Modeling web browsing at the user level
How well does the Inverse Gaussian model fit?
Key

Juana Sanchez

INSTRUCTIONS FOR FORMAT:

This project should be written like the first one. You should write a small report with sections

I.- Introduction which includes the main question (is the inverse gaussian a good model for....?)
II. Data
III. Summary of the Analysis with the graphs and results that are relevant, only
IV. Conclusions.
V. References (this handout and your teacher).
Appendix with commands and results.

To fill in all those sections, you need to read this handout, and find all the questions I ask along. Some are asked at the beginning, but others are within some discussion, so unless you read everything you will miss them. You do not need to copy and paste the paragraphs I mention below, either.
DO NOT ANSWER in fhe following format :
Question 1, question 2, etc... The questions are there to organize your analysis.

Be selective in the output and what you choose from the information given below.

Concepts used in this activity: Histogram, (you could do other graphs if you want...), summary statistics, regression, maximum likelihood estimators, maximum likelihood estimates, numerical methods of finding mle (computer done), cumulative distributions, qq-plots.

Grading: based on how well you follow format, and on how well you explain your answers supporting them with the statistics and output.

1.- Introduction

Web server log data consists of millions of lines like these lines from the ucla stats server, one line for each click or page visited. Some visitors are robots, or the spiders of search engines, like the Googlebot seen below. Some robots are good or bad, but we don't care about this here.

61.149.137.109 - - [01/Jun/2004:00:00:12 -0700] "GET /index.php?vol=2 HTTP/1.1" 200 32896
"http://www.jstatsoft.org/index.php?vol=1" "$
64.68.82.14 - - [01/Jun/2004:00:00:21 -0700] "GET /~cochran HTTP/1.0" 200 2991 "-" "Googlebot/2.1
(+http://www.googlebot.com/bot.html)"
127.0.0.1 - - [01/Jun/2004:00:01:30 -0700] "GET /server-status" 200 17082 "-" "-"
80.232.169.174 - - [01/Jun/2004:00:02:22 -0700] "GET /v06/i06/codes/mingcv.m HTTP/1.1" 200 1806 "-" "tfqsgmsnnpurmbwmgjdyglyogxdpwe"

There are many engineering issues in processing this data in a way that we end up only having the sequence of visits by each individual user. E-commerce web server logs are more interesting, because they have more information about the user than academic sites do. Above, for example, we only have the IP addresses of the visitors.

Computer Scientists are being called more and more frequently to provide computer log data like that above that can be used by statisticians and e-commerce people to find out how users interact with the a web site (see, for example, Sen and Hansen (2003), Cadez et al. (forthcoming), Huberman et al (1998)). 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 the variables investigated in the references provided above. They have become standard metrics in this area of research.

2.- Data Description

The data we will use in this activity has already been processed once and converted to numerical data that we can, in principle, process even more to prepare it for the analysis we want. It is not data from UCLA, either. The processed 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) --logs that look like those above. 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 unique pages 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
After processing logs like those above, the data set you see in this URL looks like this

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

or, symbolically, by assigning numbers to each click and substituting,

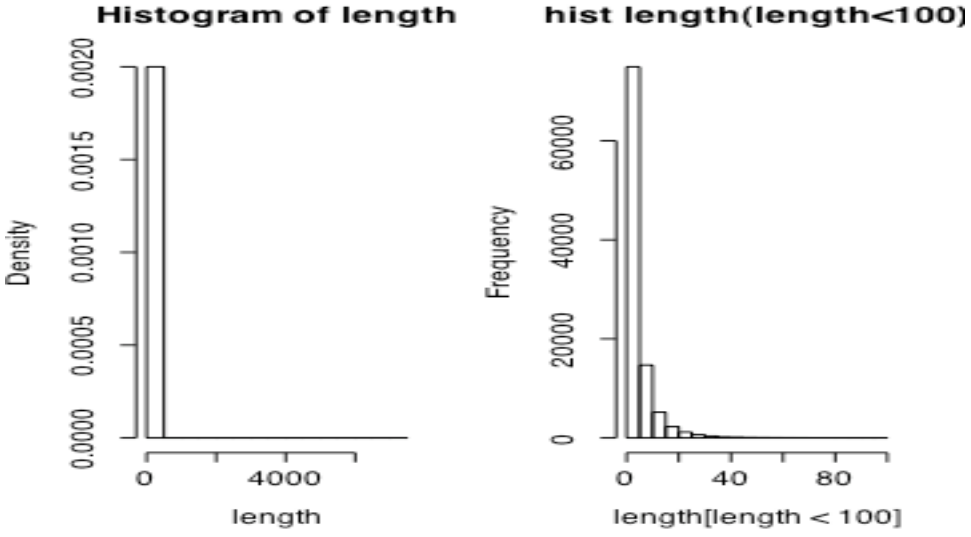
User session 1: 1, 1
User session 2: 2
User session 3: 3, 2, 2, 4, 2, 2, 2, 3, 3

The above is saying that the first user session entered the msnbc.com web site via the frontpage, and hit two links in the frontpage. User session 3, however, started in the tech page, moved to the news page and hit two links there, then to local, then to news again and hit two links there, then to tech again, and hit two links there. The reader should visit the msnbc.com web site to understand that the user can move from any page category to another due to the frame structure of that web Site. This aspect makes this web Site different from Sites used in other research.

The above sequences were processed further by J.Sanchez to obtain the variable length that you will use in this activity. The "length" variable will measure how many clicks did each user do after entering the web page. For instance, user 1 clicked twice, user 2 once, user 3 clicked 9 times.

In this report we only use the data for 100000 visitors. The histogram of length or number of clicks by visitors can be seen in Figure 1. We can see that if we leave all the observations, it is hard to see detail, due to some observations being too large. So we plot for length less than 100 and we can see more detail.

Figure 1.- Histogram of number of pages visited by msnbc visitors.



The histogram says that most users visit very few pages and very few users vitis a lot of pages within the MSNBC web site. The are a few users who click a huge number of times, and these are probably not humans, they are probably robots from search engines.

The median length is 2, the mean is 4.779. The third quartile is 6. So 25% of the visitors visit more than 6 pages. The maximum is 7033.00 indicating that the people that generated the original data set at the kdd archive probably did not remove all the robots. A box plot will give us a better feeling of the extent of the outliers…

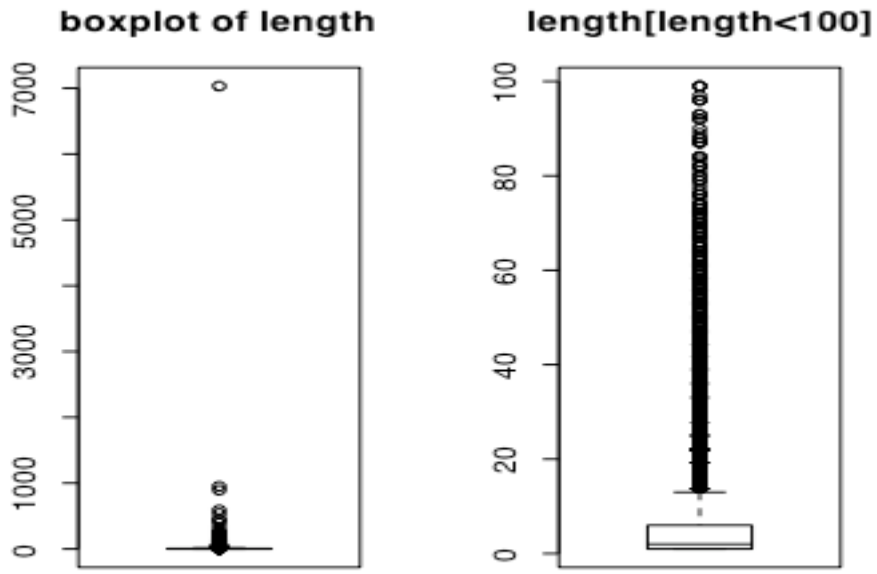


Figure 2.- Box plots of the number of pages visited by visitors to msnbc

Notice how the large observation is an extreme outlier, outside the whole data set. We should remove it for further analysis.

When we look at the distribution of length for values less than 100, we see many outliers, but these are too many to be considered outliers. What we have here is a situation where thick tails is the norm, rare is not so rare in this internet data analysis. And these should be kept and modeled properly. Most of the models we learned will not work for this kind of data sets.

To remove the extreme outlier I typed
>length=length[length<3000]

3.- Data Analysis

Huberman et al.(1998) derive the probability P(L) of the number of

links L that a user follows in a web site. They determined that this distribution is given asymptotically by the two-parameter inverse Gaussian Distribution

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} e^{\left[-\frac{\lambda(L-\mu)^2}{2\mu^2 L}\right]}$$

with mean E(L)=μ and variance Var(L)=μ3/λ, where λ 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." 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 length being surfed."

It can be shown that the cumulative distribution function of the inverse Gaussian distribution is

$$F(L,\mu,\lambda) = \Phi\left[\sqrt{\frac{\lambda}{L}}\left(\frac{L}{\mu}-1\right)\right] + e^{\frac{2\lambda}{\mu}}\Phi\left[\sqrt{\frac{\lambda}{L}}\left(\frac{L}{\mu}+1\right)\right]$$

where Φ is the standard normal distribution function.

The theoretical equations for the maximum likelihood estimator of λ and μ in the inverse Gaussian distribution given above can be derived as shown in Figure 3 (see next page). Using our data set, we compute the quantities necessary to substitute in those formulas and find the mle estimates for our data set in closed form. This will be the model that is most likely to have generated the data (assuming that an inverse Gaussian model is a good model, which is a big assumption).

The MLE estimates are

$\hat{\mu}_{mle} = 4.708917$

$\hat{\lambda}_{mle} = 3.089086$

Alternatively, we can use a numerical routine in R to find the MLE estimates directly without mathematics.

•

$$
\begin{aligned}
P(L_i) &= \sqrt{\frac{\lambda}{2\pi L_i^3}} \exp[\frac{-\lambda(L_i-\mu)^2}{2\mu^2 L_i}] \\
\Longrightarrow P(L) = P(L_1, \cdots, L_n) &= \prod_{i=1}^{n} \{\sqrt{\frac{\lambda}{2\pi L_i^3}} \exp[\frac{-\lambda(L_i-\mu)^2}{2\mu^2 L_i}]\} \\
&= \lambda^{\frac{n}{2}} (2\pi)^{-\frac{n}{2}} (\prod_{i=1}^{n} L_i)^{-\frac{3}{2}} \exp\{\sum_{i=1}^{n} \frac{-\lambda(L_i-\mu)^2}{2\mu^2 L_i}\} \\
\Longrightarrow \log P(L) &= \frac{n}{2}\log\lambda - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log L_i - \frac{\lambda}{2}\sum_{i=1}^{n}\frac{(L_i/\mu - 1)^2}{L_i}
\end{aligned}
$$

$$
\Longrightarrow \begin{cases}
\frac{\partial \log P(L)}{\partial \lambda} = \frac{n}{2\lambda} - \frac{1}{2}\sum_{i=1}^{n}\frac{(L_i/\mu - 1)^2}{L_i} = 0 \\
\frac{\partial \log P(L)}{\partial \mu} = -\frac{\lambda}{2}\sum_{i=1}^{n}\frac{2(\frac{L_i}{\mu} - 1)\cdot(-\frac{L_i}{\mu^2})}{L_i} = 0
\end{cases}
$$

$$
\Longrightarrow \begin{cases}
\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} L_i \\
\hat{\lambda} = \frac{n}{\sum_{i=1}^{n}\frac{(L_i/\hat{\mu} - 1)^2}{L_i}}
\end{cases}
$$

Figure 3.- Theoretical derivation of the MLE estimators for the parameters of the Inverse Gaussian model.

The MLE estimates obtained directly with the numerical routine are:

$\hat{\mu} = 4.708917 \qquad se_{\hat{\mu}} = 0.01838$

$\hat{\lambda} = 3.089093 \qquad se_{\hat{\lambda}} = 0.013814$

which are identical to the ones we found with the closed form mathematical solution.
-2 log L: 468712.4

To see how good is this model, Huberman et al.(1998) and Sen and Hansen(2003) compare the cumulative distribution function implied by the model to the empirical cumulative distribution function derived from the data. Then they use a quantile-quantile against the fitted distribution. *We will do this here. We find the cumulative distribution of the length values which are less than 100, and the cumulative distribution implied by the model just estimated and superimpose them. We do also a q-q plot and determine how good is the fit based on that. Figure 4 shows these plots.*

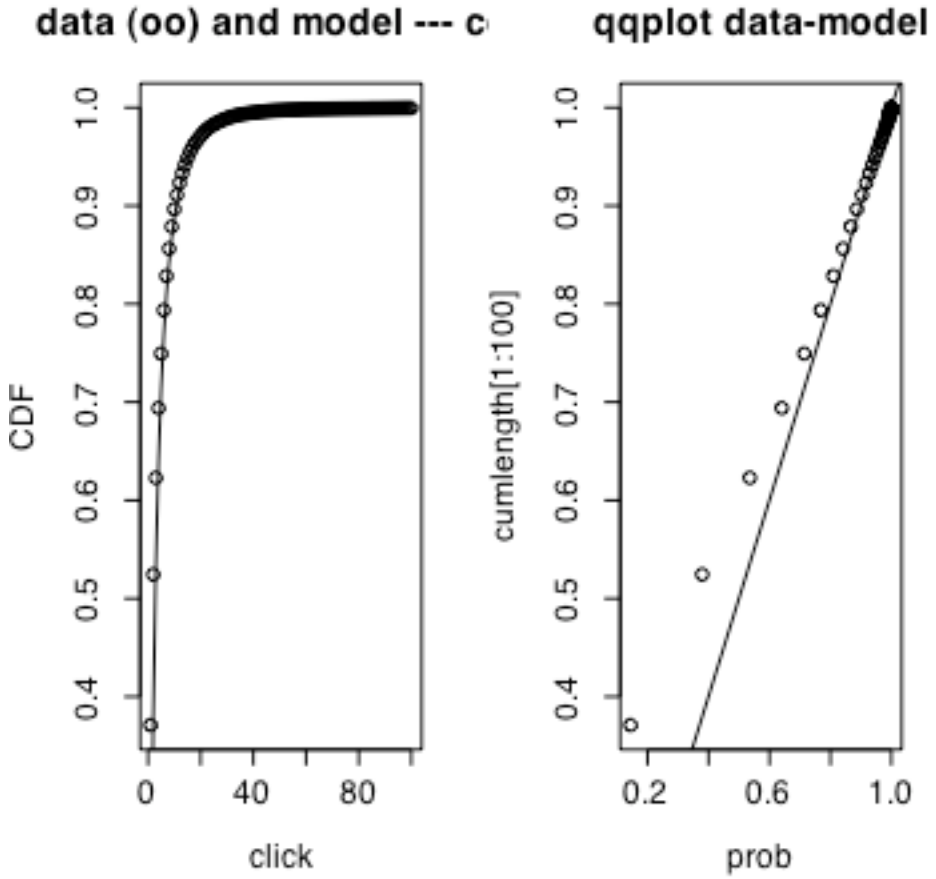


Figure 4.- CDF of the data and the model fit overimposed (left) and qq plot of the data and model fit (right)

It is obvious from the plots in Figure 4 that the inverse Gaussian model does not fit the data well for small length of visits. The cumulative distribution plot shows that a little ambiguously because the line seems to be on the dots. The qq plot makes it more obvious. To see that we could also see this in the cumulative distribution, we repeated the plot for only smaller values of the length variable. Figure 5 shows more clearly that for short visits, the inverse Gaussian model cumulative distribution is not fitting very well. Because it doesn't fit for small values of length and most of the visits have small length, we can say that the model does not fit well for most of the data, and hence we should question its validity for this data set.

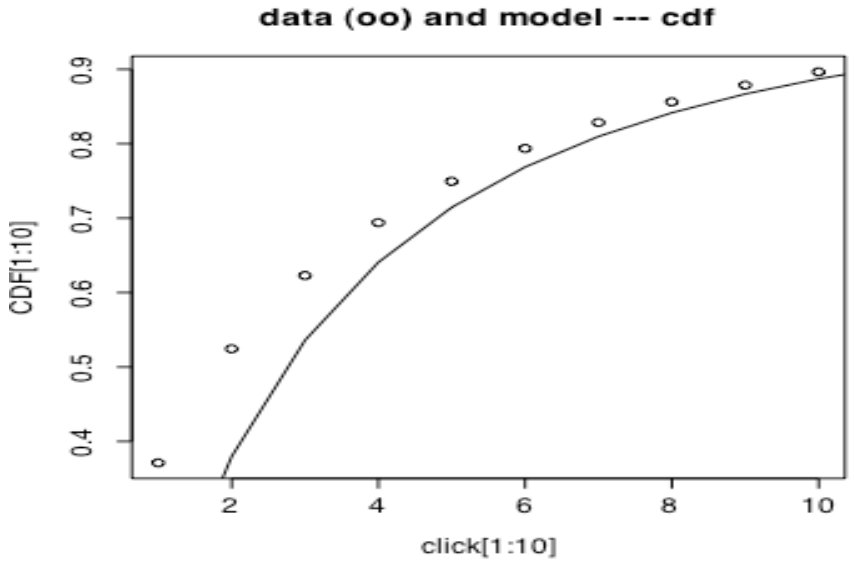


Figure 5.- Blow up of the cumulative distribution plots for the data and the model fit.

If we take logs in both sides of the inverse gaussian formula we obtain an expression such that a plot of log(P(L)) vs log(L) shows a straight line whose slope approximates -3/2 for small values of L and large values of the variance. This could be another way of determining whether the model is good for these data.

Figure 6 shows the plot of the log of the frequency and the log of the length and the regression line fitting these two. We can see that the fit is not very good. Looking at the results of the regression fit, we find that the slope is –1.989. We fitted the line to all values of length. Looking at Figure 6 we can see that if we had fitted it only to the small values of length (i.e. fit only the first 20 values, like we did in the cumulative distribution of Figure 5), the slope would –1.598, which is much closer. However, as we show in Figure 7, the relation is not linear, and therefore, this should be interpreted with lots of caution.

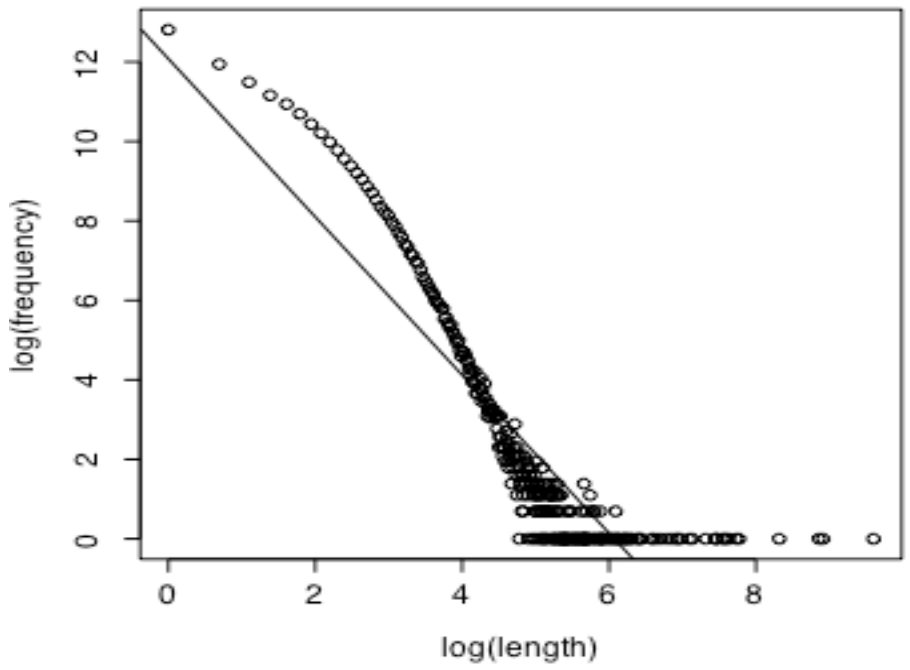


Figure 6.- Log-log plot of frequency versus length to check for the theoretical result that the slope of the regression line is –1.5

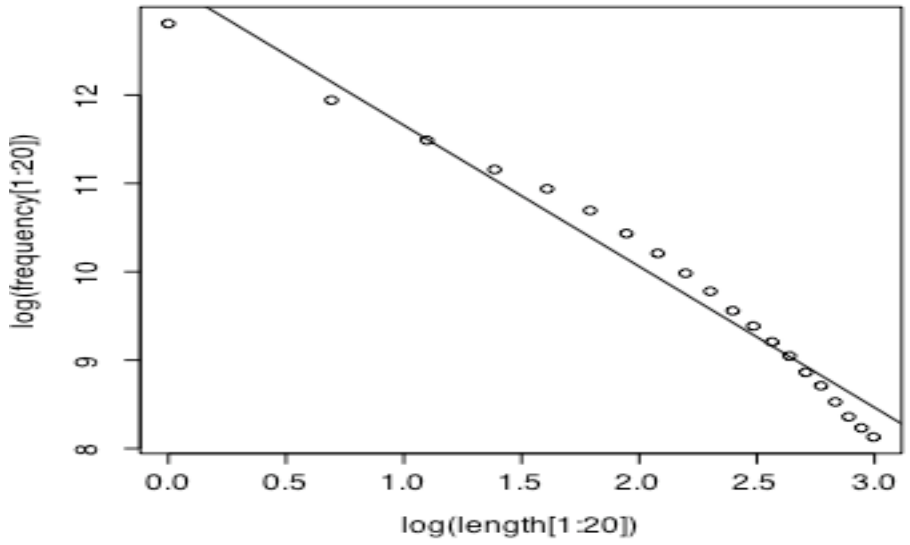


Figure 7.- The same as 6 but fitting only the first 20 values of the frequency distribution.

5.- Conclusions

With the data that they used, Sen and Hansen did not find much support for the inverse Gaussian model. They used the tools we used in this report. We found in our analysis of the msnbc data set that the inverse Gaussian model does not fit well the length of visits (or number of pages clicked) by visitors to the msnbc web site. We illustrated this lack of fit by first fitting the model using maximum likelihood and then observing the cdf implied by the model and the cdf of the data (or empirical cdf). We also compared the quantiles of the data and the quantiles implied by the fitted model in a qq plot. Finally we investigated the implication of the Inverse Gaussian model that the slope of the log-log relation between frequency and length was –1.5. ***All of the above methods showed that the inverse Gaussian model does not fit the length data well when length is small. Since about 75% of the observations have length values that are small, we are saying that the model does not fit well 75% of the data. That is pretty bad !!!! So the Inverse Gaussian model is not a good model.***

It could be that the
model is good for the msnbc data set, so it is worth trying further
and see what conclusions we reach.

\end{verbatim}

\vspace{0.2in}

\begin{flushleft}

{\bf Question 5}. Based on your results above, would you recommend the
inverse Gaussian model for the length of visits (or number of links
that a user visits) in the msnbc.com data set? Why? Explain your reasons and back them up with the analysis done above.

\end{flushleft}

\vspace{0.2in}

\vspace{0.2in}

{\bf References}

\vspace{0.2in}

(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 \begin{verbatim} http://kdd.ics.uci.edu \end{verbatim} 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) Sanchez, J. Internet Data Analysis for the Undergraduate Statistics Curriculum. Forthcoming, Journal of Statistics Education.

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

Appendix. R commands used

```
install.packages("SuppDists")
install.packages("stats4")

msnbc = read.table("http://www.stat.ucla.edu/~jsanchez/teaching/stat110A
/reading/msnbclength.txt",nrow=100000,header=T)
length = msnbc[,2]  # the length variable is the second column
n=length(msnbc[,2])  #sample size
hist(length,prob=T)   # get a feeling for the distribution and summaries
hist(length,prob=T, main="hist length<100")
summary(length)
boxplot(length,main="boxplot of length")
boxplot(length[length<100],main="length[length<100]")
length=length[length<3000]
mu=mean(length)                 ## mu
temp=(((length/mu)-1)^2)/length
lambda=n/sum(temp)              ##lambda
lengthsum=sum(length)
lengthtable=table(length)       ## frequency of length
```

```
lengthfrequency=table(length)/n   ## relative frequency of length, empirical pdf
cumlength=cumsum(lengthfrequency)   ##cumulative distribution of the data
library(stats4)   # need this library to do mle
#now we write the negative of the log likelihood of the data (double check with your
notes)
fn=function(mu,lambda)
{
 -(   (n/2) *log(lambda)-(n/2)*log(2*pi) -(3/2)*sum(log(length))-((lambda)/2)*
 sum((((length/mu)-1)**2)/length))
 }
est=mle(minuslogl=fn,start=list(mu=0.1,lambda=0.1))
summary(est)

library(SuppDists)
l=seq(1, 100)
prob=pinvGauss(l, mu, lambda)        ##calculate the distributon of length
##put CDF on same graph to check the fit
##lines for inverse Gaussian
click=l
CDF=cumlength[1:100]
 plot(click, CDF, type="p",main="data (oo) and model --- cdf")
lines(click, prob)

##(Q-Q plot)
qqplot(prob, cumlength[1:100])   ##same as above
abline(0,1)

#blow up
plot(click[1:10], CDF[1:10], type="p",main="data (oo) and model --- cdf")
##point for empirical data
lines(click, prob)
freqtable = read.table("http://www.stat.ucla.edu/~jsanchez/teaching/stat110A/reading/
frequencytable.txt",header=TRUE)
length=freqtable$length
frequency=freqtable$frequency
reg=lm(log(frequency)~log(length))  #regression to see if we get -3/2 slope
reg
plot(log(length),log(frequency))  #will do a plot on the log-log scale.
abline(reg)
plot(log(length[1:20]),log(frequency[1:20]))  #will do a plot on the log-log scale.
 abline(reg)
```