# CSSS/SOC/STAT 536:
# Logistic Regression and Log-linear Models

# Introduction to Contingency Tables

Christopher Adolph*
University of Washington, Seattle

November 16, 2006

*Assistant Professor, Department of Political Science and Center for Statistics and the Social Sciences.

# Plan for today, and next week

We're still talking about count data

But from today, we will restrict our attention to data that can be put in tables

$\rightarrow$ datasets with *only* categorical variables

Methods so far have been general enough to handle continuous and categorical data, so they still apply

But we will develop special techniques—and a different language—for tabular data

These methods are used often in medicine and sociology

# Plan for today, and next week

Today we introduce tables for count data, called *contingency tables*

Main goal today is to get a handle on the language, to aid your reading

The language may seem very different from what you are accustomed to, but connections will emerge

Next week, we'll talk about regression models for contingency tables

# Contingency tables

Let's start with the simplest possible table, a $2 \times 2$

|  | $Y$ |  | Sum |
|---|---|---|---|
| $X$ | **1** | **2** |  |
| **1** | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| **2** | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| Sum | $n_{.1}$ | $n_{.2}$ | $n$ |

In each cell, we have a count, $n_{ij}$

The sum over any given row, $i$, is $\sum_j n_{ij} = n_{i.}$

The sum over any given column, $j$, is $\sum_i n_{ij} = n_{.j}$

The grand sum is simply $n$

Note we haven't said anything about independent and dependent variables

# Contingency tables

| | $Y$ | | Sum |
|---|---|---|---|
| $X$ | **1** | **2** | |
| **1** | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| **2** | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| Sum | $n_{.1}$ | $n_{.2}$ | $n$ |

Suppose that any given observation could fall into any cell

Another way of saying this is that *ex ante*, the row and column sums are not fixed but random

If we assume each of the cells is made up of iid draws from some discrete distribution, we get . . .

# $2 \times 2$ Tables: Rows and Columns free

| | $Y$ | | Sum |
|---|---|---|---|
| $X$ | **1** | **2** | |
| **1** | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1.}$ |
| **2** | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2.}$ |
| Sum | $\pi_{.1}$ | $\pi_{.2}$ | $1.0$ |

The probability of an observation falling into cell $i, j$ is $\pi_{ij}$

$\{\pi_{ij}\}$ is the joint distribution of the data

Marginal distributions are sums over rows or columns

$$\pi_{i.} = \sum_j \pi_{ij} \qquad \pi_{.j} = \sum_i \pi_{ij}$$

The conditional distribution—probability of falling in $j$ given being in $i$—is

$$pi_{j|i} = \pi_{ij}/\pi_{i.}$$

# Probability distributions for contingency tables

What would be an appropriate probability distribution for Table 1?

Let's assume the total $n$ is unbounded

and the events in each cell are iid

The Poisson! We'll talk about this at greater length next week

Now let's consider another case

# $2 \times 2$ **Tables: Rows fixed, Columns free**

| | $Y$ | | Sum |
|---|---|---|---|
| $X$ | **1** | **2** | |
| **1** | $\pi_{1\|1}$ | $\pi_{1\|2}$ | $1.0$ |
| **2** | $\pi_{2\|1}$ | $\pi_{2\|2}$ | $1.0$ |
| Sum | $\pi_{.1}$ | $\pi_{.2}$ | $1.0$ |

Suppose the row marginals of our table are fixed ahead of time

We might be conducting an experiment, placing half our subjects in a control group, and half in a treatment group

In this case, $Y$ is a response variable

Cell entries are conditional probabilities, $\pi_{j|i}$, and sum to $1$ by rows

# Probability distributions for contingency tables

Table 2 (rows fixed and columns free) should remind you of the multinomial logit model

In fact, you could rewrite any such table to be an MNL, by breaking each cell up into its constituent individuals

In tabular form, each cell is a sum of trials, but there could be more than two outcomes in a row (or table, if only $n$ is fixed, and not $n_{i.}$)

So we'll need something like the binomial, but able to handle more than just two outcomes

# Probability distributions for contingency tables

In this case, an appropriate distribution is the multinomial, which is a generalization of the binomial to $k$ categories

Recall the binomial

$$f_{Bin}(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{N-y}, \quad y = 0, 1, 2, \ldots, n$$

$$\mathrm{E}(y) = n\pi \qquad \mathrm{Var}(y) = n\pi(1-\pi)$$

The multinomial is quite similar, but allows different $\pi$'s for each category

$$f_{Multin}(y_j|\pi_j) = \binom{n}{y_1 \cdots y_k} \pi_1^{y_1} \cdots \pi_k^{y_k}, \quad y_j = 0, 1, 2, \ldots, n, \quad \sum_j y_j = n$$

$$\mathrm{E}(y_j) = n\pi_j \qquad \mathrm{Var}(y_j) = n\pi_j(1-\pi_j)$$

# Independence of $X$ and $Y$

We aren't going to to go too far with the Multinomial

We'll use it to consider one simple hypotheses:

That Y does not depend on X, or vice versa

We say that $X$ and $Y$ are independent if knowing $X$ doesn't change $\mathrm{E}(Y)$

Formally, two categorical variables are independent if

$$\pi_{ij} = \pi_{i.}\pi_{.j} \qquad \forall i, j$$

In other words,

$$\begin{aligned} \pi_{j|i} &= \pi_{ij}/\pi_{i.} \\ &= \pi_{i.}\pi_{.j}/\pi_{i.} \\ &= \pi_{.j} \end{aligned}$$

The conditional prob that $Y = j$ is just the marginal probability that $Y = j$; $X = i$ is irrelevant

# Testing Independence

Suppose we have a $k$-vector drawn from the multinomial distribution.

$$n = \{n_1, \ldots, n_k\}, \qquad \text{with} \quad \sum_i n_i = n$$

The multinomial draws a series of binomial variables given a series of probabilities. Let the vector $\pi_{i0}$ be our null hypothesis

$$H_0 : \pi_i = \pi_{i0}$$

which will try to reject in favor of any other vector of probabilities
We can choose several test statistics:

$$\text{Pearson } X^2 = \sum_{i=1}^{k} \frac{(n_i - \mu_i)^2}{\mu_i}$$

where $\mu_i$ is our expectation under the null, $\mu_i = n\pi_{i0}$

If $X^2$ is true and $n$ is large, then $X^2 \sim \chi^2(k-1)$,
$\rightarrow$ we reject independence for large $X^2$.

# Testing Independence

Alternatively, we can use a likelihood ratio test for independence

We calculate

$$\mathrm{LR} = \frac{\text{maximum likelihood under } H_0}{\text{maximum likelihood under any other } \boldsymbol{\pi}}$$

which is commonly transformed to

$$\begin{aligned} G^2 &= -2 \ln \mathrm{LR} \\ &= 2 \sum_{i-1}^{k} n_i \ln \frac{n_i}{\mu_i} \\ &\sim \chi^2(k-1) \end{aligned}$$

Note that $G^2 = 0$ when the model is a perfect predictor $(n_i = \mu_i)$

Any other model is worse $(G^2 > 0)$

$G^2$ is often called the *deviance*

# Testing Independence

Let's apply these test statistics to a $2 \times 2$ table, under independence
(The same formulas will work for $I \times J$ tables)

Under independence, the MLE of the cell entries is

$$\begin{aligned} \hat{\mu}_{ij} &= n\hat{\pi}_{i.}\hat{\pi}_{.j} \\ &= \frac{n_{i.}n_{.j}}{n} \end{aligned}$$

We plug this into the formulas for $X^2$ and $G^2$.

The degrees of freedom for the test is $n - p - 1$, where $p$ is the number of parameters

In the $2 \times 2$ case, $\mathrm{d.f.} = 4 - 2 - 1 = 1$

(We've estimated one row probability and one column probability)

# Testing Independence

If we apply these methods to the a dataset on cancer

| | Eventual result | | |
|---|---|---|---|
| PCR reading | Relapse | No relapse | Sum |
| Traces of cancer | 30 | 45 | 75 |
| No cancer | 8 | 95 | 103 |
| Sum | 38 | 140 | 178 |

Leukemia patients were tested for precursors of relapse using polymerase chain reaction (PCR); three years later, their health was recorded

$n = 178$ is fixed, so we assume the cells follow a multinomial

Research question: Did the PCR predict cancer status? Or are they independent?

(Source: Simonoff 2003, *Analyzing Categorical Data*; an accessible supplement to your readings in Agresti)

# Testing Independence

We're assuming independence, so forget the cell entries; we just need the marginals

| | Eventual result | | |
|---|---|---|---|
| PCR reading | Relapse | No relapse | Sum |
| Traces of cancer | | | 75 |
| No cancer | | | 103 |
| Sum | 38 | 140 | 178 |

# Testing Independence

We need the marginal probabilities, so divide by $n = 178$

| | Eventual result | | |
|---|---|---|---|
| PCR reading | Relapse | No relapse | Sum |
| Traces of cancer | | | 0.42 |
| No cancer | | | 0.58 |
| Sum | 0.21 | 0.79 | 1 |

# Testing Independence

Under independence, the MLE of the cells is the product of the marginal probabilities

| | Eventual result | | |
|---|---|---|---|
| PCR reading | Relapse | No relapse | Sum |
| Traces of cancer | 0.09 | 0.33 | 0.42 |
| No cancer | 0.12 | 0.46 | 0.58 |
| Sum | 0.21 | 0.79 | 1 |

# Testing Independence

Finally, we multiply by $n = 178$ to get expected cell counts under independence

| | Eventual result | | |
|---|---|---|---|
| PCR reading | Relapse | No relapse | Sum |
| Traces of cancer | 16.01 | 58.99 | 75 |
| No cancer | 21.99 | 81.01 | 103 |
| Sum | 0.21 | 140 | 178 |

Compare to the data:

| | Eventual result | | |
|---|---|---|---|
| PCR reading | Relapse | No relapse | Sum |
| Traces of cancer | 30 | 45 | 75 |
| No cancer | 8 | 95 | 103 |
| Sum | 38 | 140 | 178 |

# Testing Independence

Is independence a good fit?

$$
\begin{aligned}
G^2 &= 2\sum_{i-1}^{k} n_i \ln \frac{n_i}{\mu_i} \\
&= 27.40 \qquad p < 0.0000\ldots
\end{aligned}
$$

No; the deviance test rejects independence.

PCR did predict cancer recurrence

But then, one could tell just looking at the data and predictions . . .

# Testing Independence

Similar tests exist for tables with no fixed sums, or with fixed rows.

But to check for independence in tables with fixed rows and columns, we'll need a different method.

# The Lady Tasting Tea

R.A. Fisher, the famous statistician, had a friend who claimed to be able to tell by taste whether tea had been added to milk, or milk to tea

Fisher ran an experiment, with 8 cups of tea, to see if this was true

Suppose the result was

| | Guess | | |
|---|---|---|---|
| Truth | Tea first | Milk first | Sum |
| Tea first | 3 | 1 | 4 |
| Milk first | 1 | 3 | 4 |
| Sum | 4 | 4 | 8 |

(what is fixed in this table? How many free variables are there?)

# Fisher's Exact Test

To see whether his friend's choices were more unusual than picking at random, Fisher proposed a test of *independence*.

He showed that a $2 \times 2$ table with fixed rows and columns is distributed *hypergeometric*

$$\mathrm{P}(n_{11} = x) = \frac{\binom{n_{1.}}{x}\binom{n_{2.}}{n_{.1}-x}}{\binom{n}{n_{.1}}}$$

writing out the combinatorics, we get

$$\mathrm{P}(n_{11} = x) = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

Note this doesn't depend on any random parameters, just $n$, which is known

That makes this an "exact" test

The $p$-value we calculate is valid even for tiny samples;
it does not depend on asymptotic assumptions

Tends to be conservative

# Fisher’s Exact Test

So how unlikely was the taster’s performance? What is the probability it would have happened by chance?

$$\begin{aligned} \mathrm{P}(n_{11} \geq 3) &= \mathrm{P}(n_{11} = 3) + \mathrm{P}(n_{11} = 4) \\ &= \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} \\ &= 0.229 + 0.014 \\ &= 0.243 \end{aligned}$$

Note that the test takes discrete jumps as the number of events increases

(It’s unknown how many cups Fisher’s friend correctly classified;
we do know Fisher was convinced, given his test)

# Fisher's Exact Test

Fisher's Exact Test extends to larger tables $(I \times J)$, but not easily.

We make use in this case of a generalized version of the geometric distribution, known as the multiple hypergeometric

One can still order all possible tables for most to least likely, then compute the probability of a table less likely than the realized table

Even with modern computing, this quickly gets unmanagable

Monte Carlo is one solution

1. randomly sample tables
2. compute their probability under the multiple hypergeometric distribution
3. and rank your table against the sorted draws

# The need for models of contingency tables

So why all the emphasis on independence?

Independence not usually an interesting question

An exception: controlled experiments

But generally, it's more interesting to know

- effect sizes
- contextual effects
- predicted values

The usual quantities of interest we get from modeling.

Modeling will tend to involve larger and/or higher dimensional tables

Like any modeling exercise, models of tabular data may be misleading if misspecified

# 3+ Dimensional Tables: $I \times J \times K \times \ldots$

Tables can have more than two dimensions.

We can "nest" rows or columns to show the further dimensions

Or we can collapse over dimensions to reduce back to an $I \times J$ table

It is customary to warn students of the dangers of this move.

. . . Though as social scientists, I suspect the hazards will come as no surprise

# An example: Discrimination?

Suppose the (fictional) University of Tlon is sued for discriminatory hiring

Both sides stipulate that

- the best candidate can be determined uniquely
- should always be hired
- is equally likely to be male or female

The case turns on whether the University hired male and female candidates at the same rate

# An example: Discrimination?

Here is the data for the university's "eclectic" departments

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |

The plaintiffs point out that in each dept, a greater % of men were hired:

| Departments | Men | | Women |
|---|---|---|---|
| Ancient Egyptian Algebra | 25% | $>$ | 20% |
| Navajo Cryptography | 80% | $>$ | 75% |

# An example: Discrimination?

“But wait!” says the defense. “Look at the *totals*”

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |
| Total | 6 | 13 | 7 | 13 |

“We actually hired *more* women at a higher rate than men!”

The plaintiffs in a lawsuit point out that in each dept, a greater % of men were hired:

| Departments | Men | | Women |
|---|---|---|---|
| Ancient Egyptian Algebra | 25% | > | 20% |
| Navajo Cryptography | 80% | > | 75% |
| Both departments | 46% | < | 54% |

What’s going on here?

# Simpson's Paradox

The Departments are different. Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting this variable

But it is correlated with the outcome *and* with our explanatory variable

This is omitted variable bias, also known as Simpson's Paradox

It's no different from the OVB you already know;
with contingency tables we must still condition on confounding variables

(The linear regression parallel to Simpson's Paradox:
treating a 2-D scatterplot of $Y$ on $X$ as sufficient for all data analysis)

# Presentation and EDA

One hazard of working with tabular data is that it encourages lazy presentation

**Contention:**

For any data table, there exists a superior graphic or sentence, except when

- Presenting *small* quantities of data whose *precise* values are worth comparing (Ex: a striking $2 \times 2$ table; a small linear regression with no interactions)
- Reporting large quantities of data whose precise values are *worth looking-up* (Census tables; election returns; the phonebook)
- Reporting exact data for *replication* (Mostly obsolete in the electronic era, but . . . )

# Presentation and EDA

Devising good graphics takes creativity, attention to the reader, and hard work

Coming up with a good table requires the same things

Don't just slap down rows and columns in arbitrary order

Choose order, dimensions, scale, and nesting to highlight relationships

Avoid excessive digits (even excessive significant digits)

Always consider whether tables are the right choice

Sometimes all you need is a good paragraph

# Titanic Example

A 4-D table groups all persons on the *Titanic* by gender, age, class, and survival

| | Adult | | | | Child | | | |
|---|---|---|---|---|---|---|---|---|
| | Male | | Female | | Male | | Female | |
| | Died | Lived | Died | Lived | Died | Lived | Died | Lived |
| 1st class | 118 | 57 | 4 | 140 | 0 | 5 | 0 | 1 |
| 2nd | 154 | 14 | 13 | 80 | 0 | 11 | 0 | 13 |
| 3rd | 387 | 75 | 89 | 76 | 35 | 13 | 17 | 14 |
| Crew | 670 | 192 | 3 | 20 | — | — | — | — |

What do the “—”’s mean?

What patterns leap out?

Any suggestions for improving the table?

What graphical alternatives to this table could we try?

# Titanic Example: Mosaic plots

Most graphics we've discussed in class are focused on *presentation*

Let's consider a different role for graphics: exploration (EDA)

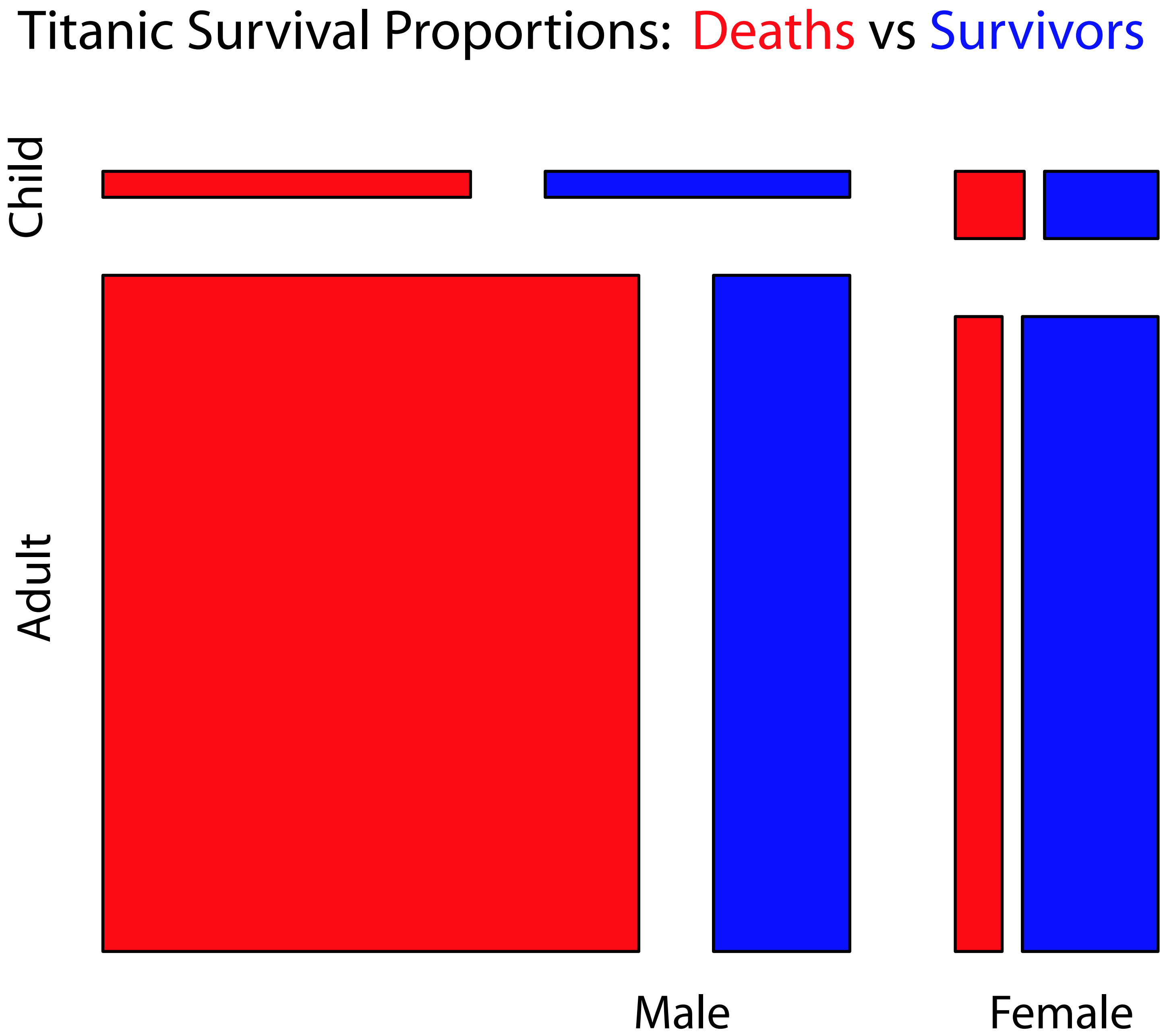
An exploratory tool for contingency tables: Mosaic plots

Mosaic plots draw rectangles for each cell,
with height and width showing the proportions of observations falling in the cell

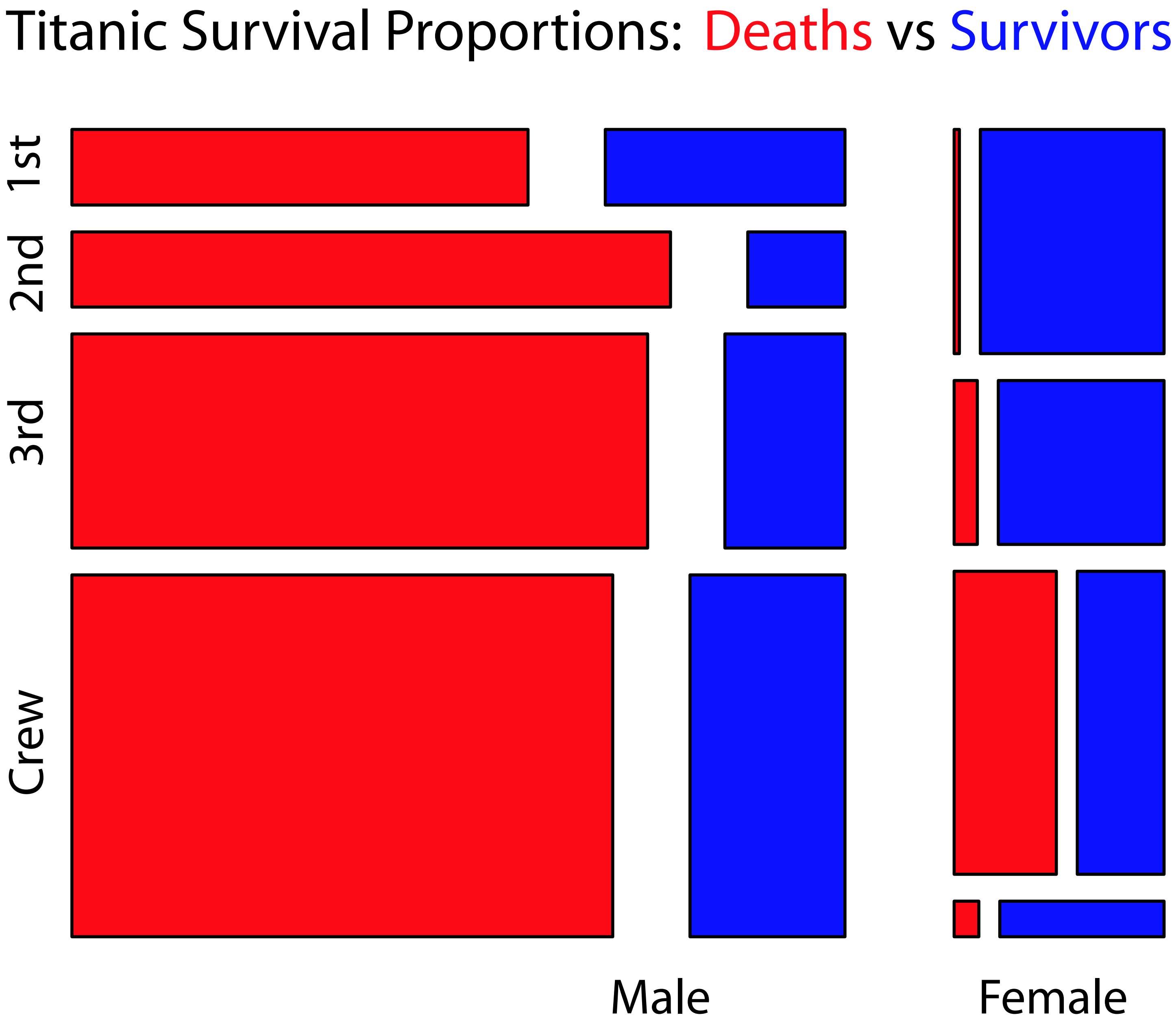
The plots take some effort to understand at first . . .

(How to: `mosaicplot` in the `R` library `graphics`)

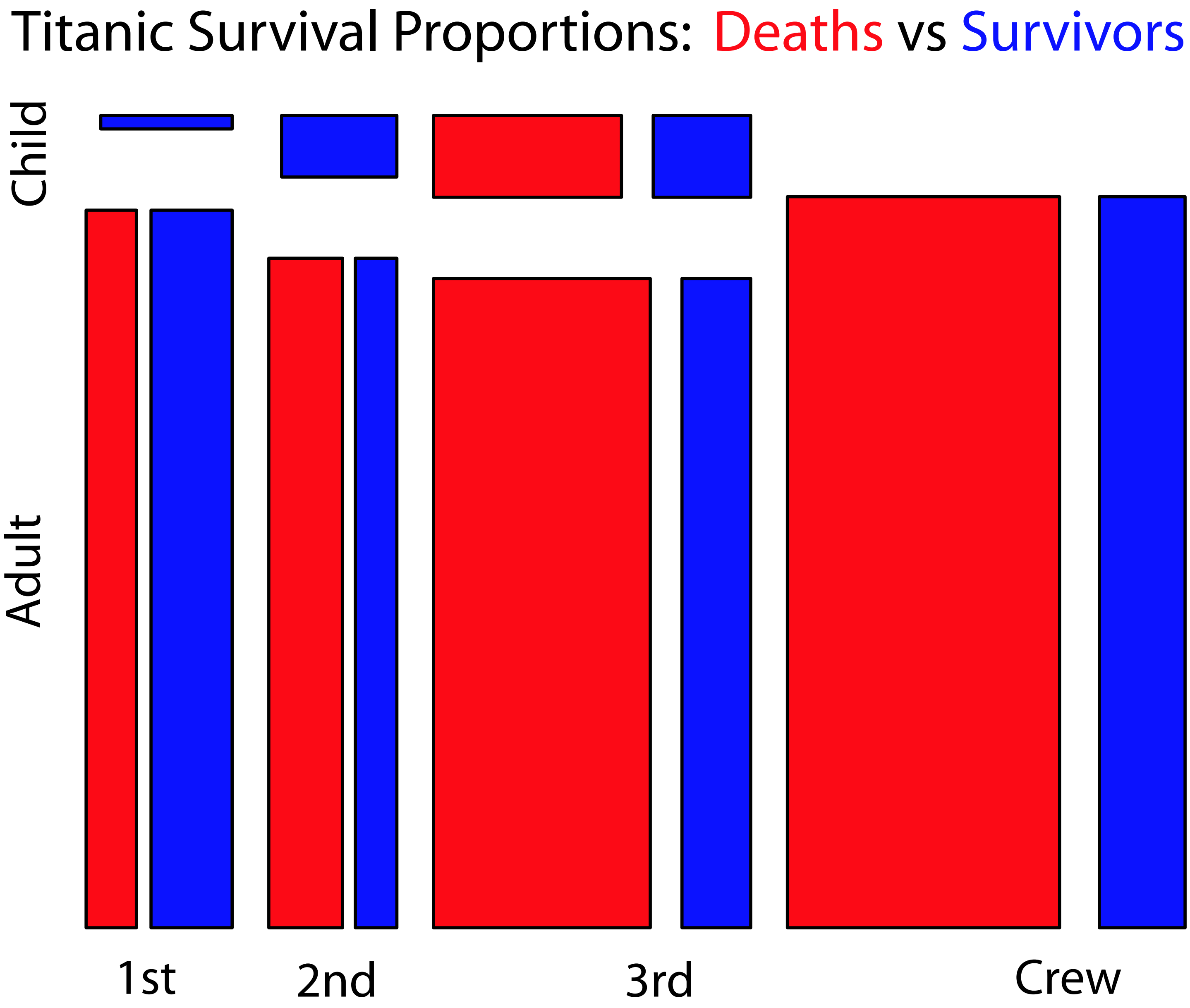
# Mosaic: Age, Gender, and Survival



# Mosaic: Class, Gender, and Survival

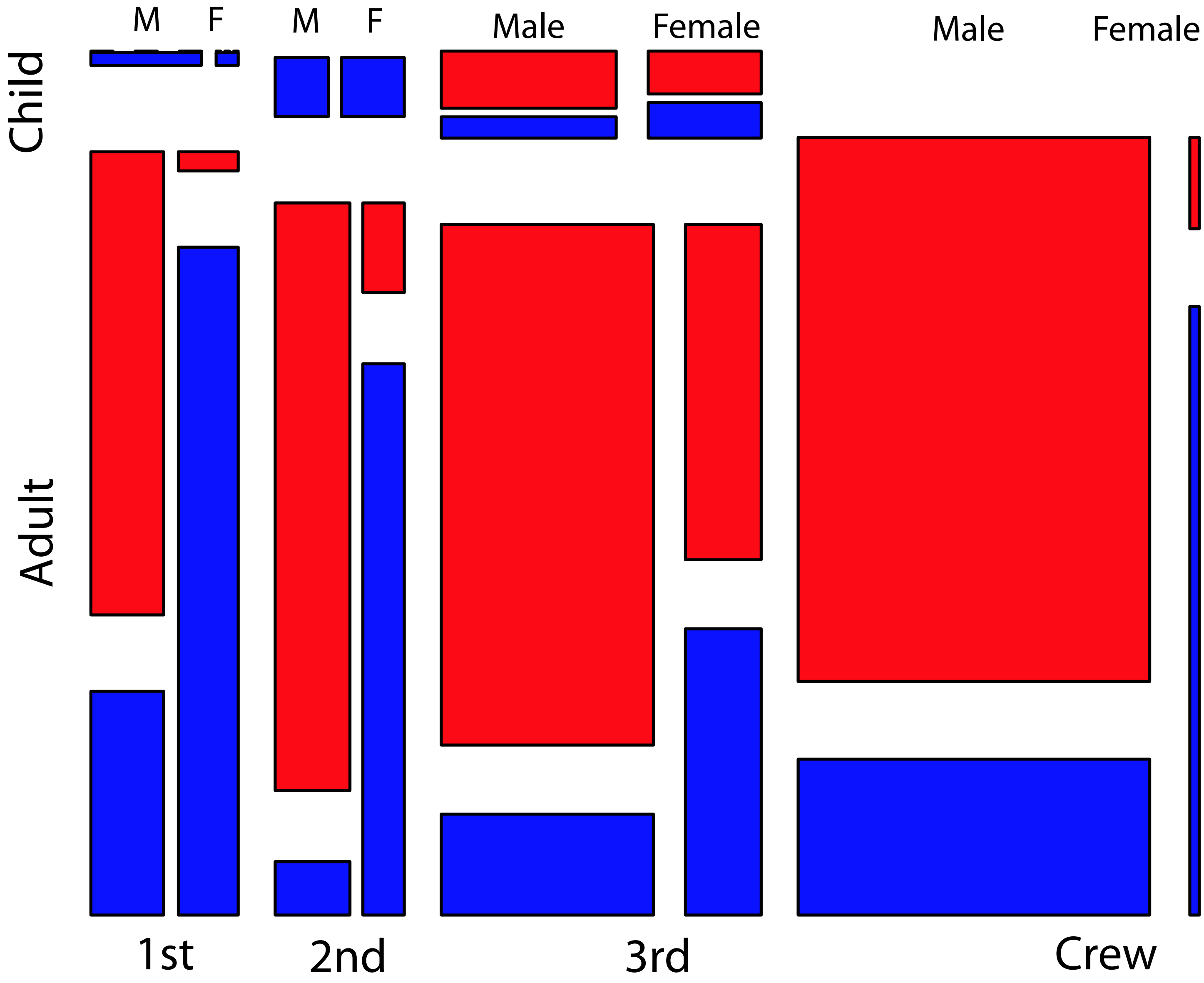


# Mosaic: Age, Class, and Survival



# Mosaic: Age, Class, Gender, and Survival

Titanic Survival Proportions: Deaths vs Survivors


M
F
M
F
Male
Female
Male
Female
Child
Adult
1st
2nd
3rd
Crew