Titanic data description

The ship Titanic sank in 1912 with the loss of most of its passengers. Details can be obtained on 1309 passengers and crew on board the ship Titanic. The main use of this data set is Chi-squared and logistic regression with survival as the key dependent variable. Summary statistics for the categorical variables can be demonstrated and the cost of the ticket (fare) is very skewed so it can be used to demonstrate skewed data and differences between means and medians etc.

The titanic data has also been linked to numerous articles in the press including this one:


Australasia

More Britons than Americans died on Titanic 'because they queued'

http://www.independent.co.uk/news/world/australasia/more-britons-than-americans-died-on-titanic-because-they-queued-1452299.html

This is a great example of misleading statistics as nationality is significant with Chi-squared but not after controlling for class in a logistic regression model. Most Americans were in 1st class which was one of the main factors influencing survival.

Variable name	Variable label	Data type	Value labels
pclass	Class	Ordinal	1 = 1st, 2= 2nd, 3 = 3rd
survived		Binary (Nominal)	0 = Died, 1 = Survived
Residence	Country of Residence	Nominal	0 = American, 1 = British, 3 = Other
Name		String	
age		Scale	
sibsp	Number of siblings/ spouses	Scale (Discrete)	
parch	Number of parents/ children on board	Scale (discrete)	
Ticket	Ticket number	String	
fare	Price of ticket	Scale	
Cabin	Cabin number	String	
Embarked	Where passenger embarked	String	
Boat	Boat identification (if rescued)	String	
Body	Body number (if died)	ID	
Home.dest	Home town	tring	
Gender	Gender	Binary (Nominal)	

Data contributed by Ellen Marshall, Sheffield Hallam University

Possible research questions:

Technique	Possible research questions		
1. Recoding variables	Identifying children	Identifying those travelling alone	
2. Bar/ pie charts	Class/gender/nationality	Class and nationality	Class and gender
3. Contingency tables	Is there a relationship between class and survival	Is there a relationship between gender and survival	Is there a relationship between nationality and survival?
4. Chi-squared	Is there a relationship between class and survival	Is there a relationship between gender and survival	Is there a relationship between nationality and survival?
5. Logistic regression	Predicting probability of survival using any independent variables.		
6. Skewed data	Fare is heavily skewed. Comparisons of fare by survival/nationality demonstrate differences in mean/medians etc		
7. Kruskall-Wallis	Are there differences in the amount paid for tickets by nationality?		