Biostatistics

Applications in Medicine

Nuno Sepúlveda, 27.11.2023

Syllabus

1. General review
   a. What is Biostatistics?
   b. Population/Sample/Sample size
   c. Type of Data – quantitative and qualitative variables
   d. Common probability distributions
   e. Work example – Malaria in Tanzania

2. Applications in Medicine
   a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve,Rogal-Gladen estimator
   b. Estimation of treatment effects - generalized linear models
   c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model

3. Applications in Genetics, Genomics, and other 'omics data
   a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
   b. Methylation association studies – M versus beta values, estimation of biological age
   c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial

4. Other Topics
   a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
   b. Serological analysis – Gaussian (skew-normal) mixture models
   c. Advanced sample size and power calculations

Parametric analysis

versus

Non-parametric analysis

Parametric analysis

Non-parametric analysis

Non-parametric methods

Comparison of different survival curves

Log-rank test
Peto-Peto test

Semi-parametric regression

Cox's proportional hazard model

Comparison of different survival curves

Two treatments under comparison

Time to clinical response

$$H_0 : S_1(t) = S_2(t) \text{ versus } H_0 : S_1(t) \neq S_2$$

Log-rank test as a Mantel-Haenszel test for categorical data

Mantel-Haenszel test

Analysis of the association in K x 2 x 2 contingency tables (an extension of Fisher's exact test to K tables 2 x 2).

Stratum	Treatment	Responded	Not Responded
1	A		
	B		
2	A		
	B		
3	A		
	B		

In stratum i

$$\Delta_i = \frac{\pi_{1i}(1-\pi_{2i})}{(1-\pi_{1i})\pi_{2i}}$$

$\pi_{1i} =$ prob. of response to treatment 1

$\pi_{2i} =$ prob. of response to treatment 2

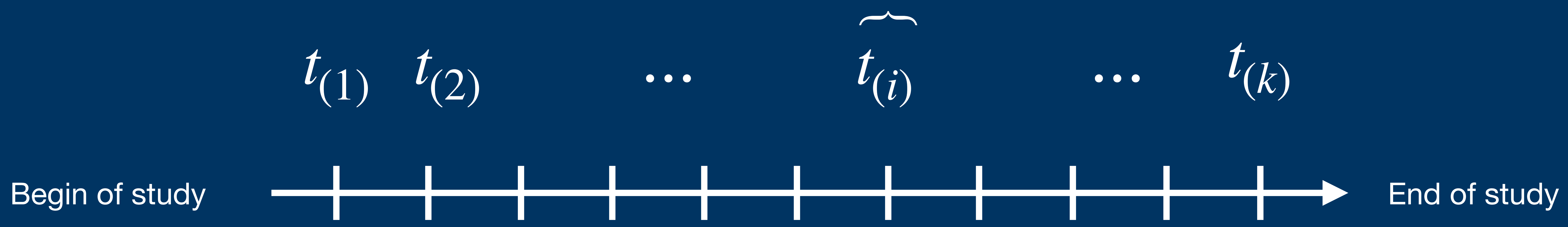
$H_0 : \Delta_1 = \cdots = \Delta_K = 1\,(t)$ versus $H_1 : \exists_{i,j} \Delta_i \neq \Delta_j = 1$

under the assumption of $\Delta_1 = \cdots = \Delta_K = \Delta$

Mantel & Haenszel (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. Journal of the National Cancer Institute 22, 719-748

Log-rank test

Adaptation of the classical Mantel-Haenszel test for k x 2 x 2 contigency tables where k is the number of different timepoints in which it was observed the event of interest



Basic idea

There are k 2 x 2 tables like this one

Group	Number of “deaths” at $t_{(i)}$	Number of “survivors” beyond $t_{(i)}$	Total
1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Conditional probalitity (see Fisher's exact test)

$H_0 : S_1(t) = S_2(t)$ versus $H_1 : S_1(t) \neq S_2(t)$

$H_0 : \pi_{1i} = \pi_{2i} = \pi$ versus $H_1 : \pi_{1i} \neq \pi_{2i}$

$\pi_{1i} =$ probability of "death" at time $t_{(i)}$ in group 1

$\pi_{2i} =$ probability of "death" at time $t_{(i)}$ in group 2

$d_{li} \,|\, \pi_{li}, n_{li} \rightsquigarrow$ Binomial$(n = n_{li}, \pi = \pi_{li}), l = 1{,}2$

$d_i \,|\, \pi_{li}, n_{li}, H_0 \rightsquigarrow$ Binomial$(n = n_i, \pi = \pi_i)$

Basic idea

Calculate the distribution of d_{1i} conditional to the total marginals

Group	Number of "deaths" at $t_{(i)}$	Number of "survivors" beyond $t_{(i)}$	Total
1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Conditional probability (see Fisher's exact test)

$$d_{1i} \,|\, d_i, n_i, n_{1i}, H_0 \rightsquigarrow \text{Hypergeometric}(N = n_i, M = d_i, n = n_{1i})$$

$$P\left[d_{1i} = d \,|\, d_i, n_i, n_{1i}, H_0\right] = \frac{\binom{d_i}{d}\binom{n_i - d_i}{n_{1i} - d}}{\binom{n_i}{n_{1i}}}$$

$$E\left[d_{1i} \,|\, d_i, n_i, n_{1i}, H_0\right] = n_{1i}\frac{d_i}{n_i} \qquad Var\left[d_{1i} \,|\, d_i, n_i, n_{1i}, H_0\right] = n_{1i}\frac{d_i}{n_i}(1 - \frac{d_{d_i}}{n_i})\frac{n_i - n_{1i}}{n_i - 1}$$

Test statistic

Incorporating information from k 2 x 2 contingency tables

$$U = \sum_{i=1}^{k} \left(d_{1i} - e_{1i}\right)$$

$$e_{1i} = E\left[d_{1i} \,|\, d_i, n_i, n_{1i}, H_0\right] = n_{1i}\frac{d_i}{n_i}$$

$$E\left[U \,|\, H_0\right] = 0$$

$$v_{1i} = Var\left[d_{1i} \,|\, d_i, n_i, n_{1i}, H_0\right]$$

$$= n_{1i}\frac{d_i}{n_i}\left(1 - \frac{d_{d_i}}{n_i}\right)\frac{n_i - n_{1i}}{n_i - 1}$$

$$Var\left[U \,|\, H_0\right] = \sum_{i=1}^{k} v_{1i}$$

Log-rank test

For large samples

$$Q = \frac{U - \overbrace{E(U)}^{=0}}{\sqrt{var(U)}} \mid H_0 \rightsquigarrow \text{Normal}(\mu = 0, \sigma = 1)$$

$$Q* = \frac{U^2}{var(U)} \mid H_0 \rightsquigarrow \chi^2_{(1)}$$

$$p = P\left[Q* > q_{obs} \mid H_0\right]$$

Decision rule

$$\begin{cases} \text{do not reject } H_0, & \text{if } p > \alpha \\ \text{reject } H_0, & \text{otherwise} \end{cases}$$

A general class of non-parametric tests

$$Q^* = \frac{\left[\sum_{i=1}^{k} w_i \left(d_{1i} - e_{1i}\right)\right]^2}{\sum_{i=1}^{k} w_i^2 v_{1i}} \,|\, H_0 \rightsquigarrow \chi^2_{(1)}$$

for large samples

Choices of the weights

$w_i = 1$, log-rank test

$w_i = \sqrt{n_i}$, Tarone-Ware test

$w_i = n_i$, Gehan test

$w_i = \prod_{j:t_(j) \leq t_(j)} \left(1 - \frac{d_j}{n_j + 1}\right)$, Peto-Peto test

A general statistic for non-parametric tests

$$Q^* = \frac{\left[\sum_{i=1}^{k} w_i \left(d_{1i} - e_{1i}\right)\right]^2}{\sum_{i=1}^{k} w_i^2 v_{1i}} \,|\, H_0 \rightsquigarrow \chi^2_{(1)}$$

for large samples

Choices of the weights

$w_i = 1$, log-rank test

$w_i = \sqrt{n_i}$, Tarone-Ware test

$w_i = n_i$, Gehan test

$w_i = \prod_{j:t_{(j)} \leq t_{(j)}} \left(1 - \frac{d_j}{n_j + 1}\right)$, Peto-Peto test

Harrington-Fleming class of tests (Survival package)

$$Q^* = \frac{\left[\sum_{i=1}^{k} w_i \left(d_{1i} - e_{1i}\right)\right]^2}{\sum_{i=1}^{k} w_i^2 v_{1i}} \,|\, H_0 \rightsquigarrow \chi^2_{(1)}$$

for large samples

$$w_i = \hat{S}^{\rho}_{t_{(i)}}(1 - \hat{S}_{t_{(i)}})^{\gamma},$$

$\hat{S}_{t_{(i)}}$ = Kaplan-Meier estimates for the common survival function

$\gamma = 0 \Rightarrow w_i = \hat{S}^{\rho}_{t_{(i)}}$

$\rho = 0 \Rightarrow w_i = 1$ - log-rank test

$\rho = 1 \Rightarrow w_i = \hat{S}_{t_{(i)}}$ - Peto-Peto test

Exercise 1: rituximab clinical trial data

survival package (analysis) survminer package (plotting)

Plot survival curves (surfit command) of time to treatment response for:

(i) males versus females
(ii) patients with and without an infection disease trigger
(iii) patients with and without family history of autoimmune diseases

Compared with the curves for each case using log-rank and Peto-Peto tests (survdiff function from survival package)

Draw your conclusions.

Non-parametric methods

Comparison of different survival curves

Log-rank test
Peto-Peto test

Semi-parametric regression

Cox’s proportional hazard model

Cox’s proportional hazard model

$$h_{x_{ij}}(t) = h_0\,(t)\,e^{\sum_{j=1}^{p} \beta_j x_{ij}}$$

“All models are wrong, some are useful.”

George Box (1976)

John Wiley
https://rss.onlinelibrary.wiley.com › doi › j.2517-6161.1...

Regression Models and Life-Tables - Cox - 1972

by DR Cox · 1972 · Cited by 60930 — **Cox**, **D. R.** (1959). The **analysis** of exponentially distributed **life**-times with two types of failure. **J. R. Statist**. **Soc**. **B**, 21, 411–421. **Cox**, **D. R.**...

D. R. Cox (1972). Regression models and life-tables. J. R. Stat. Soc. Ser. B. Stat. Methodol. 34, 187–220

Cox's proportional hazard model

$$\log \frac{h_{x_{ij}}(t)}{h_0(t)} = \sum_{j=1}^{p} \beta_j x_{ij}$$

Let be two individuals i and k with covariates $\left\{ x_{ij} \right\}$ and $\left\{ x_{kj} \right\}$

$$\frac{h_{x_{ij}}(t)}{h_{x_{kj}}(t)} = e^{\sum_{j=1}^{p} \beta_j (x_{ij} - x_{kj})}$$

Interpretation of the coefficients

Let be two individuals i and k with covariates $\left\{x_{ij}\right\}$ and $\left\{x_{kj}\right\}$

$\left\{x_{ij}\right\}$ and $\left\{x_{kj}\right\}$ are only different at x_{1k} and x_{2k} in one unit

$$\frac{h_{x_{ij}}(t)}{h_{x_{kj}}(t)} = e^{\beta_j}$$

Relative risk when one changes one unit in covariate k while maintaining the remaining covariates fixed

Estimation

$$t_{(1)} < \ldots < t_{(k)}, k < n$$

$$R_i = R(t_{(i)}) = \left\{ j : t_j \geq t_{(i)} \right\}$$

$$D = \left\{ i : t_{(i)} \right\}$$

Maximisation of the following function

$$L(\beta_1, \ldots, \beta_p) = \prod_{i \in D} \frac{e^{\sum_{j=1}^{p} \beta_j x_{(i)j}}}{\sum_{l \in R_i} e^{\sum_{j=1}^{p} \beta_j x_{lj}}}$$

(numerical methods)

A theoretical note

$$L(\beta_1, \ldots, \beta_p) = \prod_{i \in D} \frac{e^{\sum_{j=1}^{p} \beta_j x_{(i)j}}}{\sum_{l \in R_i} e^{\sum_{j=1}^{p} \beta_j x_{lj}}}$$

Partial likelihood (because it is independent of the baseline hazard function)

This is not a likelihood function in the strict sense as it does not represent the probability of a given event.

A theoretical note

True likelihood

$$h(t) = \frac{f(t)}{S(t)} \Leftrightarrow f(t) = S(t) \times h(t)$$

$$L(\beta_1, \cdots \beta_p, h_0(t)) = \prod_i (h_0(t) e^{\sum_j \beta_j x_{ij}} S_0(t))$$

$$L(\beta_1, \ldots, \beta_p; h_0(t)) = L(\beta_1, \ldots, \beta_p) \times \prod_{i \in D} \left(h_0(t) \sum_{l \in R_i} e^{\sum_j \beta_j x_{lj}} \right) \times \prod_{i=1}^{n} S_0(t)^{exp\left\{ \sum_j x_{ij} \right\}}$$

Exercise 2: rituximab clinical trial data

Fit a Cox's proportional hazard model including the following covariates

(i) gender
(ii) age
(iii) history of autoimmune diseases
(iv) disease duration

Draw your conclusions.

Model selection and comparison

$M_1 \subset M_2$

$$M_1 : \log \frac{h_{x_{ij}}(t)}{h_0(t)} = \beta_1 x_{i1} + \cdots + \beta_1 x_{ip}$$

$$M_2 : \log \frac{h_{x_{ij}}(t)}{h_0(t)} = \beta_1 x_{i1} + \cdots + \beta_1 x_{ip} + \beta_1 x_{i,p+1} + \cdots + \beta_1 x_{i,p+m}$$

Are AIC or BIC applicable?

Analysis of Residuals

Cox-Snel Residual

$$r_i = e^{\hat{\beta}_1 x_{i1} + \cdots \hat{\beta}_p x_{ip}} \hat{H}_0(t)$$

$r_i \rightsquigarrow$ Exponential(1)

$\hat{H}_0(t)$ is the estimated cumulative baseline hazard

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

$$\hat{H}_0(t) = -\log \hat{S}_0(t)$$

$$H(t) = -\log S(t)$$