Internet Usage Activity 2
More on Who is using the Internet?
Chi Square tests of Independence

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?

Activity

In today's activity, 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

In this lab, we will look at Chi-squared tests for independence to see if two categorical variables are related or not. We will focus on the demographic variables and look at two by two interrelations between the variables. We will look at the following pairs of variables:

- Age (converted to categories) vs years on internet
- Race vs years on internet
- Income vs years on internet
- Disability vs years on internet
- Marital status vs income
- Education vs income
- Education vs years on internet
- Sex vs years on internet

Since the variable "age" is quantitative, we first categorize it by coding in the following way:
· gen ageclass=age
· replace ageclass=8 if age≥70
· replace ageclass=7 if age≥60 & age<70
· replace ageclass=6 if age≥50 & age<60

· replace ageclass=5 if age≥40 & age<50
· replace ageclass=4 if age≥30 & age<40
· replace ageclass=3 if age≥20 & age<30
· replace ageclass=2 if age≥10 & age<20
· replace ageclass=1 if age≥5 & age<10

It is important also to convert non-responses to missing values: · mvdecode ageclass, mv(0)

Thus we classify "age" to eight categories, with "." represents missing categories when respondents failed to provide valid age.
Question 2: Find the frequency distribution for age classes. Which age group accounts for the largest proportion of internet users?

The commands you can use are
· hist ageclass
· sort ageclass
· by ageclass: sum ageclass
· tabulate ageclass
· tabulate ageclass, plot
Similar methods can be used to manipulate other variables.

Choose one or two other variables to repeat above commands.

#Codebook
· label define ageclass1 1 "age ≥ 5 & age< 10" 2 "age ≥ 10 & age < 20" 3 "age ≥ 20 & age < 30" 4 "age ≥ 30 & age < 40" 5 "age ≥ 40 & age < 50" 6 "age ≥ 50 & age < 60" 7 "age ≥ 60 & age < 70" 8 "age ≥ 70"
· label value ageclass ageclass1

· 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

· label define race1 0 “not say” 1“White” 2 “Hispanic” 3 “Asian” 4 “Black” 5 “Latino” 6 “Indigenous” 7 “American” 8 “Korean” 99 “other”
· label value race race1
· mvdecode race, mv(0)

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

· label define maritalstatus1 0 “not say” 1 “divorced” 2 “living with another” 3 “married” 4 “separated” 5 “single” 6 “widowed”
· label value maritalstatus maritalstatus1
· mvdecode maritalstatus, mv(0)

· 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

· label define gender1 0 “female” 1 “male”
· label value gender gender1

To explore the pairwise interrelation as well as cross-tabs tables, use following commands:
· tabulate ageclass yearsoninternet, row column cell chi2
· tabulate race yearsoninternet, row column cell chi2
· tabulate householdincome yearsoninternet, row column cell chi2
· tabulate maritalstatus householdincome, row column cell chi2
· tabulate educationattainment householdincome, row column cell chi2
· tabulate educationattainment yearsoninternet, row column cell chi2
· tabulate gender yearsoninternet, row column cell chi2

Question 3: What can you conclude about independence test from Stata outputs? Are the

pairwise variables independent?

Question 4: Interpret the cross-tabs tables?

The pair "disability vs years on internet" is somehow distinct from other pairs. There are six disability variables in the data set, corresponding to 6 types of disability ("cognitive", "hearing", "motor", "notimpaired", "notsay", and "vision") . One way of handling this is to set up the cross-tabs tables and perform Chi-squared test for each disability variable vs "years on internet".

· tabulate disability_cognitive yearsoninternet, row column cell chi2
· tabulate disability_hearing yearsoninternet, row column cell chi2
· tabulate disability_motor yearsoninternet, row column cell chi2
· tabulate disability_notimpaired yearsoninternet, row column cell chi2
· tabulate disability_notsay yearsoninternet, row column cell chi2
· tabulate disability_vision yearsoninternet, row column cell chi2

Question 5: What can you say about the interrelations between "disabilities" and "years on internet"? Are they all the same?

Another way is to consider combining the 6 types into one single variable "disability", with "0" indicating no disabilities, "1" for "cognitive", "2" for "hearing", "3" for "motor", "4" for "notimparied", "5" for "notsay", and "6" for "vision", respectively. Then set up the cross-tabs table and perform Chi-squared independence test using the same introduced above.
· gen disability= disability_cognitive
· replace disability=2 if disability_hearing=1
· replace disability=3 if disability_motor=1
· replace disability=4 if disability_notimpaired=1
· replace disability=5 if disability_notsay=1
· replace disability=6 if disability_vision=1
· tabulate disability yearsoninternet, row column cell chi2

Question 6: What can you say about the interrelation between "disability" vs "years on internet"? Is it the same as type-wise interrelations? Is Chi-squared test appropriate here?