axioms

MDPI

Article

Proportional Odds Hazard Model for Discrete Time-to-Event Data

Maria Gabriella Figueiredo Vieira [1], Marcílio Ramos Pereira Cardial [2], Raul Matsushita [1] and Eduardo Yoshio Nakano [1,*]

[1] Department of Statistics, University of Brasilia, Campus Darcy Ribeiro, Asa Norte, Brasília 70910-900, Brazil; raulmta@unb.br (R.M.)

[2] Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos 13566-590, Brazil; marciliocardial@usp.br

* Correspondence: nakano@unb.br

Abstract: In this article, we present the development of the proportional odds hazard model for discrete time-to-event data. In this work, inferences about the model's parameters were formulated considering the presence of right censoring and the discrete Weibull and log-logistic distributions. Simulation studies were carried out to check the asymptotic properties of the estimators. In addition, procedures for checking the proportional odds assumption were proposed, and the proposed model is illustrated using a dataset on the survival time of patients with low back pain.

Keywords: proportional odds regression model; discrete censored data; survival analysis

MSC: 62N01

check for updates

Citation: Vieira, M.G.F.; Cardial, M.R.P.; Matsushita, R.; Nakano, E.Y. Proportional Odds Hazard Model for Discrete Time-to-Event Data. *Axioms* **2023**, *12*, 1102. https://doi.org/10.3390/axioms12121102

Academic Editor: Jean Renaud Pycke

Received: 31 October 2023
Revised: 28 November 2023
Accepted: 1 December 2023
Published: 6 December 2023

1. Introduction

The proportional hazards model [1] is a regression model widely used in survival data analysis, whose main characteristic is that the covariates act multiplicatively on the hazard function. However, this characteristic cannot be met when survival times are discrete (intrinsically discrete or grouped into intervals) since the hazard function is limited in the interval (0,1). According to [2], the use of statistical methods that are specially designed for discrete times has many advantages. Indeed, [3] illustrated through simulation studies and application to real data that it is inadvisable to use a continuous model to analyze discrete data.

Given this situation and the importance of correctly treating discrete data to effectively model discrete time-to-event survival data, since the aforementioned model would not always be the most suitable, the proportional odds hazard model has been used with some frequency in the literature for this purpose. This model is an alternative version proposed by [1] to be used when the time-to-event data are discrete, with the covariates having a multiplicative effect on the odds hazard.

A comprehensive study of the model in which various link functions are considered is presented in [2], and the semiparametric extensions that the model can take on in [4]. Applications of this model are given in [5–8].

The popularization of this model is due, in part, to the fact that users do not invest effort in reporting the baseline hazard, which receives less attention in these studies. However, according to [9], the behavior of the hazard function is of potential medical interest because it is directly related to the course of a disease. To estimate this hazard function informatively (i.e., smoothly), a parametric model may be appropriate. In this context, parametric models in which the response variable is discrete to inform the baseline hazard of the model efficiently become fundamental, and in recent years a large number of research articles dealing with discrete distributions arising from the discretization of distributions of continuous random variables in a survival analysis context have emerged among these are: discrete Weibull distribution (DW) in [10,11], discrete Weibull geometric in [12], exponentiated discrete Weibull (EDW) in [13], discrete Gumbel in [14], discrete Burr in [15] and discrete log-logistic in [16].

This work aims to formulate the proportional odds hazard model considering the discrete Weibull and discrete log-logistic distributions as baseline distributions, as well as the estimation via maximum likelihood of the model's parameters for right-censored data. The Weibull distribution was chosen due to its popularity in modeling survival data and the log-logistic distribution, allowing model data with non-monotonic hazards. The quality of the model's fit was assessed using simulation studies. Finally, the proposed methodology was illustrated using a data set whose response variable is the number of unsuccessful sessions before pain relief or reduction in patients with low back pain [17].

2. Discrete Random Variables for Time-to-Event Data

Let T be a discrete random variable that takes on non-negative integer values ($T = 0, 1, 2, \ldots$), whose distribution function, survival function and hazard function are defined, respectively, by $p(t) = P(T = t), \quad S(t) = P(T > t)$ and $h(t) = P(T = t|T \geq t), \quad t = 0, 1, \ldots$. Other relationships can be established from the functions mentioned, such as:

$$S(t) \quad = \quad P(T > t) = \left(\sum_{k=\lfloor t \rfloor + 1}^{\infty} P(T = k) \right) \mathbb{I}_{\{t=0,1,2,\ldots\}}, \tag{1}$$

$$h(t) = \left(\frac{p(t)}{S(t) + p(t)} \right) \mathbb{I}_{\{t=0,1,2,\ldots\}}, \tag{2}$$

$$\begin{aligned} p(t) &= \left([1 - S(0)]^{\mathbb{I}_{\{t=0\}}} [S(t-1) - S(t)]^{\left(1 - \mathbb{I}_{\{t=0\}}\right)} \right) \mathbb{I}_{\{t=0,1,2,\ldots\}} \\ &= \left([h(0)]^{\mathbb{I}_{\{t=0\}}} [h(t) S(t-1)]^{\left(1 - \mathbb{I}_{\{t=0\}}\right)} \right) \mathbb{I}_{\{t=0,1,2,\ldots\}} \end{aligned} \tag{3}$$

and

$$S(t) = \left(\prod_{k=0}^{\lfloor t \rfloor} [1 - h(k)] \right) \mathbb{I}_{\{t=0,1,2,\ldots\}}. \tag{4}$$

In Equations (1) and (4), $\lfloor t \rfloor$ denotes the largest integer less than or equal to t. More details on the functions and relationships presented can be found in [2].

2.1. Discrete Weibull Distribution (DW)

The discrete Weibull distribution (DW) was first proposed by [10]. Denoted by $T \sim DW(q, \eta)$, $0 < q < 1$ and $\eta > 0$, its probability function is given by:

$$p_{dw}(t|q, \eta) = (q^{t^\eta} - q^{(t+1)^\eta}) \mathbb{I}_{\{t=0,1,2,\ldots\}}. \tag{5}$$

The survival and hazard functions of the DW, obtained from Equations (1) and (2), are expressed respectively by:

$$S_{dw}(t|q, \eta) \quad = \quad \left(q^{(\lfloor t \rfloor + 1)^\eta} \right) \mathbb{I}_{\{t \geq 0\}} \tag{6}$$

and

$$h_{dw}(t|q, \eta) \quad = \quad \left(\frac{q^{t^\eta} - q^{(t+1)^\eta}}{q^{t^\eta}} \right) \mathbb{I}_{\{t=0,1,2,\ldots\}}. \tag{7}$$

In Equation (6), $\lfloor t \rfloor$ denotes the largest integer less than or equal to t.

According to [11], the DW hazard function has different shapes that are directly linked to its shape parameter η, i.e., when $\eta > 1$, the hazard function is strictly increasing; $\eta < 1$, the hazard function is strictly decreasing; $\eta = 1$, the hazard function is constant, in which case the DW is reduced to the geometric distribution, which is a discrete analog of the exponential distribution [3].

Discrete Log-Logistic Distribution (DLL)

Let T be a discrete random variable that follows a discrete log-logistic distribution (DLL) which is the discrete analog of the continuous log-logistic distribution, with some important results presented by [16], with parameters $\alpha > 0$ and $\eta > 0$, denoted by $T \sim DLL(\alpha, \eta)$, the probability, survival, and hazard function are given by:

$$p_{dll}(t|\alpha,\eta) = \left(\frac{1}{1+(t/\alpha)^{\eta}} - \frac{1}{1+[(t+1)/\alpha]^{\eta}}\right)\mathbb{I}_{\{t=0,1,2,\dots\}}, \tag{8}$$

$$S_{dll}(t|\alpha,\eta) \quad = \quad \left(\frac{1}{1+[(\lfloor t \rfloor+1)/\alpha]^{\eta}}\right)\mathbb{I}_{\{t\geq 0\}}, \tag{9}$$

and

$$h_{dll}(t|\alpha,\eta) \quad = \quad \left(1 - \frac{1+(t/\alpha)^{\eta}}{1+[(t+1)/\alpha]^{\eta}}\right)\mathbb{I}_{\{t=0,1,2,\dots\}}. \tag{10}$$

In Equation (9), $\lfloor t \rfloor$ denotes the largest integer less than or equal t. According to [14], the DLL is a particular case of the discrete Burr distribution studied by [15], which is the discrete analog of the continuous Burr distribution.

3. Materials and Methods

3.1. Proportional Odds Hazard Model for Discrete Time-to-Event

Let T be a discrete non-negative random variable that represents the time until the occurrence of the event of interest follows the proportional odds hazard model (POHM) if [1]:

$$\frac{h(t|\boldsymbol{z})}{1-h(t|\boldsymbol{z})} = \exp\{\boldsymbol{z}^{'}\boldsymbol{\beta}\}\frac{h_0(t)}{1-h_0(t)}, \tag{11}$$

where $h_0(\cdot)$ is the baseline hazard function and $\boldsymbol{\beta}^{'} = (\beta_1, \dots, \beta_p)$ is the vector of coefficients associated with the vector of covariates $\boldsymbol{z}^{'} = (z_1, \dots, z_p)$.

Note that the intercept β_0 does not appear in the linear predictor because the baseline hazard function, $h_0(t)$, absorbs this constant term. This model is a discrete version of the Cox proportional hazards model to cover the possibility of an appreciable number of draws.

From expression (11), it is possible to establish the hazard function in the presence of covariates:

$$h(t|\boldsymbol{z}) = \left(\frac{\exp\{\boldsymbol{z}^{'}\boldsymbol{\beta}\}h_0(t)}{1+(\exp\{\boldsymbol{z}^{'}\boldsymbol{\beta}\}-1)h_0(t)}\right)\mathbb{I}_{\{t=0,1,2,\dots\}}. \tag{12}$$

From (4) and (12), the survival function in the presence of covariates can be written as:

$$S(t|\boldsymbol{z}) = \left(\prod_{k=0}^{\lfloor t \rfloor}\left[\frac{1-h_0(k)}{1+(\exp\{\boldsymbol{z}^{'}\boldsymbol{\beta}\}-1)h_0(k)}\right]\right)\mathbb{I}_{\{t\geq 0\}}, \tag{13}$$

where, $\lfloor t \rfloor$ denotes the largest integer less than or equal to t.

Furthermore, using expressions (3) and (4), the probability function in the presence of covariates is:

$$\begin{aligned} p(t|\boldsymbol{z}) &= \left(\left[\frac{\exp\{\boldsymbol{z}'\boldsymbol{\beta}\} h_0(0)}{1 + (\exp\{\boldsymbol{z}'\boldsymbol{\beta}\} - 1) h_0(0)} \right]^{\mathbb{I}_{\{t=0\}}} \right. \\ &\times \left. \left[\frac{\exp\{\boldsymbol{z}'\boldsymbol{\beta}\} h_0(t)}{1 + (\exp\{\boldsymbol{z}'\boldsymbol{\beta}\} - 1) h_0(t)} \prod_{k=0}^{t-1} \left[\frac{1 - h_0(k)}{1 + (\exp\{\boldsymbol{z}'\boldsymbol{\beta}\} - 1) h_0(k)} \right] \right]^{(1-\mathbb{I}_{\{t=0\}})} \right) \mathbb{I}_{\{t=0,1,2,\ldots\}}. \qquad (14) \end{aligned}$$

To estimate the parameters of the proportional odds hazard model, consider an observed random sample $(t_1, t_2, \ldots, t_n)$ with its respective censoring indicators $(\delta_1, \delta_2, \ldots, \delta_n)$, where $\delta_i = 1$ if t_i is a failure time and $\delta_i = 0$ if is a right-censored time and $\boldsymbol{z}_{\boldsymbol{i}}' = (z_{i1}, z_{i2}, \ldots, z_{ip})$ the covariates vector of individual i, $i = 1, 2, \ldots, n$. The model's likelihood function, where $\boldsymbol{\xi}$ represents the vector of parameters of the baseline distribution, is given by:

$$\begin{aligned} L(\boldsymbol{\xi}, \boldsymbol{\beta}; \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{z}) &\propto \prod_{i=1}^{n} \left[\frac{\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} h_0(t_i|\boldsymbol{\xi})}{1 + (\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} - 1) h_0(t_i|\boldsymbol{\xi})} \prod_{k=0}^{t_i-1} \left[\frac{1 - h_0(k|\boldsymbol{\xi})}{1 + (\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} - 1) h_0(k|\boldsymbol{\xi})} \right] \right]^{(1-\mathbb{I}_{\{t_i=0\}})\delta_i} \\ &\times \left[\frac{\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} h_0(0|\boldsymbol{\xi})}{1 + (\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} - 1) h_0(0|\boldsymbol{\xi})} \right]^{\mathbb{I}_{\{t_i=0\}}\delta_i} \left[\prod_{k=0}^{t_i} \left[\frac{1 - h_0(k|\boldsymbol{\xi})}{1 + (\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} - 1) h_0(k|\boldsymbol{\xi})} \right] \right]^{(1-\delta_i)} \qquad (15) \end{aligned}$$

Applying the logarithm to the likelihood function (15), we get:

$$\begin{aligned} \ell(\boldsymbol{\xi}, \boldsymbol{\beta}; \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{z}) &= \sum_{i=1}^{n} \left\{ \left(1 - \mathbb{I}_{\{t_i=0\}}\right) \delta_i \left[\log \left[\frac{\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} h_0(t_i|\boldsymbol{\xi})}{1 + (\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} - 1) h_0(t_i|\boldsymbol{\xi})} \right] + \sum_{k=0}^{t_i-1} \log \left[\frac{1 - h_0(k|\boldsymbol{\xi})}{1 + (\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} - 1) h_0(k|\boldsymbol{\xi})} \right] \right] \right\} \\ &+ \sum_{i=1}^{n} \left\{ \mathbb{I}_{\{t_i=0\}} \delta_i \log \left[\frac{\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} h_0(0|\boldsymbol{\xi})}{1 + (\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} - 1) h_0(0|\boldsymbol{\xi})} \right] \right\} \\ &+ \sum_{i=1}^{n} \left\{ (1 - \delta_i) \sum_{k=0}^{t_i} \log \left[\frac{1 - h_0(k|\boldsymbol{\xi})}{1 + (\exp\{\boldsymbol{z}_{\boldsymbol{i}}'\boldsymbol{\beta}\} - 1) h_0(k|\boldsymbol{\xi})} \right] \right\} + c, \qquad (16) \end{aligned}$$

where c is a constant that does not depend on $\boldsymbol{\xi}$ and $\boldsymbol{\beta}$.

The likelihood equation is given by:

$$\boldsymbol{U}(\boldsymbol{\vartheta}) = \frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = 0. \qquad (17)$$

Thus, the value $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\beta}})$, that satisfies Equation (17), is the maximum likelihood estimator of the POHM, which under appropriate regularity conditions has a multivariate normal asymptotic distribution with mean $\boldsymbol{\vartheta}$ and variance and covariance matrix given by:

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\vartheta}}) = \left[- \frac{\partial^2 \ell(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^T} \bigg|_{\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}} \right]^{-1} = \left[-J(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}} \right]^{-1}. \qquad (18)$$

The $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\beta}})$ and the observed matrix $J(\boldsymbol{\vartheta})$ can be obtained numerically using computational optimization methods using the Newton-Raphson type algorithm, which provides an accurate numerical approximation for this matrix. From these results, it is possible to construct confidence intervals for the parameters and carry out significance tests on the POHM covariates.

When considering the model presented in (11), by assigning the baseline hazard function to the hazard function of DW (7), DW with $\eta = 1$ and DLL (10), we obtain the proportional odds hazard model: discrete Weibull (POHM-DW), geometric (POHM-G) and discrete log-logistic (POHM-DLL), which will be studied in the following subsections.

3.2. Verification of the Proportional Odds Hazard Assumption

The model proposed in (11), assumes that the odds hazard for two individuals are proportional. Considering a discrete non-negative random variable T and a dichotomous covariate z that takes on the values 0 and 1, the model assumes that:

$$\frac{h(t|z=1)}{1-h(t|z=1)} = \theta\frac{h(t|z=0)}{1-h(t|z=0)}, \tag{19}$$

where $h(\cdot)$ is the hazard function and θ is the proportionality constant that does not depend on t. Let $g_l(t)$ be the odds hazard function of an individual with covariate $z = l; l = 0, 1$, expressed by:

$$g_l(t) = \frac{h(t|z=l)}{1-h(t|z=l)}, \quad l = 0, 1. \tag{20}$$

The function $G_l(.)$ is, in turn, the cumulative odds hazard function given by:

$$G_l(t) = \sum_{u=0}^{t} g_l(u) = \sum_{u=0}^{t} \frac{h(u|z=l)}{1-h(u|z=l)}, \quad l = 0, 1. \tag{21}$$

Note that, under the assumption of odds proportional hazard, expressions (19) and (21), it follows that:

$$G_1(t) = \theta G_0(t). \tag{22}$$

Applying the logarithm to both sides of the equality in (22), we get:

$$\log(G_1(t)) = \log(\theta) + \log(G_0(t)). \tag{23}$$

Therefore, the relationship between $\log(G_1(t))$ and $\log(G_0(t))$ is a straight line with the angular coefficient, m_1, equal to 1 and the linear coefficient $m_0 = \log(\theta)$, i.e.,

$$\log(G_1(t)) = m_0 + m_1 \log(G_0(t)). \tag{24}$$

Thus, the assumption of proportional odds hazard can be verified graphically by fitting a simple regression line with an angular coefficient, m_1, equal to one (fixed). In this way we can plot the graph of points formed by the coordinates $(\log(G_0(t)), \log(G_1(t)))$, and the expected behavior is that the points formed by the coordinates are close to this regression line.

Graphical analysis is very informative, and for a given assessment for decision-making to be complete, it is advisable to have a measure of evidence. Thus, when considering expression (24), a hypothesis test can be used to check whether the odds hazards are proportional to each other. Thus, if $t_{(j)}$, with $j = 1, 2, \ldots, J$, is the j-th distinct time observed (censored or uncensored), the verification can be conducted by testing the hypothesis that the angular coefficient of the straight line is different from one ($m_1 \neq 1$). Thus, the hypotheses of interest are described by:

$$H_0 : m_1 = 1 \quad vs. \quad H_1 : m_1 \neq 1. \tag{25}$$

The statistical test of the hypothesis (25) is given by:

$$M = \frac{\widehat{m_1} - 1}{\sqrt{\frac{\sum_{j=1}^{J}(x_j - \bar{x})^2}{(J-2)\sum_{j=1}^{J}(y_j - \bar{y})^2}}}, \tag{26}$$

where $\widehat{m_1} = \dfrac{J\sum_{j=1}^{J} x_j y_j - \sum_{j=1}^{J} x_j \sum_{j=1}^{J} y_j}{J\sum_{j=1}^{J} x_j^2 - \left(\sum_{j=1}^{J} x_j\right)^2}$, $\bar{x} = \dfrac{\sum_{j=1}^{J} x_j}{J}$ and $\bar{y} = \dfrac{\sum_{j=1}^{J} y_j}{J}$ with $x_j = \log(O_0(t_j))$ and $y_j = \log(O_1(t_j))$. Assuming normality of $\log(O_1(t))$, M follows a Student's t distribution with $J-2$ degrees of freedom.

The procedures for checking the assumption of proportional odds hazard presented here can be easily extended to categorical covariates with three or more levels, comparing each level of the covariate two by two. In the case of numerical covariates, the same method can be adopted when categorizing the covariate to be verified.

In case of lack of proportionality, the POHM might not perform optimally. In these cases, other regression models for discrete data can be considered (see for examples Equations (27) and (28)).

4. Simulation Study

This section describes a simulation study to evaluate the behavior of the maximum likelihood estimators of the POHM-DW and POHM-G models. The study was conducted using data simulated in the R software [18], and the survival times of these models were generated using the inverse transformation method. For more details, see [19].

The survival time samples were simulated, considering two covariates: a numerical covariate, Z_1, with a standard normal distribution and a dichotomous covariate, Z_2, generated from a Bernoulli distribution with a probability of success $p = 0.5$, the various parameters used take into account the baseline hazard of a WD and geometric distribution (particular case of WD considering $\eta = 1$), more specifically the parameters of the two scenarios are shown in Table 1.

Table 1. Scenarios used in the simulations.

Scenario	q	η	β_1	β_2	Model
S_1	0.9	1.50	2.0	1.0	POHM-DW
S_2	0.5	1.00	2.0	1.0	POHM-G

To assess the behavior of the parameter estimators, the histograms of the parameter estimates of the different scenarios resulting from 1000 Monte Carlo replications will be evaluated for different sample sizes, i.e., n = 30, 50, 100, 250 and 500.

The mean of the parameter estimates, the mean squared error (MSE), and the coverage probability (CP) are shown alongside the above graphs. To construct the confidence intervals for calculating the CP, a confidence level of 0.95 was used. In addition, for the parameters of the probability distributions (q and η), which are limited in parametric space, it is interesting to transform them to make them unrestricted. The appropriate transformations were made to the following parameters to construct the confidence intervals, as described by [13].

The results from 1000 Monte Carlo replication that refer to the estimator q, η, β_1 and β_2 are shown in Figures 1–4 respectively.

When evaluating the estimators in general, it can be seen that the mean estimates are approximately equal to the respective true parameter values, regardless of the scenario and sample size. For the estimators referring to the baseline distribution, it can be seen that the mean values of the parameter estimates are concentrated close to the true parameter values, and as the sample size increases, the mean estimates of the MSEs become closer to zero, and the coverage probabilities converge to the adopted confidence level of 0.95.

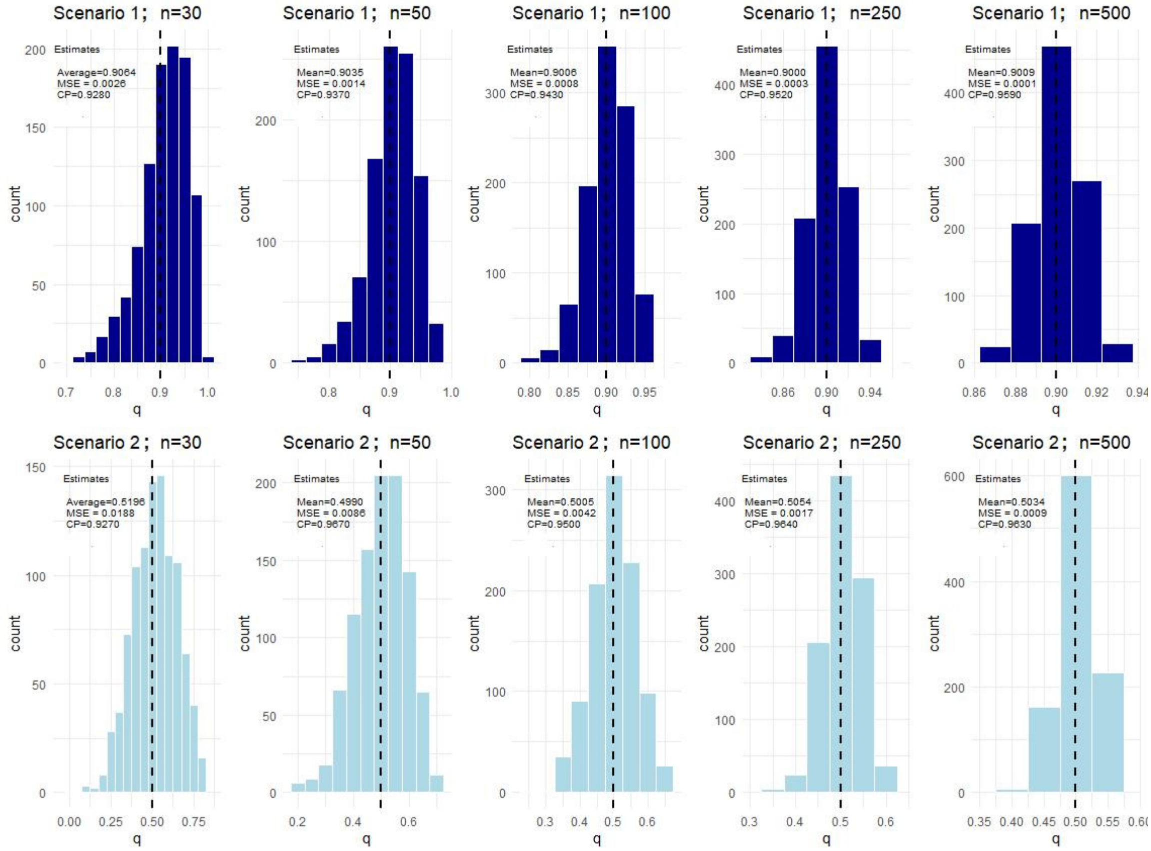


Figure 1. Results from 1000 Monte Carlo replications for the parameter q.

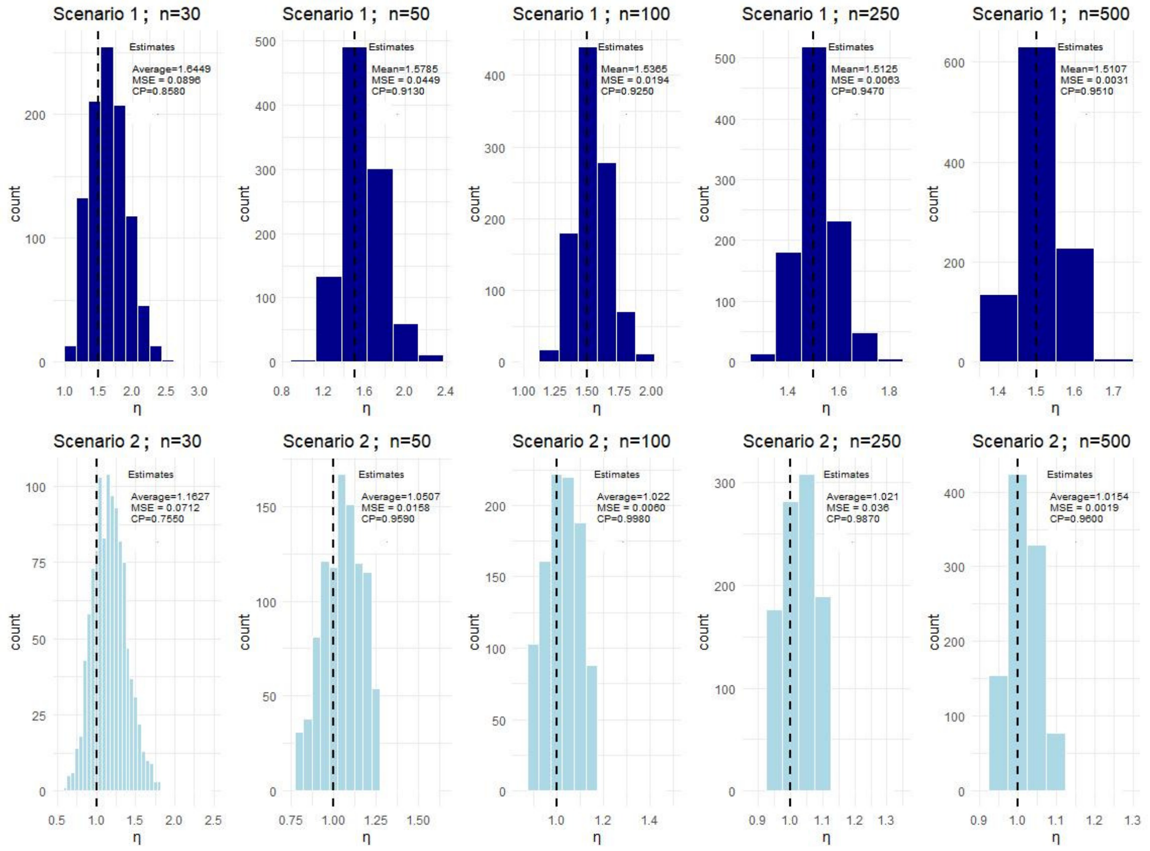


Figure 2. Results from 1000 Monte Carlo replications for the parameter η.

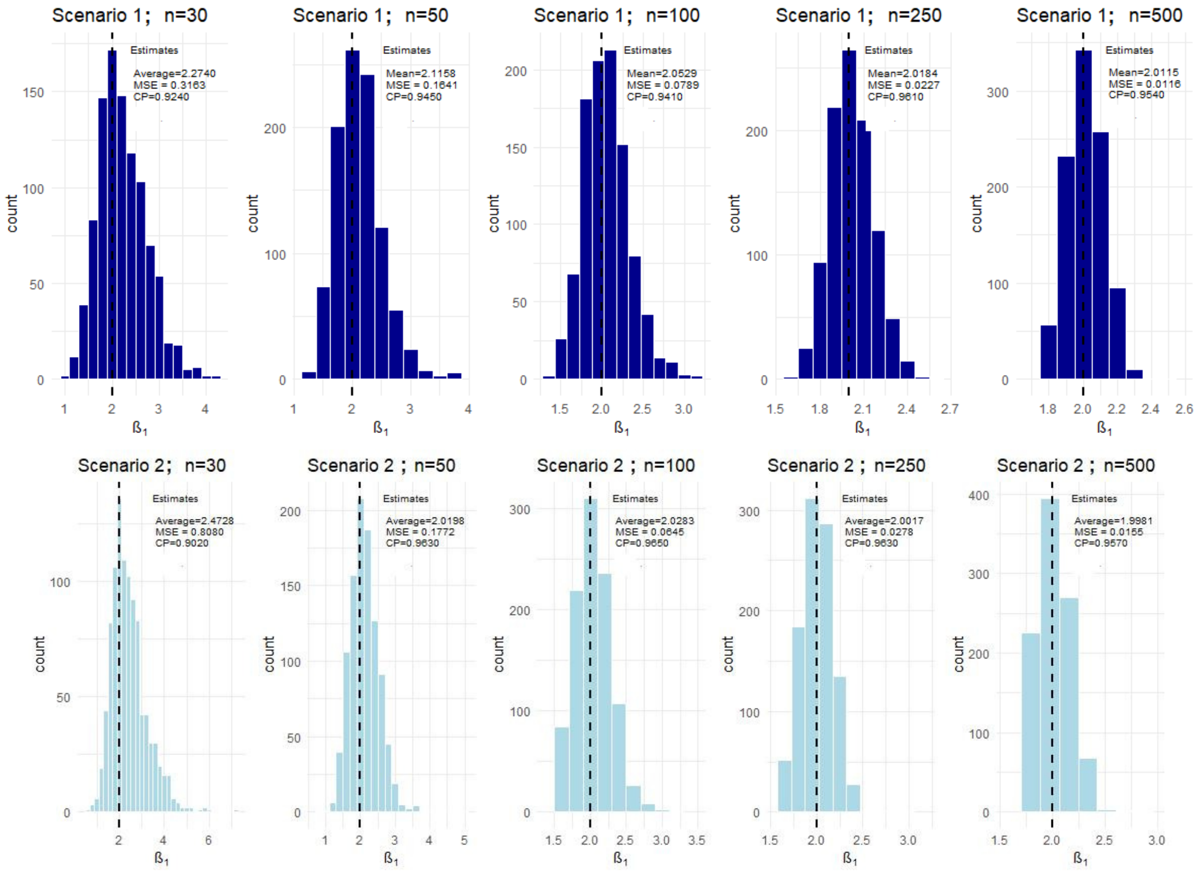


Figure 3. Results from 1000 Monte Carlo replications for the parameter β_1.

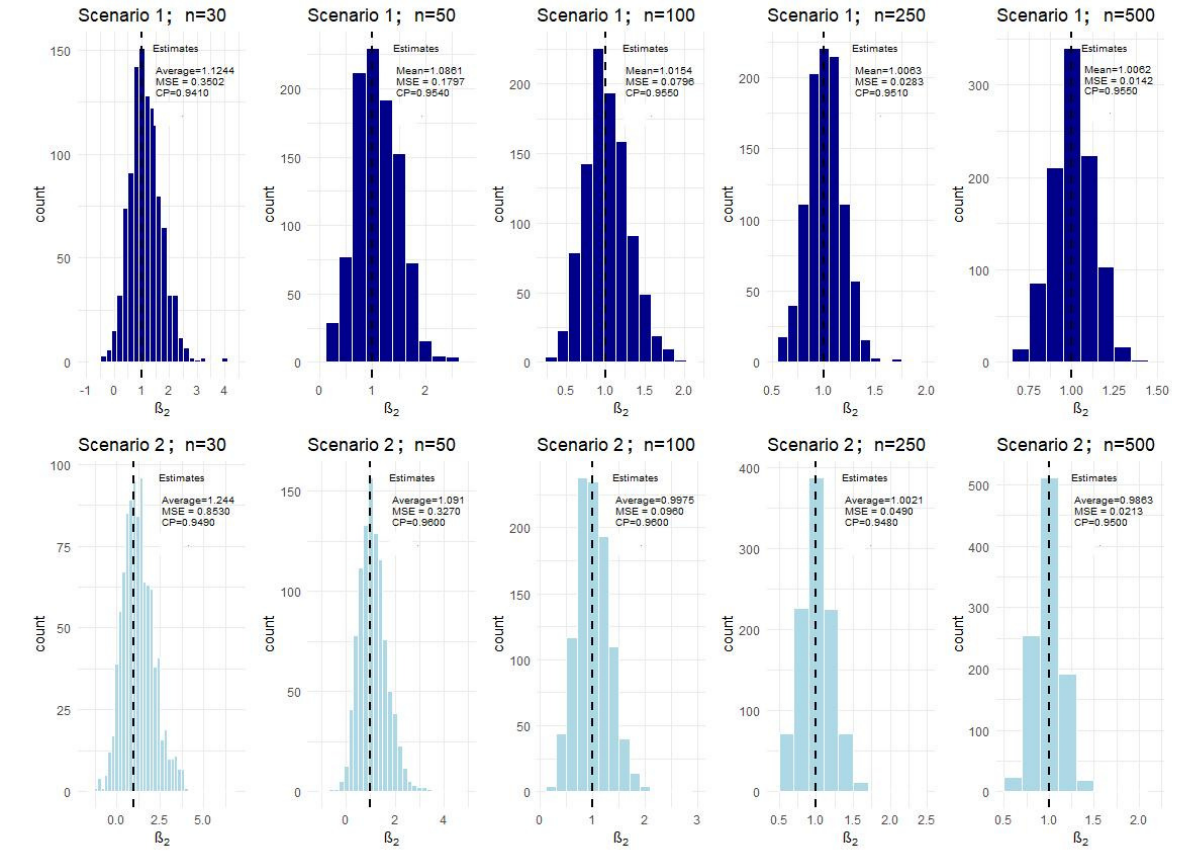


Figure 4. Results from 1000 Monte Carlo replications for the parameter β_2.

For the estimators related to the covariates, where β_1 is associated with the numerical variable and β_2 associated with the dichotomous variable, similar behavior can be observed between the two and, in turn, satisfactory performance concerning the estimates and distributions of the data, just like the estimators referring to the baseline distribution.

When evaluating the estimators for the scenarios, it can be seen that the first scenario is associated with a circumstance in which the discrete Weibull distribution is adopted as the baseline distribution and the second in which the geometric distribution is adopted (η = 1), it can be seen from the estimates and graphs presented that both baseline distributions are suitable for modeling discrete time-to-event data.

The entire evaluation up to this point has been carried out without censoring. Therefore, considering the same scenarios and sample sizes in Table 2 shows the estimates (average of parameter estimates, mean square error (MSE) and coverage probability (CP)) considering censoring percentages of 5, 10 and 30%. These estimates are the result of 1000 Monte Carlo replications.

Table 2. Average estimates, MSE and CP of the POHM-DW and POHM-G parameters considering the simulation scenarios and different sample sizes and censoring percentages.

n	Scen.	Cens. Perc.	q			η			β_1			β_2		
			Average	MSE	CP	Average	MSE	CP	Average	MSE	CP	Average	MSE	CP
30	S_1	5%	0.9097	0.0025	0.9100	1.6238	0.0825	0.8350	2.1556	0.2825	0.9110	1.0513	0.4044	0.9470
	S_2		0.5541	0.0187	0.9340	1.1186	0.0622	0.8280	2.1624	0.5490	0.8940	1.0735	0.7050	0.9410
	S_1	10%	0.9129	0.0023	0.9150	1.6038	0.0801	0.8500	2.0721	0.2510	0.9230	1.0090	0.4077	0.9390
	S_2		0.5889	0.0226	0.9130	1.0836	0.0578	0.8650	1.9374	0.4286	0.8590	0.9500	0.6541	0.9300
	S_1	30%	0.9296	0.0026	0.8820	1.5611	0.0970	0.9010	1.8336	0.2723	0.8720	0.8186	0.4750	0.9200
	S_2		0.6958	0.0493	0.6830	0.9668	0.0608	0.9360	1.4057	0.6068	0.6110	0.5878	0.7387	0.8940
50	S_1	5%	0.9075	0.0013	0.9340	1.5563	0.0380	0.8940	2.0476	0.1225	0.9350	1.0420	0.2177	0.9470
	S_2		0.5505	0.0114	0.9330	1.0499	0.0249	0.9010	1.9943	0.2079	0.9070	1.0536	0.3455	0.9400
	S_1	10%	0.9128	0.0013	0.9320	1.5460	0.0370	0.9060	1.9825	0.1242	0.9280	1.0032	0.2190	0.9450
	S_2		0.5876	0.0158	0.8770	1.0258	0.0237	0.9270	1.8185	0.2195	0.8420	0.9682	0.3290	0.9280
	S_1	30%	0.9309	0.0018	0.8770	1.5027	0.0435	0.9360	1.7629	0.1697	0.8480	0.8395	0.2562	0.9230
	S_2		0.7037	0.0472	0.4060	0.9471	0.0294	0.9590	1.3543	0.5223	0.4350	0.6763	0.3798	0.8980
100	S_1	5%	0.9069	0.0007	0.9410	1.5288	0.0153	0.9290	2.0139	0.0604	0.9390	1.0045	0.0848	0.9510
	S_2		0.5414	0.0060	0.9180	1.0215	0.0102	0.9320	1.9417	0.0913	0.9050	0.9766	0.1119	0.9460
	S_1	10%	0.9119	0.0008	0.9210	1.5223	0.0164	0.9280	1.9644	0.0637	0.9270	0.9718	0.0867	0.9410
	S_2		0.5763	0.0100	0.8230	1.0069	0.0103	0.9510	1.8175	0.1163	0.8290	0.9234	0.1127	0.9360
	S_1	30%	0.9312	0.0014	0.7800	1.4917	0.0220	0.9490	1.7748	0.1127	0.7970	0.8686	0.1066	0.9280
	S_2		0.6992	0.0417	0.0138	0.9563	0.0157	0.9450	1.4367	0.3828	0.3600	0.7579	0.1660	0.8810
250	S_1	5%	0.9060	0.0003	0.9430	1.5036	0.0062	0.9430	1.9614	0.0256	0.9300	0.9845	0.0299	0.9570
	S_2		0.5429	0.0036	0.8490	1.0037	0.0044	0.9380	1.8817	0.0478	0.8520	0.9536	0.0493	0.9480
	S_1	10%	0.9104	0.0004	0.9070	1.4928	0.0066	0.9390	1.9401	0.0326	0.8790	0.9546	0.0388	0.9340
	S_2		0.5778	0.0074	0.5900	0.9489	0.0046	0.9420	1.7501	0.0932	0.6200	0.8940	0.0580	0.9100
	S_1	30%	0.9303	0.011	0.5400	1.4579	0.0098	0.9420	1.7052	0.1092	0.4910	0.8492	0.0595	0.8710
	S_2		0.6927	0.0348	0.0020	0.9248	0.0108	0.8630	1.3568	0.4352	0.0230	0.7080	0.1303	0.7000
500	S_1	5%	0.9057	0.0002	0.9210	1.5006	0.0033	0.9430	1.9480	0.0148	0.9000	0.9700	0.0224	0.9340
	S_2		0.5404	0.0027	0.7380	0.9929	0.0023	0.9460	1.8399	0.0427	0.6870	0.9125	0.0371	0.8980
	S_1	10%	0.9102	0.0002	0.8620	1.4864	0.0035	0.9570	1.8845	0.0246	0.7870	0.9405	0.0244	0.9140
	S_2		0.5750	0.0660	0.3100	0.9817	0.0023	0.9650	1.7225	0.0873	0.3060	0.8496	0.0486	0.8160
	S_1	30%	0.9301	0.0010	0.2810	1.4489	0.0073	0.8830	1.6718	0.1195	0.1380	0.8368	0.0476	0.7660
	S_2		0.6902	0.0369	0.0010	0.9120	0.0105	0.6450	1.2752	0.5362	0.0000	0.6307	0.1590	0.2790

In the presence of censoring, it can be seen that the higher the percentage of censoring, the greater the deviations of the estimates from the true value of the parameter. This behavior is expected since the higher the percentage of censoring, the more the empirical distribution of the simulated data differs from the theoretical distribution used to generate the data. The probability of coverage, which has a confidence level of 0.95, reinforces this statement. Note that as the amount of censoring increases, the greater the differences between the CP and the confidence level stipulated for constructing the intervals.

Another pertinent aspect is that, even with this shift in the true value of the parameter, the distribution of the estimators, even in the presence of censoring, is similar to the estimators in the absence of censoring (see, for example, Figure 5, which shows the estimator of β_1, considering 30% censoring, which has the lowest CP values among the estimators).

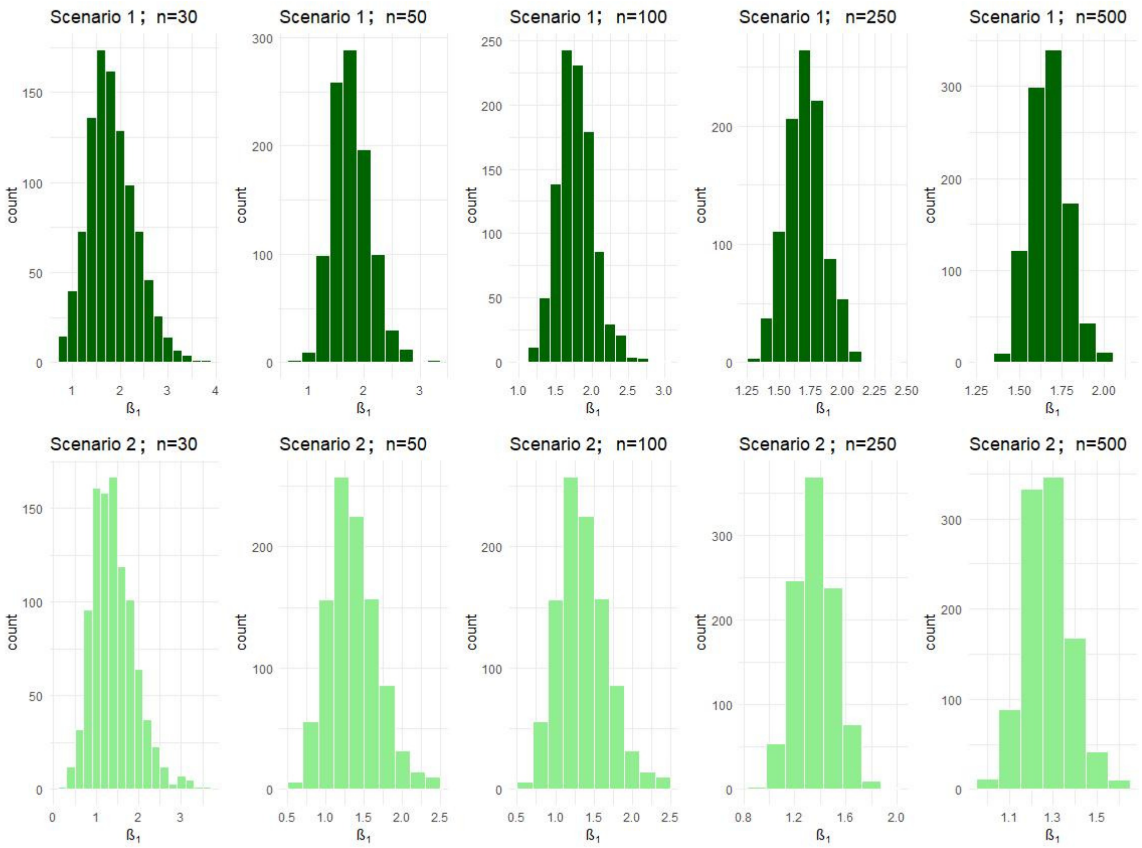


Figure 5. Results from 1000 Monte Carlo replications for the parameter β_1 (30% censoring).

Therefore, it can be seen from the results of the estimates and histograms, regardless of the scenario, censoring percentage, or sample size, that the shape of the empirical distribution of the estimators suggests adherence to the normal distribution. Thus, this distribution can be used for interval parameter estimation. As a result, hypothesis tests approximated by a normal distribution to verify the significance of the covariate can also be used in applications.

5. Application

Chronic low back pain is a major public health problem, as it can affect the quality of life and daily activities. Low back pain is also responsible for high rates of absenteeism from work.

The data set used in this study comes from [17], whose time-to-event is the number of unsuccessful sessions before the session that reduced or relieved the low back pain. Here, $t = 0$ represents the patient who would have had pain relief in the very first session.

Observations were considered censored when the patient's follow-up was interrupted for some reason unrelated to the event of interest in the study or after 11 unsuccessful sessions. Table 3 shows the number of patients who experienced a reduction or relief of low back pain and the number and percentage of censored patient observations per number of unsuccessful sessions.

Table 3. Patients with low back pain relief by number of unsuccessful sessions.

Time-to-Event	Patients with Relief	Number of Censures	% of Censures
0	85	0	0.00%
1	29	0	0.00%
2	7	1	0.67%
3	5	1	0.67%
4	4	0	0.00%
5	6	0	0.00%
6	1	0	0.00%
7	3	0	0.00%
9	1	0	0.00%
11	2	5	3.33%
Total	143	7	4.67%

Souce: [17].

In addition to the number of unsuccessful sessions, the data set includes information on the various characteristics of the 150 patients (6 covariates). The covariates age, body mass index (BMI), and duration of pain were originally quantitative and were categorized. The patients were divided into two age groups, one for individuals aged up to 50 and the other aged 50 or over; into two BMI groups, non-obese (BMI less than 30) and obese (BMI greater than or equal to 30); into two pain time groups, one with less than five years of pain and the other with five years or more of pain. This information is summarized in Table 4.

Table 4. Summary of study covariates by category.

Covariates	Categories	Frequency	%
Treatment	Placebo Active	50 100	33.33% 66.67%
Sex	Female Male	115 35	76.67% 23.33%
Age	Up to 50 years 50 years and over	64 84	44.00% 56.00%
BMI	Up to 30 30 or more	108 42	72.00% 28.00%
Pain duration	5 years or more Less than 5 years	93 57	62.00% 38.00%
Medicines	Yes No	115 35	76.67% 23.33%

The application data was then adjusted using POHM-G, POHM-DW and POHM-DLL. Initially, these multiple models were adjusted to check the significance of their covariates ($H_0 : \beta_1 = 0$ to $H_0 : \beta_6 = 0$). The p-value results of the multiple models are shown in Table 5.

According to the results in Table 5, the covariates treatment and medicines are significant (at a significance level of 5%) in all three models.

On the other hand, the other covariates are not significant and would not influence the relief or reduction of the patient's back pain. The significance test was therefore carried out by adjusting only the significant covariates, and the results are shown in Table 6.

The results in Table 6 show that the covariates treatment and medicines influence the relief or reduction of patients' low back pain. Therefore, taking these covariates into account, the study to verify the assumption of proportional odds hazard will be carried out using the methods presented in Section 3.2.

The assumption of proportional odds hazard will be verified for the data set, observing this proportionality between the levels of the covariate treatment and the covariate medicines and for each of the levels of these two covariates, using the graph: $log(G_0(t)) \times log(G_1(t))$ and the hypothesis test proposed in (25).

Table 5. Test of significance of covariates for multiple POHM-G, POHM-DW and POHM-DLL.

Identification		***p*-Value**		
Variable	**Parameter**	**POHM-G**	**POHM-DW**	**POHM-DLL**
Treatment [1]	β_1	5×10^{-5}	0.0035	0.0050
Sex [2]	β_2	0.9124	0.9601	0.9889
Age [3]	β_3	0.2915	0.4440	0.4272
BMI [4]	β_4	0.6500	0.7280	0.7357
Pain duration [5]	β_5	0.0129	0.0879	0.1098
Medicines [6]	β_6	0.0600	0.0396	0.0397

Reference level of the covariates: [1] = Placebo; [2] = Female; [3] = Up to 50 years; [4] = Up to 30; [5] = 5 years or more; [6] = Yes.

Table 6. Test of significance of covariates for POHM-G, POHM-DW and POHM-DLL for significant covariates.

Identification		***p*-Value**		
Variable	**Parameter**	**POHM-G**	**POHM-DW**	**POHM-DLL**
Treatment [1]	β_1	2×10^{-5}	0.0028	0.0045
Medicines [2]	β_2	0.0010	0.0172	0.0202

Reference level of the covariates: [1] = Placebo; [2] = Yes.

Since five consecutive tests were carried out, the Bonferroni correction will be used to correct the probability of incorrectly rejecting the null hypothesis, and thus the significance level will be $0.05/5 = 0.01$. The results are shown in Figure 6, assuming: z_1 = level of the covariate treatment ($z_1 = 0$: placebo; $z_1 = 1$: active) and z_2 = level of the covariate medicines ($z_2 = 0$: yes; $z_2 = 1$: no).

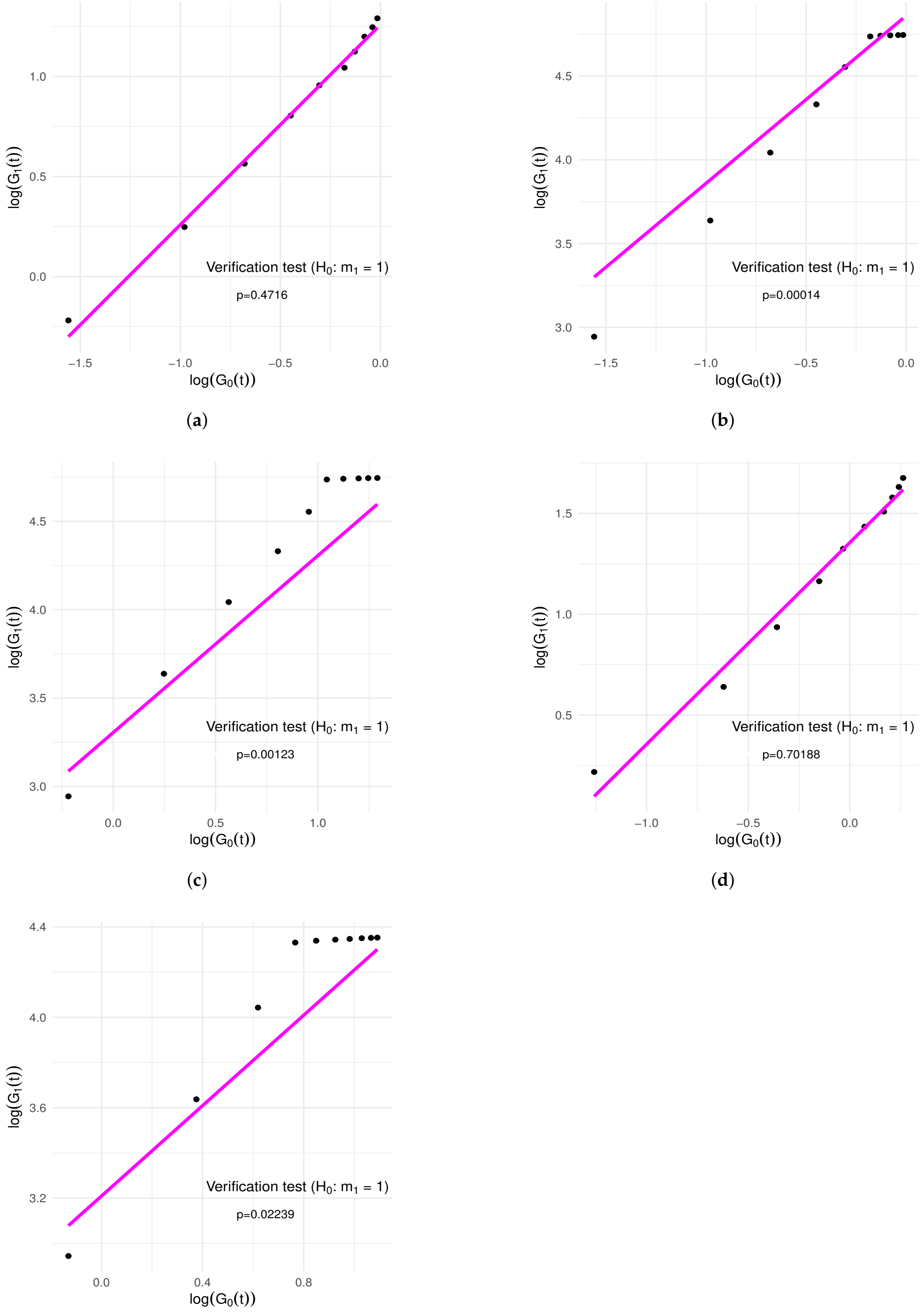
It is important to note that the number of tests to be carried out would be eight, that is, four levels of covariates combined two by two, totaling six tests plus the two levels within the covariates. However, one of the covariate levels ($z_1 = 0$; $z_2 = 1$) has a limited number of observations (<10), making it inadequate to construct graphs and test hypotheses.

The test results shown in Figure 6 show that the proportional odds hazard assumption was not rejected for 3 of the five levels of covariates considered in this study (given a significance level of 1%). The $log(G_0(t)) \times log(G_1(t))$ graphs shown corroborate that the proportional odds hazard was not rejected in the hypothesis tests, as the points are close to the fitted regression line.

The fact that most of the two-by-two levels studied have proportional odds hazard indicates that the data under study have proportional odds hazard, which justifies using this methodology in this data set.

Thus, for the POHM-G, POHM-DW, and POHM-DLL models as a whole, considering the two significant covariates, the point and interval estimates of their parameters were calculated and are shown in Table 7.

The estimates in Table 7, provide an interpretation of the odds hazard for the different categories of the covariates under study. Taking the POHM-DW model and the treatment covariate as an example. Since $\exp\{\beta_1\}$ represents the ratio of the odds hazard of the different groups, constant over time, considering that the covariate medicines is constant. Assuming the group of patients with active treatment ($z_1 = 1$). In this context, the odds hazard for patients on active treatment is $\exp\{0.7153\} = 2.0448$ times the odds hazard for patients on placebo treatment.



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure 6. Verification of the proportional hazards assumption for the covariates treatment (z_1) and medicines (z_2). (**a**) ($z_1 = 0$; $z_2 = 0$) × ($z_1 = 1$; $z_2 = 0$); (**b**) ($z_1 = 0$; $z_2 = 0$) × ($z_1 = 1$; $z_2 = 1$); (**c**) ($z_1 = 1$; $z_2 = 0$) × ($z_1 = 1$; $z_2 = 1$); (**d**) treatment; (**e**) medicines.

Table 7. POHM-G, POHM-DW and POHM-DLL joint parameter estimates—significant covariates.

POHM-G					
Variable	Parameter	Estimate	Standard Error	CI (95%)	p-value
-	q	0.2359	0.0325	[0.1722; 0.2996]	-
Treatment [1]	β_1	0.9657	0.2277	[0.5195; 1.4412]	2×10^{-5}
Medicines [2]	β_2	0.9967	0.3048	[0.3993; 1.5942]	0.0010
POHM-DW					
Variable	Parameter	Estimate	Standard Error	CI (95%)	p-value
-	q	0.5827	0.0574	[0.4701; 0.6953]	-
-	η	0.5887	0.0670	[0.4573; 0.7201]	-
Treatment [1]	β_1	0.7153	0.2390	[0.2468; 1.1838]	0.0028
Medicines [2]	β_2	0.7553	0.3169	[0.1342; 1.3764]	0.0172
POHM-DLL					
Variable	Parameter	Estimate	Standard Error	CI (95%)	p-value
-	α	1.4299	0.3606	[0.7231; 2.1367]	-
-	η	0.9568	0.1253	[0.7111; 1.2024]	-
Treatment [1]	β_1	0.6875	0.2419	[0.2135; 1.1615]	0.0045
Medicines [2]	β_2	0.7366	0.3171	[0.1151; 1.3582]	0.0202

Reference level of the covariates: [1] = Placebo; [2] = Yes.

Therefore, the odds hazard of the patient having active treatment is 1.0448 times greater than the odds hazard of the patient having placebo treatment ($z_1 = 0$). In this circumstance, the odds hazard for patients who do not use medication is 1.1282 times greater than the odds hazard for patients who do use medication.

The same interpretation can be made for the other models with different numerical values. However, the odds hazard remain higher for active treatment and patients not taking medication.

To assess the fit of the models to the data, the survival graphs of the Kaplan-Meier estimator (K-M) [20] and the survival curves of the models under study were drawn for each of the covariate levels to analyze the set of graphs and interpret their overall fit (Figure 7).

Figure 7 shows that the models fit the data well, with the survival estimates of these models always being close to the empirical estimates, with a positive highlight for the POHM-DLL and POHM-DW models, where the survival estimates are closer to the survival estimates of the Kaplan-Meier estimator.

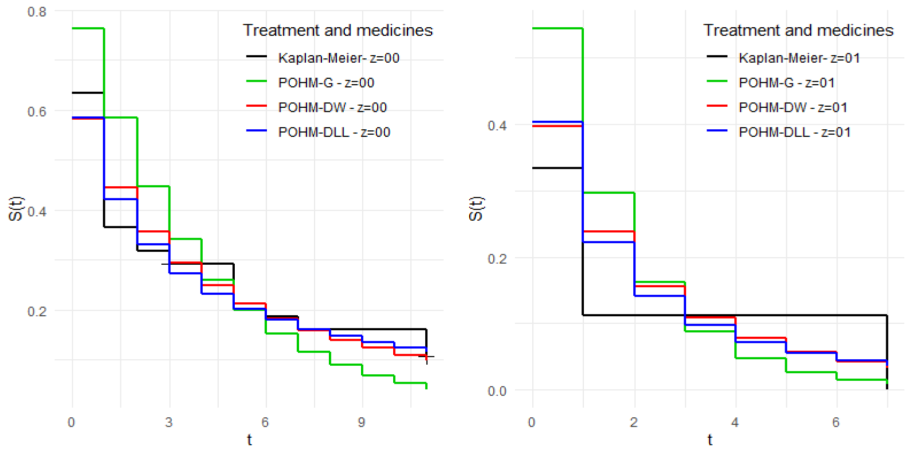


Figure 7. *Cont.*

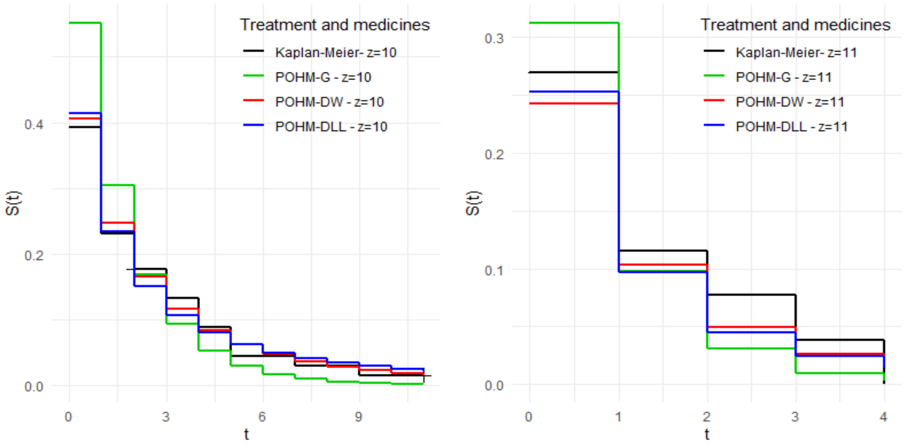


Figure 7. Fitting the models to the low back pain data by level of the covariates treatment and medicines.

In order to compare with pre-existing discrete models in the literature, regression models were fitted taking into account the DW (expressions (5)–(7)) and geometric distribution (DW with $\eta = 1$) with the covariates associated in the parameter q through a logit link function, i.e.,

$$q = \frac{\exp\{z^{'}\boldsymbol{\beta}\}}{1+\exp\{z^{'}\boldsymbol{\beta}\}}. \tag{27}$$

These models will be referred to respectively as the discrete Weibull regression model (DWRM) and the geometric regression model (GRM).

In addition, we also consider the DLL (expressions (8)–(10)) with the covariates associated in the parameter α through a logarithmic link function, i.e.,

$$\alpha = \exp\{z^{'}\boldsymbol{\beta}\}. \tag{28}$$

This model will be called the discrete log-logistic regression model (DLLRM).

Note, through Figure 8, that for levels $z_1 = 1$; $z_2 = 0$ and $z_1 = 1$; $z_2 = 1$, the so-called traditional models behaved similarly to the model under study. However, for the other levels, the estimates of these models are more distant from the empirical estimates compared to the model under study, providing indications that the proportional odds hazard structure for discrete data provides a better fit to the data when compared to traditional discrete models.

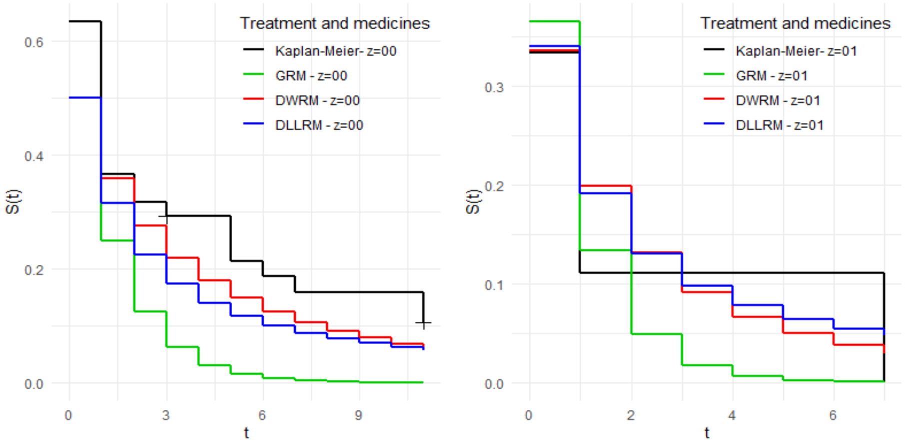


Figure 8. *Cont.*

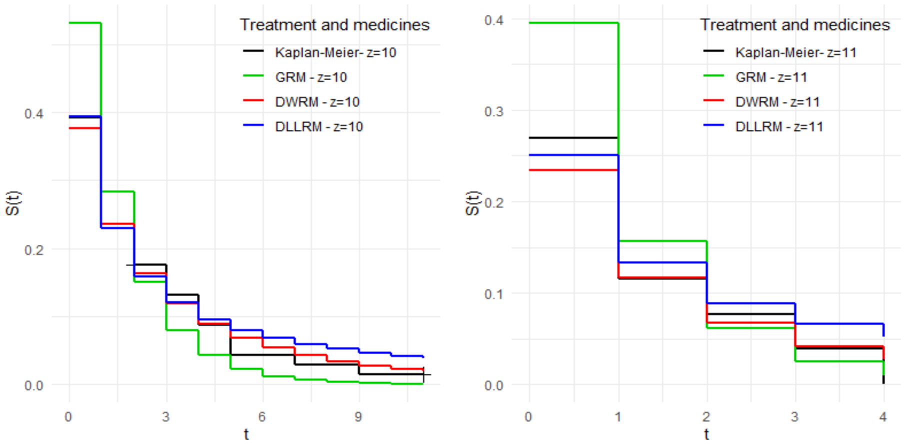


Figure 8. Fitting traditional discrete models to low back pain data by level of covariates treatment and medicines.

6. Conclusions

The proportional odds hazard model (POHM) presented in this paper is a regression model developed for discrete data that has been used in the literature because it has the advantage of facilitating the interpretation of its coefficients without having to worry about using the baseline hazard. However, in certain studies, correctly informing the baseline hazard is essential.

In this study, the POHM was formulated considering the following distributions: discrete Weibull of [10], geometric and discrete log-logistic. The inferential process was developed in a survival analysis context using the maximum likelihood estimation method. The results obtained on simulated data showed evidence of the asymptotic properties of the estimators for two different baseline distributions. Furthermore, the model proposed by adopting three different baseline distributions (geometric, discrete Weibull and discrete log-logistic) showed a good fit to the real data set, demonstrating that the estimation method developed and the use of baseline distributions for discrete random variables used to develop the POHM is a highly viable alternative for modeling discrete survival data with covariates.

Author Contributions: Conceptualization, M.G.F.V., M.R.P.C., R.M. and E.Y.N.; methodology, M.G.F.V., M.R.P.C., R.M. and E.Y.N.; software, M.G.F.V. and M.R.P.C.; validation, M.G.F.V., M.R.P.C., R.M. and E.Y.N.; formal analysis, M.G.F.V., M.R.P.C., R.M. and E.Y.N.; investigation, M.G.F.V., M.R.P.C., R.M. and E.Y.N.; resources, M.G.F.V. and E.Y.N.; data curation, M.G.F.V. and M.R.P.C.; writing—original draft preparation, M.G.F.V., M.R.P.C., R.M. and E.Y.N.; writing—review and editing, M.G.F.V., M.R.P.C., R.M. and E.Y.N.; visualization, M.G.F.V., M.R.P.C., R.M. and E.Y.N.; supervision, E.Y.N.; project administration, E.Y.N.; funding acquisition, E.Y.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES)—Finance Code 001, Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF)—TOA 443/2022, National Council for Scientific and Technological Development (CNPq), Editais de Auxílio Financeiro DPI/DPG/UnB, DPI/DPG/BCE/UnB and PPGEST/UnB.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CP	Coverage probability
DW	Discrete Weibull distribution
DLL	Discrete log-logistic distribution
DLLRM	Discrete log-logistic regression model
DWRM	Discrete Weibull regression model
GRM	Geometric regression model
MSE	Mean squared error
POHM	Proportional odds hazard model
POHM-G	Proportional odds hazard model geometric
POHM-DW	Proportional odds hazard model discrete Weibull
POHM-DLL	Proportional odds hazard model discrete log-logistic

References

1. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser. (Methodol.)* **1972**, *34*, 187–202. [CrossRef]
2. Tutz, G.; Schmid, M. *Modeling Discrete Time-to-Event Data*; Springer: New York, NY, USA, 2016.
3. Nakano, E.Y.; Carrasco, C.G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *Trends Comput. Appl. Math.* **2006**, *7*, 91–100. [CrossRef]
4. Berger, M.; Schmid, M. Semiparametric regression for discrete time-to-event data. *Stat. Model.* **2018**, *18*, 322–345. [CrossRef]
5. Vallejos, C.A.; Steel, M.F. Bayesian survival modelling of university outcomes. *J. R. Stat. Soc. Ser. Stat. Soc.* **2017**, *180*, 613–631. [CrossRef]
6. Heyard, R.; Timsit, J.-F.; Held, L. COMBACTE-MAGNET consortium. Validation of discrete time-to-event prediction models in the presence of competing risks. *Biom. J.* **2020**, *62*, 643–657. [CrossRef] [PubMed]
7. Zhou, X.-D.; Wang, Y.-J.; Yue R.-X. Optimal designs for discrete-time survival models with random effects. *Lifetime Data Anal.* **2021**, *27*, 300–332. [CrossRef] [PubMed]
8. Groll, A.; Tutz, G. Variable selection in discrete survival models including heterogeneity. *Lifetime Data Anal.* **2017**, *23*, 305–338. [CrossRef] [PubMed]
9. Royston, P.; Parmar, M.K.B. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat. Med.* **2002**, *21*, 2175–2197. [CrossRef] [PubMed]
10. Nakagawa, T.; Osaki, S. The discrete Weibull distribution. *IEEE Trans. Reliab.* **1975**, *24*, 300–301. [CrossRef]
11. Vila, R.; Nakano, E.Y.; Saulo, H. Theoretical results on the discrete Weibull distribution of Nakagawa and Osaki. *Statistics* **2019**, *53*, 339–363. [CrossRef]
12. Jayakumar, K.; Babu, M.G. Discrete Weibull geometric distribution and its properties. *Commun.-Stat.-Theory Methods* **2018**, *47*, 1767–1783. [CrossRef]
13. Cardial, M.R.P.; Fachini-Gomes, J.B.; Nakano, E.Y. Exponentiated discrete Weibull distribution for censored data *Braz. J. Biom.* **2020**, *38*, 35–56.
14. Chakraborty, S. Generating discrete analogues of continuous probability distributions-A survey of methods and constructions. *J. Stat. Distrib. Appl.* **2015**, *2*, 6. [CrossRef]
15. Krishna, H.; Pundir, P.S. Discrete Burr and discrete Pareto distributions. *Stat. Methodol.* **2009**, *6*, 177–188. [CrossRef]
16. Para, B.A.; Jan, T.R. Discrete version of log-logistic distribution and its applications in genetics. *Int. J. Mod. Math. Sci* **2016**, *14*, 407–422.
17. Corrêa, J.B.; Costa, L.O.P.; Oliveira, N.T.B.; Lima, W.P.; Sluka, K.A.; Liebano, R.E. Effects of the carrier frequency of interferential current on pain modulation and central hypersensitivity in people with chronic nonspecific low back pain: A randomized placebo-controlled trial. *Eur. J. Pain* **2016**, *20*, 1653–1666. [CrossRef]
18. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019. Available online: https://www.R-project.org/ (accessed on 1 October 2023).
19. Ross, S.M. *Simulation*; Academic Press: Cambridge, MA, USA, 2022.
20. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.