Do we need more probability distributions?
(Some remarks on criteria for model building)

Adelchi Azzalini

Università di Padova, Italia

SEIO 2010, A Coruña
'Skewing Symmetry 25 Years on'
14th September 2010

Foreword

- thanks for this session!
- a non-technical talk
- not quite a ‘skew-distributions’ talk (. . . not explicitly)

The quest for probability distributions

Q do we need more probability distributions?

A_0 yes, however we already have lots to choose from

A_1 of course we do! . . . to build better statistical models

Q what is a 'better' model?

A_0 one which fits the data better

A_1 . . . well, not only that

Desiderata for a statistical/probabilistic model

Cox (1997):
Desiderata for a probabilistic model can be formulated in various ways. One such is the following (. . .):

1. the model should establish a link with underlying substantive knowledge or theory;
3. the model should be consistent with or suggest a possible process that might have generated the data;
6. the fit to data should be adequate.

Additional desideratum: mathematical tractability.

About available distributions, $d = 1$

- in $d = 1$ case, a plethora of distributions are available
- still, in practice, the choice is often driven by motivations other than C-1 and C-3
- model building may require to develop new probability distributions
- aim at model building, not not mere data fitting

About available distributions, $d > 1$

- in $d > 1$ case much fewer 'native' distributions
- however we have tools to generate more, such as:
  - mixtures of a basic type (usually normals)
  - copulae (very popular!)
  - perturbation of a symmetric density f_0 via $f(x) = 2\, f_0(x)\, G\{w(x)\}$
  - ...
- hence very flexible families can be generated

Caveat faber: flexibility is not the Holy Grail

- Flexibility of a distribution *may* lead to good data fit,
- . . . but it may also lead to unidentifiable model.
- Even technically identifiable models may hide problems:
  - may have a nearly flat log-likelihood
  - or a log-likelihood with many local maxima
- Use of a ultra-general mechanisms requires caution
  - danger of blind 'distributional automat' style of use
  - easy to adopt an over-complicated model
  - then a model which we do no really understand
  - and/or with above risks in inferential stage

(NB: this concerns both classical and Bayesian inference)

Caveat faber: other issues with 'distributional automats'

- recall earlier desiderata C-1 and C-3,
  hardly compatible with 'distributional automats'
- the problem may require certain formal properties to hold, e. g.
  - closure under marginalization of some components,
    (may require that $d = 5, \ldots, d = 1$ distributions are of same family)
  - or, vice versa, dimensionality upgrade must be seamless
  - the final target is, say, a linear combination of the variables,
    hence *its* distribution must be known (portfolio selection)
  - hence simple distributional properties are required

Conclusions

We need to work with probability distributions

- whose properties we understand
  both on the probability and the inferential statistics side,
- are linked to some 'physical' form of genesis,
- enjoy tractable formal properties,
- *and* are flexible.

No single family of distributions can satisfy all these requirements for *all* problems of interest.

Hence more study of probability distributions is required,
both for existing ones and for new constructions.

Reference

- Cox (1997). Int. Statist.Rev. 65, 261–290.