Contingency Table - Count Data

Topics:

- Contingency table and mosaic plot
- Examples
  - Passenger survival on Titanic
  - UC Berkeley admissions
- Simpson's paradox

Contingency Tables

- is a table of counts.
- Rows and columns are labelled with values of categorical variable (factor).
- Value in each cell is the count of number of cases in that combination of categories

Constructing a contingency table in R

```
drugz <- data.frame(
expand.grid(treatment=c("drugz","placebo"),result=c("sicker","better")),count=c(20,60,80,40))
drugz
```

```
##   treatment result count
## 1     drugz sicker    20
## 2   placebo sicker    60
## 3     drugz better    80
## 4   placebo better    40
```

```
drugztable <- xtabs(formula = count ~ treatment + result, data = drugz)
drugztable
```

```
##          result
## treatment sicker better
##   drugz       20     80
##   placebo     60     40
```

Got it! drugztable is now a special data type representing a **contingency table**.

100 patients with some disease were treated with drug Z, and 100 patients were treated with a placebo.

Nearly always, we are interested in interactions or correlations between the variables.

The table above suggests that drug Z might cure people, not completely reliably.

3 ways to represent contingency tables in R

1. Case form:
   - Data frame (like a database table) with 2 factors Treatment and Result
   - This needs to be counted!
2. Frequency form:
   - Data frame (like a database table) with factors Treatment, Result and one numeric column
   - Needs to be converted to table form using xtabs function
3. Table form:
   - A labelled multidimensional array.
   - This is the form we need in order to use the function mosaic

For details on how to create and convert different forms, see: https://cran.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf

Shaded according to statistical significance

A simple and effective visualization technique for contingency tables is the **mosaic plot**.

```
#install.packages("vcd")
#install.packages("vcdExtra")
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.0.5
```

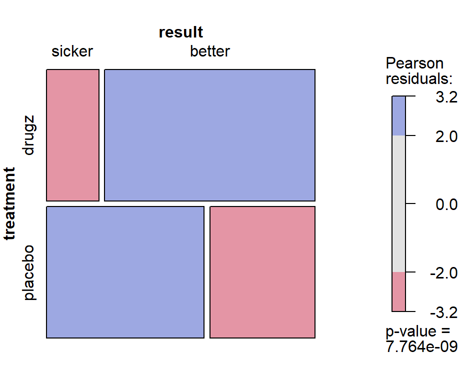
```
## Loading required package: grid
```

```
library(vcdExtra)
```

```
## Loading required package: gnm
```

```
library(gcookbook)
library(datasets)

mosaic(~treatment+result, drugztable, shade=TRUE,split_vertical=c(FALSE,TRUE))
```



The shade argument causes mosaic to shade the plot according to statistical significance of deviation from independence.

Constructing a Contingency Table in R

```
GSS <- data.frame( expand.grid(sex=c("female", "male"), party=c("dem", "indep", "rep")), count=c(279,165,73,47,22
5,191))
GSS
```

```
##      sex party count
## 1 female   dem   279
## 2   male   dem   165
## 3 female indep    73
## 4   male indep    47
## 5 female   rep   225
## 6   male   rep   191
```

```
gsstable <- xtabs( formula = count ~ sex + party, data = GSS)
dimnames( gsstable )
```

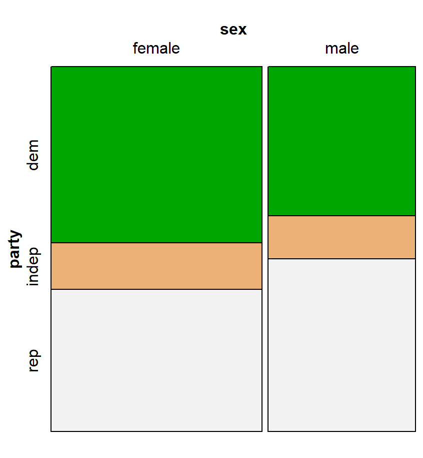
```
## $sex
## [1] "female" "male"
##
## $party
## [1] "dem"   "indep" "rep"
```

```
gsstable
```

```
##         party
## sex      dem indep rep
##   female 279    73 225
##   male   165    47 191
```

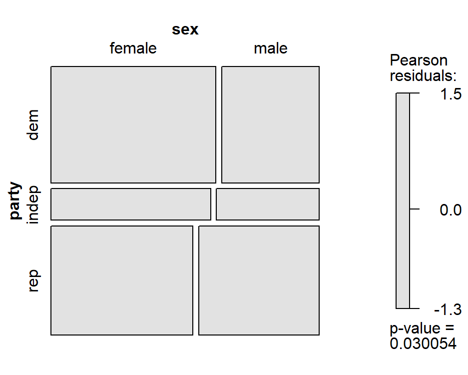
Social survey data

```
mosaic( ~party + sex, data=gsstable, highlighting=TRUE,highlighting_fill=terrain.colors(3))
```



Assessing statistical significance

```
mosaic( ~party + sex, data=gsstable, shade=TRUE)
```



No individual box is significantly larger or smaller than expected, by itself.

Examples: Passenger survival on Titanic

Passenger survival: 4 dimensional contingency table giving

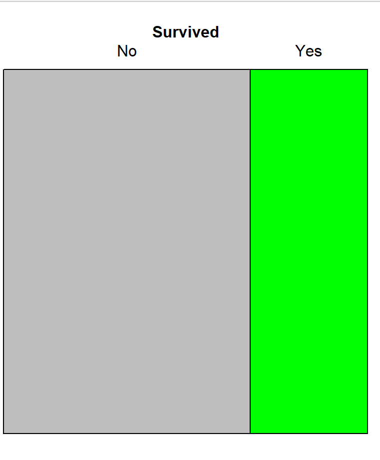
1. Class,
2. Age(adult/child),
3. Sex, and
4. whether Survived

```
Titanic
```

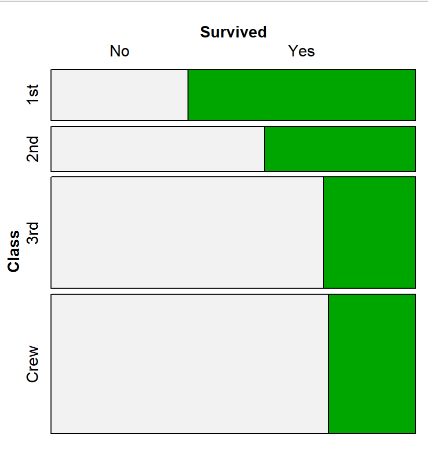
```
## , , Age = Child, Survived = No
##
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
##
## , , Age = Adult, Survived = No
##
##       Sex
## Class  Male Female
##   1st   118      4
##   2nd   154     13
##   3rd   387     89
##   Crew  670      3
##
## , , Age = Child, Survived = Yes
##
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
##
## , , Age = Adult, Survived = Yes
##
##       Sex
## Class  Male Female
##   1st    57    140
##   2nd    14     80
##   3rd    75     76
##   Crew  192     20
```

Titanic: overall survival

```
mosaic(~Survived, data=Titanic, split_vertical=c(TRUE),highlighting_fill=c("grey","green"), highlighting=TRUE)
```

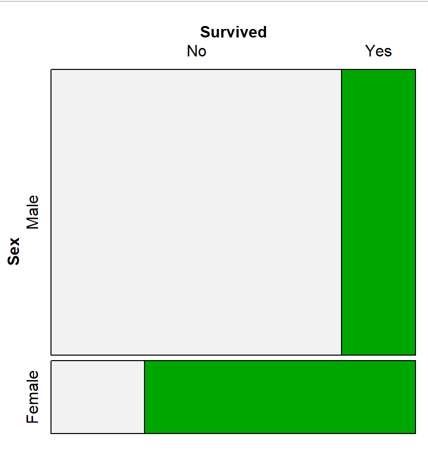



```
mosaic(~Survived + Class, data=Titanic, split_vertical=c(TRUE,FALSE,FALSE), highlighting_fill=rev(terrain.colors(
2,alpha=1)),highlighting=TRUE)
```



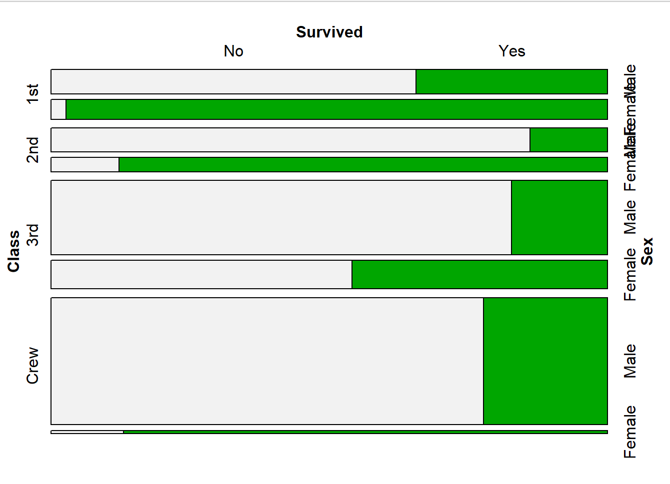
Titanic: survival rates of men and women

```
mosaic(~Survived + Sex, data=Titanic[,,"Adult",], split_vertical=c(TRUE,FALSE), highlighting_fill=rev(terrain.col
ors(2)), highlighting=TRUE)
```



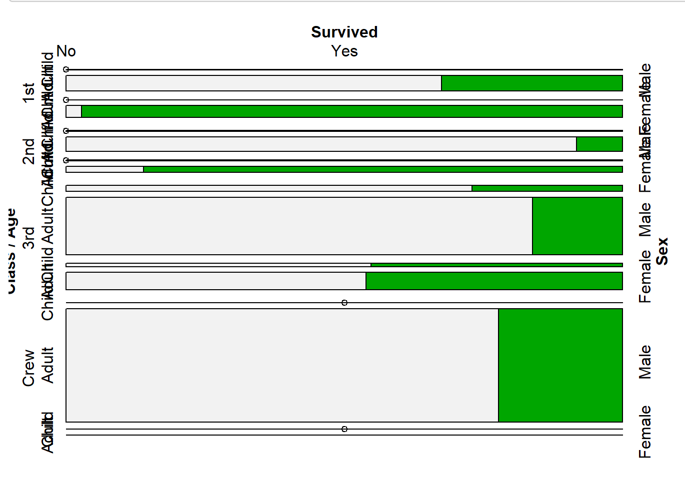
Titanic: survival by gender and class

```
mosaic(~Survived + Class + Sex, data=Titanic, split_vertical=c(TRUE,FALSE,FALSE), highlighting_fill=rev(terrain.c
olors(2,alpha=1)),highlighting=TRUE)
```



With children separated from adults … hard to read now for general display, but ok for exploration

```
mosaic(~Survived + Class + Sex + Age, data=Titanic, split_vertical=c(TRUE,FALSE,FALSE, FALSE), highlighting_fill=
rev(terrain.colors(2,alpha=1)),highlighting=TRUE)
```



Selecting children only

```
dimnames(Titanic)
```

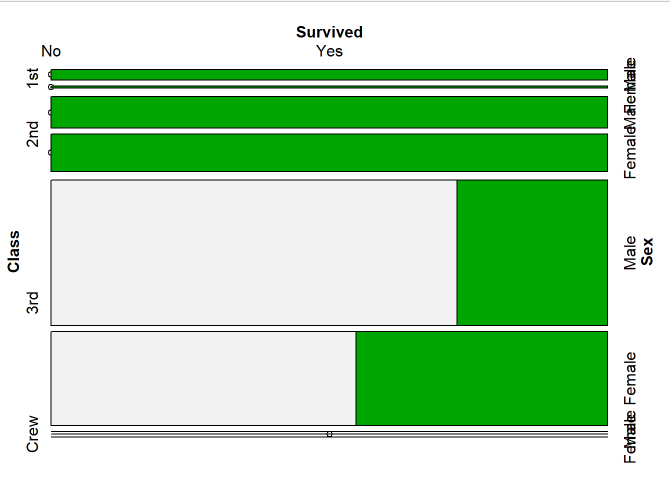
```
## $Class
## [1] "1st"  "2nd"  "3rd"  "Crew"
##
## $Sex
## [1] "Male"   "Female"
##
## $Age
## [1] "Child" "Adult"
##
## $Survived
## [1] "No"  "Yes"
```

```
Titanic[ , , "Child", ]
```

```
## , , Survived = No
##
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
##
## , , Survived = Yes
##
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
```

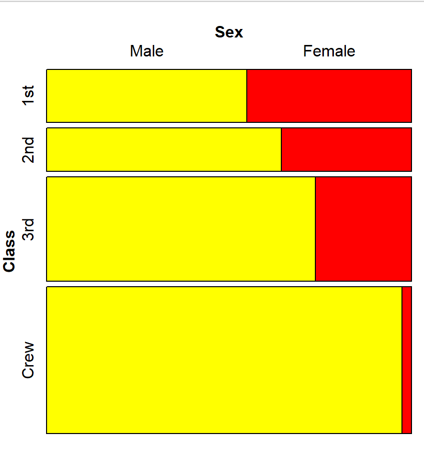
Titanic: survival of children

```
mosaic(~Survived + Class + Sex, data=Titanic[,,"Child",], split_vertical=c(TRUE,FALSE,FALSE), highlighting_fill=r
ev(terrain.colors(2)), highlighting=TRUE)
```



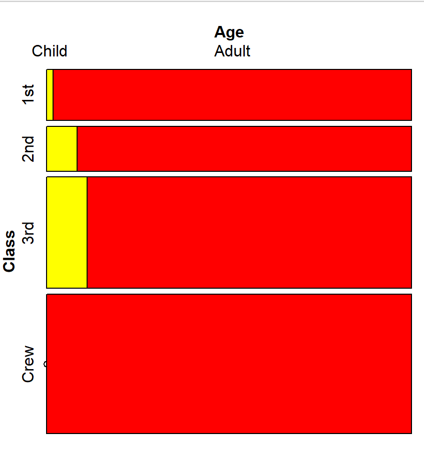
Titanic: proportions of women among adults, by class

```
mosaic(~Sex + Class, data=Titanic[,,"Adult",], split_vertical=c(TRUE,FALSE), highlighting_fill=rev(heat.colors(2
)), highlighting=TRUE)
```



Titanic: proportions of children, by class

```
mosaic(~Age + Class, data=Titanic, split_vertical=c(TRUE,FALSE), highlighting_fill=rev(heat.colors(2)), highlight
ing=TRUE)
```

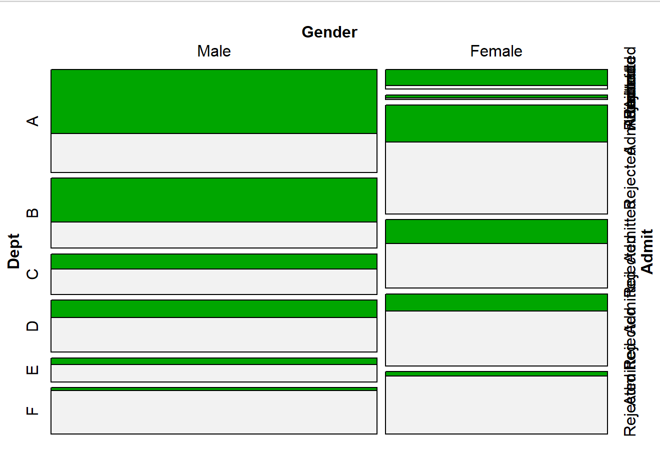


Titanic: some qualitative conclusions

- The legend is that the 1st class passengers were saved while the lower classes and the crew perished.
- The visualisation shows a strong interaction of class and survival.
- But there is also a strong interaction of survival with gender and age: a higher proportion of women and children of all classes survived than did even the men in 1st class.
- Among men, the order of survival rates was 1st class, crew, 3rd class, 2nd class.
- Among women, the order of survival rates was 1st class, 2nd class and crew, 3rd class.
- The survival rate of female children was greater than for male children.
- The apparent greater survival rate of 1st class passengers may be partly because a greater proportion of 1st class passengers were women.
- Strikingly, the third-class children had much lower survival rates than first and second-class children. What happened? Were many of the third-class children younger, or was there some other reason?

Example: UC Berkeley Admissions

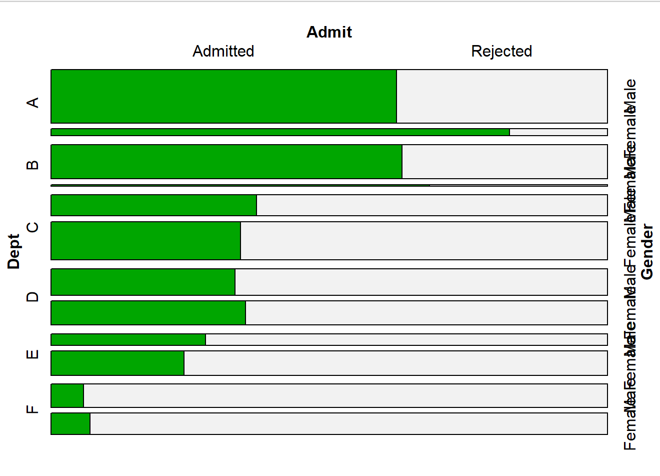
```
mosaic( ~Admit + Gender + Dept, data=UCBAdmissions, highlighting=TRUE, highlighting_fill=terrain.colors(2))
```



UCB admissions by Department

and Sex

```
mosaic( ~Admit + Dept + Gender, data=UCBAdmissions, highlighting=TRUE, highlighting_fill=terrain.colors(2),split_
vertical=c(TRUE,FALSE,FALSE))
```



UC Berkely admissions: qualitative conclusions

- When analysed by department, two out of six departments had slightly higher admissions rates for men than for women; in four departments, admissions rates for women were higher than for men.
- Women mostly applied to departments with low admissions rates; on aggregate, therefore, the admission rate for women was lower than for men.

Simpson's Paradox

- Simpson's Paradox occurs when trends that appear when a dataset is separated into groups reverse when the data are aggregated. – This result is often encountered in social science and medical science statistics.
- More statistically, it says that the apparent relationship between two variables can change in the light or absence of a third variable (cofounding variable).

Three types of questions

1. Visualisation
   - how can we see any interactions?
2. Pattern or chance? Statistical significance
   - could the apparent interaction be just luck?
   - do we have enough observations to reliably identify an interaction?

3.Investigation: confounding factors

- what is the explanation of the interaction?