36-350: Data Mining

Handout 21
November 5, 2003

Mining contingency tables

The trick for contingency tables is to look at them as probability tables, with associated error bars. A line chart shows if the variables are dependent, just like a χ^2 test, and also shows the form of the dependence. Judging independence is exactly like judging if there is an interaction term in a regression.

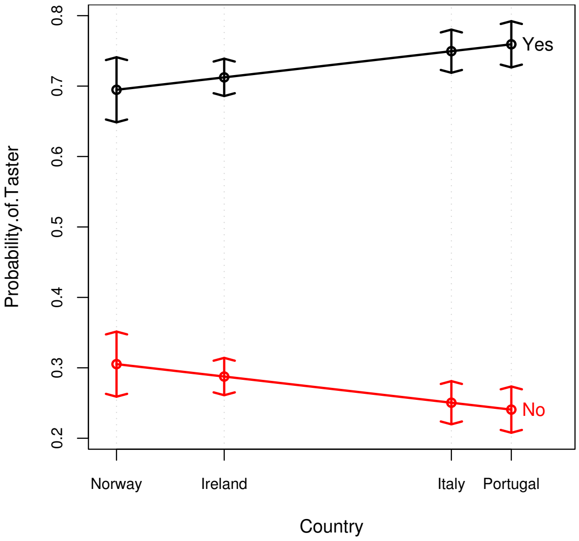
Depicting error in probability estimates

- Standard error $= \sqrt{p(1-p)/n}$
- Error bar = 1.64 × Standard error (for 95% confidence in a bar-to-bar comparison)

Number of people who can taste PTC (Moore&McCabe exercise 9.23):

```
      Country
Taster Ireland Portugal Norway Italy
   Yes 558     345      185    402
   No  225     109       81    134
```

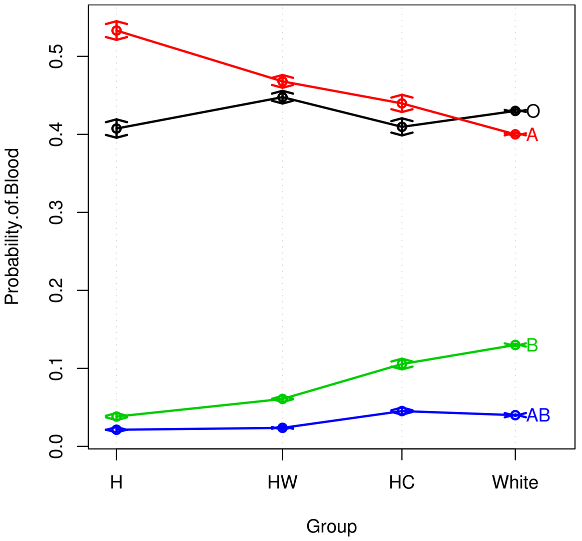
Countries are grouped automatically by the linear profiles method. Since the probability of `Taster` does not vary across country, the variables are independent. (p-value is 0.11.)



Blood types in Hawaii (Moore&McCabe exercise 9.24):

```
     Group
Blood H     HW    HC    White
   O   1903  4469  2206 53759
   A   2490  4671  2368 50008
   B    178   606   568 16252
   AB    99   236   243  5001
```

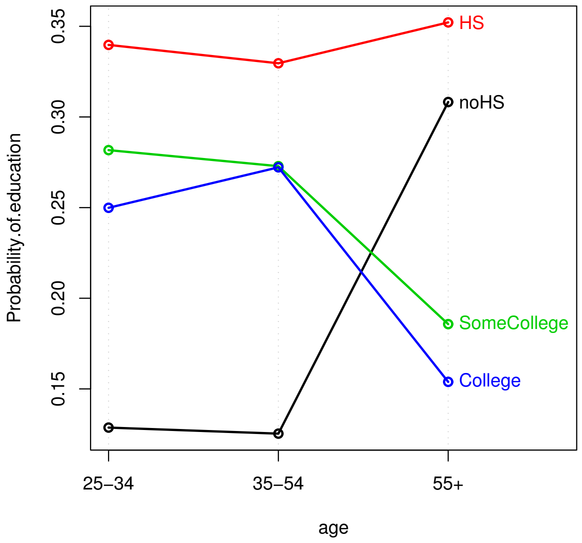
There is a dependence, and we can see the trends. (p-value is 0.)

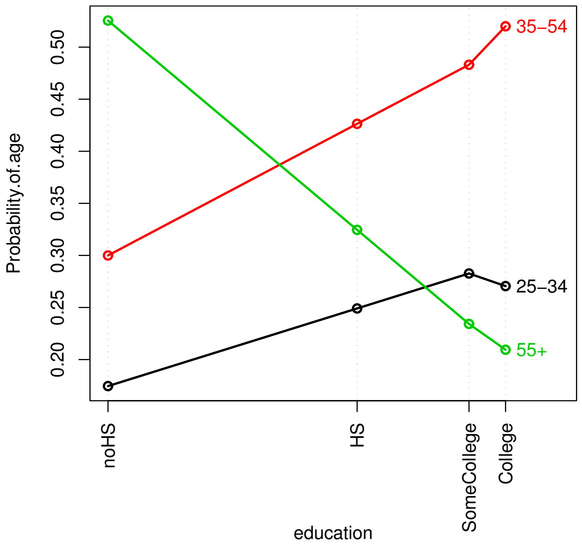


Contingency tables usually arrive with too many categories, and need to be simplified to be understood. Line charts suggest categories to merge. If two categories of X give the same probability for Y, then there is no need to keep them separate. For example, in the table below, we can merge ages whose probability of education is the same, or education levels whose probability of age is the same. In the line chart, look for flatness in the profiles, or use the linear profiles method to automatically cluster the categories.

Americans, 1995 (in thousands) (Moore&McCabe table 2.14):

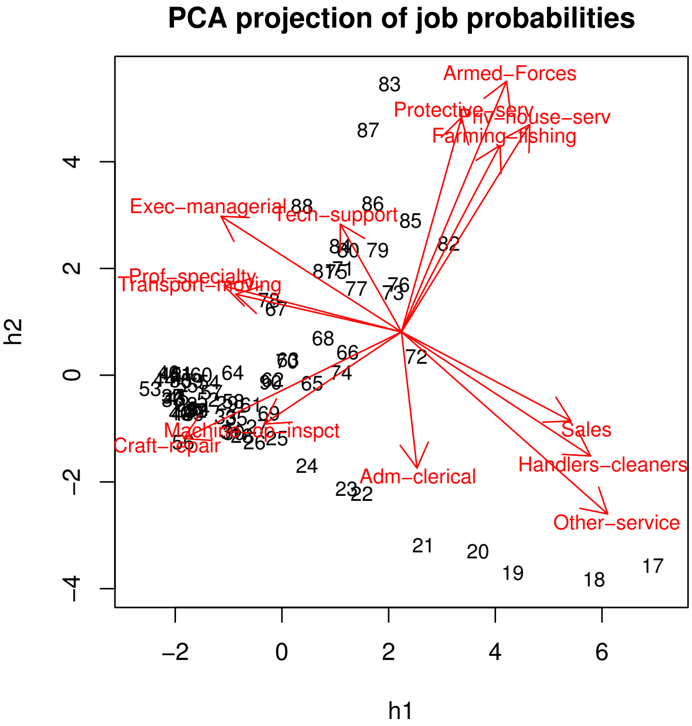
```
             age
education     25-34 35-54 55+
  noHS         5325  9152 16035
  HS          14061 24070 18320
  SomeCollege 11659 19926  9662
  College     10342 19878  8005
```



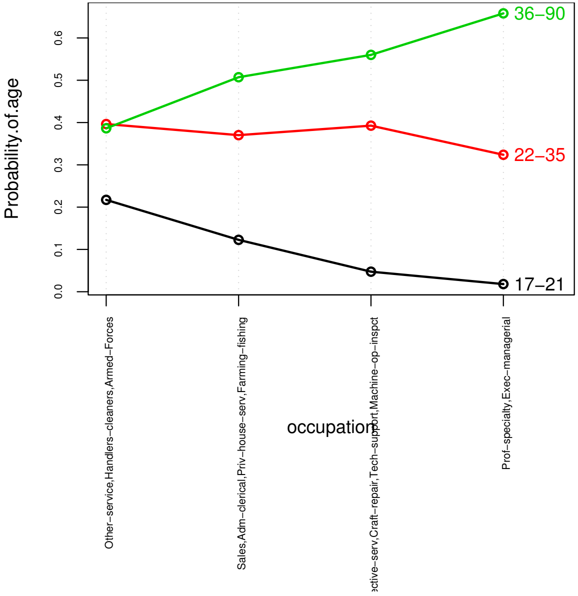


Another trick for contingency tables is to apply ideas from retrieval (weeks 1 and 2), particularly clustering rows/columns by Euclidean distance (after normalization).

Consider a contingency table of people cross-classified by 14 jobs and 73 ages. Each age is described by a vector of job probabilities, and each job is described by age probabilities.



Cluster the jobs, then the ages, and repeat until the merging trace says to stop. The result is 4 job groups and 3 age groups:



Museum patron profiling

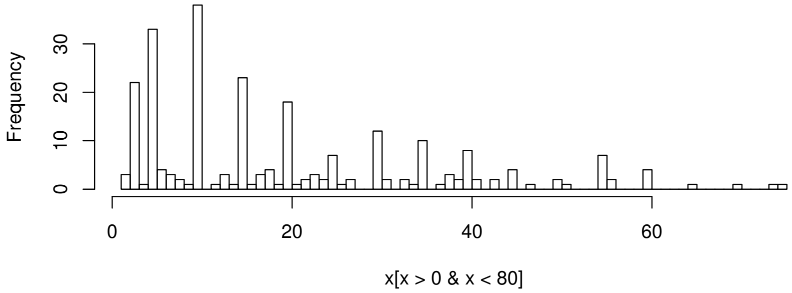
We want to profile visitors to a museum according to how much time they spend at exhibits. (The data came from a project on "smart museums" at MIT.)

According to museum theory, visitors tend to be either "busy," "greedy," or "selective." Busy visitors spend a small amount of time at each exhibit. Greedy visitors spend a lot of time at every exhibit. Selective visitors spend time at a small set of exhibits. It would be useful to automatically classify visitors into these three types so that the museum can adapt its exhibits, either on a time-of-day or day-of-week schedule or dynamically as visitors move through the museum.

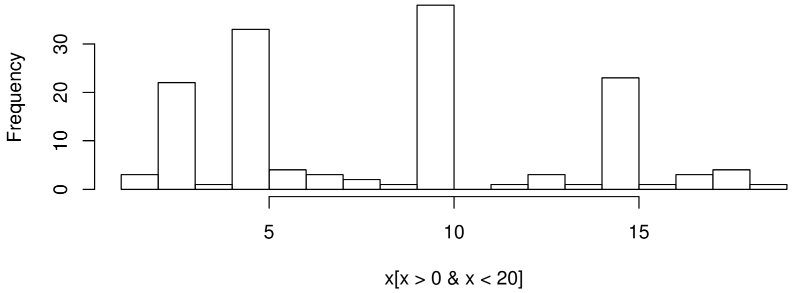
Time durations have been recorded for 50 visitors and 12 different exhibits, giving a data matrix with 50 rows and 12 columns. In this data, all cells have been filled in, but in practice we won't necessarily have time measurements for each visitor at all exhibits. How can we classify visitors?

If we abstract the time durations into "zero," "short," and "long," then we can make a three-bin probability histogram for each visitor class. A new visitor can be classified by comparing their exhibit times to the visitor class distributions. This is the same method we used to classify text and images.

The remaining question is how to define "short" and "long." To start, consider a histogram of all time measurements:



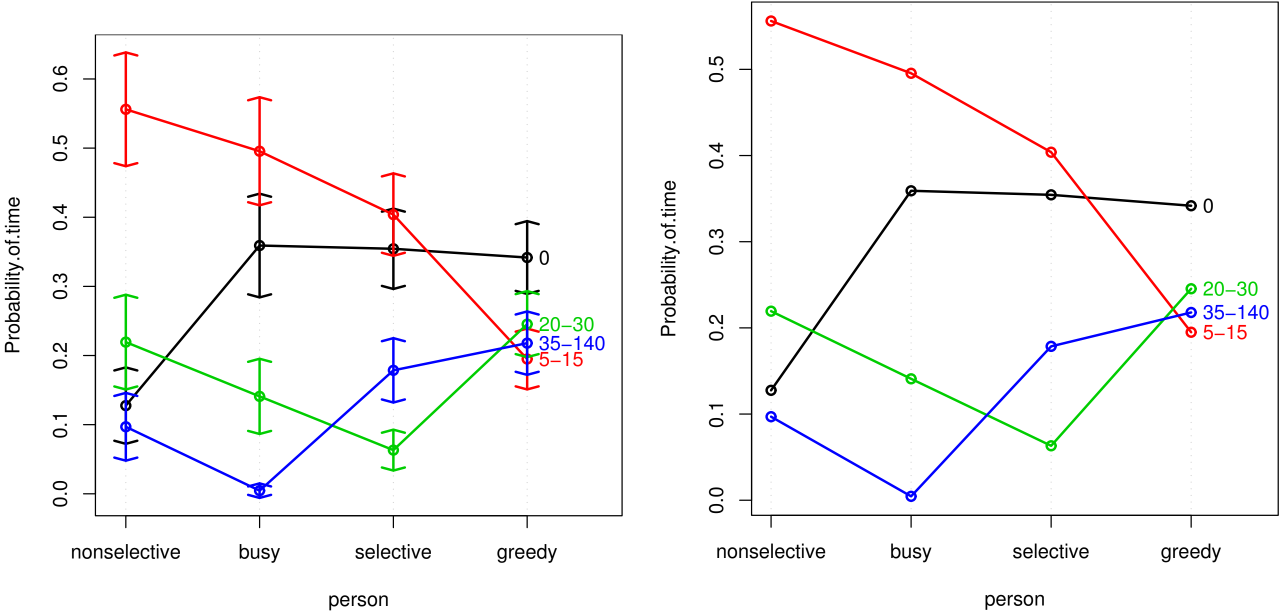
It appears that there are clusters around 10, 30, and 55 seconds. However, the histogram has an unusual spiky structure, which is clearer when we zoom in:



Apparently, some of the people recording the times rounded them to multiples of five, and some did not. We wouldn't have known this without looking at the data. The only way to fix this is to round all of the data to multiples of five.

Now we have a contingency table of 50 rows and 20 columns which reports, for each customer and each time duration (in multiples of 5 seconds), the number of times the customer spent that much time at an exhibit. In other words, each row is a histogram of durations for a given customer.

Reducing the number of time categories is exactly the type of merging we did above, and can be done by clustering the columns. Then cluster the rows to get patron groups. The merging trace suggests four groups for each, leading to a 4 × 4 reduced table:



This shows that there are actually four customer groups, not three. (All remaining differences are significant.) A "nonselective" person skips exhibits less frequently, usually spends a short amount of time, but sometimes a long time at an exhibit. The remaining groups fit the canonical "busy", "selective", and "greedy" archetypes. They skip exhibits at roughly the same rate. Busy people never spend a long amount of time, selective people rarely spend a medium amount of time, and greedy people spend unusually long amounts of time.

References

[1] David S. Moore and George P. McCabe. *Introduction to the Practice of Statistics, 3rd ed.* W. H. Freeman and Company, 1999.

[2] Thomas Minka. "Judging significance from error bars," CMU Tech Report, 2002. `http://www.stat.cmu.edu/~minka/papers/minka-errorbars.pdf`