Lab Class 2:Data, and “tricky distributions”

Nick Watkins

May 16, 2012

1 Today’s session

The first part of today’s session is looking at some examples of datasets with the tools discussed yesterday.

The second part will be given by Cosma, and will involve the “tricky” estimation problem of distributions with fat tails but finite variance, like the truncated Pareto, lognormal and stretched exponential.

As in sheet 1 these instructions assume you have some familiarity with running Matlab, editing scripts etc. Please talk to me or the other people on hand, if you need help with this, and use the “help” and “doc” commands e.g. “help normrnd”, and “doc normrnd”.

I am also assuming you’ve save the commands used yesterday in a script file that begins with:

```
close all
clear all
```

2 Loading datasets

Loading text-based data into Matlab can be done with the “load” command, e.g.

```
load words.txt
```

try plotting these as

```
 figure(1)
plot(1:size(words,1),words,'*')
xlabel('word identifier')
ylabel('number of occurences')
```

and loglog

```
 figure(2)
loglog(1:size(words,1),words,'*')
xlabel('rank')
ylabel('frequency')
```

Why have I changed the axis labels ?

Now use MLE:

```
[alpha, xmin, L]=plfit(words)
plplot(words,xmin,alpha);
title('Zipf dataset: MLE of power law tail')
grid on
```

What do you get for α and x_{min} ?

Also look at the same data with the histogram, the normplot (and the normplot of log(data)), and other diagnostics. Would you have made the same inference in these cases ?

Now repeat for the fires, and E. Coli, datasets. What inference would you make for these ?

For more details on the datasets, see the page at Aaron Clauset's site that we have linked to.