

# Biostatistics

## Applications in Medicine

Nuno Sepúlveda, 13.11.2023

# 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

Prevent

Diagnose

Medicine

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

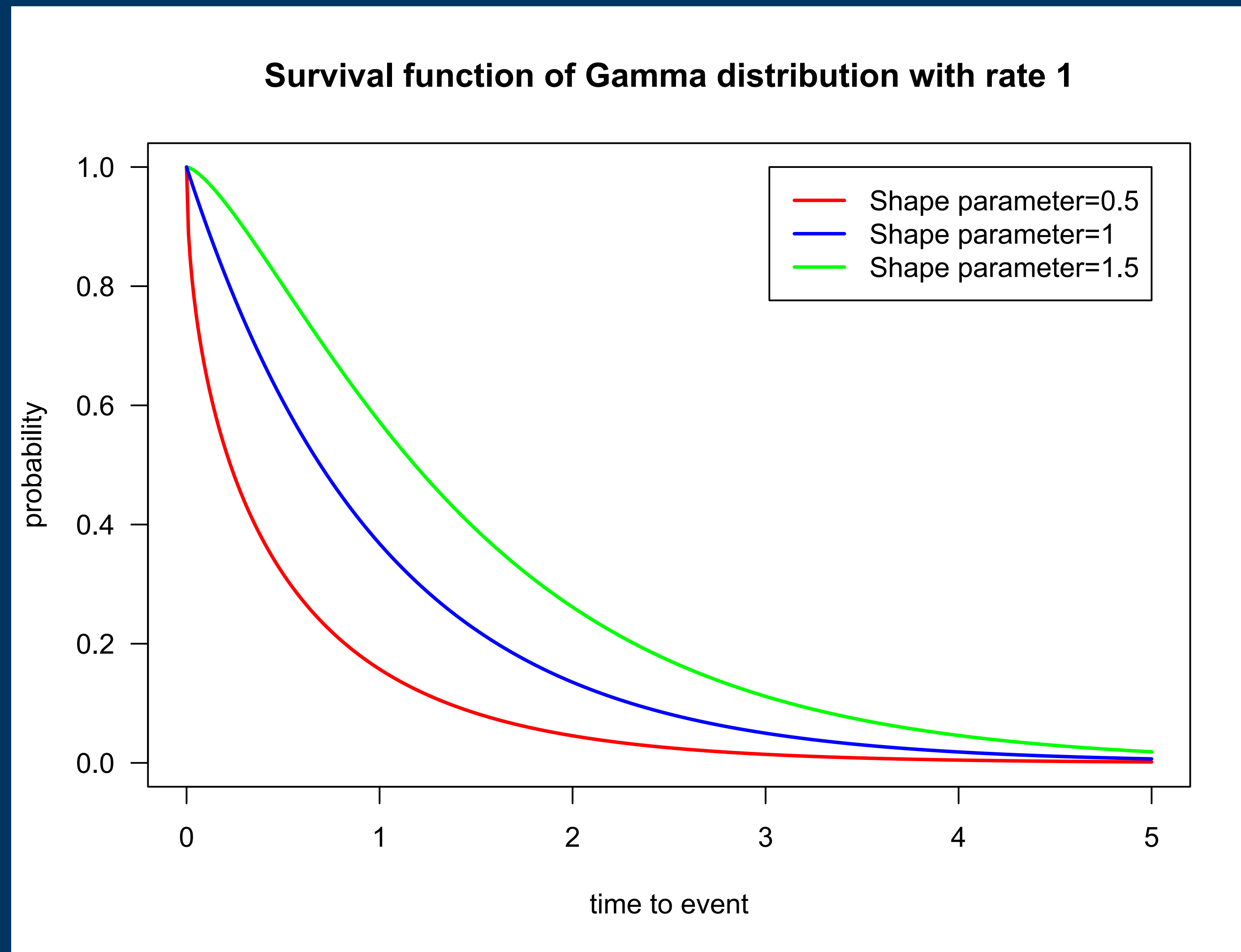
$$S(t) = P(T > t), \quad t \geq 0$$

$$S(t) = 1 - F(t), \quad t \geq 0$$

$S$  is strictly a decreasing (continuous) function

$$S(0) = 1 \text{ and } S(+\infty) = 0$$

# Example





## Hazard function (formal definition)

$$h(t) = \lim_{dt \rightarrow 0+} \frac{P[t \leq T < t + dt \mid T \geq t]}{dt}$$

“Instant” risk of the event occurring at time  $t$

## Hazard function (more practical definition)

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

Hazard function is simply the ratio between the probability density function and the survival function

## Two interesting relationships between probability density function, survival function and hazard function

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

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

(by the fundamental theorem of calculus)

## Exercise 0

Use the practical 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: 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

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

## Weibull distribution

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

Shape parameter

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

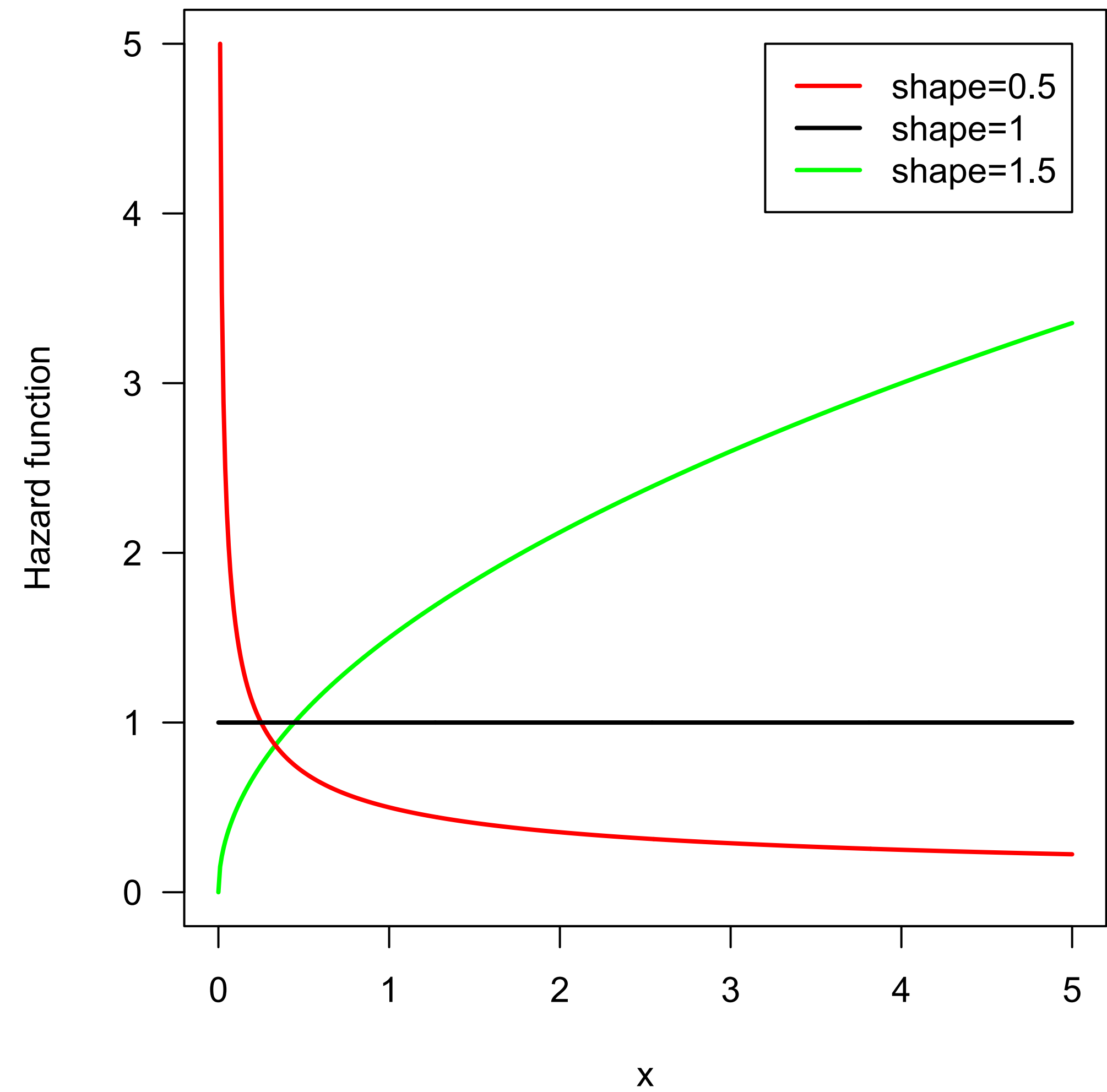
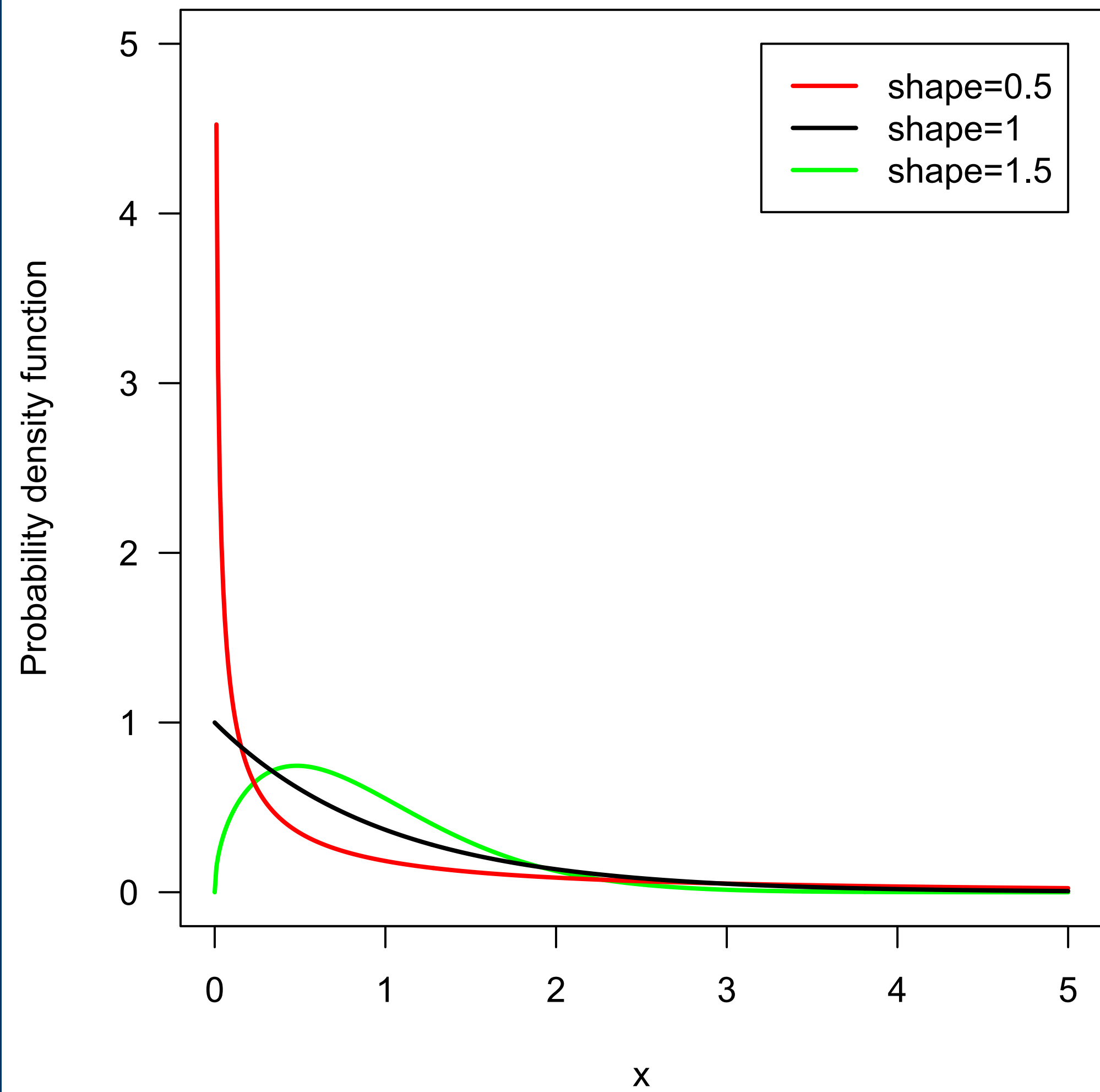
$$F_{\gamma,\lambda}(t) = 1 - e^{-(t/\lambda)^\gamma}, \quad t > 0$$

Scale parameter

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

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

# Weibull distribution in Survival Analysis





# Weibull distribution and its relationship with the Exponential distribution

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

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

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

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

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

$$t_1, \dots, t_n \quad \hat{\gamma} \text{ and } \hat{\lambda}$$

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

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

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

Make the plot

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

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

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: 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)

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