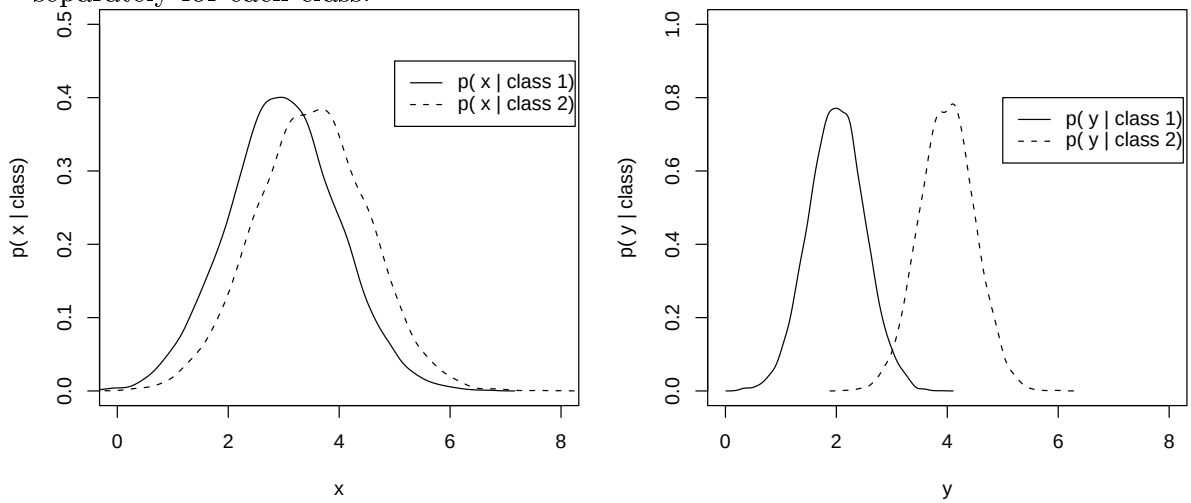
36-350: Data Mining

Homework 3
Date: September 9, 2001 **Due: start of class September 16, 2001**

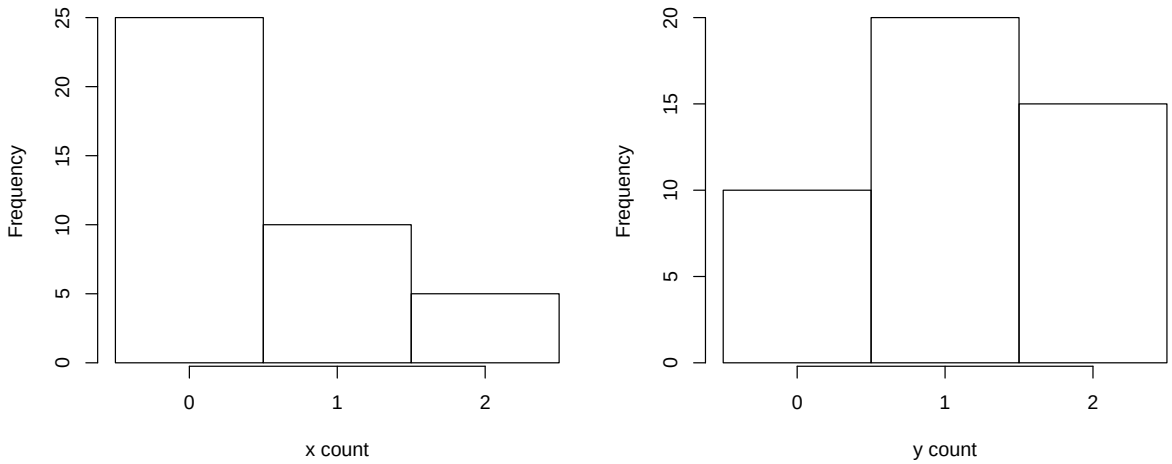
---

1. Suppose there are two colors x and y, and you want to know which is more discriminative of two known image classes. Each image has a count for x and a count for y, which vary between images. The distribution of the count for x and for y is depicted below, separately for each class:



   Which color is more discriminative, and why?

2. Suppose there are two words x and y, and you want to know which is more likely to discriminate documents into subgroups (the subgroups are as yet unknown). Each document has a count for x and a count for y. The distribution of the count for x, over the entire dataset, is depicted below, along with the distribution of the count for y.



(a) Which word has higher "document frequency" (occurs in the most documents)?

(b) Which word has higher entropy? (Hint: which distribution is closer to uniform?)

(c) Which word has higher variance?

(d) Which word is more likely to discriminate?

3. Consider the following subtable for "suddenly":

```
> subtable(xp,"auto","suddenly")
          suddenly not suddenly
auto             0          611
not auto         2          694
```

Under the independent-word assumption, the next word will be "suddenly" with a certain probability p, which may be different for the two classes.

(a) Estimate the probability of "suddenly" in "auto" and "not auto". To avoid problems with small counts, add 0.5 to every cell in the table first.

(b) What is the ratio $p(\text{suddenly}|\text{auto})/p(\text{suddenly}|\text{not auto})$? This is the factor by which the odds of being "auto" increases after observing 'suddenly'.

4. In lab, you found two words where the chi-square measure and odds-ratio measure disagreed about the words' relevance. For each word, based on its subtable of counts, explain why the measures disagree.

5. In lab, you found a word where the odds-ratio and hedged odds-ratio measures disagreed about the word's relevance. For each word, based on its subtable of counts, explain why the measures disagree.

6. In order to use the chi-square or odds-ratio measure for finding discriminative colors, we had to assume that pixels in an image are colored independently within each class. Explain why this is an unrealistic assumption.

7. Suppose that the independent-word assumption is not true. Is it possible for a word to be discriminative, even though it occurs the same number of times in each class? For example, if the subtable of counts is

```
          present   not present
class 1        31           323
class 2        31           323
```

(Hint: think of other properties of a distribution besides its mean.)