Long Tail Analysis 1

Scott Hendrickson

- **Assumption : Showing that a given distribution is the "correct" distribution is not a reasonable goal. Why look at alternative distribution functions? To understand the dynamics of rare and unexpected behavior.**

Power Law Model

Many "Long tail" descriptions refer to the "Power Law" representation of a long-tail distribution. Assuming the power law, the probability distribution function (pdf) is f(x)=a x ^k, where x is the rank of the object and k and a are the parameters of the model. The cumulative distribution function (cdf) is the integral of the pdf.

Books sales in the US in 2004 provide a good example case (data taken from Kalvevi Kilkki's "A practical model for analyzing long tails"),

BooksTable =

Rank	Cumlative Volume (Copies per year)	Cumlative Share (%)
10	17 396 510	2.6
32	31 194 809	4.6
96	53 447 300	8
420	100 379 331	15.1
1187	152 238 166	22.9
24 234	432 238 757	65
91 242	581 332 371	87.4
294 180	650 880 870	97.8
1 242 185	665 227 287	100

;

```
data = Part[BooksTable, 2 ;;, {1, 3}];
xvals = Part[data, All, 1];
smoothXvals = Map[#^3 &, Range[3, (Max[xvals])^(1 / 3), Round[(Max[xvals])^(1 / 3) / 100, 1]]];
```

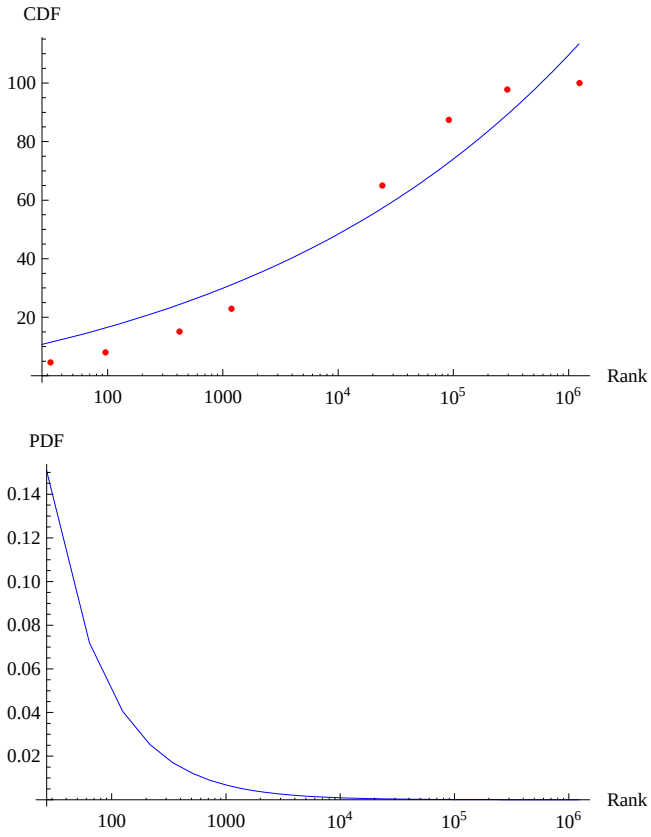
The power law model :

```
PowerPDF[x_] := a x^ k
PowerCDF[x_] := Integrate[PowerPDF[y], {y, 1, x}]
PowerSol = FindFit[data, PowerCDF[z], {a, k}, z]
```

```
{a → 2.54918, k → -0.858148}
```

This is the solution to fitting the data to the power law distribution and gives the parameters a and k of the model.

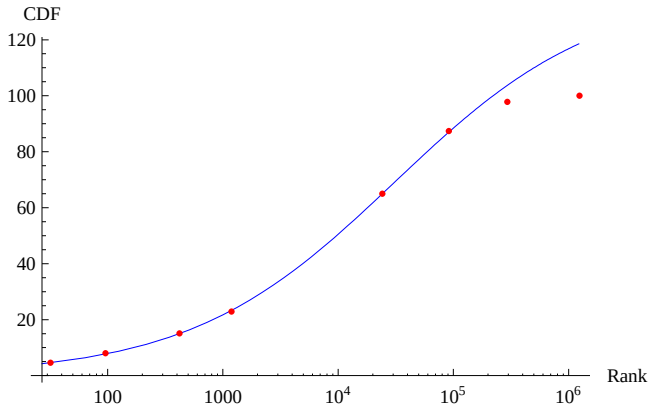
```
c = Table[{x, PowerCDF[x]}, {x, smoothXvals}] /. PowerSol;
p = Table[{x, PowerPDF[x]}, {x, smoothXvals}] /. PowerSol;
Show[{ListLogLinearPlot[c, PlotStyle → Blue, AxesLabel → {"Rank", "CDF"}, Joined → True],
  ListLogLinearPlot[data, PlotStyle → Red, PlotRange → All]}]
ListLogLinearPlot[p, PlotStyle → Blue, AxesLabel → {"Rank", "PDF"},
 Joined → True, PlotRange → All]
```

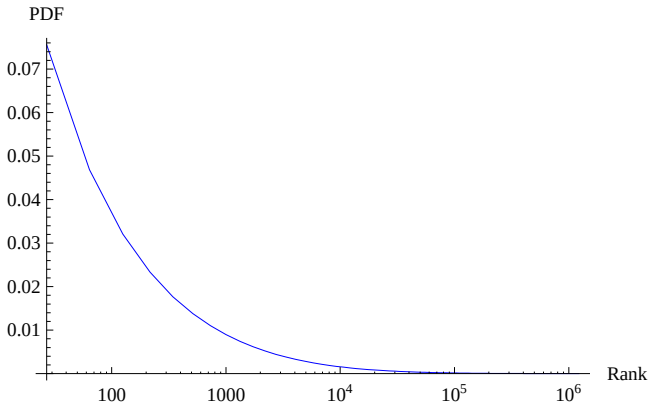


Kalevi Kilkki' s Three - parameter Model

Referring to Kalevi Kilkki' s long tail paper ("A practical model for analyzing long tails")--K proposes an alternate form of long tail function. This function of three parameters provides an anchor point at the (rank) middle of the distribution and adjusts the distribution anti-symmetrically about this anchor point. Here's an example (using K's fit for the book data shown above):

```
ltCDF[x_] := 100 b / ((N50 / x)^a + 1)
ltPDF[x_] := ltCDF[x] - ltCDF[x - 1]
ltc = Table[{y, ltCDF[y]}, {y, smoothXvals}] /. {b → 1.38, a → 0.49, N50 → 30 714};
ltd = Table[{y, ltPDF[y]}, {y, smoothXvals}] /. {b → 1.38, a → 0.49, N50 → 30 714};
Show[{ListLogLinearPlot[ltc, PlotStyle → Blue, AxesLabel → {"Rank", "CDF"}, Joined → True],
  ListLogLinearPlot[data, PlotStyle → Red, PlotRange → All]}]
ListLogLinearPlot[ltd, PlotStyle → Blue, AxesLabel → {"Rank", "PDF"},
 Joined → True, PlotRange → All]
```





Kilkki is trying to describe more precisely the variations between the simple power law from data. In the appendix, he compares the power law and his 3-parameters model show that he can fit the head data more consistently. This leads K to write, "...it is apparent that a power-law function is not suitable for modeling long tails."

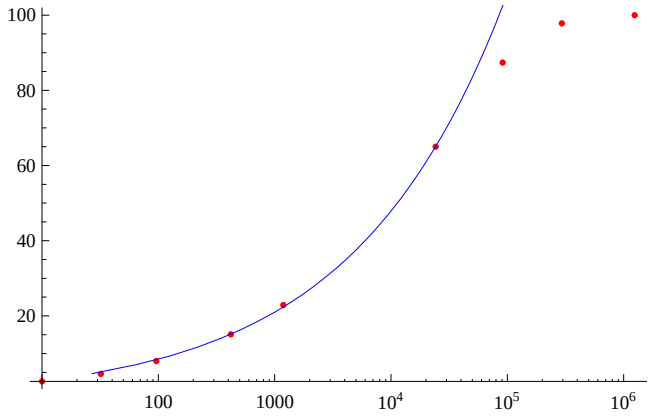
Admittedly, the variations between model and data for the power law seem fairly consistent in the data from V's paper. However, the head and tail deviations may have different explanations.

An Alternative Approach (that conserves the power law)

What if we fit to the head, omitting the tail where (for "understood" economic reasons, see "The Long Tail" by Chris Anderson) the apparent turn of the data from theory appears significant. Here is a fit of the power law to the head of the book data,

```
headData = data[[ ;; 6]];
headxvals = Part[headData, All, 1];
headPowerSol = FindFit[headData, PowerCDF[z], {a, k}, z]
chead = Table[{x, PowerCDF[x]}, {x, smoothXvals}] /. headPowerSol;
Show[{ListLogLinearPlot[data, PlotStyle → Red, PlotRange → All],
  ListLogLinearPlot[chead, PlotStyle → Blue, AxesLabel → {"Rank", "CDF"}, Joined → True]}]
```

```
{a → 0.779943, k → -0.667259}
```



What alternate explanations do we have for tail deviations? IOW, what is the justification for ignoring the tail or considering it a secondary characteristic?

- Inadequate population size to make tail dynamics subject to the same processes that build the head? These distributions are often referred to as "scalable" or "scale-free". But we know that all "scale-free" systems in nature have "boundaries" (the fractal nature of the shore line does not scale out to the whole earth, which is remarkably smooth at the largest scale) where scale-free dynamics aren't applicable modeling techniques.

- Economics-driven decisions are made farther up the tail (all the way?) than we tend to think. It is often stated that "cost of acquisition"--find and purchase, for books are responsible for the lower frequency of books in the end of the tail. The business proposition is stated as find a way to lower the buyer's costs of finding and purchasing obscure titles, and you have a new "long-tail" business (e.g. amazon.com, netflix.com).

- While a some studied dynamical processes support a power law, there doesn't seem to be any clear motivation for V's refinements. In fact, if assigning deviations to the tail is appropriate as proposed above, then V's formula "hides" latent demand by making the turn appear at higher ranks.

Future Work and Projects for Investigating Claims

This suggests a couple of projects :

PROJECT 1: develop network dynamical model that shows power law, then add economic model elements/parameters that produce observed tail deviation

PROJECT 2: How to characterize the trail and to express the concept of "latent demand" in a way useful to designing a business?

So what benefit do we get from V' s 3 - parameter fit? And, if little or none, does it make sense to "assign" all the deviation to the tail? questions :

```
DateString[]
```

```
Tue 31 Jul 2007 10:22:07
```