

Displaying & Describing Categorical Data

Chapter 2

Jason Bryer (jason@bryer.org)
epsy530.bryer.org

Titanic Data

```
load("../Data/titanic.Rda")
head(titanic, n = 3)
```

	pclass	survived	name	sex	age	sibsp	parch	
1	First	Yes	Allen, Miss. Elisabeth Walton	female	29.00	0	0	
2	First	Yes	Allison, Master. Hudson Trevor	male	0.92	1	2	
3	First	No	Allison, Miss. Helen Loraine	female	2.00	1	2	

	ticket	fare	cabin	embarked	boat	body	home.dest
1	24160	211.3	B5	S	2	NA	St Louis, MO
2	113781	151.6	C22 C26	S	11	NA	Montreal, PQ / Chesterville, ON
3	113781	151.6	C22 C26	S		NA	Montreal, PQ / Chesterville, ON

- **Who?** People on the Titanic
- **What?** Survival status, class
- **When?** April 14, 1912
- **Where?** North Atlantic
- **How?** [Vanderbilt University](#)
- **Why?** Historical interest

Frequency Table

A frequency table is a table whose first column displays each distinct outcome and second column displays that outcome's frequency.

```
table(titanic$pclass)
```

First	Second	Third
323	277	709

Relative Frequency Table

A relative frequency table (also referred to as a proportional table) is a table whose first column displays each distinct outcome and second column displays that outcome's relative frequency.

```
prop.table(table(titanic$pclass)) * 100
```

First	Second	Third
24.68	21.16	54.16

Contingency Tables

A contingency table is a table that displays two categorical variables and their relationships.

	No	Yes	Total
First	123	200	323
Second	158	119	277
Third	528	181	709
Total	809	500	1309

Marginal Distribution

The distribution of either variable alone is the marginal distribution. In the table above we have the marginal distribution of class on the right column and the marginal distribution of survival on the bottom row.

Table of Percents

```
prop.table(table(titanic$class, titanic$survived)) * 100
```

	No	Yes
First	9.396	15.279
Second	12.070	9.091
Third	40.336	13.827

Conditional Distributions

You need to be careful how you define the percentages. Do the sum of all cells equal 100, or the sum of each column, or the sum of each row.

```
prop.table(table(titanic$pclass, titanic$survived), 1) * 100
```

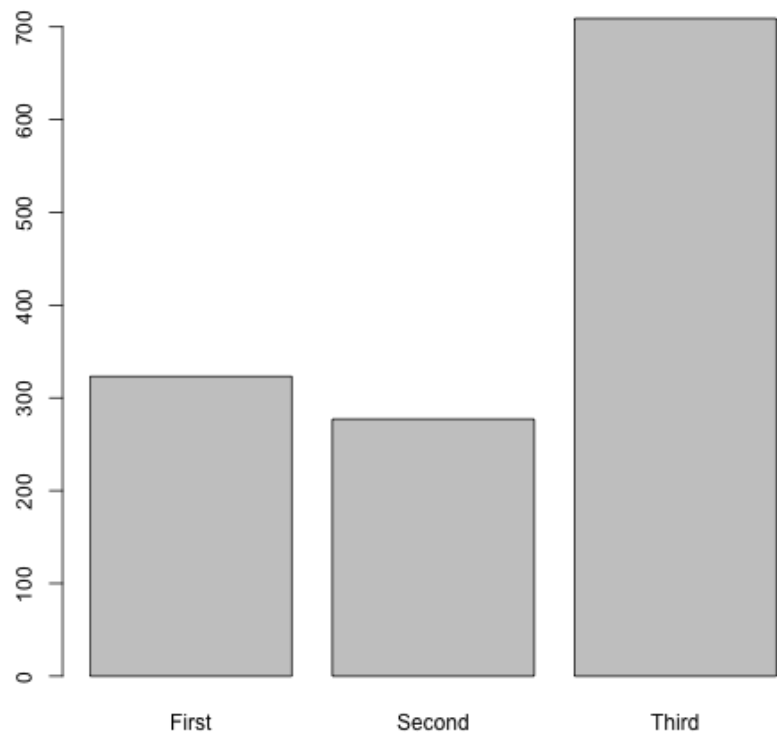
	No	Yes
First	38.08	61.92
Second	57.04	42.96
Third	74.47	25.53

Pie Charts

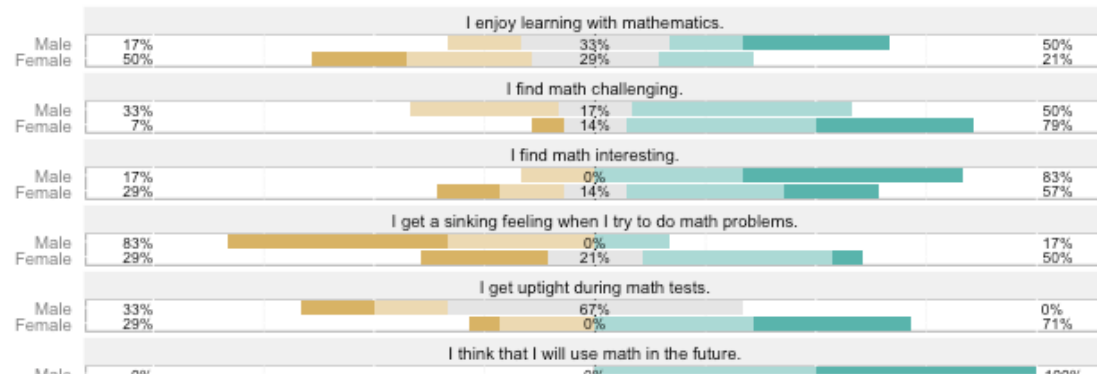
“There is no data that can be displayed in a pie chart, that cannot be displayed BETTER in some other type of chart.”

-- John Tukey


```
plot(titanic$class)
```



Grouped Likert Results



12/15

Simpson's Paradox

Berkeley gender bias case

GENDER	APPLICANTS	ADMITTED
Men	8442	44%
Women	4321	35%

In the above table it appears there is a bias against women. However, including department it appears the bias against women disappears, and in fact there are several advantages for women.

DEPARTMENT	MEN Applicants	MEN Admitted	WOMEN Applicants	WOMEN Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%

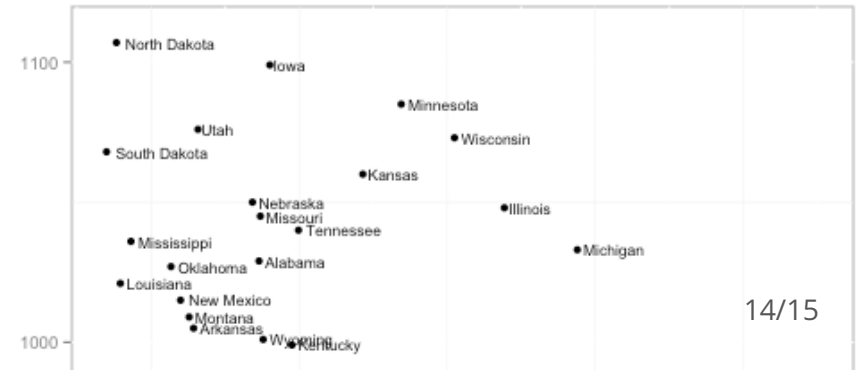
Simpson's Paradox

Teacher salary's and SAT Scores
SAT data including:

- `state` - the state whose SAT score is used.
- `salary` - estimated average annual salary of teachers in public schools in 1994-95 school year (in thousands of dollars).
- `frac` - the fraction of eligible students taking the SAT in 1994-95.

Guber, D.L. (1999), Getting what you pay for: the debate over equity in public school expenditures, *Journal of Statistics Education* 7(2).

See also ?`SAT` for more information.



14/15

Simpson's Paradox

Let's now include the fraction of eligible students who took the SAT.

