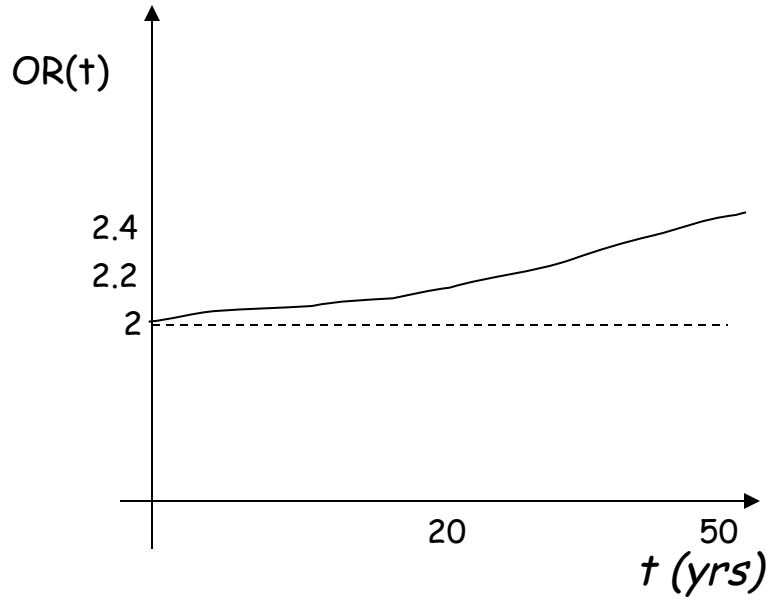
$$OR(t) \approx RR(t) \approx IRR(t) \approx RH(t) \approx K$$

when comparing two groups

Comparison of *OR(t)* and *RH(t)* with proportional hazards



typical values for CHD



Introducing Covariates into *RH*

Suppose you consider risk factor X, and wish to compare two levels of exposure, X_1 and X_0

$$\log\left(\frac{p}{1-p} \mid X\right) = a + bX$$ *(logistic regression)*

$$\log(OR : X_1 \; vs \; X_0) = b(X_1 - X_0)$$

Cox Proportional Hazards Model (1972):

$$\log(RH : X_1 \; vs \; X_0) = c(X_1 - X_0)$$

$$h(t \mid X) = h_0(t)e^{cX}$$

Cox Proportional Hazards Model

$$h(t \mid X) = h_0(t)e^{cX}$$

$h_0(t)$: *Baseline hazard function (X = 0)*

$$\log(RH : X_1 \; vs \; X_0) = c(X_1 - X_0)$$

c: log (Relative hazard associated with unit increase in *X*)

Sir David Cox

Fitting the Proportional Hazards Model

We use a modified version of the likelihood function, called the partial likelihood, and maximize it to find estimates of c *and* $h_{0,}$ *and estimates of their sampling variation*

Usual testing procedures (e.g. Wald tests, likelihood ratio tests) are available as in other regression models

Logistic regression fit

. logit chd69 diage disbp dismoke dichol dibpat0

chd69	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
diage	.5370678	.1367423	3.93	0.000	.2690577	.8050778
disbp	.733672	.1464785	5.01	0.000	.4465794	1.020765
dismoke	.5510715	.1372644	4.01	0.000	.2820382	.8201049
dichol	.9433532	.1495272	6.31	0.000	.6502853	1.236421
dibpat0	.7612871	.1430818	5.32	0.000	.4808519	1.041722
_cons	-4.402261	.2058241	-21.39	0.000	-4.805668	-3.998853

In terms of Odds ratios

logit chd69 diage disbp dismoke dichol dibpat0, or

chd69	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
diage	1.710982	.2339637	3.93	0.000	1.308731	2.23687
disbp	2.082714	.305073	5.01	0.000	1.562957	2.775316
dismoke	1.735111	.2381691	4.01	0.000	1.325829	2.270738
dichol	2.56858	.3840724	6.31	0.000	1.916087	3.443268
dibpat0	2.14103	.3063425	5.32	0.000	1.617452	2.834094

Fitting the Proportional Hazards Model

stcox diage disbp dismoke dichol dibpat0, nohr

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
diage	.5273534	.1268302	4.16	0.000	.2787708	.775936
disbp	.6822427	.1391327	4.90	0.000	.4095476	.9549377
dismoke	.5282569	.1287685	4.10	0.000	.2758752	.7806386
dichol	.9072686	.1429656	6.35	0.000	.6270613	1.187476
dibpat0	.737248	.135597	5.44	0.000	.4714828	1.003013

In terms of Relative Hazards

stcox diage disbp dismoke dichol dibpat0

	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
diage	1.694442	.2149064	4.16	0.000	1.321504	2.172625
disbp	1.978309	.2752475	4.90	0.000	1.506136	2.598509
dismoke	1.695974	.218388	4.10	0.000	1.317683	2.182866
dichol	2.477546	.3542038	6.35	0.000	1.872101	3.278795
dibpat0	2.090175	.2834214	5.44	0.000	1.602368	2.726485

Comparison of Logistic and Proportional Hazards Model

Logistic

chd69	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
diage	1.710982	.2339637	3.93	0.000	1.308731	2.23687
disbp	2.082714	.305073	5.01	0.000	1.562957	2.775316
dismoke	1.735111	.2381691	4.01	0.000	1.325829	2.270738
dichol	2.56858	.3840724	6.31	0.000	1.916087	3.443268
dibpat0	2.14103	.3063425	5.32	0.000	1.617452	2.834094

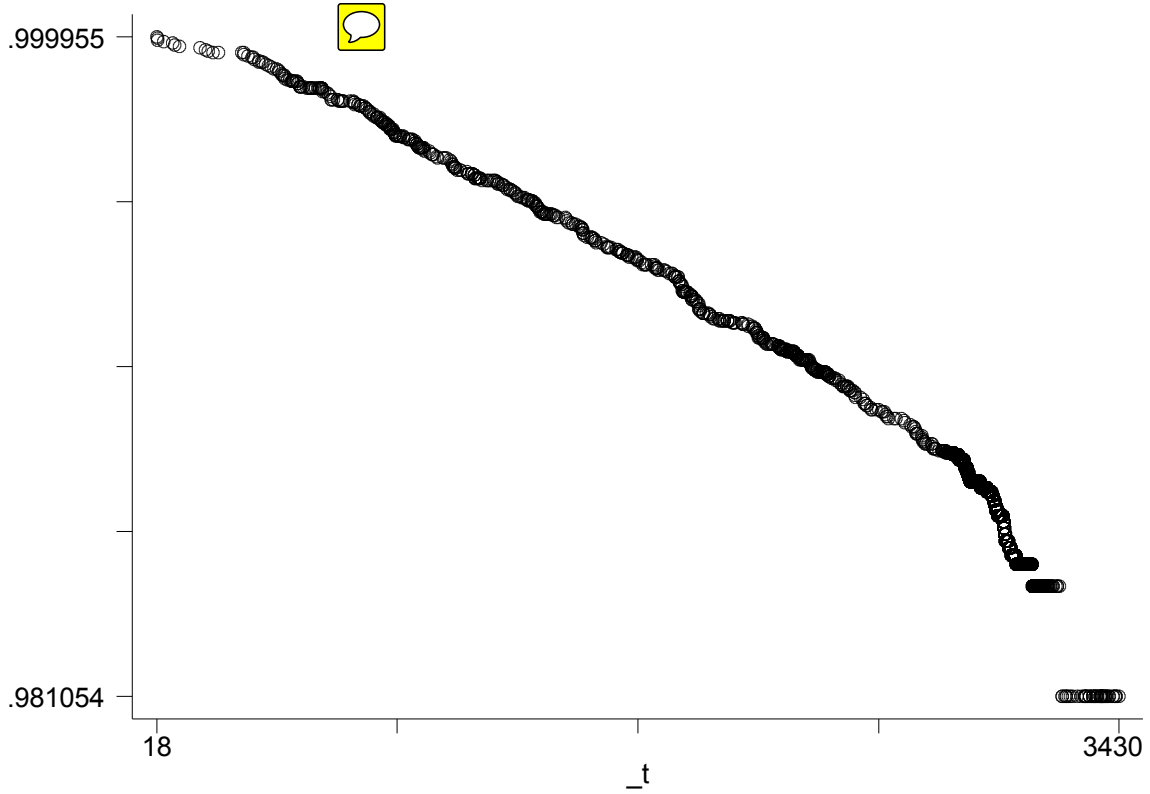
Proportional Hazards

	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
diage	1.694442	.2149064	4.16	0.000	1.321504	2.172625
disbp	1.978309	.2752475	4.90	0.000	1.506136	2.598509
dismoke	1.695974	.218388	4.10	0.000	1.317683	2.182866
dichol	2.477546	.3542038	6.35	0.000	1.872101	3.278795
dibpat0	2.090175	.2834214	5.44	0.000	1.602368	2.726485

Baseline Survival Function Estimate, $S_0(t)$

stcox diage disbp dismoke dichol dibpat0, basesurv(S)

graph S _t



When Does It Make a Difference in fitting the Proportional Hazards Model?

How does the Cox Model work?

Consider the simplest case using a single factor at two levels: $X = 1$ and $X = 0$

First, order all incident event times, irrespective of X

Logrank Test

At each *death* point construct a 2x2 table:

	Dead	Alive	Total
Treat 1			
Treat 2			
Total			

Then treat as set of independent 2x2 tables in Cochran-Mantel-Haenszel test

Survival in months for haemophil.

<41 years of age	>41 years of age
2	1
3+	1
6	1
6	1
7	2
10+	3
15	3
15	9
16	22
27	
30	
32	

At time = 1 month

2
3+
6
6
7
10+
15
15
16
27
30
32

	Dead	Alive	Total
< 41 years	0	11	11
> 41 years	4	5	9
Total	4	16	20

1
1
1
1
2
3
3
9
22

At time = 2 month

2
3+
6
6
7
10+
15
15
16
27
30
32

	Dead	Alive	Total
< 41 years	1	10	11
> 41 years	1	4	5
Total	2	14	12

1
1
1
1
2
3
3
9
22

Note: Proportional Hazards assumption is equivalent to assumption of "no interaction", necessary for appropriate use of Cochran-Mantel-Haenszel test

```
. sts test age

  Log-rank test for equality of survivor functions

       |    Events
 age   |  observed      expected
-------+-------------------------
 0     |        10         14.67
 1     |         9          4.33
-------+-------------------------
Total  |        19         19.00

           chi2(1) =       8.02
           Pr>chi2 =     0.0046
```

Example: Western Collaborative Group Study

sts test dichol

Log-rank test for equality of survivor functions

dichol	Events observed	Events expected
0	67	129.54
1	190	127.46
Total	257	257.00

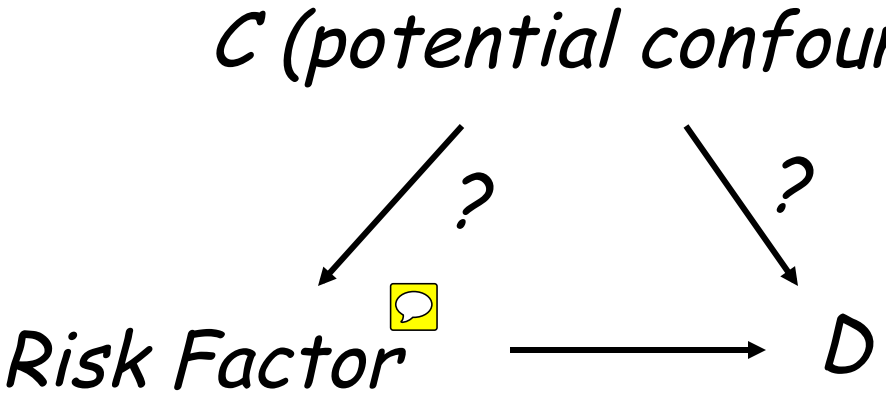
chi2(1) = 60.95
Pr>chi2 = 0.0000

Interpretation: Stratification on Time at Risk

Heuristically, we are treating time at risk as a potential confounding variable

Proportional hazards assumption means that there is no change in RH over the levels of time at risk (i.e. no time--covariate interaction)

When Does "Time at Risk" Confound?



Conditions for confounding:

(1) Time at Risk and D are associated (almost always true)

and

(2) Time at Risk and Risk Factor are associated (sometimes true)

“Time at Risk” as a Confounder

- Time-dependent Covariates
- Differential Loss to Follow-up

Differential Loss to Follow-up

This is easiest to see if you think of new individuals supplying the "risk set" at different times:

Suppose outcome is CHD, risk factor (F) is smoking.

Without loss to follow-up: at first "risk period" 100 in risk with F and 100 at risk without F. Same at second later risk period.

With loss to follow-up: at first "risk period" 100 in risk with F and 100 at risk without F. At second risk period, 100 at risk without F, but only 75 at risk with F (smoking killed off, from other causes, 25 of those who would normally have still been at risk at the second time)

Differential Loss to Follow-up (cont'd)

Without loss to follow-up

	First Risk Period	Second Risk Period
F	100	100
not F	100	100

With loss to follow-up

	First Risk Period	Second Risk Period
F	100	75
not F	100	100

Differential loss to follow-up has induced a relationship between F and time at risk

Stanford Heart Transplant Data

- Data is from Crowley and Hu (1977)
- Patients are admitted to program when need for heart transplant is determined
- Patients then wait for suitable donor heart (lasts from a few days to years)
- Some patients die before suitable donor heart is found
- All patients followed to death

Stanford Heart Transplant data (time dependent covariate)

```
stcox transplant

        failure _d:  status
  analysis time _t:  t1
                id:  patno

Cox regression -- Breslow method for ties

No. of subjects =          103                     Number of obs   =       172
No. of failures =           75
Time at risk    =      31954.1
                                                   LR chi2(1)      =     25.75
Log likelihood  =   -285.44037                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |
          _d | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
transplant |   .2674327   .0652523    -5.41   0.000     .1657774    .4314233
------------------------------------------------------------------------------
```

```
stcox posttran

        failure _d:  status
  analysis time _t:  t1
               id:  patno

Cox regression -- Breslow method for ties

No. of subjects =          103                     Number of obs   =       172
No. of failures =           75
Time at risk    =      31954.1
                                                   LR chi2(1)      =      0.17
Log likelihood  =   -298.22883                     Prob > chi2     =    0.6778

------------------------------------------------------------------------------
          _t |
          _d | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    posttran |   1.132577   .3408133     0.41   0.679     .6279502    2.042725
------------------------------------------------------------------------------
```

Summary---Lessons Learned

- Poisson regression cannot handle varying hazards or varying incidence rate ratios
- Survival analysis techniques allow estimation of hazard and survival function over time
- Proportional hazards model allows study of the effect of risk factors on time to outcome
- Proportional hazards model often similar to simple logistic regression analysis based on occurrence of outcome
- Proportional hazrds valuable with time-dependent risk factors and differential loss to follow-up

References

- *The Statistical Analysis of Failure Time Data*, J. D. Kalbfleisch & R. Prentice, 1980, Wiley
- *Statistical Analysis of Epidemiologic Data*, Steve Selvin, 1991, Oxford University Press
- *Modelling Survival Data in Medical Research*, D. Collett, 1994, Chapman & Hall