k-Modified Power Series Distributions

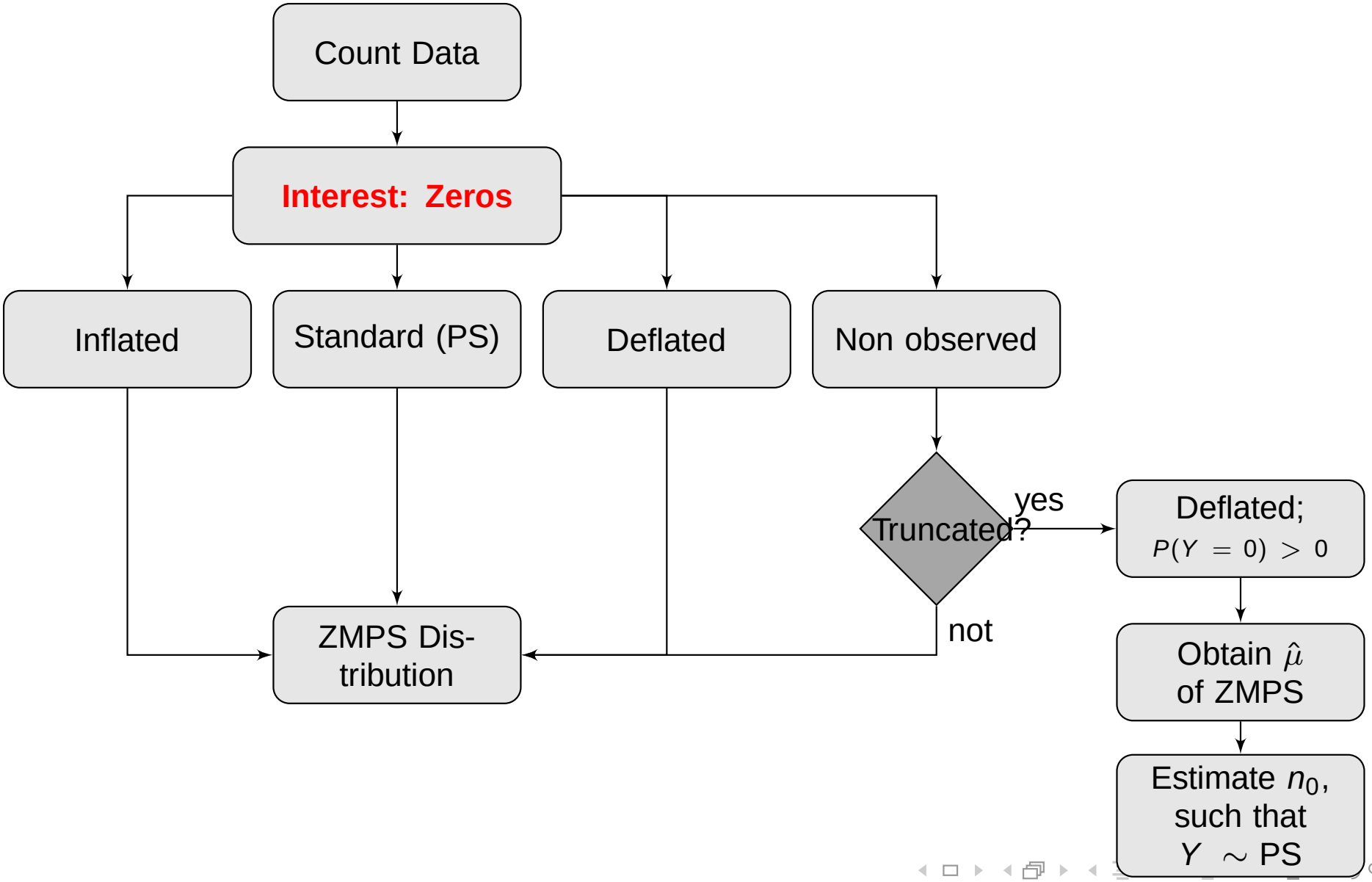
Katiane S. Conceição

SME/ICMC/USP

Abril/2017

Introduction

Motivation



Introduction

Motivation

Some Examples:

1. Notification of a disease by city;
2. Number of defective components in a batch;
3. Number of left-handed individuals in a classroom;
4. Goals scored by a particular team in a match;

Introduction ZMPS Distribution

PS Distributions:

1. Poisson;
2. Generalized Poisson;
3. Geometric;
4. Binomial ($m > 1$);
5. Negative Binomial;
6. Generalized Negative Binomial.

Introduction

ZMPS Distribution

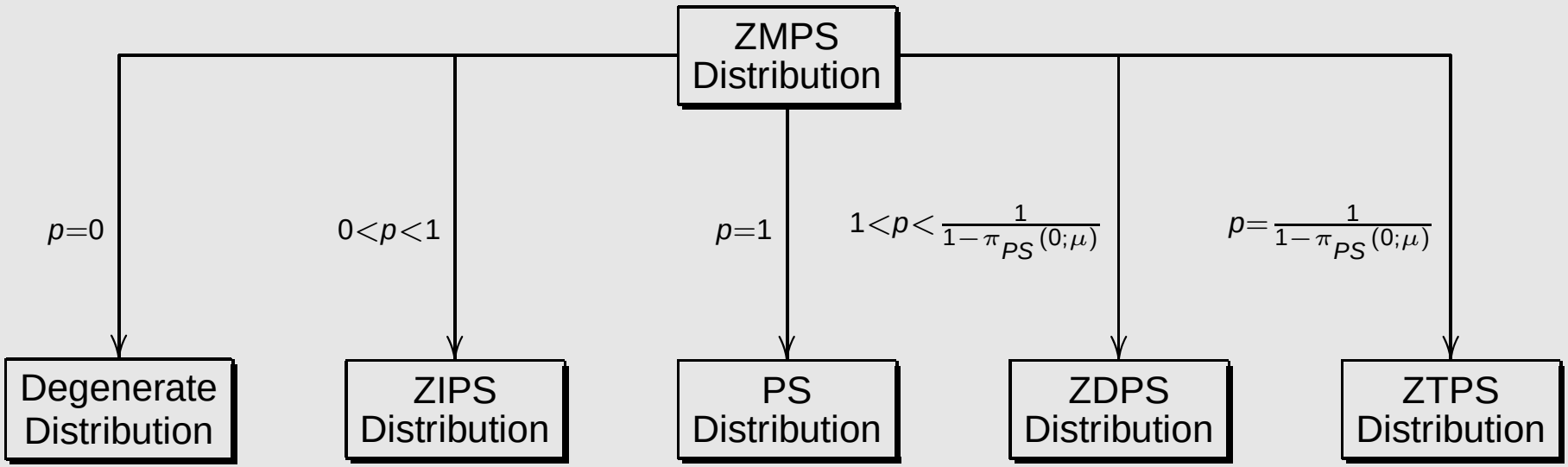


Figura: Diagram of the particular cases of the ZMPS distribution.

Introduction Idea

- There are situations in which the discrepancy between the observed and expected frequencies, according to a PS distribution, occur in some observation not equal to zero.

Introduction

Idea

- There are situations in which the discrepancy between the observed and expected frequencies, according to a PS distribution, occur in some observation not equal to zero.
- Pandey (1965) described a situation using the Inflated Poisson distribution. The number of flowers produced by *Primula veris* plants (excessive number of plants with 8 flowers) is observed.

Introduction

Idea

- There are situations in which the discrepancy between the observed and expected frequencies, according to a PS distribution, occur in some observation not equal to zero.
- Pandey (1965) described a situation using the Inflated Poisson distribution. The number of flowers produced by *Primula veris* plants (excessive number of plants with 8 flowers) is observed.
- Datasets with deflated value k are obtained from experiments in which the value k is observed with frequency significantly lower than expected in a PS distribution.

Introduction — Idea

- An example of k-deflated dataset is the process of counting the number of children in poor families and some of these are part of a social program and receive financial assistance by number of children (few families with only 1 child).

Introduction — Idea

- An example of k-deflated dataset is the process of counting the number of children in poor families and some of these are part of a social program and receive financial assistance by number of children (few families with only 1 child).

- Another example is the process of counting the amount of a particular product purchased per customer in a supermarket. An bid that proposes a decrease in the price of the unit of this product, if the customer chooses to take over k units (with $k > 1$, deflation of k items sold).

Introdution

Objective

Objetivo:

- Extend the idea of modification of zero observation (ZMPS distributions) to any k observation.
  - Thus, we have the k-Modified Power Series (k-MPS) family.
- We will consider the uniparametric distributions belonging to the PS family for the modification at point k.

k-MPS Distribution

Probability Mass Function:

$$\pi_{k-MPS}(y;\mu,p) = (1-p)I_{\{k\}}(y) + p\pi_{PS}(y;\mu),\ y \in \mathcal{A}_S,$$

where:
$\pi_{PS}(y;\mu) = \frac{a(y)[g(\mu)]^y}{f(\mu)}$ is the uniparametric Power Series (PS) distribution associated;
p is the parameter responsible for modifying the probability of observation k, with

$$0 \leq p \leq \frac{1}{1-\pi_{PS}(k;\mu)};$$

and $I(y)$ is the indicator function, with

$$I(y) = \begin{cases} 1, & \text{if } y = k \\ 0, & \text{if } y \neq k \end{cases}.$$

k-MPS Distribution

- The distributions in the ZMPS family are:

1. k-Modified Poisson (k-MP);
2. k-Modified Geometric (k-MG);
3. k-Modified Binomial (k-MB);
4. k-Modified Borel (k-MBo);
5. k-Modified Borel-Tanner (k-MBT);
6. k-Modified Haight (k-MH).

k-MPS Distribution PS Distribution

Tabela: Some distributions of PS family.

PS	Distribution	$f(\mu)$	$g(\mu)$	$a(y)$	$\mathcal{A}_s$
P	Poisson	e^{μ}	μ	$\frac{1}{y!}$	$\{0,1,\ldots\}$
G	Geometric	$1+\mu$	$\frac{\mu}{1+\mu}$	1	$\{0,1,\ldots\}$
B	Binomial	$\left(\frac{m}{m-\mu}\right)^m$	$\frac{\mu}{m-\mu}$	$\binom{m}{y}$	$\{0,1,\ldots,m\}$
Bo	Borel	$1-\frac{1}{\mu}$	$\left(1-\frac{1}{\mu}\right)e^{-1+\frac{1}{\mu}}$	$\frac{y^{y-2}}{(y-1)!}$	$\{1,2,\ldots\}$
BT	Borel-Tanner	$\left(1-\frac{m}{\mu}\right)^m$	$\left(1-\frac{m}{\mu}\right)e^{-1+\frac{m}{\mu}}$	$\frac{my^{y-m-1}}{(y-m)!}$	$\{m,m+1,\ldots\}$
H	Haight	$\frac{\mu-1}{2\mu-1}$	$\frac{\mu(\mu-1)}{(2\mu-1)^2}$	$\frac{(2y-2)!}{y!(y-1)!}$	$\{1,2,\ldots\}$

k-MPS Distribution | Particular Cases

- Different values of p lead to different k-MPS distribution, as can be seen by considering the proportion of additional or of missing Observation k, given by

$$\begin{aligned} \pi_{k-MPS}(k;\mu,p) - \pi_{PS}(k;\mu) &= 1 - p + p\pi_{PS}(k;\mu) - \pi_{PS}(k;\mu) \\ &= (1-p)(-\pi_{PS}(k;\mu)). \end{aligned}$$

Particular Cases of k-MPS(μ, p) Distribution

1. If $p = 0$, k-MPS is the degenerate distribution with all mass at observation k.
2. For all $0 < p < 1$, k-MPS is the k-Inflated Power Series (k-IPS) distribution.
3. If $p = 1$, k-MPS is the PS distribution.
4. For all $1 < p < \frac{1}{1-\pi_{PS}(k;\mu)}$, k-MPS is the k-Deflated Power Series (k-DPS) distribution.
5. If $p = \frac{1}{1-\pi_{PS}(k;\mu)}$, k-MPS is the k-Subtracted Power Series (k-SPS) distribution, with probability mass function given by

$$\pi_{k-SPS}(y;\mu) = \frac{\pi_{PS}(y;\mu)}{1-\pi_{PS}(k;\mu)} \left\{1 - I_{\{k\}}(y)\right\}.$$

k-MPS Distribution | Particular Cases

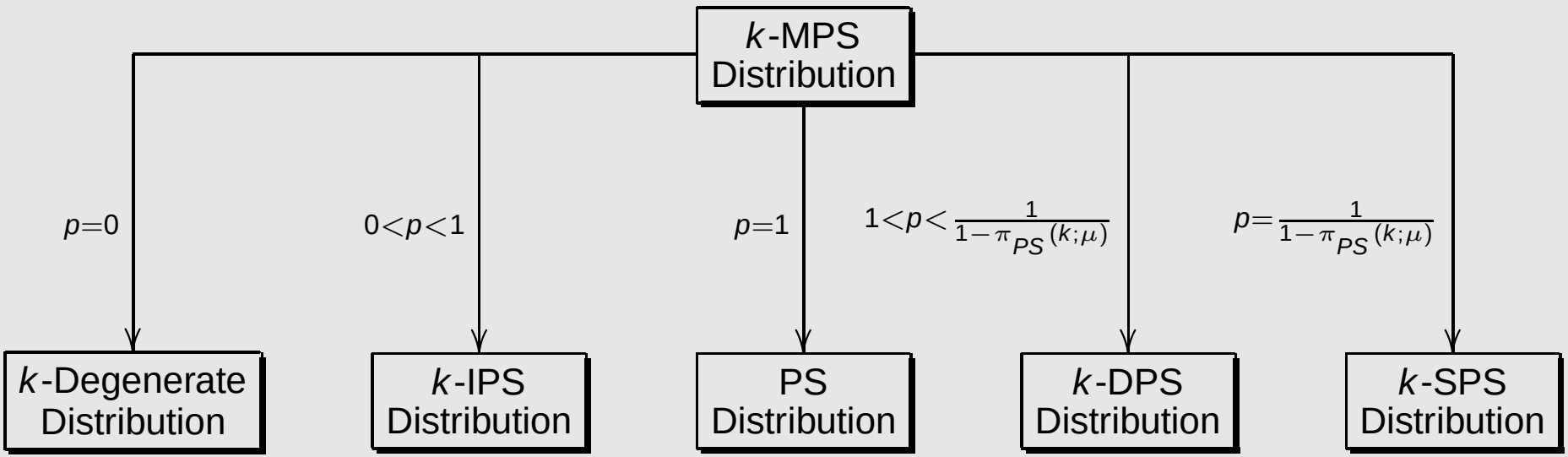


Figura: Diagram of the particular cases of k-MPS distribution.

Observation: The ZMPS distribution is a particular case of the k-MPS distribution when $k = 0$.

Characterizing the k-MPS Distribution

- Consider:
  - $Y \sim k\text{-MPS}(\mu, p)$.
  - $A_s = \{s, s+1, s+1, \ldots\}$, the support of Y.
  - $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ a random sample of Y.
  - b a real constant ($b \in \mathbb{R}$).

Distribution Function:

$$\begin{aligned} F_Y(b) = P(Y \leq b) &= \sum_{y:y\leq b} \pi_{k-MPS}(y;\mu,p) \\ &= (1-p)\sum_{y:y\leq b} I_{\{k\}}(y) + pF(b), \end{aligned}$$

where $F(b)$ is the distribution function of the PS distribution associated with Y at point b. Thus, we have:

$$F_Y(b) = \begin{cases} 0, & \text{se } b < s \\ pF(b), & \text{se } b < k \\ 1 - p(1-F(b)), & \text{se } b \geq k \end{cases}.$$

Probability Generating Function:

$$\begin{aligned} G_Y(z) = E[z^Y] &= \sum_{y=s}^{\infty} z^y \pi_{k-MPS}(y;\mu,p) \\ &= (1-p)z^k + pG(z), \end{aligned}$$

for all z, $|z| \leq 1$, such that $G(z)$ exists, where $G(z)$ is the probability generating function of the PS distribution associated with Y .

Characteristic Function:

$$\begin{aligned}\varphi_Y(t) = E[e^{itY}] &= \sum_{y=s}^{\infty} e^{ity}\pi_{k-MPS}(y;\mu,p) \\ &= (1-p)e^{itk} + p\varphi(t),\end{aligned}$$

where $\varphi(t)$ is the characteristic function of the PS distribution associated with Y.

Moment Generating Function:

$$\begin{aligned} \mathcal{M}_Y(t) = E[e^{tY}] &= \sum_{y=s}^{\infty} e^{ty} \pi_{k-MPS}(y; \mu, p) \\ &= (1-p)e^{tk} + p\mathcal{M}(t), \end{aligned}$$

for all t such that $\mathcal{M}(t)$ exists, where $\mathcal{M}(t)$ is the moment generating function of the PS distribution associated with Y.

- The r-th moment of the k-MPS family, denoted by $\mu^r_{k-MPS} = E(Y^r)$, is obtained by evaluating the r-th derivative of $\mathcal{M}_Y(t)$ at $t = 0$, defined as

$$\begin{aligned} \mu^r_{k-MPS} = E(Y^r) &= \frac{\partial^r}{\partial t^r}\mathcal{M}_Y(t)\Big|_{t=0} \\ &= (1-p)k^r + p\mathcal{M}^{(r)}(t)\Big|_{t=0} \\ &= (1-p)k^r + p\mu^r, \ \ \forall\, r \geq 1, \end{aligned}$$

where μ^r is the r-th populational moment of the PS distribution associated with Y.

Mean and Variance

Mean:

$$\mu_{k-MPS} = E(Y) = (1-p)k + p\mu.$$

Variance:

$$\sigma^2_{k-MPS} = Var(Y) = p\{\sigma^2 + (1-p)(k-\mu)^2\},$$

where σ^2 is the variance of the PS distribution associated with Y.

k-MPS Distribution $\times$ Hurdle Version:

The probability mass function $\pi_{k-MPS}(y;\mu,p)$ can be written as a hurdle distribution by

$$\pi_{k-MPS}(y;\mu,\omega) = (1-\omega)I_{\{k\}}(y) + \omega\pi_{k-SPS}(y;\mu),\ y \in A_S,$$

where $\omega = p(1-\pi_{PS}(k;\mu))$ is the parameter responsible for modifying the probabilities of k-SPS distribution, such that

$$0 \leq \omega \leq 1.$$

k-MPS Distribution | Hurdle Version

- With the k-MPS(μ, ω) distribution, We have:
  1. the event $y = k$ occurs with probability $1 - \omega$; and
  2. the event $y \neq k$ occurs with probability $\omega \pi_{k-SPS}(y; \mu)$
- That is, one that produces positive observations from a k-SPS distribution and another that produces only observations k.

Particular Cases of k-MPS(μ, ω) Distribution

1. If $\omega = 0$, $\pi_{k-MPS}(y; \mu, \omega)$ is the degenerate distribution with all mass at observation k.
2. For all $0 < \omega < 1 - \pi_{PS}(k; \mu)$, $\pi_{k-MPS}(y; \mu, \omega)$ is the k-IPS distribution.
3. If $\omega = 1 - \pi_{PS}(k; \mu)$, $\pi_{k-MPS}(y; \mu, \omega)$ is the PS distribution.
4. For all $1 - \pi_{PS}(k; \mu) < \omega < 1$, $\pi_{k-MPS}(y; \mu, \omega)$ is the k-DPS distribution.
5. If$\omega = 1$, $\pi_{k-MPS}(y; \mu, \omega)$ is the k-SPS distribution.

Estimation of the Parameters

- Let Y be a random variable with k-MPS distribution.
- Consider $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ a random sample of Y e the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ of $\boldsymbol{Y}$.
- Denote by $\mathcal{D} = (\boldsymbol{y}, n, n_j)$ the vector of information related to $\boldsymbol{y}$, where n is the total number of observations and n_j, $j = s, s+1, s+2, \ldots,$, is the number of observations j in vector $\boldsymbol{y}$.
- For estimation of the parameters, we consider the method **Method of Maximum Likelihood**.

Estimation of the Parameters log-Likelihooh Function

k-MPS(μ, p):

The natural logarithm of the likelihood function is given by

$$\ell(\mu, p; \mathcal{D}) = n_k \log\left(1 - p + p\pi_{PS}(k; \mu)\right) + \sum_{j=s; s\neq k}^{\infty} n_j \log\left(p\pi_{PS}(j; \mu)\right),$$

such that $n = n_k + \sum_{j=s; j\neq s}^{\infty} n_j$.

O MLE for p parameter is given by

$$\hat{p} = \frac{n - n_k}{n(1 - \pi_{PS}(k; \hat{\mu}))}.$$

Estimation of the Parameters log-Likelihooh Function

k-MPS(μ, ω):

The natural logarithm of the likelihood function is given by

$$\begin{aligned}\ell(\mu, \omega; \mathcal{D}) &= n_k \log(1-\omega) + (n - n_k)\log(\omega) + \sum_{j=s; j\neq k}^{\infty} n_j \log\left(\frac{\pi_{PS}(j;\mu)}{1-\pi_{PS}(k;\mu)}\right) \\ &= \ell_1(\mu; \mathcal{D}) + \ell_2(\omega; \mathcal{D}).\end{aligned}$$

O MLE for ω parameter is given by

$$\hat{\omega} = \frac{n - n_k}{n}.$$

Estimation of the Parameters

Theorem:

If Y is a random variable with k-MPS distribution, then the maximum likelihood estimator of parameter μ for this distribution is equal to the maximum likelihood estimator obtained for the parameters of the k-SPS distribution associated, considering only the observations different of k in $\boldsymbol{y}$.

Leptospirosis Notifications in Cities of Bahia State

- We consider the dataset from leptospirosis notifications reported in cities of Bahia State in 2004;
- Bahia State has 417 cities;
- Two cities, Barrocas and Luís Eduardo MagalhŸães, were excluded from the study because they were built recently (less than 10 years old).
- Only 50 cities reported notification of the disease.

Aplications: Real Data — Leptospirosis Notifications

Tabela: Frequency distribution of leptospirosis notifications in cities of Bahia State in 2004.

y_i	0	1	2	3	4	5	6	11	12	14	16	111
f_i	365	29	7	3	2	2	1	2	1	1	1	1

Aplications: Real Data — Leptospirosis Notifications

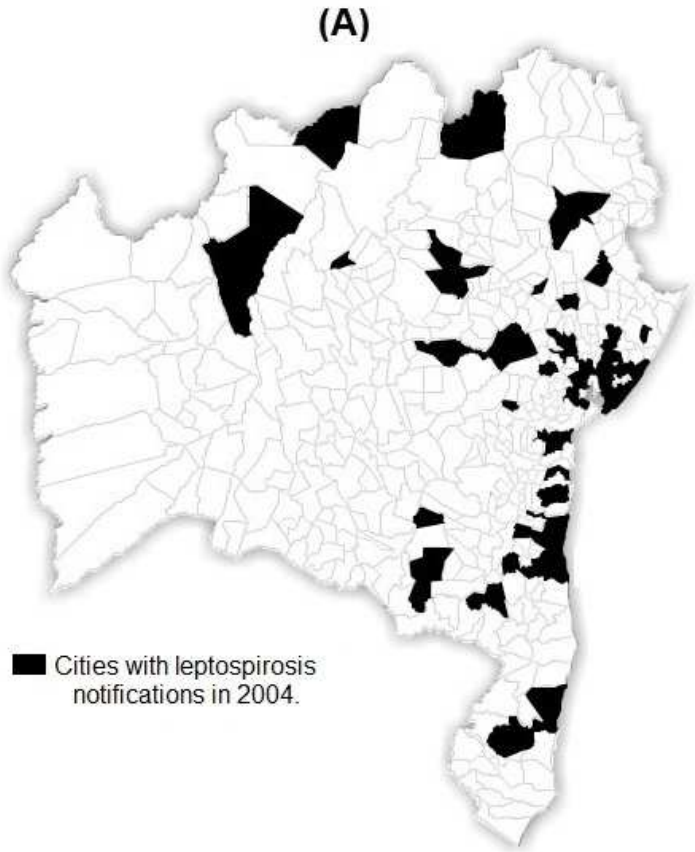


Figura: Map of Bahia State. **(A)** Cities that submitted notifications in 2004.

Aplications: Real Data — Leptospirosis Notifications

Tabela: Maximum likelihood estimates of the parameters of the k-MP distribution adjusted for the leptospirosis notification dataset.

Data	$\hat{\mu}$ (IC 95%)	$\hat{p}$ (IC 95%)	$\hat{n}_0$		AIC	BIC
			k-M	P		
Complete	4,98 (4,36; 5,61)	0,12 (0,09; 0,15)	362	3	1117,98	1126,03
Removed obs. 415	2,66 (2,16; 3,15)	0,13 (0,09; 0,16)	336	29	571,92	579,97

Modification only for $k = 0$.

Aplications: Real Data — Leptospirosis Notifications

- **Cities at Risk:**
  - That group consists of cities which reported notifications of the disease in 2004 and/or the year before (2003).
  - The risk group is composed only by cities that have Human Development Index (HDI) less that 0.66.
  - Thus, we have a sub-sample of leptospirosis notifications which occurred in cities at risk with low HDI in 2004, whose zeros are related only to cities that had cases in the previous year (2003), but not presented in the following year (in 2004).

Aplications: Real Data — Leptospirosis Notifications

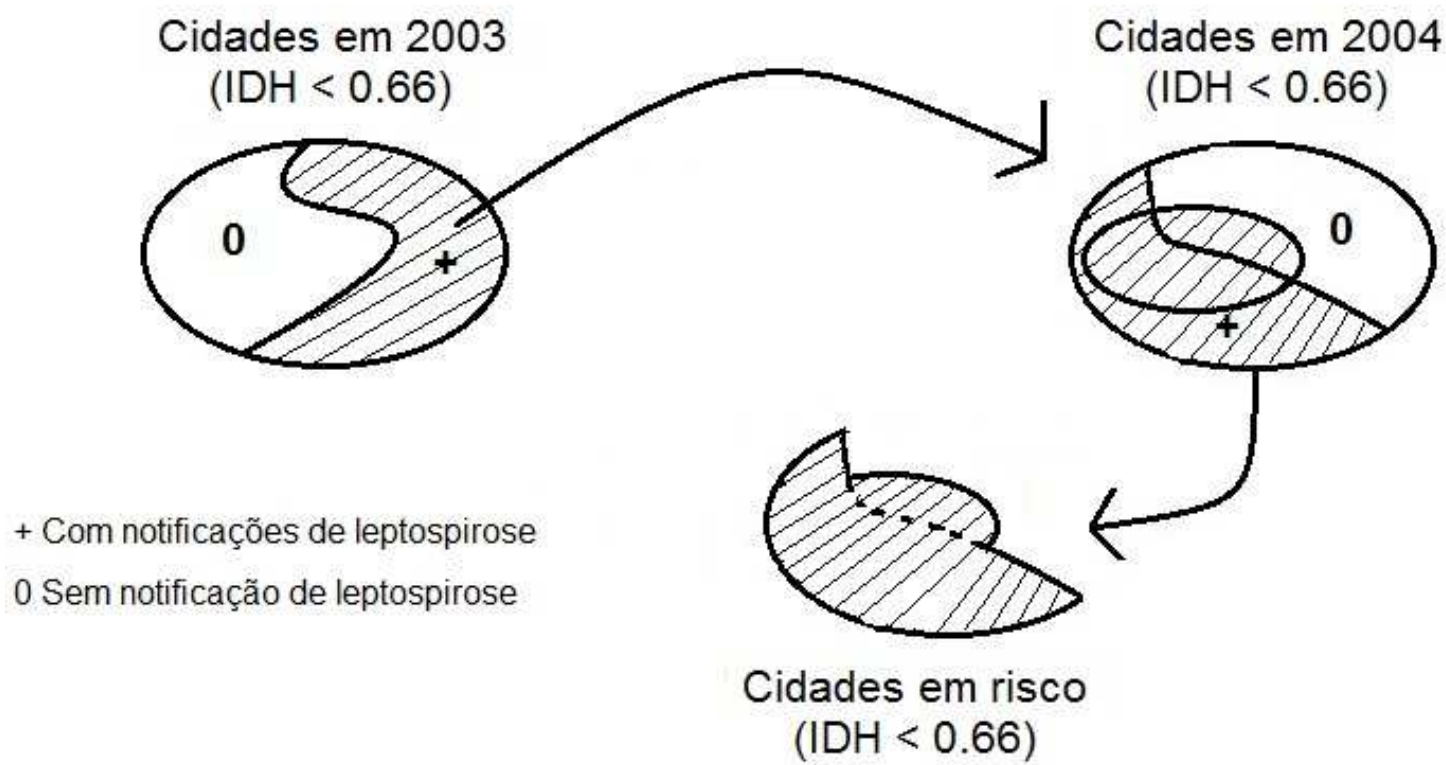


Figura: Diagram representing the sub-sample of cities at risk in 2004, with HDI $<$ 0.66.

Aplications: Real Data — Leptospirosis Notifications

- The sub-sample consists of 13 cities in 2003 and 30 cities in 2004 that reported notifications;
- Only 4 cities of this sub-sample reported notification in both years;
- Therefore, the resulting sample contains notifications from 39 cities.

Aplications: Real Data — Leptospirosis Notifications

Tabela: Frequency distribution of leptospirosis notifications in cities at risk of Bahia State in 2004.

y_i	0	1	2	3
f_i	9	25	4	1

Aplications: Real Data — Leptospirosis Notifications

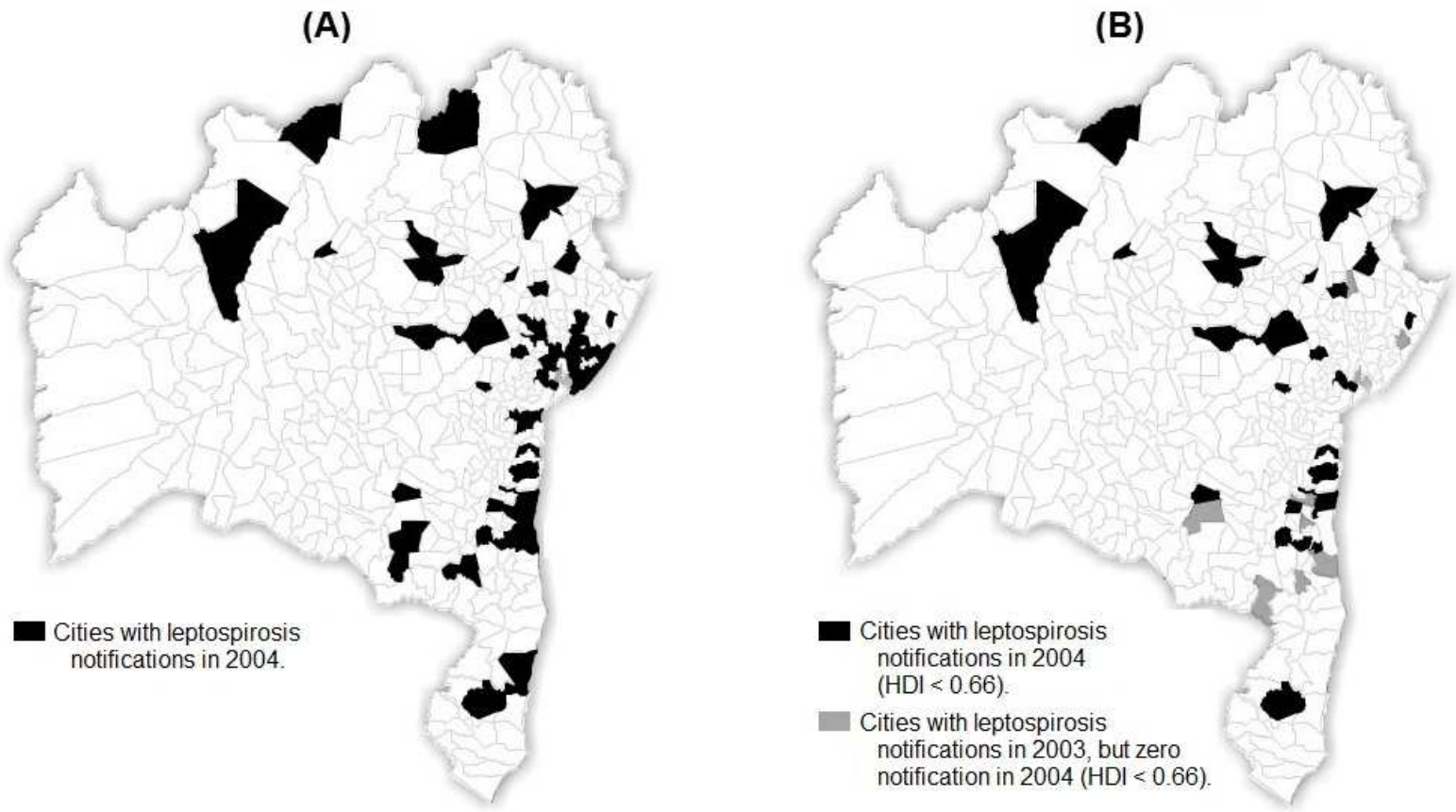


Figura: Map of Bahia State. **(A)** Cities that submitted notifications in 2004. **(B)** Cities in the risk group in 2004, with HDI<0.66.

Aplications: Real Data — Leptospirosis Notifications

Tabela: Maximum likelihood estimates of the parameters of the k-MP distribution adjusted for the leptospirosis notification dataset in cities at risk, with HDI < 0.66.

k	$\hat{\mu}$ (IC 95%)	$\hat{p}$ (IC 95%)	$\hat{n}_k$		AIC	BIC
			k-M	P		
0	0,38 (0,08; 0,67)	2,45 (2,03; 2,87)	-18	27	78,64	81,96
1	0,86 (0,47; 1,25)	0,56 (0,33; 0,80)	11	14	78,78	82,11

Aplications: Real Data

Leptospirosis Notifications

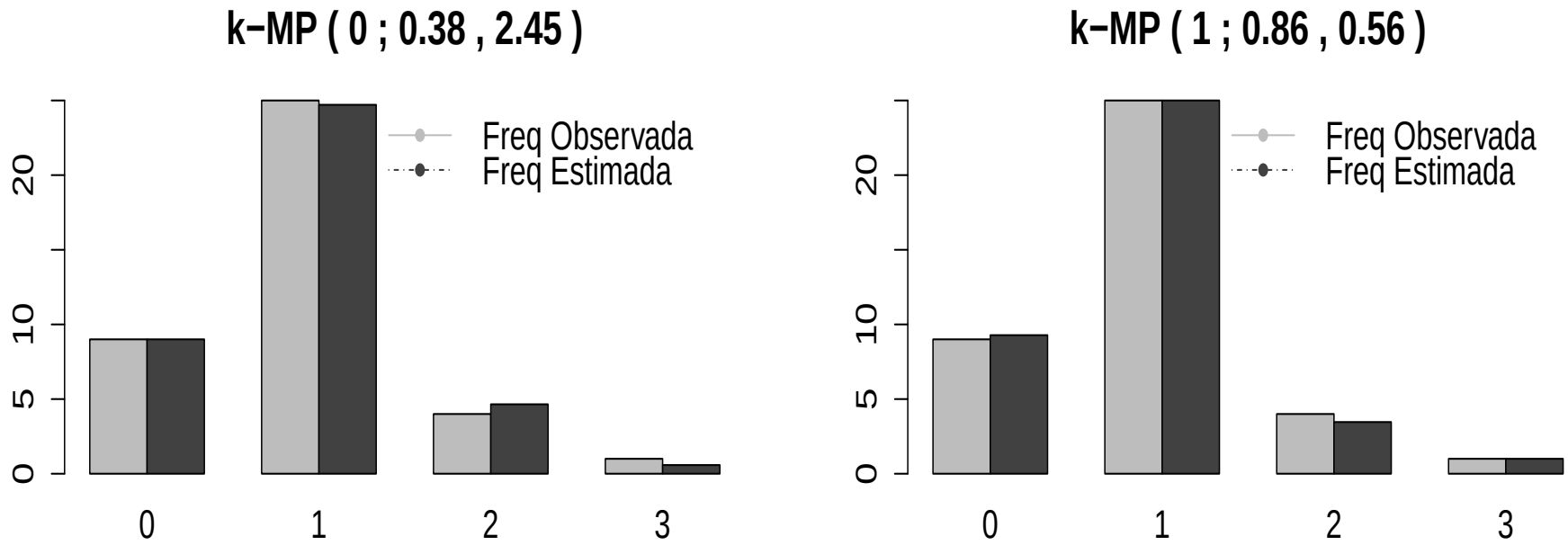


Figura: Observed and expected frequency under the leptospirosis notifications in cities at risk, with HDI < 0.66 in 2004, considering the adjusts of k-MP distribution, with $k = 0$ e 1.

Aplications: Real Data — Number of Goals

Number of Goals in Soccer Matches

- We considered a dataset regarding the number of goals scored by Barcelona FC in all the matches with the Real Madrid CF between the years 1955 and 2015.
- These teams faced each other 131 times. Of these matches, 68 games resulted in victory of Barcelona , 62 games resulted in victory of Real Madrid and only 1 draw.

Aplications: Real Data — Number of Goals

Tabela: Frequency distribution of the number of goals scored by Barcelona in all matches with Real Madrid in the 1955 to 2015 period.

y_i	0	1	2	3	4	5	6	Total
f_i	27	39	34	21	5	3	2	131

Aplications: Real Data — Number of Goals

Tabela: Parameter estimates μ and p of the k-MP distribution, adjusted to the number of goals scored by Barcelona in all matches with Real Madrid between 1955 and 2015.

k	$\hat{\mu}$ (IC 95%)	$\hat{p}$ (IC 95%)	$\hat{n}_k$		AIC	BIC
			k-M	P		
0	1,70 (1,41; 1,99)	0,97 (0,89; 1,06)	3	24	427,85	433,60
1	1,63 (1,40; 1,86)	1,03 (0,92; 1,15)	-3	42	427,94	433,69
2	1,65 (1,43; 1,87)	1,00 (0,90; 1,10)	0	34	428,17	433,92

Aplications: Real Data — Number of Goals

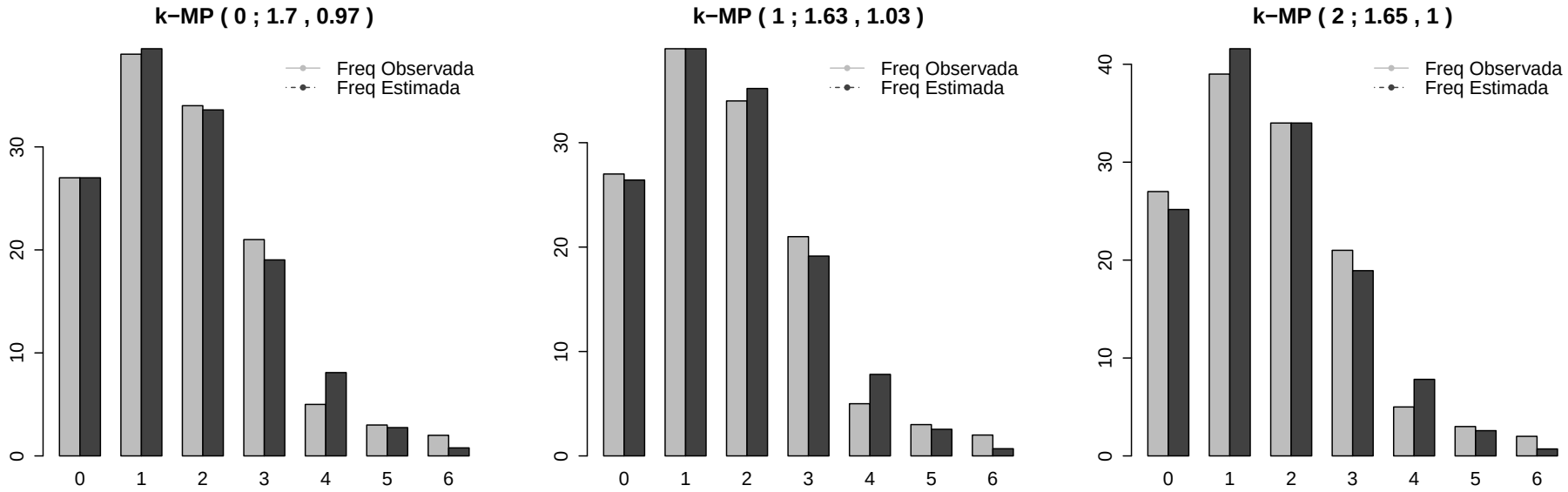


Figura: Observed and expected frequency of the number of goals scored by Barcelona in all matches with Real Madrid from 1955 to 2015, considering the adjustments from k-MG to $k = 0, 1$ and 2.

Global Temperature Variation

- We consider the dataset referring to the annual variation of the global temperature (in degree Celsius, $^{\circ}C$) in the period between 1958 and 2008.

- This variation was obtained with the deviation between the average annual temperatures and the average of the annual temperatures of the first 20 years considered in this study (1958 to 1977).

Applications: Real Data — Global Temperature Variation

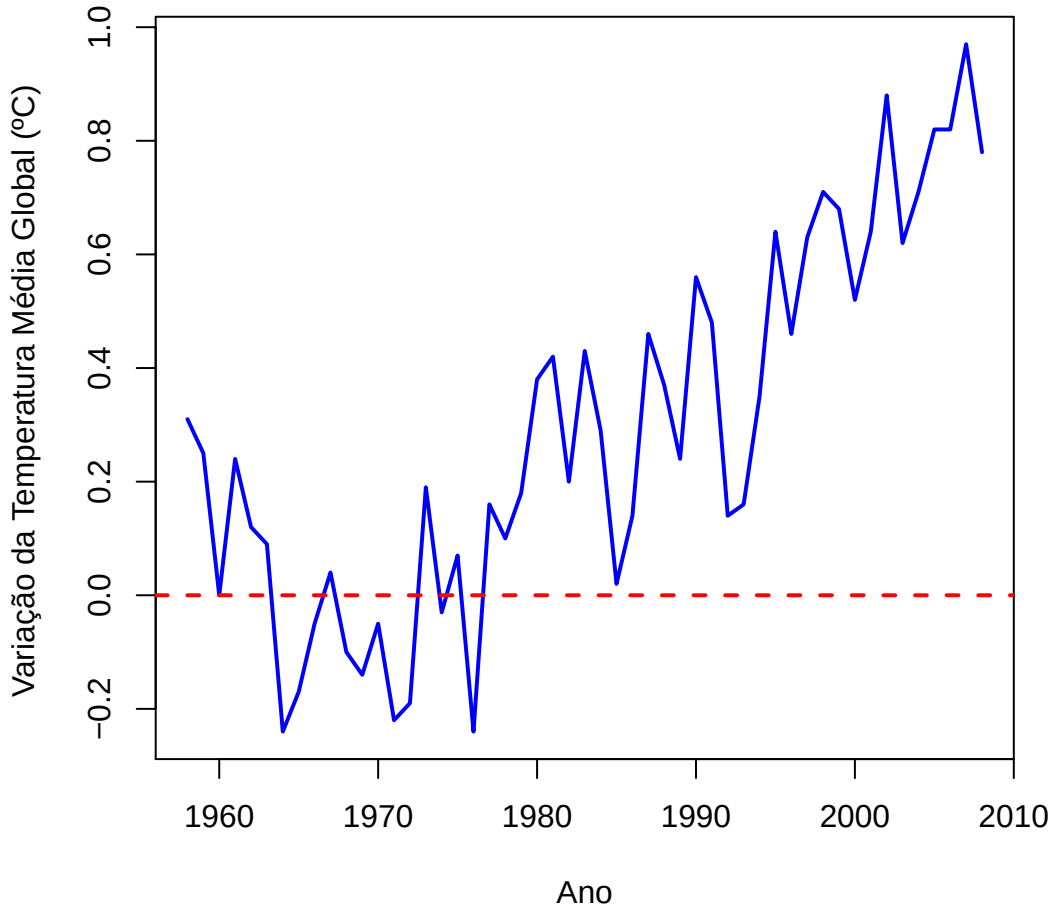


Figura: Variation ($^{\circ}C$) of the average global temperature in the period between 1958 and 2008.

Applications: Real Data — Global Temperature Variation

- ***Bernoulli* experiment:** We verified whether a positive variation (success, a relative increase of the mean global temperature) or a negative variation (failure, a relative decrease of the mean global temperature).

- From the *Bernoulli* process, we define a random variable Y as the **number of consecutive years in which there was a negative average annual temperature variation (failure) until the occurrence of a positive variation** (success).

Applications: Real Data — Global Temperature Variation

Tabela: Frequency distribution of the number of consecutive years with negative variation of the average annual temperature until the occurrence of a positive variation in the period between 1958 and 2008.

y_i	0	1	3	5	Total
f_i	34	3	1	1	39

- It is natural to assume a **Geometric distribution** for Y.

Applications: Real Data — Global Temperature Variation

Tabela: Parameter estimates μ and p of the k-MG distribution, adjusted to the dataset of the number of consecutive years with negative variation of the average annual temperature until the occurrence of a positive variation between 1958 and 200.

k	$\hat{\mu}$ (IC 95%)	$\hat{p}$ (IC 95%)	$\hat{n}_k$		AIC	BIC
			k-M	G		
0	1,20 (0; 2,62)	0,24 (0,04; 0,43)	16	18	49,03	52,36
1	0,38 (0,14; 0,61)	1,15 (1,05; 1,26)	-5	8	52,95	56,28

Applications: Real Data — Global Temperature Variation

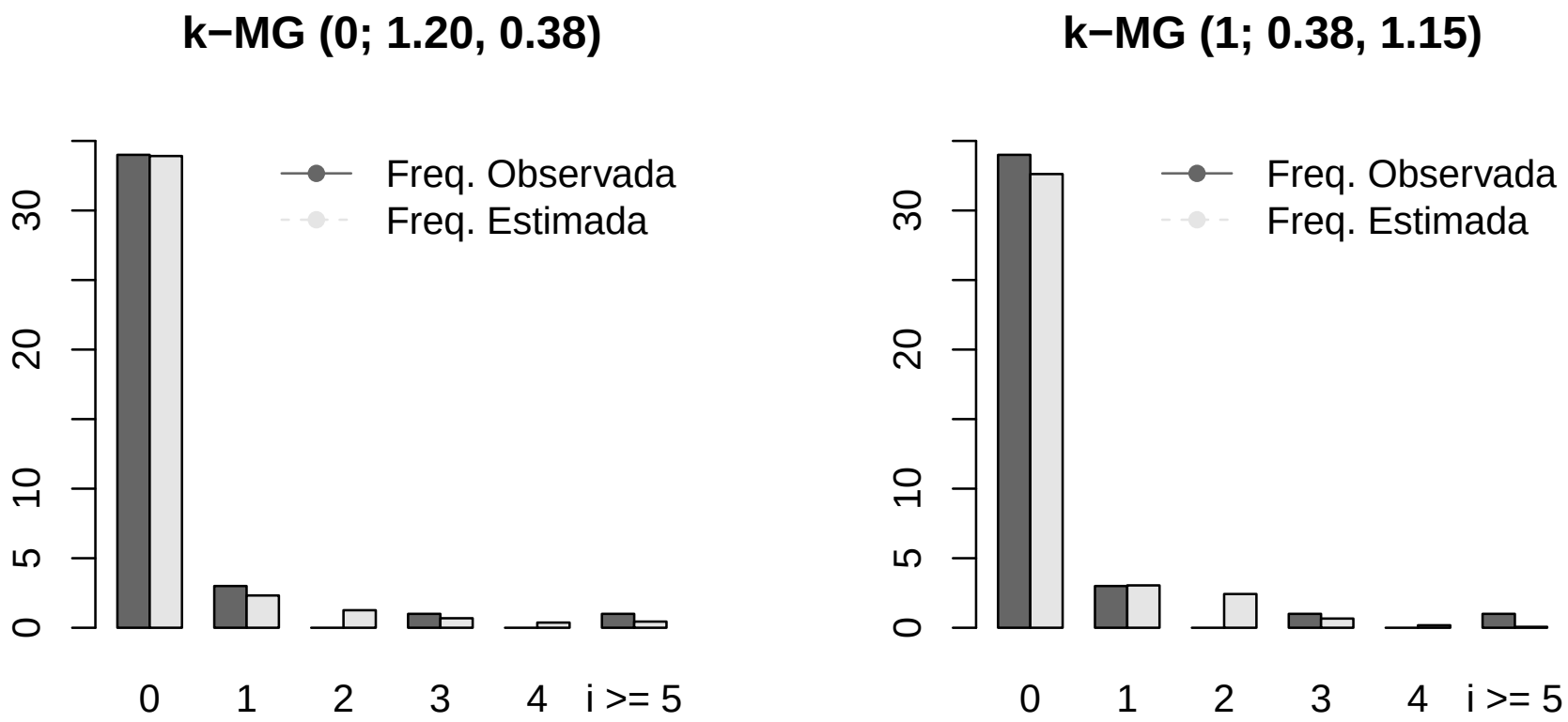


Figura: Observed and expected frequency of the number of consecutive years with negative variation of the average annual temperature until the occurrence of a positive variation in the period between 1958 and 2008, considering the adjustments of k-MG to $k = 0$ and 1.

Meaning of technical words in Statistics

- We evaluated the knowledge of undergraduate students in Statistics regarding the translation into Portuguese of the following technical terms in English: “*Average*”(média) e “*Standard Deviation*”(desvio padrão).

- 59 students who were already enrolled in the undergraduate course for at least one year were asked about the translation of the words.

Applications: Real Data — Meaning of technical words in Statistics

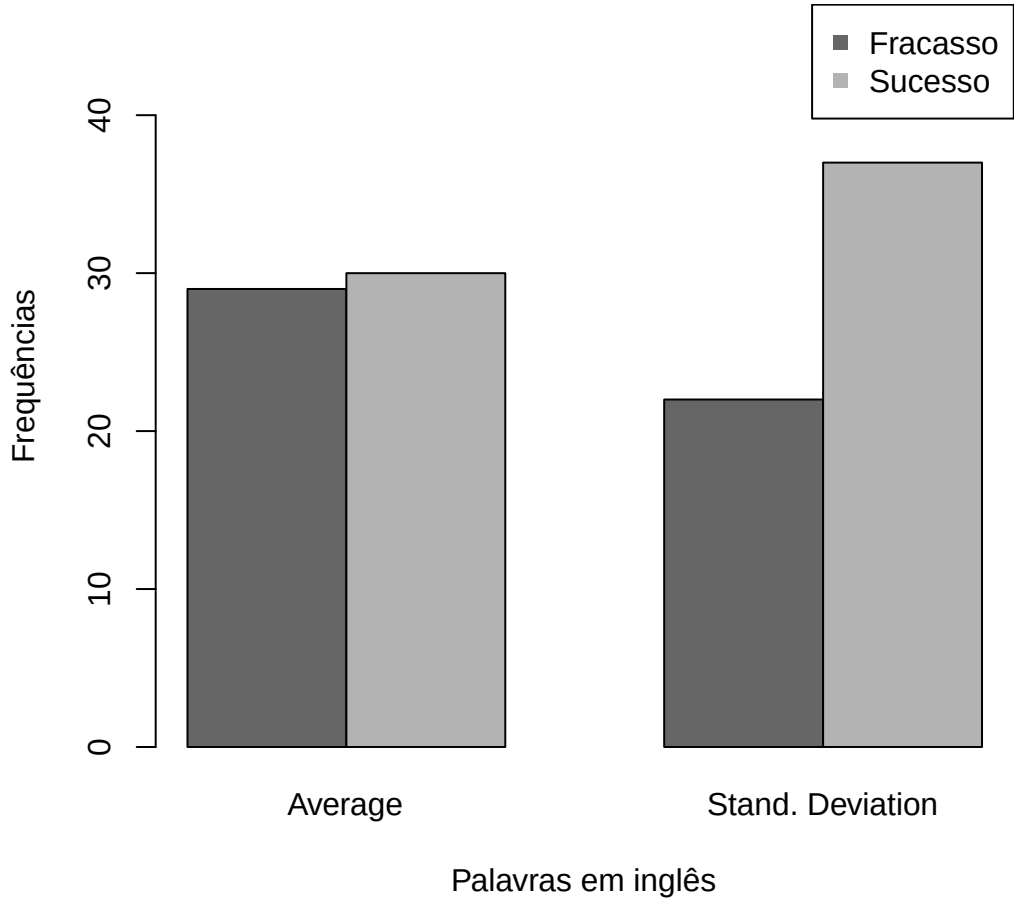


Figura: Frequency distribution of the number of correct answers (success) and error (failure) in translations of technical words in English made by students.

Applications: Real Data — Meaning of technical words in Statistics

- It was verified that the correct number of each word by a student consists of a *Bernoulli* experiment with the same probability of success $\theta \in (0, 1)$.
- In this way, we can define the random variable of interest Y as the total number of correct answers obtained by a student.

Applications: Real Data — Meaning of technical words in Statistics

Tabela: Frequency distribution of the number of correct answers in translations of technical words in English made by students.

y_i	0	1	2	Total
f_i	18	15	26	59

Applications: Real Data — Meaning of technical words in Statistics

Tabela: Parameter estimates μ and p of the k-MB distribution, adjusted to the dataset of the number of correct answers in translations of technical words in English made by students.

k	$\hat{\mu}$ (IC 95%)	$\hat{p}$ (IC 95%)	$\hat{n}_k$		AIC	BIC
			k-M	B		
0	1,55 (1,33; 1,77)	0,73 (0,61; 0,86)	15	3	130,43	134,59
1	1,09 (0,94; 1,24)	1,48 (1,26; 1,70)	-14	29	130,43	134,59
2	0,59 (0,30; 0,87)	0,61 (0,47; 0,75)	21	5	130,43	134,59

Applications: Real Data — Meaning of technical words in Statistics

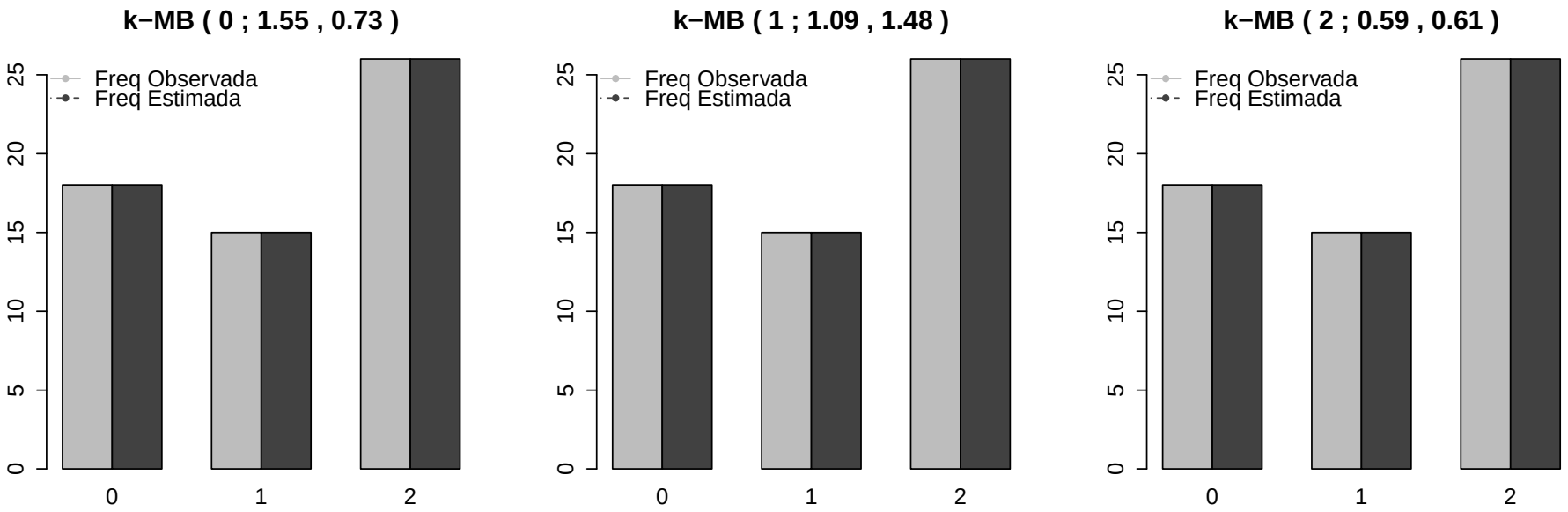


Figura: Observed and expected frequency of the total number of correct translations of the words, considering the adjustments of the k-MB for $k = 0, 1$ and 2.

Final comments

- In general, the k-MPS distributions can be used to analyze k-inflation counted datasets, k-deflated or even for datasets coming from a standard PS distribution.
- The maximum likelihood method was used to obtain the parameter estimators of the biparametric k-MPS distributions, which proved to be efficient.
- In the analysis of real datasets, we obtained quite satisfactory results, allowing us to classify the type of k-modification existing in each real dataset.

Final comments

- The adequacy of the adjusted distribution to the dataset was observed, mainly, when we compared the observed frequencies and expected frequencies.
- Therefore, it is recommended to use the k -MPS distributions as an alternative to the standard discrete distributions, since prior knowledge about the type of k-modification present in the data is not necessary.

Future Research Proposals

- Consider the method of moments to adjust the k-MPS biparametric distributions and compare the performance of this with the method presented here;
- To introduce the k-MPS distribution in the context of regression models;
- Consider the k-MPS models for spatially correlated datasets.

Some References

Angel, J. K. (2008). Air Resources Laboratory, National Oceanic and Atmospheric Administration. `http:www.noaa.gov`. Acessado em Fevereiro de 2016.

Conceição, K. S. (2013). *Modelos Séries de Potência Zero-Modificados*. Tese de doutorado, Programa de Pós-Graduação em Estatística, Universidade Federal de São Carlos, São Carlos/SP.

Consul, P. C. (1990). New Class of Location-Parameter Discrete Probability Distributions and Their Characterizations. *Communications in Statistics - Theory and Methods*, 19 (12), 4653-4666.

Cordeiro, G. M.; Andrade, M. G.; de Castro, M. (2009). Power Series Generalized Nonlinear Models. *Computational Statistics & Data Analysis*, 53, 1155-1166.

Pandey, K. N. (1965). Generalized inflated Poisson distribution. J*ournal of Science and Research Banaraes Hindu University*, 15(2), 157-162.

Thanks

THANK YOU!