Relationships of Categorical Variables: Who Survived on the Titanic?

We'll look at information for 2201 passengers and crew on the *Titanic*. For each person, the following is known:

```
class:      first, second, third, or crew
age:        adult or child
sex:        female or male
survived:   no or yes
```

This data may contain errors. For more background, see the article by Robert Dawson in the on-line *Journal of Statistics Education*, volume 3, number 3, available at

```
http://www.amstat.org/publications/jse/
```

The data is available as a worksheet in the "lecture-data" directory.

Distributions for each Variable

Before looking at relationships between variables, we can look at the distribution of each variable separately.

Here are tables for the *Titanic* data, giving counts and percentages:

```
          CLASS          |          AGE
                         |
        count   %        |        count   %
                         |
   1st    325  14.77     | adult  2092  95.05
   2nd    285  12.95     | child   109   4.95
   3rd    706  32.08     |
   crew   885  40.21     |    ALL 2201 100.00
                         |
    ALL 2201 100.00      |
-----------------------------------------------
           SEX           |       SURVIVED
                         |
        count   %        |       count   %
                         |
female   470  21.35      |  no   1490  67.70
male    1731  78.65      |  yes   711  32.30
                         |
    ALL 2201 100.00      |   ALL 2201 100.00
```

Two-Way Relationships: Age and Sex

Here is a *two-way table*, showing how age and sex are related. In terms of counts:

```
ROWS: age     COLUMNS: sex

       female    male      ALL

adult     425    1667      2092
child      45      64       109

ALL       470    1731      2201
```

In terms of percentage of all cases:

```
ROWS: age     COLUMNS: sex

       female    male      ALL

adult   19.31   75.74    95.05
child    2.04    2.91     4.95

ALL     21.35   78.65   100.00
```

The distributions for each variable separately are given in the margins: Hence they are sometimes called *marginal* distributions.

Conditional Distributions: Looking at Sex Given Age

We can highlight the different proportions of males and females among children and adults by looking at the *conditional distribution* of sex given age (child or adult):

```
ROWS: age     COLUMNS: sex

       female    male      ALL

adult   20.32   79.68   100.00
child   41.28   58.72   100.00

ALL     21.35   78.65   100.00
```

Why are more children boys (58.72%) than girls (41.28%)? Could it just be chance?

When all conditional distributions are the same, the variables are said to be *independent*. Knowing one tells you nothing about the other. (Even for independent variables, percentages in a sample are usually not *exactly* the same.)

Did Social Class Affect Survival?

Let's look at the conditional distributions for survival given class:

```
ROWS: class     COLUMNS: survived

         no      yes      ALL

1st    37.54    62.46   100.00
2nd    58.60    41.40   100.00
3rd    74.79    25.21   100.00
crew   76.05    23.95   100.00

ALL    67.70    32.30   100.00
```

Why did more of the upper class survive?

- Crew discriminated against lower-classes?
- Lower-class people are not too bright, and tend to panic?
- Other explanations?

Why was the proportion of the crew who survived so low?

Did the Crew Sacrifice Themselves?

Before concluding that the low survival rate for the crew was due to their saving passengers, let's look at a possible "lurking" variable: sex.

```
CONTROL: sex =  female
ROWS: class     COLUMNS: survived

          no      yes      ALL

1st     2.76    97.24   100.00
2nd    12.26    87.74   100.00
3rd    54.08    45.92   100.00
crew   13.04    86.96   100.00

ALL    26.81    73.19   100.00

CONTROL: sex =  male
ROWS: class     COLUMNS: survived

          no      yes      ALL

1st    65.56    34.44   100.00
2nd    86.03    13.97   100.00
3rd    82.75    17.25   100.00
crew   77.73    22.27   100.00

ALL    78.80    21.20   100.00
```

How is This Reversal Possible? Simpson's Paradox

The proportions of the crew who survived was greater than the proportion for third-class passengers for *both* males and females, but *overall* it was less.

We can see why by looking at the conditional distribution of sex given class:

```
ROWS: class     COLUMNS: sex

     female      male      ALL

1st   44.62     55.38   100.00
2nd   37.19     62.81   100.00
3rd   27.76     72.24   100.00
crew   2.60     97.40   100.00

ALL   21.35     78.65   100.00
```

Does this information alter your interpretation of any of the other comparisons?