

COM-Poisson 分布について

—過小分散も過大分散も表現するポアソン分布—

長塚 豪己[†]

(受付 2024 年 11 月 28 日; 改訂 2025 年 2 月 7 日; 採択 2 月 12 日)

要 旨

Poisson 分布は、カウントデータの統計モデルとして、最もよく知られた分布である。しかし、Poisson 分布には、平均と分散が等しいという意味での等分散性 (equi-dispersion) が課されている。応用の場では、多くのカウントデータに、分散が平均より大きい過分散性 (over-dispersion)、時には、分散が平均より小さい過小分散性 (under-dispersion) が見られる。それぞれの場合において、負の二項分布や幾何分布等がモデルとして適宜検討される。Conway と Maxwell によって生み出された Conway-Maxwell-Poisson (COM-Poisson) 分布は、Poisson 分布の一般化分布の一つである。生存時間解析と信頼性解析の分野においては、治癒率モデルに関する研究において、近年、COM-Poisson 分布を用いた報告が多く見られ、注目を集めている。COM-Poisson 分布は、パラメータ数が少なく (2 パラメータ) シンプルなモデルでありながら、広い範囲の過分散、過小分散を柔軟に表現でき、かつよく知られた幾何分布を含み、極限分布としてベルヌーイ分布を持つ点が特徴的である。これらの特徴が、他のポアソン分布の一般化モデルと一線を画す。しかし、それらメリットの代償として、確率関数の規格化定数に無限級数を含み、推定、検定においては困難をきたす。これまで、規格化定数の各種近似法や、規格化定数の計算を回避する推定が研究されてきた。本稿では、COM-Poisson 分布、並びに近年提案された手法も含む代表的な COM-Poisson 分布の推定法について紹介する。

キーワード：カウントデータ、Poisson 分布、最尤法、規格化定数。

1. はじめに

既定の期間、あるいは空間内におけるイベントの発生数を表すデータであるカウントデータは、古くから広い分野で利用されてきた。古典的な例としては、伝染病や負傷による兵士の死亡数 (Nightingale, 1858)、コレラによる死者数 (Snow, 1858) が挙げられる。近年でもその有用性は衰えることなく、その利用例は、実店舗やオンラインストアでの来客数や購入数、SNS の利用数や閲覧数、TV やオンライン動画サービスにおける視聴数、等々と枚挙にいとまがない。

カウントデータのモデリングに最も広く用いられてきた分布はポアソン分布であるといえる。ポアソン分布、あるいは関連分布によるカウントデータのモデリングの応用範囲は広く、古典的な例は、プロイセン騎兵隊における 20 年間にわたる馬蹴りによる年間死亡者数にポアソン分布をあてはめた例 (von Bortkewitsch, 1898; Hanley and Bhatnagar, 2022) であり、近年でも、円形の競馬場におけるスタート位置毎の勝った馬の数や斑点病にかかった植物の葉

[†] 中央大学 理工学部: 〒112-8551 東京都文京区春日 1-13-27; hideki07@kc.chuo-u.ac.jp

のグレード毎の数にポアソン分布をあてはめて解析を行った例等例は無数にある (Wani and Ahmad, 2024). しかし, ポアソン分布はシンプルで極めてよく知られた分布であるが, 多くの実データはポアソン分布の基礎となる仮定, すなわち平均と分散が等しいという意味での等分散性を満たさない.

Conway と Maxwell によって生み出された Conway-Maxwell-Poisson (COM-Poisson) 分布 (Conway and Maxwell, 1962) は, Poisson 分布の一般化分布の一つである. 生存時間解析と信頼性解析の分野においては, 治癒率モデルに関する研究において, 近年, COM-Poisson 分布を用いた報告が多く見られ, 注目を集めている. COM-Poisson 分布は, パラメータ数が少なく (2 パラメータ) シンプルなモデルでありながら, 広い範囲の過分散, 過小分散を柔軟に表現でき, かつよく知られた幾何分布を含み, 極限分布としてベルヌーイ分布を持つ点が特徴的である. これらの特徴が, 他のポアソン分布の一般化モデルと一線を画す. しかし, それらメリットの代償として, 確率関数の規格化定数に無限級数を含み, 推定, 検定においては困難をきたす. これまで, 規格化定数の各種近似法や, 規格化定数の計算を回避する推定が研究されてきた. 本稿では, COM-Poisson 分布, 並びに近年提案された手法も含む代表的な COM-Poisson 分布の推定法について紹介する.

本稿は, 以下のように構成されている. 第 2 節では, COM-Poisson 分布について概説する. 第 3 節では, 代表的な推定法や近年提案された新しい推定法を紹介する. 第 4 節では, COM-Poisson 分布の応用例を紹介する. 第 5 節で, 本稿を締めくくる.

2. COM-Poisson 分布

2.1 分布の概要

(Ω, \mathcal{F}, P) を, 本論文で扱う確率変数が定義された確率空間とする. さらに, \mathbb{N} は自然数全体を表す集合, $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$ と定義する. 確率変数 X が, 次の確率質量関数を持つとき, X は, パラメータ λ, ν の COM-Poisson 分布 (CMP(λ, ν)) で表す) にしたがうという:

$$f_{\text{CMP}}(x) := P(X = x) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)}, \quad x \in \mathbb{N}_0,$$

ただし, $\lambda > 0, \nu \geq 0$ はパラメータで,

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu},$$

である.

平均パラメータが λ であるポアソン分布 $\text{Poi}(\lambda)$ の確率質量関数

$$f_{\text{P}}(x) = \frac{\lambda^x}{x!} e^{-\lambda} \propto \frac{\lambda^x}{x!},$$

において, $x!$ に, 過分散, 過小分散を表現する形状パラメータ ν をべき乗し, あらためて標準化定数を $Z(\lambda, \nu)$ にとったのが, f_{CMP} と考えられる. よって, $\text{CMP}(\lambda, \nu)$ は, $\text{Poi}(\lambda)$ の一般化されたモデルであり, $\nu = 1$ のとき, $\text{Poi}(\lambda)$ となる. $\nu < 1$ のとき, 過分散モデル (分散が平均より大きいモデル, over-dispersion model), $\nu > 1$ のとき, 過小分散モデル (分散が平均より小さいモデル, under-dispersion model), $\nu = 1$ のとき, 平均と分散が等しいモデル (equi-dispersion model) になる. 図 1 に, 各パラメータ値における $\text{CMP}(\lambda, \nu)$ の確率質量関数のグラフを示す. $\text{CMP}(\lambda, \nu)$ は, 前述のように $\nu = 1$ のときポアソン分布, $\nu = 0, \lambda < 1$ のとき幾何分布, $\nu \leftarrow \infty$ のときベルヌーイ分布となる (詳細については, 表 1 を参照のこと). 表 2 に, $\text{CMP}(\lambda, \nu)$ の確率質量関数, 分布関数, k 次積率 ($k \in \mathbb{N}$), 積率母関数, 特性関数, 確率母

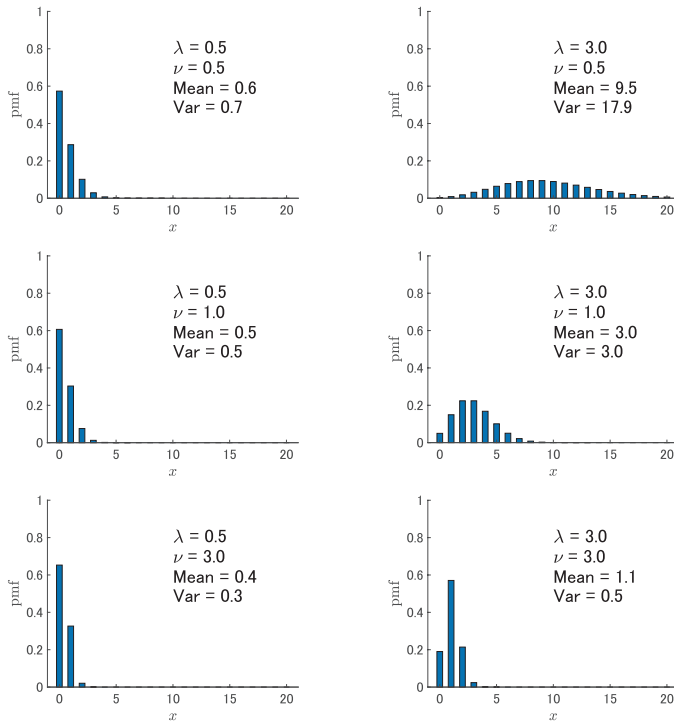
図 1. CMP(λ, ν) の確率質量関数 (pmf) のグラフ (Mean は平均, Var は分散を表す).

表 1. COM-Poisson 分布の特殊例.

λ, ν	$Z(\lambda, \nu)$	確率質量関数	分布名 (パラメータ)
$\nu = 1$	e^λ	$\frac{\lambda^x e^{-\lambda}}{x!}, x \in \mathbb{N}_0$	ポアソン (λ)
$\nu = 0, \lambda < 1$	$1/(1 - \lambda)$	$(1 - \lambda) \lambda^x, x \in \mathbb{N}_0$	幾何 ($p = 1 - \lambda$)
$\nu \rightarrow \infty$	$1 + \lambda$	$\begin{cases} \frac{1}{1+\lambda}, & x = 0 \\ \frac{\lambda}{1+\lambda}, & x = 1 \\ 0, & x \in \mathbb{N} \setminus \{1\} \end{cases}$	ベルヌーイ ($p = \frac{\lambda}{1+\lambda}$)

表 2. CMP(λ, ν) の確率質量関数 (pmf), 分布関数 (CDF), k 次積率 (k 's moment), 積率母関数 (MGF), 特性関数 (CF), 確率母関数 (PGF).

pmf	CDF	k 's moment	MGF	CF	PGF
$\frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)}$	$\sum_{k=0}^x \frac{\lambda^k}{(k!)^\nu} \frac{1}{Z(\lambda, \nu)}$	$\sum_{j=0}^{\infty} \frac{j^k \lambda^j}{(j!)^\nu Z(\lambda, \nu)}$	$\frac{Z(e^t \lambda, \nu)}{Z(\lambda, \nu)}$	$\frac{Z(e^{it} \lambda, \nu)}{Z(\lambda, \nu)}$	$\frac{Z(t \lambda, \nu)}{Z(\lambda, \nu)}$

関数を示す. なお, COM-Poisson 分布の由来と導出については, Sellers (2023, pp.22–24) に詳しい.

2.2 標準化定数の計算

COM-Poisson 分布は、平均と分散が等しいという意味での等分散性の制約に縛られず、過分散性も過小分散性も表現できる柔軟なモデルである一方、規格化定数は閉形式でないことから、計算困難の問題を持つ。この規格化定数の計算困難を解決するために、いくつかのアプローチが提案されてきた。その一つは、以下の式で与えられる有限和近似である (Shmueli et al., 2005; Sellers et al., 2019) :

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \approx \sum_{j=0}^M \frac{\lambda^j}{(j!)^\nu}, \quad M \in \mathbb{N},$$

ただし、 M は、十分大きくとった適当な自然数である。

さらに、以下の漸近展開式を利用した近似も提案されている (Minka et al., 2003; Shmueli et al., 2005; Sellers et al., 2019) :

$$Z(\lambda, \nu) = \frac{\exp(\nu \lambda^{1/\nu})}{\lambda^{(\nu-1)/(2\nu)} (2\pi)^{(\nu-1)/2} \sqrt{\nu}} \left\{ 1 + O\left(\frac{1}{\lambda^{1/\nu}}\right) \right\}, \quad \nu \leq 1, \quad \text{あるいは, } \lambda > 10^\nu.$$

他の近似法も含め、 $Z(\lambda, \nu)$ の近似法については、Sellers (2023, pp.66–69) に詳しい。

3. 推定法の紹介

3.1 一般化最小二乗法

$X \sim \text{CMP}(\nu, \lambda)$ のとき、

$$\frac{P(X = x-1)}{P(X = x)} = \frac{x^\nu}{\lambda}, \quad x \in \mathbb{N},$$

であるから、両辺対数を取って、

$$(3.1) \quad \ln \frac{P(X = x-1)}{P(X = x)} = -\ln \lambda + \nu \ln x,$$

が得られる。(3.1) の右辺は、 $-\ln \lambda$ を切片、 ν を傾き、 $\ln x$ を説明変数としたときの単回帰式と見ることができる。左辺の $\ln \frac{P(X=x-1)}{P(X=x)}$ において、 $P(X=x-1)$, $P(X=x)$ を、それぞれに対応する相対頻度を用いてノンパラメトリックに推定することにより、正規確率紙や、ワイブル確率紙のような COM-Poisson 確率紙 (図 2) と、一般化最小二乗法によるパラメータ推定法が提案されている (Shmueli et al., 2005)。

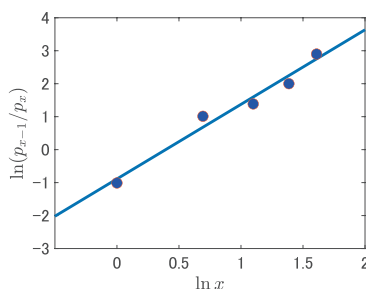


図 2. COM-Poisson 確率紙の例 (図中の点は、 $X \sim \text{CMP}(\lambda, \nu)$ の観測値 x について、 $(\ln x, \ln \frac{\hat{p}_{x-1}}{\hat{p}_x})$ を座標とする点 (ただし、 $x=0$ は除く)、直線は、それらの点に対し、一般化最小二乗法を用いてあてはめた直線である)。

$X_i \stackrel{\text{i.i.d.}}{\sim} \text{CMP}(\nu, \lambda)$, $i \in \{1, \dots, n\}$, とする. このとき, $p_x := P(X = x)$, $x \in \mathbb{N}_0$, 及びその相対頻度を, \hat{p}_x とおく. つまり, $\hat{p}_x = \frac{\#\{i: X_i = x\}}{n}$, $x \in \mathbb{N}_0$, である (なお, 実データに適用する場合には, \hat{p}_x が 0 とならないよう補正が必要である. 著者が知る限り, 補正の方法について言及した論文はほとんどない).

すると, デルタ法により, $\ln \frac{\hat{p}_{x-1}}{\hat{p}_x}$ に関する漸近分散, 漸近共分散が以下のように得られる:

$$(3.2) \quad \text{var} \left(\ln \frac{\hat{p}_{x-1}}{\hat{p}_x} \right) \approx \frac{1}{n p_x} + \frac{1}{n p_{x-1}}, \quad \text{cov} \left(\ln \frac{\hat{p}_{x-1}}{\hat{p}_x}, \ln \frac{\hat{p}_x}{\hat{p}_{x+1}} \right) \approx -\frac{1}{n p_x}.$$

(3.2)において, p_x に, \hat{p}_x を代入し, さらに, (i, j) 成分を $\text{cov} \left(\ln \frac{\hat{p}_{i-1}}{\hat{p}_i}, \ln \frac{\hat{p}_j}{\hat{p}_{j+1}} \right)$ とする分散共分散行列を \widehat{M} とおく. さらに, $\mathbf{y} = \left(\ln \frac{\hat{p}_1}{\hat{p}_2}, \dots, \ln \frac{\hat{p}_{n-1}}{\hat{p}_n} \right)'$, $\mathbf{x}'_i = (1, \ln i)$, $\mathbf{X} = (\mathbf{x}'_2, \dots, \mathbf{x}'_n)'$ とおくと, $\boldsymbol{\xi} = (-\ln \lambda, \nu)$ の“分散共分散行列をその推定値で置き換えたときの一般化最小二乗推定量” (FGLS; feasible generalized least squares estimators) $\hat{\boldsymbol{\xi}}$ が, 以下で得られる:

$$\hat{\boldsymbol{\xi}}' = \arg \inf_{\boldsymbol{\theta} \in \Theta} (\mathbf{y} - \mathbf{X} \boldsymbol{\xi})' \widehat{M}^{-1} (\mathbf{y} - \mathbf{X} \boldsymbol{\xi}) = (\mathbf{X}' \widehat{M}^{-1} \mathbf{X})^{-1} \mathbf{X}' \widehat{M}^{-1} \mathbf{y},$$

ただし, Θ はパラメータ空間であり, $\Theta = \{(\lambda, \nu) : \lambda > 0, \nu > 0, \text{あるいは } 0 < \lambda < 1 \text{ かつ } \nu = 0\}$ である. 以上より, $\boldsymbol{\theta}$ の FGLS 推定量 $\hat{\boldsymbol{\theta}}_{FGLS}$ が以下で得られる:

$$\hat{\boldsymbol{\theta}}_{FGLS} = \begin{pmatrix} e^{-\hat{\xi}_{(1)}} & \hat{\xi}_{(2)} \end{pmatrix},$$

ただし, $\hat{\xi}_{(i)}$ は, $\hat{\boldsymbol{\xi}}$ の第 i 成分である. 有限の n において $\ln \frac{\hat{p}_{x-1}}{\hat{p}_x}$ は正規分布に従わないので, $\hat{\boldsymbol{\theta}}$ は, 最良線形不偏推定量 (BLUE) といった最適性が保証された推定量ではない. 一見有用そうに見える手法であるが, 本手法の実データへの適用例は, 筆者の知る限り極めて少なく Shmueli et al. (2005) で示されているのみである. Shmueli et al. (2005) での適用例は, 頻度が 0 である数値を取り除いた作為データによる適用例しか示されていない. 実際, 相対頻度が 0 となる点を含むデータでは, 本手法はそのままでは適用できないので工夫が必要である.

3.2 最尤推定法

$X_i \stackrel{\text{i.i.d.}}{\sim} \text{CMP}(\nu, \lambda)$, $i \in \{1, \dots, n\}$, であり, x_i , $i \in \{1, \dots, n\}$, をこれらの実現値とする. このとき, $\mathbf{x} = (x_1, \dots, x_n)$ に基づく, $\boldsymbol{\theta} = (\lambda, \nu)$ の対数尤度関数 $l_{\mathbf{x}}(\boldsymbol{\theta})$ は, 以下で与えられる:

$$(3.3) \quad l_{\mathbf{x}}(\boldsymbol{\theta}) = \sum_{i=1}^n x_i \ln \lambda - \nu \sum_{i=1}^n \ln(x_i!) - n \ln Z(\lambda, \nu).$$

因子分解定理により,

$$\left(\sum_{i=1}^n x_i, \prod_{i=1}^n x_i! \right),$$

が, $\boldsymbol{\theta}$ の十分統計量であることが, 直ちに導かれる.

$l_{\mathbf{x}}(\boldsymbol{\theta})$ の λ, ν それぞれについての 1 階偏導関数を 0 とおくことで, 以下の尤度方程式が得られる:

$$(3.4) \quad \begin{cases} \frac{\partial l_{\mathbf{x}}(\boldsymbol{\theta})}{\partial \lambda} = \frac{\sum_{i=1}^n x_i}{\lambda} - n \frac{\frac{\partial Z(\lambda, \nu)}{\partial \lambda}}{Z(\lambda, \nu)} = 0, \\ \frac{\partial l_{\mathbf{x}}(\boldsymbol{\theta})}{\partial \nu} = -\sum_{i=1}^n \ln(x_i!) - n \frac{\frac{\partial Z(\lambda, \nu)}{\partial \nu}}{Z(\lambda, \nu)} = 0, \end{cases}$$

ただし,

$$(3.5) \quad \frac{\partial Z(\lambda, \nu)}{\partial \lambda} = \sum_{j=0}^{\infty} \frac{j \lambda^{j-1}}{(j!)^\nu} = \frac{Z(\lambda, \nu)}{\lambda} \sum_{j=0}^{\infty} \frac{j \lambda^j}{(j!)^\nu Z(\lambda, \nu)} = \frac{Z(\lambda, \nu) E(X)}{\lambda},$$

$$(3.6) \quad \frac{\partial Z(\lambda, \nu)}{\partial \nu} = - \sum_{j=0}^{\infty} \frac{\lambda^j \ln(j!)}{(j!)^\nu} = -Z(\lambda, \nu) \sum_{j=0}^{\infty} \frac{\ln(j!) \lambda^j}{(j!)^\nu Z(\lambda, \nu)} = -Z(\lambda, \nu) E(\ln(X!)).$$

さらに, (3.5) と (3.6) を, (3.4) に代入して, $E(X)$ と $E(\ln X!)$ について解くと, 以下の式が得られる:

$$(3.7) \quad \begin{cases} E(X) = \lambda \bar{x}, \\ E(\ln X!) = \overline{\ln x!}, \end{cases}$$

ただし, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\overline{\ln x!} = \frac{1}{n} \sum_{i=1}^n \ln(x_i!)$ である. 尤度方程式 (3.7) は, 陽には解けないので, 数値的解法を用いて解く必要がある. 尤度方程式の解の存在条件は, Bedbur et al. (2023) において示されている.

3.3 Tomio and Nagatsuka (2022) の方法

$X_i \stackrel{\text{i.i.d.}}{\sim} \text{CMP}(\nu, \lambda)$, $i = 1, \dots, n$, であり, x_i , $i \in \mathbb{N}$, をこれらの実現値であるとする. このとき, $S_1 = s_1$ (ただし, $S_1 = \sum_{i=1}^n X_i$, $s_1 = \sum_{i=1}^n x_i$) が与えられた元での, $\mathbf{x} = (x_1, \dots, x_n)$ に基づく ν の条件付き対数尤度関数は, 以下で与えられる.

$$l_{\mathbf{x}|s_1}(\nu) = \nu \ln \left(\frac{s_1}{x_1, \dots, x_n} \right) - \ln C_{s_1}(\nu),$$

ただし,

$$C_{s_1}(\nu) = \sum_{z_1=0}^{s_1} \sum_{z_2=0}^{s_1-z_1} \cdots \sum_{z_{n-1}=0}^{s_1-\sum_{i=1}^{n-2} z_i} \left(\frac{s_1}{z_1, \dots, z_n} \right)^\nu,$$

である.

また, $S_2 = s_2$ (ただし, $S_2 = \prod_{i=1}^n (X_i!)$, $s_2 = \prod_{i=1}^n (x_i!)$) が与えられた元での, $\mathbf{x} = (x_1, \dots, x_n)$ に基づく ν の条件付き対数尤度関数は, 以下で与えられる.

$$l_{\mathbf{x}|s_2}(\lambda) = \left(\sum_{i=1}^n x_i \right) \ln \lambda - \ln \sum_{\mathbf{z} \in \Omega_n(s_2)} \lambda^{\sum_{i=1}^n z_i},$$

ただし, $\Omega_n(s) = \{(z_1, \dots, z_n) \in \mathbb{N}_0^n : \prod_{i=1}^n z_i! = s\}$, $s \in \mathbb{N}_0$, である.

$l_{\mathbf{x}|s_1}(\nu)$ の ν についての 1 階偏導関数 $\frac{\partial}{\partial \nu} l_{\mathbf{x}|s_1}(\nu)$ と, $l_{\mathbf{x}|s_2}(\lambda)$ の λ についての 1 階偏導関数 $\frac{\partial}{\partial \lambda} l_{\mathbf{x}|s_2}(\lambda)$ は,

(3.8)

$$\frac{\partial}{\partial \nu} l_{\mathbf{x}|s_1}(\nu) = \ln \left(\frac{s_1}{x_1, \dots, x_n} \right) - \sum_{z_1=0}^{s_1} \cdots \sum_{z_{n-1}=0}^{s_1-\sum_{i=1}^{n-2} z_i} \left(\frac{s_1}{z_1, \dots, z_n} \right)^\nu \ln \left(\frac{s_1}{z_1, \dots, z_n} \right) \bigg/ C_{s_1}(\nu),$$

(3.9)

$$\frac{\partial}{\partial \lambda} l_{\mathbf{x}|s_2}(\lambda) = - \frac{\sum_{\mathbf{z} \in \Omega_n(s_2)} (\sum_{i=1}^n z_i - \sum_{i=1}^n x_i) \lambda^{\sum_{i=1}^n z_i - 1}}{\sum_{\mathbf{z} \in \Omega_n(s_2)} \lambda^{\sum_{i=1}^n z_i}},$$

と求められ、これらをそれぞれ 0 とおいて、尤度方程式が得られる．尤度方程式の解の存在について、以下の定理が得られている (Tomio and Nagatsuka, 2022)．

定理 1 (Tomio and Nagatsuka, 2022). 任意の $\nu \in \mathbb{R}$ と, $\mathbf{x} = (x_1, \dots, x_n)$, $x_i \in \mathbb{N}_0$, $n \in \mathbb{N} \setminus \{1\}$, について, $s_1 = \sum_{i=1}^n x_i$ が与えられた元での ν の条件付き尤度方程式 $\frac{\partial}{\partial \nu} l_{\mathbf{x}|s_1}(\nu) = 0$ の解は, \mathbf{x} の要素の内, 少なくとも $n-1$ 個の要素が 0 でないか, $\mathbf{x} = \arg \max_{\mathbf{z} \in D(s_1)} \binom{s_1}{z_1, \dots, z_n}$ でない限り存在し, 一意である. ただし, $D(s) = \{(z_1, \dots, z_n) : \sum_{i=1}^n z_i = s, z_i \in \mathbb{N}_0\}$ である. \mathbf{x} の要素の内, 少なくとも $n-1$ 個の要素が 0 の場合には, 解は $-\infty$ に発散し, $\mathbf{x} = \arg \max_{\mathbf{z} \in D(s_1)} \binom{s_1}{z_1, \dots, z_n}$ の場合には, 解は ∞ に発散する.

定理 2 (Tomio and Nagatsuka, 2022). 任意の $\lambda \in \mathbb{R}$ と, $\mathbf{x} = (x_1, \dots, x_n)$, $x_i \in \mathbb{N}_0$, $n \in \mathbb{N} \setminus \{1\}$, について, $s_2 = \prod_{i=1}^n (x_i!)$ が与えられた元での λ の条件付き尤度方程式 $\frac{\partial}{\partial \lambda} l_{\mathbf{x}|s_2}(\lambda) = 0$ の解は, $\mathbf{x} = \arg \min_{\mathbf{z} \in \Omega_n(s_2)} \sum_{i=1}^n z_i$ でないか, $\mathbf{x} = \arg \max_{\mathbf{z} \in \Omega_n(s_2)} \sum_{i=1}^n z_i$ でない限り存在し, 一意である. $\mathbf{x} = \arg \min_{\mathbf{z} \in \Omega_n(s_2)} \sum_{i=1}^n z_i$ の場合には, 解は 0 に収束し, $\mathbf{x} = \arg \max_{\mathbf{z} \in \Omega_n(s_2)} \sum_{i=1}^n z_i$ の場合には, 解は ∞ に発散する.

ただし、尤度方程式 (3.8), (3.9) に含まれる和の計算は、NP 困難問題であり、 n がある程度大きくなると計算困難に陥る．この推定法が適用できるのは、多くとも n が 12 程度のデータであることが (Tomio and Nagatsuka, 2022) によって指摘されている．なお、同様の問題は、第 4.2 で紹介する sCOM-Poisson 分布の確率質量関数の計算においても起こる．

4. COM-Poisson 分布の応用

本節では、COM-Poisson 分布の応用例として、Cure rate モデルへの応用と、確率過程モデルへの応用について紹介する．

4.1 COM-Poisson 分布の治癒率 (Cure rate) モデルへの応用

生存時間分析や信頼性データ解析の分野において、治癒率 (cure rate) モデル (長期間生存モデル, long-term survival model と呼ばれる) は、広く研究されてきたが、COM-Poisson 分布に基づく治癒率モデルは、Rodrigues et al. (2009) において提案された．そして、これまで多くの研究が行われてきた (Balakrishnan and Pal, 2013; Wiangnak and Pal, 2018; Majakwara and Pal, 2019; He and Emura, 2019; Balakrishnan et al., 2024)．

ここでは、まず、治癒率モデルについて概説する．その後、COM-Poisson 分布に基づく治癒率モデルについて紹介する．

4.1.1 治癒率モデルについて

生存時間分析や信頼性工学では、対象についてあるイベント (死亡, 疾病の発症, 故障, 等) が生じるまでの時間に興味があることが多い．しかし、このようなイベントが、一つの要因により生じるとは限らず、多くの場合、複数の要因に起因して生じる．例えば、ある患者を観察中、心疾患による死亡までの時間に興味があったにも拘らず、脳疾患により死亡する場合がある．このような時、興味があった心疾患による死亡までの時間のデータは、完全には得られない (このとき得られるのは、打ち切りデータ)．このように、興味がある要因によって生じるイベントの観察において、その他の要因に起因するイベントによって観察が打ち切りになってしまうリスクを、競合リスクと呼ぶ．

ある要因に起因するイベントが一度起きた後、二度と起こらないことがありうる．例えば、ある疾病の治療後、その疾病の再発が起こりえない場合や、あるバグによるソフトウェアの

エラーを修復後、そのバグによるエラーの再発が起こりえない場合である。これらを「治癒(cure)」と表現し、治癒する割合を「治癒率(cure rate)」と呼ぶ。治癒率を取り入れた生存関数で表現される治癒率モデル(cure rate model)は、がん臨床試験のデータを解析する際に、これまでよく用いられてきた。これは、乳がん、非ホジキンリンパ腫、白血病、前立腺がん、メラノーマ、頭頸部がん等の疾病においては、患者のかなりの割合が治癒する(再発がない)とされているためである(Ibrahim et al., 2001)。「治癒」とは、生物医学分野においては、その名の通り治癒した、あるいは病気の再発がないことを意味し、工学分野においては、故障やエラーの再発がないことを意味する。

これまで研究されてきた治癒率モデルは、混合治癒率モデル(mixture cure rate model)と促進時間治癒率モデル(promotion time cure rate model)に大別される。

混合治癒率モデル(mixture cure rate model)は、治癒した対象と、イベントが起こりうる対象に関する混合モデルである(Boag, 1949; Berkson and Gage, 1952)。混合治癒率モデルでは、時点 y における累積分布関数と生存関数は、それぞれ以下のように表される(Boag, 1949; Berkson and Gage, 1952; Ibrahim et al., 2001)。

$$\begin{aligned} F_p(y) &= (1 - p_0)F_s(y), \\ S_p(y) &= 1 - (1 - p_0)F_s(y) = p_0 + (1 - p_0)S_s(y), \end{aligned}$$

ただし、 $F_p(\cdot)$ と $S_p(\cdot)$ はそれぞれ、対象全体の分布関数と生存関数を示し、 $F_s(\cdot)$ と $S_s(\cdot)$ は、それぞれ、イベントが起こりうる対象のみの分布関数と生存関数を示す。また、 p_0 は、全体に対する治癒する対象の割合で、治癒率を表す。

促進時間治癒率モデル(promotion time cure rate model)は、イベント発生に関連する M 個の競合リスク(競合する要因)を仮定し、さらに M はポアソン分布に従う確率変数とするモデルである(Chen et al., 1999)。 W_j を j 番目の競合リスクによるイベント発生までの時間(つまり「生存時間」とし、 $j = 1, \dots, M$, とする。 $M = m$ が与えられたとき、 W_j は独立かつ同一の分布にしたがい、分布関数を $F(\cdot)$ 、生存関数を $S(\cdot)$ とする。また、 $W_j, j = 1, \dots, m$, は M と独立であるとする。

競合リスクの数は、観測できない場合が多いので M は潜在変数であると考えるのが妥当である。また、対象の生存時間は、競合リスクにおける潜在故障時間モデルの仮定(Cox and Oakes, 1984)の下で、 $Y = \min(W_0, W_1, W_2, \dots, W_M)$ (W_0 については後で説明)で与えられるため、一つの競合リスク j における生存時間 W_j の観測値は得られるが、それ以外の観測値は得られない。よって、 $W_j, j = 1, \dots, m$, も潜在変数と考えるのが妥当である。なお、 W_0 は、治癒したイベントの生存時間を表し、これは、再発しないイベントの生存時間であるから、 $P(W_0 = \infty) = 1$ である。 $p_0 := P(M = 0)$ は、全体に対する治癒する対象の割合で、治癒率を表す。

M が平均 λ のポアソン分布にしたがう時、促進時間治癒率モデルにおける生存関数は次のように表される：

$$S_p(y) = \exp(-\lambda F(y)).$$

また、イベントが起こりうる対象のみの生存関数 $S_s(\cdot)$ は、以下で与えられる：

$$S_s(y) = \frac{\exp(-\lambda F(y)) - \exp(-\lambda)}{1 - \exp(-\lambda)}.$$

なお、 $S_p(y)$ と $S_s(y)$ の導出については、Chen et al. (1999)を参照のこと。

このとき、 $S_p(y)$ と $S_s(y)$ の間には、以下の関係が成り立つ：

$$S_p(y) = \exp(-\lambda) + \{1 - \exp(-\lambda)\} S_s(y).$$

これは、混合治癒モデルの生存時間関数でもある。以上より、 M がポアソン分布にしたがうと仮定すると、促進時間治癒率モデルは、混合治癒率モデルにもなり、治癒率モデルを統一的に表現できる利点が生まれる。

4.1.2 COM-Poisson 分布に基づく治癒率モデル

前節で説明した促進時間治癒率モデルと混合治癒率モデルの統一モデルにおいて、 M が従う分布を COM-Poisson 分布に拡張した治癒率モデルが、Rodrigues et al. (2009) により提案された。このモデルでは、 M に、より柔軟なモデルを仮定している一方、促進時間治癒率モデルと混合治癒率モデルの統一モデルになるとの特徴は失われていることに注意されたい。以下にこのモデルについて、詳述する。

改めて定義を行う。 M は、競合リスクの種類数を表す確率変数で、 M の確率質量関数を $f_M(m) = P(M = m)$, $m \in \mathbb{N}_0$, とする。 $M = m$ が与えられたもとで、確率変数列 W_j , $j = 1, \dots, m$, を考える。 W_j は、 j 番目の競合リスクによる生存時間を表す確率変数で、互いに独立、かつ M とも独立で、同一の分布に従うとする。

競合リスクにおける潜在故障時間モデルの仮定 (Cox and Oakes, 1984) の下で、競合リスク数 M や、競合リスク j による生存時間 W_j は潜在変数であり、対象全体の生存時間 Y は以下で与えられる：

$$Y = \min\{W_0, W_1, \dots, W_M\},$$

ただし、 W_0 は、治癒したイベントの生存時間を表し、これは、再発しないイベントの生存時間であるから、 $P(W_0 = \infty) = 1$ である。このようなリスクの存在は、 $P(M = 0) > 0$ であることを示唆する。 $p_0 := P(M = 0)$ は治癒率である。 $M \sim \text{CMP}(\lambda, \nu)$ の仮定の下で、治癒率は、

$$p_0 = f_{\text{CMP}}(0) = \frac{1}{Z(\lambda, \nu)},$$

となる。さらに、 W_j , $j \in \mathbb{N}$, が従う分布の生存関数を $S(y) := P(W_j \geq y)$ とすると、 Y の生存関数 $S_p(\cdot)$ は、

$$S_p(y) = P(Y \geq y) = \sum_{m=0}^{\infty} P(M = m) S(y)^m = \frac{Z(\lambda S(y), \nu)}{Z(\lambda, \nu)},$$

とシンプルな形で書ける。さらに、 Y の確率密度関数 $f_p(\cdot)$ は、

$$f_p(y) = -\frac{\partial}{\partial y} S_p(y) = \frac{1}{Z(\lambda, \nu)} \frac{f(y)}{S(y)} \sum_{j=1}^{\infty} \frac{j(\lambda S(y))}{(j!)^\nu},$$

となる。なお、 $S(y)$ として、通常、パラメトリック分布の生存関数が選ばれ、特に生存時間分析での生存時間や信頼性工学における寿命を表すのによく用いられるワイブル分布、対数正規分布、ガンマ分布が良く用いられる (Balakrishnan and Pal, 2013, 2015b, 2016)。ワイブル分布とガンマ分布の一般化分布である一般化ガンマ分布を用いた例もみられる (Balakrishnan and Pal, 2015a)。 $\lambda, \nu, S(y)$ が持つパラメータ、並びに後述する共変量モデルにおける回帰係数の推定には、最尤法が用いられる場合がほとんどである (Rodrigues et al., 2009; Balakrishnan and Pal, 2013, 2015a, 2015b)。

4.1.3 共変量を持つ治癒率モデル

共変量を持つ治癒率モデルについても、これまで研究がなされてきた。治癒率モデルに共変

量を導入する方法として、最初に考えられるのは、GLIM(一般化線形モデル)の枠組みにおいてリンク関数を対数関数に取る方法、つまり、 $\mathbf{x}'_i = (1, x_{i1}, \dots, x_{ip})$ を i 番目の対象の共変量、 $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_p)$ として、パラメータ λ を、 $\exp(\mathbf{x}'_i \boldsymbol{\beta})$ におきかえた共変量モデルである (Balakrishnan and Pal, 2012; Cancho et al., 2012; Rodrigues et al., 2009). しかし、 $\nu = 0$ のとき、COM-Poisson 分布は幾何分布となり、 $\lambda < 1$ とならなければいけないが、上記の方法だと、 $\exp(\mathbf{x}'_i \boldsymbol{\beta})$ が 1 以上になりうる。そこで、考え出されたのが、対象が i のときの治癒率 $p_0^{(i)}$ と共変量 \mathbf{x}_i の間のリンク関数をロジット関数にとった

$$p_0^{(i)} = \frac{1}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})} = \frac{1}{Z(\lambda, \nu)}$$

とするモデルである (Balakrishnan and Pal, 2013, 2015a, 2015b; Balakrishnan and Feng, 2018).

COM-Poisson 分布に基づく、共変量を持つ治癒率モデルの適用例として、がんの再発に関するデータへの適用例 (Rodrigues et al., 2009) が良く知られている。この適用例では、悪性がんの一種である皮膚黒色腫を持つ患者に、薬剤(インターフェロンアルファ-2b)を利用したときの治療効果を評価するため、腫瘍の厚さを共変量とし、患者の生存時間データに対しモデルが適用されている。

4.2 COM-Poisson 分布の確率過程モデルへの応用

ポアソン分布は、再生性を持つためポアソン分布に基づくレヴィ過程である定常ポアソン過程を構築できる。一方、COM-Poisson 分布は、 $\nu = 1$ つまり、ポアソン分布の時を除いて再生性を持たないばかりか、 $\nu = 0$, あるいは $\nu = 1$ でない限り、無限分解可能 (infinite divisible) ではない (Geng and Xia, 2022)。よって、特殊ケースを除き、COM-Poisson 分布に基づくレヴィ過程を構築することができない。

しかし、COM-Poisson 分布の代わりに、COM-Poisson 分布にしたがう *i.i.d.* 確率変数の和が従う分布である sCOM-Poisson 分布 (Sum of COM-Poisson 分布) (Sellers et al., 2017) は、再生性を持ち、これに基づくレヴィ過程を構築することが可能である (Zhu et al., 2017)。

$X_i \stackrel{i.i.d.}{\sim} \text{CMP}(\nu, \lambda)$, $i = 1, \dots, n$, のとき、 $S = \sum_{i=1}^n X_i$ が従う分布は、sCOM-Poisson 分布と呼ばれ、以下の確率質量関数を持つ:

$$f_{\text{sCMP}}(s) = \frac{\lambda^s}{(s!)^\nu Z(\lambda, \nu)^n} \sum_{x_1 + \dots + x_n = s} \binom{s}{x_1 \dots x_n}, \quad s = 0, \dots, n.$$

パラメータは、 $\lambda > 0, \nu \geq 0, n \in \mathbb{N}$ であり、これらのパラメータを持つ sCOM-Poisson 分布を、本稿では、sCMP(λ, ν, n) で表す。

計数過程 $\{N(t), t \geq 0\}$ が以下の条件を満たす時、 $\{N(t), t \geq 0\}$ を、パラメータ λ, ν の定常 COM-Poisson 過程と呼ぶ (Zhu et al., 2017) :

- (1) $N(0) = 0$, *a.s.*
- (2) $N(s+t) - N(s) \sim \text{sCMP}(\lambda, \nu, n)$.
- (3) 独立、かつ定常増分を持つ。

パラメータ λ, ν の定常 COM-Poisson 過程は、 $\nu = 1$ のとき、パラメータ λ の定常ポアソン過程となるので、定常ポアソン過程の一般化モデルである。

ただし、sCOM-Poisson 分布の確率質量関数に含まれる $\sum_{x_1 + \dots + x_n = s} \binom{s}{x_1 \dots x_n}$ の計算は、 n がある程度大きくなると計算困難に陥る。このモデルを実データ解析に用いるには、この計算困難問題解決する必要があるが、著者の知る限り未解決である。

5. 最後に

本稿では、COM-Poisson 分布、並びに近年提案された手法も含む代表的な COM-Poisson 分布の推定法について紹介してきた。COM-Poisson 分布の区間推定や仮説検定については、Sellers (2023) でいくつか触れられているものの、点推定に比べると研究はまだまだ少ない。COM-Poisson 分布の区間推定、仮説検定において用いることができるピボットが見つかっていないため、推定量の漸近性質を用いた近似法、並びにブートストラップ法が有効であると考えられる。ちなみに、著者は現在、規格化定数 $Z(\lambda, \nu)$ に依存しない新たな条件付き尤度関数を導出し、それに基づく点推定、区間推定、並びに検定法を提案した論文を投稿中である。この手法は、サンプル数 n が大きくても実行可能であり、実データ解析に問題なく適用可能である。COM-Poisson 分布の区間推定や仮説検定の方法も含めて、別の機会があれば紹介するかもしれない。

謝 辞

有意義かつ適切な助言を頂いた 3 名の匿名査読者に深謝いたします。本研究の一部は、JSPS 科研費 基盤研究(C) (24K07952)の助成を受けました。ここに記して感謝の意を表します。

参 考 文 献

- Balakrishnan, N. and Feng, T. (2018). Proportional odds under Conway-Maxwell-Poisson cure rate model and associated likelihood inference, *Statistics, Optimization & Information Computing*, **6**, 305–334.
- Balakrishnan, N. and Pal, S. (2012). EM algorithm-based likelihood estimation for some cure rate models, *Journal of Statistical Theory and Practice*, **6**, 698–724.
- Balakrishnan, N. and Pal, S. (2013). Lognormal lifetimes and likelihood-based inference for flexible cure rate models based on COM-Poisson family, *Computational Statistics & Data Analysis*, **67**, 41–67.
- Balakrishnan, N. and Pal, S. (2015a). An EM algorithm for the estimation of parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood- and information-based methods, *Computational Statistics*, **30**, 151–189.
- Balakrishnan, N. and Pal, S. (2015b). Likelihood inference for flexible cure rate models with gamma lifetimes, *Communications in Statistics — Theory and Methods*, **44**, 4007–4048.
- Balakrishnan, N. and Pal, S. (2016). Expectation maximization-based likelihood inference for flexible cure rate models with Weibull lifetimes, *Statistical Methods in Medical Research*, **25**, 1535–1563.
- Balakrishnan, N., Barui, S. and Milienos, F. S. (2024). Piecewise linear approximations of baseline under proportional hazards based COM-Poisson cure models, *Communications in Statistics — Simulation and Computation*, **53**, 1135–1160.
- Bedbur, S., Kamps, U. and Imm, A. (2023). On the existence of maximum likelihood estimates for the parameters of the Conway-Maxwell-Poisson distribution, *Latin American Journal of Probability and Mathematical Statistics*, **20**, 561–575.
- Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association*, **47**, 501–515.
- Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society, Series B*, **11**, 15–53.
- Cancho, V. G., de Castro, M. and Rodrigues, J. (2012). A Bayesian analysis of the Conway-Maxwell-Poisson cure rate model, *Statistical Papers*, **53**, 165–176.

- Chen, M. H., Ibrahim, J. G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association*, **94**, 909–919.
- Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates, *Journal of Industrial Engineering*, **12**, 132–136.
- Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall, London.
- Geng, X. and Xia, A. (2022). When is the Conway-Maxwell-Poisson distribution infinitely divisible?, *Statistics and Probability Letters*, **181**, <https://doi.org/10.1016/j.spl.2021.109264>.
- Hanleya, J. A. and Bhatnagar, S. (2022). The “Poisson” distribution: History, reenactments, adaptations, *The American Statistician*, **76**, 363–371.
- He, Z. and Emura, T. (2019). The COM-Poisson cure rate model for survival data-computational aspects, *Journal of the Chinese Statistical Association*, **57**, 1–42.
- Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian Survival Analysis*, Springer, New York.
- Majakwara, J. and Pal, S. (2019). On some inferential issues for the destructive COM-Poisson-generalized gamma regression cure rate model, *Communications in Statistics — Simulation and Computation*, **48**, 3118–3142.
- Minka, T. P., Shmueli, G., Kadane, J. B., Borle, S. and Boatwright, P. (2003). Computing with the COM-Poisson distribution, Technical Report 776, Department of Statistics, Carnegie Mellon University, Pittsburgh.
- Nightingale, F. (1858). *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army, Founded Chiefly on the Experience of the Late War*, Harrison of Sons, London.
- Rodrigues, J., de Castro, M., Cancho, V. G. and Balakrishnan, N. (2009). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data, *Journal of Statistical Planning and Inference*, **139**, 3605–3611.
- Sellers, K. F. (2023). *The Conway-Maxwell-Poisson Distribution*, Cambridge University Press, Cambridge.
- Sellers, K. F., Swift, A. W. and Weems, K. S. (2017). A flexible distribution class for count data, *Journal of Statistical Distributions and Applications*, **4**, <https://doi.org/10.1186/s40488-017-0077-0>.
- Sellers, K. F., Lotze, T. and Raim, A. (2019). COM-PoissonReg: Conway-Maxwell-Poisson (COM-Poisson) regression, version 0.7.0., <https://cran.r-project.org/web/packages/COM-PoissonReg/index.html> (最終アクセス日 2025 年 2 月 15 日).
- Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution, *The Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **54**, 127–142.
- Snow, J. (1858). *On Chloroform and Other Anaesthetics and Their Action and Administration*, John Churchill, London.
- Tomio, Y. and Nagatsuka, H. (2022). A conditional maximum likelihood estimation of the COM-Poisson distribution and its uniqueness and existence, *Total Quality Science*, **7**, 137–148.
- von Bortkewitsch, L. (1898). *Das Gesetz der Kleinen Zahlen*, Teubner, Leipzig.
- Wani, M. K. and Ahmad, P. B. (2024). One-inflated zero-truncated Poisson distribution: Statistical properties and real life applications, *Annals of Data Science*, **71**, <https://doi.org/10.1007/s40745-024-00526-3>.
- Wiangnak, P. and Pal, S. (2018). Gamma lifetimes and associated inference for interval-censored cure rate model with COM-Poisson competing cause, *Communications in Statistics — Theory and Methods*, **47**, 1491–1509.
- Zhu, L., Sellers, K. F., Morris, D. S. and Shmueli, G. (2017). A generalized stochastic process for count data, *The American Statistician*, **71**, 71–80.

On COM-Poisson Distribution

—An Extended Poisson Model for Under- and Over-dispersed Count Data—

Hideki Nagatsuka

Department of Data Science for Business Innovation, Chuo University

The Poisson distribution is one of the most popular distributions for modeling count data. While the Poisson distribution is a simple model and easy to treat, its applicability is often limited due to the equi-dispersion assumption. The Conway-Maxwell-Poisson (abbreviated as COM-Poisson) distribution, provided by Conway and Maxwell in 1962, is a more flexible model for modeling count data, which generalizes the Poisson distribution by adding a parameter to model over-dispersion and under-dispersion. In the survival analysis and reliability analysis, there have been many papers in recent years on cure rate models using the COM-Poisson distribution, which has attracted a lot of attention. The COM-Poisson distribution is a simple model with only two parameters, but it can flexibly be fitted a wide range of overdispersion and underdispersion phenomena and includes the Poisson and geometric distributions as special cases, and the Bernoulli distribution as a limiting case. These features distinguish it from other generalized models of the Poisson distribution. However, the price of these benefits is that the normalization constant of the probability function includes an infinite series, which makes estimation and testing difficult. So far, various approximation methods for the normalization constant or estimation methods that avoid calculating the normalization constant have appeared in the literature. In this paper, we introduce the COM-Poisson distribution and its prominent estimation methods for the COM-Poisson distribution, including recently proposed methods appeared in recent years.