



Athens, July 1-3, 2019

Alternative count regression models for modeling football outcomes

Alessandro Barbiero

Department of Economics, Management and Quantitative Methods

University of Milan, Italy

alessandro.barbiero@unimi.it

Introduction

- ⚽ The **bivariate Poisson model**, with independent components, was probably the first used in football data analysis for modeling the outcome of a game (number of goals scored by the two competing teams) due to its ease of use and interpretation.
- ⚽ Later, more complex models allowing for **non-null correlation** were explored, since real data often show a slight but non-negligible positive correlation between the numbers of goals scored by the two teams; or allowing for **overdispersion** and **excess in $(0 - 0)$ draws**, which usually characterize football outcomes (diagonally-inflated models).
- ⚽ In this work, we introduce a **discrete counterpart** to the continuous Weibull model and use it to model the number of scored goals in a match; we control dependence between scored goals of the two competing teams through **copulas**. In this direction, two contributions have already been presented by Boshkanov et al. (2017) and Barbiero (2018).

Type I DW

A continuous Weibull rv T has probability density function given by

$$f_t(t; \lambda, \beta) = \lambda \beta t^{\beta-1} e^{-\lambda t^\beta} \quad t > 0,$$

with $\lambda, \beta > 0$, and cdf

$$F_t(t; \lambda, \beta) = 1 - e^{-\lambda t^\beta}.$$

If we consider the rv $Y = \lfloor T \rfloor$, where $\lfloor T \rfloor$ its pmf defined on \mathbb{N} is given by

$$p(y; q, \beta) = F_t(y+1) - F_t(y) = e^{-\lambda y^\beta} - e^{-\lambda (y+1)^\beta} = q^{y^\beta} - q^{(y+1)^\beta}$$

with $q = e^{-\lambda}$, and then $0 < q < 1$. The corresponding cdf is

$$F(y; q, \beta) = 1 - q^{(y+1)^\beta} \quad y \in \mathbb{N}.$$

Note that $p(0) = 1 - q \quad \forall \beta > 0$

Type I DW

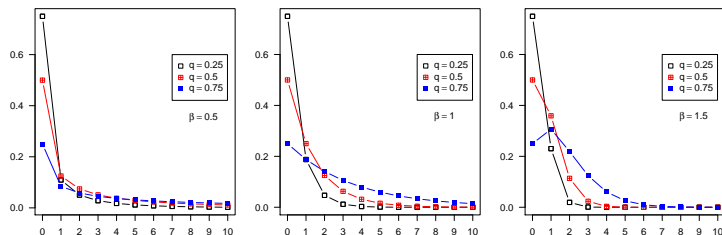


Figure: Graphs of the pmf of type I DW for some values of q and β

$$\mathbb{E}(Y) = \sum_{y=0}^{\infty} q^{(y+1)\beta} < \mathbb{E}(T) = \left(-\frac{1}{\log q}\right)^{\frac{1}{\beta}} \Gamma\left(1 + \frac{1}{\beta}\right) < \mathbb{E}(Y) + 1$$

$$\mathbb{E}(Y^2) = 2 \sum_{y=0}^{\infty} y q^{(y+1)\beta} + \mathbb{E}(Y)$$

→ for a fixed value of q , $\mathbb{E}(Y)$ decreases with β ;

→ for a fixed value of β , $\mathbb{E}(Y)$ increases with q

Type I DW

Table: Expected value and Variance for the type I DW, for several combinations of q and β . **Overdispersion in red colour**

	$\beta = 1.25$		$\beta = 1.5$		$\beta = 1.75$		$\beta = 2$		$\beta = 2.25$		$\beta = 2.5$	
q	mean	var	mean	var	mean	var	mean	var	mean	var	mean	var
0.5	0.78	1.00	0.67	0.64	0.61	0.47	0.56	0.38	0.54	0.32	0.52	0.29
0.6	1.12	1.64	0.93	0.95	0.81	0.65	0.74	0.49	0.69	0.40	0.66	0.34
0.7	1.64	2.93	1.30	1.53	1.11	0.96	0.98	0.68	0.90	0.52	0.84	0.42
0.8	2.60	6.21	1.96	2.83	1.60	1.60	1.38	1.04	1.22	0.74	1.11	0.57
0.9	5.14	20.61	3.55	7.61	2.72	3.69	2.23	2.12	1.91	1.37	1.68	0.96

DW regression

For match i in a round-robin tournament, we observe the final result (Y_{1i}, Y_{2i}) , and we assume

$$Y_{1i} \sim \text{DW}(q_{1i}, \beta_{1i})$$

$$Y_{2i} \sim \text{DW}(q_{2i}, \beta_{2i})$$

- ⚽ The first parameter q of the type I DW model, that can be interpreted as the probability of scoring, can be related to explanatory variables \mathbf{x}_i through a complementary log-log link function:

$$\log(-\log(q_i)) = \boldsymbol{\alpha}'\mathbf{x}_i.$$

or the logit function $\log[q_i/(1 - q_i)] = \boldsymbol{\alpha}'\mathbf{x}_i$.

- ⚽ Additionally, even the second parameter β can be related to explanatory variables \mathbf{z}_i through the following natural link function (remember that β takes only positive values):

$$\log(\beta_i) = \boldsymbol{\gamma}'\mathbf{z}_i.$$

Introducing dependence via copulas

We can accommodate dependence between the numbers of goals scored by the two competing teams in a football match resorting to copulas.

Once a bivariate copula $C(u_1, u_2; \theta)$ is selected, the joint cdf of the number of goals scored by home and away team in match i is

$$F(y_{1i}, y_{2i}) = C(F_{1i}(y_{1i}), F_{2i}(y_{2i}); \theta),$$

so that the joint pmf is derived as

$$P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) = F(y_{1i}, y_{2i}) - F(y_{1i} - 1, y_{2i}) - F(y_{1i}, y_{2i} - 1) + F(y_{1i} - 1, y_{2i} - 1).$$

Modelling correlation between the number of scored and conceded goals

From among the multitude of parametric bivariate copulas, we pick Clayton's copula, belonging to the so-called Archimedean family. The expression of the one-parameter Clayton copula is


$$C(u_1, u_2) = \max \left\{ (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, 0 \right\}, \quad \theta \in (-1, +\infty) \setminus \{0\}.$$

It is a **comprehensive** copula, i.e., it can model various kinds of dependence, ranging from

- ⊛ comonotonicity in the limit as $\theta \rightarrow +\infty$,
- ⊛ independence if $\theta \rightarrow 0$, and
- ⊛ countermonotonicity if $\theta \rightarrow -1$

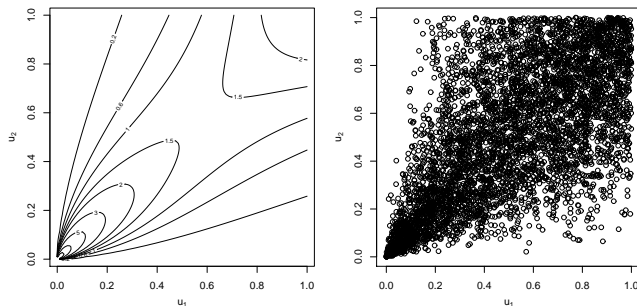
The coefficient of lower tail dependence, defined as

$$\lambda_L = \lim_{u \rightarrow 0^+} P(X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)) = \lim_{u \rightarrow 0^+} C(u, u)/u,$$

for the Clayton copula with $\theta > 0$ is $\lambda_L = 2^{-1/\theta} > 0$. 

Clayton copula

Figure: Contour density plot and scatter plot of N observations from a Clayton copula



The Clayton copula may be a suitable candidate since it may capture the empirical frequency of 0-0 draws better than standard stochastic models.

Possible regression models

- ⚽ **parameter q for the home team h_i** depending on the attack ability of the home team and defense ability of the away team, and on the home effect;
- ⚽ **parameter q for the away team a_i** depending on the attack ability of the away team and defense ability of the home team:

$$\begin{cases} \log[-\log(q_{1i})] &= \mu^{(q)} + \text{home}^{(q)} + \text{att}_{h_i}^{(q)} + \text{def}_{a_i}^{(q)} \\ \log[-\log(q_{2i})] &= \mu^{(q)} + \text{att}_{a_i}^{(q)} + \text{def}_{h_i}^{(q)} \end{cases}$$

- ⚽ **parameter β :**
 - ⌚ constant
 - ⌚ depending on the team only (not on the competing team)
 - ⌚ depending on the team abilities and home effect as q
- ⚽ **parameter θ :** constant
- ⚽ parameters can be estimated via **full maximum likelihood method**: initial common estimates for q and β can be obtained as $\arg \max \ell(q, \beta; y_1, \dots, y_n)$, with $y_i \sim \text{DW}(q, \beta)$, considering the total 760 scores, or via the method of proportion; and setting home and team effects equal to zero

Serie A 2018-2019



Table: Table of results

	Ata	Bol	Cag	Chi	Emp	Fio	Fro	Gen	Int	Juv	Laz	Mil	Nap	Par	Rom	Sam	Sas	Spa	Tor	Udi
Ata	***	4-1	0-1	1-1	0-0	3-1	4-0	2-1	4-1	2-2	1-0	1-3	1-2	3-0	3-3	0-1	3-1	2-1	0-0	2-0
Bol	1-2	***	2-0	3-0	3-1	0-0	0-4	1-1	0-3	0-1	0-2	0-0	3-2	4-1	2-0	3-0	2-1	0-1	2-2	2-1
Cag	0-1	2-0	***	2-1	2-2	2-1	1-0	1-0	2-1	0-2	1-2	1-1	0-1	2-1	2-2	0-0	2-2	2-1	0-0	1-2
Chi	1-5	2-2	0-3	***	0-0	3-4	1-0	0-0	1-1	2-3	1-1	1-2	1-3	1-1	0-3	0-0	0-2	0-4	0-1	0-2
Emp	3-2	2-1	2-0	2-2	***	1-0	2-1	1-3	0-1	1-2	0-1	1-1	2-1	3-3	0-2	2-4	3-0	2-4	4-1	2-1
Fio	2-0	0-0	1-1	6-1	3-1	***	0-1	0-0	3-3	0-3	1-1	0-1	0-0	0-1	1-1	3-3	0-1	3-0	1-1	1-0
Fro	0-5	0-0	1-1	0-0	3-3	1-1	***	1-2	1-3	0-2	0-1	0-0	0-2	3-2	2-3	0-5	0-2	0-1	1-2	1-3
Gen	3-1	1-0	1-1	2-0	2-1	0-0	0-0	***	0-4	2-0	2-1	0-2	1-2	1-3	1-1	1-1	1-1	1-1	0-1	2-2
Int	0-0	0-1	2-0	2-0	2-1	2-1	3-0	5-0	***	1-1	0-1	1-0	1-0	0-1	1-1	2-1	0-0	2-0	2-2	1-0
Juv	1-1	2-0	3-1	3-0	1-0	2-1	3-0	1-1	1-0	***	2-0	2-1	3-1	3-3	1-0	2-1	2-1	2-0	1-1	4-1
Laz	1-3	3-3	3-1	1-2	1-0	1-0	1-0	4-1	0-3	1-2	***	1-1	1-2	4-1	3-0	2-2	2-2	4-1	1-1	2-0
Mil	2-2	2-1	3-0	3-1	3-0	0-1	2-0	2-1	2-3	0-2	1-0	***	0-0	2-1	2-1	3-2	1-0	2-1	0-0	1-1
Nap	1-2	3-2	2-1	0-0	5-1	1-0	4-0	1-1	4-1	1-2	2-1	3-2	***	3-0	1-1	3-0	2-0	1-0	0-0	4-2
Par	1-3	0-0	2-0	1-1	1-0	1-0	0-0	1-0	0-1	1-2	0-2	1-1	0-4	***	0-2	3-3	2-1	2-3	0-0	2-2
Rom	3-3	2-1	3-0	2-2	2-1	2-2	4-0	3-2	2-2	2-0	3-1	1-1	1-4	2-1	***	4-1	3-1	0-2	3-2	1-0
Sam	1-2	4-1	1-0	2-0	1-2	1-1	0-1	2-0	0-1	2-0	1-2	1-0	3-0	2-0	0-1	***	0-0	2-1	1-4	4-0
Sas	2-6	2-2	3-0	4-0	3-1	3-3	2-2	5-3	1-0	0-3	1-1	1-4	1-1	0-0	0-0	3-5	***	1-1	1-1	0-0
Spa	2-0	1-1	2-2	0-0	2-2	1-4	0-3	1-1	1-2	2-1	1-0	2-3	1-2	1-0	2-1	1-2	0-2	***	0-0	0-0
Tor	2-0	2-3	1-1	3-0	3-0	1-1	3-2	2-1	1-0	0-1	3-1	2-0	1-3	1-2	0-1	2-1	3-2	1-0	***	1-0
Udi	1-3	2-1	2-0	1-0	3-2	1-1	1-1	2-0	0-0	0-2	1-2	0-1	0-3	1-2	1-0	1-0	1-1	3-2	1-1	***

Summary of results

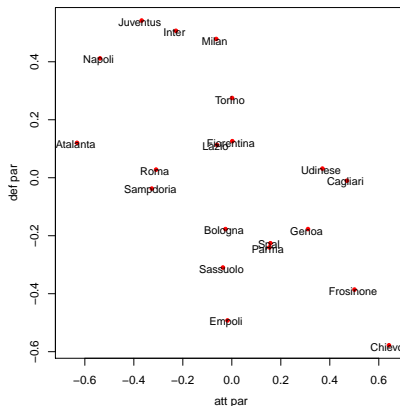
Table: Summary results for several DW regression models. They all include home effect and attack and defense abilities parameters for q

model id	description	n.par	ℓ_{\max}	AIC	$\hat{\theta}$
1	indep, β const	41	-1060.5	2202.9	
2	indep, β team	60	-1043.7	2207.5	
3	indep, β home att def	80	-1036.3	2232.7	
4	dep, β const	42	-1056.2	2196.5	0.2329 **
5	dep, β team	61	-1039.8	2201.6	0.2257 *
6	dep, β home att def	81	-1032.5	2226.9	0.2212 *

NB: AIC for the bivariate Poisson with independent components (40 parameters): 2203.1

Some results: model 1

Figure: q parameter estimates plots; $\hat{\mu}^{(q)} = -1.144$, $\widehat{home}^{(q)} = -0.2687$, $\hat{\beta} = 1.850$. Better attack ability: \Leftarrow ; better defense ability \Uparrow



Some results: model 1

Interpretation of the attack ability estimate for **Juventus**, $\text{att}_{Juve}^{(q)} = -0.3681$:

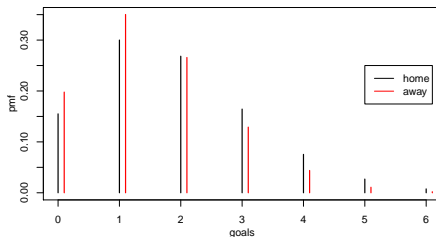
Setting $\text{def}^{(q)} = 0$ for the competing team:

$$q_{\text{home},0} = \exp(-\exp(-1.144 - 0.2687 - 0.3681)) = 0.845$$

$$q_{\text{away},0} = \exp(-\exp(-1.144 - 0.3681)) = 0.802$$

Estimated probability of scoring at home when playing against an average (defensive ability) team: 0.845

Estimated probability of scoring away when playing against an average (defensive ability) team: 0.802



Some results: model 1

Table: Actual and expected number of goals for/against and points

team	real			expected		
	pts	GF	GA	pts	GF	GA
Juventus (C)	90	70	30	74.6	65.9	31.8
Napoli	79	74	36	76.0	73.8	35.2
Atalanta	69	77	46	71.9	77.9	44.1
Inter	69	57	33	70.2	59.6	33.0
Milan	68	55	36	65.2	52.9	34.1
Roma	66	66	48	61.0	62.2	48.1
Torino	63	52	37	58.7	50.1	40.3
Lazio	59	56	46	56.4	52.2	45.6
Sampdoria	53	60	51	59.9	62.8	50.5
Bologna	44	48	56	48.3	50.3	56.6
Sassuolo	43	53	60	45.4	50.4	62.2
Udinese	43	39	53	43.1	37.3	49.3
Spal	42	44	56	42.4	43.7	59.0
Parma	41	41	61	42.2	43.8	59.6
Cagliari	41	36	54	39.7	34.3	51.0
Fiorentina	41	47	45	55.0	49.9	45.3
Genoa	38	39	57	39.7	38.8	57.3
Empoli (R)	38	51	70	40.5	49.3	70.7
Frosinone (R)	25	29	69	30.7	32.8	66.7
Chievo ^{-3pts} (R)	17	25	75	24.1	28.7	76.4

NB: Fiorentina 10W 11D 17L, GD +2

Some results: model 2

Letting β depending on team, we get among all the new estimates

$\hat{\mu}^{(q)} = -1.195$ and $\widehat{\text{home}}^{(q)} = -0.2957$, and the estimates of β for each (scoring) team:

Atalanta	1.760	Lazio	2.189
Bologna	1.622	Milan	2.035
Cagliari	2.051	Napoli	1.996
Chievo	1.459	Parma	1.788
Empoli	1.971	Roma	2.340
Fiorentina	1.391	Sampdoria	1.637
Frosinone	1.155	Sassuolo	1.906
Genoa	1.970	Spal	1.905
Inter	1.754	Torino	2.091
Juventus	3.095	Udinese	1.917

Interpretation and direct comparison of β 's are not straightforward! They must be made for the same value of q !

Some results: model 2

Scored goals' distribution for Frosinone:

goals	0	1	2	3	4
frequency	21	10	3	3	1

$$\bar{y} = 0.763, \sigma_y^2 = 1.128$$

Fitting a type I DW to it, we would get:

$$\begin{cases} \hat{q}_{ML} = 0.4534 \\ \hat{\beta}_{ML} = 1.080 \end{cases}$$

Scored goals' distribution for Juventus:

goals	0	1	2	3	4
frequency	3	10	16	8	1

$$\bar{y} = 1.842, \sigma_y^2 = 0.870$$

Fitting a type I DW to it, we would get:

$$\begin{cases} \hat{q}_{ML} = 0.9423 \\ \hat{\beta}_{ML} = 2.923 \end{cases}$$

Some results: model 2

Table: Actual and expected number of goals for/against and points

team	real			expected		
	pts	GF	GA	pts	GF	GA
Juventus	90	70	30	84.7	70.1	31.6
Napoli	79	74	36	77.3	74.6	35.4
Atalanta	69	77	46	69.5	77.5	45.4
Inter	69	57	33	68.6	59.4	33.2
Milan	68	55	36	66.5	53.9	34.5
Roma	66	66	48	64.6	64.4	48.3
Torino	63	52	37	60.2	51.3	40.7
Lazio	59	56	46	59.0	54.1	45.7
Sampdoria	53	60	51	56.4	61.6	51.9
Bologna	44	48	56	48.0	49.2	53.5
Sassuolo	43	53	60	45.1	50.6	62.6
Udinese	43	39	53	42.0	37.7	51.5
Spal	42	44	56	43.3	44.1	57.6
Parma	41	41	61	41.5	43.5	59.7
Cagliari	41	36	54	39.6	35.2	52.5
Fiorentina	41	47	45	49.5	46.9	45.2
Genoa	38	39	57	39.7	39.2	57.9
Empoli	38	51	70	41.0	49.6	69.9
Frosinone	25	29	69	26.9	29.7	67.8
Chievo	17(20)	25	75	23.9	28.0	75.6

The model guesses the first positions correctly!

Some results: model 4

Table: Actual and expected points and ranks

	real pts	real rank	exp pts	exp rank
Juventus	90	1	76.1	2
Napoli	79	2	76.8	1
Atalanta	69	3	72.0	3
Inter	69	4	70.4	4
Milan	68	5	65.2	5
Roma	66	6	61.4	6
Torino	63	7	59.1	8
Lazio	59	8	56.5	9
Sampdoria	53	9	59.6	7
Bologna	44	10	47.3	11
Sassuolo	43	11	45.1	12
Udinese	43	12	41.9	13
Spal	42	13	41.7	14
Parma	41	14	41.1	15
Cagliari	41	15	38.2	18
Fiorentina	41	16	53.6	10
Genoa	38	17	38.9	17
Empoli	38	18	39.1	16
Frosinone	25	19	28.4	19
Chievo	17(20)	20	22.3	20

Some results: model 5

Table: Actual and expected points and ranks

	real pts	real rank	exp pts	exp rank
Juventus	90	1	85.6	1
Napoli	79	2	77.9	2
Atalanta	69	3	69.6	3
Inter	69	4	69.1	4
Milan	68	5	66.4	5
Roma	66	6	64.8	6
Torino	63	7	61.3	7
Lazio	59	8	58.5	8
Sampdoria	53	9	56.2	9
Bologna	44	10	47.0	11
Sassuolo	43	11	45.2	12
Udinese	43	12	40.9	14
Spal	42	13	42.5	13
Parma	41	14	40.5	15
Cagliari	41	15	37.8	18
Fiorentina	41	16	49.0	10
Genoa	38	17	38.5	17
Empoli	38	18	39.6	16
Frosinone	25	19	25.7	19
Chievo	17(20)	20	22.5	20

Some results: model 5

Table: Number of actual and expected draws

	0-0	1-1	2-2	3-3	other	total
actual	34	44	20	10	0	108
expected	28.9	49.0	20.4	3.1	0.4	101.8

The number of 0-0 (and 3-3!) is still underestimated...

Some results: model 5

Table: Expected probabilities for the outcomes of the match Juventus-Napoli:

Juv-Nap	0	1	2	3	≥ 4
0	0.0307	0.0257	0.0109	0.0028	0.0005
1	0.1036	0.1623	0.0864	0.0238	0.0039
2	0.0855	0.1842	0.1129	0.0326	0.0054
3	0.0220	0.0532	0.0345	0.0101	0.0017
≥ 4	0.0013	0.0032	0.0021	0.0006	0.0001

$\rightarrow \text{cor}(Y_1, Y_2) = 0.127.$




Under the hypothesis of independence, the probability of 0-0 would be $0.0171 < 0.0307$.

Some results: model 5

Expected probabilities: 1: 49% X: 31.6% 2: 19.4%

Corresponding odds: 1: 2.04 X: 3.16 2: 5.15

Figure: Odds from some betting companies

	1	X	2
 Sisal	<u>2.10</u>	<u>3.30</u>	<u>3.90</u>
 Unibet	<u>2.10</u>	<u>3.20</u>	<u>4.00</u>
 Sisal Matchpoint	<u>2.10</u>	<u>3.30</u>	<u>3.65</u>
 NetBet	<u>2.05</u>	<u>3.20</u>	<u>3.75</u>
 bwin	<u>2.05</u>	<u>3.30</u>	<u>3.80</u>

Future research

- ⚽ using other discrete counterparts of the continuous Weibull rv:
 - ⌚ type II DW, which has a finite or infinite support according to the value of the second parameter β of the continuous distribution;
 - ⌚ type III DW, which has a complicated expression for pmf
- ⚽ using other link functions
- ⚽ using other dependence structures
- ⚽ using other sets of regressors (but these and previous finding tend to discourage the use of too many regressors)
- ⚽ applying the DW models to other championships

Future research

- ⚽ using other discrete counterparts of the continuous Weibull rv:
 - ⌚ type II DW, which has a finite or infinite support according to the value of the second parameter β of the continuous distribution;
 - ⌚ type III DW, which has a complicated expression for pmf
- ⚽ using other link functions
- ⚽ using other dependence structures
- ⚽ using other sets of regressors (but these and previous finding tend to discourage the use of too many regressors)
- ⚽ applying the DW models to other championships

Thanks for your time and attention!

Main references

- ⊛ Alessandro Barbiero. Discrete Weibull regression for modeling football outcomes. *International Journal of Business Intelligence and Data Mining*, 2018.
- ⊛ Georgi Boshnakov, Tarak Kharrat, and Ian G. McHale. A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458-466, 2017.
- ⊛ Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381-393, 2003.
- ⊛ Dimitris Karlis and Ioannis Ntzoufras. Robust fitting of football prediction models. *IMA Journal of Management Mathematics*, 22(2):171-182, 2011.
- ⊛ Alan J Lee. Modeling scores in the premier league: is Manchester United really the best? *Chance*, 10(1):15-19, 1997.
- ⊛ Michael J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109-118, 1982.
- ⊛ Ian McHale and Phil Scarf. Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, 61(4):432-445, 2007.
- ⊛ Toshio Nakagawa and Shunji Osaki. The discrete Weibull distribution. *IEEE Transactions on Reliability*, 24(5):300-301, 1975.