Applied population analysis, week 4

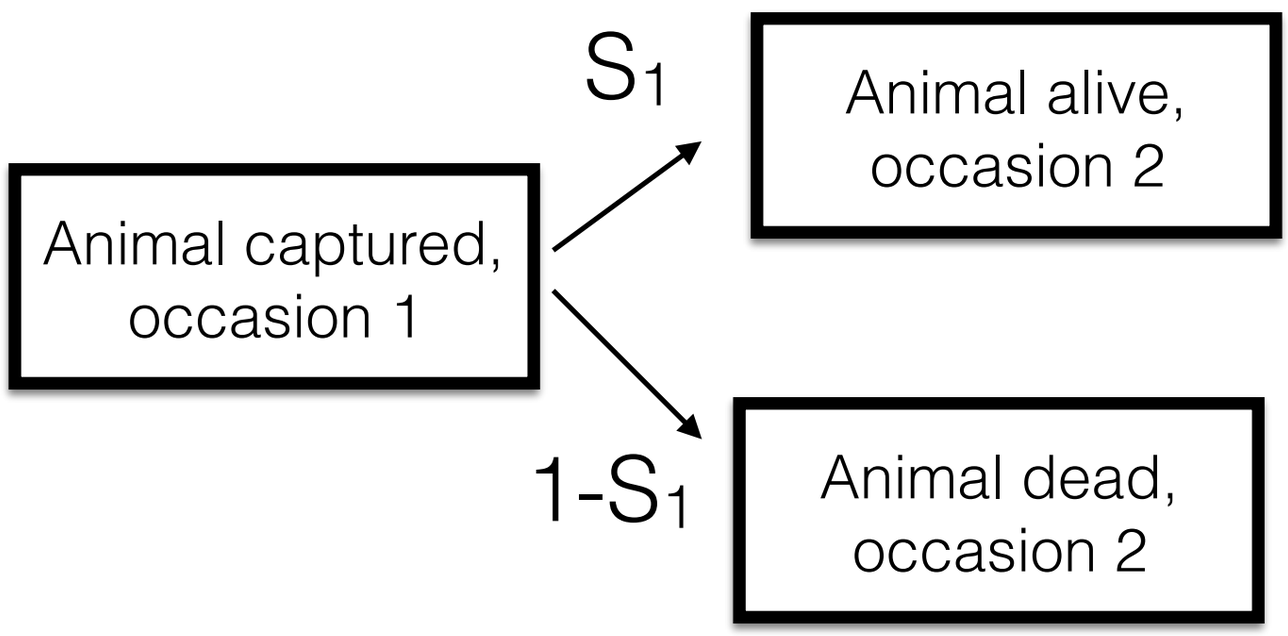
WLF 504, Fall 2019

Today:

- Go over basic bootstrapping challenge, and extension “prediction with covariates”
- More info on bootstrapping
- Basic intro to survival curves, distributions, and analyses
- This week’s coding challenge: bootstrap blackduck survival, 3 versions.

Known fates data

- When we know the fates of all our animals…
- We can eliminate the detection probability (p), woo! The probabilities are so much simpler.



Known fates data

Type of model	Encounter history (CH)	Probability
Known Fates	10 10 10 10	$S_1S_2S_3S_4$
	10 10 11 00	$S_1S_2(1\text{-}S_3)$
	10 11 00 00	$S_1(1–S_2)$
	11 00 00 00	$(1–S_1)$
	10 00 00 10	$S_1S_2S_3S_4$
	00 00 10 11	$S_3(1–S_4)$
	10 00 00 00	S_1 ←

Tagged occasion 1, unknown fate, **censored**

How to calculate variance of the product of survivals?

- We often estimate survival (daily, weekly) in intervals. What if we want to know the mean and the variance of some longer time period, say a month or a year?
- We can multiply the survival rates together:
- And we can determine the variance…
- But what do we mean by “variance?”

Method 2: Bootstrapping

- Re-sample our data (individuals) with replacement
- If we have groups, we must re-sample within groups so that number in each group remains the same
- Coding challenge this week: We will do very simple bootstrapping functions “manually” and with the “boot” function in program R

Predictions using covariate models

- Extension assignment to the coding challenge for last week, for those that get through bootstrapping challenge easily/quickly

- Using the observed range of a covariate, what Si do we predict? CI's?

What's a bootstrap?

- Monte Carlo simulation = repeatedly creating/sampling random data in some way to estimate something… very broad category!

- Technically, a bootstrap is a special type of Monte Carlo simulation, used to estimate some characteristics of the sampling distribution (for example, mean and confidence limits). They are often referred to as seperate things though.

- Parametric bootstrap- resampling using our modeled distribution (mean and standard error, w/ parameter distribution specified).

- Non-parametric bootstrap- resampling our individual animals, no disribution specified.

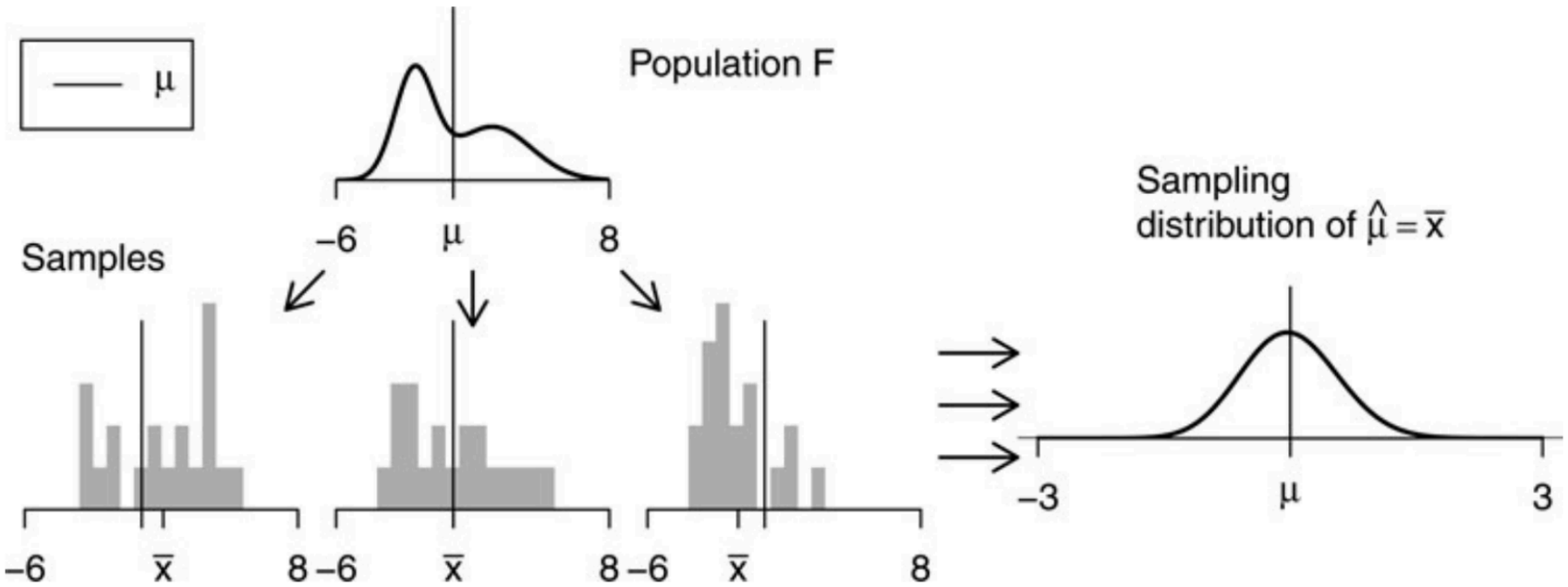


Figure 4. Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution.

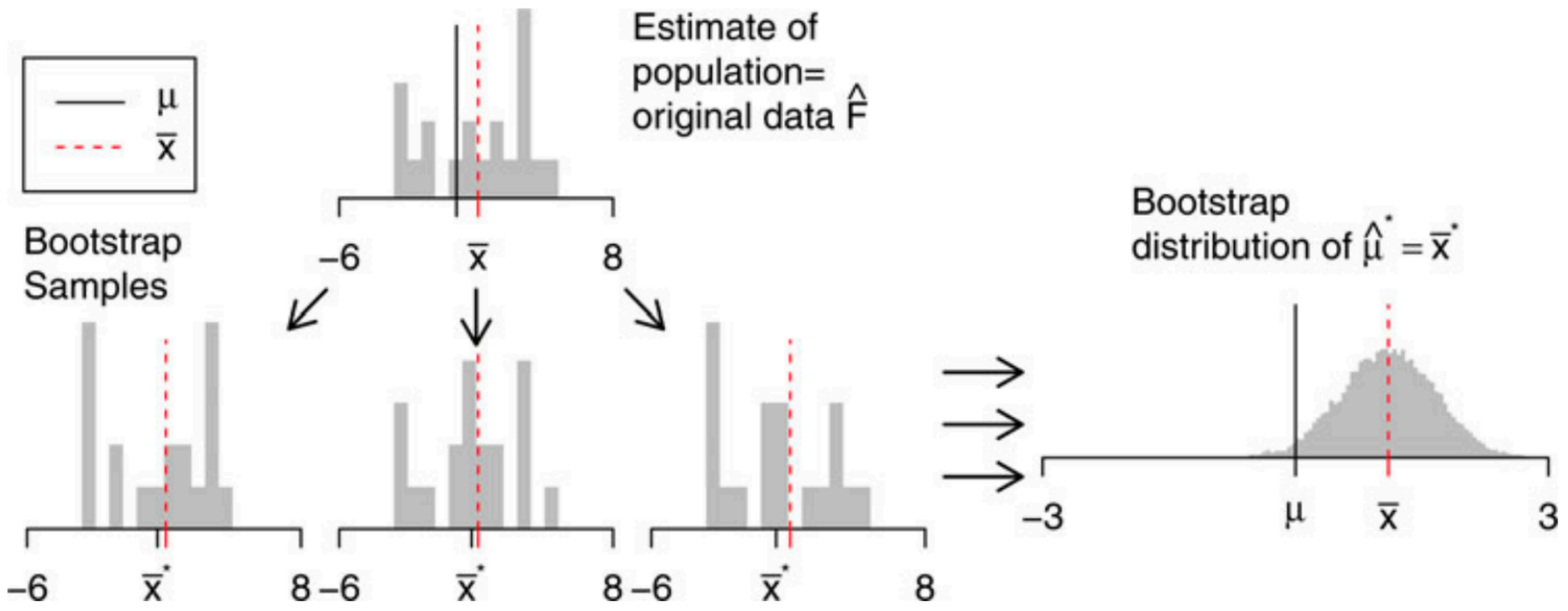


Figure 5. Bootstrap world. The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{x}$), not the parameter (μ).

Non-parametric bootstrap

	Real World	Bootstrap World
true distribution	F	$\widehat{F}_n$
data	$X_1, \ldots, X_n$ IID F	$X_1^*, \ldots, X_n^*$ IID $\widehat{F}_n$
empirical distribution	$\widehat{F}_n$	F_n^*
parameter	$\theta = t(F)$	$\widehat{\theta}_n = t(\widehat{F}_n)$
estimator	$\widehat{\theta}_n = t(\widehat{F}_n)$	$\theta_n^* = t(F_n^*)$

Real world: data is collected from a true distribution
Bootstrap world: data is collected from an empirical distribution (our sampled data).
Example: our survival data is a sample from the population's survival function, and is also the sample we "collect data" from to bootstrap.

Non-parametric bootstrap

	Real World	Bootstrap World
true distribution	F	$\widehat{F}_n$
data	$X_1, \ldots, X_n$ IID F	$X_1^*, \ldots, X_n^*$ IID $\widehat{F}_n$
empirical distribution	$\widehat{F}_n$	F_n^*
parameter	$\theta = t(F)$	$\widehat{\theta}_n = t(\widehat{F}_n)$
estimator	$\widehat{\theta}_n = t(\widehat{F}_n)$	$\theta_n^* = t(F_n^*)$

Real world: the data is a sample from the true distribution
Bootstrap world: we re-sample the empirical distribution (our data) with replacement.
Example: We sample n individuals from the pop. We bootstrap by sampling n individuals from our original sample of n, with replacement.

Non-parametric bootstrap

	Real World	Bootstrap World
true distribution	F	$\widehat{F}_n$
data	$X_1, \ldots, X_n$ IID F	$X_1^*, \ldots, X_n^*$ IID $\widehat{F}_n$
empirical distribution	$\widehat{F}_n$	F_n^*
parameter	$\theta = t(F)$	$\widehat{\theta}_n = t(\widehat{F}_n)$
estimator	$\widehat{\theta}_n = t(\widehat{F}_n)$	$\theta_n^* = t(F_n^*)$

Real world: We now have an empirical distribution by sampling, dependent on our sampling size
Bootstrap world: we now have an empirical distribution of the original empirical distribution!
Example: a collection of survival data sets we simulated w/ resampling

Non-parametric bootstrap

	Real World	Bootstrap World
true distribution	F	$\widehat{F}_n$
data	$X_1, \ldots, X_n$ IID F	$X_1^*, \ldots, X_n^*$ IID $\widehat{F}_n$
empirical distribution	$\widehat{F}_n$	F_n^*
parameter	$\theta = t(F)$	$\widehat{\theta}_n = t(\widehat{F}_n)$
estimator	$\widehat{\theta}_n = t(\widehat{F}_n)$	$\theta_n^* = t(F_n^*)$

Real world: There is some true value of a parameter, theta, that depends on the true distribution
Bootstrap world: There is an estimated parameter, theta-hat, that depends on the empirical (original sample) distribution
Example: In the real world, we want to learn about St, true cumulative survival. In the bootstrap world we want to learn about S-hat t, estimated cumulative survival.

Non-parametric bootstrap

	Real World	Bootstrap World
true distribution	F	$\widehat{F}_n$
data	$X_1, \ldots, X_n$ IID F	$X_1^*, \ldots, X_n^*$ IID $\widehat{F}_n$
empirical distribution	$\widehat{F}_n$	F_n^*
parameter	$\theta = t(F)$	$\widehat{\theta}_n = t(\widehat{F}_n)$
estimator	$\widehat{\theta}_n = t(\widehat{F}_n)$	$\theta_n^* = t(F_n^*)$

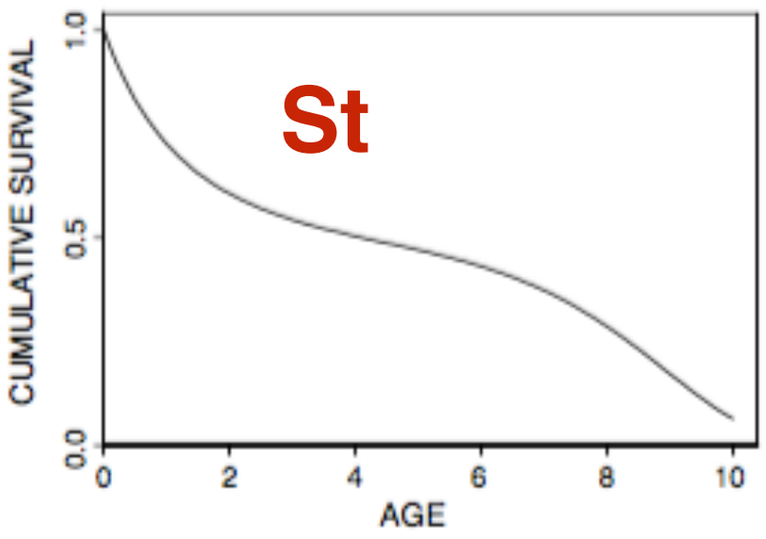
Real world: We estimate a value of a parameter, theta-hat, that depends on the sampled distribution.

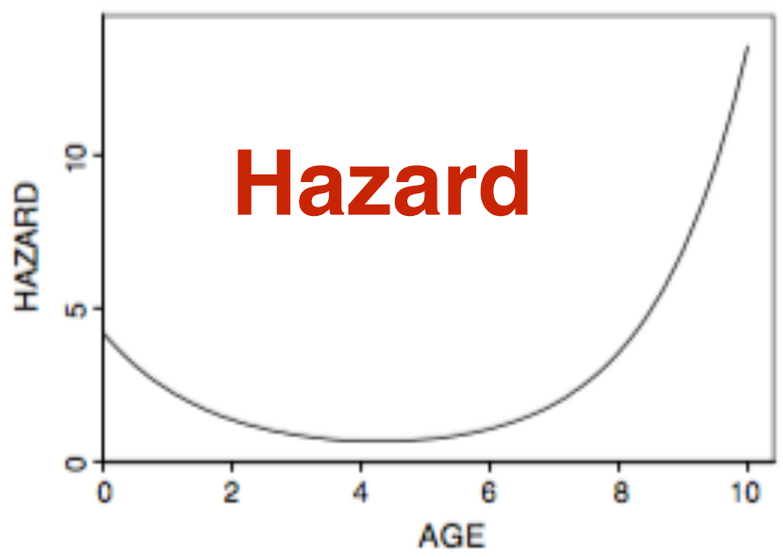
Bootstrap world: There is an estimated parameter, bootstrapped theta, that depends on the distribution of theta-hats.

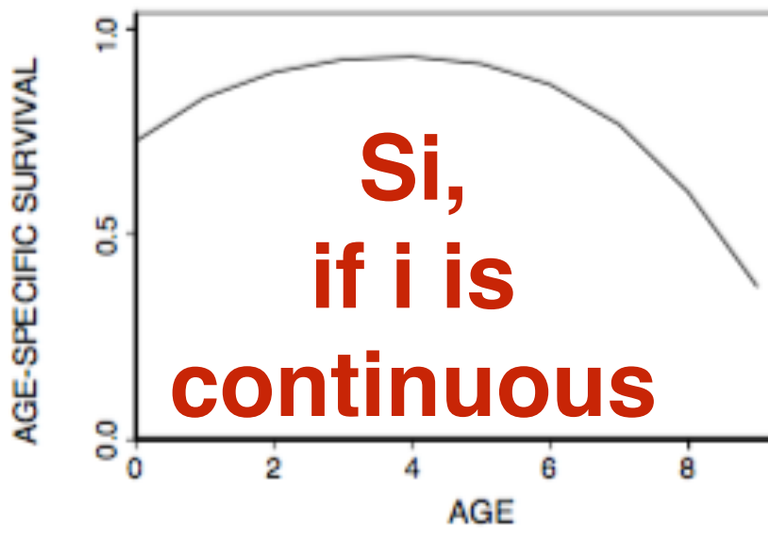
Example: In the real world, we estimate S-hat t, estimated cumulative survival. Bootstrapping, we estimate St bootstrap, a collection of S-hat t estimates

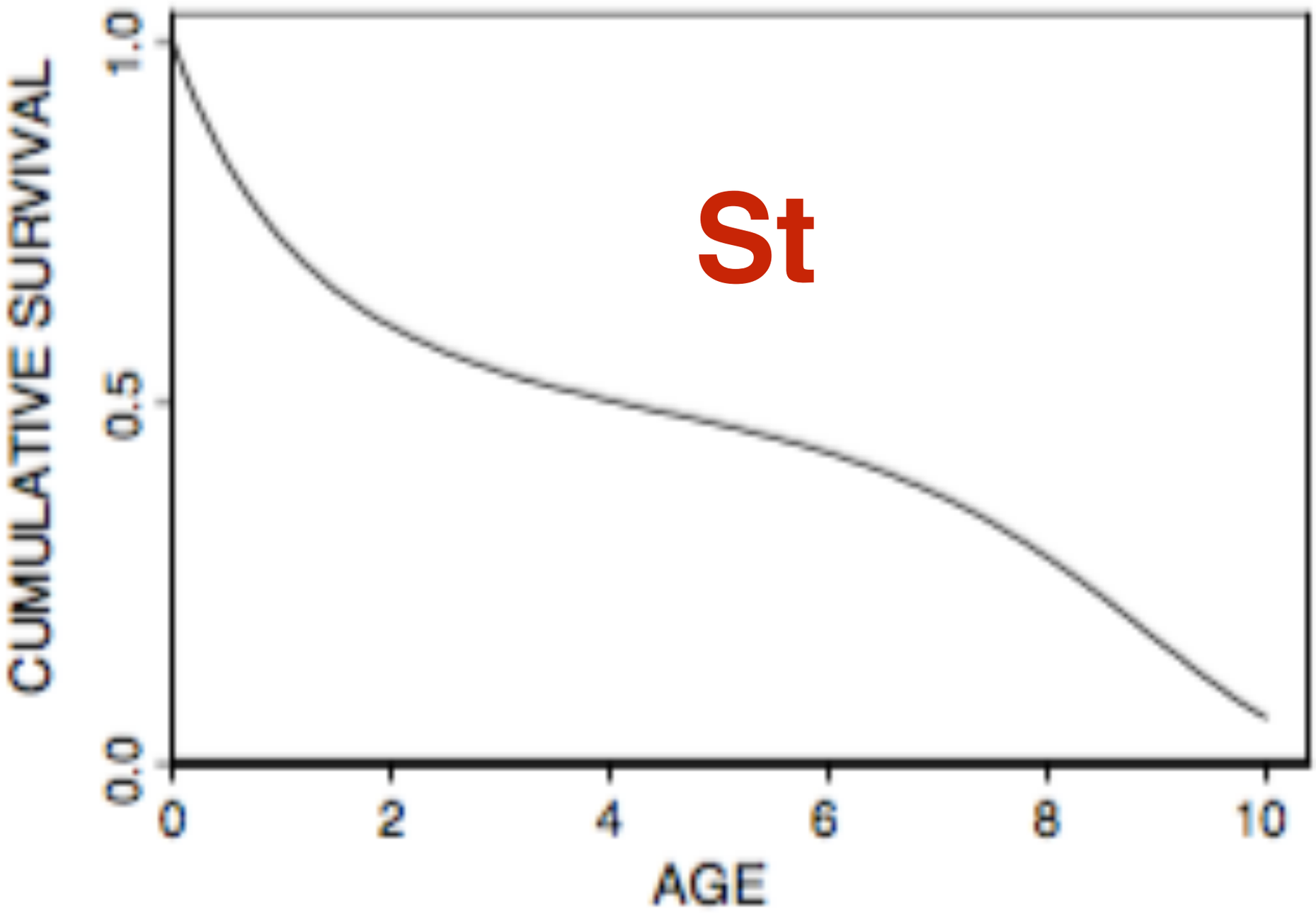
Time-to-event models

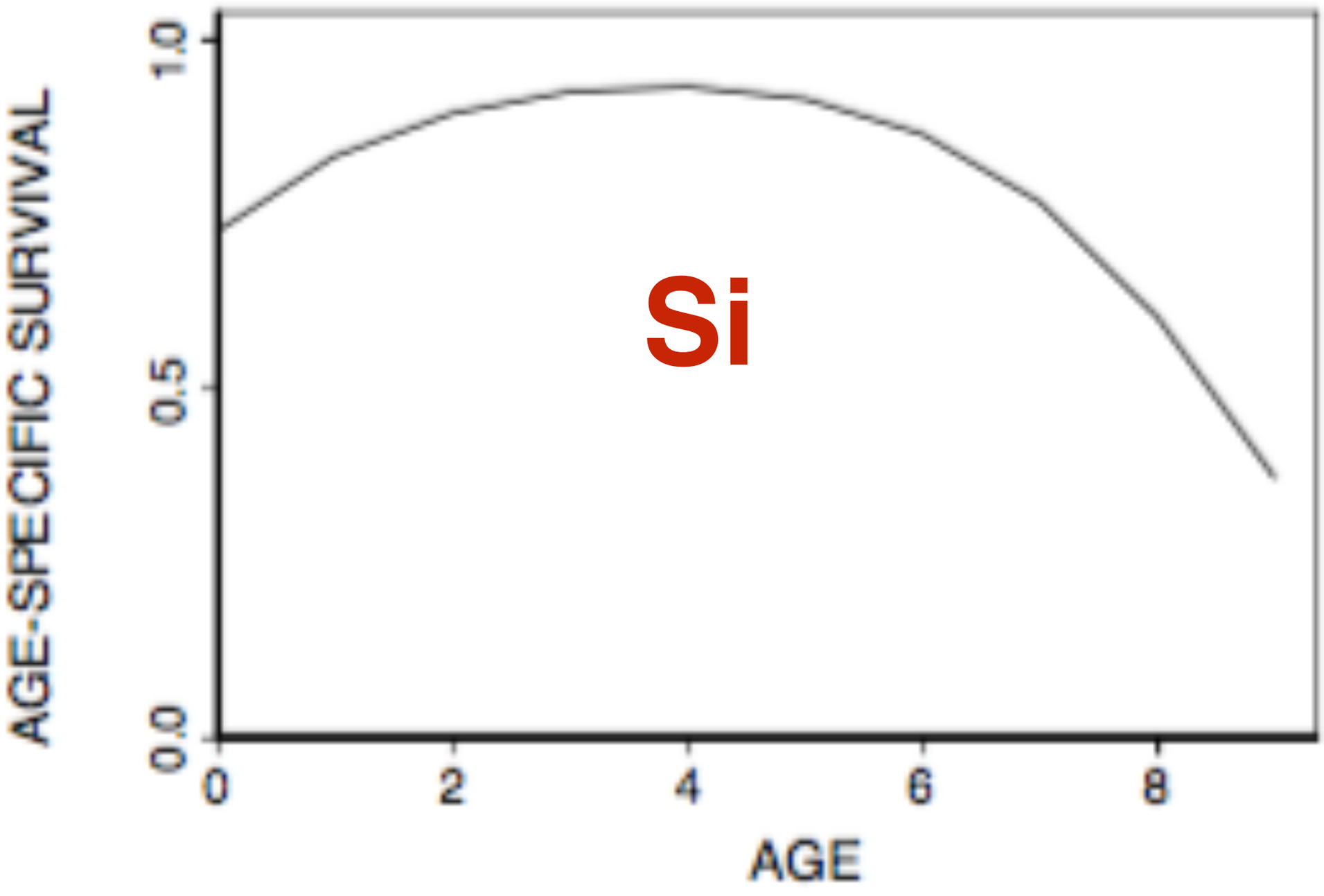
- Instead of thinking about survival as a series of binomial outcomes, Si, that cumulatively produce St…
- We can instead think about survival as the inverse of failure… and failures happen at different times
- —>distribution of failure times and a “hazard rate”



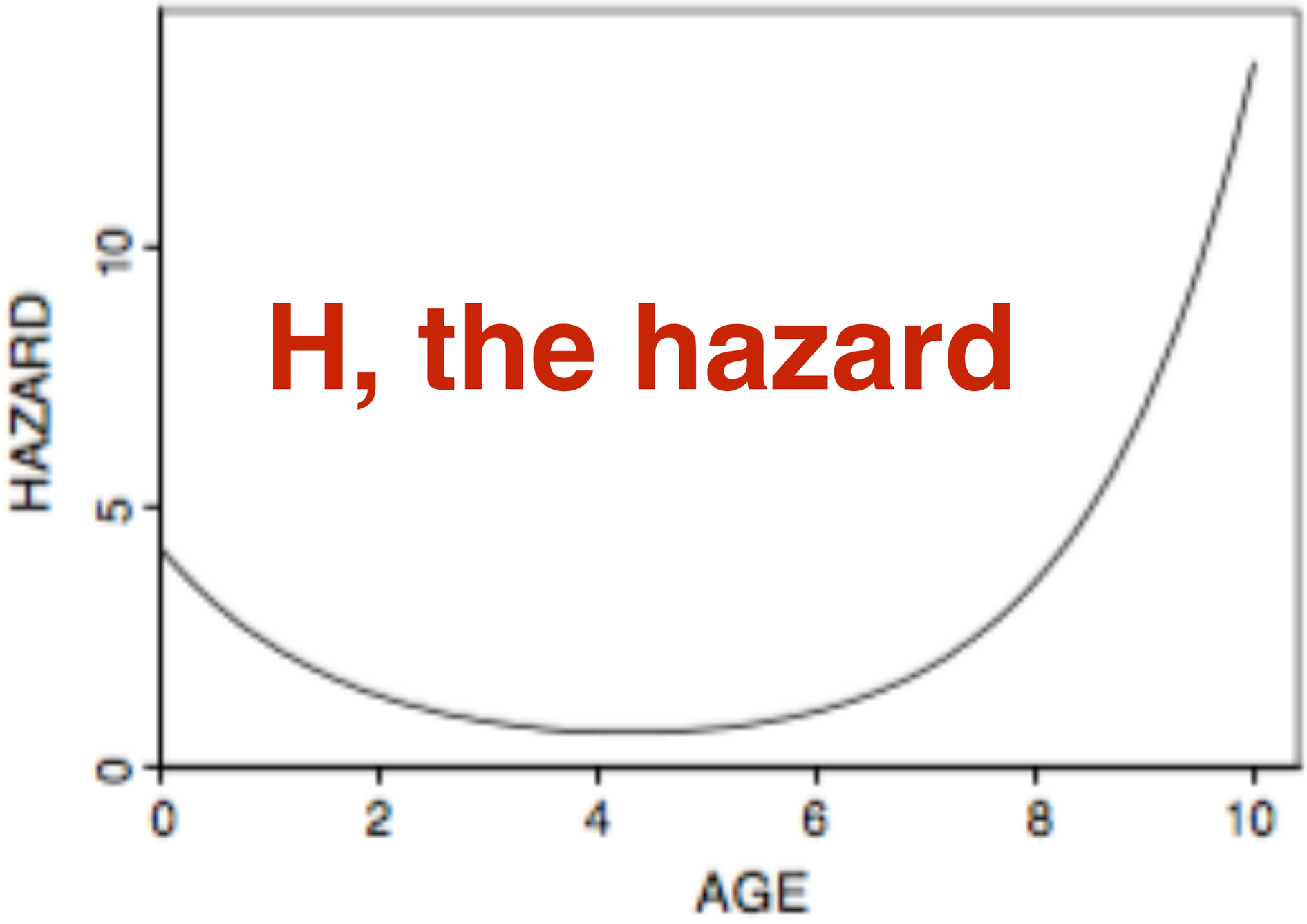





St
CUMULATIVE SURVIVAL
1.0
0.5
0.0
0
2
4
6
8
10
AGE


AGE-SPECIFIC SURVIVAL
1.0
0.5
0.0
0
2
4
6
8
AGE
Si

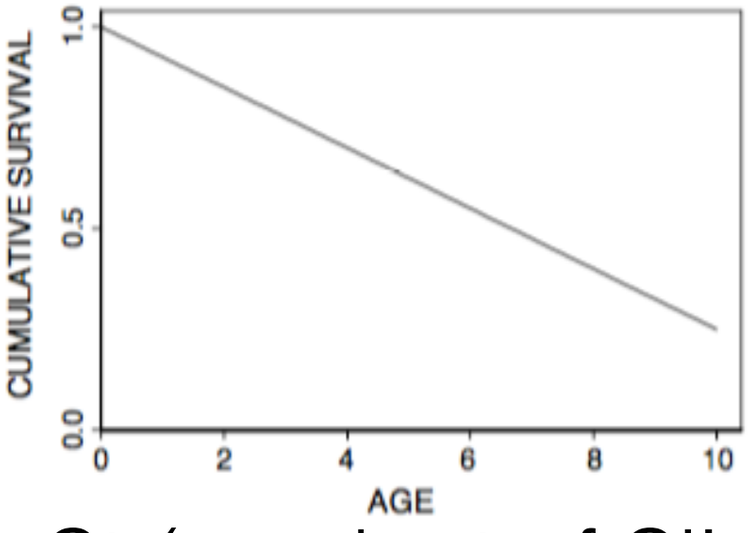
H, the hazard



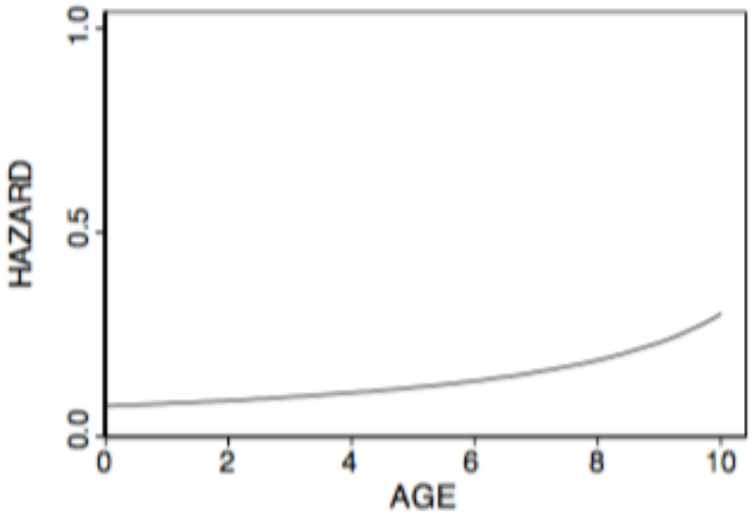
Different distributions of survival times describing life history…

To maintain a constant cumulative survival, the % dying in each age class becomes > each yr —> decreasing age-specific survival probability over time.

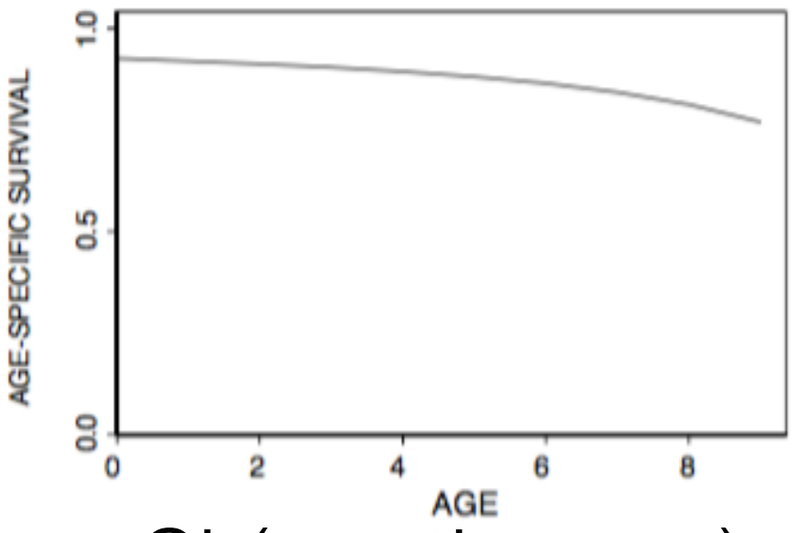
Uniform distribution



St (product of Si's)



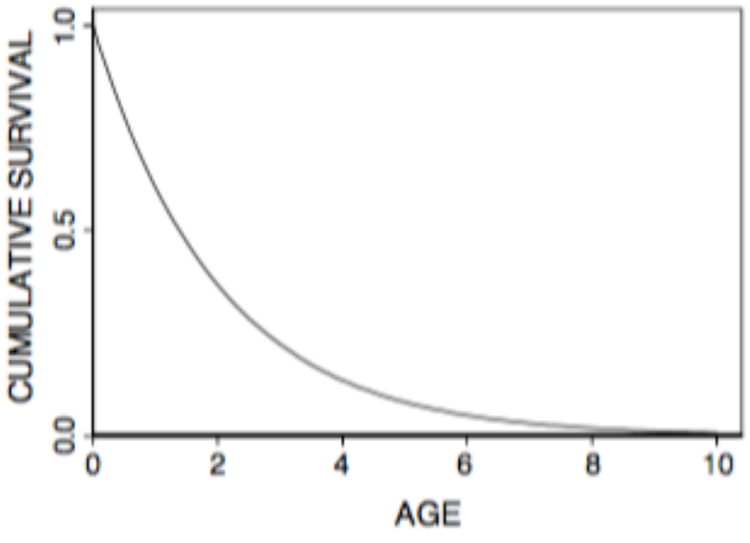
Hazard



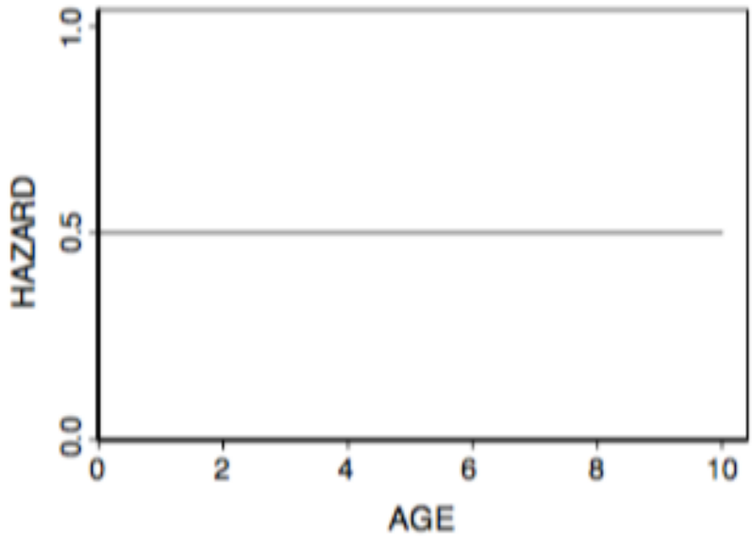
Si (continuous)

No matter how long an individual has lived, the probability of dying within any interval remains constant

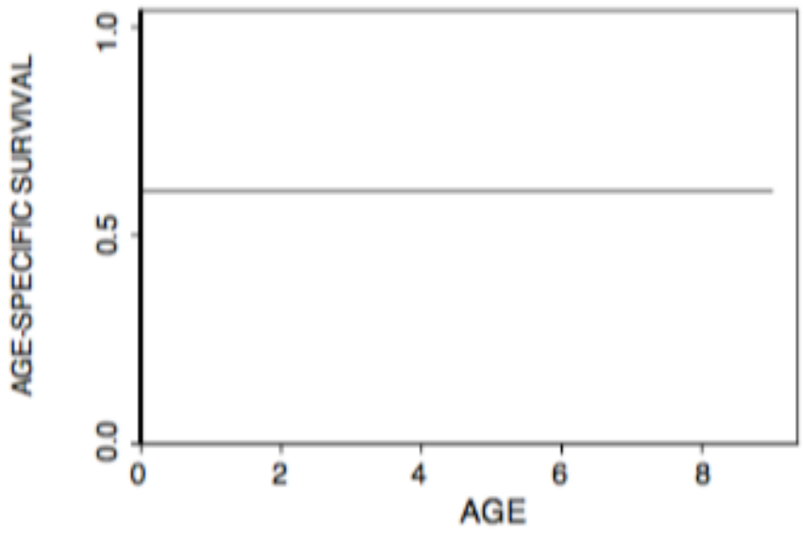
Exponential distribution



St (product of Si's)



Hazard

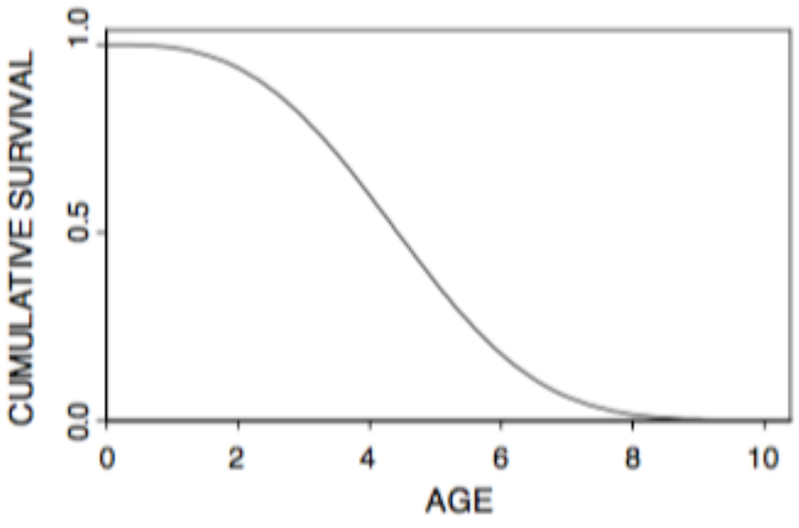


Si (continuous)

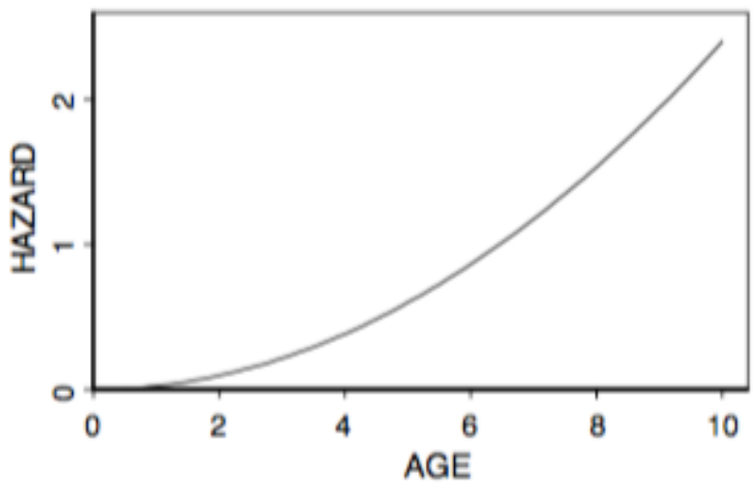
More flexible form of exponential, w/2 shape parameters.

The hazard function increases over the lifetime, with a corresponding decrease in age-specific survival.

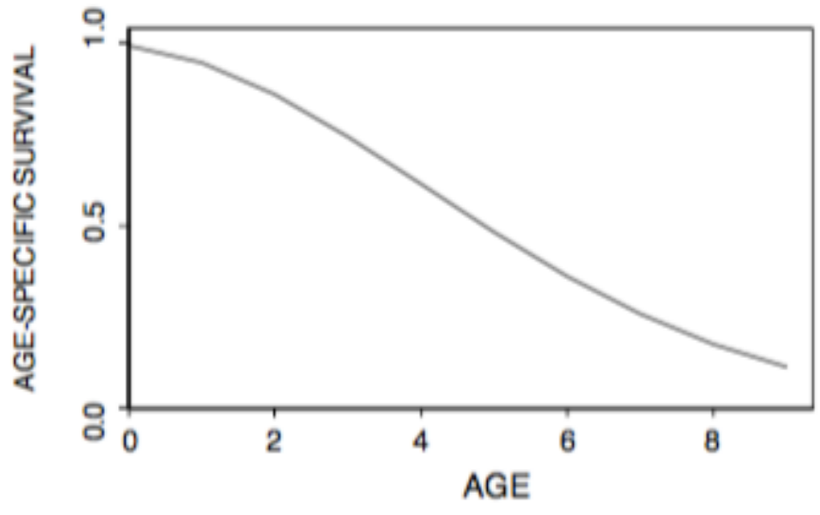
Weibull distribution



St (product of Si's)



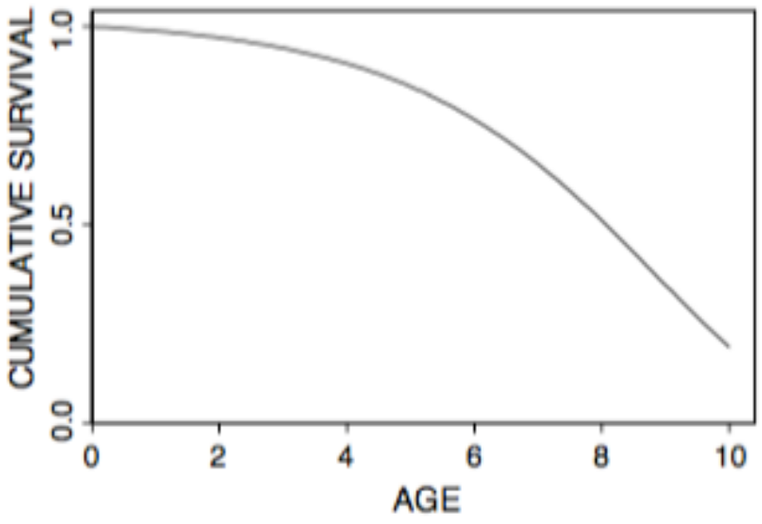
Hazard



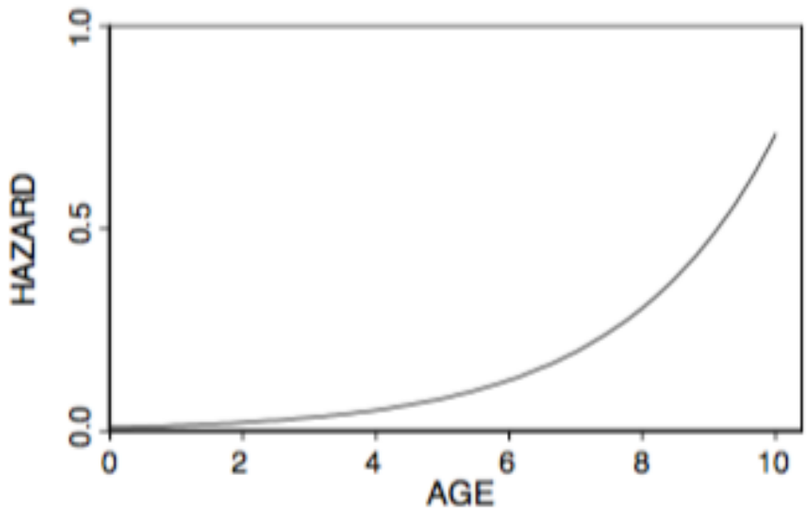
Si (continuous)

a more rapid decrease in survival later in life that does not level off

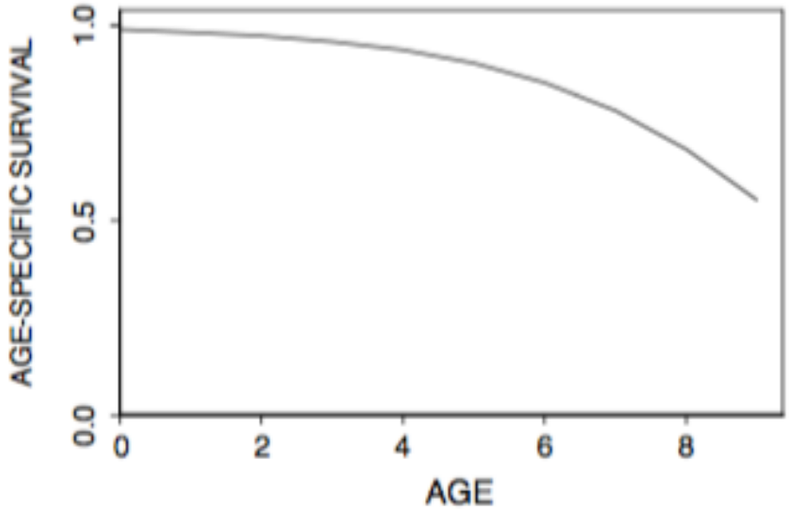
Gompertz distribution



St (product of Si's)

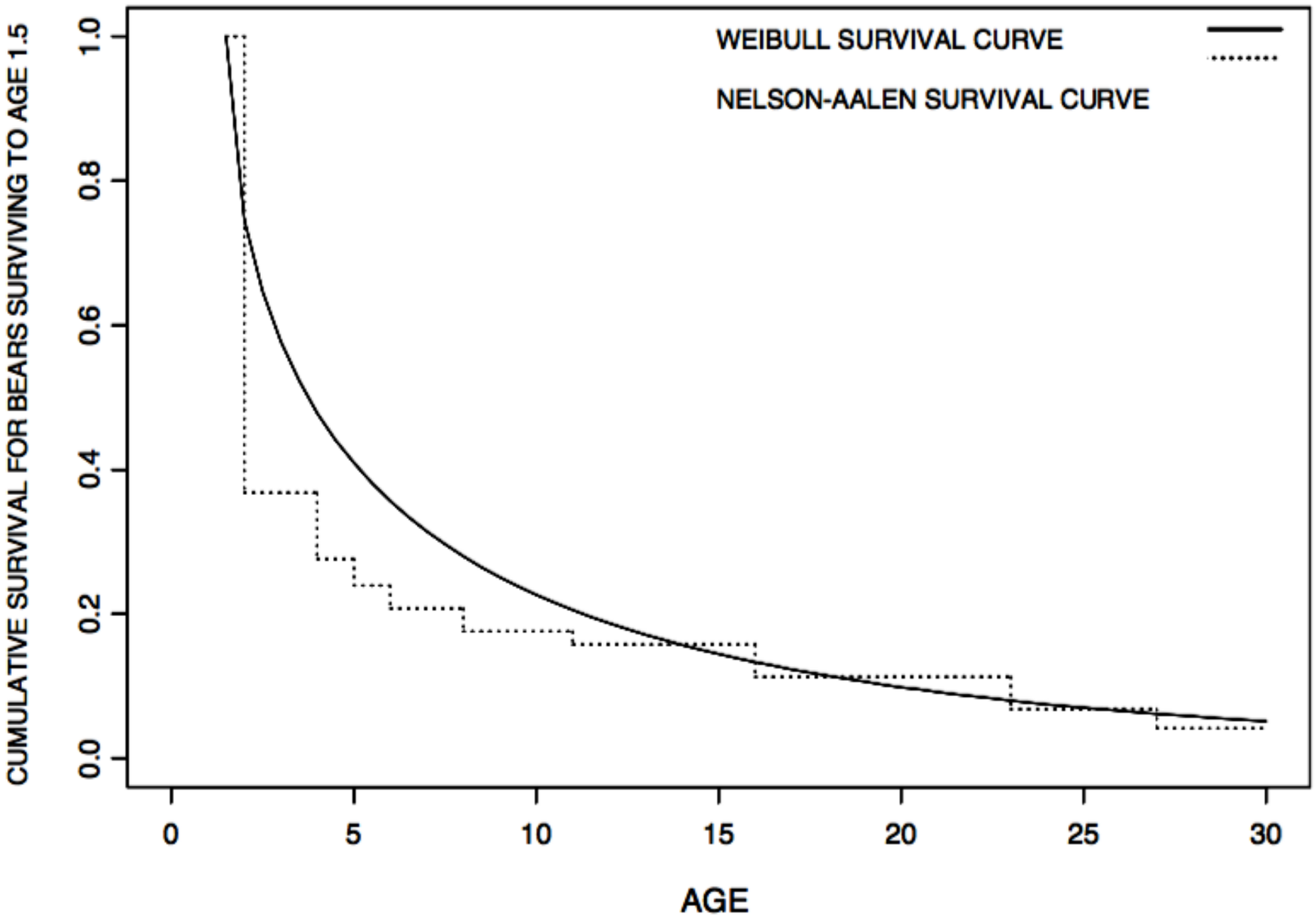


Ht



Si (continuous)

Parametric or non-parametric survival curves?



Parametric or non-parametric survival curves?

Non-parametric and semi-parametric options, if you fail to meet parametric distr. assumptions:

- Kaplan-Meier or Nelson-Alen estimator of S(t)
- Cox proportional hazards regression: KM survival curve is response variable, regression with covariates/ predictors.
- Cause specific mortality regressions: Cumulative incidence funcitions.
- I can provide readings to those who are interested.

Coding challenge