An R package for the Generalized Hermite distributions

David Moriña[1], Manuel Higueras[2,3], Pere Puig[3]

[1]CREAL
[2]Biological Effects Department, PHE
[3]Departament de Matemàtiques, UAB

I JCE SEB

Facultat de Matemàtiques, Universitat de València

20^{th} January 2015

Generalized Hermite distribution

- It is a two-parameter count data family distribution.
- These distributions are closed under convolutions.
- They satisfy that the sample mean is the maximum likelihood estimator of the population mean.
- They hold multi-modality, overdispersion (the variance greater than the mean), and zero inflation.
- A Generalized Hermite distribution of order m is represented by

$$X_1 + mX_2,$$

  where X_1 and X_2 are Poisson independent variables with respective intensities a and b, and $m \geq 2 \in \mathbb{Z}$.
- Notation: Herm(a, b, m).

Package `hermite`

Let $\mathcal{Y}$ be a Herm(a, b, m), x a count, p a probability, n a positive integer and X a sample of counts, the user can:

- Calculate the probability mass function for a given count x: `dhermite(x, a, b, m)`.
- Calculate the cumulative distribution function for a given count: `phermite(x, a, b, m, lower.tail)`.
- Calculate the quantile function for a given probability p: `qhermite(p, a, b, m, lower.tail)`.
- Simulate a sample of size n of $\mathcal{Y}$: `rhermite(n, a, b, m)`.
- Estimate the maximum likelihood for parameters a, b and m for a given sample X: `mle.hermite(X, a, b, m)`.
- It is available at CRAN repository:
  http://cran.r-project.org/web/packages/hermite/index.html

Function dhermite

- This function calculates the probabilities using the recurrence relation

$$p_k = \frac{\mu}{k(m-1)}\left(p_{k-m}(d-1) + p_{k-1}(m-d)\right), k \geq m,$$

where $p_k = P(Y = k)$ and the first values can be computed as

$$p_k = p_0 \frac{\mu^k}{k!}\left(\frac{m-d}{m-1}\right)^k, k = 1, \ldots, m-1,$$

and, $\mu = a + mb$ is the population mean and

$$d = \frac{a + m^2 b}{a + mb}$$

is the dispersion index (the ratio of the variance to the mean).

- In case that a or b are greater than 20 the probability of Y taking k counts is approximated using an Edgeworth expansion of the distribution function, $P(Y = k) = F_H(k) - F_H(k-1)$.

Function `phermite`

If a or $b > 20$, the distribution function is approximated by means of an Edgeworth expansion, using the following expression

$$F_H(k) \approx \Phi(k^*) - \phi(k^*) \cdot \left(\frac{1}{6}\gamma_1 He_2(k^*) + \frac{1}{24}\gamma_2 He_3(k^*) + \frac{1}{72}\gamma_1^2 He_5(k^*) \right),$$

where ϕ is the typified normal density function, $He_n(k)$ are the nth-degree probabilists' Hermite polynomials, k^* is the typified continuous correction of k

$$k^* = 1 + \frac{1}{24(a + m^2 b)} \cdot \frac{k + 0.5 - a - mb}{\sqrt{a + mb}},$$

and γ_1 and γ_2 are respectively the skewness and the excess kurtosis of $\mathcal{Y}$.

Function `qhermite`

When the parameters a or b are over 20, a Cornish-Fisher expansion is used to approximate the quantile function. The Cornish-Fisher expansion uses the following expression

$$y_p \approx \left(u_p + \tfrac{1}{6}\gamma_1 He_2(u_p) + \tfrac{1}{24}\gamma_2 He_3(u_p) - \tfrac{1}{36}\gamma_1^2(2u_p^3 - 5u_p)\right) \cdot \sqrt{a + m^2 b} + a + mb,$$

where u_p is the p quantile of the typified normal distribution.

Function `rhermite`

This function uses the fact that

$$\mathcal{Y} = X_1 + mX_2$$

where $X_1 \sim \mathrm{Pois}(a)$ and $X_2 \sim \mathrm{Pois}(b)$ independent, to simulate Generalized Hermite counts.

Function `mle.hermite`

- **Proposition.** Let $x_1, \ldots, x_n$ be a random sample from a generalized Hermite population with fixed m. Then, the maximum likelihood equations have solution if and only if $\frac{n^{(m)}}{\bar{x}^m} > 1$, where $\bar{x}$ is the sample mean and $n^{(m)}$ is the m-th order sample factorial moment, $n^{(m)} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i(x_i - 1) \cdots (x_i - m + 1)$.
- If no initial value is supplied for the order m,

$$\hat{m} = \left| \frac{s^2/\bar{x} - 1}{1 + log(p0)/\bar{x}} \right|.$$

- This function returns a list with five elements: the estimates of parameters a and b, the maximum likelihood value, the Hessian matrix, the likelihood ratio test statistic (for the Poisson assumption), and the p-value of this test.
- For the next version of package `hermite`, this function will be replaced by function `glm.hermite`, which will fit Generalized Hermite regression models.

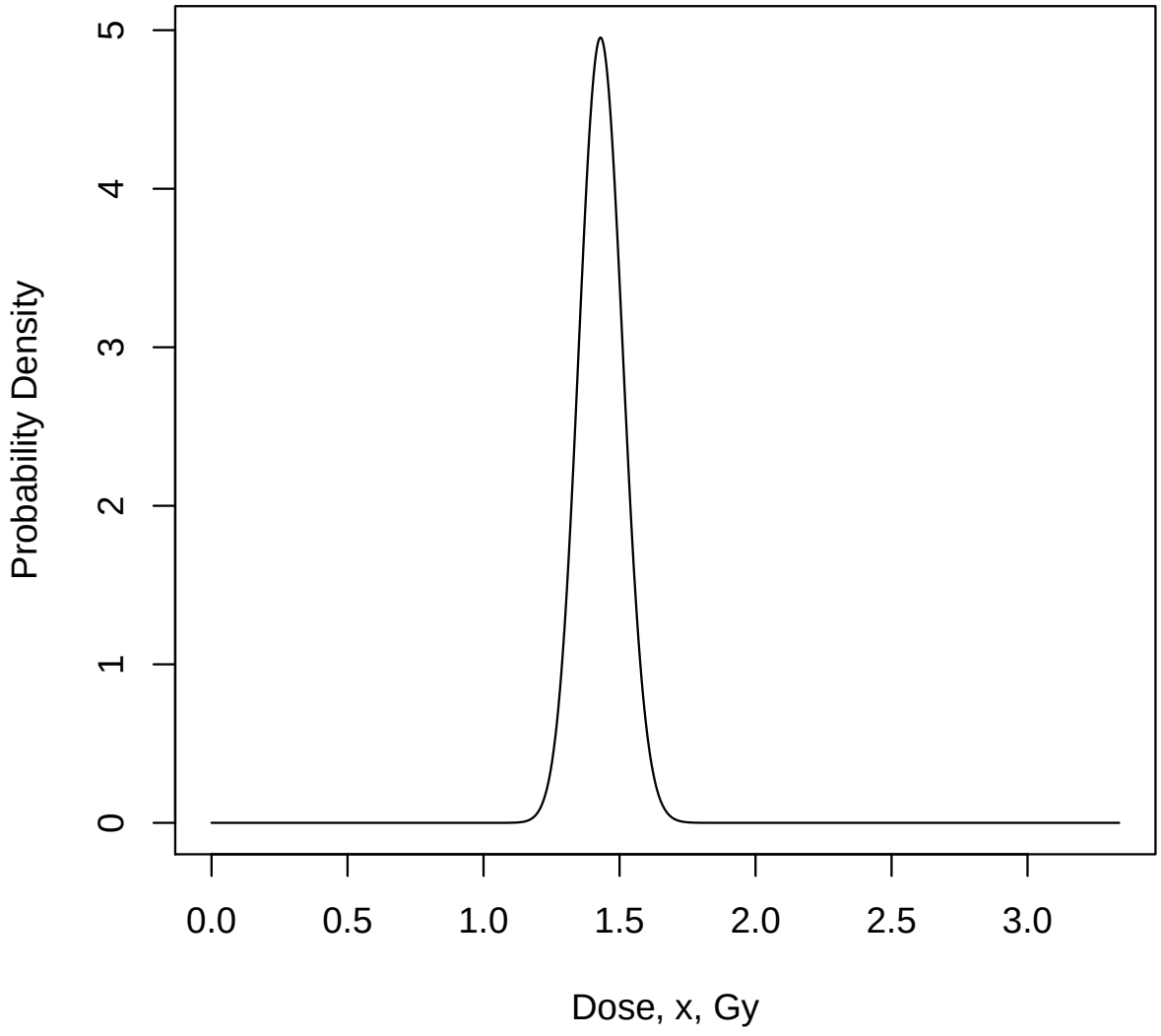
Example. Higueras et al 2015 (I)

The Bayesian-type estimation of the absorbed dose by Cobalt-60 gamma rays after the *in vitro* irradiation of a sample of blood cells is given by a density proportional to the probability mass function of a Hermite distribution taking 102 counts whose mean and variance are functions of the dose x, respectively $\mu(x) = 45.939x^2 + 5.661x$ and $v(x) = 8.913x^4 - 22.553x^3 + 69.571x^2 + 5.661x$.

Mode	Expected	SD	95% CI
1.430	1.432	0.081	(1.275, 1.591)

Table : Statistics summary of the resulting dose density.

Example. Higueras et al 2015 (II)



Bibliography

- Moriña D, Higueras M, Puig P, Oliveira M. The `R` Package `hermite` (in preparation).
- Puig P (2003). Characterizing Additively Closed Discrete Models by a Property of Their Maximum Likelihood Estimators, with an Application to Generalized Hermite Distributions. *Journal of the American Statistical Association*, **98**(463), 687–692.
- Higueras M, Puig P, Ainsbury EA, Rothkamm K (2015). A new Inverse Regression Model Applied to Radiation Biodosimetry. *Proceedings of the Royal Society A*. DOI: 10.1098/rspa.2014.0588.

Acknowledgements

This work was partially funded by the grant MTM2012-31118 and by the grant UNAB10-4E- 378 co-funded by FEDER “A way to build Europe”.