From sand grains to tomatoes: applications of size distributions

Dr Nick Fieller

Department of Probability & Statistics

University of Sheffield

Visiting

Dipartimento di Scienze Statistiche,

Università di Perugia

- **Outline:**
  - **Background:–**
    - Areas of application
      ¿ why measure particle sizes?
  - **Size definitions:–**
    - ¿ what is 'size' of a non-spherical body?
  - **Models:–**
    - Kolmogorov's law of breakage & the log-hyperbolic family
  - **Examples:–** Oronsay middens
  - **Extensions:–** Mixtures, 'marrying'/shape, number-weight relationships,…..

- **Background**
  - **Sand particles: P**article **S**ize **D**istribution (**PSD**) reflects depositional process — e.g. transported by wind or water
  - **Blood platelets:** PSD a diagnostic for coronary disease
  - **Fuel droplets:** PSD determines combustion properties
  - **Coal Dust / Flour Dust:** — Explosive??

- **Asteroids / Meteorites:** planetary (disintegration) processes
- **Diamonds:** — Value
- **Snow granules in arctic regions:–** heat transfer
- **Quartz inclusions in pottery:—** provenance studies
- **Tomatoes:—** commercial interests, quality, value, packaging problems

Sand grains
Blood Platelets
Fuel Droplets
Dust
Asteroids(?)

} Formed by degradative processes from some parent (+ subsequent selection & mixing)

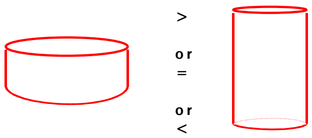
Diamonds
Crystals
Snow granules
Raindrops
Tomatoes

} Formed by **aggregative** processes + ???

¿ Implications for modelling the size distributions?

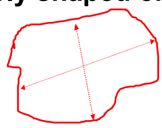
- **Measuring Techniques & Size**
  - Size is not a unique property:

Which is bigger?



(Same volumes)

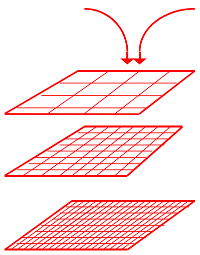
¿ Irregularly shaped objects?

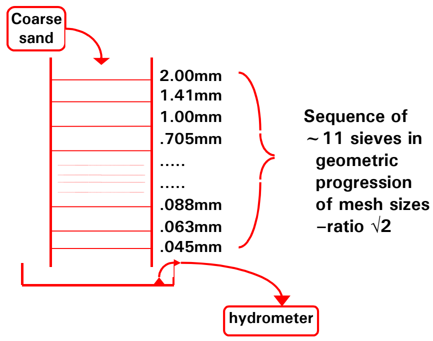


- **Geometrical Properties:**
  - Maximum / minimum diameter / volume/ surface area
- **Proxy measures:**
  - Settling velocity in viscous fluid / electrical resistance / light scattering processes

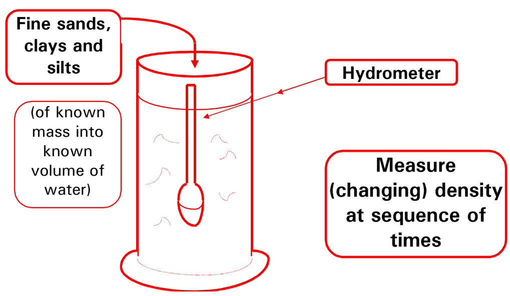
— all for individual objects — not practical for sand.

- Instead use a sequence of sieves of progressively smaller mesh size:





- Weigh material entrapped in each sieve
- **Data**: proportions [by weight] in size classes:-
  [0,0.045), [0.045, 0.063), ...,[2.0, ∞]



- As coarse particles settle out density changes.
- (2nd order) Stokes' Law allows calculation of

  proportions (by mass) **of**

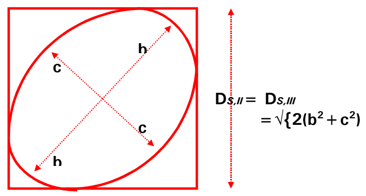
  particles > calculated radius

  (similar grouped data to sieves)

- Note: Strictly the hydrometer gives the size of a sphere which behaves in the same way as the particle — *the Stokes' Diameter*
  - for simple geometrical objects, e.g. spheroids this can be converted t a geometrical size *(details available on request)*

- Sieves measure approximately the 2nd principle diameter of particles.
  - For example, an ellipsoid with principle semi–diameters *a>b>c* can just pass through a square mesh of size *s=b[2(1+c/b)]½ ~2b*

- *s=b[2(1+c/b)]½ ~2b*



(independent of longest diameter a)

- Other methods:
  - Coulter Counter (electrical resistance (proportional to 'size')
  - Laser light scattering
  - Laser light sensors (i.e. 'projected area')
  - etc.
- **Important Conclusion:**
  - **SIZE is that property which is measured by the technique used**

(size is only unambiguous for spheres)

- This does not matter [too much] if only one technique is used to measure a set of objects
- But convenient to use models for the size distribution which allow easy transformation from one size property to another.

- e.g. Sieves can only measure as fine as 63 microns — finer fractions need to use Stokes' method.
- The two methods measure different size properties and these need to be ***'married'*** together. The relationship between them depends on the ***shape*** of the particles (in fact it **defines** their shape)
  - (see later)

- **Models**
  - Notes
    (i) most methods determine 'size' only to within a size class
    (ii) (usually) particles in a size-class are *weighed* and not *counted*

so relative proportions are available by mass and not relative frequency

- (certainly for sand, usually also for even tomatoes, not for diamonds)

- Consequence is that it is natural to model the 'mass-size' distribution and not the 'frequency-size'
  - [— absence of sample size gives difficulties in testing models]
  - We can convert between the two if the *size–weight relationship* is known.
  - Sometimes particles are both weighed and counted e.g. Diamonds, tomatoes (rarely) and this allows estimation of the size–weight relationship (see later)

- Possibilities:

1) Log–Normal

- i.e. mass–log(size) ~ Normal
  — follows from '*Law of Breakage'*
- i.e. random breakage of a distribution of parent particles results in a log-normal distribution *(Kolmogorov 1941)*
  - {Breakage →
    multiplicative on size with factor <1
    so log size → additive log(breakage)
    and the Central Limit Theorem}

- However, empirical evidence suggests that 'generally' log–normal is not sufficiently skew nor sufficiently 'peaked'
- Nevertheless, law of breakage suggests use of a model 'not too far' from log–normal
  - (i.e. family to include log-normal as a special case)
    - This is provided by:

2) Log-Hyperbolic (Barndorff-Nielsen, 1977)

- This depends on 4 parameters:

$$LH(\alpha, \beta, \delta, \mu)$$

- If $\alpha=\beta$, then as $\delta\to\infty$ $(\alpha,\beta,\delta,\mu)\to LN(\mu,\sigma^2)$
- (also, random mixing of $LN(\mu,\sigma^2)$ with a Γ-dist$^{n.}$ on σ gives LH)

However, major computational difficulties in fitting this — likelihood surface near-flat in direction of δ).

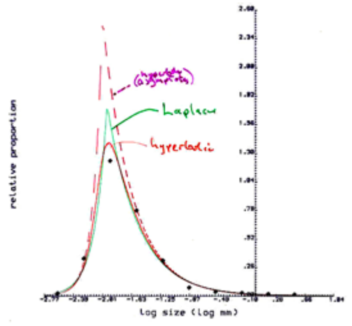
- Pragmatic alternative is

3) Log–skew–Laplace

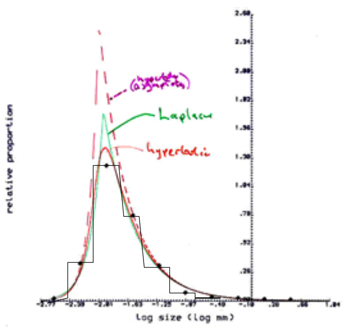
which depends on 3 parameters LsL(α,β,μ) derived from LH(α,β,δ,μ) as $\delta \to 0$

(essentially 'back–to–back' exponentials on log–size)

- I claim that this model 'works' reasonably well in a wide class of problems
  - fits observed data adequately
  - produces sensible answers to practical problems.



Beach sample EC-C15, histogram and Fitted densities

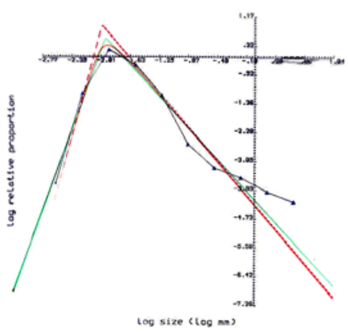


Beach sample EC-C15, histogram and Fitted densities

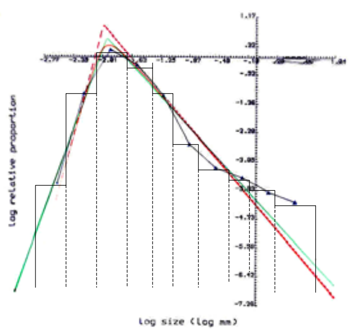
Note: When plotting the frequency curves of these distributions it is convenient to use ***log scales for both horizontal and vertical axes*** since then the distributions have easily recognisable forms:

- LN(μ,σ^2) $\to$ parabola
- LH(α,β,δ,μ) $\to$ hyperbola
- LSL(α,β,μ) $\to$ pair of straight lines

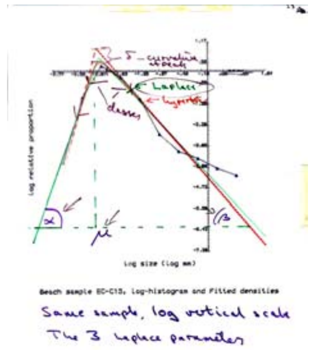
- Plotting histograms with logarithmic vertical scales is a useful exploratory analysis and diagnostic.



Beach sample EC-C15, log-histogram and Fitted densities



Beach sample EC-C15, log-histogram and Fitted densities

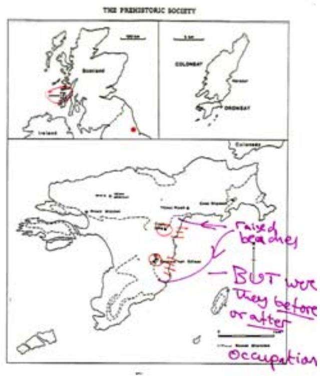


- **Example:– The Oronsay Middens**
  - Oronsay:- A small Scottish Island
- Evidence of Mesolithic occupation (~7,000b.p) from 6 well-preserved shell-middens (rubbish tips)
- Particular interest in 2 of these
  Cnoc Coig (CC) &
  Caisteal nan Gillean II (CNG II)

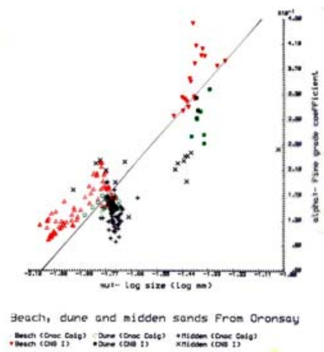
- Data: Sand samples from
  - (a) Each midden (Above, Inside, Below)
  - (b) Nearby Beach
  - (c) Nearby Dune
    - various transects from upper → lower

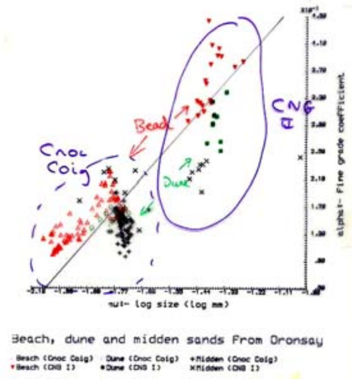
226 samples in all

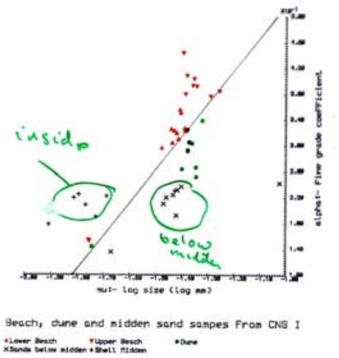
- **Were the Prehistoric middens situated on *beach* or *dune* environments ?**

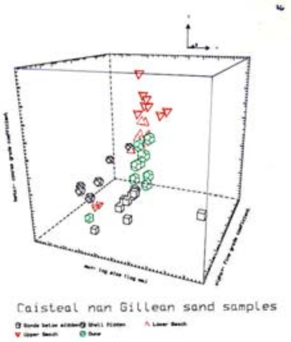


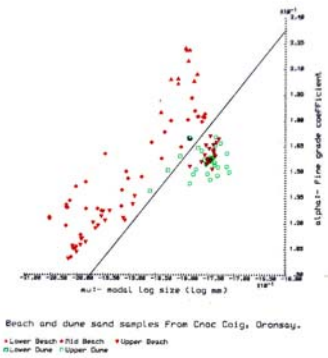
- **Problem:** Can we distinguish between beach & dune sand on the basis of particle size distributions and can we classify unknown midden sand as one or the other?
  - Preliminary analysis suggest log-skew-Laplace models LsL(α,β,μ) fit the observed data well.
- Obtain estimates of α_i, β_i and μ_i for each sample and examine scatterplots.

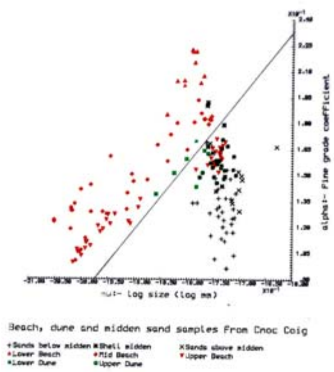












- **Conclude:**
  - Sand from the two sites can be distinguished
  - beach & dune sand can be distinguished
  - middens were situated on **dune** environments
  - beach sand was introduced to the middens (on hands & feet & shells)
  - dune environments returned after middens were formed

Note that there is no other **contemporary** evidence available to answer this question

- **Tomatoes:–**
  size–weight relationships
  - In the *environmentally challenged* UK tomatoes are grown under glass
  - Tomatoes are a major commercial crop
  - Tomatoes are sold by weight packaged by volume
  - Tomatoes are not individually weighed and measured but only in aggregate by grouped size class

- Relationship between weight & size and how this varies with growing season and conditions are of major commercial interest

  Unit of sale:— weight

  Unit of packaging:— size

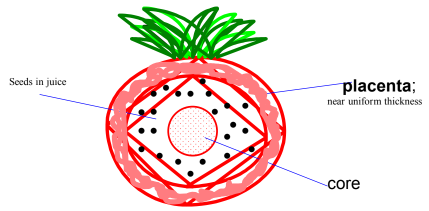
- Expect:

  weight $\propto$ **[size]** ^{b}

  for some b —

  perhaps $b \approx 3$ if size= diameter

But tomatoes have non-uniform density



Thickness of placenta is near constant for different sized tomatoes

- **So**

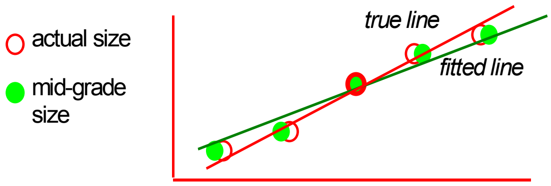
  weight of placenta $\propto$ **[size]** ^{2}

  weight of seeds & core $\propto$ **[size]** ^{3}

  So expect that **weight** $\propto$ **[size]** ^{b}

  with **2< b < 3**

- **Why not regression?**
  - If weight $\propto$ [size] ^{b} then we have
    log(weight)= a + b log(size)
- But
  - typically only 5 or 6 non-zero size classes for any one sample
  - systematic ***bias*** since sizes are not uniformly distributed over size classes



mid-point of class interval **underestimates** mean class size

- if class intervals < overall mean and
- **overestimates** if class intervals > overall mean

- regression estimate of *b* is **biased downwards**

- **Alternative Approach**
  - If we have both weights & numbers in each size class then we can fit LsL models separately to weights and numbers
  - Relationship between weight-size and number-size distributions depends upon the allometric relationship between weight and size

- **Relationships**
  - If $L(\alpha_w, \beta_w, \mu_w)$ is the weight-size distribution & $L(\alpha_N, \beta_N, \mu_N)$ is the number-size distribution

    and if **weight ∝ [size]** $^{\boldsymbol{b}}$

    then $\alpha_w^{-1} = \alpha_N^{-1} + \boldsymbol{b}$

    $\beta_w^{-1} = \beta_N^{-1} - \boldsymbol{b}$
  - So, we can estimate ***b*** by fitting Log skew Laplace models to both weight- and number-size distributions from the same sample

- **Data**
  - From a controlled experiment at HRI Wellesbourne
  - Tomatoes graded, weighed and counted from a 3×split-split-plot design
  - Factors include
    - Season
    - Growing conditions (North/South facing etc)

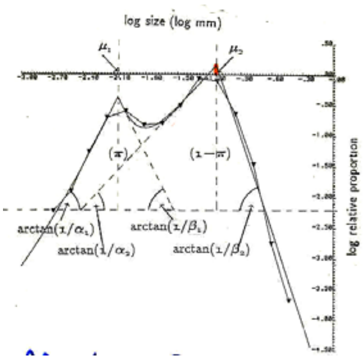
- **Results**
  - log skew Laplace models fit appreciably better than log Normal (and log hyperbolic)
  - 'best overall guess' ***b*= 2.2 :—**
    weight ∝[size] $^{2.2}$
  - values of ***b*** are generally ***lower***
    - early in season
    - south side of glass house
  - little variation in ***b*** with other growing factors (truss thinning, side shoot removal)

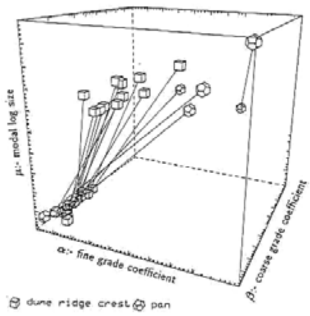
- **Alternative approaches**
  - Fit models simultaneously to number-size and weight size distributions, estimating ***b*** directly as an additional parameter
  - Bias corrected regression
    - i.e. estimated interval mean from fitted distributions

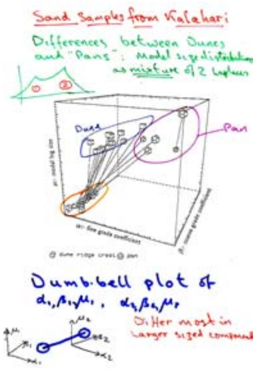
- **Extensions: (1) Covariate models**
  - Allow α, β & μ to depend on factors or continuous covariates: —

    gives direct assessment of factors on size distribution.
    - Implemented by Andy Lynch for discrete factors and continuous covariates
      - (now at University of Newcastle)
  - Further work on dependent samples (i.e. time series)

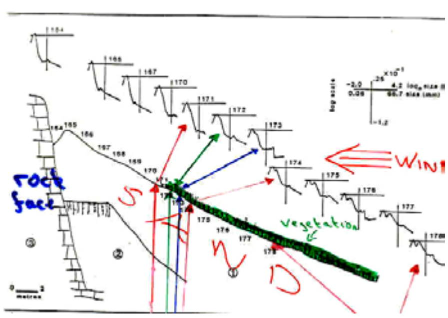
- **Extensions (2): Mixtures**
  - Many samples from desserts show bimodality:
    - (two competing depositional processes)
- Model by mixture distribution:
  - $\pi_1 LsL(\alpha_1,\beta_1,\mu_1)+(1-\pi_2)LsL(\alpha_2,\beta_2,\mu_2)$
  - Gives 7 parameters (1 mixing, 3 pairs of α,β and μ)
  - (6 parameters can be displayed in 3d dumbbell-plots)
    - (example: Kalahari dessert pan sediments)

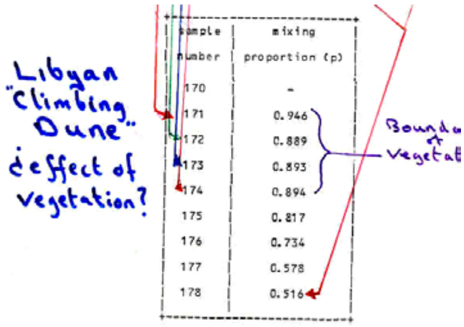




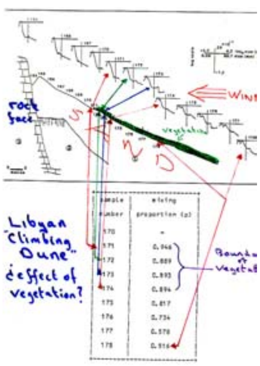


- **Example: Libyan climbing dune:–**
  - Bimodal distributions at lower vegetated end of slope,
  - Unimodal distributions at upper non-vegetated end
- Where does influence of vegetation stop?





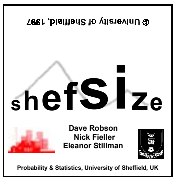
sample number	mixing proportion (p)
170	-
171	0.946
172	0.889
173	0.893
174	0.894
175	0.817
176	0.734
177	0.578
178	0.516



- **Computing Aspects**
  - Models were fitted using shefSize which fits
    - Log Normal
    - Log skew Laplace
    - Log hyperbolic
    - Log skew Laplace mixtures

  models to grouped data

  Available from
  http://www.shef.ac.uk/nickfieller



- Acknowledgments
  - Eleanor Stillman (Sheffield)
  - Walter Olbricht (Bayreuth)
  - Andy Lynch (Newcastle)

- References

  - Fieller, Flenley & Olbricht (1992), *Applied Statistics*, ***41,*** 127-146.
  - Fieller, Gilbertson & Olbricht, *Nature*, **311,** 648-651