stcp-karadimitriou-categoricalR

The following resources are associated:

The dataset '*Titanic.csv*', R script file 'Rscript-cat', 'Chi-squared test in R' resource

Summarising categorical variables in R

Dependent variable: Categorical

Independent variable: Categorical

Data: On April 14th 1912 the ship the Titanic sank. Information on 1309 of those on board will be used to demonstrate summarising categorical variables.

After saving the 'Titanic.csv' file somewhere on your computer, open the data, call it TitanicR and define it as a data frame.

```
TitanicR<-data.frame(read.csv('...\\Titanic.csv',header=T,sep=','))
```

Attaching the data means that variables can be referred to by their column name
`attach(TitanicR)`

	pclass	survived	Residence	name	sex
1	3	0	0	Abbing, Mr. Anthony	male
2	3	0	0	Abbott, Master. Eugene Joseph	male
3	3	0	0	Abbott, Mr. Rossmore Edward	male
4	3	1	0	Abbott, Mrs. Stanton (Rosa Hunt)	female
5	3	1	2	Abelseth, Miss. Karen Marie	female
6	3	1	0	Abelseth, Mr. Olaus Jorgensen	male
7	2	0	2	Abelson, Mr. Samuel	male

R needs to know which variables are categorical variables and the labels for each value which can be specified using the `factor` command.

`variable<-factor(variable,c(category numbers),labels=c(category names)).`
The values are as follows: survival (0=died, 1=survived), Gender (0 = male, 1 = female), class (1st, 2nd, 3rd) and Country of Residence (Residence=American, British, Other).

```
survived<-factor(survived,c(0,1),labels=c('Died','Survived'))
pclass<-factor(ï..pclass,c(1,2,3),labels=c('First','Second','Third'))
Residence<-
factor(Residence,levels=c(0,1,2),labels=c('American','British','Other'))
Gender<-factor(Gender,levels=c(0,1),labels=c('Male','Female')
```

Research question: Did class affect survival?

© Sofia Maria Karadimitriou and Ellen Marshall
University of Sheffield

Reviewer: Paul Wilson
University of Wolverhampton

When summarising categorical data, percentages are usually preferable to frequencies although they can be misleading for very small sample sizes. Frequency tables can be produced using the `table()` command and proportions using the `prop.table ()` command. Here the frequencies and percentages of survival are calculated.
To calculate frequencies use the table command and give the table a name (SurT here).

```
SurT<-table(survived)
```

To view the table, type the name.

```
SurT
```

To add totals to the table, use the `addmargins()` command.

```
addmargins(SurT)
```

```
survived
    Died Survived      Sum
     809      500     1309
```

To calculate proportions from the frequency table.

```
prop.table(SurT)
```

Reduce the number of decimal places using the `round` function.

```
round(prop.table(SurT),digits=2)
```

```
survived
    Died Survived
    0.62     0.38
```

To produce percentages rounded to whole numbers.

```
round(100*prop.table(SurT),digits=0)
```

```
survived
    Died Survived
      62       38
```

The summary tables show that 500 of the 1309 passengers (38%) survived.

To break down survival by class, a cross tabulation or contingency table is needed. To produce a contingency table of frequencies, use the table command and give the table a name e.g. cross.

```
cross<-table(survived,class)
```

To add row and column totals to the table, use the `addmargins()` command.

```
addmargins(cross)
```

```
          class
survived   First Second Third  Sum
  Died       123    158   528  809
  Survived   200    119   181  500
  Sum        323    277   709 1309
```

To produce a contingency table containing proportions, use the `prop.table()` command. To calculate row proportions use `prop.table(cross, 1)` and to calculate column proportions use `prop.table(cross, 2)` then multiply by 100 to get percentages. Choose either row or column percentages carefully depending on the research question. Here percentages dying within each class are of interest so use column percentages. It would

be misleading to use row percentages (percentage those who died who were travelling in 3rd class) as there were more people in 3rd class.

To produce column percentages rounded to 0 decimal places
`round(prop.table(cross,2)*100,digits=0)`

```
> round(100*prop.table(cross,2),digits=0)
          class
survived   First Second Third
  Died        38     57    74
  Survived    62     43    26
```

It is clear from the percentages that the percentage of those dying increased as class lowered. 38% of passengers in 1st class died compared to 74% in 3rd class.

Bar Charts

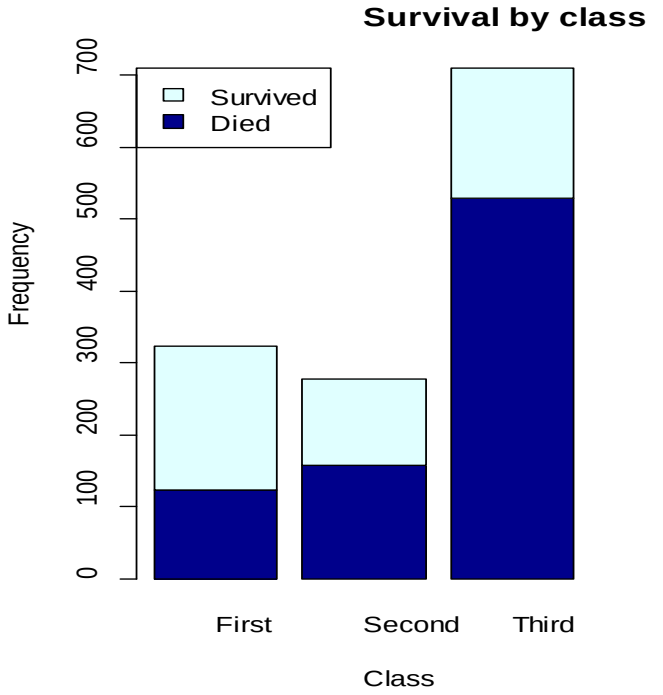
To display the information from the cross-tabulation graphically, use either a stacked or clustered (multiple) bar chart. To produce a stacked bar chart of contingency table '`cross`' with different colours for those dying/ surviving and a legend to identify the groups use:

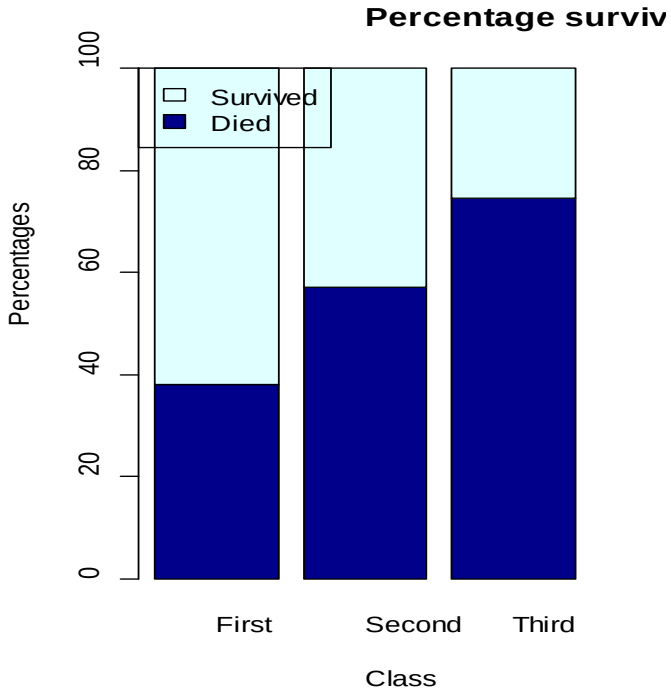
```
barplot(cross, xlab='Class',ylab='Frequency',main="Survival by class",
col=c("darkblue","lightcyan")
,legend=rownames(cross), args.legend = list(x = "topleft"))
```

To give a title to the plot use the `main=''` argument and to name the x and y axis use the `xlab=''` and `ylab=''` respectively.
Colours are changed through the `col` command e.g. `col=c("darkblue","lightcyan")`
Choose one light and one dark colour for black and white printing.
`Legend` assigns a legend to identify what each colour represents. The `args.legend` argument specifies the location of the legend e.g *'bottomright', 'topleft'* etc.)

It's not always clear if there are differences when there are different frequencies within each group so comparing percentages is often better.
To use percentages instead of frequencies on the bar chart, just change the table name `cross` to `prop.table(cross,2)`. However, it is not possible to display the percentages on the graph.

Ask for more information about the options for the barplot command
`?barplot`

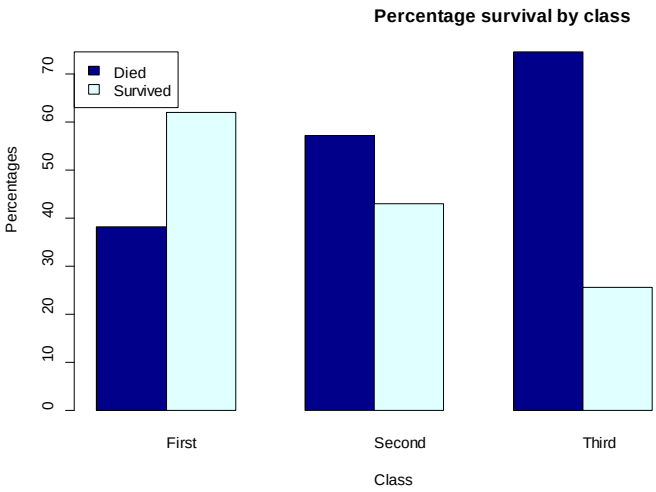




The charts show the frequencies and percentages of those dying and surviving within each class. The differences between classes are clearer on the percentage chart. It is clear from the percentages and bar chart that the percentage of those dying increased as class lowered. 38% of passengers in 1st class died compared to 74% in 3rd class.

Alternatively, produce a clustered bar chart by adding `beside=T` into the barplot command

```
barplot(prop.table(cross,2)*100,
xlab='Class',ylab='Percentages',mai
n="Percentage survival by
class",beside=T,col=c("darkblue","l
ightcyan"),
legend=rownames(cross), args.legend
= list(x = "topleft"))
```



Tips on reporting

Do not include every possible chart and frequency.

Think back to the key question of interest and answer this question.

Briefly talk about every chart and table you include but don't discuss every number if the table is included.

Percentages should be rounded to whole numbers unless you are dealing with very small numbers e.g. 0.01%