# Some goodness-of-fit tests for the Poisson distribution

## with applications in Biodosimetry

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

**C.H. Weiß**

Department of Mathematics & Statistics,
Helmut Schmidt University, Hamburg

MATH STAT

**P. Puig**

Departament de Matemàtiques,
Universitat Autònoma de Barcelona

# Outline & Full Paper

- New characterizations of Poisson distribution
- New Goodness-of-Fit (GoF) test for Poisson distribution
- Performance analysis by simulations
- Several examples of applications in Biodosimetry

**Full paper:**

Puig & Weiß (2020) Some goodness-of-fit tests for the Poisson distribution with applications in Biodosimetry. *Computational Statistics and Data Analysis* **144**, 106878.

($\rightarrow$ open access)

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

# Some characterizations of Poisson distribution

Motivation & Approach

**Common approach in developing GOF tests:**

utilize a characterization of considered distribution family, see Nikitin (2017) for recent survey.

**Example:** normal distribution characterized by equality

$$X \stackrel{d}{=} a \cdot X + b \cdot Y \quad \text{with } a, b \in (0,1) \text{ satisfying } a^2 + b^2 = 1,$$

which only holds iff the i. i. d. and centered r. v. $X, Y$ are normally distributed, $X, Y \sim \mathrm{N}(0, \sigma^2)$ (Nikitin, 2017, p. 13).

Among count r. v., i. e., if $X$ has range $\mathbb{N}_0 = \{0, 1, \ldots\}$, the Poisson distribution with mean $\lambda > 0$, Poi($\lambda$), constitutes the "discrete normal" distribution.

**Idea:** There should be characterization analogous to $X \stackrel{d}{=} a \cdot X + b \cdot Y$ for "continuous normal" distribution. This could be utilized for constructing GoF tests.

However, multiplications in $X \stackrel{d}{=} a \cdot X + b \cdot Y$ would destroy the integer nature of Poisson r. v. Thus, integer substitute to multiplication required.

## Some characterizations of Poisson distribution



Most popular substitute:

**binomial thinning** operator (Steutel & van Harn, 1979):

$$\alpha \circ X \;\sim\; \text{Bin}(X, \alpha) \text{ with } \alpha \in (0;1);$$

has *integer range* $\{0, \ldots, X\}$

and behaves *multiplicative* in mean: $E[\alpha \circ X] \;=\; E[\alpha \cdot X]$.

Furthermore, binomial thinning preserves Poisson property:

$$X \sim \text{Poi}(\lambda) \quad \Rightarrow \quad \alpha \circ X \sim \text{Poi}(\alpha \cdot \lambda).$$

**Idea:** use binomial thinning to construct Poisson identity.

# Some characterizations of Poisson distribution

**Theorem 1:** Let $X_1$ and $X_2$ be i. i. d. count r. v.,

and let $Y_\alpha = \alpha \circ X_1 + (1-\alpha) \circ X_2$.

Then, $X_i \sim$ Poi$(\lambda)$ iff any of following conditions hold:

(a) $Y_\alpha$ has same distribution as $X_i$ for all $\alpha \in (0,1)$;

(b) $X_i$ has a first-order moment, and

$Y_\alpha$ has same distribution as $X_i$ for a certain $\alpha \in (0,1)$.

**Proof:** considers probability generating function (pgf)

$\phi_X(s) = E[s^X]$, where $\phi_{\alpha \circ X}(s) \;=\; \phi_X(1-\alpha+\alpha\, s)$,

and utilizes Cauchy's functional equation.

See details in Puig & Weiß (2020). #

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH
STAT

# New GoF tests of Poisson distribution

Definition & Properties

# New GoF tests of Poisson distribution



Since binomial thinning is random operator, test statistics cannot be constructed based on $X \stackrel{d}{=} \alpha \circ X_1 + (1-\alpha) \circ X_2$.

Thus, compare pgfs of left- and right-hand side, as identity

$$\phi(s) \; = \; \phi(1-\alpha+\alpha s)\, \phi\big(\alpha + (1-\alpha)s\big),$$

holds for Poisson distribution.

Tests statistics relying on discrepancies

$$\Big\| \phi(t) \; - \; \phi(1-\alpha+\alpha t)\, \phi\big(\alpha + (1-\alpha)t\big) \Big\|$$

using $L^1$-, $L^2$-, or $L^\infty$-norm for $\|\cdot\|$.

If members of **LC class** as alternatives (Puig & Kokonendji, 2018), i. e., count r. v. having a *log-convex pgf* in $[0,1]$, then following refinements hold.

**Theorem 2:** Let $\phi(t)$ be pgf of r. v. from LC-class. Then, for all $t, \alpha \in [0,1]$,

$$\phi(t) \ \geq \ \phi(1-\alpha+\alpha t)\,\phi\big(\alpha+(1-\alpha)t\big).$$

**Proposition 1:** Consider $g(\alpha) = \phi(t) - \phi\big(1+\alpha(t-1)\big)\,\phi\big(t-\alpha(t-1)\big)$, where $t \neq 1$ is fixed and $\phi(t)$ is pgf of non-Poisson r. v. from LC-class. Then, $g(\alpha)$ is maximized for $\alpha = 1/2$.

For GoF test statistics relying on above pgf (in)equalities, replace pgf by empirical pgf (epgf) defined as

$$\widehat{\phi}(s) = \frac{1}{n}\sum_{i=1}^{n} s^{X_i} = \frac{1}{n}\sum_{j=0}^{m} f_j\, s^j,$$

where $f_j = \{\#X_i : X_i = j\}$ and $m = \max\{X_1, \ldots, X_n\}$; see Gürtler & Henze (2000) for further epgf-based GoF tests.

If testing against LC-alternatives,

consider signed discrepancy and evaluate at $\alpha = 1/2$:

$$\widehat{\Delta}_1 \;=\; \int_0^1 \left( \widehat{\phi}(t) - \left[ \widehat{\phi}\left( \frac{t+1}{2} \right) \right]^2 \right) dt\,,$$

$$\widehat{\Delta}_\infty \;=\; \max_{t\in[0,1]} \left\{ \widehat{\phi}(t) - \left[ \widehat{\phi}\left( \frac{t+1}{2} \right) \right]^2 \right\}.$$

If testing against more general alternatives,

consider absolute discrepancy instead:

$$\hat{\Delta}_1^* \;=\; \int_0^1 \left| \hat{\phi}(t) - \left[\hat{\phi}\left(\frac{t+1}{2}\right)\right]^2 \right| \, dt\,,$$

$$\hat{\Delta}_2 \;=\; \int_0^1 \left( \hat{\phi}(t) - \left[\hat{\phi}\left(\frac{t+1}{2}\right)\right]^2 \right)^2 \, dt\,,$$

$$\hat{\Delta}_\infty^* \;=\; \max_{t\in[0,1]} \left\{ \left| \hat{\phi}(t) - \left[\hat{\phi}\left(\frac{t+1}{2}\right)\right]^2 \right| \right\}.$$

If **additional weighting scheme** "$\cdot\, t^a$" as recommended by Gürtler & Henze (2000), we denote $\hat{\Delta}_{1,a}$, $\hat{\Delta}_{1,a}^*$ and $\hat{\Delta}_{2,a}$.

**Implementation of new GoF tests:**

- **Computation of test statistics** $\hat{\Delta}_1, \ldots$:

  Although one could explicitly solve integrals in $\hat{\Delta}_1, \hat{\Delta}_2$,

  so integrals would turn to sums,

  most efficient implementation in R by numerical integration

  (and using numerical optimization for $\hat{\Delta}_\infty$).

- **Computation of critical values**: (. . . )

## Implementation of new GoF tests:

- **Computation of test statistics** $\widehat{\Delta}_1, \ldots$: (. . . )
- **Computation of critical values**:

  Under Poisson null, $(X_1, \ldots, X_n)\big|S$ is multinomial

  with parameters $(S,\ 1/n, \ldots, 1/n)$,

  see González-Barrios et al. (2006),

  where $S = \sum_{i=1}^{n} X_i$ is sufficient statistic.

  Thus, percentiles of statistics' exact distribution

  by Monte–Carlo simulation

  (better accuracy with more replications).

# Conclusions



- Comprehensive simulation study, see Puig & Weiß (2020). Power of new tests often better than for selected competitors. Particularly good performance of $\widehat{\Delta}_1^{(*)}$ and $\widehat{\Delta}_{1,5}^{(*)}$.
- Several examples from Biodosimetry, where important to identify whether distribution of chromosome aberrations from patient's blood sample is Poisson or not (essential for dose estimation and to evaluate extension of irradiation).

# Outlook



- We developed GoF tests based on Poisson identity

$$\begin{aligned} X &\overset{d}{=} \alpha \circ X_1 + (1-\alpha)\circ X_2 \\ \Leftrightarrow \qquad \phi(s) &= \phi(1-\alpha+\alpha s)\,\phi\big(\alpha + (1-\alpha)s\big). \end{aligned}$$

  Another Poisson identity (Weiß & Aleksandrov, 2020):

  **Stein–Chen identity** $\quad E\big[X\cdot f(X)\big] \;=\; \lambda\cdot E\big[f(X+1)\big]$,

  see Aleksandrov et al. (2021) for GoF tests.

- **Research in progress:**

  GoF tests based on **binomial Stein identity**,

  see **talk by B. Aleksandrov in "Time Series" section:**

  *"Novel goodness-of-fit tests for binomial count time series"*.

# Thank You

# for Your Interest!

**Christian H. Weiß**

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

`weissc@hsu-hh.de`

**Puig & Weiß (2020) Some goodness-of-fit tests for the Poisson distribution with applications in Biodosimetry.**
***Computational Statistics and Data Analysis* 144, 106878.**
**($\rightarrow$ open access)**

Aleksandrov et al. (2021) Goodness-of-fit tests ... *Stat. Neerl.*, in press.

González-Barrios et al. (2006) Goodness of fit ... *Metrika* **64**, 77–94.

Gürtler & Henze (2000) Recent and classical ... *JSPI* **90**, 207–225.

Nikitin (2017) Tests based on characterizations ... *ACUTM* **21**, 3–24.

Puig & Kokonendji (2018) Non-parametric estimation ... *SJS* **45**, 347–365.

Steutel & van Harn (1979) Discrete analogues ... *Ann. Probab.* **7**, 893–899.

Weiß & Aleksandrov (2020) Computing ... Stein–Chen Identities. *TAS*.