mathematics

MDPI

Article

cpd: An R Package for Complex Pearson Distributions

María José Olmo-Jiménez *,†, Silverio Vílchez-López † and José Rodríguez-Avi †

Department of Statistics and Operations Research, University of Jaén, 23071 Jaén, Spain
* Correspondence: mjolmo@ujaen.es; Tel.: +34-953-211909
† These authors contributed equally to this work.

Abstract: The complex Pearson (CP) distributions are a family of probability models for count data generated by the Gaussian hypergeometric function with complex arguments. The complex triparametric Pearson (CTP) distribution and its biparametric versions, the complex biparametric Pearson (CBP) and the extended biparametric Waring (EBW) distributions, belong to this family. They all have explicit expressions of the probability mass function (pmf), probability generating function and moments, so they are easy to handle from a computational point of view. Moreover, the CTP and EBW distributions can model over- and underdispersed count data, whereas the CBP can only handle overdispersed data, but unlike other well-known overdispersed distributions, the overdispersion is not due to an excess of zeros but other low values of the variable. Finally, the EBW distribution allows the variance to be split into three uniquely identifiable components: randomness, liability and proneness. These properties make the CP distributions of interest in the modeling of a great variety of data. For this reason, and for trying to spread their use, we have implemented an R package called cpd that contains the pmf, distribution function, quantile function and random generation for these distributions. In addition, the package contains fitting functions according to the maximum likelihood. This package is available from the Comprehensive R Archive Network (CRAN). In this work, we describe all the functions included in the cpd package, and we illustrate their usage with several examples. Moreover, the release of a plugin in order to use the package from the interface R Commander tries to contribute to the spreading of these models among non-advanced users.

Keywords: count data models; overdispersion; underdispersion; R package

MSC: 60-04

Citation: Olmo-Jiménez, M.J.; Vílchez-López, S.; Rodríguez-Avi, J. cpd: An R Package for Complex Pearson Distributions. *Mathematics* **2022**, *10*, 4101. https://doi.org/10.3390/math10214101

Academic Editor: Manuel Franco

Received: 21 September 2022
Accepted: 31 October 2022
Published: 3 November 2022

1. Introduction

The use of discrete distributions to model count data is widely illustrated in the literature. The first model, which describes the pure random case for an infinite range, is the Poisson distribution. This is a uniparametric model which assumes that data have equidispersion; that is to say, the variance is equal to the mean. Nevertheless, in real studies, data often exhibit overdispersion (i.e., the variance is greater than the mean) and less often exhibit underdispersion (i.e., the variance is less than the mean). For these situations, a great variety of models has been developed, with many of them obtained from the Poisson distribution. Among them, we find well-known models, such as the negative binomial (NB) [1], univariate generalized Waring (UGW) [2,3], generalized Poisson (GP) [4], zero-inflated [5] or hurdle models [6], as well as many other new models (see, for instance, [7–10]). One of these new models is the complex triparametric Pearson distribution with parameters a, b and γ, which are denoted by $CTP(a, b, \gamma)$. This distribution belongs to the family of discrete distributions generated by the Gaussian hypergeometric function when the two first parameters are complex conjugated numbers (i.e., ${}_2F_1(a + ib, a - ib; \gamma; 1)$, where i is the imaginary unit), and it has been widely studied in [11,12]. Two particular cases with two parameters have also been developed: the complex biparametric Pearson (CBP) distribution [13,14] and the extended bivariate Waring (EBW) distribution [15,16]. It is interesting

to take into account the fact that the CTP and EBW models can handle both over- and underdispersion, whereas the CBP model can only handle overdispersion. Nevertheless, the overdispersion of these three models, unlike other well-known overdispersed models, is not due to an excess of zeros but other low values of the variable. Specifically, the CBP model is useful when there is overdispersion, and the probability of zero is similar to that of a Poisson distribution [14]. In addition to the fact that the CTP and EBW models can be underdispersed, another advantage is that they do not have computational problems since there are explicit expresions of the pmf, pgf and moments, as they do occur in other models such as the CMP, HP or GP models [12]. All these properties make the CP distributions of interest in the modeling of a great variety of data.

For these reasons, it is essential to facilitate their use, which is accomplished through their implementation in different statistical software. In this sense, R, the free software environment for statistical computing and graphics [17], not only allows for using the most common distributions to compute the probabilities and quantiles or generate random numbers but also model the data. Thus, for example, the stats package contains functions for handling many standard univariate probability distributions, and the extraDistr package adds more univariate and multivariate distributions to the list. In the MASS package, the maximum likelihood modeling of several models is available via the fitdistr functions, and the fitdistrplus package implements several methods for fitting univariate parametric distributions. In addition, there are also specific built-in functions related to these aspects in other R packages. To sum up, this allows us to propose different models for a given data set, estimating the corresponding parameters and, in addition, comparing them to choose the more adequate one. However, all of these packages include the CP distributions, making their use inaccessible to most researchers.

In trying to solve this problem, we implemented an R package called cpd for the CP distributions. Thus, in this work, we present and describe a package which allows for obtaining the pmf, distribution function, quantile function and random number generation for the three distributions. Moreover, this package offers the possibility of estimating the parameters by the maximum likelihood method and also provides several goodness-of-fit tests and graphics as well as additional fit criteria. The package is available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=cpd (accessed on 1 September 2022). In addition, we implemented a plugin for the R Commander GUI that allows non-advanced R users to work with these models without using R code. The plugin is also a package called RcmdrPlugin.cpd available from CRAN at https://CRAN.R-project.org/package=RcmdrPlugin.cpd (accessed on 1 September 2022).

The remainder of this paper is organized as follows. Section 2 reviews the definitions and properties of the CTP, CBP and EBW distributions. In Section 3, the functions of the cpd package are detailed, including several examples to illustrate their use. In the final section, this paper concludes with a summary of the main characteristics of the package implemented.

2. Complex Pearson Distributions

2.1. Brief Description and Properties

The complex triparametric Pearson (CTP) distribution was first developed in [11]. It is a triparametric discrete distribution of an infinite range generated by the Gaussian hypergeometric function ${}_2F_1$ with complex parameters, so it belongs to the Gaussian hypergeometric distributions (GHD) family [1]. Specifically, X follows a $CTP(a, b, \gamma)$ distribution when its pmf has the following expression:

$$f(x) = f_0 \ \frac{(a+ib)_x(a-ib)_x}{(\gamma)_x}\frac{1}{x!}, \quad x = 0, 1, \ldots \tag{1}$$

where i is the imaginary unit, $a, b \in \mathbb{R}$ and $\gamma > max\{0, 2a\}$. $(\alpha)_r$ is the Pochhammer symbol (also known as a rising factorial) defined as $\Gamma(\alpha + r)/\Gamma(\alpha)$, with $\Gamma(\cdot)$ as the gamma function and f_0 as the normalizing constant given by

$$ {}_2F_1(a + ib, a - ib; \gamma; 1)^{-1} = \frac{\Gamma(\gamma - a - ib)\Gamma(\gamma - a + ib)}{\Gamma(\gamma)\Gamma(\gamma - 2a)}. $$

An alternative expression of Equation (1) in terms of the gamma function is

$$ f(x) = C \cdot \frac{\Gamma(a + ib + x)\Gamma(a - ib + x)}{\Gamma(\gamma + x)} \frac{1}{x!}, \quad x = 0, 1, \ldots \tag{2} $$

where C is the normalizing constant

$$ C = \frac{\Gamma(\gamma - a - ib)\Gamma(\gamma - a + ib)}{\Gamma(\gamma - 2a)\Gamma(a + ib)\Gamma(a - ib)}. $$

The probability generating function (pgf) is given by

$$ G(t) = \frac{\Gamma(\gamma - a - ib)\Gamma(\gamma - a + ib)}{\Gamma(\gamma)\Gamma(\gamma - 2a)} \; {}_2F_1(a + ib, a - ib; \gamma; t), \quad t \in \mathbb{R}. \tag{3} $$

Aside from its pmf, the model also has explicit expressions of the mean μ and the variance σ^2:

$$ \mu = \frac{a^2 + b^2}{\gamma - 2a - 1}, \quad \sigma^2 = \mu \frac{\mu + \gamma - 1}{\gamma - 2a - 2}, $$

which exist if $\gamma > 2a + 1$ and $\gamma > 2a + 2$, respectively.

This is unimodal with the mode in $\left[\frac{(a-1)^2+b^2}{\gamma-2a+1}\right]$ when $\frac{(a-1)^2+b^2}{\gamma-2a+1} \notin \mathbb{Z}$, where $[\cdot]$ is the integer part; otherwise, the distribution has two consecutive modes in the values:

$$ \frac{(a-1)^2 + b^2}{\gamma - 2a + 1} - 1 \text{ and } \frac{(a-1)^2 + b^2}{\gamma - 2a + 1}. $$

Then, if $a^2 + b^2 < \gamma$, then there is only one mode in zero. As a consequence, the pmf is J-shaped or bell-shaped. Moreover, the CTP is skewed to the right since its third central moment is always positive. For further details about the model, see [11,12].

One of the main properties of the CTP distribution is that it can be underdispersed, equidispersed or overdispersed. In particular, if $a \geq 0$, then the CTP is always overdispersed. Thus, the model has a great versatility in the modeling of count data, especially when the overdispersion of the data is due to a higher frequency of non-zero values. In addition, the fact is that having explicit expressions of the pmf, mean and variance prevents the computational problems of other well-known models that cope with over- and underdispersion, such as the Conway–Maxwell–Poisson [18] (CMP) and the hyper-Poisson [19] (HP) models.

The CTP model is a generalization of the complex biparametric Pearson ($CBP(b, \gamma)$) distribution, since the latter appears when $a = 0$ [13]. This model has the advantage of having one less parameter, but it is always overdispersed. It can be compared to an NB distribution, except for the fact that the probability of zero is less than that provided by an NB model and similar to that of a Poisson distribution.

The EBW distribution has two parameterizations: $EBW(\alpha, \rho)$ with $\alpha, \rho = \gamma - 2\alpha > 0$ and $EBW(\alpha, \gamma)$ with $\alpha < 0$ and $\gamma > 0$. This model can also be seen as a particular case of the $CTP(a, b, \gamma)$ distribution when $b = 0$ and as a particular case of the $UGW(a, k, \rho)$ distribution when $a = k = \alpha > 0$. In fact, given a UGW distribution, there exists an EBW distribution that is very close to the former with the benefit of having one less parameter. In addition, the EBW distribution allows the variance to be split into three uniquely

identifiable components: randomness, liability and proneness (see more details in [16]), solving the existing indeterminacy in the *UGW* model. Specifically, if $X \sim EBW(\alpha, \rho)$, then these components are

$$\sigma^2 = \frac{\alpha^2}{\rho - 1} + \frac{\alpha^2(\alpha + 1)}{(\rho - 1)(\rho - 2)} + \frac{\alpha^3(\alpha + \rho - 1)}{(\rho - 1)^2(\rho - 2)}. \quad (4)$$

2.2. Maximum Likelihood Estimation

Under the i.i.d. sample assumption, the parameters a, b and γ are by default estimated by maximizing the log-likelihood function, defined as

$$\ln L_{x_1,\ldots,x_n}(a, b, \gamma) = \sum_{j=1}^{n} \ln f(x_j|a, b, \gamma) = 2Re[\ln \Gamma(\gamma - a + ib)] - 2Re[\Gamma(a + ib)] \quad (5)$$
$$+ 2Re[\Gamma(a + ib + x)] - \ln \Gamma(\gamma - 2a) - \ln \Gamma(\gamma + x) - \ln x!$$

with $x_1, \ldots, x_n$, where n is the number of observations of the variable $X \sim CTP(a, b, \gamma)$. It should be noticed that $\Gamma(z) = \overline{\Gamma(\overline{z})}$, where z and $\overline{z}$ are conjugate complex numbers.

The log-likelihood for the *CBP* distribution is obtained when $a = 0$, and that for the *EBW* distribution is obtained when $b = 0$.

3. Using the cpd Package

3.1. Overview

The cpd package provides the functions to compute the probability mass function, distribution function, quantile function and random generation for the complex triparametric Pearson (CTP), complex biparametric Pearson (CBP) and extended biparametric Waring (EBW) distributions. In addition, the package contains maximum-likelihood fitting functions for these models.

The source code is available from the Comprehensive R Archive Network (CRAN) repository (https://CRAN.R-project.org/package=cpd, accessed on 1 September 2022), with all the information about its functions and parameters in the package's help file. It can be installed and loaded by typing the following commands in R:

```
R> install.packages(‘‘cpd’’)
R> library(cpd)
```

The package is open-source, so it is also available from GitHub (https://github.com/ujaen-statistics/cpd, accessed on 1 September 2022), where updates and comments can be submitted.

Specifically, the cpd package allows for computing the probability mass, distribution and quantile funtions of a CBP distribution through the following respective code:

```
dcbp(x, b, gamma)
pcbp(q, b, gamma, lower.tail = TRUE)
qcbp(p, b, gamma, lower.tail = TRUE)
```

where x is a vector of the non-negative integer values, q is a vector of the quantiles, p is a vector of the probabilities and b and gamma are the parameters of the distribution. In the pcbp and qcbp functions, the argument lower.tail has to be specified to consider $P(X \leq x)$ (if it is TRUE) or $P(X \geq x)$ (if it is FALSE). It is also possible to generate n random numbers from a CBP distribution with parameters b and gamma using the rcbp function, whose sentence is rcbp(n, b, gamma).

For the CTP distribution, the probability mass, distribution and quantile functions, as well as the random generation, are analogous:

```
dctp(x, a, b, gamma)
```

```
pctp(q, a, b, gamma, lower.tail = TRUE)
qctp(p, a, b, gamma, lower.tail = TRUE)
rctp(n, b, gamma)
```

In the case of the EBW distribution, there are two possible parameterizations depending on the sign of its first parameter α. Thus, if $\alpha < 0$, then the usual parametrization is $EBW(\alpha, \gamma)$ with $\gamma > 0$, whereas if $\alpha > 0$, then the usual parametrization is $EBW(\alpha, \rho)$ with $\rho = \gamma - 2\alpha > 0$. (This constraint guarantees the existence of the probability distribution.) Then, the corresponding probability mass, distribution and quantile functions, together with the random generation, take into account these two parameterizations:

```
debw(x, alpha, gamma, rho)
pebw(q, alpha, gamma, rho, lower.tail = TRUE)
qebw(p, alpha, gamma, rho, lower.tail = TRUE)
rebw(n, alpha, gamma, rho)
```

Moreover, the cpd package provides functions for fitting the CTP, CBP and EBW distributions to count data by the maximum likelihood method. These functions are fitctp, fitcbp and fitebw, respectively. Thus, the usage for the fitcbp function is

```
fitcbp(x, bstart = NULL, gammastart = NULL, method = ‘‘L-BFGS-B’’,
+ control = list(), ...),
```

for the CTP distribution

```
fitctp(x, astart = NULL, bstart = NULL, gammastart = NULL,
+ method = ‘‘L-BFGS-B’’, control = list(), ...)
```

as well as for the EBW distribution

```
fitebw(x, alphastart = NULL, gammastart = NULL, rhostart = NULL,
+ method = ‘‘L-BFGS-B’’, control = list(), ...)
```

These fitting functions estimate the distribution parameters by maximizing the log-likelihood function given in Equation (5) using the optim function of the stats package with “L-BFGS-B” [20] as the default fitting method, which allows box constraints. Other alternative methods of the optim function include “Nelder–Mead”, “BFGS”, “CG” and “SANN” (see the function help information for more details). Nonlinear minimization using a Newton-type algorithm is also possible (see the nlm function of the stats package). If the fitting method is not “L-BFGS-B”, then the parameters have to be reparameterized as $\alpha = e^{\alpha_0}$ and $\rho = e^{\rho_0}$ or $\alpha = -e^{\alpha_0}$ and $\gamma = e^{\gamma_0}$ with α_0, ρ_0 and $\gamma_0 \in \mathbb{R}$, in order to satisfy the corresponding constraints in each model. In this case, the standard errors provided by the fitting function are for the estimates of α_0 and ρ_0 or γ_0. The starting values for the optimization process are the estimates obtained by the method of the moments, unless the user introduces other values. These estimates are obtained by solving the system of equations:

$$\begin{pmatrix} m'_1 & -m'_1 & -1 \\ m'_2 & -m'_2 - m'^2_1 & -m'_1 + 1 \\ m'_3 & -m'_3 - 2m'_2 - m'_1 & -m'_2 - 2m'_1 - 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ m_2 \\ 2m_3 + m_2 \end{pmatrix} \quad (6)$$

where m'_r is the rth sample raw moment, $\theta_1 = \widehat{\gamma} - 1$, $\theta_2 = 2\widehat{a}$ and $\theta_3 = \widehat{a}^2 + \widehat{b}^2$.

In the case of the EBW distribution, the method of moments could provide two sets of starting values. In such a case, the optimization process would be carried out twice (one with each set of starting values), and the solution with less AIC will be shown. These fitting functions return S3 objects of the classes fitCBP, fitCTP and fitEBW for which the print, summary and plot methods are provided.

The summary of an object of the classes fitCBP, fitCTP and fitEBW provides the ML parameter estimates, their standard errors and the statistic and p values of the Wald test to

check if the parameters are significant. This summary also shows the loglikelihood, AIC and BIC values, as well as the results for the χ^2 goodness of fit test and the Kolmogorov–Smirnov test for discrete variables [21,22]. Finally, the correlation matrix of the parameter estimates also appears.

In addition, when the ML estimate of α is positive, the function varcomp—applied to an object of the class fitEBW—allows us to obtain the decomposition of the variance in the fitted EBW model (see the components in Equation (4)). This fact is useful to know the origin of the data variability.

The plot of an object of class fitCBP, fitCTP or fitEBW provides, by default, the observed and theoretical frequencies against the values of the variable, the CDF plot of both the empirical distribution and the fitted distribution or a PP plot representing the empirical distribution function evaluated at each data point (y axis) against the fitted distribution function (x axis).

3.2. Examples

3.2.1. Probability Mass, Distribution, Quantile and Random Generation Functions

We illustrate the use of the probability mass, distribution and quantile functions and how to generate random numbers from the CP distributions.

First, we consider $X \sim CBP(3, 2.5)$, and we compute $P(X = 0), P(X = 1)$ and $P(X = 2)$:

```
R> library(cpd)
R> cpd::dcbp(c(0, 1, 2), 3, 2.5)
[1] 0.02985882 0.10749176 0.15355965
```

The following sentences allow for computing $P(X \leq 3)$, $P(X \leq 5)$ and $P(X > 2)$:

```
R> cpd::pcbp(c(3, 5), 3, 2.5)
[1] 0.4387825 0.6528353
R> cpd::pcbp(c(2), 3, 2.5, lower.tail = FALSE)
[1] 0.7090898
```

To obtain the quartiles of X and the 95th percentile, the sentence and the R output are

```
R> cpd::qcbp(c(0.25, 0.5, 0.75, 0.95), 3, 2.5)
[1]  2  4  7 17
```

Finally, to generate 300 numbers from X, we type

```
R> set.seed(123)
R> x <- cpd::rcbp(300, 3, 2.5)
R> table(x)
x
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 26
 5 28 54 48 35 27 24 14 14  6  4  9  6  4  1  2  1  2  4  1  2  2  1  1
 28 29 30 45
  1  1  2  1
```

Figure 1a shows the bar plot of the random generated data which is obtained with the code

```
R> barplot(table(x), xlab = ‘‘values’’, ylab = ‘‘frequencies’’)
```

We can observe that the mode of these data is at a value of two, and they exhibit overdispersion ($\overline{x} = 5.676667 < s^2 = 32.21952$).

Next we consider $Y \sim CTP(-1.5, 2, 2)$, which is underdispersed, and we compute $P(X = 0), P(X = 1), P(X = 2), P(X = 3)$:

```
R> cpd::dctp(c(0:3),-1.5, 2, 2)
[1] 0.1331089 0.4159654 0.2946422 0.1043524
```

Now, the cumulative probabilities $P(X \leq 1), P(X \leq 3)$ and $P(X > 1)$ are

```
R> cpd::pctp(c(1,3), −1.5, 2, 2)
[1] 0.5490744 0.9480690
R> cpd::pctp(1, −1.5, 2, 2, lower.tail = FALSE)
[1] 0.4509256
```

The quartiles and 95th percentile are obtained as follows:

```
R> cpd::qctp(c(0.25, 0.5, 0.75, 0.95), −1.5, 2, 2)
[1] 1 1 2 4
```

To generate 500 random values from Y, we use the code

```
R> set.seed(123)
R> y <- cpd::rctp(500, −1.5, 2, 2)
```

The frequency table is

```
R> table(y)
y
  0   1   2   3   4   5   6   7  10
 57 227 142  41  21   8   2   1   1
```

Additionally, Figure 1b shows a bar plot of the random generated data. The mode of these data is at a value of one, and they exhibit underdispersion ($\overline{y} = 1.574 > s^2 = 1.367259$).

Finally, to conclude this section, let us consider $X_1 \sim EBW(\alpha = 2, \rho = 5)$ and $X_2 \sim EBW(\alpha = -1.2, \gamma = 0.75)$, with the first being overdispersed and the second being underdispersed, and let us compute $P(X_i = 0), P(X_i = 1), P(X_i = 2), P(X_i = 3)$ and $P(X_i = 4), i = 1, 2$:

```
R> cpd::debw(c(0:4), 2, rho = 5)
[1] 0.53571429 0.23809524 0.10714286 0.05194805 0.02705628
R> cpd::debw(c(0:4), −1.2, gamma = 0.75)
[1] 0.3396452713 0.6521189210 0.0074527877 0.0005781556 0.0001248816
```

The cumulative probabilities $P(X_i \leq 2)$, $P(X_i \leq 4)$ and $P(X_i > 2), i = 1, 2$ are obtained as follows:

```
R> cpd::pebw(c(2,4), 2, rho = 5)
[1] 0.8809524 0.9599567
R> cpd::pebw(2, 2, rho = 5, lower.tail = FALSE)
[1] 0.1190476
R> cpd::pebw(c(2,4), −1.2, gamma = 0.75)
[1] 0.9981288 0.9998974
R> cpd::pebw(3, −1.2, gamma = 0.75, lower.tail = FALSE)
[1] 0.0002048643
```

The corresponding quartiles and 99th percentile are given by

```
R> cpd::qebw(c(0.25, 0.5, 0.75, 0.99), 2, rho = 5)
[1] 0 0 1 8
R> cpd::qebw(c(0.25, 0.5, 0.75, 0.99), −1.2, gamma = 0.75)
[1] 0 1 1 1
```

To generate 1000 random values of X_1 and X_2, we type the code

```
R> set.seed(123)
R> x1 <- cpd::rebw(1000, 2, rho = 5)
R> x2 <- cpd::rebw(1000, −1.2, gamma = 0.75)
```

The respective frequency tables are

```
R> table(x1)
x1
  0   1   2   3   4   5   6   7   8   9  10  11  18
542 236 105  51  25  16   8   6   5   3   1   1   1
R> table(x2)
x2
  0   1   2   3
332 657  10   1
```

In addition, the bar plots of these two datasets may be seen in Figure 1c,d. The modes of these data were zero and one, respectively. The first dataset was overdispersed ($\overline{x}_1 = 0.975 < s_1^2 = 2.657032$), and the second one was underdispersed ($\overline{x}_2 = 0.68 > s_2^2 = 0.2438438$).

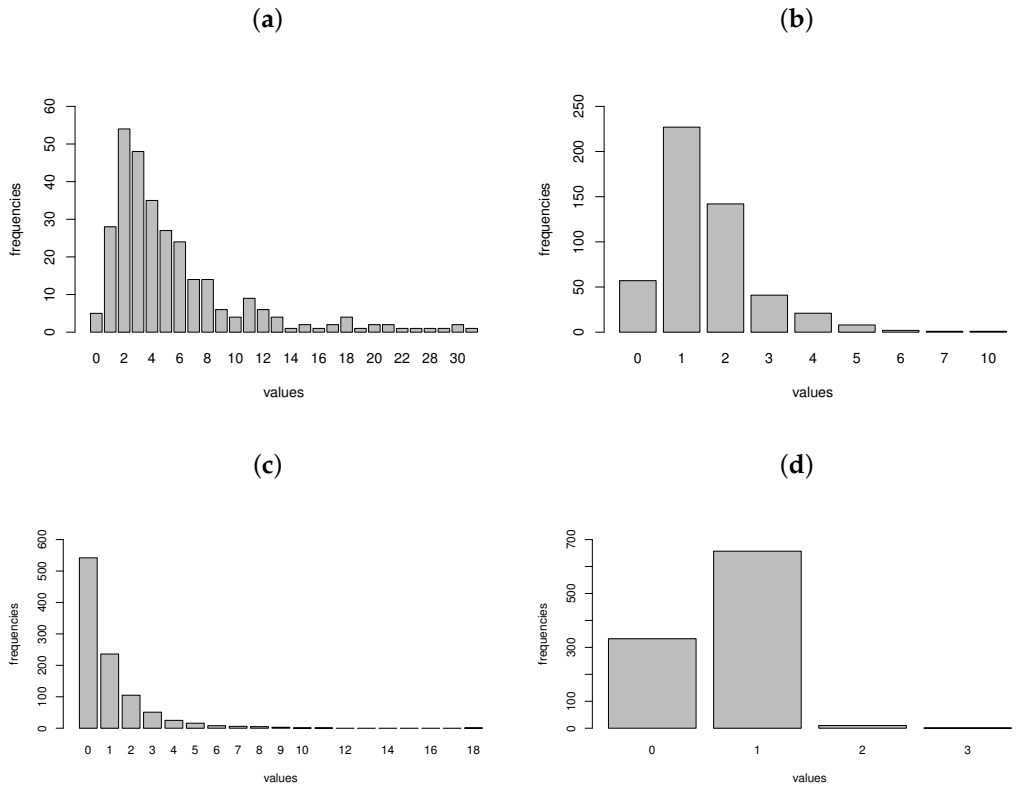

Figure 1. Bar plots of the random generated data from (**a**) $CBP(3, 2.5)$, (**b**) $CTP(-1.5, 2, 2)$, (**c**) $EBW(2, 5)$ and (**d**) $EBW(-1.2, 0.75)$.

3.2.2. Fitting Functions

To illustrate the use of the fitting functions, we provide three examples: two overdispersed and one underdispersed.

The first data set refers to the number of fire outbreaks by municipality in the region of Andalusia (Spain). Data were obtained from the Nature Databank of the Ministry of the Environment (Spain) and counting the number of fire outbreaks from 2001 to 2014. A fire outbreak was defined as a wildfire whose total area was less than one hectare. Moreover, municipalities whose forest land was zero were removed from the data. A description of these data appears in Table 1, which contains the mean, variance, quartiles, minimum and maximum of the data. It is clear that these data exhibit overdispersion since $s^2 > \overline{x}$.

Table 1. Descriptive summary of data.

	$\bar{x}$	s^2	Q_1	Q_2	Q_3	**Min**	**Max**
Fire outbreaks	3.528	28.964	1	2	4.75	0	56
Schools	1.431	10.3807	0	1	1	0	48
Syllables-1	1.889	0.910	1	2	2	0	4

We fit a CBP model to the data, considering as the initial values the estimates by the method of moments:

```
R> fireoutbreaks.cbp <- cpd::fitcbp(fireoutbreaks)
```

The output shows the ML estimates and their standard errors in parentheses:

```
     b           gamma
  1.486206      1.494944
(0.08849089)  (0.12183708)
```

Using the summary method, the output is more complete. The argument grouping = TRUE is the setting for grouping in classes with an expected frequency greater than or equal to five in the χ^2 goodness of fit test, since the default value is FALSE:

```
R> summary(fireoutbreaks.cbp, grouping = TRUE)

Parameters:
        Estimate    Std. Error    z-value      Pr(>|z|)
b       1.486206    0.08849089    16.79502   2.654186 × 10^−63
gamma   1.494944    0.12183708    12.27003~1.312062 × 10^−34

Loglikelihood: −1637.21   AIC: 3278.43   BIC: 3287.5

Goodness-of-fit tests:
Chi-2: 60.05902 (p-value: 5.11525627536094 × 10^−7)
Kolmogorov-Smirnov: 0.04732388 (p-value: 0.033)

Correlation Matrix:
              b     gamma
b     1.0000000 0.9296264
gamma 0.9296264 1.0000000
```

The AIC for the CBP fit is lower than the AIC related to an usual NB fit for these data:

```
R> library(MASS)
R> fireoutbreaks.nb <- MASS::fitdistr(fireoutbreaks, ‘‘negative binomial’’)
R> fireoutbreaks.nb
     size          mu
 0.80061445   3.52752644
(0.05537946) (0.16624492)
R> AIC(fireoutbreaks.nb)
[1] 3292.707
```

Next we model the data using the CTP distribution. However, the method of moments does not provide any estimates, so we introduce starting values for these parameters:

```
R> fireoutbreaks.ctp <- cpd::fitctp(fireoutbreaks)
Error in fitctp(fireoutbreaks) :
    The method of moments does not provide any estimates. Enter
    initial values for the~parameters.

R> fireoutbreaks.ctp <- cpd::fitctp(fireoutbreaks, astart = 0, bstart = 1,
+ gammastart = 2.1)
R> summary(fireoutbreaks.ctp, grouping = TRUE)

Parameters:
       Estimate   Std. Error    z-value    Pr(>|z|)
a      1.880214    0.8151401   2.306615   0.021076319
b      1.579993    0.5102531   3.096489   0.001958269
gamma  6.441561    2.1065484   3.057874~0.002229130

Loglikelihood: −1623.73   AIC: 3253.45   BIC: 3260.53

Goodness-of-fit tests:
Chi-2: 23.21167 (p-value: 0.0569114065884523)
Kolmogorov-Smirnov: 0.01796262 (p-value: 0.746)

Correlation Matrix:
               a          b      gamma
a      1.0000000 −0.9502083  0.9952855
b     −0.9502083  1.0000000 −0.9193242
gamma  0.9952855 −0.9193242  1.0000000
```

Once again, the fitted model was improved compared with the previous ones.

Finally, we carry out an EBW fit:

```
R> fireoutbreaks.ebw <- cpd::fitctp(fireoutbreaks)
R> summary(fireoutbreaks.ebw, grouping = TRUE)

Parameters:
        Estimate    Std. Error    z-value      Pr(>|z|)
alpha   2.749528    0.1621672    16.954903   1.770541 × 10^−64
rho     3.139183    0.3171861     9.896977   4.290494 × 10^−23
gamma   8.638240    0.6293458    13.725746~7.119262 × 10^−43

Loglikelihood: −1624.12   AIC: 3252.25   BIC: 3261.32

Goodness-of-fit tests:
Chi-2: 24.05791 (p-value: 0.0641165517298056)
Kolmogorov-Smirnov: 0.01778027 (p-value: 0.773)

Correlation Matrix:
          alpha       rho
alpha 1.0000000 0.9247998
rho   0.9247998 1.0000000
```

This is the most accurate of the four fits using the AIC. Moreover, the goodness-of-fit tests show that the EBW distribution is a reasonable model for the fire outbreak data. Figure 2 includes the observed and expected frequencies, CDFs and PP plots for this fit obtained with the following sentences:

```
R> plot(fireoutbreaks.ebw)
R> plot(fireoutbreaks.ebw, plty = ‘‘CDF’’)
R> plot(fireoutbreaks.ebw, plty = ‘‘PP’’)
```

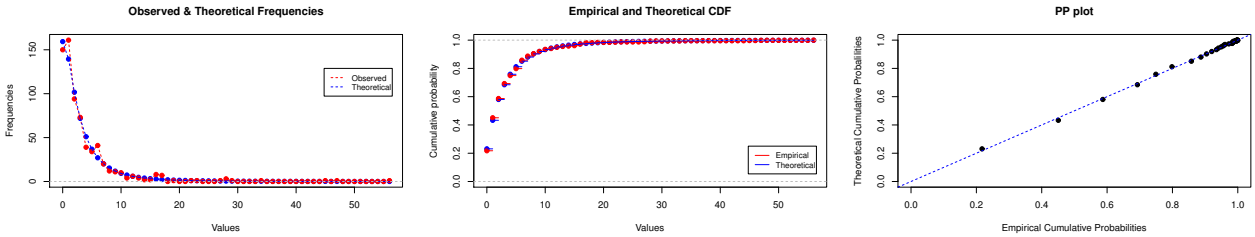

Figure 2. Observed and expected frequencies, CDFs and PP plots for the EBW fit of fire outbreak data.

As $\widehat{\alpha} = 2.7495 > 0$, the absolute value and the ratio of the variance components of the EBW fit can be obtained by typing the command

```
R> cpd::varcomp(fireoutbreaks.ebw)

               Value      Ratio
Randomness  3.534015  0.1019654
Liability  11.631922  0.3356107
Proneness  19.493035  0.5624239
```

The results indicate that 10.1965% of the variability in fire outbreaks was due to ramdomness, 33.5611% was due to liability (which refers to the general and external conditions of the municipality), and 56.2424% was due to proneness (related to the specific and internal characteristics of the municipality).

The second data set refers to the number of compulsory secondary schools by municipality in Andalusia (Spain) in the academic year 2020–2021. Data have been obtained through the Multiterritorial Information System of Andalusia (SIMA). The main descriptive statistics for these data appear in Table 1. These data reveal severe overdispersion caused by a value of one, as can be seen from the table of frequencies:

```
R> table(cs_schools)

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  18  20
201 437  67  31  17   6   6   2   3   2   2   1   2   1   1   1~1

 26  28  47  48
  1   1   1   1
```

First, we fit an NB model:

```
R> schools.nb <- MASS::fitdistr(cs_schools, ‘‘negative binomial’’)
R> schools.nb

    size          mu
1.19336385   1.43057377
(0.10212557) (0.06330097)

R> AIC(schools.nb)

2584.456
```

As expected, the results for the CTP fit (with initial values $a = 0, b = 1$ and $\gamma = 0.5$) improved remarkably with respect to the previous ones:

```
R> schools.ctp <- cpd::fitctp(schools, astart = 0, bstart = 1,
+ gammastart = 0.5)
R> summary(schools.ctp)

Parameters:
        Estimate    Std. Error    z-value      Pr(>|z|)
a      −0.5141782   0.03519091   −14.611108  2.386050 × 10^−48
b       0.4165253   0.09994774     4.167431  3.080518 × 10^−5
gamma   0.2020829   0.05620251     3.595620~3.236198 × 10^−4

Loglikelihood: −1058.91   AIC: 2123.81   BIC: 2131.14

Goodness-of-fit tests:
Chi-2: 40.81333 (p-value: 0.649840570375861
Kolmogorov-Smirnov: 0.007734697 (p-value: 0.932)

Correlation Matrix:
                a          b       gamma
a       1.0000000 −0.8066776 −0.7935876
b      −0.8066776  1.0000000  0.9637505
gamma  −0.7935876  0.9637505  1.0000000
```

The goodness-of-fit tests also support the adequacy of the model fitted. Figure 3 includes the observed and expected frequencies, CDFs and PP plots for the CTP fit obtained with the code sentences

```
R> plot(schools.ctp)
R> plot(schools.ctp, plty = ''CDF'')
R> plot(schools.ctp, plty = ''PP'')
```

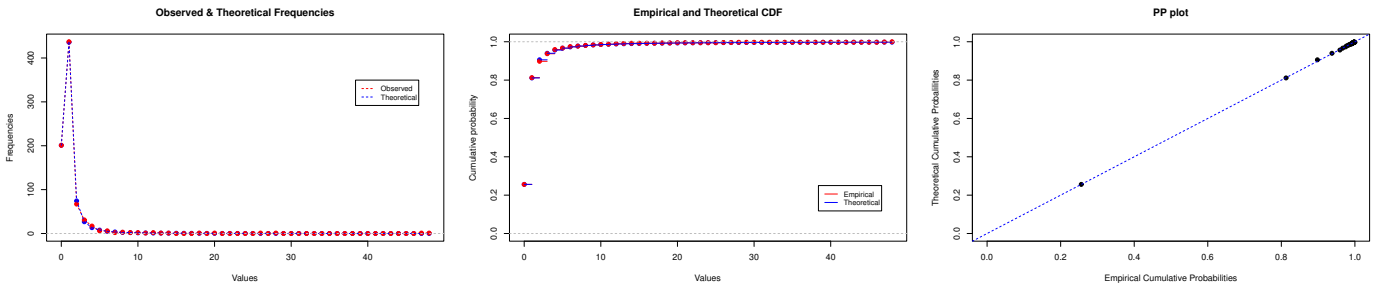

Figure 3. Observed and expected frequencies, CDFs and PP plots for the EBW fit of compulsory secondary school data.

In this example, it makes no sense to fit a CBP or a EBW model, since the parameters a and b in the CTP are statistically significant.

The third data set contained lengths of words (numbers of syllables) in a Slovak poem [23]. A description of these data appears in Table 1, where we considered the response variable $X - 1$, as though the data were generated by adding one to the distribution. These types of data related to word length often exhibit underdispersion [24].

As the CBP model is always overdispersed, it made no sense to fit it to these data, so we fitted a CTP model, considering as the initial values the estimates by the method of moments. In this example, the argument grouping is missing, so it is set to FALSE by default, since there are not enough degrees of freedom to group the classes with an expected frequency greater than or equal to five. Thus, the code is

```
R> syllables.ctp <- cpd::fitctp(syllables)
R> summary(syllables.ctp)
```

and the output

```
Parameters:
        Estimate     Std. Error      z-value      Pr(>|z|)
a     −5.7319721810   0.8127761   −7.0523387550 1.759352 × 10^−12
b      0.0008637403   6.0264402    0.0001433251 9.998856 × 10^−1
gamma  6.9298746437   3.3472112    2.0703428298~3.842025 × 10^−2

Loglikelihood: −159.33   AIC: 324.67   BIC: 328.19

Goodness-of-fit tests:
Chi-2: 0.6018625 (p-value: 0.437868271837147)
Kolmogorov-Smirnov: 0.01181905 (p-value: 0.994)

Correlation Matrix:
                a            b        gamma
a      1.000000000  0.002883042 −0.970588862
b      0.002883042  1.000000000 −0.001203176
gamma −0.970588862 −0.001203176  1.000000000
```

Now, we model these data using EBW distribution by typing

```
R> syllables.ebw <- cpd::fitebw(syllables)
R> summary(syllables.ebw)

Parameters:
        Estimate    Std. Error    z-value      Pr(>|z|)
alpha  −5.731882     0.812740    −7.052541   1.756792 × 10^−12
gamma   6.929558     3.347026     2.070363~3.841840 × 10^−2

Loglikelihood: −159.33   AIC: 322.67   BIC: 328.19

Goodness-of-fit tests:
Chi-2: 0.6019236 (p-value: 0.74010605961245)
Kolmogorov-Smirnov: 0.01181588 (p-value: 0.994)

Correlation Matrix:
           alpha      gamma
alpha  1.0000000 −0.9705887
gamma −0.9705887  1.0000000
```

Let us notice that the EBW fit was the best one according to the AIC and goodness-of-fit tests.

In addition, we used the sentences

```
R> plot(syllables.ebw)
R> plot(syllables.ebw, plty = ‘‘CDF’’)
R> plot(syllables.ebw, plty = ‘‘PP’’)
```

to obtain the plots in Figure 4.

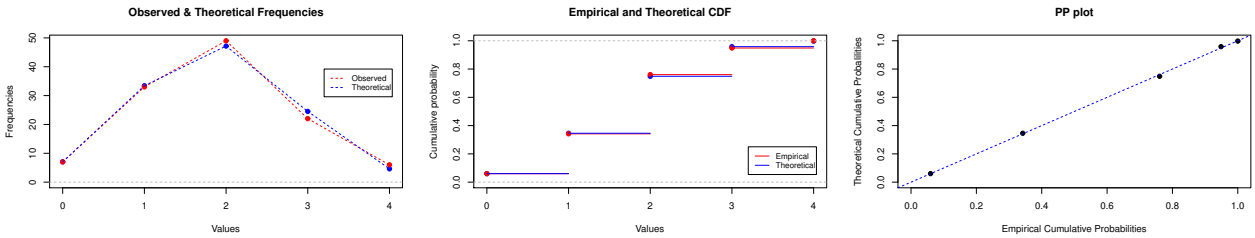

Figure 4. Observed and expected frequencies, CDFs and PP plots for the CTP fit of Slovak poem data.

4. Conclusions

The cpd package has been designed for computing probabilities and quantiles as well as for generating random numbers from the CBP, CTP and EBW distributions. These functions have also been included in a plugin for the GUI R Commander with the aim of facilitating their use by non-advanced R users. In addition, the package contains fitting functions to obtain the ML estimates of their parameters. In this way, we give more visibility to these models, which allows for modeling overdispersed data in which the overdispersion is not due to a value of zero but to low values of the variable (1, 2, ...) and also underdispersed data. Thus, the probability of zero in the CBP is lower than in the corresponding Poisson with the same mean, so the CBP can be seen as an adequate model for overdispersed data without too many zeros. Regarding the CTP and EBW models, they do not have the computational problems of other well-known models for both over- and underdispered data such as the GP, the CMP or the HP.

Author Contributions: Data curation, J.R.-A.; Formal analysis, M.J.O.-J. and S.V.-L.; Investigation, M.J.O.-J., S.V.-L. and J.R.-A.; Methodology, M.J.O.-J., S.V.-L. and J.R.-A.; Software, M.J.O.-J. and S.V.-L.; Supervision, M.J.O.-J., S.V.-L. and J.R.-A.; Writing—original draft, M.J.O.-J. and J.R.-A.; Writing—review & editing, M.J.O.-J., S.V.-L. and J.R.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data in Section 3.2.2 have been obtained from https://www.miteco.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/ (accessed on 1 September 2022) (Example 1) and https://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/index2-en.htm (accessed on 1 September 2022) (Example 2).

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

References

1. Johnson, N.L.; Kemp, A.W.; Kotz, S. *Univariate Discrete Distributions*, 3rd ed.; Wiley: New York, NY, USA, 2005.
2. Irwin, J.O. The generalized Waring distribution. Part I. *J. R. Stat. Soc. Ser. A* **1975**, *138*, 18–31. [CrossRef]
3. Rodríguez-Avi, J.; Conde-Sánchez, A.; Sáez-Castillo, A.J.; Olmo-Jiménez, M.J. A new generalization of the Waring distribution. *Comput. Stat. Data Anal.* **2007**, *51*, 6138–6150. [CrossRef]
4. Joe, H.; Zhu, R. Generalized Poisson Distribution: The Property of Mixture of Poisson and Comparison with Negative Binomial Distribution. *Biom. J.* **2005**, *45*, 219–229. [CrossRef] [PubMed]
5. Vieira, A.M.C.; Hinde, J.P.; Demetrio, C.G.B. Zero-inflated proportion data models applied to a biological control assay. *J. Appl. Stat.* **2000**, *27*, 373–389. [CrossRef]
6. Conceição, K.S.; Louzada, F.; Andrade, M.G.; Helou, E.S. Zero-modified power series distribution and its Hurdle distribution version. *J. Stat. Comput. Simul.* **2017**, *87*, 1842–1862. [CrossRef]
7. Sáez-Castillo, A.J.; Conde-Sánchez, A. Detecting over- and under-dispersion in zero inflated data with the hyper-Poisson regression model. *Stat. Pap.* **2017**, *58*, 19–33. [CrossRef]
8. da Silva, W.B.; Ribeiro, A.M.T.; Conceição, K.S.; Andrade, M.G.; Neto, F.L. On Zero-Modified Poisson-Sujatha Distribution to Model Overdispersed Count Data. *Austrian J. Stat.* **2018**, *47*, 1–19. [CrossRef]

9. Bonat, W.H.; Jørgensen, B.; Kokonendji, C.C.; Hinde, J.; Demétrio, C.G.B. Extended Poisson–Tweedie: Properties and regression models for count data. *Stat. Model.* **2018**, *18*, 24–49. [CrossRef]
10. Satheesh Kumar, C.; Harisankar, S. On some aspects of a general class of Yule distribution and its applications. *Commun. Stat.-Theory Methods* **2019**, *49*, 1–11. [CrossRef]
11. Rodríguez-Avi, J.; Conde-Sánchez, A.; Sáez-Castillo, A.J.; Olmo-Jiménez, M.J. A triparametric discrete distribution with complex parameters. *Stat. Pap.* **2004**, *45*, 81–95. [CrossRef]
12. Olmo-Jiménez, M.J.; Rodríguez-Avi, J.; Cueva-López, V. A review of the CTP distribution: A comparison with other over- and underdispersed count data models. *J. Stat. Comput. Simul.* **2018**, *88*, 2684–2706. [CrossRef]
13. Rodríguez-Avi, J.; Conde-Sánchez, A.; Sáez-Castillo, A.J. A new class of discrete distributions with complex parameters. *Stat. Pap.* **2003**, *44*, 67–88. [CrossRef]
14. Rodríguez-Avi, J.; Olmo-Jiménez, M.J. A regression model for overdispersed data without too many zeros. *Stat. Pap.* **2017**, *58*, 749–773. [CrossRef]
15. Cueva-López, V.; Olmo-Jiménez, M.J.; Rodríguez-Avi, J. EM algorithm for an extension of the Waring distribution. *Comput. Math. Methods* **2019**, *1*, e1046. [CrossRef]
16. Cueva-López, V.; Olmo-Jiménez, M.J.; Rodríguez-Avi, J. An over- and underdispersed biparametric extension of the Waring distribution. *Mathematics* **2021**, *9*, 170. [CrossRef]
17. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
18. Sellers, K.F.; Borle, S.; Shmueli, G. The COM-Poisson model for count data: A survey of methods and applications. *Appl. Stoch. Model. Bus. Ind.* **2012**, *28*, 104–116. [CrossRef]
19. Sáez-Castillo, A.J.; Conde-Sánchez, A. A hyper-Poisson regression model for overdispersed and underdispersed count data. *Comput. Stat. Data Anal.* **2013**, *61*, 148–157. [CrossRef]
20. Byrd, R.H.; Lu, P.; Nocedal, J.; Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208. [CrossRef]
21. Conover, W.J. A Kolmogorov goodness-of-fit test for discontinuous distributions. *J. Am. Stat. Assoc.* **1972**, *67*, 591–596. [CrossRef]
22. Gleser, L.J. Exact power of goodness-of-fit tests of Kolmogorov type for discontinuous distributions. *J. Am. Stat. Assoc.* **1985**, *80*, 954–958. [CrossRef]
23. Wimmer, G.; Kohler, R.; Grotjahn, R.; Altmann, G. Toward a theory of word length distributions. *J. Quant. Ling.* **1994**, *1*, 98–106. [CrossRef]
24. DjurasErnst, G.; Stadlober, S. *Text and Language: Structures Function Interrelations Quantitative Perspectives*; Chapter Modeling Word Length Frequencies by the Singh-Poisson Distribution; Praesens Verlag: Wien, Austria, 2010; pp. 37–46.