Visualising Categorical Data

To visualise categorical (or qualitative) variables, we will be using the Berkeley Admissions dataset.This dataset is included in the R datasets package. It contains graduate school applicants to the six largest departments at University of California, Berkeley in 1973. To learn more about the dataset, use the help function.

Why is this dataset interesting? This dataset is useful for demonstrating how we can visualise categorical data using fourfold plots and cotab plots. It is also an example of Simpson's paradox. This can be explained as a phenomenon where a trend appears in groups of data but disappears or reverses when combined with another group of data.

In the UCBAdmissions dataset, when we look at the **Admit** and **Gender** variables, there appears to be bias towards the number of men being admitted, with women having a lower acceptance rate overall. When we compare **Admit** and **Gender** with **Dept**, this bias disappears and we can see that the admission rates are similar for males and females in most departments, except A.

We will explore the data using the **vcd** (Visualising Categorical Data) package.

```
# Load the dataset
data(UCBAdmissions)

# Help page
?UCBAdmissions
```

Now let's look at the structure of the dataset.

```
dim(UCBAdmissions)
```

```
## [1] 2 2 6
```

```
dimnames(UCBAdmissions)
```

```
## $Admit
## [1] "Admitted" "Rejected"
##
## $Gender
## [1] "Male"   "Female"
##
## $Dept
## [1] "A" "B" "C" "D" "E" "F"
```

```
str(UCBAdmissions)
```

```
##  table [1:2, 1:2, 1:6] 512 313 89 19 353 207 17 8 120 205 ...
##  - attr(*, "dimnames")=List of 3
##   ..$ Admit : chr [1:2] "Admitted" "Rejected"
##   ..$ Gender: chr [1:2] "Male" "Female"
##   ..$ Dept  : chr [1:6] "A" "B" "C" "D" ...
```

The applicants are classified by **Admit** (either Admitted or Rejected), **Gender** (either Male or Female) and **Department** (A to F). The data forms a 3-way table (2 x 2 x 6).

First, let’s examine the relationship between **Admit** and **Gender** using a two-way frquency table.

```
UCB.GA <- margin.table(UCBAdmissions, c(1,2))
UCB.GA
```

```
##           Gender
## Admit      Male Female
##   Admitted 1198    557
##   Rejected 1493   1278
```

There seems to be a difference between the number of females and males that are admitted. Let’s create a cross table using the **gmodels** package.

```
univ <- apply(UCBAdmissions, c(1,2), sum)
univ
```

```
##           Gender
## Admit      Male Female
##   Admitted 1198    557
##   Rejected 1493   1278
```

```
prop.table(univ, 2)
```

```
##           Gender
## Admit           Male    Female
##   Admitted 0.4451877 0.3035422
##   Rejected 0.5548123 0.6964578
```

We can see that the proportion of males admitted is 44.5%, compared to 30.4% of females. It seems there may be some bias here. Let’s take a closer look at the odds ratio.

What is the odds ratio?

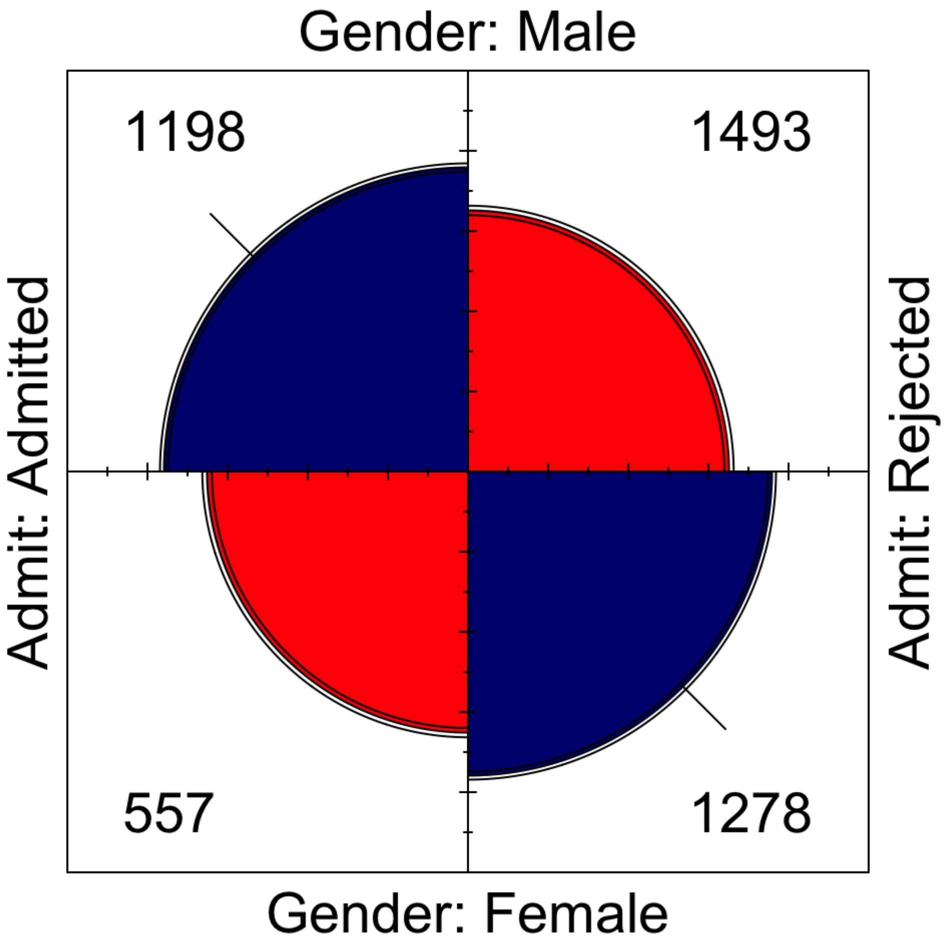
The odds ratio is the probability of success over failure. In this case, it is the probability that one is admitted versus the probability of being rejected by department, given their gender.

Let’s look a two-way plot first of **Admit** and **Gender**, ignoring the department that they applied to.

```
# load the vcd package
library(vcd)
```

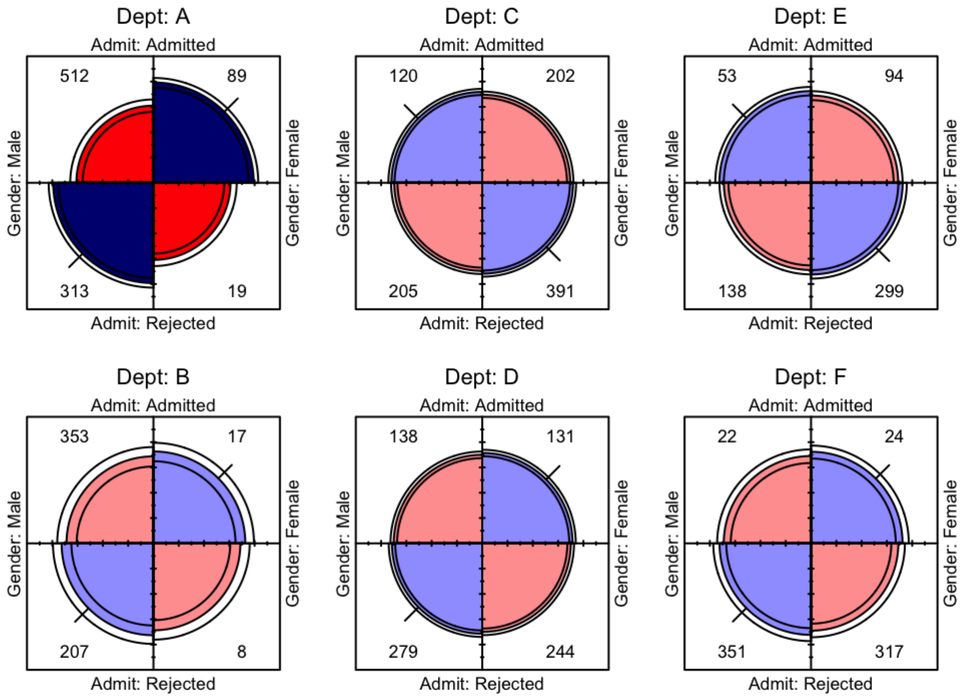
```
## Loading required package: grid
```

```
# fourfold plot
UCB <- aperm(UCBAdmissions, c(2,1,3))
fourfold(margin.table(UCB, c(1, 2)))
```



Now let’s look at a three-way plot.

```
fourfold(UCBAdmissions, mfrow=c(2,3))
```



As we can see, when department is excluded there is quite a different story being told. Admission is the response variable, whilst gender and department are explanatory variables.

When we look at the admission of males and females by department, we can that the admission of males and females is quite similar, with the exception of department A. Within department A, more females are admitted than males. Proportionally, more females are also accepted with departments B, D and F.

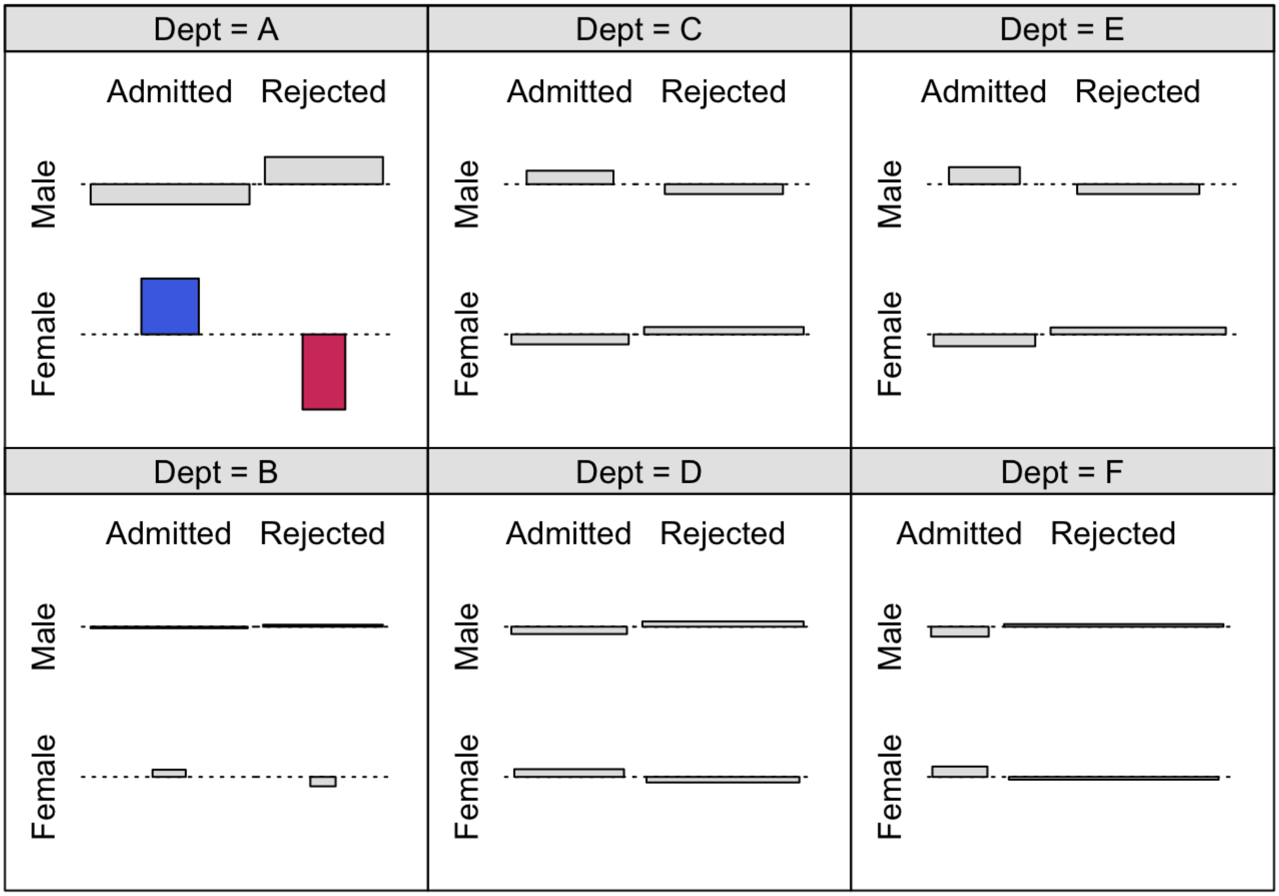
To check this, let's create another frequency table with the proportion of males and females accepted by department.

```
ftable(round(prop.table(UCBAdmissions, c(2,3)), 2),
       row.vars="Dept", col.vars = c("Gender", "Admit"))
```

```
##      Gender     Male              Female         
##      Admit  Admitted Rejected Admitted Rejected
## Dept                                         
## A               0.62     0.38     0.82     0.18
## B               0.63     0.37     0.68     0.32
## C               0.37     0.63     0.34     0.66
## D               0.33     0.67     0.35     0.65
## E               0.28     0.72     0.24     0.76
## F               0.06     0.94     0.07     0.93
```

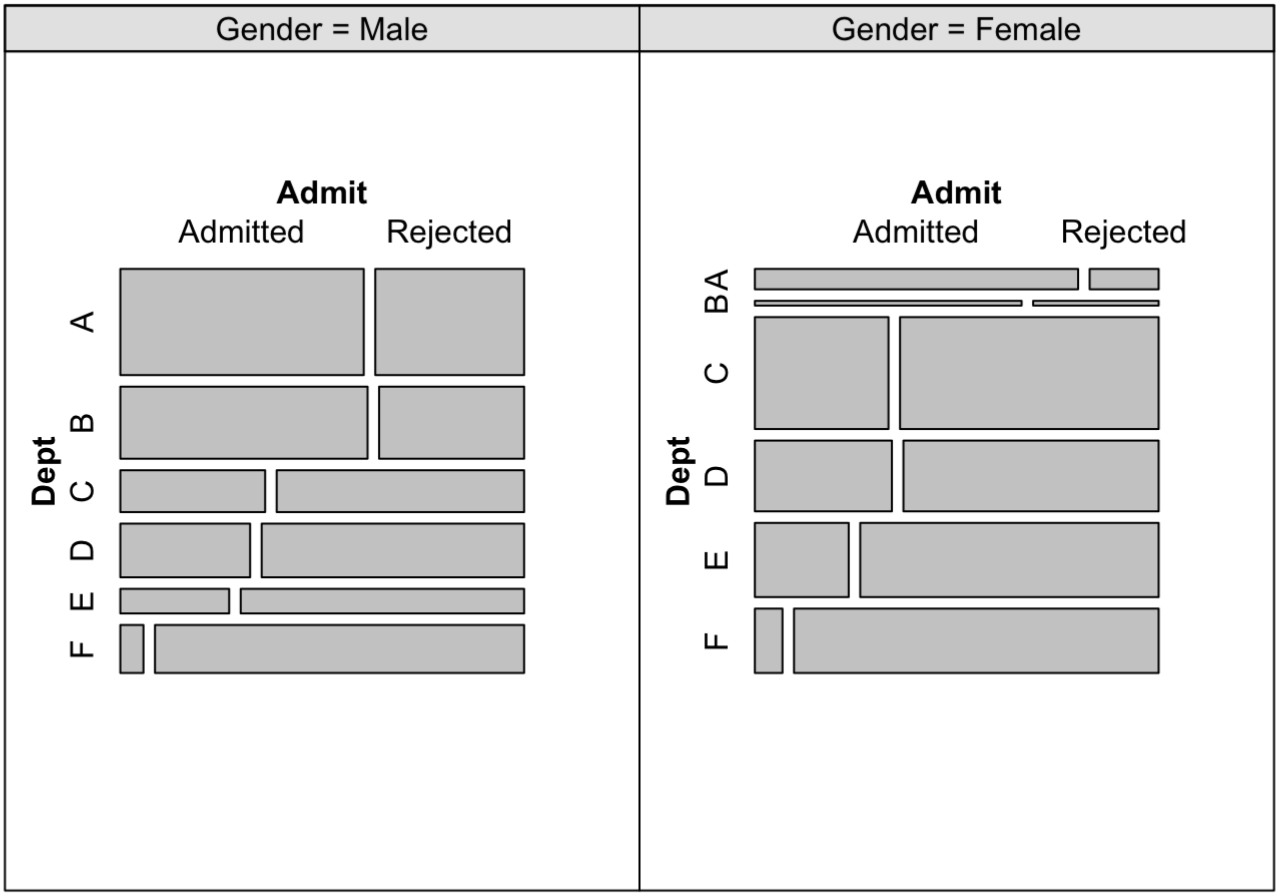
We can also look at the residuals by department using a cotabplot. This highlights the large residuals in department A.

```
cotabplot(aperm(UCBAdmissions, c(2,1,3)), panel = cotab_coindep, shade = TRUE,
          legend = FALSE,
          panel_args = list(type = "assoc", margins = c(2,1,1,2), varnames = FALSE))
```



Now, let’s look at this split by gender.

```
UCB_G <- structable(Gender ~ Dept + Admit, data =UCBAdmissions)
cotabplot(UCB_G)
```



This allows us to clearly see that there are more males who apply to departments A and B, where more females apply for departments C, D, E, and F. By visualising the data we are able to better understand the relationship between the three variables: Admit, Gender and Dept.

If we take another look at the proportional values, we can see that departments A and B admit approximately 2 in 3 applicants of either gender, whilst departments C:F admit 1 in 3 or fewer.

```
admit <- apply(UCBAdmissions, c(2,3,1), sum)
admit
```

```
## , , Admit = Admitted
## 
##         Dept
## Gender     A   B   C   D  E  F
##   Male   512 353 120 138 53 22
##   Female  89  17 202 131 94 24
## 
## , , Admit = Rejected
## 
##         Dept
## Gender     A   B   C   D   E   F
##   Male   313 207 205 279 138 351
##   Female  19   8 391 244 299 317
```