Biostatistics

Applications in Medicine

Nuno Sepúlveda, 10.11.2025

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, log-rank test, Kaplan-Meier curve, 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

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

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

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

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

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

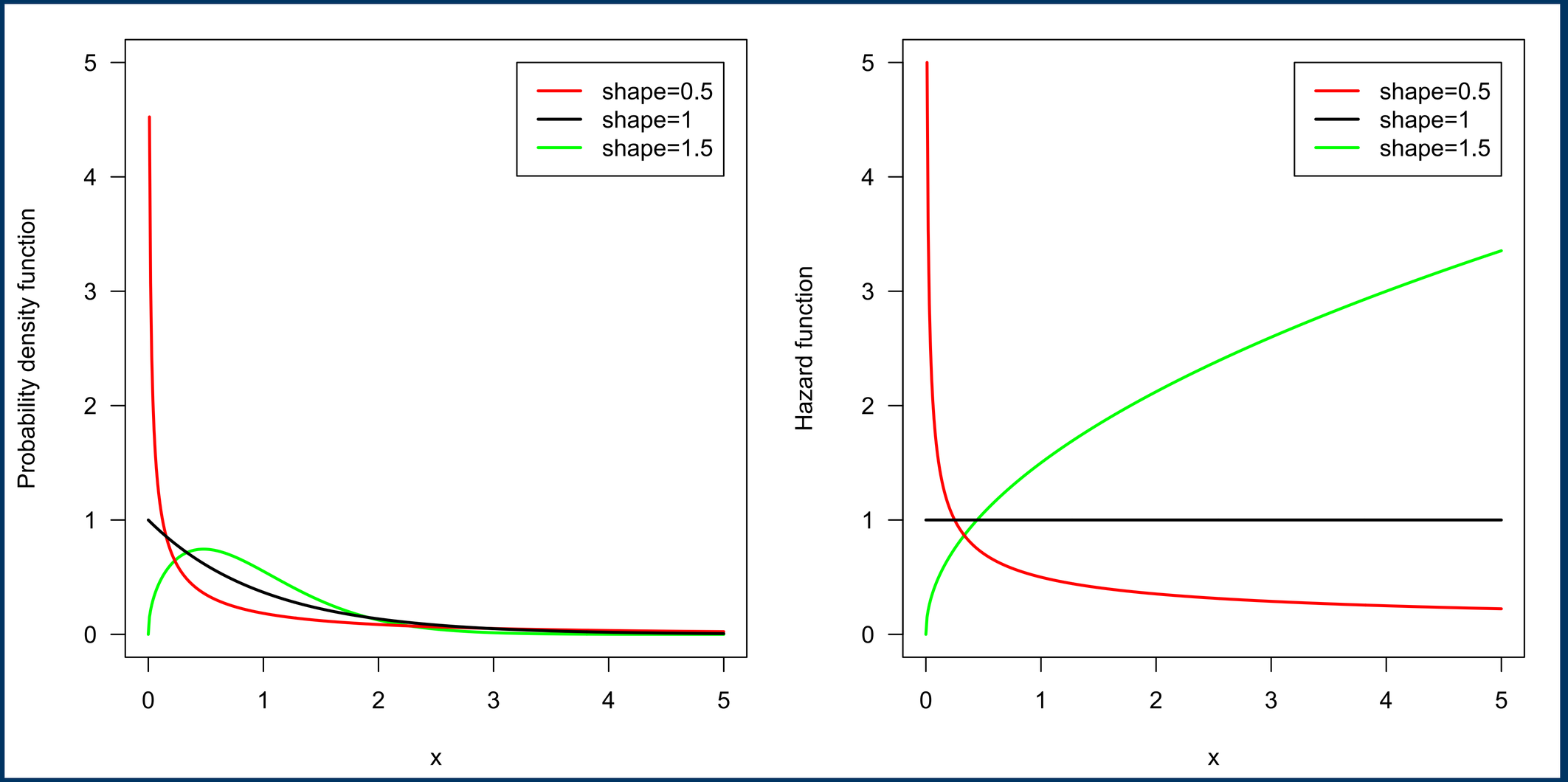
Shape parameter

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

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:

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

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)

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

Weibull regression model as a proportional hazard model

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

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

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

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 7)

Weibull regression model is not a generalized linear model

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

Homework!

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

Incomplete data

Complete data

Non-parametric analysis

Incomplete data

Complete 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

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

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

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

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

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)

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

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

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

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

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

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

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

Decision rule

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

Exercise:
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

Compare survival curves of males versus females using Kolmogorov-Smirnov test and log-Rank test. Compare the results.

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.