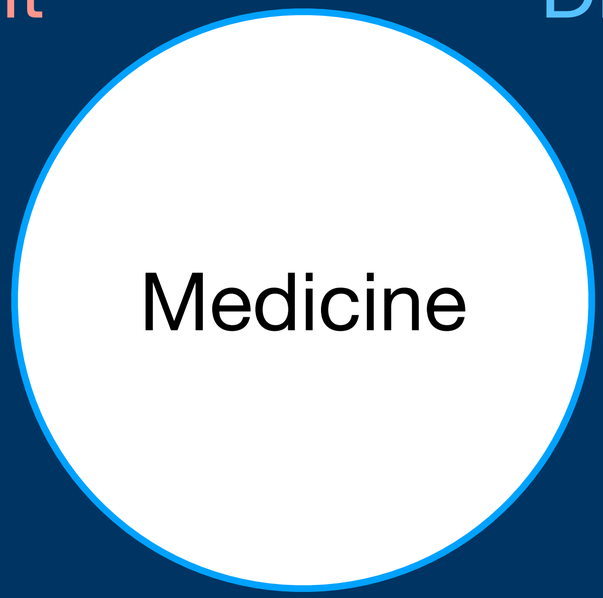
Biostatistics

Applications in Medicine

Nuno Sepúlveda, 13.10.2023

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



Prevent

Diagnose

Improve

Treat

Develop

Survival or time-to-event analysis

Endpoint: time to event

Examples of endpoints

time to death in cancer patients (hence, survival analysis)

time to first symptomatic infection after vaccination

time to hospital discharge

time to a positive diagnosis of a chronic disease

time to clearance of infection

What parametric distributions could be used to analyse this random variable?

T = random variable that represents the time when the event of interest occurs

$$T \rightsquigarrow \ ?$$

Survival function

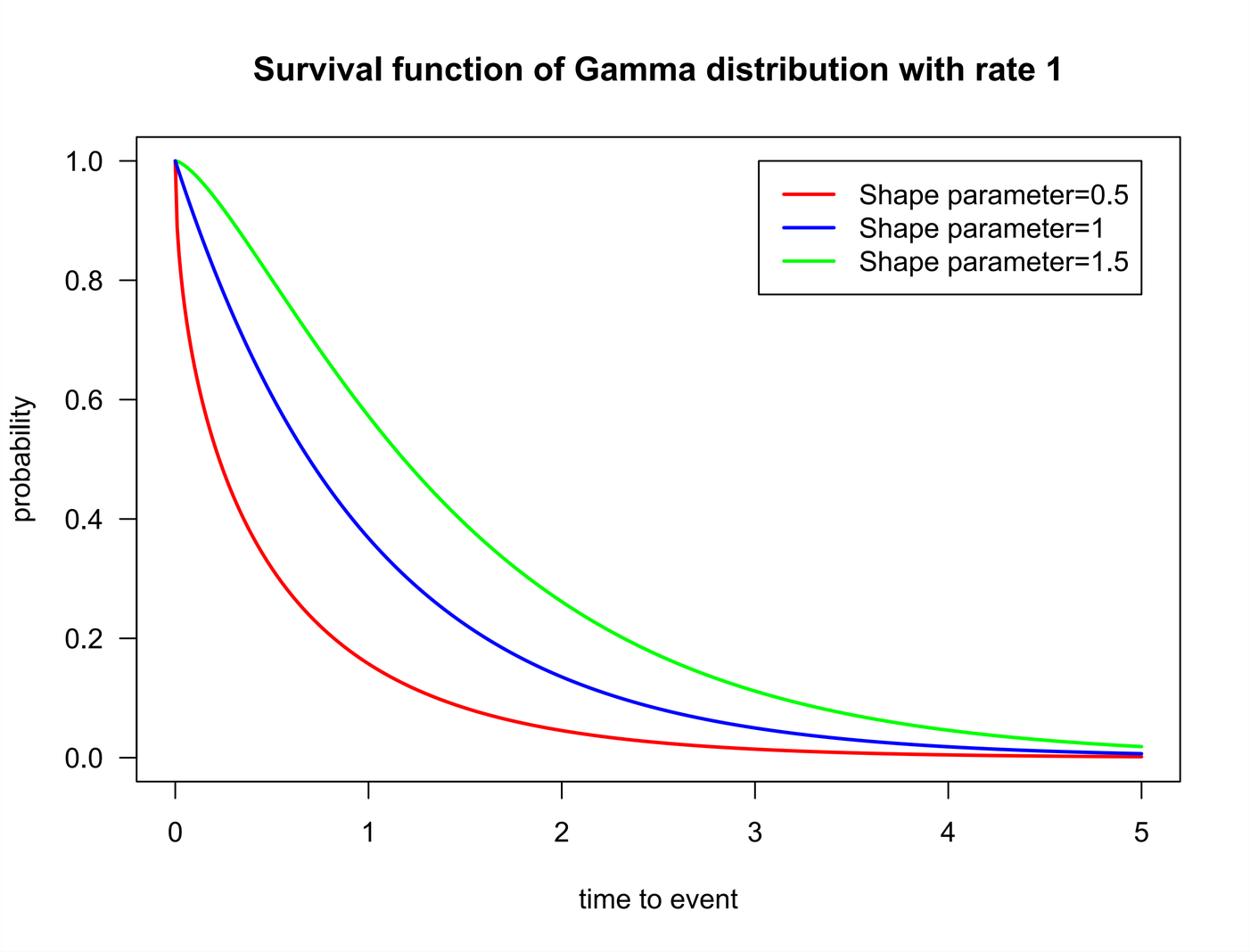
$$S_\theta(t) = P\left(T > t \,|\, \theta\right),\ t \geq 0$$

$$S(t) = 1 - F_{T,\theta}(t),\ t \geq 0$$

S is strictly a decreasing (continuous) function

$S_\theta(0) = 1$ and $S_\theta(+\infty) = 0$

Example


Survival function of Gamma distribution with rate 1
Shape parameter=0.5
Shape parameter=1
Shape parameter=1.5
probability
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
time to event

Hazard function (formal definition)

$$h_\theta(t) = \lim_{dt \to 0+} \frac{P[t \leq T < t + dt \,|\, T \geq t]}{dt}$$

“Instant” risk of the event occurring at time t

Hazard function (more practical definitions)

$$h_\theta(t) = \frac{f(t)}{S(t)}$$

$$f_\theta(t) = \lim_{dt \to 0+} \frac{P[t \leq T < t + dt]}{dt} = -S'(t)$$

$$h(t) = -\frac{S'(t)}{S(t)} \Leftrightarrow S(t) = e^{-\int_0^t h(x)dx}$$

(by the fundamental theorem of calculus)

Exercise 0

Use the definition of hazard function and plot the hazard functions of the following distributions:

Exponential distribution with rate parameter =1

Gamma distribution with shape parameter = 0.5 and rate parameter =1

Gamma distribution with shape parameter = 1.5 and rate parameter =1

What is your interpretation of these hazard functions?

Discussion

What is the qualitative aspect of the hazard function for time to death in humans?

Exercise 1: COVID19

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

time to end of symptoms

time to negative PCR test

Package MASS

Fit exponential, gamma, lognormal, and weibull distributions to each endpoint

Select the best model to each endpoint and plot the corresponding survival and hazard functions

Compare the survival and hazard functions and draw your conclusions

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.

Weibull distribution

$$f_{\gamma,\lambda}(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1} e^{-(t/\lambda)^{\gamma}},\ t > 0$$

$$F_{\gamma,\lambda}(t) = 1 - e^{-(t/\lambda)^{\gamma}},\ t > 0$$

$$h_{\gamma,\lambda}(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1},\ t > 0$$

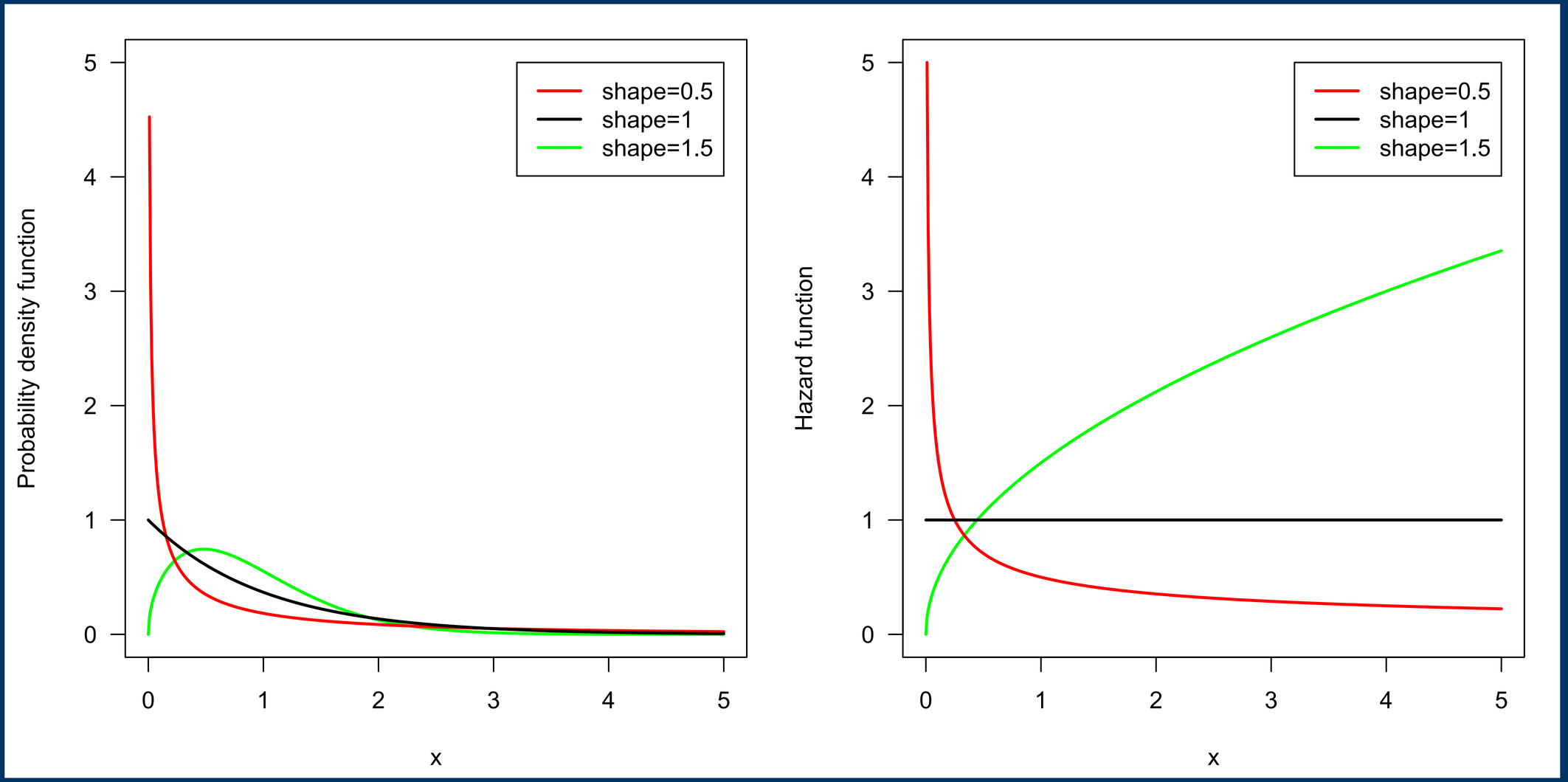
Shape parameter

$\gamma \in (0, +\infty)$

Scale parameter

$\lambda \in (0, +\infty)$

Weibull distribution in Survival Analysis



Weibull distribution and its relationship with the Exponential distribution

$$T \mid \lambda \rightsquigarrow \text{Exponential}(\lambda) \Rightarrow X^{\gamma} \rightsquigarrow \text{Weibull}(\gamma, \lambda)$$

$$T \rightsquigarrow \text{Exponential}(1) \Rightarrow \left(\frac{X}{\lambda}\right)^{\gamma} \rightsquigarrow \text{Weibull}(\gamma, \lambda)$$

$$\gamma = 1 \Rightarrow T \mid \lambda \rightsquigarrow \text{Exponential}(\lambda)$$

Why is this relationship important?

Weibull distribution and its relationship with the Gumbel distribution

$$T|\gamma,\lambda \rightsquigarrow \text{Weibull}(\gamma,\lambda) \Rightarrow \log T|\gamma,\lambda \rightsquigarrow \text{Gumbel}(\mu=\frac{\log\lambda}{\gamma},\sigma=\frac{1}{\gamma})$$

$$T|\mu,\sigma \rightsquigarrow \text{Gumbel}(\mu,\sigma) \Rightarrow e^{T}|\mu,\sigma \rightsquigarrow \text{Weibull}(\lambda=e^{\frac{\mu}{\sigma}},\gamma=\frac{1}{\sigma})$$

Why is this relationship important?

How to assess the adequacy of the Weibull distribution in a given data set?

Visualisation method

Do you know any method?

How to assess the adequacy of the Weibull distribution in a given data set?

Visualisation method

$$F_{\gamma,\lambda}(t) = 1 - e^{-(t/\lambda)^\gamma}$$

$$1 - F_{\gamma,\lambda}(t) = e^{-(t/\lambda)^\gamma}$$

$$\log(1 - F_{\gamma,\lambda}(t)) = -(t/\lambda)^\gamma$$

$$\log(-\log(1 - F_{\gamma,\lambda}(t))) = -\gamma \log \lambda + \gamma \log t$$

$t_1, \ldots, t_n$ $\hat{\gamma}$ and $\hat{\lambda}$

$\hat{F}(t_i)$ = empirical cumulative distributions

Make the plot

$\log t_i$ versus $\log(-\log(1 - \hat{F}(t_i)))$

Interpretation:
If the Weibull distribution fits well the data, the plot should look like a straight line

How to assess the adequacy of the Weibull distribution in a given data set?

Formal Hypothesis testing

Kolmogorov-Smirnov test

What are the null and alternative hypotheses?

What is the decision rule of the test?

Eventual problems?

Exercise 2: COVID19

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

time to end of symptoms

time to negative PCR test (Homework)

Assess the adequacy of the Weibull distributions to model “time to end of symptoms” using the visualisation method and a formal hypothesis testing

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.

Weibull regression model

Log-linear formulation (similar to linear regression)

$$\log T_i = \beta_0 + \sum_j \beta_j x_{ij} + \sigma_0 \epsilon_i$$

$$\epsilon_i \,|\, \rightsquigarrow \text{Gumbel}(\mu = 0, \sigma = 1)$$

$$\log T_i \rightsquigarrow \text{Gumbel}\left(\mu = \beta_0 + \sum_j \beta_j x_j, \sigma = \sigma_0\right)$$

(see slide 17)

$$T_i \rightsquigarrow \text{Weibull}\left(\lambda = \exp\left\{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}\right\}, \gamma = \frac{1}{\sigma}\right)$$

Weibull regression model as a proportional hazard model

$$T_i \rightsquigarrow \text{Weibull}\left(\lambda = \exp\left\{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}\right\}, \gamma = \frac{1}{\sigma}\right)$$

$$h_{\gamma,\lambda}(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1}, \; t > 0$$

$$h_{\gamma,\left\{\beta_j\right\}}(t) = \frac{1}{\sigma e^{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}}}\left(\frac{t}{e^{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}}}\right)^{\frac{1}{\sigma}-1}$$

Weibull regression model as a proportional hazard model

$$h_{\gamma,\left\{\beta_j\right\}}(t) = \frac{1}{\sigma e^{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}}} \left(\frac{t}{e^{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}}} \right)^{\frac{1}{\sigma}-1}$$

$$= \frac{1}{\sigma e^{\frac{\beta_0}{\sigma}}} \left(\frac{t}{e^{\frac{\beta_0}{\sigma}}} \right)^{\frac{1}{\sigma}-1} \left(\frac{1}{e^{\frac{\sum_j \beta_j x_j}{\sigma}}} \right)^{\frac{1}{\sigma}}$$

$$= \frac{1}{\sigma e^{\frac{\beta_0}{\sigma}}} \left(\frac{t}{e^{\frac{\beta_0}{\sigma}}} \right)^{\frac{1}{\sigma}-1} e^{-\frac{\sum_j \beta_j x_j}{\sigma^2}}$$

Estimation and statistical validation

Maximum likelihood estimation using numerical methods (e.g., Newton-Raphson)

$$\left\{\hat{\beta}_j, j = 0, \ldots, p\right\}, \hat{\sigma}$$

Validation of the model

Standardized residuals: $\hat{e}_i = \dfrac{\log t_i - \hat{\log} t_i}{\hat{\sigma}}$

they should follow a Gumbel distribution with mu=0 and sigma=1

Cox-Snel residuals: $\tilde{e}_i = \left(t_i e^{-\hat{\log} t_i}\right)^{1/\hat{\sigma}}$

they should follow a Exponential distribution with parameter 1 (see slide 16)

Weibull regression model is not a generalized linear model

Weibull distribution does not belong to the exponential family of distributions.

Homework!

Exercise 3: COVID19

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.