

Comments on Clauset's 'Trends and Fluctuations' paper



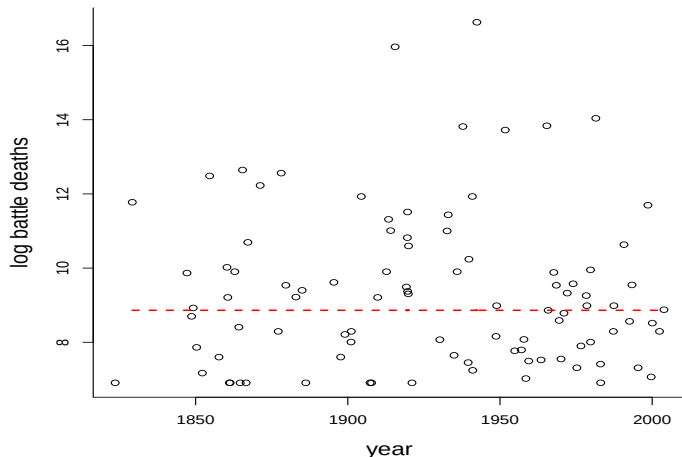
Nils Lid Hjort

Department of Mathematics, University of Oslo

Towards a More Peaceful World [insert '!' or '?' here],
January 2018

Wars over time, $\log(\text{battle deaths})$:

95 wars, from 1823 to 2003, with $z \geq 1000$. Clauset's [main story](#) telling picture, with $y = \log z$: Does the world look stationary?



Some (out of many) themes

Plan:

- A comments on Clauset's data, models, methods;
- B scaled data (with other models, and other conclusions);
- C other statistical issues.

Some (out of many) key words:

- ▶ data (coding, definitions, moving targets)
- ▶ statistical commentary
- ▶ more careful modelling
- ▶ searching for a change-point
- ▶ more refined extreme value model
- ▶ scaling issues: z_i/w_i with w_i world population for war i
- ▶ Some additional analysis for z_i/w_i
- ▶ A (partial) list of further statistical issues

A: Clauset's data, models, methods

Main story is being told via (x_i, z_i) for 95 wars (from 1823 to 2003), with

x_i = time of onset and z_i = number of battle deaths.

Classical issues, 'history and political science' vs. 'statistical modelling, interpretation, uses':

- ▶ definitions might be changing under our feet
- ▶ each dramatic data point has its own history
- ▶ interpretation and usefulness of statistical models

Almost regardless of subtleties: statisticians (data scientists, number crunchers) can use data bases to

- (a) look for features, 'significant effects',
- (b) help interpret the past,
- (c) predict the future.

The stationary view

A simple (but decently robust) **stationary model** world view:

- (i) between-onset-times $d_i = x_i - x_{i-1}$ are i.i.d.,
- (ii) the sizes $y_i = \log z_i$ are i.i.d.

There are certain specialisations of (i) and (ii).

Ideal Poisson process like world:

- (i) the between-times d_i are $\text{Expo}(\lambda)$,
- (ii) the z_i have power-law tails.

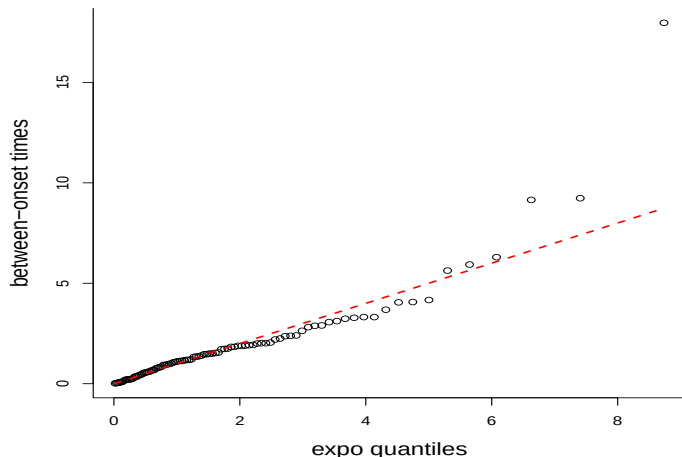
To (ii): That z_i above threshold z_0 have density c/z^α for $z \geq z_0$ is equivalent to y_i above threshold y_0 having density $\theta \exp\{-\theta(y - y_0)\}$, i.e.

$$y_i - y_0 \sim \text{Expo}(\theta),$$

with $\theta = \alpha - 1$.

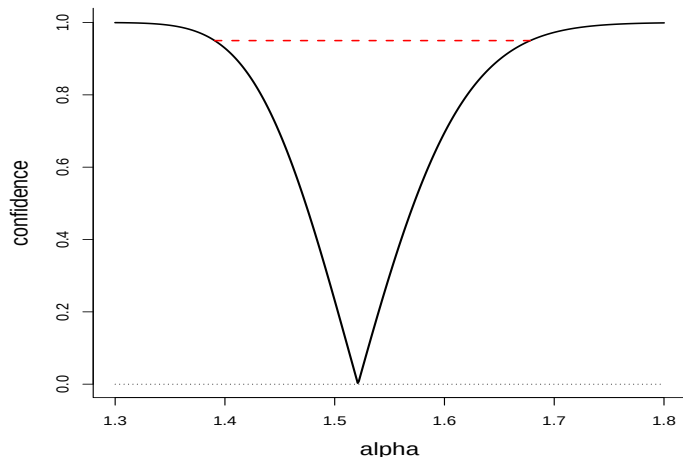
Modelling between-onsets times

The [simple Poisson process view](#): the $d_i = x_i - x_{i-1}$ are i.i.d. and $\text{Expo}(\lambda)$. Answer: **not quite**. Better with a mixed-out Expo, like $f(d) = \int \lambda \exp(-\lambda d) \pi(\lambda) d\lambda$.



Precision of α estimation

To assess the α in c/z^α for z above threshold, I go to $y = \log z$ above threshold y_0 and exponentiality. The optimal confidence curve $cc(\alpha)$:



More accurate estimation of (threshold, tail index)?

Starting with battle deaths z_i for 95 wars, we care for (z_0, α) , the threshold and the tail index. Clauset finds threshold = 7061 and $\alpha = 1.53$ (actually, 1.5247, so rounding to 1.52).

Transformed to $y_i = \log z_i$, the question is equivalent to how to estimate (y_0, θ) for the model that says ' y_i above y_0 behave like $\theta \exp\{-\theta(y - y_0)\}$ '. Method involves pushing the threshold to check when exponentiality appears ok (via Kolmogorov-Smirnov measures).

This is fully ok, but somewhat ad hoc, and better methods might have worked out – using more of the full information in the 95 wars, and assessing (y_0, θ) jointly.

More accurate tails?

The **power laws** $f(z) \approx c/z^\alpha$ for z above threshold are prevalent for war statistics (and elsewhere in statistics), but often ad hoc.

Can also exploit **theory of extreme value statistics**. Theorems of Gnedenko (1943) and onwards indicate:

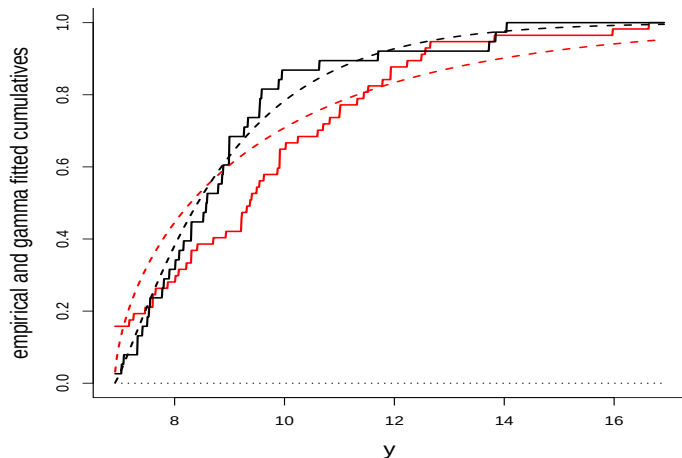
$$G(v) = 1 - (1 - a\theta v)^{1/a} \quad \text{for } v > 0$$

is a **more accurate cdf** for $v_i = y_i - y_0$ above threshold, than a mere $G(y) = 1 - \exp(-\theta v)$. This is the **limiting case** $a \rightarrow 0$, corresponding to the c/z^α behaviour (and with $\theta = \alpha - 1$).

I've fitted the COW data to this **two-parameter tail distribution**:
 $\hat{a} = 0.072$, decently close to zero (standard error 0.166).

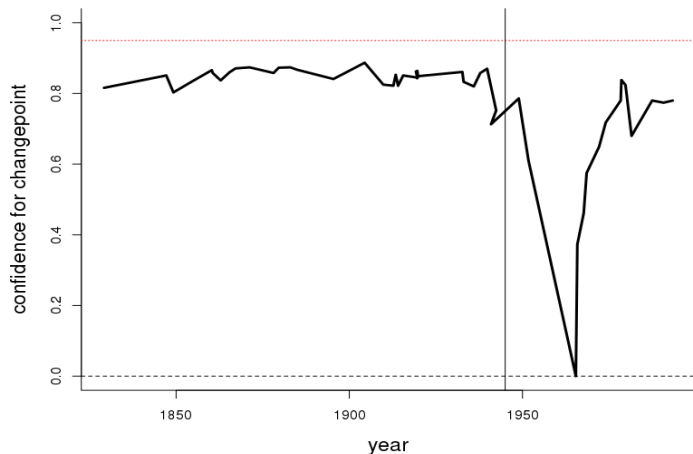
Modelling the full distribution (not only the tails)

There's a perhaps unnecessary focus on **behaviour above thresholds**. Can also model the full distribution of z (above minimum threshold 1000), or $y = \log z$ (above minimum threshold 6.908). **red** before 1945, black after 1945: **significant difference**.



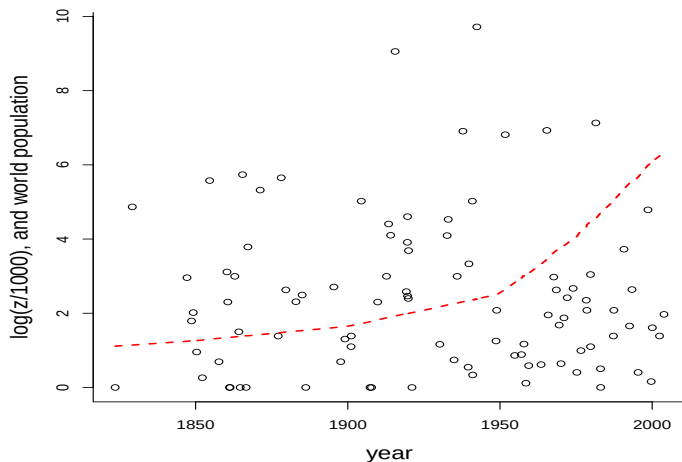
Searching for a change-point

For those $y_i = \log z_i$ above $y_0 = \log 7061$, the $v_i = y_i - y_0$ should be exponential. We can search for break point τ where $v_i \sim \text{Expo}(\theta_L)$ to the left and $v_i \sim \text{Expo}(\theta_R)$ to the right.



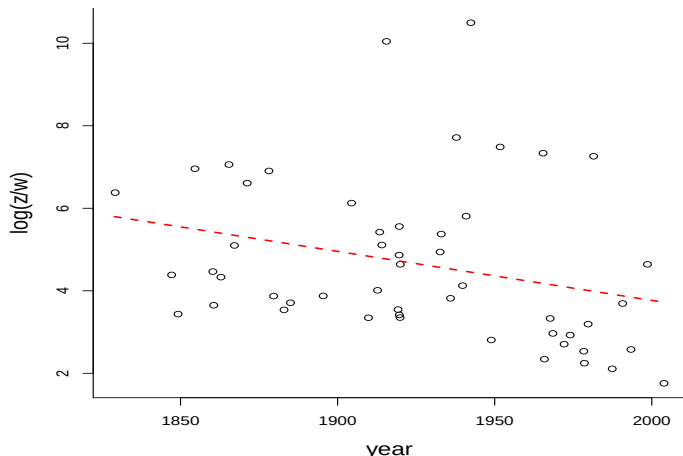
B: Scaling

Losing 1 mill is more dramatic if $w_p = 1$ billion than if $w_p = 10$ billion. This invites examining (and making models for, and predictions from) z_i/w_i rather than from z_i alone.



Power law (but negative trend) also for the scaled fatalities

By analogous arguments: $\log(z_i/w_i)$ above threshold $\text{Expo}(\theta_i)$ –
but now with **negative trend** (pvalue = 0.04).



C: Some further statistical issues

1. More careful modelling of the **full distribution** (not only of the above threshold part). Also, more accurate modelling of the tails, via extreme value statistics.
2. **Covariates**: should attempt to model distribution of z_i (or directly for $y_i = \log z_i$) via covariates $c_{i,1}, \dots, c_{i,p}$ – degree of democratisation, population sizes, degree of inter-state trading, etc. Their influence could be for **onset distribution** and/or for **size of conflict**.
3. (Focused) model selection: among candidate models (for this dataset, or for bigger conflict datasets), the best stories need the best models. Version of **FIC** (the Focused Information Criterion) can be worked out.
4. **Fine-tuned prediction**, with **confidence distribution methods** for assessing precision.
5. Meta-analysis, **combining information** across diverse sources.

(Some) references

- Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.
- Cunen, C., Hermansen, G. and Hjort, N.L. (2017). Confidence distributions for change-points and regime-shifts. *Journal of Statistical Planning and Inference*.
- Cunen, C. and Hjort, N.L. (2018). Combining information across diverse sources via [the II-CC-FF paradigm](#).
- Hjort, N.L., Holmes, C., Müller, P. and Walker, S.G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.
- Our Nature, Angels of (2011). *The Better*. Penguin Science.
- Schweder, T. and Hjort, N.L. (2016). *Confidence, Likelihood, Probability*. Cambridge University Press.