Chaos, Complexity, and Inference (36-462)

Lecture 16

Cosma Shalizi

6 March 2008

Comparing Heavy-Tailed Distributions

Further reading: Clauset *et al.* (2007)

“The Fundamental Theorem of Statistics”

per Pitman (1979)

Theorem (Glivenko-Cantelli)

Let $X_1, X_2, \ldots$ be IID with CDF F. Let $\widehat{F}_n$ be their empirical CDF from n samples.

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow[n \to \infty]{} 0$$

EXERCISE: Who was Glivenko? Who was Cantelli?
Notice that this is a KS-distance:

$$\lim_{n \to \infty} d_{KS}(\widehat{F}_n, F) \to 0$$

Goodness-of-Fit Testing

The logic: If F is the true CDF, then

$$d_{KS}(\widehat{F}_n, F) \rightarrow 0$$

but if the true CDF is $F' \neq F$, then

$$d_{KS}(\widehat{F}_n, F) \rightarrow d_{KS}(F', F) > 0$$

The data **fits** the model F when d_{KS} is small, but not if it's large
We never expect d_{KS} to be zero, even if our model is exactly right
need to know *how big* we should expect d_{KS} to be, if our model is right
p-value: probability of getting a discrepancy *at least as big* as the one we observe in the data, *if* our model is right
Lack of fit if p-value is very small

Getting p-values means getting the distribution of d_{KS}, under the assumption the model is right

For the *true* F, $d_{KS}(\widehat{F}_n, F)$ has a known distribution, which does not depend on F when n is large

$$\Pr\left(\sqrt{n} d_{KS}(\widehat{F}_n, F) \leq x\right) \rightarrow 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}$$

So: we can calculate p-values, *if* F is fixed

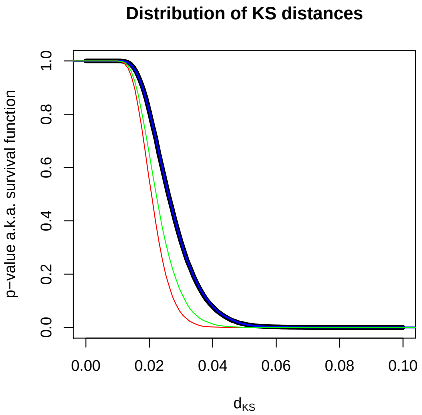
If we do *not* fix F but estimate it from the data, we cannot use the usual formula to calculate p-values

of course our estimated F is close to the data, we *made* it that way

Illustration of these points:

1. Draw $X_1, X_2, \ldots X_{1000}$ from $\mathcal{N}(0,1)$
2. Calculate d_{KS} for X vs. $\mathcal{N}(0,1)$ and $\mathcal{N}(\overline{X}, s^2_X)$
3. Repeat (1) and (2) 10,000 times to get two sampling distributions
4. Draw $Y_1, Y_2, \ldots Y_{1000}$ from $\mathrm{Exp}(1)$
5. Calculate d_{KS} for Y vs. $\mathrm{Exp}(1)$ and $\mathrm{Exp}(1/\overline{X})$
6. Repeat (4) and (5) 10,000 times to get two more sampling distributions

Results on next slide — see `16.R` on website for code
N.B.: for a given value of d_{KS}, the true p-value is smaller with estimation than without it; ignoring estimation makes you think the fit is better than it really is!



black = fixed Gaussian, red = estimated Gaussian, blue = fixed exponential, green = estimated exponential

Finding Goodness-of-Fit p-values Through Simulation

Wanted: The sampling distribution of d_{KS} when F is estimated
Problem: The probability theory is *very hard*
Solution:

1. Estimate model F_{est} from real data; calculate real $d_{KS} = d^*$
2. Use F_{est} to generate simulated data
3. Treat simulated data as if real, estimate model on it and calculate d_{KS}
4. Repeat steps (2) and (3) many times to get sampling distribution of d_{KS}
5. p-value is fraction of d_{KS} values $\geq d^*$

To get p-value accurate to $\pm\epsilon$, use $\approx \frac{1}{4\epsilon^2}$ simulations ($\Leftarrow$ binomial)

Application to Fit of Power-Law Tails

Given: n data points $x_1, \ldots x_n$

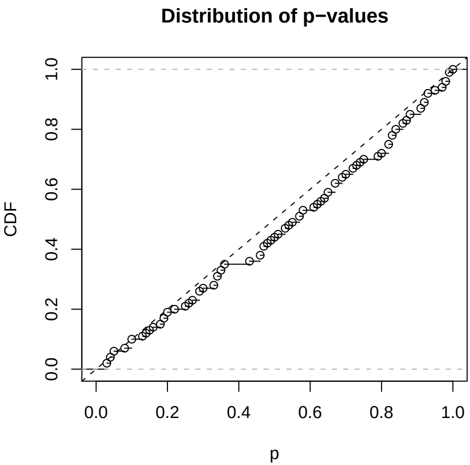
1. Estimate α and x_{min}; $n_{\mathrm{tail}} =$ # of data points $\geq x_{\mathrm{min}}$
2. Calculate d_{KS} for data and best-fit power law $= d^*$
3. Draw n random values $b_1, \ldots b_n$ as follows:
    1. with probability n_{tail}/n, draw from power-law
    2. otherwise, pick one of the $x_i < x_{\mathrm{min}}$ uniformly
4. Estimate α and x_{min} for the simulation, calculate its d_{KS}
5. Repeat many times to get distribution of d_{KS} values
6. p-value = fraction of simulations where $d \geq d^*$

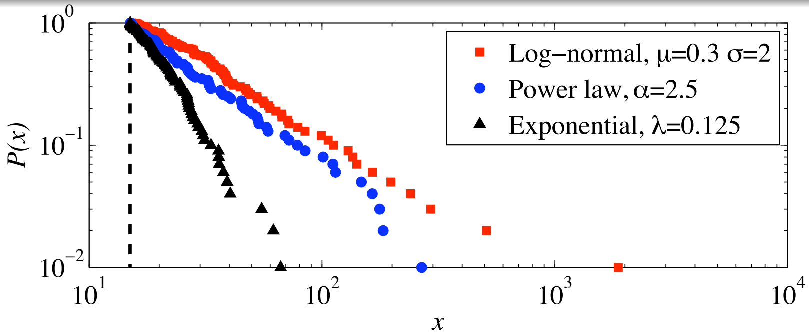
Coded as `pareto.tail.ks.test` in R file for this lecture

If the model is right and *p*-values are properly calculated, they should be $\sim \mathrm{Uniform}(0, 1)$
CDF of uniform distribution is the diagonal
Using `rpareto.tail` (random variables from a distribution with a power-law tail) and `pareto.tail.ks.test`

```
> sample.of.p.values <- replicate(100,
  pareto.tail.ks.test(rpareto.tail(1e2,1,2.5,0.5),100))
> plot(ecdf(sample.of.p.values),xlim=c(0,1),
  main="Distribution of p-values")
> abline(0,1,lty=2)
```

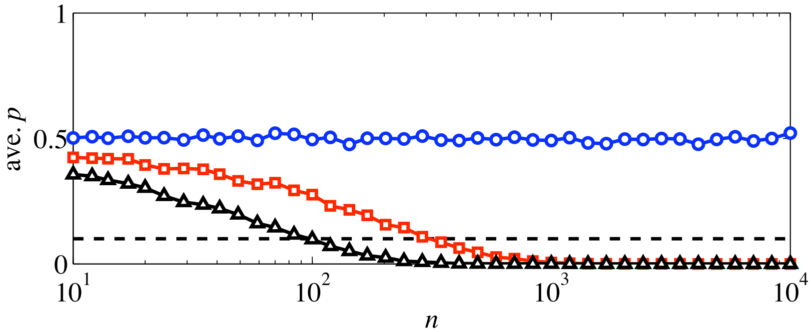
samples of size 100, 100 simulations per *p*-value, 100 replications — all comparatively small, to save time


Distribution of p−values
CDF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
p

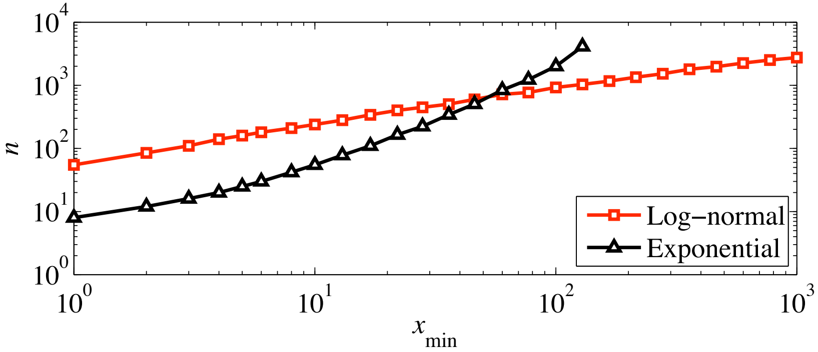


Empirical CDFs for samples of size 100 from specified distributions, with $x_{\min} = 15$

This and next two figures from Clauset *et al.* (2007)



Average *p*-values according to our procedure



Average number of samples required to make the p value < 0.1

Cautions about Goodness of Fit Tests

or: “Does this data make me look fat?”

“Your distribution doesn’t fit” But where, and enough to matter? Looking at **relative distribution** (next section) is a way to start answering that

“Your distribution fits” Would your test *notice* if it didn’t? It’s only *evidence* if it would

Remember problems with R^2 from last time

Look at previous two slides

Need to consider **power** and **severity** — much more about severity after break

Relative Distributions

After Handcock and Morris (1998, 1999)

Want to compare two distributions, not just mean/variance etc.

Specifically: $y_1, \ldots y_n$ are **comparison sample**, have either a **reference distribution** or a **reference sample** $x_1, \ldots x_m$, CDF $= F_0$

Construct **relative data**

$$r_i = F_0(y_i)$$

relative CDF:

$$G(r) = F(F_0^{-1}(r))$$

relative density

$$g(r) = \frac{f(F_0^{-1}(r))}{f_0(F_0^{-1}(r))}$$

Why do this?

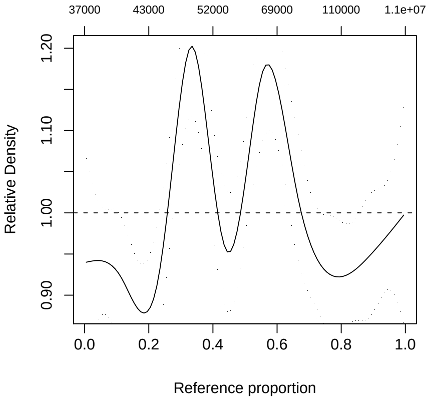
- Relative data are uniform if and only if distributions are the same
- Invariant under any monotone transformation of the data (multiplication, taking logs, etc.) *so* no loss of information except about absolute values
- Can control for covariates much more flexibly than in regression See Handcock and Morris (1999)
- $g(r) > 1 \Rightarrow$ comparison data is more likely to be close to $F_0^{-1}(r)$ than reference — tells us *where* and *how* the distributions differ

Can estimate $G(r)$ by empirical CDF of r_i
Can estimate $g(r)$ by non-parametric density estimation on r_i
`R` package: `reldist`, from CRAN

Relative Distributions with Power Laws

1. Estimate power law distribution from data
2. Use this as the reference distribution
3. Relative density *should* shoot up at right (finite maximum)

```
> http.mle <- pareto.fit(http,"find")
> F0 <- function(x) {
  ppareto(x,http.mle$xmin,http.mle$exponent) }
> F0inv <- function(p) {
  qpareto(p,http.mle$xmin,http.mle$exponent) }
> reldist(y=F0(http[http>=http.mle$xmin]),smooth=-1)
> top.ticks = c(0,0.2,0.4,0.6,0.8,F0(max(http)))
> top.tick.values = signif(F0inv(top.ticks),2)
> axis(side=3,at=top.ticks,labels=top.tick.values,
  cex.axis=0.75)
```



Relative distribution of HTTP file sizes (in kb) vs. best-fit Pareto; big spikes around $\approx 48\mathrm{kb}$ and $\approx 64\mathrm{kb}$

Likelihood-Ratio Tests for Model Selection

After Vuong (1989)

Likelihood ratio of two models θ, ψ

$$\frac{p_\psi(x_1, \ldots x_n)}{p_\theta(x_1, \ldots x_n)}$$

often easier to use **log likelihood ratio**

$$\mathcal{R}(\psi, \theta) = \log p_\psi(x_1, \ldots x_n) - \log p_\theta(x_1, \ldots x_n)$$

$\mathcal{R}(\psi, \theta) > 0$ means: the data were *more likely* under ψ than under θ

Likelihood ratio test: chose between models using $\mathcal{R}$

Distribution of Likelihood Ratios: Fixed Models

Assume $X_1, X_2, \ldots$ all IID, with true distribution μ
Fix θ and ψ; what is distribution of $\mathcal{R}(\psi, \theta)$?

$$\begin{aligned}
\mathcal{R}(\psi, \theta) &= \log p_\psi(x_1, \ldots x_n) - \log p_\theta(x_1, \ldots x_n) \\
&= \sum_{i=1}^{n} \log p_\psi(x_i) - \sum_{i=1}^{n} \log p_\theta(x_i) \\
&= \sum_{i=1}^{n} \log \frac{p_\psi(x_i)}{p_\theta(x_i)}
\end{aligned}$$

so $\mathcal{R}(\psi, \theta)$ is a sum of IID terms

Use LLN:

$$\begin{aligned} \frac{1}{n}\mathcal{R}(\psi, \theta) &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\psi(x_i)}{p_\theta(x_i)} \\ &\rightarrow \mathbf{E}_\mu\left[\log \frac{p_\psi(X)}{p_\theta(X)}\right] \\ &= D(\mu \| \theta) - D(\mu \| \psi) \end{aligned}$$

$\mathcal{R}(\psi, \theta) > 0$ tends to mean: ψ is closer (in relative entropy) to μ than θ is

Use CLT:

$$\frac{1}{\sqrt{n}}\mathcal{R}(\psi,\theta) \rightsquigarrow \mathcal{N}(\sqrt{n}(D(\mu\|\theta) - D(\mu\|\psi)), \omega^2_{\psi,\theta})$$

where

$$\omega^2_{\psi,\theta} = \mathrm{Var}\left[\log\frac{p_\psi(X)}{p_\theta(X)}\right]$$

so if the models are equally good, we get a mean-zero Gaussian
but if one is better $\mathcal{R}(\psi,\theta) \to \pm\infty$, depending

Distribution of $\mathcal{R}$ with Estimated Models

two classes of models Ψ, Θ; $\hat{\psi}, \hat{\theta}$ = ML *estimated* models
$\hat{\psi} \rightarrow \psi^*$, $\hat{\theta} \rightarrow \theta^*$: converging to **pseudo-truth**; $\psi^* \neq \theta^*$
some regularity assumptions
then everything works out as if no estimation

$$\begin{aligned}
\frac{1}{\sqrt{n}}\mathcal{R}(\hat{\psi}, \hat{\theta}) & \rightsquigarrow & \mathcal{N}(\sqrt{n}(D(\mu \| \theta^*) - D(\mu \| \psi^*)), \omega^2_{\psi^*, \theta^*}) \\
\frac{1}{n}\mathcal{R}(\hat{\psi}, \hat{\theta}) & \rightarrow & D(\mu \| \theta^*) - D(\mu \| \psi^*) \\
\widehat{\omega}^2 \equiv \mathrm{Var}_{\mathrm{sample}}\left[\log \frac{p_\psi(X)}{p_\theta(X)}\right] & \rightarrow & \omega^2_{\psi^*, \theta^*}
\end{aligned}$$

Vuong's Test for Non-Nested Model Classes

Assume all conditions from before
If the two models are really equally close to the truth,

$$\frac{\mathcal{R}}{\sqrt{n\widehat{\omega}^2}} \rightsquigarrow \mathcal{N}(0, 1)$$

but if one is better, normalized log likelihood ratio goes to $\pm\infty$, telling you which is better

Note: do not need to adjust for which model has more parameters

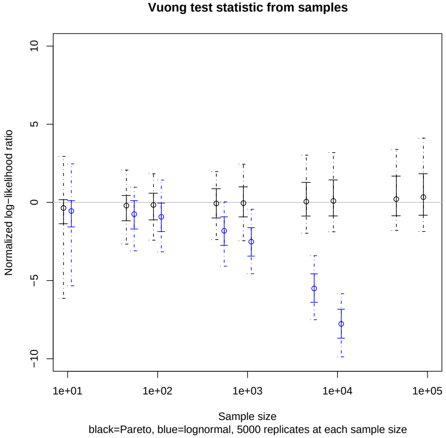
can include adjustment (AIC, BIC, ...) if it is $o(n)$ without changing asymptotics

Note: does not assume that either Ψ or Θ contains the truth

Note: does assume that $\psi^* \neq \theta^*$

Procedure

1. Estimate $\hat{\psi} \in \Psi$ and $\hat{\theta} \in \Theta$
2. Calculate $\rho_i \equiv \log p_{\hat{\psi}}(x_i)/p_{\hat{\theta}}(x_i)$
3. $\mathcal{R} = \sum \rho_i$, $\hat{\omega}^2 = \mathrm{Var}_{\mathrm{sample}}\left[\rho_i\right]$
4. $V = \mathcal{R}/\sqrt{n\hat{\omega}^2}$
5. Go with Ψ if $V \gg 0$, Θ if $V \ll 0$, no choice if $|V| \approx 0$



smallest V, 5th percentile, median, 95th percentile, max

Nested Hypotheses

$\Theta \subset \Psi$ means $\mathcal{R} \geq 0$, but now when they are equally good
$\psi^* = \theta^*$, and $\omega^2 = 0$
Can't use that argument
Can show that

$$2\mathcal{R} \rightsquigarrow \chi^2_{\dim\Psi - \dim\Theta}$$

If μ (the true distribution) $= \theta^*$ this is a classic result (Wilks, 1938), but Vuong shows it holds even under mis-specification

The Flight of the Albatross

Edwards *et al.* (2007): an exemplary paper in several senses:

- what it does
- the way it does it
- how it came about

Requires some background from more advanced probability first

Beyond the Normal Central Limit Theorem

Ordinary CLT: IID variables with finite variance $\Rightarrow$ mean is Gaussian
Reason: Gaussian is *stable* under averaging (sum of independent Gaussians is again Gaussian)
Not IID: may or may not be Gaussian (rate of mixing)
IID, infinite variance: not Gaussian, but must be another stable distribution
LÉVY (1930s): Characterization of the stable distributions
Obvious alternatives to Gaussian in CLT have power-law tails
Schroeder (1991); Embrechts and Maejima (2002); Gnedenko and Kolmogorov (1954)

Lévy Flights

Lévy flights: random walks where the distribution of step sizes has power-law tails

Gaussian random walks produce fractal patterns, but region covered grow slowly and fairly steadily (diffusion)

Lévy flights produce sparser, more irregular fractals, big leaps between clusters (anomalous diffusion)

Lévy flights are at least *good approximations* to lots of diffusion processes

possibly with some truncation to keep variances finite

Jamie Watts at British Antarctic Survey

Diomedea exulans: long-range marine predator, skims over water to scoop up fish, squid, etc.
prey are patchy so it travels *very long* distances
how long?

Experiment (1992): attach monitor to albatrosses' legs, record when the leg is in the water (to the hour); gives indication of flight length (dry == flying)

Viswanathan *et al.* (1996):

did log-log regression on binned histogram of flight times

saw straight line

concluded: power law, therefore Lévy flight

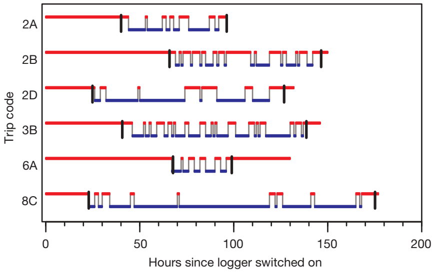
Much subsequent work on (i) replicating this kind of analysis for other animals, people, etc. and (ii) explaining why Lévy flights are a Good Thing when looking for food

... a dozen years later: new data!

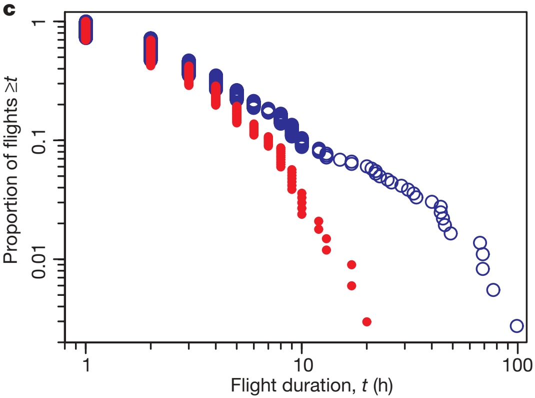
Better timer on the monitor + GPS 1/hr to tell when the birds came back to their island

Longest new trip $<$ 15 hr, at least 6% of old trips supposedly longer

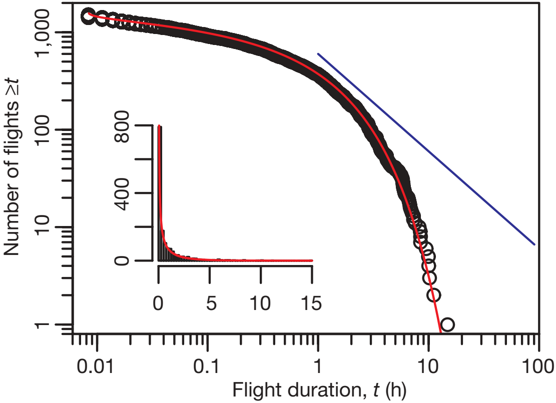
Turned out there were satellite location measurements for some of the old trips



red: dry; blue: wet; black depart from/return to island i.e., some of the really long "flights" were just spent sitting around on the island



blue: uncorrected CDF of flight times; red: corrected



CDF from newer data set + truncated power-law/gamma distribution (red) + pure power law (blue)

AIC Approach to Comparison

AIC = **Akaike Information Criterion** (Akaike, 1973)
Model class Θ, estimated with maximum likelihood ($\hat{\theta}$)

$$AIC(\Theta) = \mathcal{L}(\hat{\theta}) - d$$

d = number of parameters estimated for θ

Supposed to be unbiased estimate of *expected* likelihood on *new* data from same source

Can give relative weights for different models, prefer one with higher AIC

Could also use BIC — near-theological debates about which is better (or others)

"Practice is the sole criterion of truth": The best criterion is the one which most efficiently and reliably selects the right model; we'll come back to this

Here AIC *strongly* favors gamma distribution over power law

Conclusions

1. Apparent result was really due to an artifact (wrong flight times) and a weak analysis method (log-log regression)
2. Corrected and properly analyzed, data support gamma distribution much more strongly than the power law
3. Now the theoretical to task is to explain why, if Lévy flights are so wonderful, albatrosses (and bumblebees, and deer, and ...) do *not* take them

Morals

1. THE REAL FOUNDATIONS OF STATISTICS: You must understand your data intimately *before* you start to do statistics
2. FESS UP: The problem with the flight times was discovered by the same collaboration as made the original claims; this shows class
3. ZOMBIES: Some ideas are like zombies: they come back from the dead and they eat your brains. There are signs that Lévy flight foraging may be undead in this way (Sims *et al.*, 2008)

Akaike, Hirotugu (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In *Proceedings of the Scond International Symposium on Information Theory* (B. N. Petrov and F. Caski, eds.), pp. 267–281. Budapest: Akademiai Kiado. Reprinted in (Akaike, 1998, pp. 199–213).

— (1998). *Selected Papers of Hirotugu Akaike*. Berlin: Springer-Verlag. Edited by Emanuel Parzen, Kunio Tanabe and Genshiro Kitagawa.

Clauset, Aaron, Cosma Rohilla Shalizi and M. E. J. Newman (2007). "Power-law distributions in empirical data." *SIAM Review*, **submitted**. URL `http://arxiv.org/abs/0706.1062`.

Edwards, Andrew M., Richard A. Phillips, Nicholas W. Watkins, Mervyn P. Freeman, Eugene J. Murphy, Vsevolod Afanasyev, Sergey V. Buldyrev, M. G. E. da Luz, E. P. Raposo,

H. Eugene Stanley and Gandhimohan M. Viswanathan (2007). "Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer." *Nature*, **449**: 1044–1048. URL `http://polymer.bu.edu/hes/articles/epwfmabdrsgv07.pdf`. doi:10.1038/nature06199.

Embrechts, Paul and Makoto Maejima (2002). *Selfsimilar processes*. Princeton, New Jersey: Princeton University Press.

Gnedenko, B. V. and A. N. Kolmogorov (1954). *Limit Distributions for Sums of Independent Random Variables*. Cambridge, Massachusetts: Addison-Wesley. Translated from the Russian and annotated by K. L. Chung, with an Appendix by J. L. Doob.

Handcock, Mark S. and Martina Morris (1998). "Relative

Distribution Methods." *Sociological Methodology*, **28**: 53–97. URL http://links.jstor.org/sici?sici=0081-1750%281998%2928%3C53%3ARDM%3E2.0.CO%3B2-Z.

— (1999). *Relative Distribution Methods in the Social Sciences*. Berlin: Springer-Verlag.

Pitman, E. J. G. (1979). *Some Basic Theory for Statistical Inference*. London: Chapman and Hall.

Schroeder, Manfred (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. San Francisco: W. H. Freeman.

Sims, David W., Emily J. Southall, Nicolas E. Humphries, Graeme C. Hays, Corey J. A. Bradshaw, Jonathan W. Pitchford, Alex James, Mohammed Z. Ahmed, Andrew S. Brierley, Mark A. Hindell, David Morritt, Michael K. Musyl, David Righton, Emily L. C. Shepard, Victoria J. Wearmouth, Rory P. Wilson, Matthew J. Witt and Julian D. Metcalfe (2008). "Scaling laws of marine predator search behaviour." *Nature*, **451**: 1098–1102. doi:10.1038/nature06518.

Viswanathan, G. M., V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince and H. E. Stanley (1996). "Lévy flight search patterns of wandering albatrosses." *Nature*, **381**: 413–415. URL `http://polymer.bu.edu/hes/articles/vabmps96.pdf`.

Vuong, Quang H. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica*, **57**: 307–333. URL `http://links.jstor.org/sici?sici=0012-9682%28198903%2957%3A2%3C307%3ALRTFMS%3E2.0.CO%3B2-J`.

Wilks, S. S. (1938). "The Large Sample Distribution of the

Likelihood Ratio for Testing Composite Hypotheses." *Annals of Mathematical Statistics*, **9**: 60–62. URL http://links.jstor.org/sici?sici=0003-4851%28193803%299%3A1%3C60%3ATLDOTL%3E2.0.CO%3B2-6.