Activity 2. Web Browsing Descriptive Data Analysis

Computer Scientists are being called more and more frequently to provide computer log data that can be used to find out how users interact with the Internet (see, for example, Sen and Hansen (2003), Cadez et al. (forthcoming), Huberman et al (1998)). In addition to that, the number of Internet user surveys is growing at an amazing speed (see, for example, Wellman and Haythornthwaite (2002)). The following problem uses a data set published in the UCI KDD Archive (Kederman, 2003) to obtain data on the number of different pages visited by users who entered the msnbc.com page on September 28, 1999 and other information. The variables we analyze in this activity are usually investigated in the references provided above.

Data Description

The data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 26, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hours period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail–that is, at the level of URL, but rather, they are recorded at the levels of page category (as determined by a site administrator). The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs (bulletin board service)", "travel", "msn-news", and "msn-sports". The variable "depth" shows how many different sites are visited by users, and the variable "length" represents the actual total number of pages visited by each user (total number of clicks).

Data URL

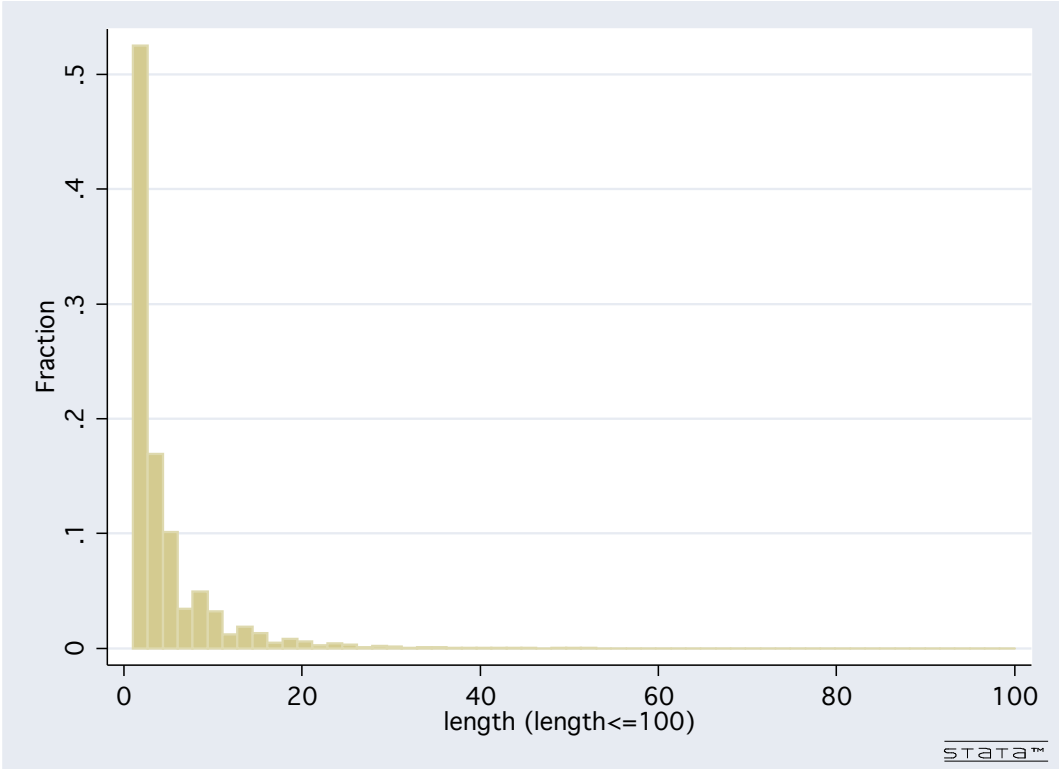
http://kdd.ics.uci.edu/databases/msnbc/msnbc.data.html

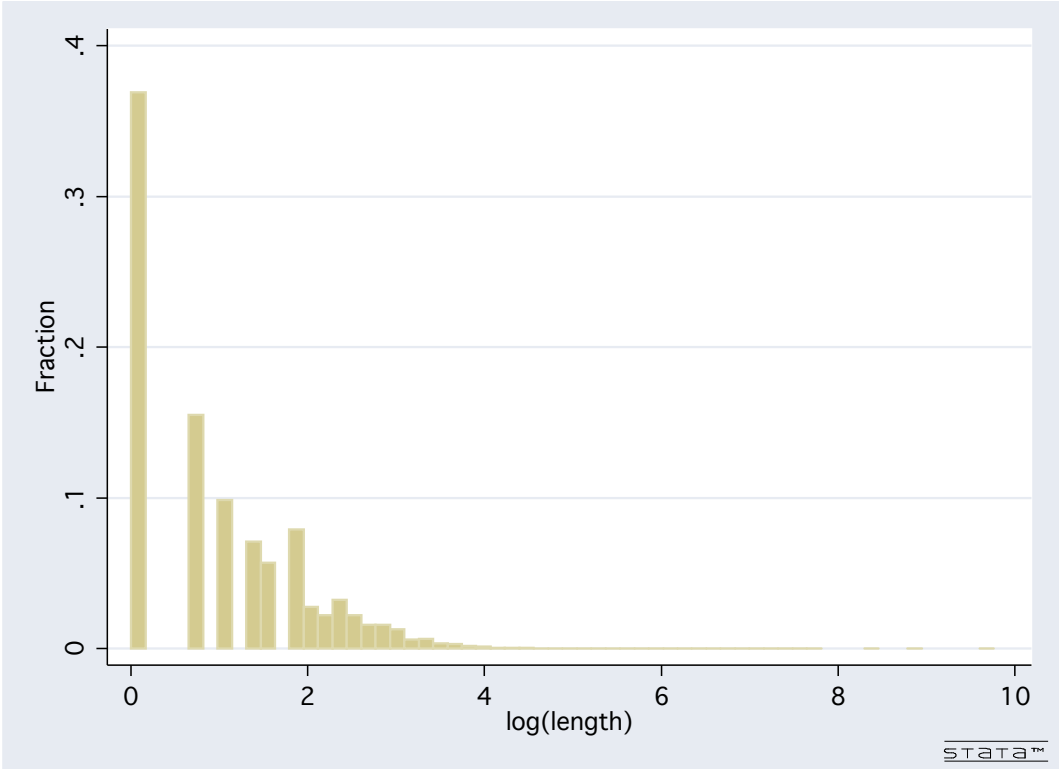
Number of users: $n = 989818$

Activity

Given below are the histograms for the variable "length", the total number of pages visited by each user, which we created from the KDD data set, and a histogram for the log of length.

Question 1: Describe the histograms for the "length" variable in terms of what it is saying about the frequency with which different lengths of visits occur in the data.


Fraction
0
.1
.2
.3
.4
.5
0
20
40
60
80
100
length (length<=100)
stata™


Fraction
0
.1
.2
.3
.4
0
2
4
6
8
10
log(length)
stata™

You may feel suspicious that we only presented the histogram for lengths less than 100 visits. The fact is that there are longer lengths in the data, and it is very hard to determine what constitutes a legitimate surfing and what is a "crawler" or "robot." Some users may be accessing the msnbc web page from the same IP address, making it appear as if one user was surfing many pages in one day. To get a feeling for the actual lengths, you may want to look at the data yourself. So access the data set:

```
.use http//www.stat.ucla.edu/~jsanchez/oid03/datasets/msnbclength.dta
```

Do a
.tabulate length
to see the frequencies of the possible lengths of visits by users, and find summary statistics
.summarize length, detail

Question 2: According to the tabulation you did and the summary statistics of the length of visits, what do you conclude about the best statistics one should use to describe these data?

Question 3: What do you suggest we do with the data if we want to maintain it in raw state, i.e., without transforming into logs and without losing its representativeness of surfing behavior? Be very specific about what criteria you use. Present the pros and cons of your proposals. Compute the new summary statistics.

References

(1) Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S. (2003). Model-based clustering and visualization of navigation patterns on a Web site. Accepted for publication on Journal of Data Mining and Knowledge Discovery, 7(4).
(2) Hansen, M.H. and Sen, R.(2003). Predicting Web User's next access based on log data. Journal of Computational and Graphical Statistics, Vol 12, No. 1, March, p. 143-155.
(3) Heckerman, D. The UCI KDD Archive

`http://kdd.ics.uci.edu`

Irvine, CA: University of California, Department of Information and Computer Science.
(4) Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M. (1998). Strong Regularities in World Wide Web Surfing. Science, Vol. 280, 3 April.

(5) Wellman, B. and Haythornthwaite edts.(2002). The Internet in Everyday Life. Blackwell Publishing.