The work of Fernando de Helguero on non-normality arising from selection

Adelchi Azzalini and Giuliana Regoli

Presented at "A Celebration of N. Balakrishnan's
30 years of Contributions to Statistics"
Hong Kong, 28–31 December 2011

Università degli Studi di Padova

Skew-symmetric/hidden-truncation models

- much recent work about distributions like
$$f(z) = \text{const} \times f_0(z)\, G(\lambda_0 + \lambda_1\, z)$$
- strongly linked to a selection mechanism
$$Z = (Z_0 | X < \lambda_0 + \lambda_1\, Z_0)$$
- main special case
$$f(z) = \text{const} \times \varphi(z)\, \Phi(\lambda_0 + \lambda_1\, z)$$
- this mechanism features marginally in several older papers
- . . . but obtained as a by-product in different type of problems
- Q: when did this construction appear first with the aim of *building an alternative to the Gaussian family*?

Fernando de Helguero (1880-1908)

Fernando de Helguero, in brief

- born in 1880 near Florence, Italy
- studied mathematics at Univ. Rome
- teacher of mathematics while student of natural sciences
- mathematics + biology $\Rightarrow$ statistics and biometry
- smart, active, excellent publications, but...
- deceased at age 28 in earthquake

Here we focus on a specific and nearly unknown of his contributions

Historical background

Distribution theory at in the beginning of 20th century

Gaussian distribution was the 'normal' one,
but wider families were called for

- Gram-Charlier and Edgeworth expansions
- mixture of normals, Pearson's work
- Pearson system of 12 families

The study of data distribution was a key target of statistics

de Helguero's view

FdH views on non-Gaussian data distributions

- principle: the role of statistics is not simply to fit data, but also to help understanding the underlying mechanism
- $\Rightarrow$ Edgeworth expansions and Pearson system ☹
- mixtures (of normals) are more helpful
- in 1908, at 4th ICM, FdH presents an innovative proposal

de Helguero's formulation

- A real phenomenon would generate a Gaussian distribution
$$X \sim \mathrm{N}(b, \sigma^2)$$
if some external 'cause' did not perturb it
- often a realistic perturbation mechanism is that a value x is censored with probability $\varphi(x)$
$\Rightarrow$ a curve "perturbed by selection"
- hence the observable distribution is proportional to
$$\mathrm{const} \times \frac{1}{\sqrt{2\,\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-b}{\sigma}\right)^2\right\}[1-\varphi(x)]$$
- $\varphi(x)$ can take various forms, the simplest is linear:
$$\varphi(x) = A(x-b) + B$$

Modern reading

$$\text{const} \times \frac{1}{\sqrt{2\,\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-b}{\sigma}\right)^2\right\} \underbrace{[1 - B - A(x-b)]}_{1-\varphi(x)}$$

- $1 - \varphi(x)$ is linear in the interval of main interest, constrained to be 0 to the left, 1 to the right (or vice versa)
- in current view, $\mathrm{N}(b, \sigma^2)$ with a uniform 'skewing factor'
- in this sense FdH anticipates the key ingredients of the current 'skew' literature

de Helguero's formulation (ctd)

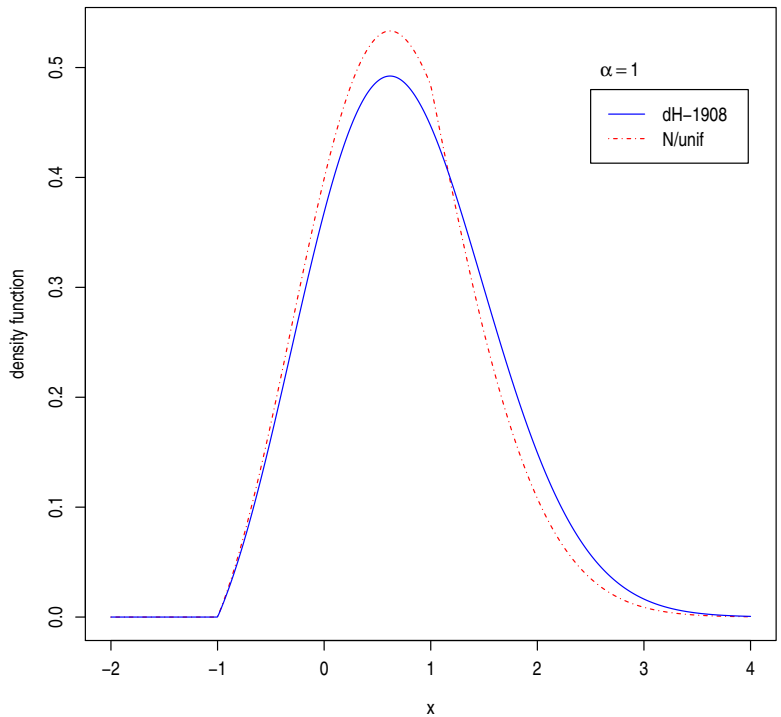
A further step:

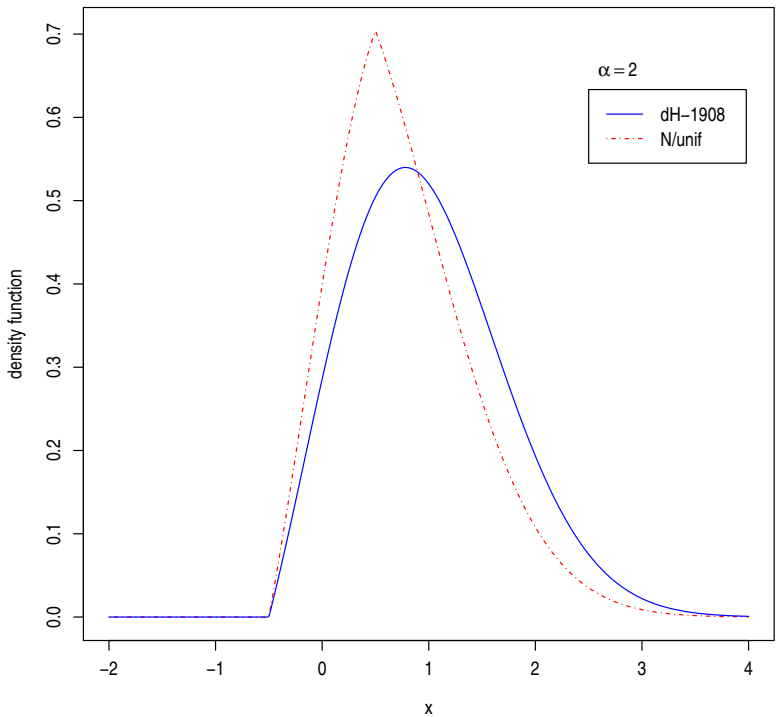
- in addition to *thinning* by censoring,
  FdH includes the possibility of *thickening* a tail
- the same mathematical expression holds
- only the 0 constraint is retained for $1 - \varphi(x)$, it can exceed 1

Operative stage:

- FdH obtains expressions of moments up to order 3
- data fitting by method of moments
- develop special tables to solve non-linear equations
- various numerical illustrations with real data

Some de Helguero's distributions





Final comments

- FdH anticipated the key idea of current 'skew distributions'
- because of his premature passing his curves "perturbed by selection" went unnoticed
- if he had survived, he would have been a major role in the Italian statistical and biometrical world, not only for this specific theme.

References

- de Helguero, F. (1909a). Sulla rappresentazione analitica delle curve abnormali. In G. Castelnuovo (Ed.), *Atti del IV Congresso Internazionale dei Matematici (Roma, 6–11 Aprile 1908)*. Available at `http://www.mathunion.org/ICM/ICM1908.3/Main/icm1908.3.0288.0299.ocr.pdf`
- de Helguero, F. (1909b). Sulla rappresentazione analitica delle curve statistiche. *Giornale degli Economisti*, XXXVIII, serie 2ª, 241–265
- de Helguero, F. (1972). *Scritti vari*. Padova: Cedam. (Collected papers)