Biostatistics

Applications in Medicine

Nuno Sepúlveda, 17.11.2025

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 - Weibull regression, Kaplan-Meier curve, log-rank test, Cox's proportional hazards model

3. Applications in Genetics, Genomics, and other 'omics data

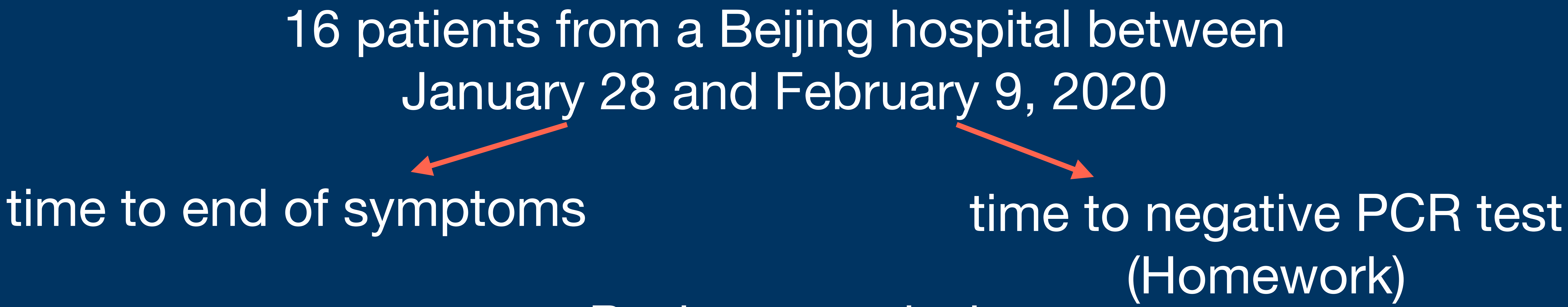
    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values, estimation of biological age
    c. Gene expression studies based on RNA-seq experiments – Tests based 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

Exercise:

data about recovery from a SARS-CoV-2 infection



16 patients from a Beijing hospital between
January 28 and February 9, 2020

time to end of symptoms

time to negative PCR test
(Homework)

Package survival

Fit a Weibull regression model with time to end of symptoms as the outcome and age and gender as the covariate

Assess the validity of the model by testing a Gumbel distribution in the residuals

Chang D, Mo G, Yuan X, et al. Time Kinetics of Viral Clearance and Resolution of Symptoms in Novel Coronavirus Infection. Am J Respir Crit Care Med. 2020;201(9):1150-1152.

Parametric analysis

versus

Non-parametric analysis

Parametric analysis

Complete data

Incomplete data

Non-parametric analysis

Complete data
Incomplete data

Non-parametric methods

Comparison of different survival curves

Log-rank test
Peto-Peto test

Kolmogorov-Smirnov test

Semi-parametric regression

Cox's proportional hazard model

Comparison of different survival curves

Two treatments under comparison

Time to clinical response

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

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

Do you know other tests

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

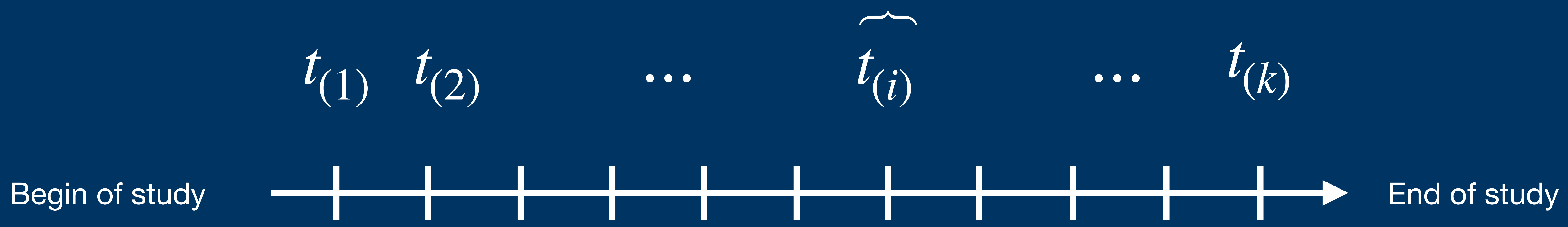
$H_0 : \Delta_1 = \cdots = \Delta_K = 1\,(t)$ versus $H_1 : \exists_{i,j} \Delta_i \neq \Delta_j = 1$

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

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

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

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)

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

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

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

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

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

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

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

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

Decision rule

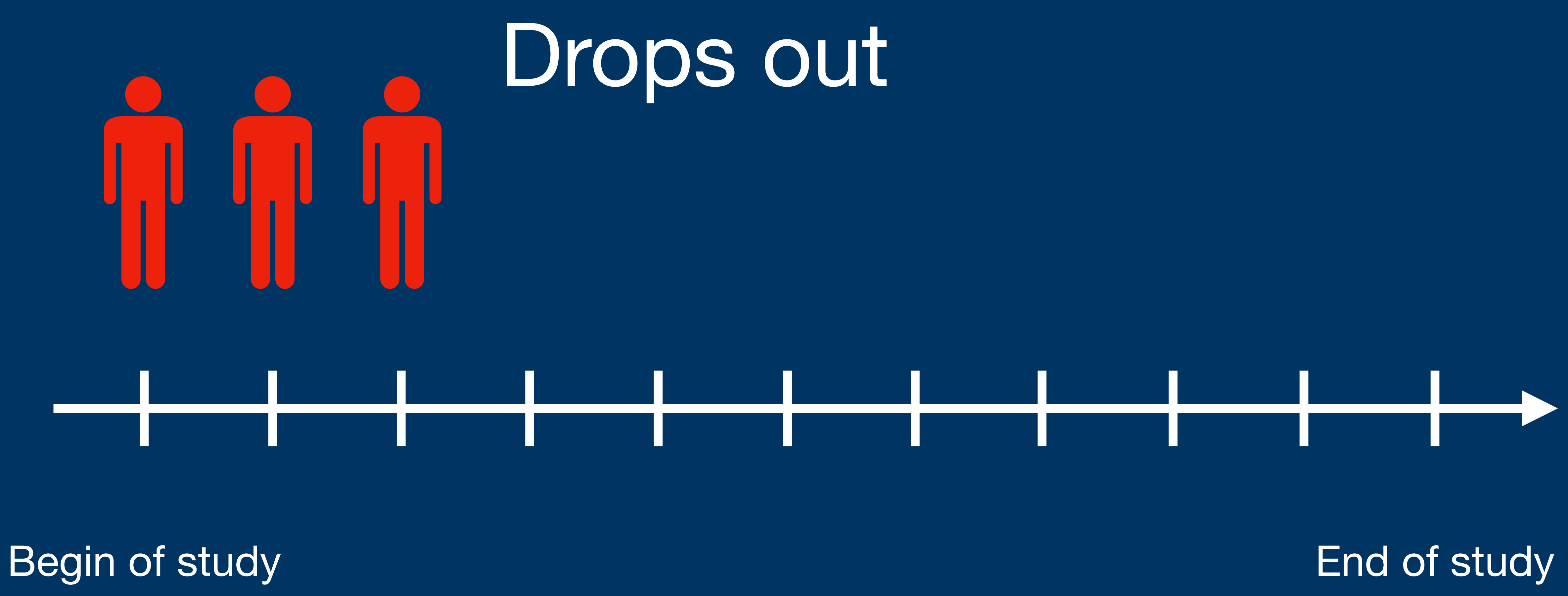
$$\begin{cases} \text{do not reject } H_0, & \text{if } p > \alpha \\ \text{reject } H_0, & \text{otherwise} \end{cases}$$

Exercise 1: rituximab clinical trial data

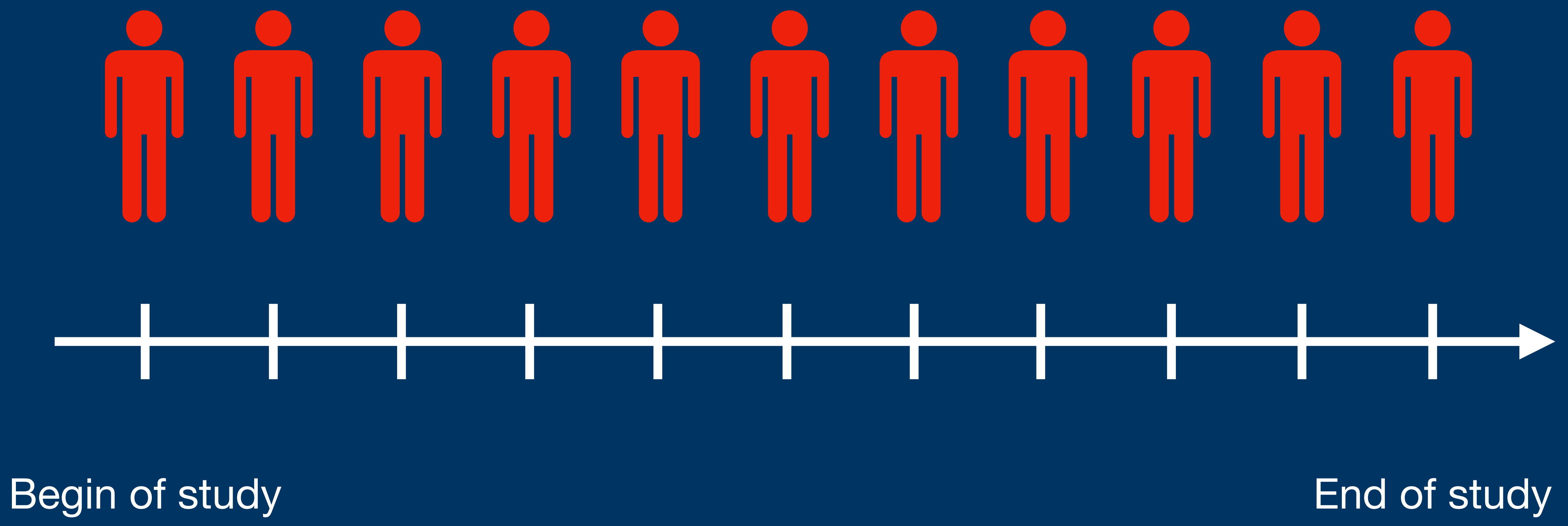
Compare the survival curves for time to first response for men versus women using log-rank (survdiff function from survival package)

Draw your conclusions.

What type of incomplete data?



What type of incomplete data?



Exercise 2: rituximab clinical trial data

Identify and quantify censored data concerning time to first treatment response

Should you consider interval censoring in this case?

Basic mathematical formulation of the problem

Right/left censored data

$$\{\mathrm{t}_i, d_i\}, i = 1,\ldots,n$$

$$\mathrm{t}_i = \begin{cases} t_i^*, & \text{if } t_i \text{ is right censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \end{cases}$$

$$\mathrm{t}_i = \begin{cases} t_i^+, & \text{if } t_i \text{ is left censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \end{cases}$$

Interval censored data

$$\{\mathrm{a}_i, b_i, d_i\}, i = 1,\ldots,n$$

$$\mathrm{a}_i = \begin{cases} t_i^*, & \text{if } t_i \text{ is interval censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \end{cases}$$

$$\mathrm{b}_i = \begin{cases} t_i^+, & \text{if } t_i \text{ is interval censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \end{cases}$$

$$d_i = \begin{cases} 0, & \text{if } t_i \text{ is censored} \\ 1, & \text{if } t_i \text{ is completely observed} \end{cases}$$

In practice

Package survival

Survival time

$$\text{time}_i = \begin{cases} t_i^*, & \text{if } t_i \text{ is right or interval censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \\ t_i^+, & \text{if } t_i \text{ is left censored} \end{cases}$$

Event indicator

$$d_i = \begin{cases} 0, & \text{if } t_i \text{ is right censored} \\ 1, & \text{if } t_i \text{ is completely observed} \\ 2, & \text{if } t_i \text{ is left censored} \\ 3, & \text{if } t_i \text{ is interval censored} \end{cases}$$

$$t_i \in \left(t_i^*, t_i^+\right)$$

$$\text{time2}_i = \begin{cases} t_i^+, & \text{if } t_i \text{ is interval censored} \\ 0, & \text{otherwise} \end{cases}$$

Likelihood function of a parametric model under different censoring mechanisms

$$T_i \,|\, \theta \rightsquigarrow F(\theta)$$

Weibull, Gamma, Lognormal, Log-logistic, etc

Right censored data

$$L\left(\theta \,|\, \{t_i, d_i\}\right) \equiv \prod_{i=1}^{n} f_\theta(t_i)^{d_i} S_\theta(t_i)^{1-d_i}$$

Left censored data

$$L\left(\theta \,|\, \{t_i, d_i\}\right) \equiv \prod_{i=1}^{n} f_\theta(t_i)^{d_i} F_\theta(t_i)^{1-d_i}$$

Interval censored data

$$L\left(\theta \,|\, \{a_i, b_i, d_i\}\right) \equiv \prod_{i=1}^{n} f_\theta(a_i)^{d_i} \left(F_\theta(b_i) - F_\theta(a_i)\right)^{1-d_i}$$

Parametric estimation

$$\hat{\theta} = \underset{\theta}{\text{argmax}} L\left(\theta \,|\, \left\{t_i, d_i\right\}\right)$$

No closed-form expressions

Numerical solutions for the maximum likelihood equations

Exercise: rituximab clinical trial data

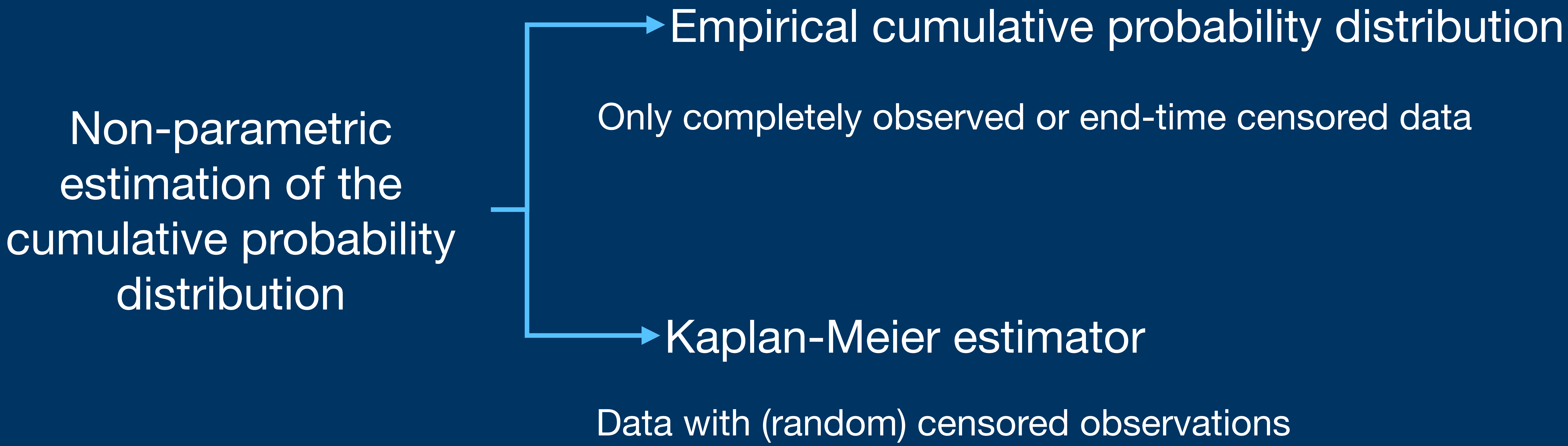
Estimate lognormal, weibull and log-logistic model to data on time to treatment response using the “survreg” function of package survival.

What is the best model for the data?

Checking the adequacy of a parametric model

Can we use the Kolmogorov-Smirnov test directly to data?

Kaplan-Meier estimator for the survival function



Non-parametric estimation of the cumulative probability distribution

Empirical cumulative probability distribution

Only completely observed or end-time censored data

Kaplan-Meier estimator

Data with (random) censored observations

Kaplan-Meier estimator for the survival function

$$\hat{S}(t) = \prod_{i:t_{(}i)\leq t} \left(1 - \frac{d_i}{n_i}\right) \qquad t \in (0, t_{\max})$$

$d_i =$ number of individuals in which the event was observed at $t_{(i)}$

$n_i =$ number of individuals without the event of interest at $t_{(i-1)}$

$\left\{ t_{(i)}, i = 1, \ldots, r \right\} =$ unique times when the event of interest was observed

Kaplan-Meier estimator for the survival function

$$\hat{S}\left(t_{(1)}\right) = 1 - \frac{d_1}{n_1}$$

$n_1 =$ number of individuals without the event of interest at time $0 = n$

$$\hat{S}\left(t_{(i)}\right) = \hat{S}\left(t_{(i-1)}\right)\left(1 - \frac{d_i}{n_i}\right)$$

Exercise 3: rituximab clinical trial data

Estimate the survival curve of time to treatment response using the Kaplan-Meier estimator

Compare the Kaplan-Meier estimated survival curve to the survival curve predicted by the best parametric model from Exercise 2.