Web browsing behavior [1]

Juana Sanchez

I need to include the namespace for the interactive components
The resolution on these plots (the ones in the original pdf) from R are really terrible. In many cases I can't even read the axis labels when I zoom in 200%.
GOAL: make the document more interactive so it's a DISCOVERY of why the inverse Gaussian is a good approximation of the distribution of the length of internet site visits by a given user.
How should I deal with the formulas in section 3? R expressions? They're pretty complicated to be expressions.
The code given in the pdf doesn't work in R
Something about the third line of code given in the pdf won't work in the xml file
I had to take out some of the quotation marks from the paras, but not all...
There is no indication for where I should find "frequencytable," which is a table I need in order to run a lot of the r:code.
i:button would be an interactive component that shows a button (named whatever the id says... or maybe another attribute called 'name' would be good) to possibly be clicked on that when clicked would do whatever is inside the i:button (which would generally be r:code.) In other words, whatever is in an i:button tag is never really looked inside of unless clicked on.
INTERACTIVE IDEA: The first histogram with the superimposed inverse gaussian distribution could instead be interactive, with the inverse gaussian line only showing up if interactively told to. This could also allow the reader to explore other distributions (also as buttons) in the process and see that in fact the invserse gaussian is a nice fit visually even in comparison to other logical options like the Normal, Geometric, and Poisson.
INTERACTIVE IDEA: For figure 2, the CDF and p-p plot side-by-side, it would be nice to see the empirical points and then the inverse gaussian appear in two separate steps (maybe even slowly) so the reader could see it happening... I guess it would be sort of like an element of surprise.

Introduction

Once a user enters a web site how many pages or links within the site does that user visit? The answer to this question may suggest actions to improve the site. If similar distributions for the number of pages visited per user are observed at different web sites, then maybe some laws can be established for all sites. Research efforts in this area are directed at finding these laws. Examples of these efforts are Hansen and Sen (2003), and Huberman, Pirolli, Pitkow, and Lukose (1998), each analyzing a different data set.

Some of the analysis done in the literature to answer that question can be illustrated with data published in the UCI KDD Archive (Heckerman). We processed these data to obtain observations on the number of different pages visited by users who entered the msnbc.com page on September 28, 1999 and other information. A random sample of this data set was used by Cadez, et al. (2003) to do cluster analysis and visualization of the patterns (order) of visits followed by users, i.e., to see the frequency of whole sequences. This is a very important question, too. But we don't look into it in this paper.
The digits="3" is fine, but unnecessary when we know the value is an integer

[1] Adapted by Jamie Julin from the JSE article Internet Data Analysis for the Undergraduate Curriculum found in the Datasets and Stories section of the Journal of Statistics Education: http://www.amstat.org/publications/jse/v13n3/datasets.sanchez.html .

The original data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 28, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hours period. Since there are 989818 users, there are 989818 sequences. This is a 22.6 MB size data set. Each event in a sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail--that is, not 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". As an example, we write below the sequence for the first three users in the data set (one line per user):

```
User 1 frontpage, frontpage
User 2: news
User 3: tech,news,news,local,news,news,news,tech,tech
```

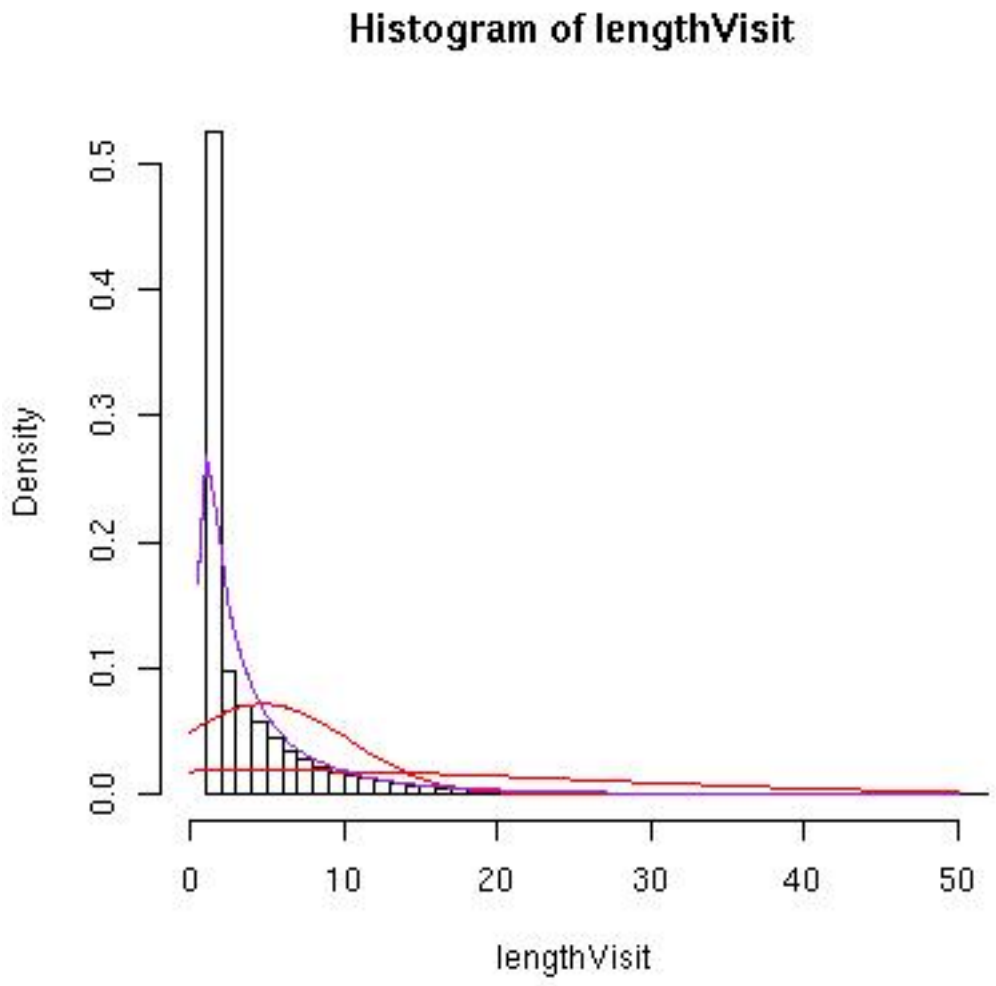
We processed the original data set to obtain the variable 'length', which represents the actual total number of links visited by each user. For example, user one has length=2, user two has length=1, and user three has length = 9.

Descriptive Analysis of the length of visits

The average number of pages visited is 4.747, the median is 2 pages, the minimum is 1 and the maximum is 14795 pages. The histogram, in Figure 1, is very skewed. *How about summary(lengthVisit)*

Figure 1. Distribution of Number of Pages Visited

```
hist(lengthVisit, xlim=c(0,50), prob=TRUE, breaks = c(1:20, seq(25, 50, 5), 100, 2
curve(dnorm(x, mean = mean(lengthVisit), sd = sd(lengthVisit[lengthVisit < 50])),
```



Want to be able to add and remove these curves

```
curve(dnorm(x, mean = mean(lengthVisit), sd = sd(lengthVisit)), col = "red", add =
```

Put a main = on the plot to provide a better title.
Need a legend or a caption to explain the curves

```
nuHat = mean(lengthVisit)
lambdaHat = 1/(mean(1/lengthVisit) - 1/nuHat)
curve(dinvGauss(x, nu = nuHat, lambda = lambdaHat), col = "purple", add = TRUE)
```

Notice that the histogram contains only values of length less than or equal to 100, excluding those users that visited more than 100 pages. The longest visits are probably crawlers or maybe different people logged into the same IP address. One of the problems with web server log data is precisely what to do with these crawlers. Should they be included, should they not? Although we did not include them all in the graphs, all the numbers were used for the computations of the statistics. An important fact to observe is that most users visit few pages, but the tails are very long, indicating that some users visit a lot of pages.

Fitting the Inverse Gaussian distribution to the length of visits

What model should we use for this behavior? Huberman, et al. (1998) and other authors, recommended an inverse Gaussian distribution for the variable length (L). This distribution has two parameters and is described by the formula

The mean and variance, where is a scale parameter. This distribution has a very long tail, which extends much farther than that of a normal distribution with comparable mean and variance. This implies a finite probability for events that would be unlikely if described by a normal distribution. Consequently, large deviations from the average number of user-clicks computed at a site will be observed (Huberman, et al. 1998, pg. 95). Another property is that because of the asymmetry of the distribution function, the typical behavior of users will not be the same as their average behavior. Thus, because the mode is lower than the mean, care must be exercised with available data on the average number of clicks, as this average overestimates the typical depth being surfed.

It can be shown that the cumulative distribution function of the inverse Gaussian distribution is where is the standard normal distribution function.

Is the inverse Gaussian really a good model for the data we have? It is instructive to follow the guidelines given in the references mentioned above to answer this question.

Theoretically, by maximizing the likelihood function, the equations for the maximum likelihood estimators (MLE) of and in the inverse Gaussian distribution given above can be found to be and

For the msnbc.com data, we find:

The inverse Gaussian with these estimates is fitted to the histogram in Figure 1. Visually, it is not a perfect fit for lower values of length, which is where the majority of the data are concentrated. And we don't show the tails, so we can not conclude from this plot that the fit is good over the whole distribution.

To see how good is this model, Huberman, et al. (1998) and Hansen and Sen (2003) compared the cumulative distribution function implied by the model to the empirical cumulative distribution function derived from the data. Then they use a probability-probability plot against the fitted distribution. We do the same with the length variable; the plots can be seen in Figure 10.The p-p plot reveals a misfit of the inverse Gaussian model to our data at the lower values of length. Hansen and Sen (2003) got similar results with the bell-labs.com data set they used.

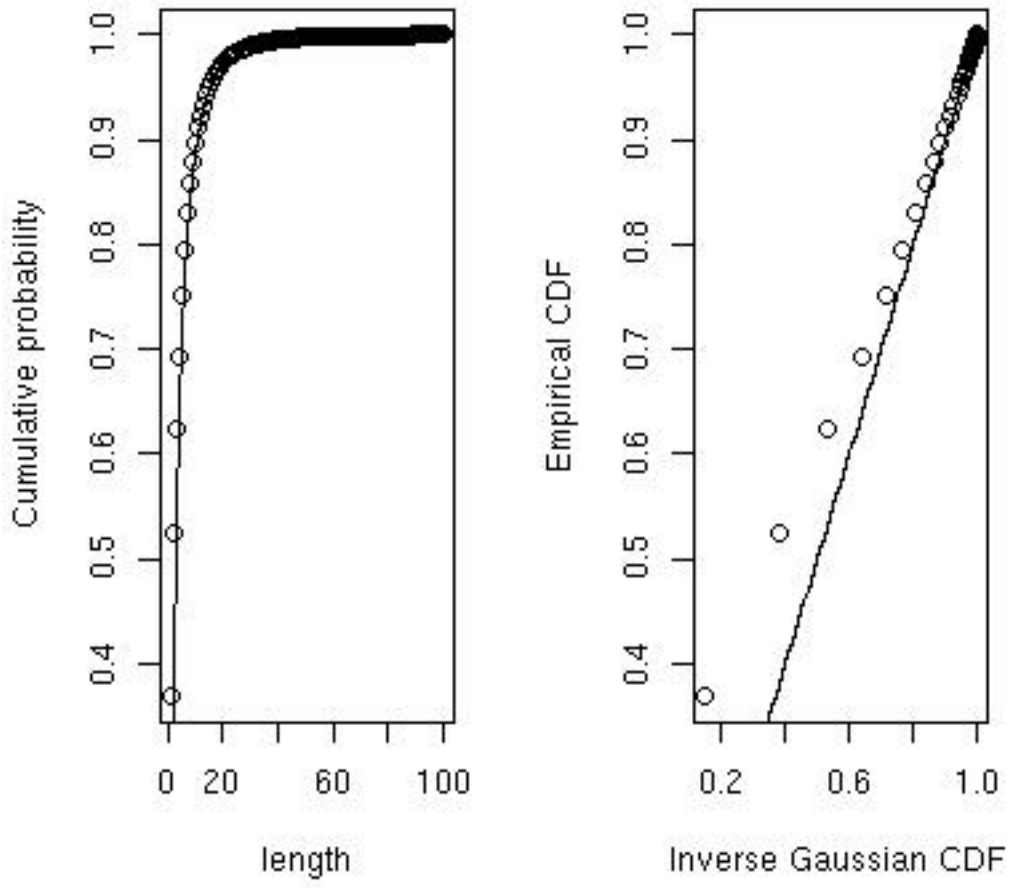
Figure 2. Quantile-Quantile and P-P plots for the Inverse Gaussian

```
n=length(lengthVisit)
lengthsum=sum(lengthVisit)
lengthtable=table(lengthVisit)
lengthfrequency=table(lengthVisit)/n
cumlength=cumsum(lengthfrequency)

l=seq(1,100)
prob=pinvGauss(l,nuHat,lambdaHat)
par(mfrow=c(1,2))
```

```
plot(l,cumlength[1:100], type="p", xlab="length", ylab="Cumulative probability")
lines(l,prob)

plot(prob,cumlength[1:100], xlab="Inverse Gaussian CDF", ylab="Empirical CDF")
lines(l/100,l/100)
par(mfrow=c(1,1))
```

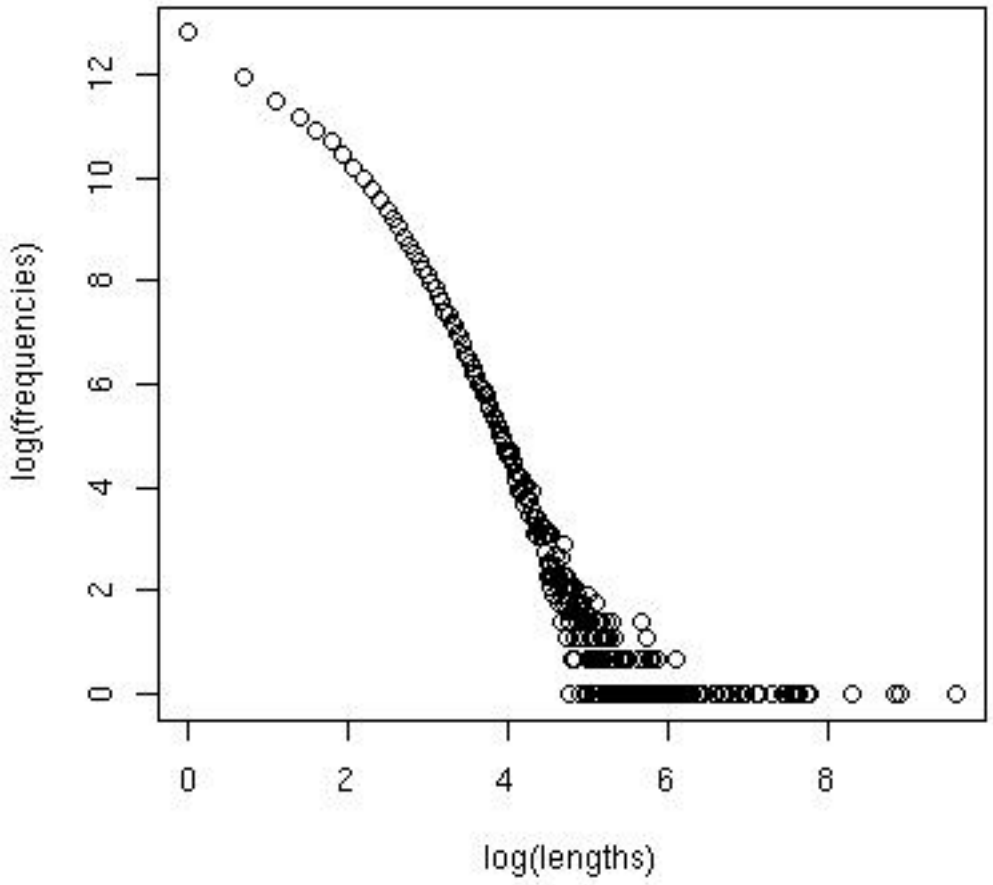


Another way of investigating whether the inverse Gaussian is a good model, is based on the following fact: If you take logs on both sides of the inverse Gaussian formula you obtain

Thus a plot of log(L) vs log(frequency) should show a straight line whose slope approximates -3/2 for small values of L and large values of the variance. This is because if we substitute for on the right hand side of the formula, i.e., which follows from the formula for the variance, the variance appears in the denominator and the mean in the numerator. For small mean, which is the case here, and large variance, which is also the case for our data, the second term is almost 0. A plot of the frequency distribution of surfing clicks on the log-log scale can be seen in Figure 11. According to this plot andthe theoretical result, the regression line for the whole range of the data has a slope of -1.9, which is not too far from -1.5, so this result holds approximately for our data.

Figure 3. Plot of frequency distribution on log-log scale

```
frequencies = table(lengthVisit)
lengths = sort(unique(lengthVisit))
plot(log(lengths),log(frequencies))
# lm(log(frequency)~log(lengthVisit))
```



The log-log plot also helps us notice that, up to a constant given by the third term, the probability of finding a group surfing at a given level scales inversely in proportion to its length, . This is a characteristic that appears in a lot of Internet data sets. We don't pursue it further here, but it is at the heart of the debate about the nature of the data and the best possible model.

The appeal of the inverse Gaussian is in its decision theoretic foundations: it is the distribution that would result if visitors to the web site were optimizing their utility (Huberman, et al. 1998). But based on the results above, would we recommend the inverse Gaussian model for the length of visits (or number of links that a user visits) in the msnbc.com data set or other web server log data? This is one of the questions still unanswered and in need of more research. Several authors have tried other distributions with other data sets, for example, the geometric, the Poisson and a power law, but none of these distributions have fit the data well, either in the upper tail or in the lower levels of length. See, for example, Baldi, Frasconi, and Smyth

(2003). A power law distribution tends to give a good fit at the tail, but it fails in fitting the lower values, which is where most of the observations are concentrated.

The sequence of user requests

Once the distribution of ``length'' is settled, the next step for researchers is to model the sequence of requests by users. Huberman, et al. (1998) model them using a simple first order Markov model. Hansen and Sen(2003) try a first and second-order Markov model, a finite mixture of first-order models, and a Bayesian approach. Cadez, et al. (2003) investigate simple Markov models for different clusters of users. The objective of these modeling attempts is to determine the best model to predict a user's next page request. Pages with higher probability of being requested can then be made more accessible.

Interested readers can experiment in class with many of these questions, either using the raw data, or three different processed data sets that we extracted from this raw data set, and that are available at the CS-STATS web site that we will be glad to provide upon request. Perhaps the reader can obtain Web server log data from the school where this material will be taught. In the latter case, be aware that raw log server data with URLs and detailed computer information needs to be converted to something like the raw data of Heckerman using Perl or similar programs. After that, you can process it further to use it for data analysis. The CS-STATS web site has some Perl scripts that could be used to that end.