Internet Usage Activity 1
Who is using the Internet?

Data Description

At the end of 2001 the number of Internet users in the world was more than 500 million (up from 16 million in 1996). The Internet has quickly become part of our lives and numerous research efforts have been made in the past to try to understand who is using it and how it is being used. The activity presented here is concerned with those issues.

The data we use to that end comes from a survey conducted by the Graphics and Visualization Unit at Georgia Tech October 10 to November 16, 1997. The full details of the survey are available at http://www.gvu.gatech.edu/user_surveys/survey-1997-10/graphs/#general. The particular subset of the survey provided here is the "general demographics" of internet users, which we have recoded as entirely numeric, with an index to codes described in http://kdd.ics.uci.edu/databases/internet_usage/changes.

The number of users participating in the survey is n= 10108.

Question 1: Browse through the web pages provided above and familiarize yourself with them. Then answer the following question: what kind of sampling, if any, was used to conduct the survey? Can you generalize the results of this survey we are about to see to the overall population of users?

Getting acquainted with some of the data and some basic summaries

In today's lab, we will explore the demographics of internet users. Begin by loading the data into Stata:

· use http://www.stat.ucla.edu/~jsanchez/oid03/datasets/usage.dta

Use the describe command to get a general feel for what is contained in the data sets.

· describe

Question 2: How many variables and observations are there in the dataset?

You can scroll line by line and page by page by typing space and return respectively. Now, we want to see the details of the dataset by typing:

· list

To stop scrolling, type:

· q

Since there are so many variables, we can concentrate on specific variables that interest us. Suppose we are interested in **gender**. Then we can type

· label define gender1 0 "female" 1 "male"
· label value gender gender1
· tabulate gender

The above commands label the values of the variable gender (which is 0 or 1) and tells us the proportion of male and females in the survey.

Question 2: Which gender predominates in the survey?

Suppose we are only interested in the variable of **yearsoninternet** for the youngest 30 people. Since missing values of age are recorded as 0, we want to exclude those 0's by typing:

· mvdecode age, mv(0)

then type:

· sort age

· label define yearsoninternet1 0 "under 6 months" 1 "6-12 months" 2 "1-3 years" 3 "4-6 years" 4 "over 7 years"
· label value yearsoninternet yearsoninternet1
· list age yearsoninternet in 1/30

Question 3: Do the data look consistent? i.e., is there anybody reporting more years using the internet than their age?

Now, let's introduce the command **summarize**. This command will give a brief numerical summary of all or any variable.

· summarize

The above command will give you a few summary statistics of all 72 variables. If you wish to see the summary of a specific variable, for example age, type:

· summarize age, detail

Question 4: What is the arithmetic mean, the median and the standard deviation of age?

If we want to summarize age for men:

· summarize age if gender==0, detail

Question 5: What if we want to summarize age for women? How different are they? Are women users younger than male users?

To look at the summary of men and women together, type:

· sort gender
· by gender: summarize age

Visual description of the data

After we have learned analyzing the summaries and details, let's begin to see the display of the data. First, let's learn the most basic type of graph, the histogram

· hist age

To change the number of bins, x-axis scale and label, and title, the command is the following:

· hist age, xscale(0,100) xlabel(0 10 to 100) bin(20) title(Age of the Respondents)

Question 6: By using different number of bins, which graph is more informative? Can you claim that it is normal?

Now, let's look at boxplots of age and plot age by gender

· sort gender
· graph box age
· graph box age, by(gender)

Question 7: Can you give a description of what you see in the histogram and the box plot? What are the graphs saying about the proportion of users of different ages

Confidence Intervals

With the sample mean of the age variable, we can generate a confidence interval for the population mean:

$$\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$$

Since the sample size is big enough, we can use the sample standard deviation. Suppose $\bar{x}$ is the sample average of any variable, s is the sample standard deviation, and n is the sample size. The following command will generate a 95 percent confidence interval.

· ci age

Question 8: Can you interpret this interval or not? Why?

If you wanted to obtain a confidence interval with different level of confidence, i.e., 90%, you can type

· ci age, level(90)

Categorical Data Analysis

For categorical variables, we can tabulate and graph the variables household income (in thousands) and education attainment respectively by the following commands:

· label define householdincome1 0 "not say" 1 "under 10 dollars" 2 "10-19 dollars" 3 "20-29 dollars" 4 "30-39 dollars" 5 "40-49 dollars" 6 "50-74 dollars" 7 "75-99 dollars" 8 "over 100 dollars"
· label value householdincome householdincome1
· mvdecode householdincome, mv(0)
· tabulate householdincome

· graph bar householdincome
· label define educationattainment1 0 "grammar" 1 "high school" 2 "professional" 3 " some college" 4 "college" 5 " masters" 6 "doctoral" 7 "special" 99 "other"
· label value educationattainment educationattainment1
· tabulate educationattainment · graph bar educationattainment

Question 9: Which income group predominates among users? Which group is the less predominant?

Question 10: What can you say about the education level of users?