Survival Analysis

DZHW Summer School Workshop

6.9.2016

Johannes Giesecke

HU Berlin

Program

Block I Basics

a) Key concepts (censoring, truncation, analysis time and functions thereof, overview of potential models)

b) Data structure and data management (wide and long data format, handling of dates, st-commands in Stata)

Block II Some Examples using Different Models

a) Non-parametric approaches

b) Semi-parametric models

c) Parametric models

d) Outlook

Literature

- Allison, Paul D. (2014). Event History and Survival Analysis. Thousand Oaks: Sage Publications
- Blossfeld, Hans-Peter/Golsch, Katrin/Rohwer, Götz (2007). Event History Analysis with Stata. Mahwah: Lawrence Erlbaum Associates
- Box-Steffensmeier, Janet M./Jones, Bradford S. (2004). Event History Modeling. A Guide for Social Scientists. Cambridge: Cambridge University Press
- Cleves, Mario/Gould, William/Gutierrez, Roberto G./Marchenko Yulia V. (2010). An Introduction to Survival Analysis Using Stata. College Station: Stata Press
- Hosmer, David W./Lemeshow, Stanley (1999). Applied Survival Analysis. Regression Modeling of Time to Event Data. New York: John Wiley & Sons
- Kleinbaum, David G./Klein, Mitchel (2011). Survival Analysis: A Self-Learning Text. Berlin: Springer Verlag

Basics

Introduction

- aim:
  - analysis of time until an (predefined) event occurs
  - analysis of distribution of events
  - investigate groups differences in time-to-event
  - examples: job search duration, employment duration, entry into first job with permanent contract, time until first substantial pay raise etc.

- synonyms:
  - hazard models, event history analysis, models for transition data, time-to-event data, survival-time data, duration data, failure time

Introduction

- data soures
  - cross-sectional surveys collecting retrospective information (e.g. German Life History Study)
  - panel studies collecting prospective information (e.g. German Socioeconomic Panel SOEP)
  - administrative data (e.g. employment information collected by Federal Emplyoment Agency)

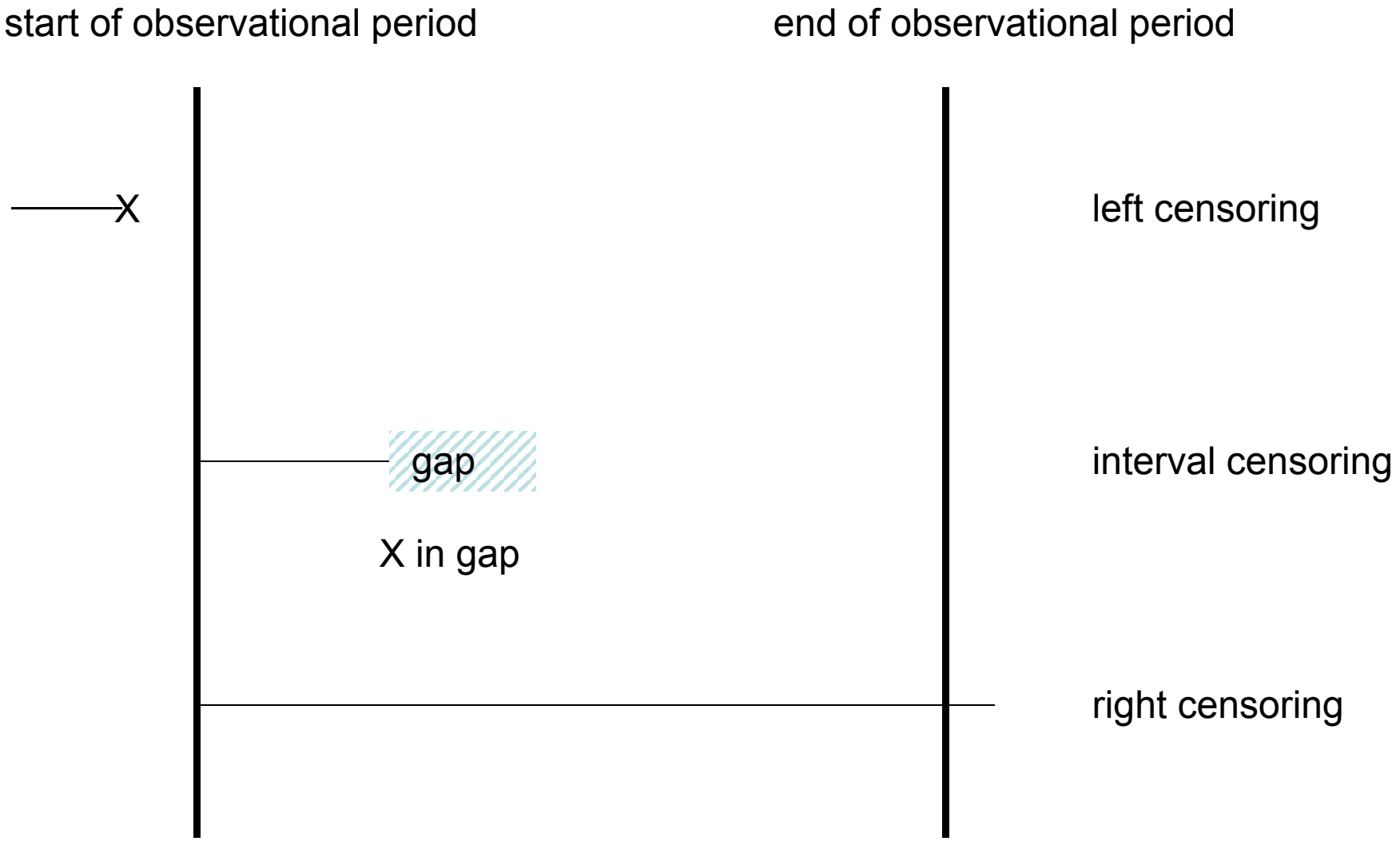
Alternative Models?

- Why not using OLS-models (dependent variable: time until an event occurs)?
  - general problem: How to deal with censored cases? (but: censored regression models)
  - more important: non-normally distributed dependent variable and time-varying variables

- Why not using logit/probit-models (dependent variable: event occurred vs. event did not occur)?
  - loss of information (chronology is lost)
  - time-varying variables cannot be integrated

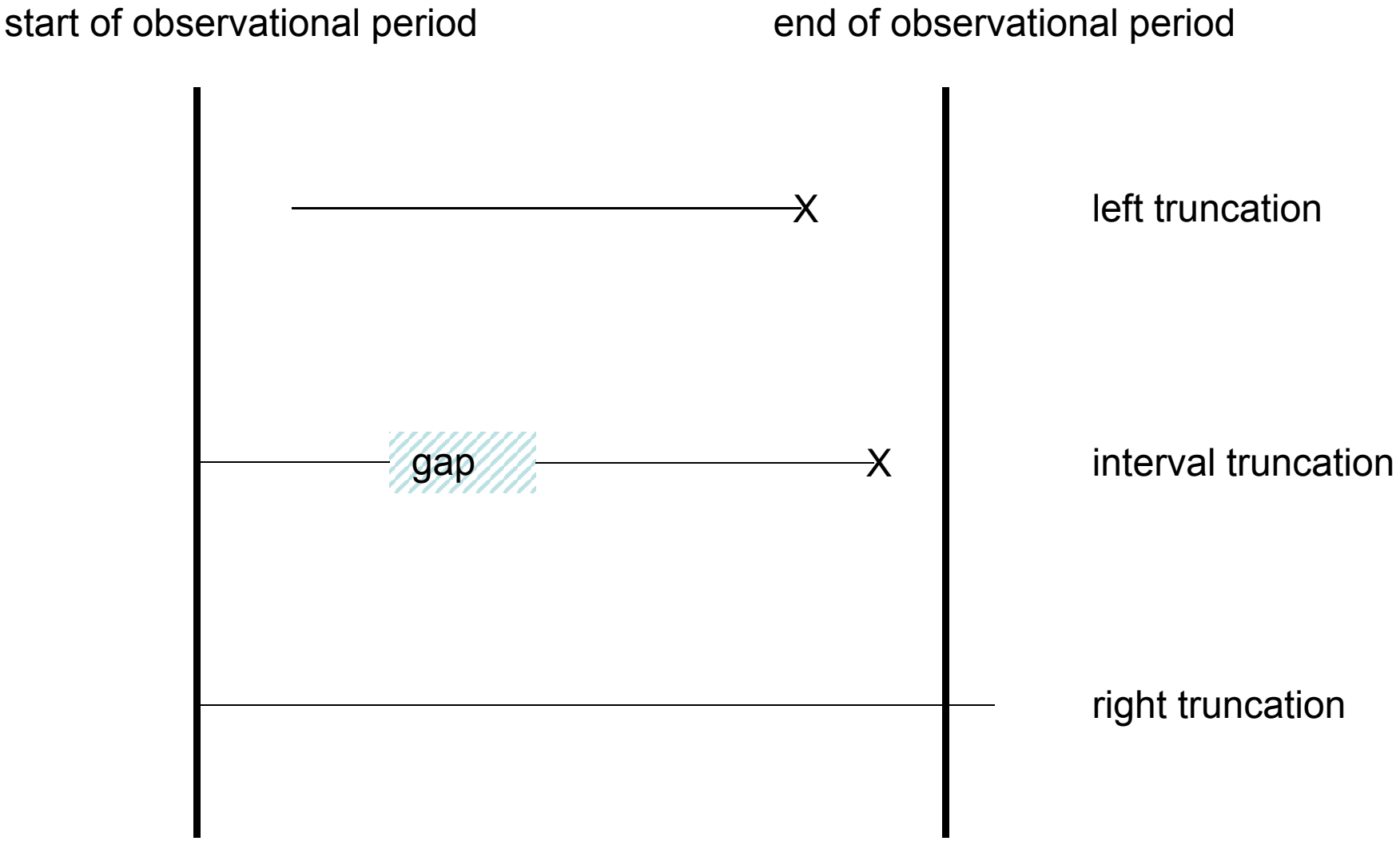
Censoring and Truncation

- censoring: event occurs, but observational unit is not observed
  - right censoring: event occurs after observational period has ended (observational unit no longer observed)
  - left censoring: event occurred before observational period started (observational unit not yet observed)
  - interval censoring: event occurred between start and end of the observational period, but no exact timing of event is known, only time interval (observational unit was not observed during time interval)


start of observational period
end of observational period
X
left censoring
gap
interval censoring
X in gap
right censoring

Censoring and Truncation

- truncation: observational unit is not observed for a certain time span
  - left truncation (delayed entry): observational unit was „at risk“ before it was actually observed for the first time
  - right truncation: observational unit UE is no longer observed after the end of the observational period, but event will finally occur; effectively not distinguishable from right censoring
  - interval truncation (gaps): observational unit is not observed for a certain time span between start and end of the observational period


start of observational period
end of observational period
left truncation
gap
interval truncation
right truncation

Continuous vs. Discrete Time

- continuous time:
  - idea: continuous time axis, no time intervals but time points (infinitely small time intervals)
  - basis of (almost) all textbooks on survival analysis

- discontinuous (discrete) time
  - grouping of originally continuous time (grouping happens during or after observation phase)
  - often neglected in textbooks, but quite important in practice

Types of Variables

- time-constant variables (e.g. sex, social origin, graduation grade)

- time-varying variables (e.g. number of job applications, type of search strategy)
  - splitting of episodes might be necessary

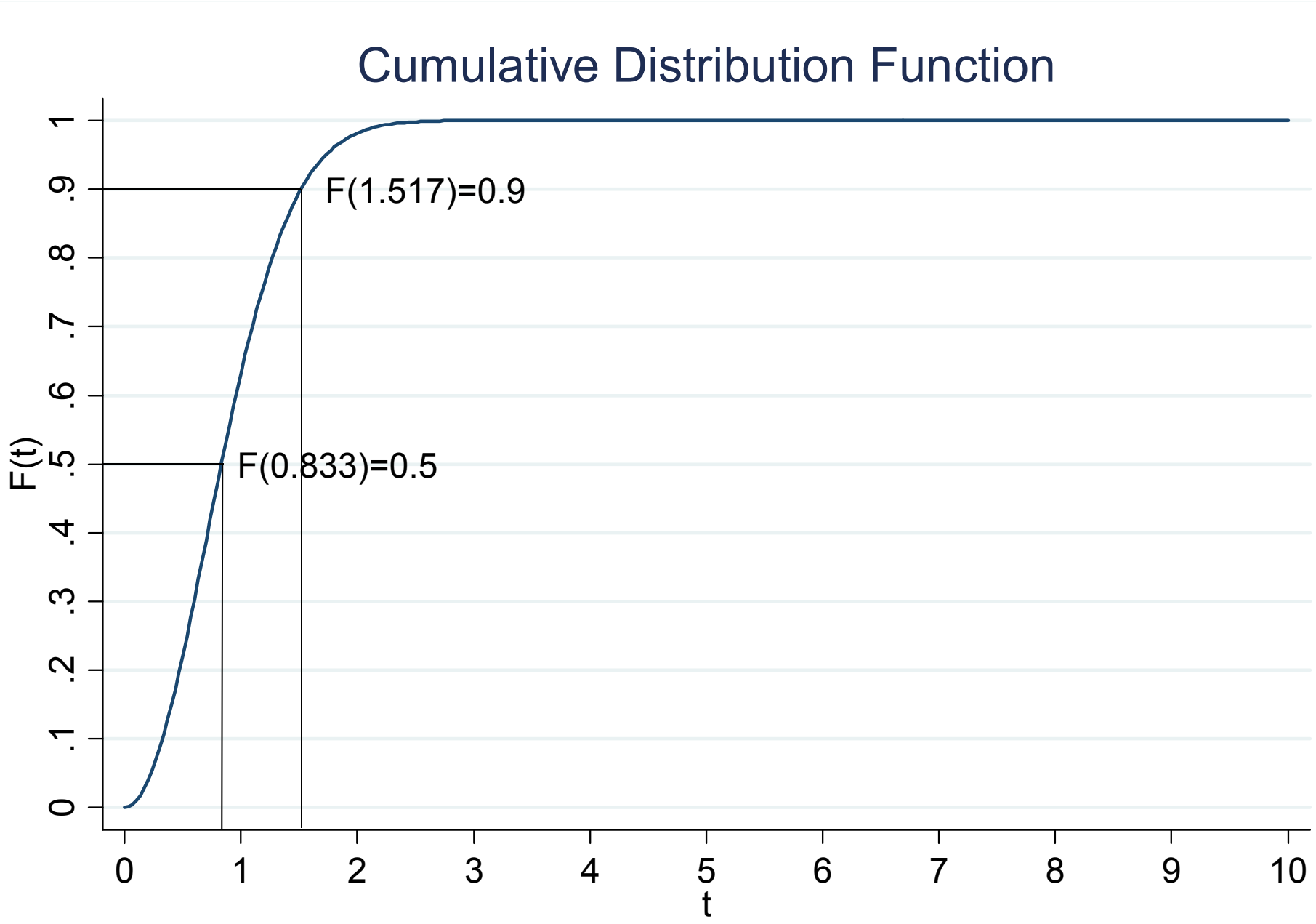
Analysis Time and Functions thereof

T non-negative random variable denoting the time until an event occurs

1. continuous time

a) $F(t)=Pr(T \leq t)$ cumulative distribution function of T

- probability that there is an event prior to t
- monotone, nondecreasing function
- also known as failure function



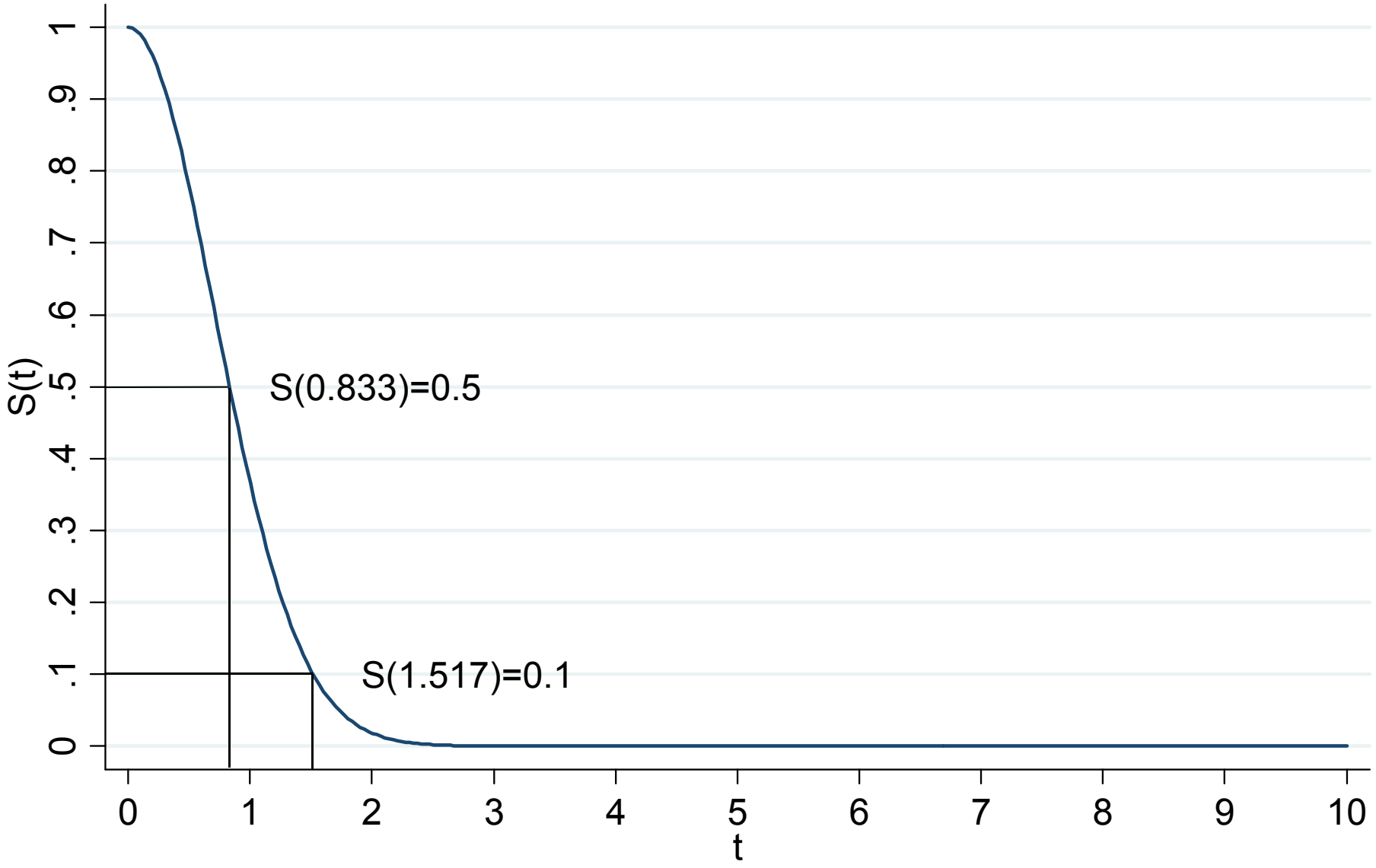
Weibull Distribution p=2

Analysis Time and Functions thereof

b) $S(t)=1-F(t)= Pr(T>t)$ survivor function

- probability that there is no event prior to t
- monotone, nonincreasing function
- $S(0)=1$, $S(t)=0$ for $t \rightarrow \infty$

Survivor Function



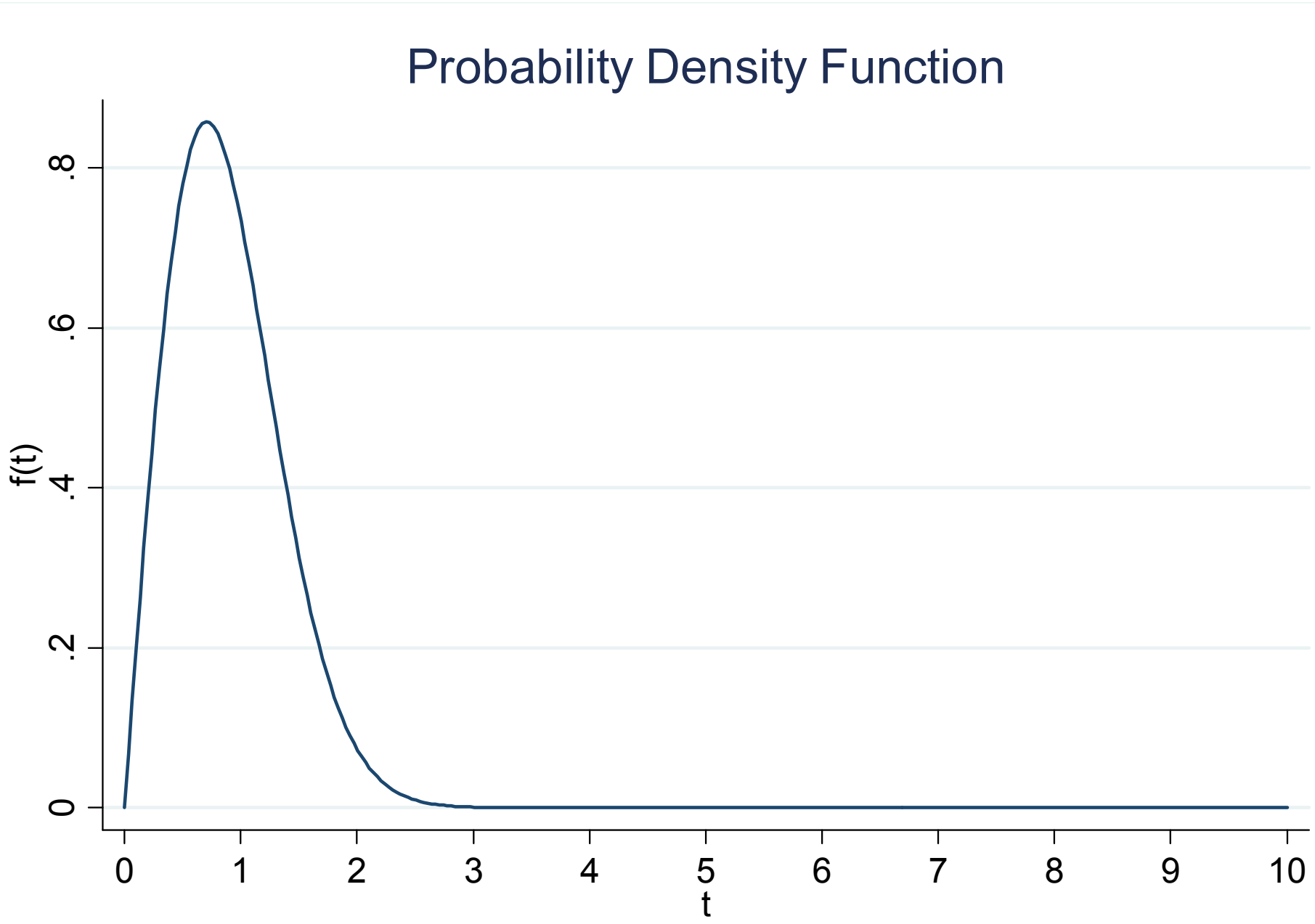
Weibull Distribution p=2

Analysis Time and Functions thereof

c) f(t) density function

- measure of concentration of events at time t
- density function, no probabilities
- note: 0 ≤ f(t), but not f(t) ≤ 1

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr\left(t \le T \le t + \Delta t\right)}{\Delta t} = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t}$$



Weibull Distribution p=2

Analysis Time and Functions thereof

d) h(t) hazard function

- measure of concentration of events at time t, <u>conditional</u> upon the subject having survived until t

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\left(t < T \leq t + \Delta t \middle| T > t\right)}{\Delta t} = \frac{f(t)}{S(t)}$$

<u>therefore:</u>
given a certain level of f(t): the smaller S(t) the higher h(t)

given a certain level of S(t): the higher f(t) the higher h(t)

also known as: hazard rate, intensity rate, failure rate, transition rate, transition intensity, risk function, mortality rate

Analysis Time and Functions thereof

characteristics of hazard function:

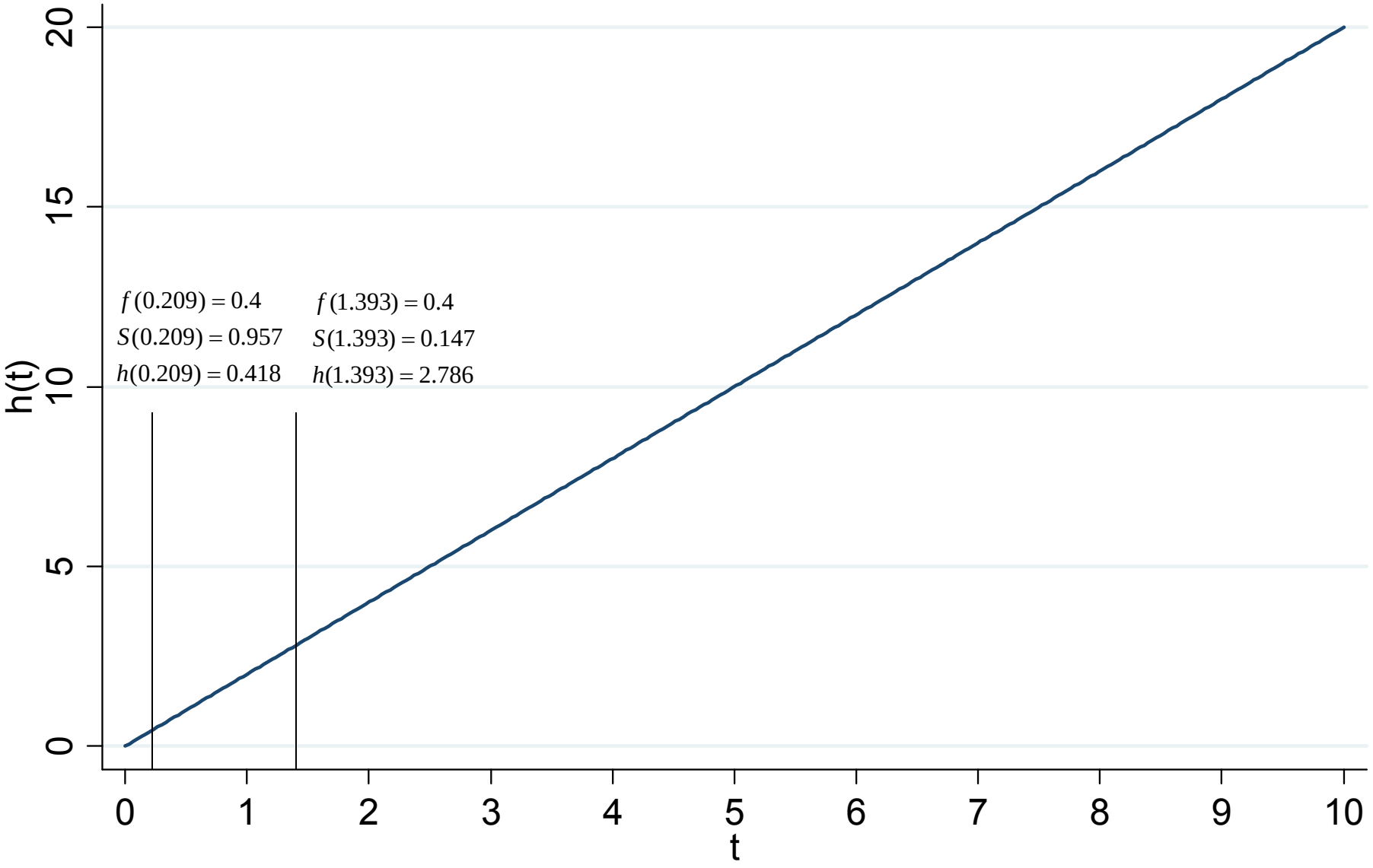
- values between 0 (no risk) to ∞ (certainty of event occurring at this particular instance)

- time-constant hazard: conditional upon the event not having occurred until t, the chances of surviving at this instance or that are all the same

- increasing hazard: increasing risk

- falling hazard: falling risk

Hazard Function



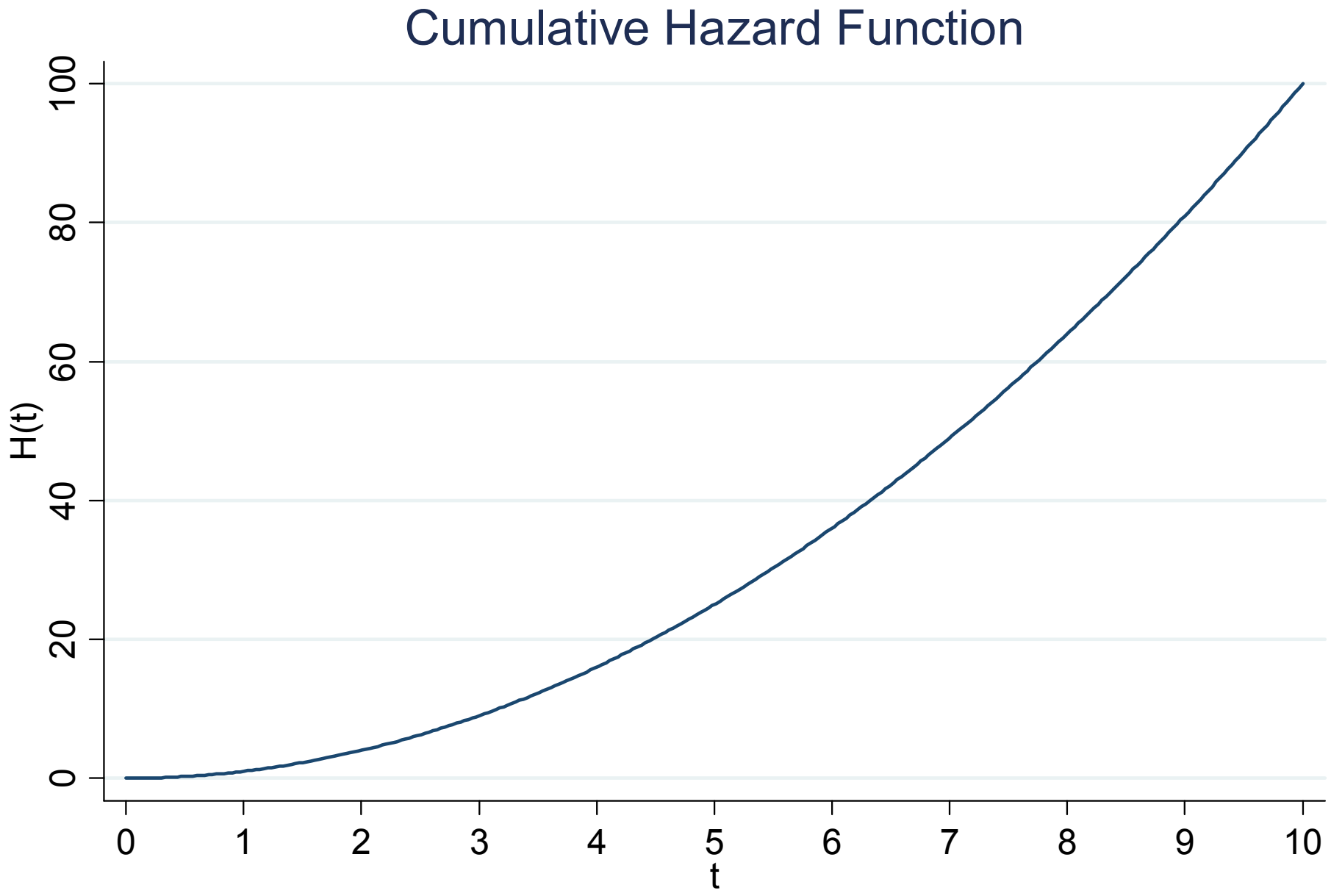
Weibull Distribution p=2

Analysis Time and Functions thereof

e) H(t) cumulative hazard function

- measure of total risk a subject has passed through until time t

$$H(t) = \int_0^t h(u)du = -\ln\left[S(t)\right]$$



Weibull Distribution p=2

Analysis Time and Functions thereof

if one particular function is known, all other functions can be determined

e.g. if $h(t) = 1$ then:

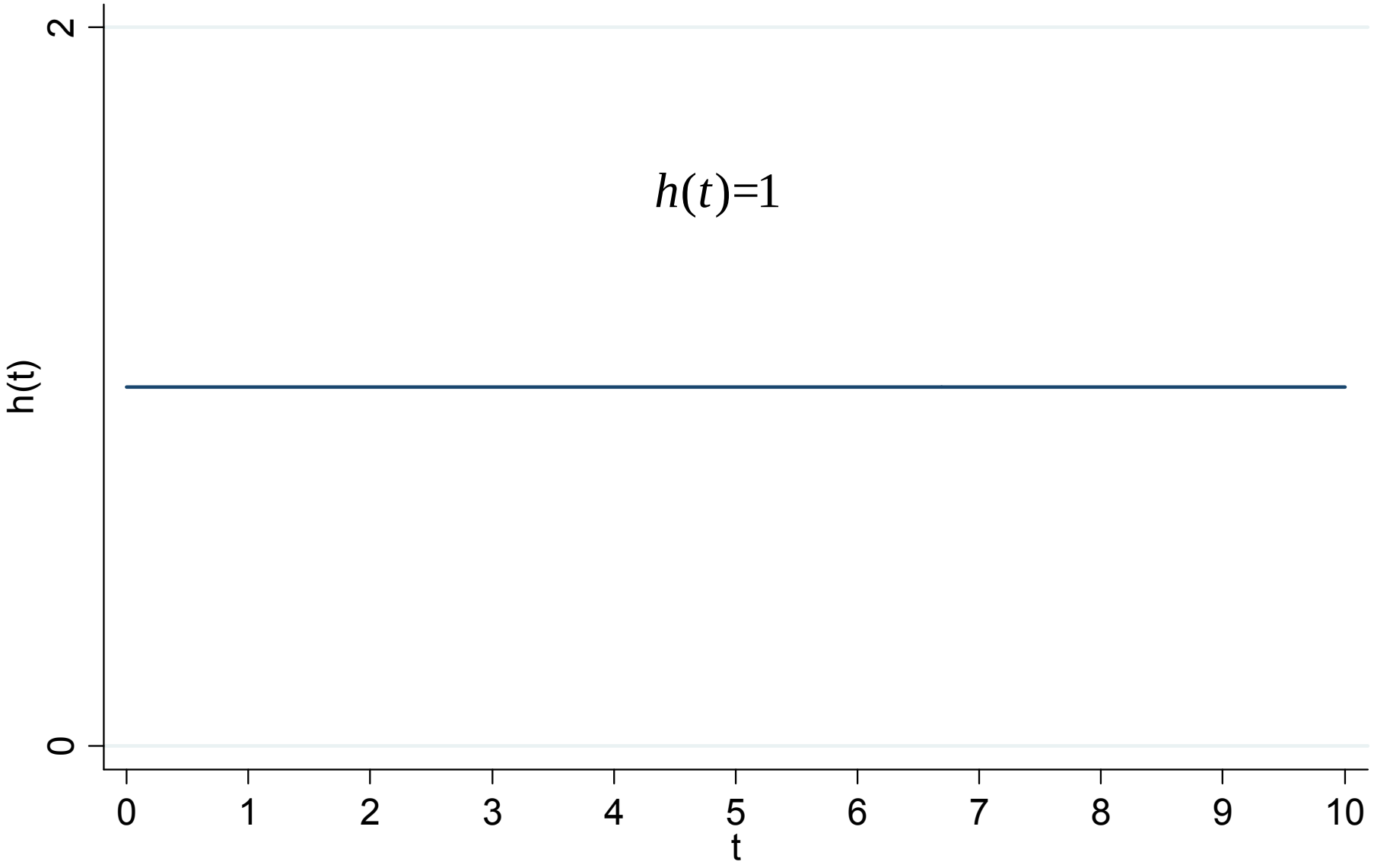
$$H(t) = \int_0^t h(u)du = t$$

$$S(t) = \exp\left[-H(t)\right] = \exp(-t)$$

$$F(t) = 1 - S(t) = 1 - \exp\left[-H(t)\right] = 1 - \exp(-t)$$

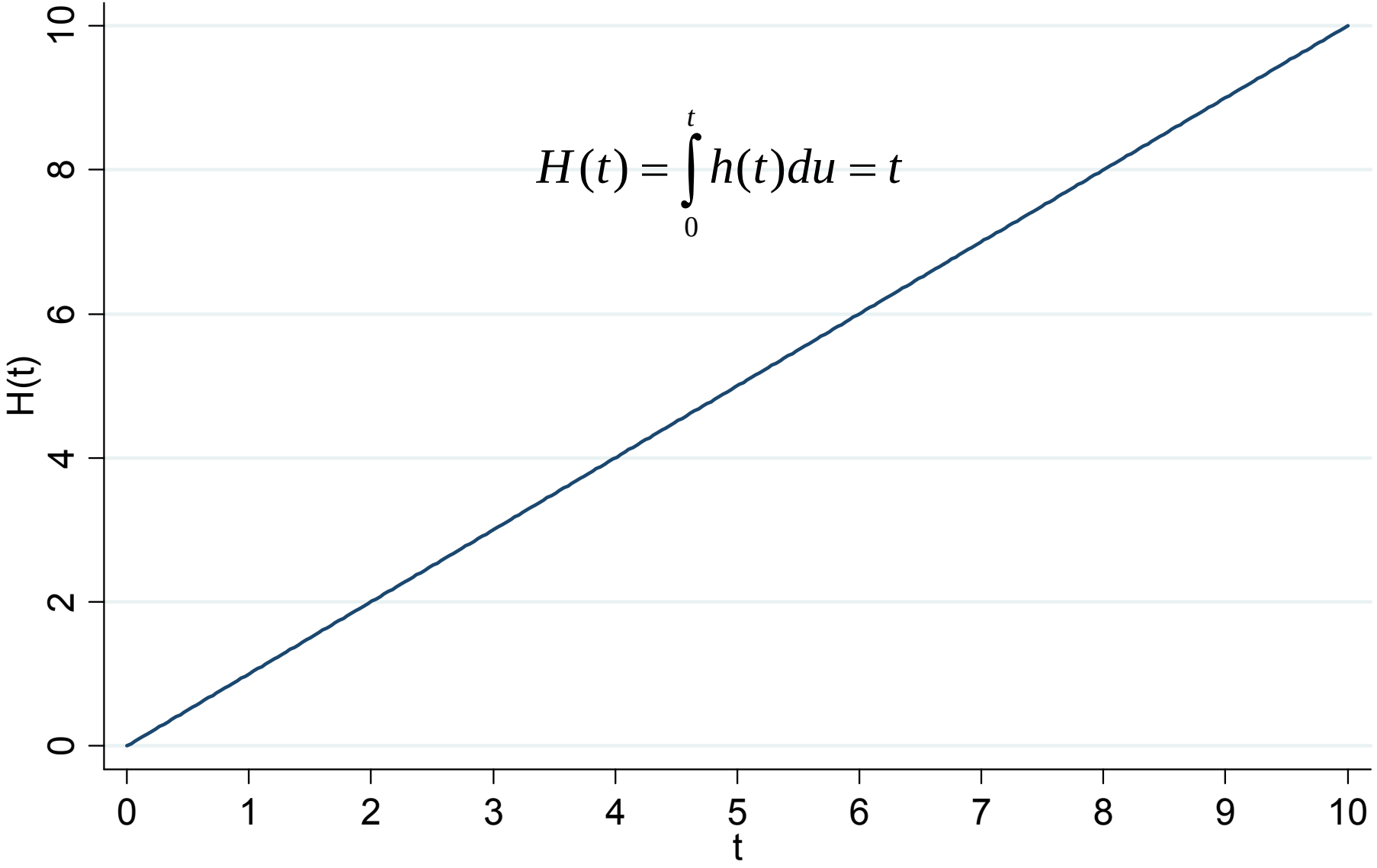
$$f(t) = h(t)S(t) = h(t)\exp\left[-H(t)\right] = \exp(-t)$$

Hazard Function



Weibull Distribution p=1

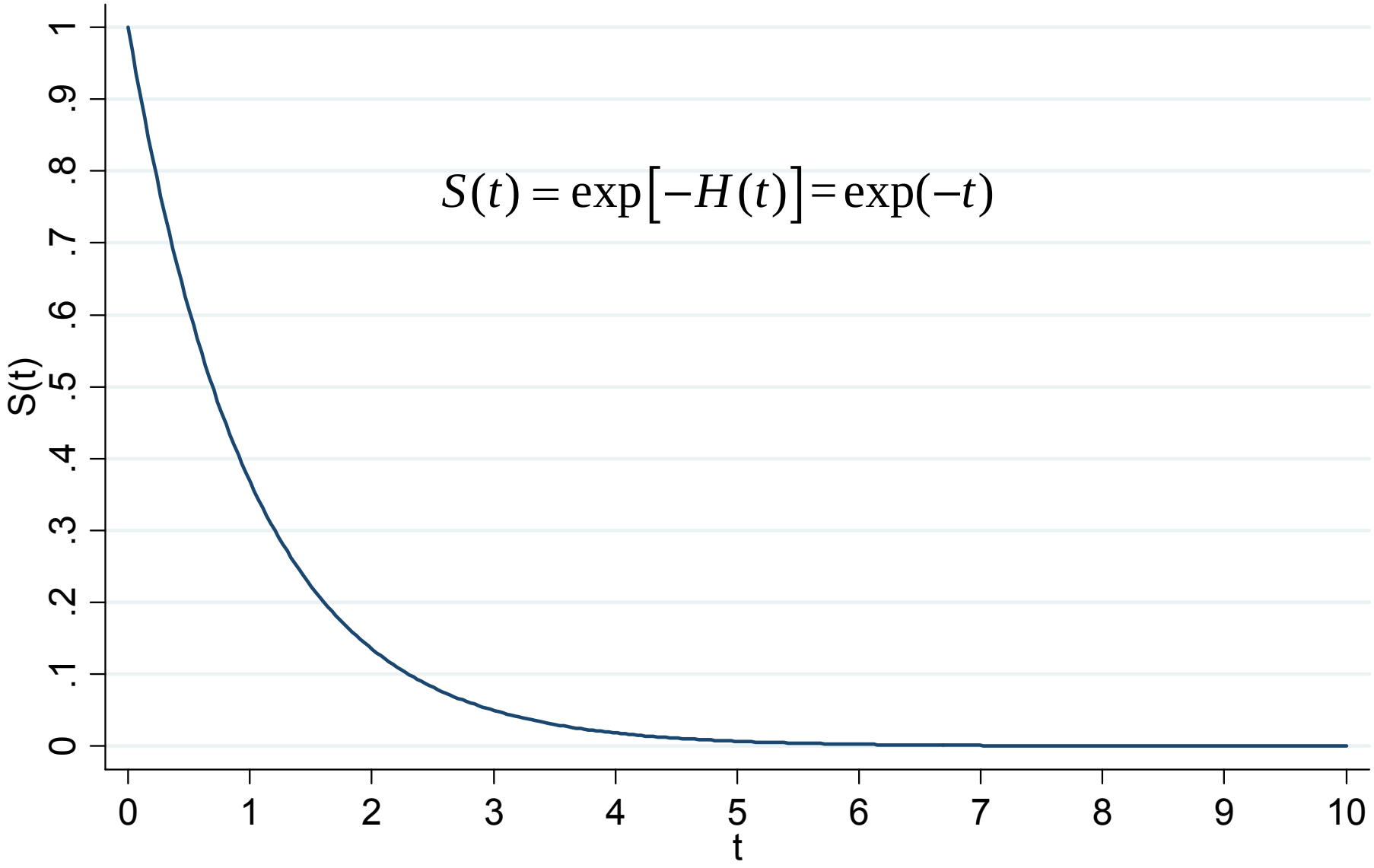
Cumulative Hazard Function



Weibull Distribution p=1

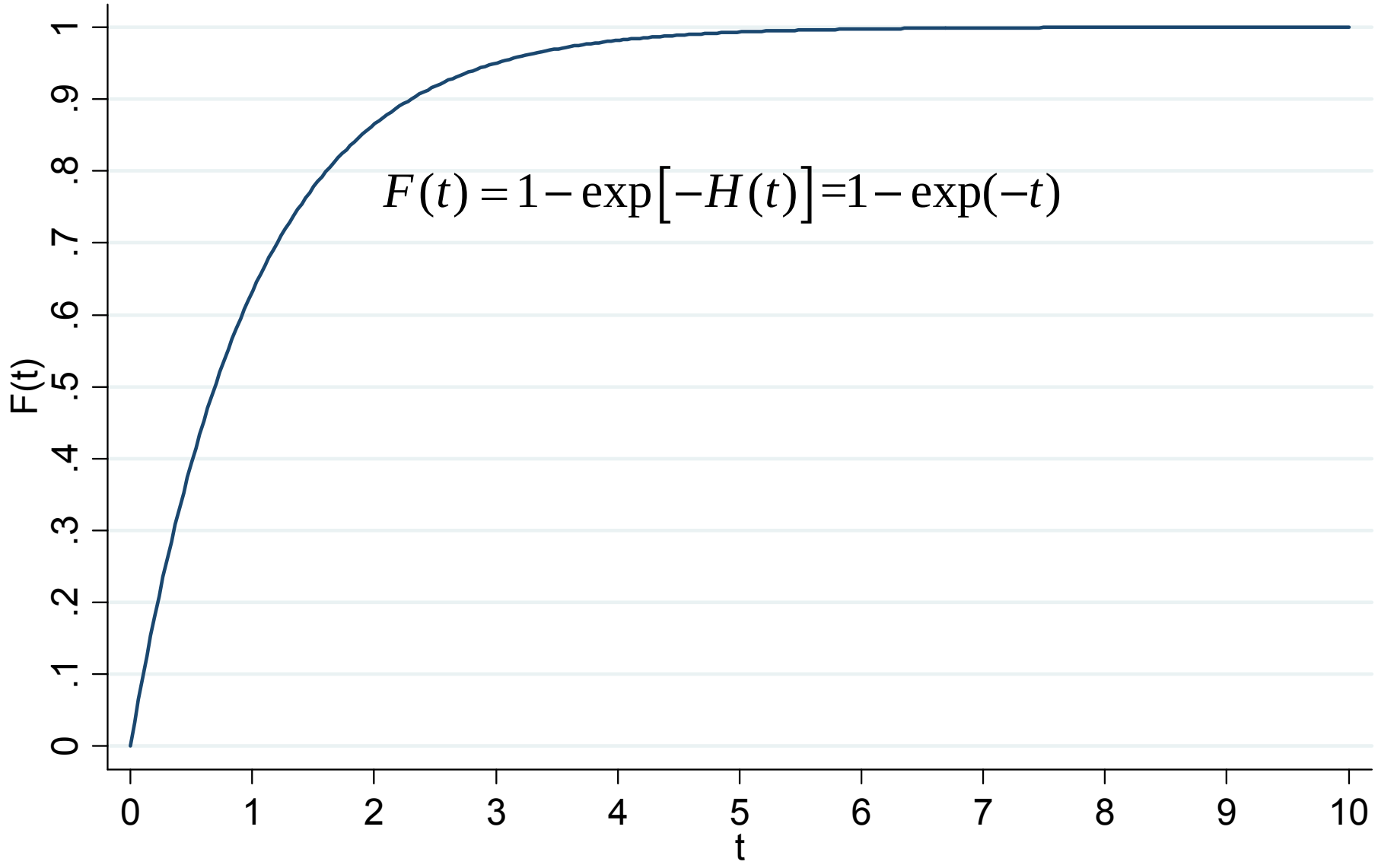
Survivor Function

$$S(t) = \exp\left[-H(t)\right] = \exp(-t)$$



Weibull Distribution p=1

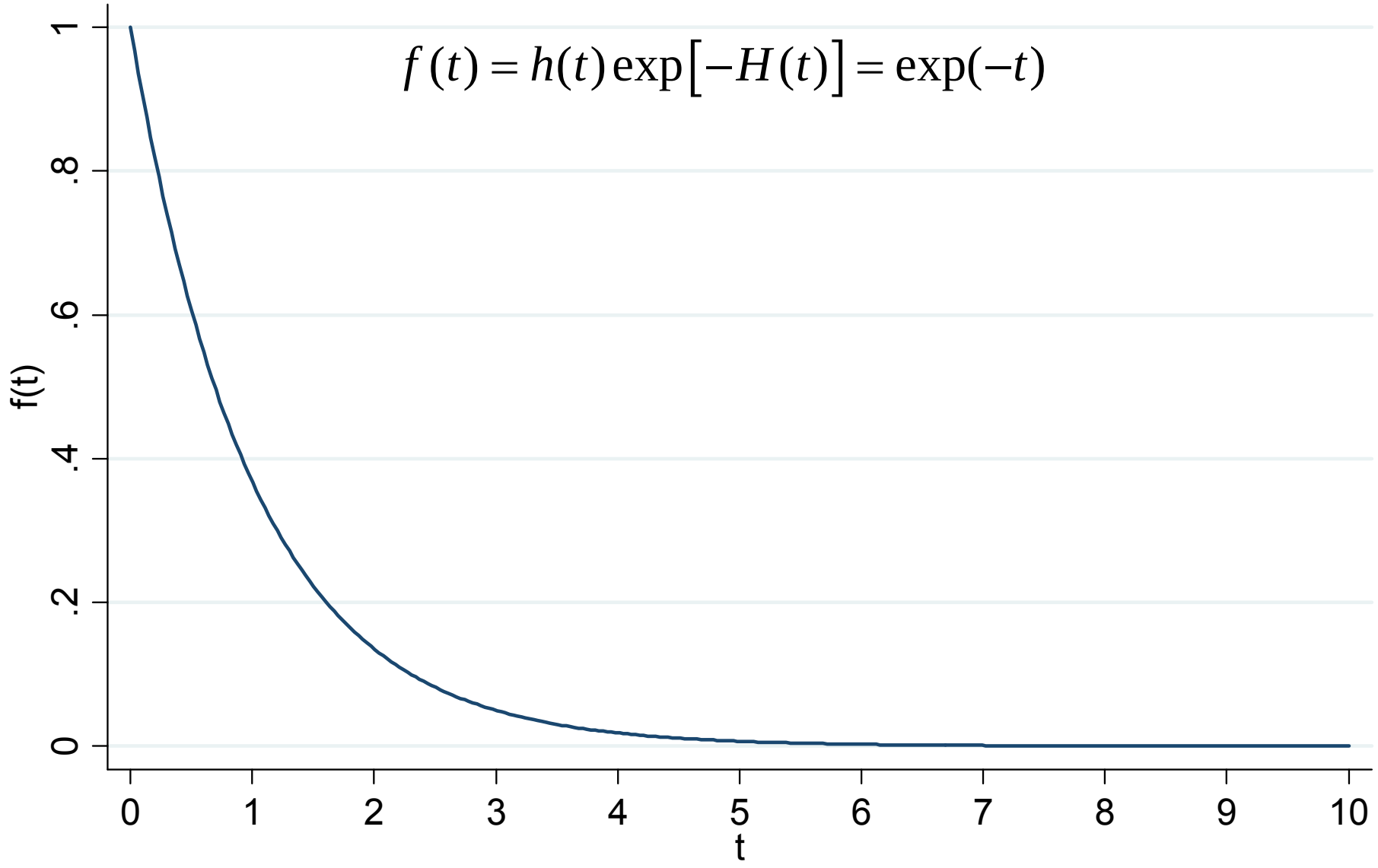
Cumulative Distribution Function



Weibull Distribution p=1

Probability Density Function

$$f(t) = h(t)\exp\left[-H(t)\right] = \exp(-t)$$



Weibull Distribution p=1

Analysis Time and Functions thereof

2. discontinuous time

idea: T still is continuous random variable, but time axis is divided into k disjoint intervals

$[0=a_0,a_1]$, $(a_1,a_2]$, $(a_2,a_3]$,…, $(a_{k-1},a_k=\infty]$

interval j: $(a_{j-1},a_j]$] – including

(– excluding

Analysis Time and Functions thereof

F(.) und S(.) are still defined as

$F(a_j)=Pr(T \leq a_j)$

- probability that an event occurs up to the end of time interval j

$S(a_j)=1-F(a_j)= Pr(T>a_j)$

- probability that no event occurs up to the end of time interval j (i.e., that subject “survives” at least until interval j+1)

Analysis Time and Functions thereof

f(.) density function of T, here: probability function

$$f(j) = \Pr\left(a_{j-1} < T \leq a_j\right) = S(j-1) - S(j)$$

probability of the event occurring in time interval j

Analysis Time and Functions thereof

interval hazard defined as:

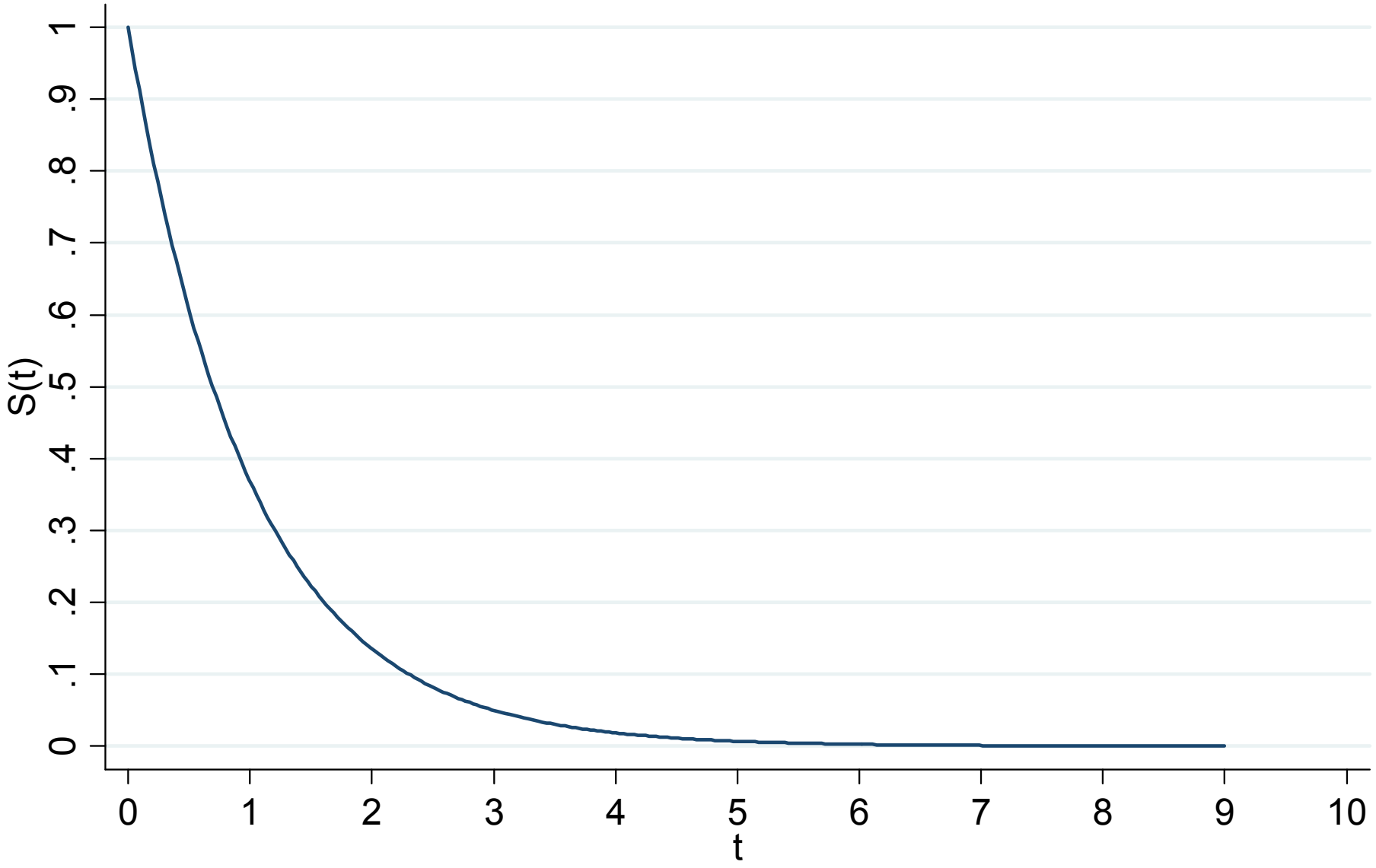
$$h(a_j) = \Pr\left(a_{j-1} < T \le a_j \middle| T > a_{j-1}\right)$$

$$= \frac{\Pr\left(a_{j-1} < T \le a_j\right)}{\Pr\left(T > a_{j-1}\right)} = \frac{S(a_{j-1}) - S(a_j)}{S(a_{j-1})} = 1 - \frac{S(a_j)}{S(a_{j-1})}$$

probability of the event occuring in time interval j, given that subject has survived to the beginning of that interval

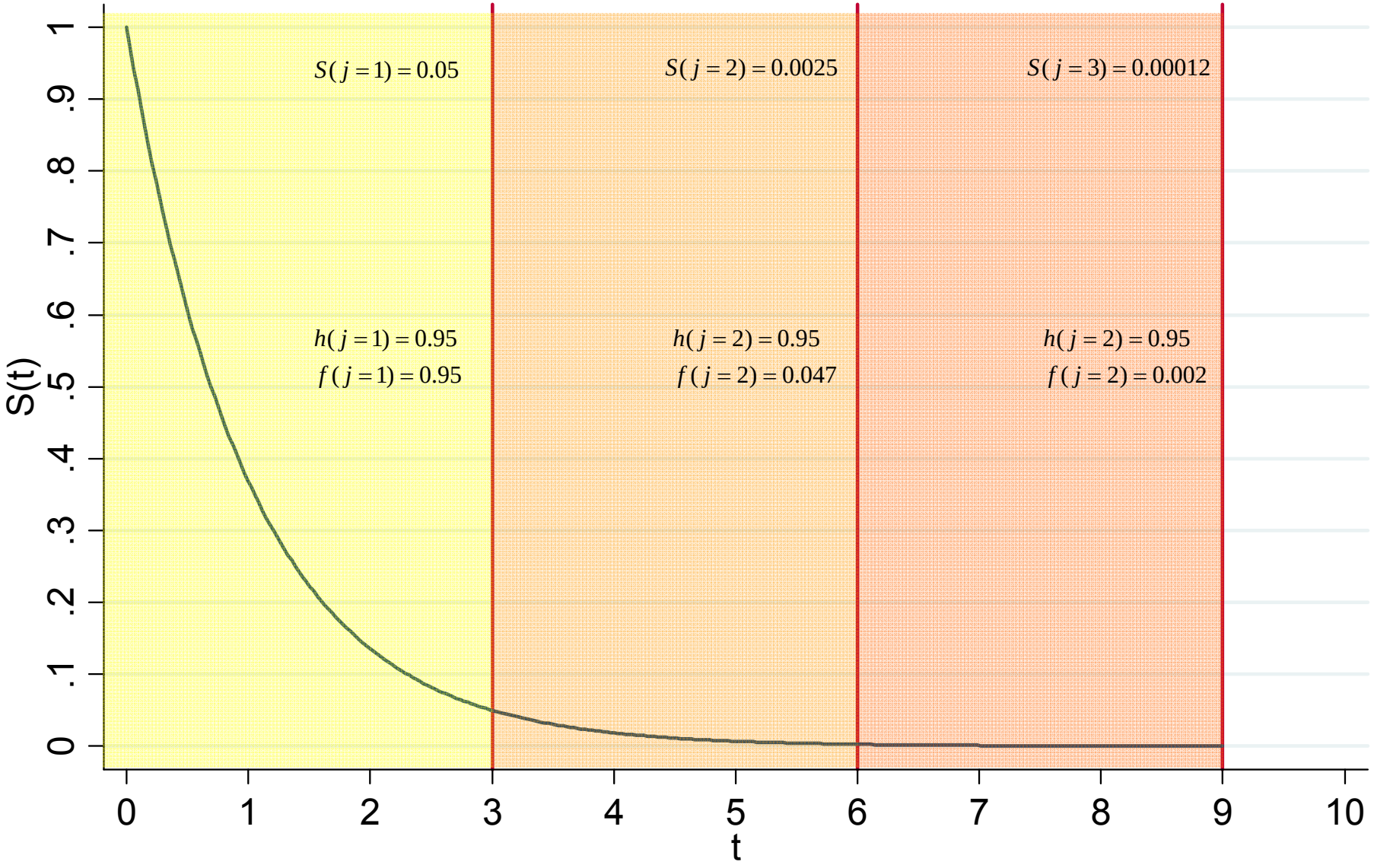
conditional probability, e.g. $0 \le h(a_j) \le 1$

Survivor Function



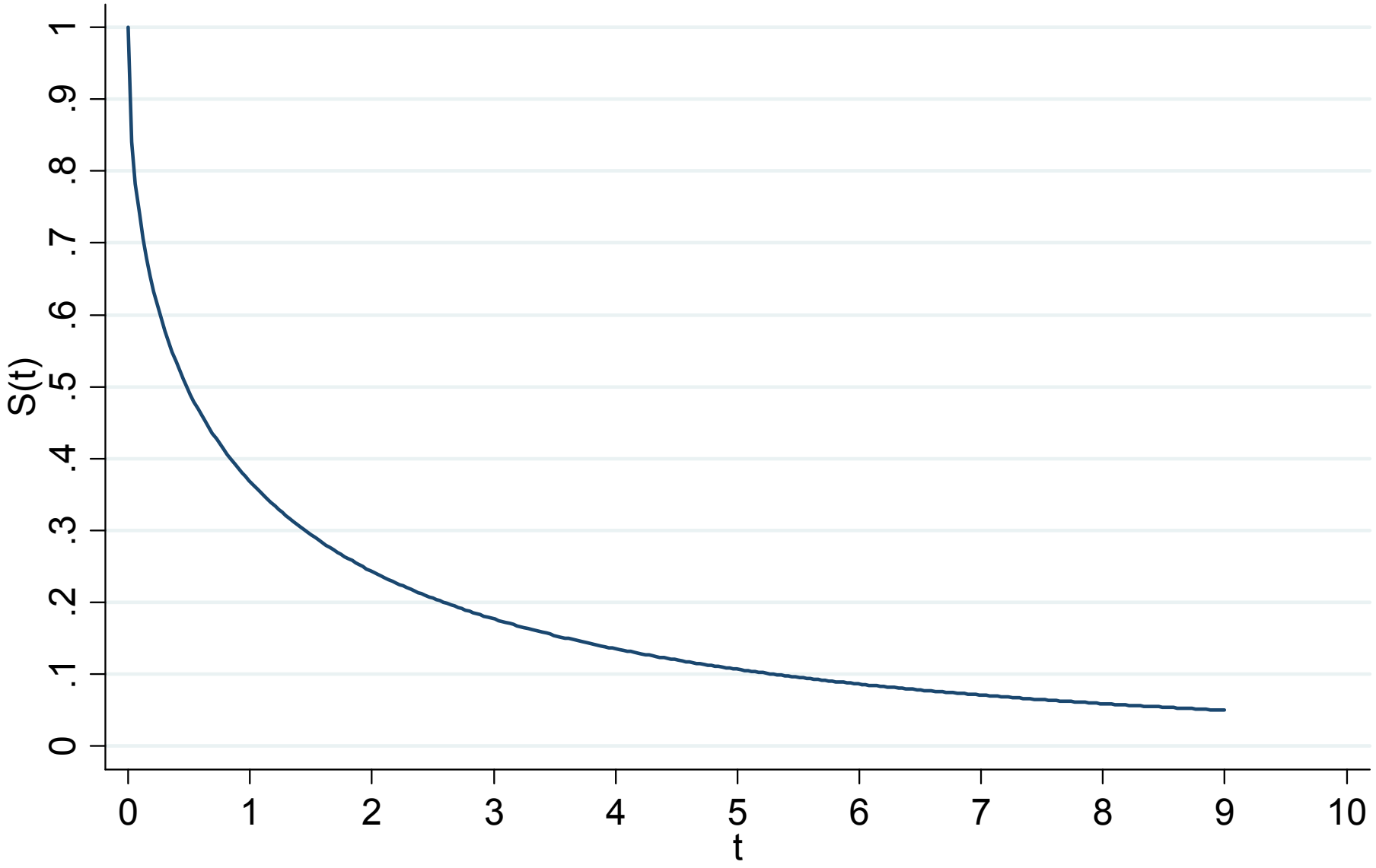
Weibull Distribution p=1

Survivor Function



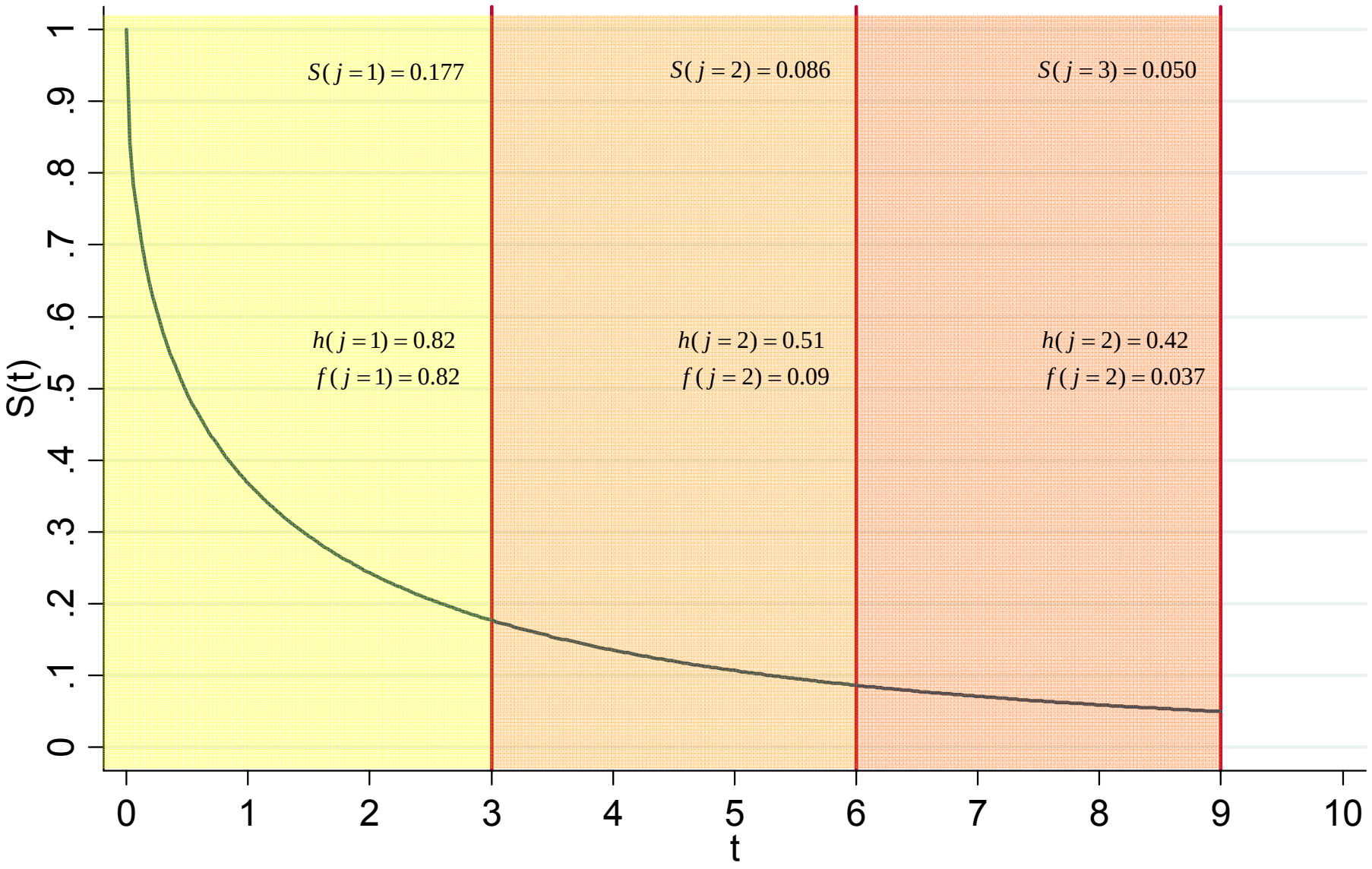
Weibull Distribution p=1

Survivor Function



Weibull Distribution p=0.5

Survivor Function



Weibull Distribution p=0.5

Continuous vs. Discrete Time

- empirically, every measurement of time is discontinuous

- decision guidance

  a) ratio of units of time intervals (days, months, years) to typical length of episodes (e.g. median, mean)

  - the smaller this ratio the more appropriate it is to use models for continuous time

  b) frequency of observational units with the same T (time until an event occurs)

  - the smaller the number of „ties" the more appropriate it is to use models for continuous time

Overview of Potential Models

we distinguish: parametric, semi-parametric and non-parametric models for event history data

- parametric and semi-parametric models: (strong) a-priori-assumptions about distribution of T and about the effects of covariates
  - parametric models: a-priori-assumption about both distribution of T and effects of covariates
  - semi-parametric models: a-priori-assumption effects of covariates
- non-parametric models: no a-priori-assumptions about distribution of T and about the effects of covariates

Overview of Potential Models

1. parametric and semi-parametric models

models have the general form:

$$h_i(t) = g\left(t, \beta_0 + \mathbf{x}_i\boldsymbol{\beta}\right)$$

a) parametric models: strong a-priori-assumptions (distribution of T, effects of covariates)

e.g. so-called proportional hazard-models

$$h_i(t) = h_0(t)\exp\left(\beta_0 + \mathbf{x}_i\boldsymbol{\beta}\right)$$

Exponential model, Weibull model, Gombertz model

Overview of Potential Models

some popular parametric models

continuous time	discontinuous time
Exponential model Weibull model Log-logistic model Log-nomal model Gombertz model Generalized Gamma model	Logistic model Complementary log-log model

Overview of Potential Models

b) semi-parametric models: a-priori-assumptions about effects of covariates, but not about distribution of T

e.g. Piecewise Constant Exponential Model

$$h_i(t) = \begin{Bmatrix} h_{01}\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right) & t \in (0, \tau_1] \\ h_{02}\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right) & t \in (\tau_1, \tau_2] \\ \vdots & \vdots \\ h_{0K}\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right) & t \in (\tau_{K-1}, \tau_K] \end{Bmatrix}$$

Overview of Potential Models

popular semi-parametric models

continuous time	discontinuous time
Piecewise Constant Exponential Model Cox Model	Piecewise Constant Logistic Model Piecewise Constant Complementary log-log Model

Overview of Potential Models

2. non-parametric models: no a-priori-Assumption about about effects of covariates or not about distribution of T

continuous time	discontinuous time
Kaplan-Meier-Estimator Nelson-Aalen-Estimator	Life Table

Data Structure and Data Management

Data Structure

intended data format: for each observational unit information on begin and end, event or cencering as well as on potential covariates should be stored

PID	t_0	t_1	Event	x_1 … x_k
1	0	1	1	2 … 78.8
2	0	3	1	1 … 54.3
3	0	7	1	1 … 89.7
4	0	10	1	2 … 66.6
5	0	15	1	1 … 44.9

Data Structure

using this data format allows to record:

a) right censoring

PID	t_0	t_1	Event	x_1 … x_k
1	0	1	1	2 … 78.8
2	0	3	1	1 … 54.3
3	0	7	1	1 … 89.7
4	0	10	1	2 … 66.6
5	0	15	0	1 … 44.9

Data Structure

using this data format allows to record:

b) left truncation (delayed entry)

PID	t_0	t_1	Event	x_1 … x_k
1	0	1	1	2 … 78.8
2	0	3	1	1 … 54.3
3	5	7	1	1 … 89.7
4	0	10	1	2 … 66.6
5	0	15	0	1 … 44.9

Data Structure

using this data format allows to record:

c) interval truncation

PID	t_0	t_1	Event	x_1 … x_k
1	0	1	1	2 … 78.8
2	0	3	1	1 … 54.3
3	5	7	1	1 … 89.7
4	0	3	0	2 … 66.6
4	7	10	1	2 … 66.6
5	0	15	0	1 … 44.9

Data Structure

using this data format allows to record:

d)time-varying covariates

PID	t_0	t_1	Event	x_1 … x_k
2	0	1	0	2 … 93.7
2	1	3	1	2 … 54.3
3	5	7	1	1 … 89.7
4	0	3	0	2 … 66.6
4	7	10	1	2 … 66.6
5	0	15	0	1 … 44.9

Data Structure

using this data format allows to record:

e)multiple events

PID	t_0	t_1	Event	x_1 … x_k
2	0	1	0	2 … 93.7
2	1	3	1	2 … 54.3
2	3	5	0	2 … 89.7
2	7	10	1	2 … 66.6
2	10	20	0	2 … 66.6
3	0	15	0	1 … 44.9

Wide vs. Long Data Format

wide format: one row per observational unit

PID	Begin	End	Event	x_1 … x_k
1	0	1	1	2 … 78.8
2	0	3	1	1 … 54.3
3	0	7	1	1 … 89.7
4	0	10	1	2 … 66.6
5	0	15	1	1 … 44.9

Wide vs. Long Data Format

wide format: one row per observational unit, even if there are multiple episodes

PID	Begin1	End1	Event1	x_{11}	…	x_{k1}	Begin2	End2	Event2	x_{12} ...
1	0	1	1	2	…	78.8	1	5	0	2
2	0	3	1	1	…	54.3	5	8	1	1
3	0	7	1	1	…	89.7	8	9	1	1
4	0	10	1	2	…	66.6	13	14	0	2
5	0	15	1	1	…	44.9	16	17	1	1

Wide vs. Long Data Format

long format: multiple rows for each oberservational unit possible

PID	Begin	End	Event	x_1 … x_k
1	0	1	1	2 … 78.8
1	1	5	0	1 … 54.3
2	0	3	1	1 … 89.7
2	5	8	1	1 … 66.6
3	0	7	1	1 … 44.9

Wide vs. Long Data Format

long format usually necessary for survival analysis (because, for example, of interval truncation, time-varying covariates, multiple events)

but also: data management much easier than in wide format

switiching from long to wide format (and vice versa) is easy in Stata (command `reshape`)

Wide vs. Long Data Format

```
. use "$home\suf97_sample.dta", clear
. keep  id_suf- job9best  einmon einjahr

. reshape long @manf @janf @mend @jend @lnoc @arve @az @std @best, i( id_suf)
j(spell_nr) string
(note: j = job1 job2 job3 job4 job5 job6 job7 job8 job9)

Data                                wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                      2797   ->   25173
Number of variables                   92   ->      21
j variable (9 values)                      ->   spell_nr
xij variables:
         job1manf job2manf ... job9manf   ->   manf
         job1janf job2janf ... job9janf   ->   janf
         job1mend job2mend ... job9mend   ->   mend
         job1jend job2jend ... job9jend   ->   jend
         job1lnoc job2lnoc ... job9lnoc   ->   lnoc
         job1arve job2arve ... job9arve   ->   arve
               job1az job2az ... job9az   ->   az
            job1std job2std ... job9std   ->   std
         job1best job2best ... job9best   ->   best
-----------------------------------------------------------------------------
```

Handling of Dates

begin and end of an episode should be stored in dates format (better interaction with Stata‘s st-commands)

conversion of existing date information into dates format:

- string-to-numeric-conversion functions, e.g. „29 5 2008“ via `gen datumsvar=date(stringvar,“DMY“)`
- date-from-numerical-components functions, e.g. `gen datumsvar=mdy(M,D,Y)`
- see *help datetime* bzw. [D] datetime for more information

Handling of Dates

```
. tab1 manf janf

-> tabulation of manf

       manf |      Freq.     Percent        Cum.
------------+-----------------------------------
 -2. filter |     17,732       70.44       70.44
   -1. k.A. |          2        0.01       70.45
     1. Jan |        920        3.65       74.10
     2. Feb |        667        2.65       76.75
    3. März |        491        1.95       78.70
     4. Apr |        662        2.63       81.33
     5. Mai |        581        2.31       83.64
    6. Juni |        443        1.76       85.40
    7. Juli |        553        2.20       87.60
     8. Aug |        688        2.73       90.33
     9. Sep |        763        3.03       93.36
    10. Okt |        838        3.33       96.69
    11. Nov |        530        2.11       98.80
    12. Dez |        303        1.20      100.00
------------+-----------------------------------
      Total |     25,173      100.00
```

Handling of Dates

janf	Freq.	Percent	Cum.
-2. Filter	17,732	70.44	70.44
-1. k. A.	2	0.01	70.45
1981	1	0.00	70.45
1986	1	0.00	70.46
1988	2	0.01	70.46
1989	1	0.00	70.47
1990	2	0.01	70.48
1991	2	0.01	70.48
1992	2	0.01	70.49
1993	4	0.02	70.51
1994	5	0.02	70.53
1995	14	0.06	70.58
1996	249	0.99	71.57
1997	2,100	8.34	79.91
1998	1,285	5.10	85.02
1999	1,066	4.23	89.25
2000	1,073	4.26	93.52
2001	870	3.46	96.97
2002	688	2.73	99.71
2003	74	0.29	100.00
Total	25,173	100.00	

Handling of Dates

```
gen begin=ym(janf,manf)
format beginn %tm

gen end=ym(jend,mend)
format ende %tm

. list  id_suf spell_nr manf janf mend jend begin end in 1/30, noobs
  +-------------------------------------------------------------------------------------------+
  | id_suf   spell_nr         manf         janf         mend         jend    beginn      ende |
  |-------------------------------------------------------------------------------------------|
  |      1          1      10. Okt         1998       9. Sep         1999   1998m10    1999m9 |
  |      1          2      10. Okt         1999      6. Juni         2000   1999m10    2000m6 |
  |      1          3      7. Juli         2000   -2. filter   -2. Filter    2000m7         . |
  |      1          4   -2. filter   -2. Filter   -2. filter   -2. Filter         .         . |
  |      1          5   -2. filter   -2. Filter   -2. filter   -2. Filter         .         . |
  |-------------------------------------------------------------------------------------------|
  |      1          6   -2. filter   -2. Filter   -2. filter   -2. Filter         .         . |
  |      1          7   -2. filter   -2. Filter   -2. filter   -2. Filter         .         . |
  |      1          8   -2. filter   -2. Filter   -2. filter   -2. Filter         .         . |
  |      1          9   -2. filter   -2. Filter   -2. filter   -2. Filter         .         . |
  |      2          1      10. Okt         1994       4. Apr         1998   1994m10    1998m4 |
  |-------------------------------------------------------------------------------------------|
  |      2          2      10. Okt         1999      7. Juli         2001   1999m10    2001m7 |
  |      2          3   -2. filter   -2. Filter   -2. filter   -2. Filter         .         . |
  ...
```

st-commands in Stata

st-commands in Stata are used to analyse survial-time data

before using other st-commands, we have to `stset` the datat

see *help st* bzw. [st] st for more information

st-commands in Stata

`stset`-command has three aims

- definition of onset of risk, definition of event(s), definition of entry and exit of subjects
- various checks if statements/data cause problems
- statements are kept for subsequent st-commands, data does not need to be stset again

stset

stset beginn, failure(firstjob=1) id(persnr) origin(time ende_studium)

means

- analysis is beginn-ende_studium, onset of risk is at t=ende_studium
- definition of event: firstjob=1
- no multiple events (single-failure data)
- multiple episodes possible (within persnr)

stset

```
. stset beginn, failure(firstjob=1) id(id_suf) origin(time ende_studium)

                id:  id_suf
     failure event:  firstjob == 1
obs. time interval:  (beginn[_n-1], beginn]
 exit on or before:  failure
    t for analysis:  (time-origin)
            origin:  time ende_studium

--------------------------------------------------------------------------
    25173  total obs.
    17734  event time missing (beginn>=.)                    PROBABLE ERROR
      103  multiple records at same instant                 PROBABLE ERROR
           (beginn[_n-1]==beginn)
      419  obs. end on or before enter()
     4527  obs. begin on or after (first) failure
--------------------------------------------------------------------------
     2390  obs. remaining, representing
     2367  subjects
     2367  failures in single failure-per-subject data
    17539  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =        74
```

stset

```
. list id_suf spell_nr end_study begin firstjob _t0 _t _d _st in 1/30, noobs

  +-------------------------------------------------------------------------+
  | id_suf   spell_nr   end_st~y     begin   firstjob   _t0   _t   _d   _st |
  |-------------------------------------------------------------------------|
  |      1       job1     1997m6   1998m10          1     0   16    1     1 |
  |      1       job2     1997m6   1999m10          0     .    .    .     0 |
  |      1       job3     1997m6    2000m7          0     .    .    .     0 |
  |      1       job4     1997m6         .          0     .    .    .     0 |
  |      1       job5     1997m6         .          0     .    .    .     0 |
  |-------------------------------------------------------------------------|
  |      1       job6     1997m6         .          0     .    .    .     0 |
  |      1       job7     1997m6         .          0     .    .    .     0 |
  |      1       job8     1997m6         .          0     .    .    .     0 |
  |      1       job9     1997m6         .          0     .    .    .     0 |
  |      2       job1     1997m8   1994m10          1     .    .    .     0 |
  |-------------------------------------------------------------------------|
  |      2       job2     1997m8   1999m10          0     .    .    .     0 |
  |      2       job3     1997m8         .          0     .    .    .     0 |
  |      2       job4     1997m8         .          0     .    .    .     0 |
  |      2       job5     1997m8         .          0     .    .    .     0 |
  |      2       job6     1997m8         .          0     .    .    .     0 |
  |-------------------------------------------------------------------------|
  ...
```

stset

after having stset the data, we can use st-commands, for example

- stdes: description of survival-time data
- stvary: information on variables

stdes

```
. stdes

        failure _d:  ereignis == 1
  analysis time _t:  (beginn-origin)
            origin:  time ende_studium
                id:  id_suf

                                |-------------- per subject --------------|
Category                  total        mean         min     median         max
-------------------------------------------------------------------------------
no. of subjects            2367
no. of records             2390    1.009717           1          1           2

(first) entry time                 .0105619           0          0          25
(final) exit time                  7.420363           1          4          74

subjects with gap             0
time on gap if gap            0           .           .          .           .
time at risk              17539    7.409801           1          4          74

failures                   2367           1           1          1           1
-------------------------------------------------------------------------------
```

stvary

```
. stvary  geschl arve

         failure _d:  ereignis == 1
   analysis time _t:  (beginn-origin)
             origin:  time ende_studium
                 id:  id_suf

               subjects for whom the variable is
                                                    never    always   sometimes
    variable |  constant    varying             missing   missing   missing
-------------+-----------------------------------------------------------------
      geschl |      2367          0                2367         0         0
        arve |      2344         23                2367         0         0
```

Some Examples using Different Models

Non-parametric Approaches

- no a-priori-assumptions about distribution of T and about the effects of covariates
- serve mainly descriptive purposes
  - estimation of survivor function: Kaplan-Meier-Estimator
  - estimation of cumulative hazard function: Nelson-Aalen-Estimator
  - for both estimators: illustration of group differences
- but also: statistical tests of group differences

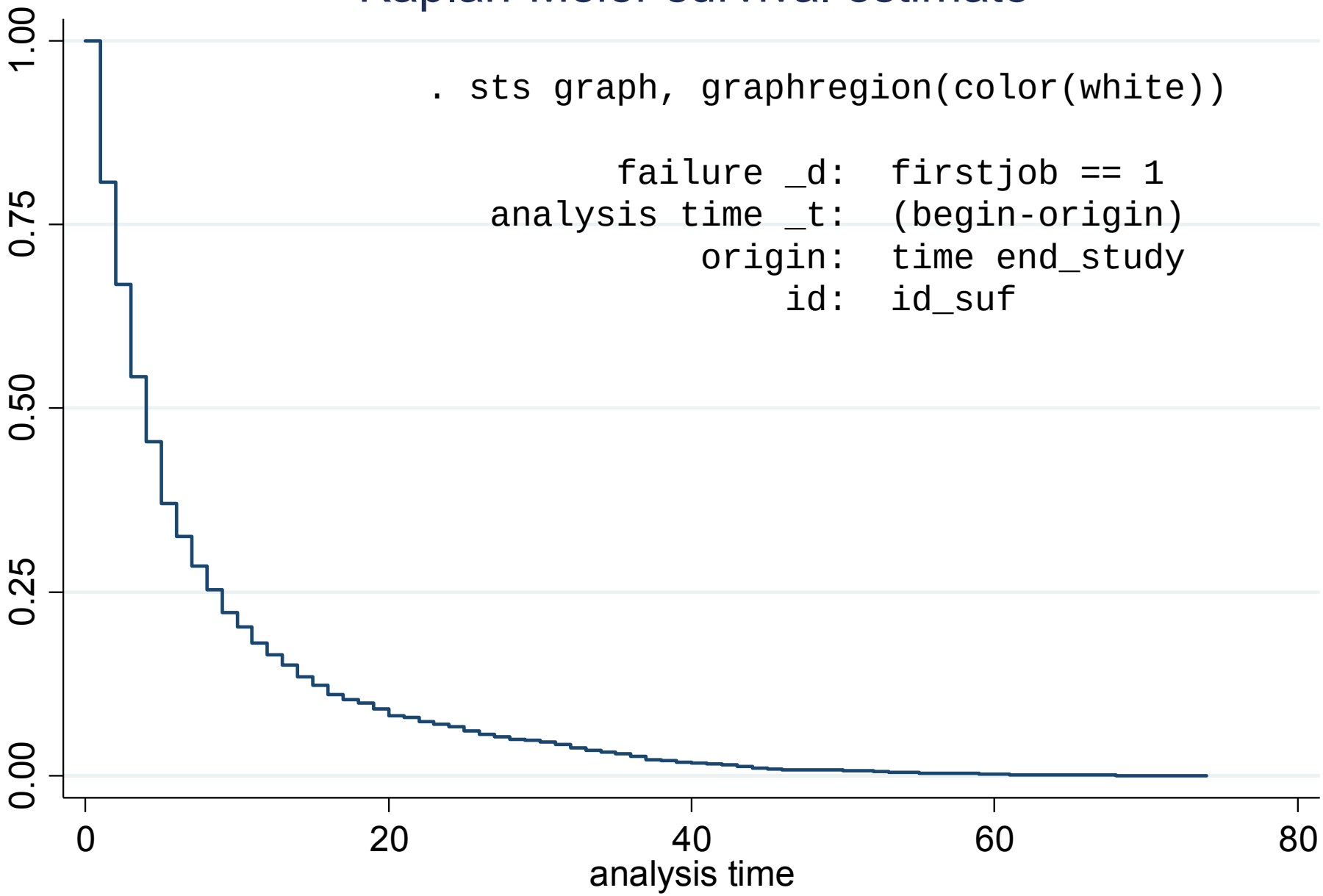
Non-parametric Approaches

```
. sts list

         failure _d:  firstjob == 1
   analysis time _t:  (begin-origin)
             origin:  time end_study
                 id:  id_suf

          Beg.          Net            Survivor      Std.
  Time    Total   Fail   Lost            Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
     1     2366    454      0              0.8081    0.0081     0.7917    0.8234
     2     1912    329      0              0.6691    0.0097     0.6497    0.6876
     3     1583    298      0              0.5431    0.0102     0.5228    0.5629
     4     1285    209      0              0.4548    0.0102     0.4346    0.4747
     5     1076    200      0              0.3702    0.0099     0.3508    0.3897
     6      876    105      0              0.3259    0.0096     0.3070    0.3448
     7      771     97      0              0.2849    0.0093     0.2668    0.3032
     8      674     74      0              0.2536    0.0089     0.2362    0.2713
     9      600     75      0              0.2219    0.0085     0.2054    0.2388
    10      525     45      0              0.2029    0.0083     0.1869    0.2193
    11      480     54      0              0.1801    0.0079     0.1649    0.1958
    12      426     38      0              0.1640    0.0076     0.1494    0.1792
...
```

Kaplan-Meier survival estimate



Non-parametric Approaches

Question: Does survivor function dffer between groups?

1. description: illustration of group differences
2. statistical tests of group differences

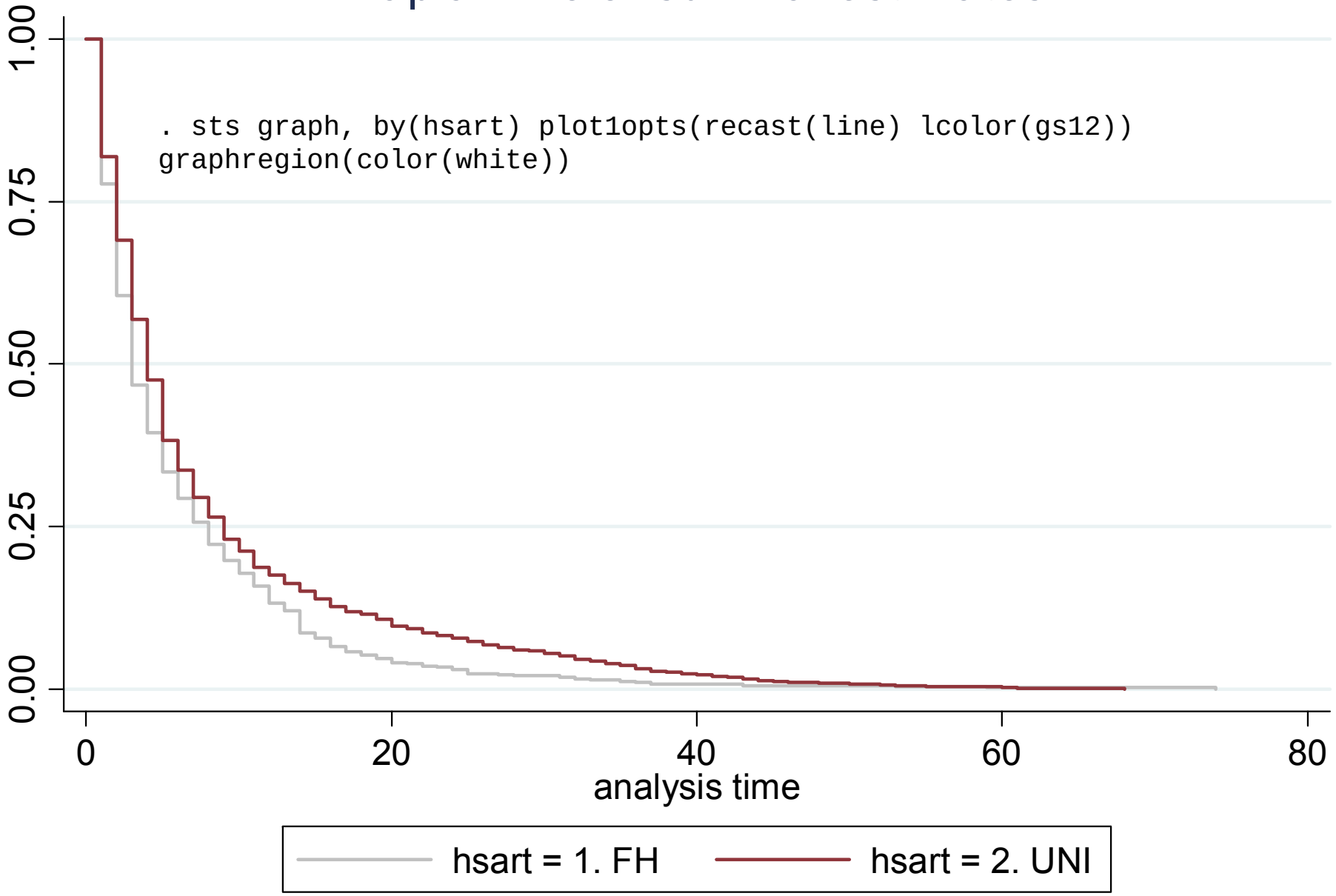
Non-parametric Approaches

1. description: illustration of group differences

sts graph, by(group)

- all Stata graph twoway-options can be used
- see *help sts graph* and [ST] sts graph for more options

Kaplan-Meier survival estimates



Non-parametric Approaches

2. statistical tests of group differences

basic idea: compare number of observed events with number of events to be expected unter H_0

- H_0:$h_1(t)=h_2(t)=\ldots=h_r(t)$
- calculation of χ^2–distributed test statistic

Non-parametric Approaches

Stata-Beispiel 7b: sts test

```
. sts test hsart
Log-rank test for equality of survivor functions

       |   Events         Events
hsart  |  observed       expected
-------+-------------------------
1. FH  |       601         521.06
2. UNI |      1766        1845.94
-------+-------------------------
Total  |      2367        2367.00

            chi2(1) =      18.82
            Pr>chi2 =     0.0000
```

Non-parametric Approaches

tests of group differences can be stratified to account for the fact that risk might also differ between other groups

sts test varname, strata(varlist)

- reduces risk of confounding
- option `detail` to display test results for single strata

```
. replace geschl=. if geschl<1
(36 real changes made, 36 to missing)

. sts test hsart, strata(geschl) detail
Stratified log-rank test for equality of survivor functions

-> geschl = 1
       |   Events         Events
hsart  |  observed       expected
-------+-------------------------
1. FH  |       402         353.33
2. UNI |       916         964.67
-------+-------------------------
Total  |      1318        1318.00

            chi2(1) =      11.32
            Pr>chi2 =     0.0008
-> Total
       |   Events         Events
hsart  |  observed    expected(*)
-------+-------------------------
1. FH  |       600         529.44
2. UNI |      1764        1834.56
-------+-------------------------
Total  |      2364        2364.00

(*) sum over calculations within geschl
            chi2(1) =      14.88
            Pr>chi2 =     0.0001

-> geschl = 2
       |   Events         Events
hsart  |  observed       expected
-------+-------------------------
1. FH  |       198         176.11
2. UNI |       848         869.89
-------+-------------------------
Total  |      1046        1046.00

            chi2(1) =       3.82
            Pr>chi2 =     0.0506
```

Semi-parametric Models

- a-priori-assumptions about effects of covariates, but not about distribution of T
- formal model:

$$h\left(t\middle|\mathbf{x}_i\right) = h_0\left(t\right)\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right)$$

$h\left(t\middle|\mathbf{x}_i\right)$ hazard at time t, given $\mathbf{x}_i$

$h_0\left(t\right)$ baseline hazard

$$\mathbf{x}_i\boldsymbol{\beta} = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K$$

- baseline hazard does not need to be specified

Semi-parametric Models

- advantage: no assumption about $h_0(t)$, therefore less error-prone
- disadvantage: not efficient compared to (correctly specified) parametric model
- In Stata: `stcox`

```
. stcox ib2.hsart ib2.geschl

         failure _d:  firstjob == 1
   analysis time _t:  (begin-origin)
             origin:  time end_study
                 id:  id_suf

Iteration 0:   log likelihood = -16189.788
Iteration 1:   log likelihood =  -16178.27
Iteration 2:   log likelihood =  -16178.25
Iteration 3:   log likelihood =  -16178.25
Refining estimates:
Iteration 0:   log likelihood =  -16178.25

Cox regression -- Breslow method for ties

No. of subjects =        2,364                 Number of obs    =       2,387
No. of failures =        2,364
Time at risk    =        17533
                                               LR chi2(2)       =       23.08
Log likelihood  =    -16178.25                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsart |
      1. FH  |   1.189472   .0569779     3.62   0.000      1.08288    1.306557
      geschl |
1. männlich  |   1.122549   .0469068     2.77   0.006     1.034278    1.218355
------------------------------------------------------------------------------
```

Parametric Models

- a-priori-assumptions about effects of covariates and about distribution of T

- fully parameterized model
- two variants of parametrization:
  - proportional hazard modelle (PH-models)
  - accelerated failure time-models (AFT-models)

Parametric Models

PH-models

- formal model:

$$h\left(t\middle|\mathbf{x}_i\right) = h_0\left(t\right)\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right)$$

$h\left(t\middle|\mathbf{x}_i\right)$ hazard at time t, given $\mathbf{x}_i$

$h_0\left(t\right)$ baseline hazard, needs to be specified

z.B. $h_0\left(t\right) = \exp\left(\beta_0\right)$ (Exponential Model)

- baseline hazard cannot be left unspecified

Parametric Models

AFT-models

- formal model:

$$\ln\left(t_i\right) = \mathbf{x}_i\boldsymbol{\beta} + \ln\left(\tau_i\right)$$

$$\mathbf{x}_i\boldsymbol{\beta} = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K$$

τ follows a certain distribution

e.g. $\tau \sim \text{Weibull}\left(\beta_0, p\right)$

```
. streg ib2.hsart ib2.geschl, distribution(gompertz)

         failure _d:  firstjob == 1
   analysis time _t:  (begin-origin)
             origin:  time end_study
                 id:  id_suf

Fitting constant-only model:

Iteration 0:   log likelihood = -3700.3899
Iteration 1:   log likelihood = -3640.7284
...
Gompertz regression -- log relative-hazard form

No. of subjects =        2,364                 Number of obs    =       2,387
No. of failures =        2,364
Time at risk    =        17533
                                               LR chi2(2)       =       28.09
Log likelihood  =   -3625.6474                 Prob > chi2      =      0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsart |
      1. FH  |   1.217352   .0582233     4.11   0.000     1.108422    1.336987
      geschl |
1. männlich  |   1.129788   .0471284     2.93   0.003     1.041093    1.226039
        cons |    .147257   .0055482   -50.84   0.000     .1367745    .1585428
-------------+----------------------------------------------------------------
      /gamma |  -.0234936    .002415    -9.73   0.000    -.0282269   -.0187604
------------------------------------------------------------------------------
```

Outlook

some important things that we did not talk about

- interactions between covariates
- time-varying effects of covariates
- multiple events
  - recurrent events
  - competing risks
- diagnostics for (semi-)parametric models