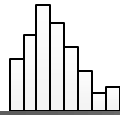


# Discrete-time survival analysis using latent variables



Presented by  
Katherine E. Masyn, Ph.D.\*

Johns Hopkins; UCLA

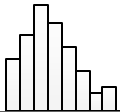
kmasyn@ucla.edu

**SAMSI 2004-05 Program on Latent Variables in  
the Social Sciences Kickoff Workshops**

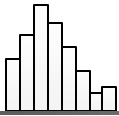
**September 12, 2004**

\*Postdoctoral fellow with Johns Hopkins SPH, Department of Mental Health  
Supported by NIMH Grant T32-MH018834

## Time-to-event data



A record of *when* an event occurs  
(relative to some “beginning”) for each  
individual in a sample, e.g., time of  
death, grade of school dropout, age of  
first alcohol use in school-aged  
children, etc.



## Time scales

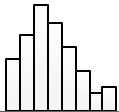
---

### ✦ *Continuous*

The “exact” time of an event for each subject is known, e.g., time of death

### ✦ *Discrete*

- 1) The timing of an event is continuous but is only recorded for an *interval* of time, e.g., grade of school dropout.
- 2) The timing of an event is itself discrete, e.g., grade retention.

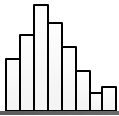


## Survival probability

---

Let  $T$  be the time interval of the event where  $T \in \{1, 2, \dots, J\}$

$S(j)$ , called the *survival probability*, is defined as the probability of “surviving” *beyond* time interval  $j$ , i.e., the probability that the event occurs after interval  $j$ :  $S(j) = P(T > j)$



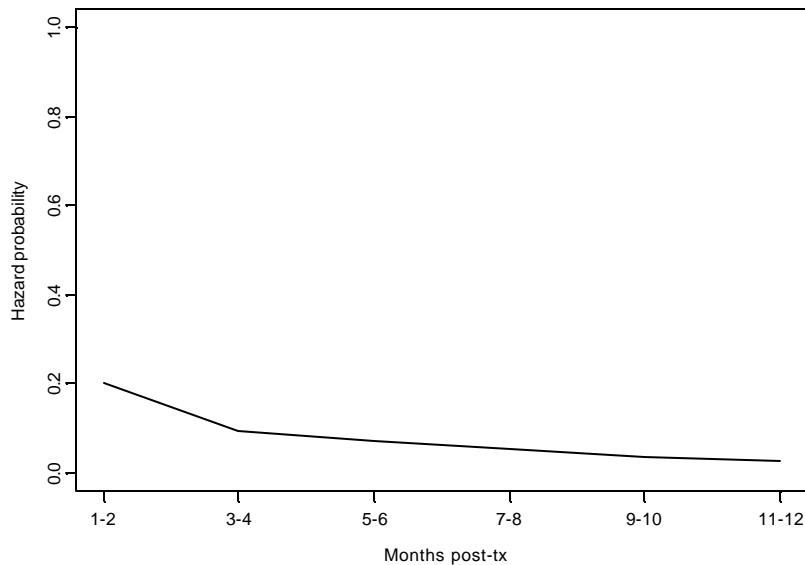
## Hazard probability

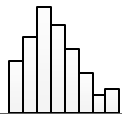
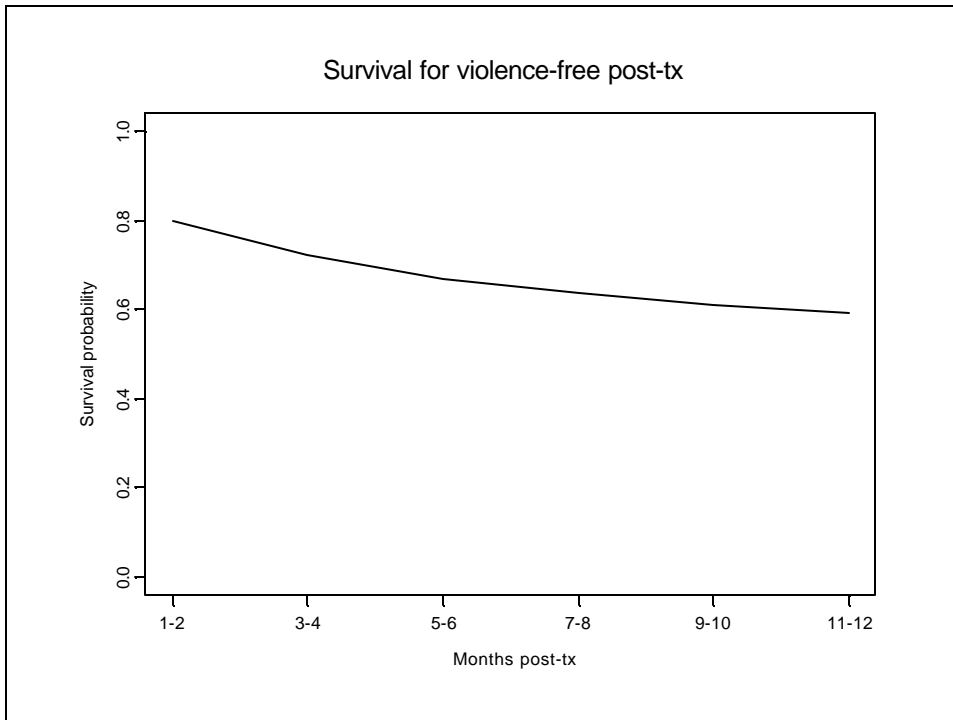
$h(j)$ , called the *hazard probability*, is defined as the probability of the event occurring in the time interval  $j$ , provided it has not occurred prior to  $j$ :

$$h(j) = P(T = j \mid T \geq j).$$

In other words,  $h(j)$  is the proportion of individuals at-risk at the beginning of time period  $j$  who experience the event sometime during period  $j$ .

Hazard for first post-tx violence

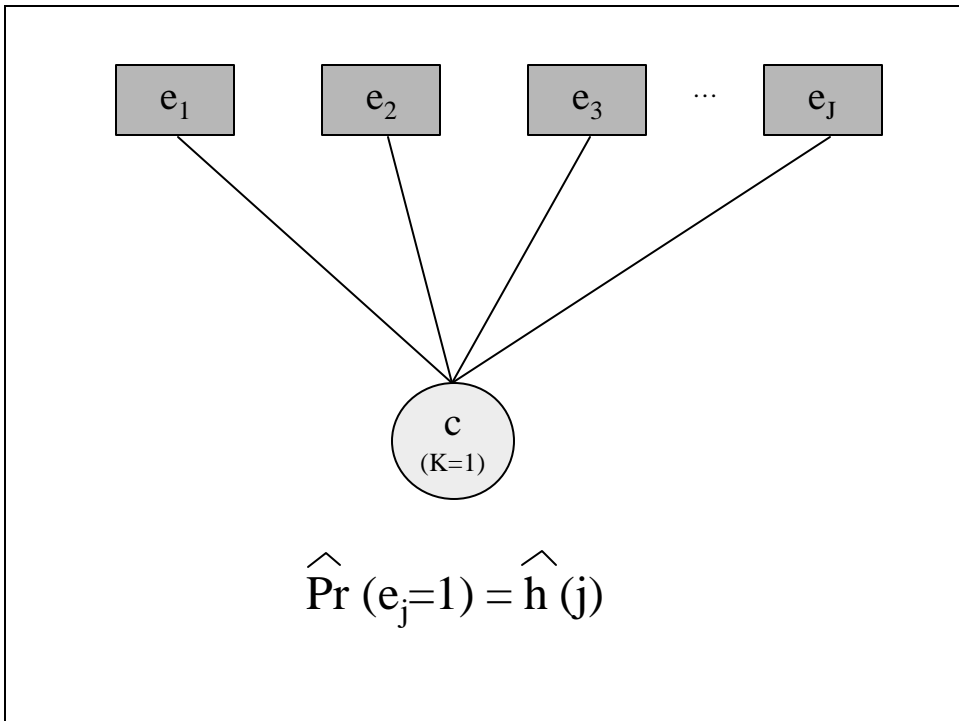


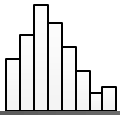


## Missing data

- ✦ The most typical survival data is right-censored and this type of *missingness* is the easiest to deal with in the data analysis.
- ✦ Right censoring occurs when a subject in the sample has *not* experienced the event of interest at the end of the observation period. It is assumed that the event eventually occurs sometime after the end of the study.
- ✦ Censoring is usually assumed to be *noninformative*. (Think: MAR)

	1	2	3	4	5	6
E	0	0	0	0	1	•
R <sup>0</sup>	1	1	1	1	1	0
E	0	0	0	•	•	•
R <sup>0</sup>	1	1	1	0	0	0
E	0	0	0	0	0	0
R <sup>0</sup>	1	1	1	1	1	1



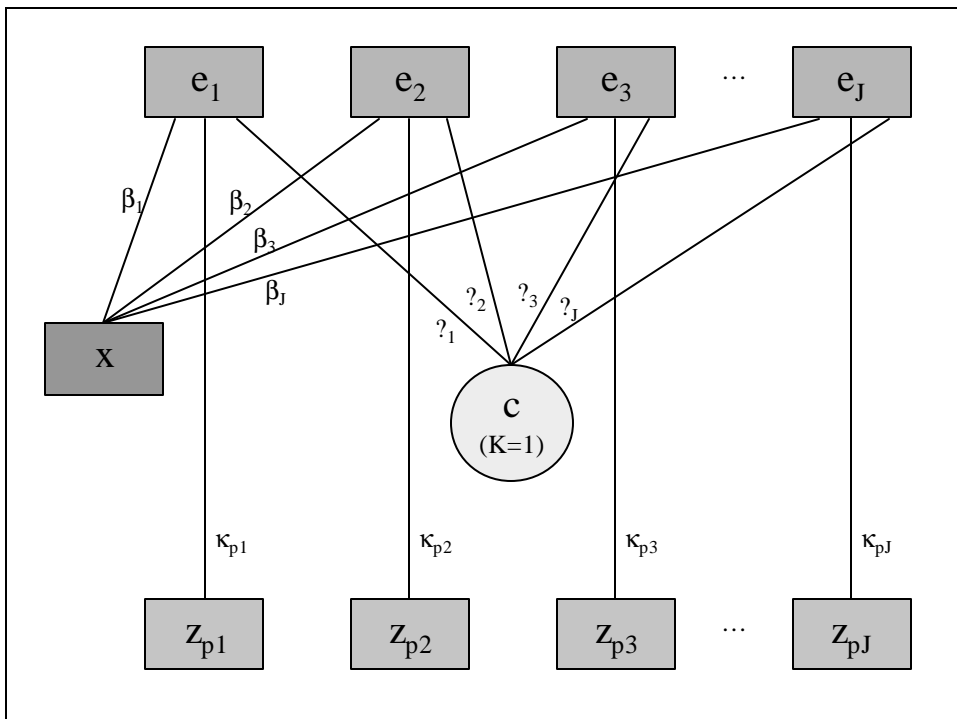


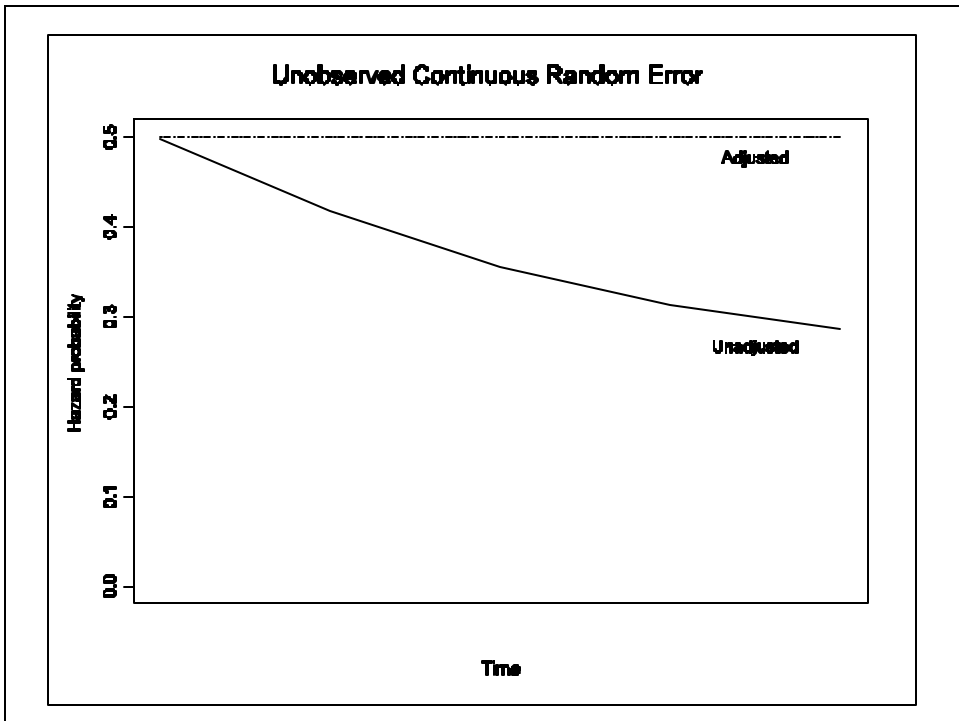
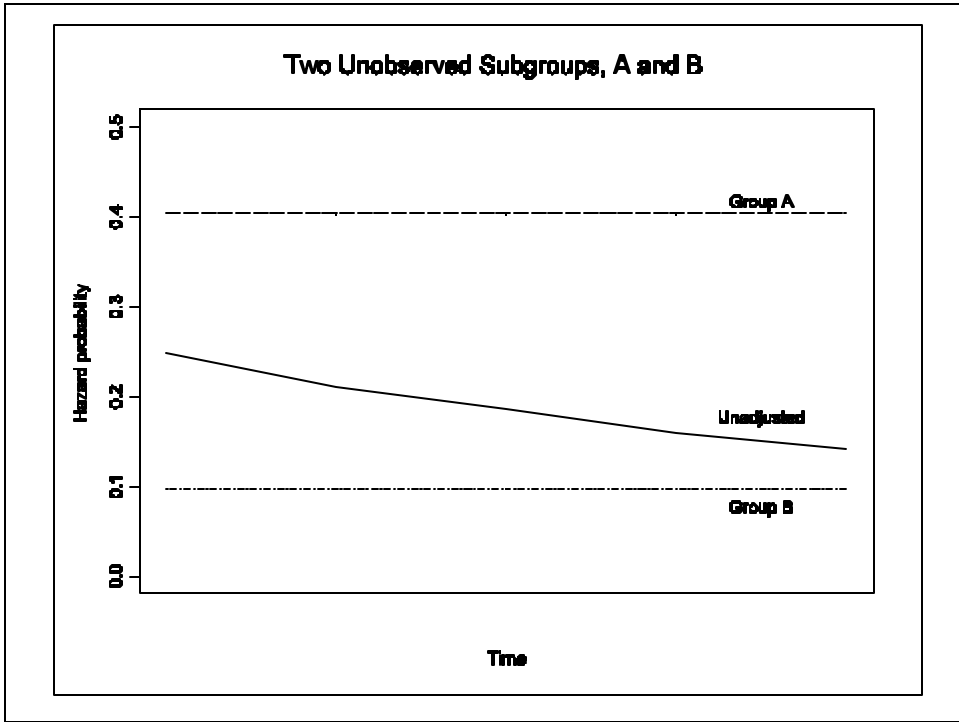
## DTSA Model w/ covariates\*

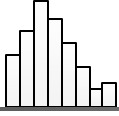
$$\text{logit } h(j) = \log\left(\frac{h(j)}{1-h(j)}\right) = -t_j + \mathbf{b}_j x + \mathbf{k}_j z_j$$

$$h(j) = \frac{1}{1 + \exp(t_j - \mathbf{b}_j x - \mathbf{k}_j z_j)}$$

\* This model, for single events with no random effects, yields identical results to the formulation in the traditional logistic regression model, á la Singer & Willet.

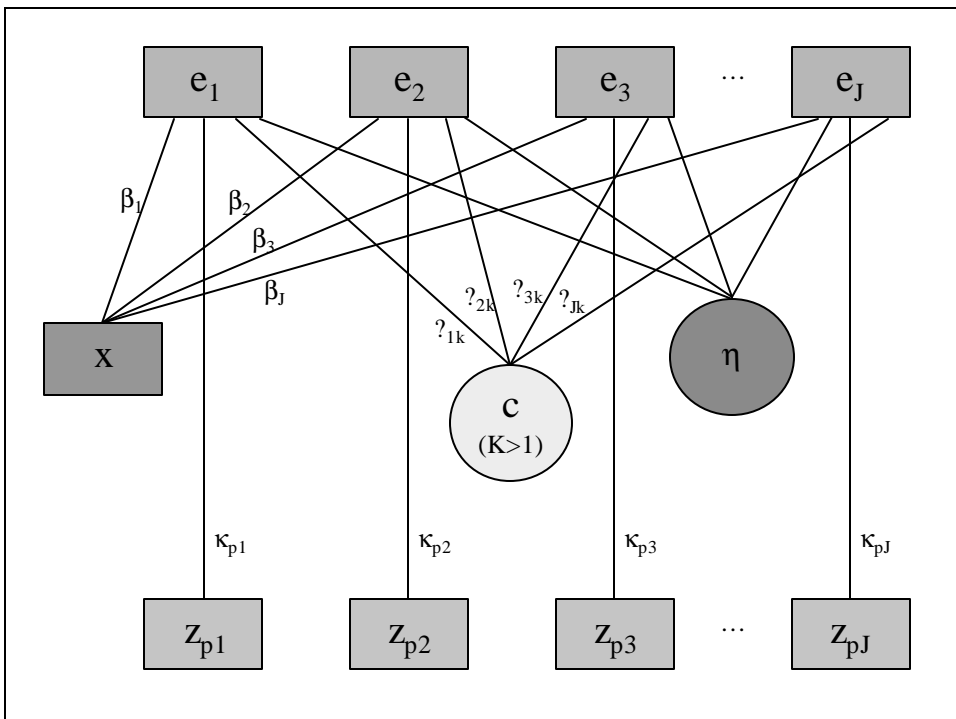


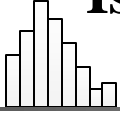




# Ignoring unobserved heterogeneity

- ✦ Baseline hazard probabilities biased downward
- ✦ Time-independent covariate effects underestimated
- ✦ Spurious time-dependent effects for observed variables

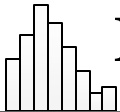




## Issues in modeling unobserved heterogeneity (frailty)

---

- ✦ Identification
- ✦ Sensitivity to parametric misspecification
- ✦ Specification of nonparametric models
- ✦ Goodness-of-fit

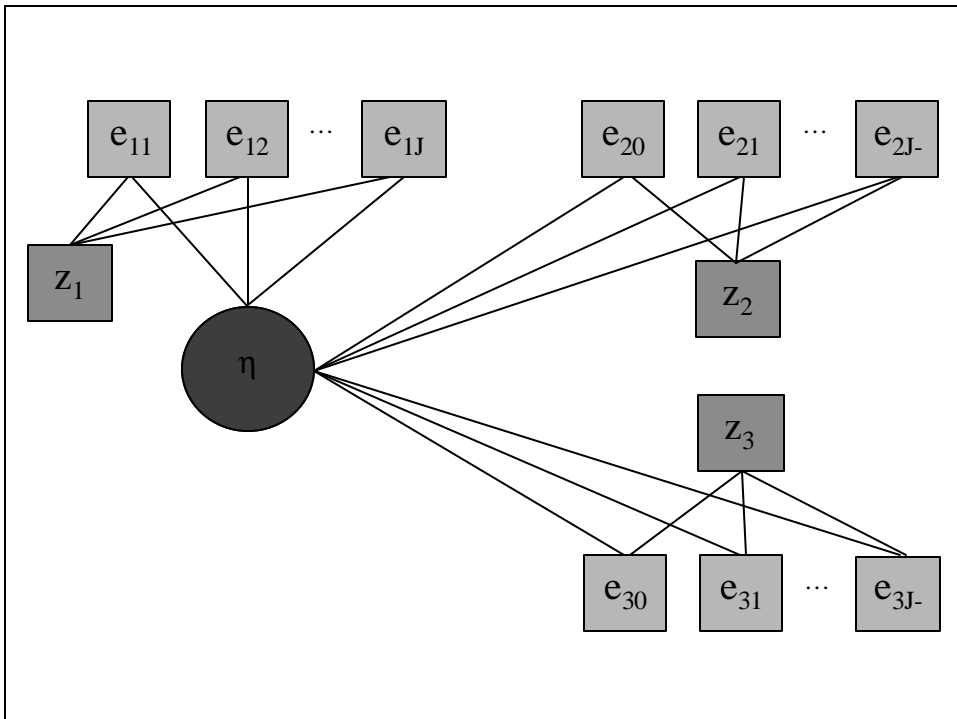
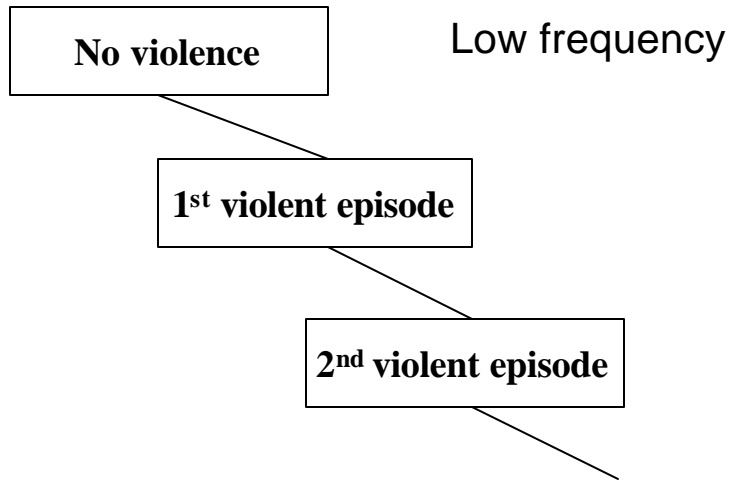


## Multivariate event histories

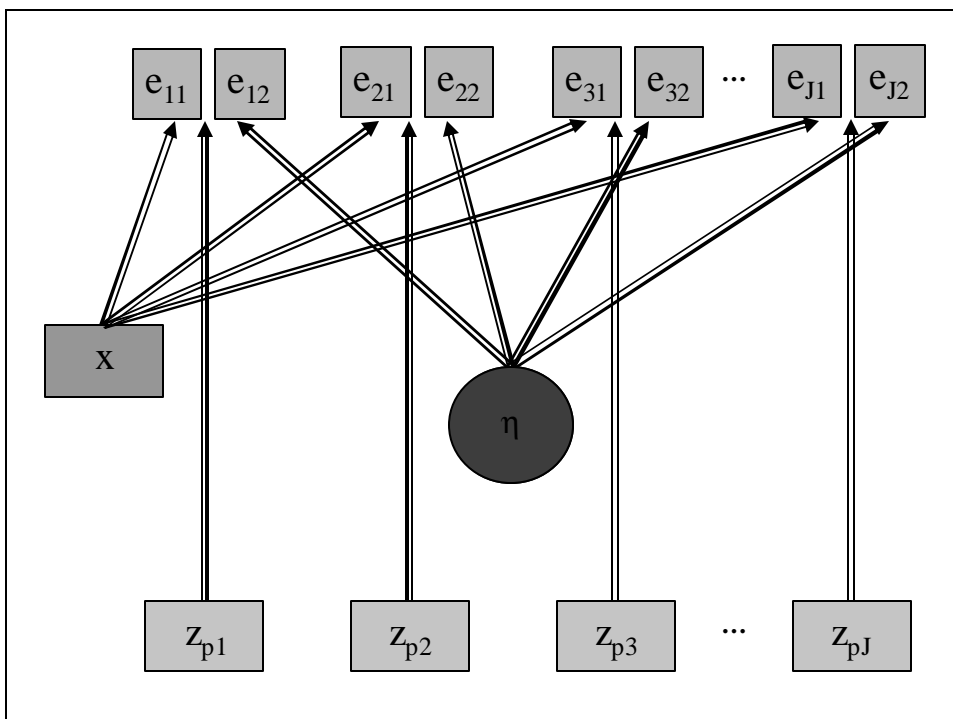
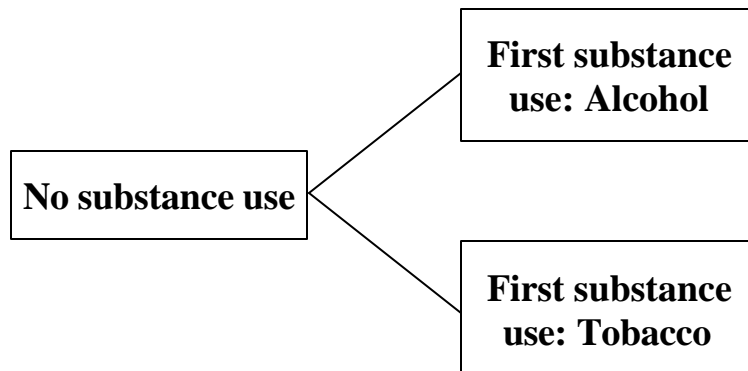
---

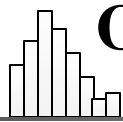
- ✦ Recurrent events (multiple spells):  
Same outcome that may occur more than once (high or low frequency)
- ✦ Competing risks:  
More than one possible outcome

# Recurrent events example



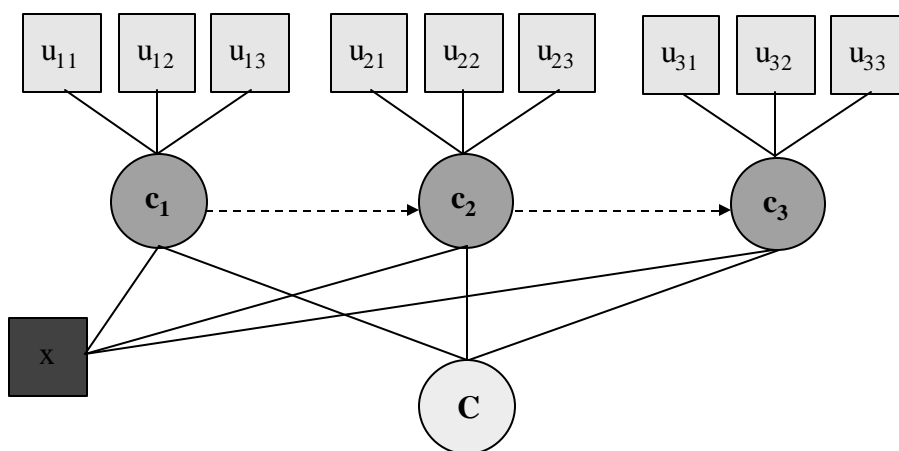
# Competing risks example

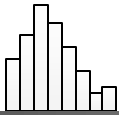




## Other sources of uncertainty

- ✦ Recall error and bias in timing of event occurrence, e.g., timing of first drink, timing of drinking-related problems
- ✦ Measurement error in event status, e.g., symptom check-list used to determine latent alcohol dependence status

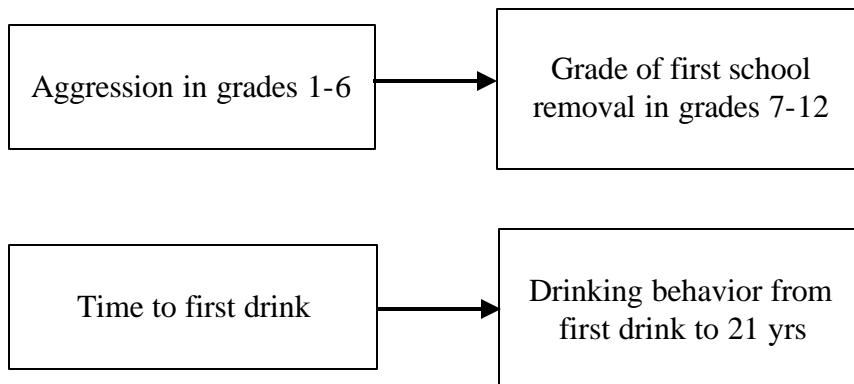


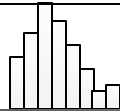


## Other extensions

- ✦ Continuous or categorical latent variable predictors of survival measured by other manifest variables, e.g., multiple survey items as measures of stress predicting time-to-event or clinical diagnostic criteria as measures of psychological profiles predicting time-to-event.
- ✦ Multilevel DTSA models
- ✦ Joint longitudinal processes

## Survival and growth combo models





## Select references

- + **Hedeker, D., Siddiqui, O., & Hu, F.B.** (2000). Random-effects regression analysis of correlated group-time survival data. *Statistical Methods in Medical Research*, 9(2), 161--179.
- + **Land, K.C., Nagin, D.S. & McCall, P.L.** (2001). Discrete-time hazard regression models with hidden heterogeneity: The semiparametric mixed Poisson regression approach. *Sociological Methods and Research*, 29(3), 342--373.
- + **Masyn, K.E.** (2003). *Discrete-time survival mixture analysis for single and recurrent events using latent variables*. Unpublished doctoral dissertation, University of California, Los Angeles. <http://www.statmodel.com/download/masyndissertation.pdf>
- + **Muthén, B. & Masyn, K.** (2004). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, in press.
- + **Singer, J.D. & Willett, J.B.** (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- + **Vermunt, J.K.** (2002). A general latent class approach to unobserved heterogeneity in the analysis of event history data. In J.A. Hagnaars & A.L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 383-407). Cambridge: Cambridge University Press.
- + **Steele, F.** (2003). A multilevel mixture model for event history data with long-term survivors: An application to an analysis of contraceptive sterilisation in Bangladesh. *Lifetime Data Analysis*, 9, 155-174.