Likelihood Ratio Inference for Survival Data in Genetics and Epidemiology

Sergey V. Malov

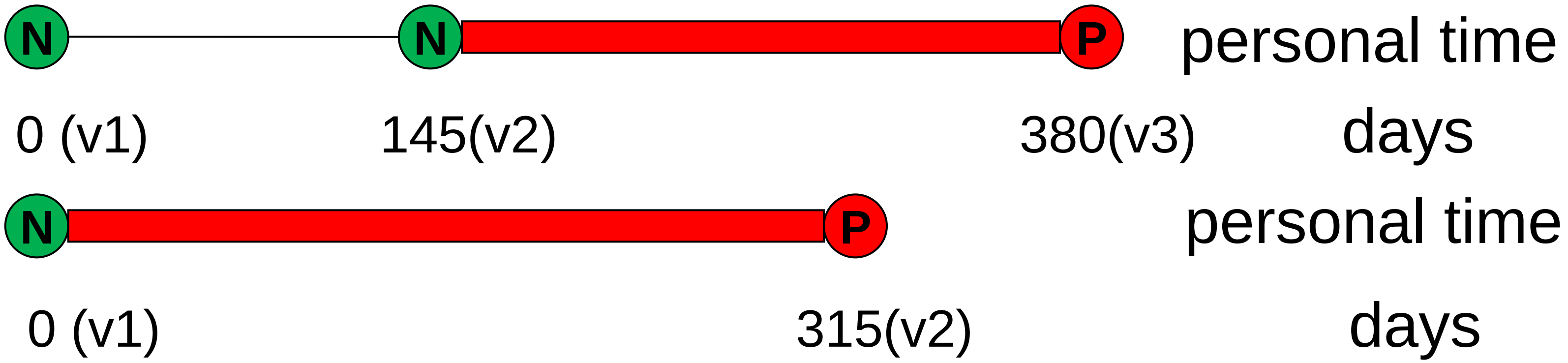
Theodosius Dobzhansky Center
for Genome Informatics,
St.-Petersburg State University

Data Description

- An association of single (or multiple) genes with an incidence (AIDS, HIV+) will be discussed.
- For more specificity we look on associations of HIV infection with genetic characters, but the methods can be used in some different cases of epidemiological research.
- Originally all individuals are HIV-positives.
- Any individual is tested on HIV infection at several sequential times.
- Necessary information for any individual: the starting time, time of the last visit when individual was HIV-negative, time of the first visit when HIV incidence were detected and genomic information.
- HIV-negative individuals at the endpoint or missed after two or more visits are assumed to be censored.

Used Data Interpretation

For infected individuals:



N N P personal time

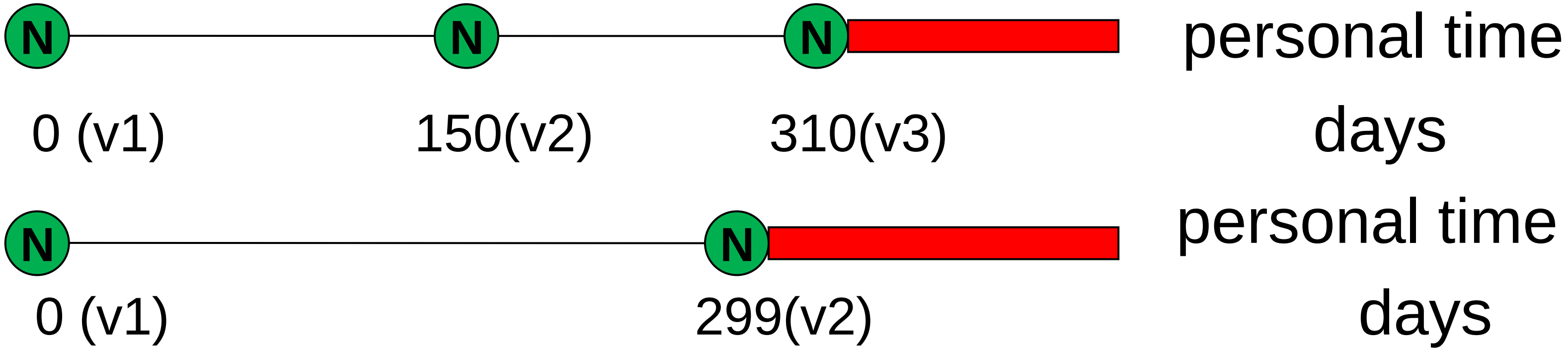
0 (v1) 145(v2) 380(v3) days

N P personal time

0 (v1) 315(v2) days

Individual can be infected anywhere in red zone

For non infected individuals:



N N N personal time

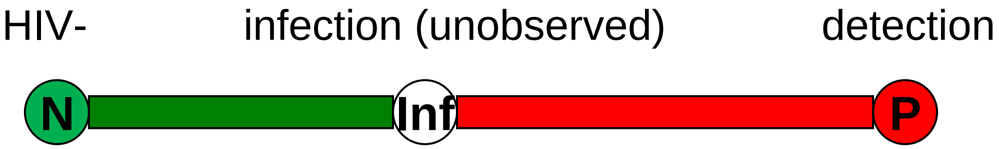
0 (v1) 150(v2) 310(v3) days

N N personal time

0 (v1) 299(v2) days

Individual was not infected until the last visit, but can be infected anywhere in red zone

Interpretation of observations



HIV- infection (unobserved) detection

N Inf P

- •Remark that the observed times are not true times of infection.
- •Substitution of true infection times by detection times lead obviously to some bias in estimation.
- •Under comparative analysis the effect of the substitution is typically vanishing.
- •In classical models of survival analysis the exact failure times are required.
- •Interval censored data approach is more applicable for the data under consideration.
- •One more important question is how to interpret the starting point.

Starting point interpretation problem

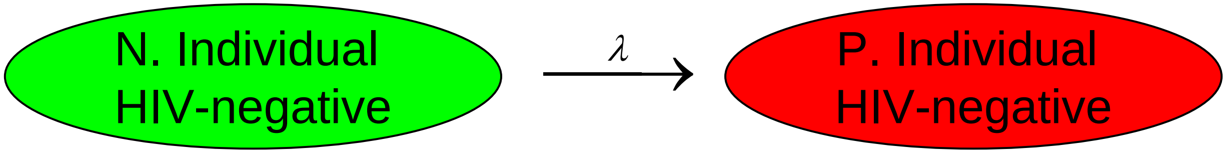
- Independence and identical distribution of observations in any group of homogeneous individuals are required to get statistical results.
- In AIDS research the starting point is not the minimal point when individual can be infected.
- Actually we deal with a left truncated distribution and in general case there is no reasons to assume that the individuals have similar truncated distribution of HIV infection time.
- Remark that whole distribution is equal to the left restricted one for the class of exponential distributions.
- For HIV+ to AIDS transition the time of HIV+ infection is correctly assumed as point 0, but the exact HIV+ infection time is not observed.

Different Models in Survival Analysis

- To specify model one may use nonparametric, semiparametric or parametric approach.
- Nonparametric methods based on minor requirements but they are not efficient, especially under interval censored data case.
- Semiparametric methods are more efficient in compare with nonpaprametric methods. The most common is Cox proportional hazards model.
- Parametric models are most convenient to use for interval censored data but strict limitations on class of possible distributions are imposed.

Birth and Death Process Approach

- For any fixed individual: exact time of transition



N. Individual
HIV-negative

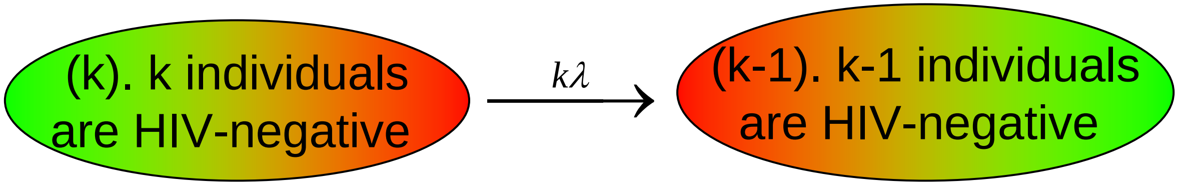
λ

P. Individual
HIV-negative

is time of first occurrence in Poisson process with intensity *λ.*

- Let *T* – be exact time of transition from N to P. T is the random variable having survival function $\mathrm{P}(T > t) = e^{-\lambda t}, \quad t \geq 0.$

For population: exact time of transition



(k). k individuals
are HIV-negative

$k\lambda$

(k-1). k-1 individuals
are HIV-negative

is time of first occurrence in Poisson process with intensity k*λ.*

- It is the so called “Pure death process” (see Karlin (1966))
- Let *T* – be exact time of transition from k to k-1. T is the random variable having survival function

$$\Pr(T > t) = e^{-\lambda kt}, \quad t \,\square\, 0.$$

Survival analysis approach

- Hazard function (*λ*(*t*)) – intensity of HIV- disease. For any individual

$$\lambda(t)dt = \frac{\Pr(T \in [t, t+dt])}{\Pr(T \geq t)}.$$

- Cumulative hazard (*Λ*(*t*)): $\Lambda(t) = \int_0^t \lambda(u)du.$
- Survival function: $S(t) = \Pr(T > t) = \mathrm{e}^{-\Lambda(t)} = \exp\left(-\int_0^t \lambda(u)du\right).$
- In case of constant hazard function *λ*(*t*) = *λ,*

$\Lambda(t) = \lambda t; \quad S(t) = \mathrm{e}^{-\lambda t}, t \,\square\, 0,$ (Exponential distribution)

it is the same as in Birth and death process model.

- If to assume $\lambda(t) = \lambda\gamma(\lambda t)^{\gamma-1}, t \geq 0,$ then

$\Lambda(t) = \left(\lambda t\right)^{\gamma}; \quad S(t) = \mathrm{e}^{-(\lambda t)^{\gamma}}, t \geq 0.$ (Weibull distribution)

- It is a generalization of the exponential model
- Other generalizations of the exponential model may be considered.

Annual Rate

- Annual rate, or percent of the population acquired HIV disease in one year

$$I = 1 - S(t)\,|_{t=1yr.} * 100\% = 1 - \exp(-\Lambda(t))\,|_{t=1yr.} * 100\%$$

- The x-year rate is

$$I(x) = (1 - \exp(-\Lambda(x))) * 100\,\%.$$

- Remark that $1 - \exp(-u) \approx u$ under small *u* and, therefore, $\Lambda(t)\,|_{t=1yr.}$ is approximately annual rate if it is small.

- Under constant hazard (exponential model),

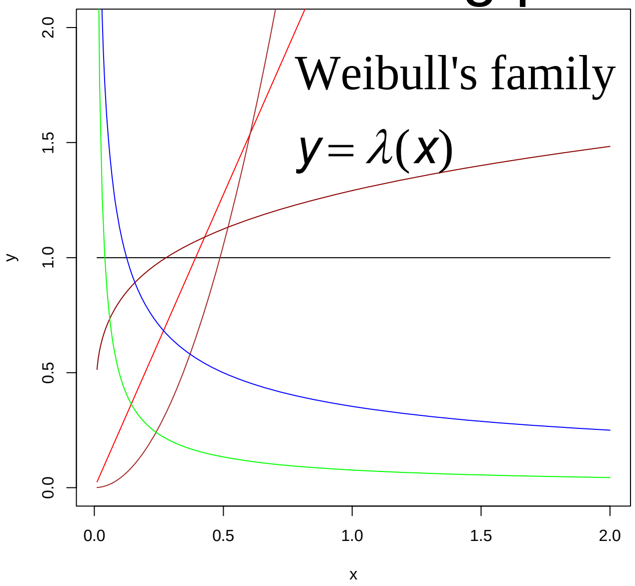
$$I = (1 - \exp(-\lambda)) * 100\%$$

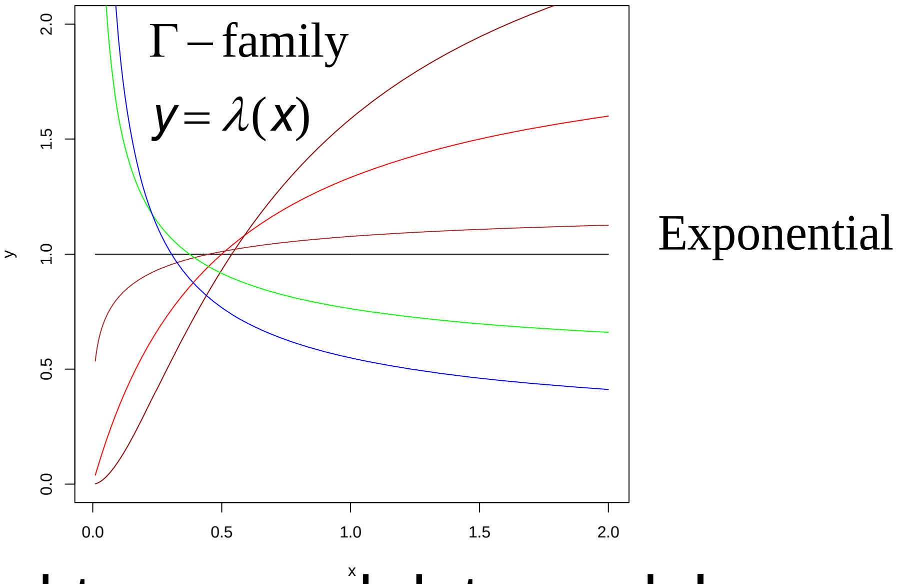
and x-year rate is calculated as follows:

$$I(x) = (1 - \exp(-\lambda x)) * 100\% = (1 - (I/100)^{x}) * 100\,\%.$$

On starting point interpretation problem

- For an exponential distribution of failure time left truncated distribution is the same as the original one, starting point can be taken as zero, but for other distributions of failure time the starting point interpretation problem is actual





- To use left truncated and right censored data model one need to specify time interval from zero to starting point.
- Results in more general Weibull’s or Γ-models can’t be interpreted correctly for any individual personally, but they can be interpreted for population of the similar form.

Maximum Likelihood Estimation

- Maximum likelihood method is efficient in wide conditions
- Let *N* be the size of population, $[L_i, U_i]$ be the personal time interval for the exact time of HIV-infection,

  δ_i indicates that *i*-th individual became HIV-positive.

Log likelihood function under the exponential model:

$$\mathrm{LL}(\overline{L}, \overline{U}, \overset{\rho}{\delta}; \lambda) = \sum_{i=1}^{N} \delta_i \log(1 - \exp(-\lambda(U_i - L_i))) - \lambda \sum_{i=1}^{N} L_i$$

- Let $\hat{\lambda}$ be the maximum likelihood estimator. Then the (estimated) annual rate is given by

$$I = 1 - \exp(-\hat{\lambda}).$$

- Brookmeyer & Goedert (1989) used more general Weibull's model, as we discussed it is not necessary lead to better interpretation in compare with the exponential model.

Homogeneity Testing

- In non homogeneous population (k-groups) it is interesting to test significance of difference in HIV-incidence rate between groups.
- Let D(j) be the disjoint sets of members of different groups, λ_j be corresponding hazards of HIV-incidence. The log likelihood function is

$$\mathrm{LL}(\overline{L},\overline{U},\overset{\rho}{\delta},\overset{\rho}{D};\overset{\rho}{\lambda}) = \sum_{j=1}^{k}\left(\sum_{i\in D(j)}\delta_i \log(1-\exp(-\lambda_j(U_i - L_i))) - \lambda_j \sum_{i\in D(j)} L_i\right)$$

- The homogeneity LR-test statistic is

$$\mathrm{LRS} = 2*(\mathrm{LL}(\overline{L},\overline{U},\overset{\rho}{\delta},\overset{\rho}{D};\overset{\rho}{\lambda}) - \mathrm{LL}(\overline{L},\overline{U};\hat{\lambda})) \sim \chi^2_{k-1}$$

  has asymptotically chi-square distribution with k-1 degrees of freedom.
- Moreover, one can use asymptotic normality of MLE to get asymptotic confidence intervals and multiple comparison methods.

Semiparametric inference

- The most commonly used semiparametric model is the proportional hazards Cox’s (1972) model

$$\lambda(x\,|\,z) = \lambda_0(x)\exp(\beta^T z)$$

where $\lambda(x\,|\,z)$ is the hazard rate function under covariate z and $\lambda_0(x)$ is some baseline hazard rate function.

- The remarkable property of Cox’s model is that hazard functions are proportional

$$\lambda(x\,|\,z_1)\,/\,\lambda(x\,|\,z_2) = \exp(\beta^T(z_1 - z_2))$$

but this property can be restrictive in general case.

- There are extensions of Cox’s model

$\lambda(x\,|\,z(\square)) = r(z(t);\theta)\,\lambda_0(x) + h(z(t);\theta)$ (AM-model,Lin&Ying (94+))

$\frac{\partial f_{z(\square)}(t)}{\partial t} = r(z(t);\theta)\,\frac{\partial f_0(t)}{\partial t} + h(z(t);\theta)$ (GAM-model, Bagdonavicius&Nikulin)

where $f_{z(\square)}(x) = H\ \ S(x\,|\,z(\square))$ and H is inverse to some survival function

- The last models allow to use time dependent covariates.

Cox’s model

- Cox’s model is more convenient for analysis in compare with the generalizations.
- The (partial) loglikelihood function is given by

$$LL(\beta) = \sum_{k=1}^{L} \left[\beta^T z_k - \log\left[\sum_{j \in R_k} \exp\left(\beta^T z_j\right) \right] \right]$$

where R_k is the number of individuals at risk at k-th sequential failure times and score equations are given by

$$U(\beta) = \sum_{k=1}^{L} \left[z_k - \sum_{j \in R_k} z_j \exp(\beta^T z_j) \Big/ \sum_{j \in R_k} \exp(\beta^T z_j) \right] = 0$$

- There are three common tests in Cox’s model:

  1. Wald’s test 2. Score-test 3. LR-test

- Remark that the likelihood function is independent of nonparametric component.

Cox model under interval censored data

- The interval censored data case considered by Finkelstein (86)
- The log likelihood function in this case is dependent on nonparametric component

$$LL(L,U;\beta,\gamma)=\sum_{i=1}^{N}\log\left[\sum_{j=1}^{m}\alpha_{ij}\left\{\exp(-\exp(\beta^{T}z_i+\gamma_{j-1}))-\exp(-\exp(\beta^{T}z_i+\gamma_{j-1}))\right\}\right]$$

where $\gamma_j=\log(-\log(s_j))$ are parameters of nonparametric component, $s_j=S_0(t_j)$, $0=t_1<\ldots<t_m<\infty$ are sequential times including all L_i and U_i, α_{ij} is the indicator of $(s_j,s_{j+1}]\subset(L_i,U_i]$.

- The Newton–Raphson method can be used to get MLE for parameters β and γ.
- To test hypothesis $H:\beta=0$ the score test is commonly used

$$U=\sum_{i=1}^{N}\sum_{j=1}^{m}\left[\frac{z_j\log\hat{p}_j\sum_{k=j}^{m}\alpha_{ik}\hat{g}_k}{\sum_{n=1}^{m}\alpha_{in}\hat{g}_n}-z_i\frac{\log\hat{p}_j}{1-\hat{p}_j}\frac{\alpha_{ij}\hat{g}_j}{\sum_{n=1}^{m}\alpha_{in}\hat{g}_n}\right]$$

$\hat{p}_j=\hat{S}_j/\hat{S}_{j-1},\ \hat{g}_j=\hat{S}_{j-1}-\hat{S}_j,\ \hat{S}_j$ are ML-estimators of $S_0(t_j)$ under *H.*

Nonparametric inference

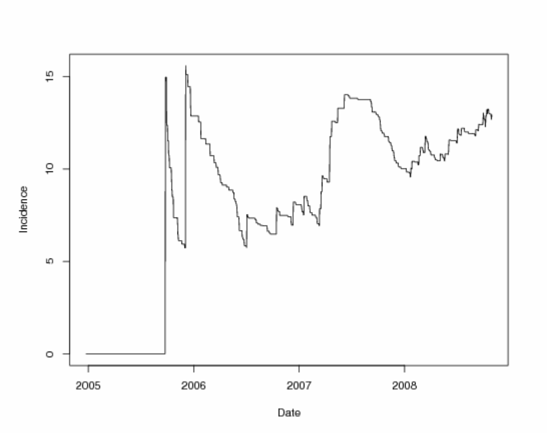
- The non parametric approach requires minimal assumptions on the distribution of the infection time.
- The starting point interpretation problem requires to consider left truncated and right censored data model.
- Common methods of nonparametric comparative analysis in right censored data case based on linear rank statistics (logrank, GW).
- Nonparametric methods of comparative analysis under interval censored data developed by Zhang et al. (2001) for two sample problem and Yuen et al. (2006) for multiple samples but they are not widely used in applications.

Choosing an original model

- What kind of model to use for analysis?
  - Exponential model is the best for interpretation, if to find that the exponential model is good associated with the original data it is reasonable to use the exponential model.
  - If the exponential model is strictly contradict to the observed data one have to use the extension – Weibull or Γ- parametric model, Cox’s model or nonparametric model.
  - Using non exponential approach one need to have in mind the starting point problem.

Further Analysis

- The exponential model can be improved by the calendar time covariate. The following picture obtained by real data. We used the exponential model with data increasingly censored at time point *T* and calculate ML-estimators by the censored data



- Linear trend by calendar time seems natural.
- For further analysis it is natural to assume that

$$\lambda_{j(i)}(t) = a_j + b(t + T_i)$$

where T_i is the difference between time of the first visit of *i*-th individual and starting time of the project.

References

- Bagdonavicius V., Nikulin M. (1997) Transfer functionals and semiparametric regression models. *Biometrika* **84**(2), 365-378.
- Brookmeyer, R. & Goedert, J.J. (1989). Censoring in an Epidemic with an Application to Hemophlia-Associated AIDS. *Biometrics* **45**(1), 325-335.
- Finkelstein D.M. (1986) A Proportional Hazards Model for Interval-Censored Failure Time Data. *Biometrics* **42** (4), 845-854.
- Karlin, S. (1966). *A First Course in Stochastic Processes*. Academic Press, New York- London 1966.
- Turnbull, B.W. (1974). Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association*, **69**(345), 169-173.
- Peto, R. (1973). Experimental Survival Curves for Interval-censored Data. *Applied Statistics*, **22**(1), 86-91.
- Yuen K-C, Shi J., Zhu L.(2006) A k-Sample Test with Interval Censored Data. *Biometrika* **93** (2), 315-328.
- Zhang Y., Liu W., Zhan Y. (2001) A Nonparametric Two-Sample Test of the Failure Function with Interval Censoring Case 2. *Biometrika* **88**(3), 677-686.