Internet Usage Activity 4
More on Who is using the Internet?

Bivariate discrete distributions

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?

Activities

In this activity, we will focus on the two variables "years on internet" vs "household income" and see the interrelation between the two variables. Please study the cross table and answer the following questions:

1. Determine the marginal distribution of each of the variables. What are these distributions saying about the income of internet users and the amount of time that users have been on the internet?

2. Determine the conditional distribution of "years on internet" given that "household income" is between $40 and $74 thousand. and the conditional distribution of "household income" given that "years on internet" is less than 6 months. What are these distributions saying about users of the internet?

3. Compare the marginal distributions and conditional distributions. Can you draw any conclusions from them about whether the two variables income and years on internet are independent?

4. Find the Chi-squared test and determine if these two are independent. Does the conclusion make sense? What could explain this?

Cross Table for Variables Years on Internet vs Household Income

Key
frequency
row percentage
column percentage
cell percentage

Years on internet	Household Income								Total
	under $10	$10-19	$20-29	$30-39	$40-49	$50-74	$75-99	over $100	
under 6 months	120	178	283	276	219	301	81	64	1,522
	7.88	11.70	18.59	18.13	14.39	19.78	5.32	4.20	100.00
	19.05	24.22	24.11	21.30	19.18	16.45	9.67	7.66	17.95
	1.41	2.10	3.34	3.25	2.58	3.55	0.96	0.75	17.95
6-12 months	84	142	285	285	241	284	111	82	1,514
	5.55	9.38	18.82	18.82	15.92	18.76	7.33	5.42	100.00
	13.33	19.32	24.28	21.99	21.10	15.52	13.25	9.81	17.85
	0.99	1.67	3.36	3.36	2.84	3.35	1.31	0.97	17.85
1-3 years	301	269	401	457	402	730	344	354	3,258
	9.24	8.26	12.31	14.03	12.34	22.41	10.56	10.87	100.00
	47.78	36.60	34.16	35.26	35.20	39.89	41.05	42.34	38.42
	3.55	3.17	4.73	5.39	4.74	8.61	4.06	4.17	38.42
4-6 years	111	124	169	204	215	342	209	217	1,591
	6.98	7.79	10.62	12.82	13.51	21.50	13.14	13.64	100.00
	17.62	16.87	14.40	15.74	18.83	18.69	24.94	25.96	18.76
	1.31	1.46	1.99	2.41	2.54	4.03	2.46	2.56	18.76
over 7 years	14	22	36	74	65	173	93	119	596
	2.35	3.69	6.04	12.42	10.91	29.03	15.60	19.97	100.00
	2.22	2.99	3.07	5.71	5.69	9.45	11.10	14.23	7.03
	0.17	0.26	0.42	0.87	0.77	2.04	1.10	1.40	7.03
Total	630	735	1,174	1,296	1,142	1,830	838	836	8,481
	7.43	8.67	13.84	15.28	13.47	21.58	9.88	9.86	100.00
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	7.43	8.67	13.84	15.28	13.47	21.58	9.88	9.86	100.00

Pearson chi2(28) = 494.3259 Pr = 0.000