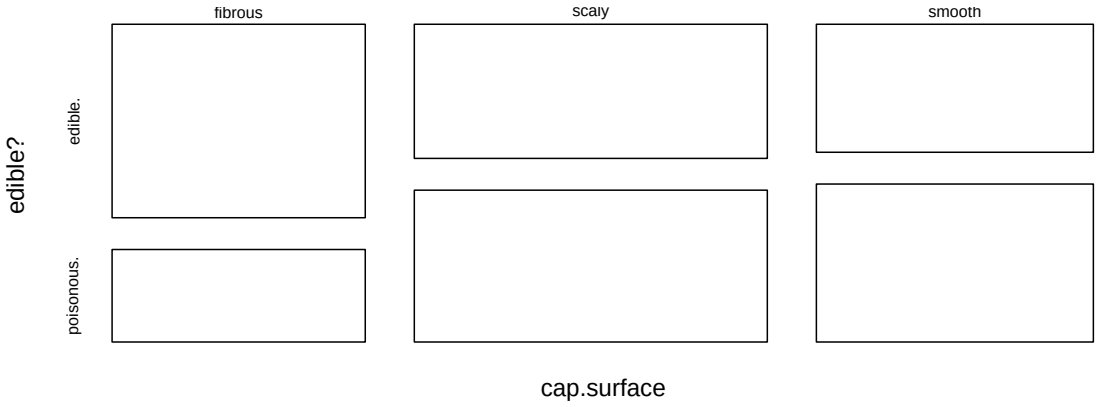
36-350: Data Mining

Homework 5
Date: September 28, 2001 **Due: start of class October 5, 2001**

---

1. (Mosaic plots) In order to determine which mushrooms are poisonous, a contingency table is made, relating the cap surface of a mushroom to whether it is poisonous or not. Below is a mosaic plot of the table.



   (a) Are these two variables likely to be independent? Note: the smallest count in the table is 502.

   (b) Where on the plot can we find the probability that a mushroom is smooth?

   (c) Where on the plot can we find the probability that a mushroom is edible?

   (d) Where on the plot can we find the probability that a scaly mushroom is edible?

   (e) Where on the plot can we find the probability that a poisonous mushroom is fibrous?

   (f) True or false: Based on the table, most poisonous mushrooms are scaly.

2. (Correspondence analysis) For the following contingency table:

1	4	7
2	5	8
3	6	9

we want to check if the optimal row scores are (1.5, 0, -1) and if the optimal column scores are (-0.25, 0, 6.25). Show your reasoning.

   (a) Is the mean of the row scores over the dataset correct?

   (b) Is the average row score for column 1 correct?

3. (Size of effects) The file `media.dat` contains the results of a questionnaire on where people get their news. The media categories are:

N_ NEWS	national newspaper
R_ NEWS	regional newspaper
MAGAZ	magazines
TVMAG	TV magazines
TV	TV news
RADIO	radio news

The job categories are:

h_ manag	high-level manager
i_ manag	intermediate-level manager
empl	employer
s_ busin	small business employee
skil	skilled labor
unsk	unskilled labor
la_ Farmer	farmer
Nowork	unemployed

(a) Download the `crosstab` package described in `day12.html`. Sort the table and make a mosaic plot, with "job" as the conditioning variable (columns). You do not have to shade the cells (though it may help interpreting the table).

(b) Describe the trend for national newspapers vs. regional newspapers.

(c) Compute the expected counts and Pearson residuals for all cells in the table. (You don't have to report them.) What cell has the most positive residual? What cell has the most negative residual? (`sort.cells` may be useful.)

(d) Compute a 68% confidence interval of the lift for all cells in the table. (You don't have to report them.) What cell has the largest lower bound on the lift? What cell has the smallest upper bound on the lift?

(e) Explain any discrepancies between the answer to part (c) and part (d).

(f) For the cell with the largest lift, explain what the lift value means about where people get their news.

(g) For the cell with the smallest lift, explain what the lift value means.