DDAR: Solutions and Hints for Exercises

2016-01-09

This document is intended as an aid to instructors who wish to use *Discrete Data Analysis with R* in a course. It contains the text of the *Exercises* sections from all chapters, together with some solutions or hints for the various problems.

## Chapter 1: Introduction

These exercises are all conceptual. There are no hints or solutions. The goal here was to get students to investigate various sources of data graphics and think about how well they communicated some facts, relationships or explanations.

## Chapter 2: Working with Cateorical Data

### Exercise 2.1

The packages vcd and vcdExtra contain many data sets with some examples of analysis and graphical display. The goal of this exercise is to familiarize yourself with these resources. You can get a brief summary of these using the function datasets() from vcdExtra. Use the following to get a list of these with some characteristics and titles.

library(vcdExtra)  
ds <- datasets(package=c("vcd", "vcdExtra"))

1. How many data sets are there altogether? How many are there in each package?

nrow() gives the number of rows in a data frame. table() for a single variable gives the frequencies for each level.

ds <- datasets(package=c("vcd", "vcdExtra"))  
nrow(ds)

## [1] 75

table(ds$Package)

##   
## vcd vcdExtra   
## 33 42

1. Make a tabular display of the frequencies by Package and class.

Use table(), but now for Package and class.

table(ds$Package, ds$class)

##   
## array data.frame matrix table  
## vcd 1 17 0 15  
## vcdExtra 3 23 1 15

1. Choose one or two data sets from this list, and examine their help files (e.g., help(Arthritis) or ?Arthritis). You can use, e.g., example(Arthritis) to run the R code for a given example.

?Arthritis #Help Files  
?Baseball #Help Files  
example(Arthritis) #Example Syntax/Analysis  
example(Baseball) #Example Syntax/Analysis

### Exercise 2.2

For each of the following data sets in the vcdExtra package, identify which are response variable(s) and which are explanatory. For factor variables, which are unordered (nominal) and which should be treated as ordered? Write a sentence or two describing substantitive questions of interest for analysis of the data. (Hint: use data(foo, package="vcdExtra") to load, and str(foo), help(foo) to examine data set foo.)

1. Abortion opinion data: Abortion

Support\_Abortion is the response, Sex and Status are binary, nominal explanatory variables. From help(Abortion), How does support for abortion depend on sex and status?

data(Abortion, package="vcdExtra")  
str(Abortion)

## table [1:2, 1:2, 1:2] 171 152 138 167 79 148 112 133  
## - attr(\*, "dimnames")=List of 3  
## ..$ Sex : chr [1:2] "Female" "Male"  
## ..$ Status : chr [1:2] "Lo" "Hi"  
## ..$ Support\_Abortion: chr [1:2] "Yes" "No"

1. Caesarian Births: Caesar

Infection is the response, Risk, Antibiotics and Planned are binary, nominal explanatory variables.

data(Caesar, package="vcdExtra")  
str(Caesar)

## table [1:3, 1:2, 1:2, 1:2] 0 1 17 0 1 1 11 17 30 4 ...  
## - attr(\*, "dimnames")=List of 4  
## ..$ Infection : chr [1:3] "Type 1" "Type 2" "None"  
## ..$ Risk : chr [1:2] "Yes" "No"  
## ..$ Antibiotics: chr [1:2] "Yes" "No"  
## ..$ Planned : chr [1:2] "Yes" "No"

1. Dayton Survey: DaytonSurvey

In DaytonSurvey, the variables cigarette, alcohol, and marijuana can all be treated as response variables. sex and race are potential explanatory variables.

data(DaytonSurvey, package="vcdExtra")  
str(DaytonSurvey)

## 'data.frame': 32 obs. of 6 variables:  
## $ cigarette: Factor w/ 2 levels "Yes","No": 1 2 1 2 1 2 1 2 1 2 ...  
## $ alcohol : Factor w/ 2 levels "Yes","No": 1 1 2 2 1 1 2 2 1 1 ...  
## $ marijuana: Factor w/ 2 levels "Yes","No": 1 1 1 1 2 2 2 2 1 1 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 2 2 ...  
## $ race : Factor w/ 2 levels "white","other": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Freq : num 405 13 1 1 268 218 17 117 453 28 ...

1. Minnesota High School Graduates: Hoyt

Status is the response, Rank, Occupation, and Sex are explanatory variables. Both Rank and Occupation are ordinal.

data(Hoyt, package="vcdExtra")  
str(Hoyt)

## table [1:4, 1:3, 1:7, 1:2] 87 3 17 105 216 4 14 118 256 2 ...  
## - attr(\*, "dimnames")=List of 4  
## ..$ Status : chr [1:4] "College" "School" "Job" "Other"  
## ..$ Rank : chr [1:3] "Low" "Middle" "High"  
## ..$ Occupation: chr [1:7] "1" "2" "3" "4" ...  
## ..$ Sex : chr [1:2] "Male" "Female"

### Exercise 2.3

The data set UCBAdmissions is a 3-way table of frequencies classified by Admit, Gender, and Dept.

1. Find the total number of cases contained in this table.

For a table object, just use sum()

data(UCBAdmissions)  
sum(UCBAdmissions)

## [1] 4526

1. For each department, find the total number of applicants.

margin.table(UCBAdmissions, 3)

## Dept  
## A B C D E F   
## 933 585 918 792 584 714

1. For each department, find the overall proportion of applicants who were admitted.

ucb.df <- as.data.frame(UCBAdmissions)  
abd <- xtabs(Freq ~ Dept + Admit, data=ucb.df)  
prop.table(abd, 1)

## Admit  
## Dept Admitted Rejected  
## A 0.64416 0.35584  
## B 0.63248 0.36752  
## C 0.35076 0.64924  
## D 0.33965 0.66035  
## E 0.25171 0.74829  
## F 0.06443 0.93557

1. Construct a tabular display of department (rows) and gender (columns), showing the proportion of applicants in each cell who were admitted relative to the total applicants in that cell.

### Exercise 2.4

The data set DanishWelfare in vcd gives a 4-way, 3 x 4 x 3 x 5 table as a data frame in frequency form, containing the variable Freq and four factors, Alcohol, Income, Status and Urban. The variable Alcohol can be considered as the response variable, and the others as possible predictors.

1. Find the total number of cases represented in this table.

This is a data set in the form of a frequency data.frame, so sum the Freq variable

data("DanishWelfare", package="vcd")  
sum(DanishWelfare$Freq)

## [1] 5144

1. In this form, the variables Alcohol and Income should arguably be considered ordered factors. Change them to make them ordered.

Use ordered() or as.ordered() on the factor variable. str() will then show them as Ord.factor.

levels(DanishWelfare$Alcohol)

## [1] "<1" "1-2" ">2"

DanishWelfare$Alcohol <- as.ordered(DanishWelfare$Alcohol)  
DanishWelfare$Income <- as.ordered(DanishWelfare$Income)  
str(DanishWelfare)

## 'data.frame': 180 obs. of 5 variables:  
## $ Freq : num 1 4 1 8 6 14 8 41 100 175 ...  
## $ Alcohol: Ord.factor w/ 3 levels "<1"<"1-2"<">2": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Income : Ord.factor w/ 4 levels "0-50"<"50-100"<..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Status : Factor w/ 3 levels "Widow","Married",..: 1 1 1 1 1 2 2 2 2 2 ...  
## $ Urban : Factor w/ 5 levels "Copenhagen","SubCopenhagen",..: 1 2 3 4 5 1 2 3 4 5 ...

1. Convert this data frame to table form, DanishWelfare.tab, a 4-way array containing the frequencies with appropriate variable names and level names.

Use xtabs() with Freq as the response.

DanishWelfare.tab <-xtabs(Freq ~ ., data = DanishWelfare)  
str(DanishWelfare.tab)

## xtabs [1:3, 1:4, 1:3, 1:5] 1 3 2 8 1 3 2 5 2 42 ...  
## - attr(\*, "dimnames")=List of 4  
## ..$ Alcohol: chr [1:3] "<1" "1-2" ">2"  
## ..$ Income : chr [1:4] "0-50" "50-100" "100-150" ">150"  
## ..$ Status : chr [1:3] "Widow" "Married" "Unmarried"  
## ..$ Urban : chr [1:5] "Copenhagen" "SubCopenhagen" "LargeCity" "City" ...  
## - attr(\*, "class")= chr [1:2] "xtabs" "table"  
## - attr(\*, "call")= language xtabs(formula = Freq ~ ., data = DanishWelfare)

1. The variable Urban has 5 categories. Find the total frequencies in each of these. How would you collapse the table to have only two categories, City, Non-city?

margin.table() handles the first part; collapse.table() is designed for the second part. It is arguable whether SubCopenhagen should be considered City or NonCity.

margin.table(DanishWelfare.tab, 4)

## Urban  
## Copenhagen SubCopenhagen LargeCity City Country   
## 552 614 594 1765 1619

DW2 <- vcdExtra::collapse.table(DanishWelfare.tab, Urban=c("City","NonCity","City","City","NonCity"))  
head(ftable(DW2))

##   
## "Urban" "City" "NonCity"  
## "Alcohol" "Income" "Status"   
## "<1" "0-50" "Widow" 10 10  
## "Married" 155 183  
## "Unmarried" 14 10  
## "50-100" "Widow" 29 7  
## "Married" 338 306  
## "Unmarried" 36 32

### Exercise 2.5

The data set UKSoccer in vcd gives the distributions of number of goals scored by the 20 teams in the 1995/96 season of the Premier League of the UK Football Association.

data("UKSoccer", package = "vcd")   
ftable(UKSoccer)

This two-way table classifies all games by the joint outcome (Home, Away), the number of goals scored by the Home and Away teams. The value 4 in this table actually represents 4 or more goals.

1. Verify that the total number of games represented in this table is 380.

data("UKSoccer", package="vcd")  
sum(UKSoccer)

## [1] 380

margin.table(UKSoccer)

## [1] 380

1. Find the marginal total of the number of goals scored by each of the home and away teams.

margin.table(UKSoccer, 1)

## Home  
## 0 1 2 3 4   
## 76 142 90 45 27

margin.table(UKSoccer, 2)

## Away  
## 0 1 2 3 4   
## 140 136 55 38 11

1. Express each of the marginal totals as proportions.

prop.table(margin.table(UKSoccer, 1))

## Home  
## 0 1 2 3 4   
## 0.20000 0.37368 0.23684 0.11842 0.07105

prop.table(margin.table(UKSoccer, 2))

## Away  
## 0 1 2 3 4   
## 0.36842 0.35789 0.14474 0.10000 0.02895

1. Comment on the distribution of the numbers of home-team and away-team goals. Is there any evidence that home teams score more goals on average?

You could find the mean number of goals, weighted by their marginal frequencies. On average, home teams score about 0.4 more goals.

weighted.mean(0:4, w=margin.table(UKSoccer,1))

## [1] 1.487

weighted.mean(0:4, w=margin.table(UKSoccer,2))

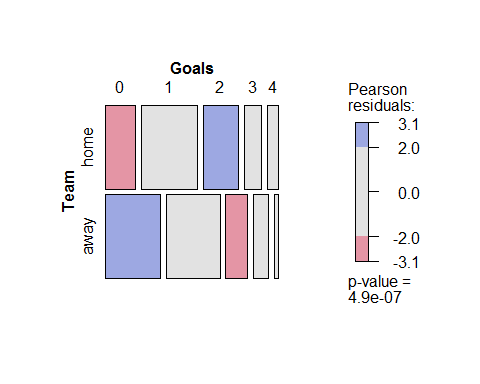
## [1] 1.063

Graphically, you could also compare the marginal frequencies in a mosaic plot, or use vcd::agreementplot().

margins <- rbind(home=margin.table(UKSoccer,1),   
 away=margin.table(UKSoccer,2))  
names(dimnames(margins)) <- c("Team", "Goals")  
margins

## Goals  
## Team 0 1 2 3 4  
## home 76 142 90 45 27  
## away 140 136 55 38 11

mosaic(margins, shade=TRUE)



### Exercise 2.6

The one-way frequency table Saxony in vcd records the frequencies of families with 0, 1, 2, 12 male children, among 6115 families with 12 children. This data set is used extensively in Chapter 3.

data("Saxony", package = "vcd")   
Saxony

Another data set, Geissler, in the vcdExtra, gives the complete tabulation of all combinations of boys and girls in families with a given total number of children (size). The task here is to create an equivalent table, Saxony12 from the Geissler data.

data("Geissler", package = "vcdExtra")   
str(Geissler)

1. Use subset() to create a data frame, sax12 containing the Geissler observations in families with size==12.

data("Saxony", package="vcd")  
data("Geissler", package="vcdExtra")  
sax12 <- subset(Geissler, size==12)  
sax12

## boys girls size Freq  
## 12 0 12 12 3  
## 24 1 11 12 24  
## 35 2 10 12 104  
## 45 3 9 12 286  
## 54 4 8 12 670  
## 62 5 7 12 1033  
## 69 6 6 12 1343  
## 75 7 5 12 1112  
## 80 8 4 12 829  
## 84 9 3 12 478  
## 87 10 2 12 181  
## 89 11 1 12 45  
## 90 12 0 12 7

1. Select the columns for boys and Freq.

sax12 <- subset(sax12, select=c("boys","Freq"))

1. Use xtabs() with a formula, Freq ~ boys, to create the one-way table.

Saxony12<-xtabs(Freq~boys, data=sax12)  
Saxony12

## boys  
## 0 1 2 3 4 5 6 7 8 9 10 11 12   
## 3 24 104 286 670 1033 1343 1112 829 478 181 45 7

1. Do the same steps again to create a one-way table, Saxony11, containing similar frequencies for families of size==11.

sax11 <- subset(Geissler, size==11, select = c("boys","Freq"))  
Saxony11 <- xtabs(Freq~boys, data=sax11)  
Saxony11

## boys  
## 0 1 2 3 4 5 6 7 8 9 10 11   
## 8 72 275 837 1540 2161 2310 1801 1077 492 93 24

### Exercise 2.7

*Interactive coding of table factors*: Some statistical and graphical methods for are implemented only for two-way tables, but can be extended to 3+-way tables by recoding the factors to interactive combinations along the rows and/or columns, in a way similar to what ftable and structable do for printed displays.

For the UCBAdmissions data, produce a two-way table object, UCB.tab2, that has the combinations of Admit and Gender as the rows, and Dept as its columns, to look like the result below:

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

1. Try this the long way: convert UCBAdmissions to a data frame (as.data.frame()), manipulate the factors (e.g., interaction()), then convert back to a table (as.data.frame()).

ucb.df$AG <- with(ucb.df, interaction(Admit, Gender, sep=":"))  
ucb <- subset(ucb.df, select = c("Dept", "AG", "Freq"))  
ucb.tab2 <- xtabs(Freq ~ AG + Dept, data=ucb)  
ucb.tab2

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

1. Try this the short way: both ftable() and structable() have as.matrix methods that convert their result to a matrix.

ucb.tab2 <- as.matrix(structable(Dept ~ Admit + Gender, data = UCBAdmissions))  
ucb.tab2

## Dept  
## Admit\_Gender A B C D E F  
## Admitted\_Male 512 353 120 138 53 22  
## Admitted\_Female 89 17 202 131 94 24  
## Rejected\_Male 313 207 205 279 138 351  
## Rejected\_Female 19 8 391 244 299 317

### Exercise 2.8

The data set VisualAcuity in vcd gives a table as a frequency data frame.

data("VisualAcuity", package = "vcd")   
str(VisualAcuity)

1. From this, use xtabs() to create two frequency tables, one for each gender.

The simplest way is to use the subset argument to xtabs().

data("VisualAcuity", package="vcd")  
va.tabm <- xtabs(Freq ~ right+left, data = VisualAcuity, subset=gender=="male")  
va.tabm

## left  
## right 1 2 3 4  
## 1 821 112 85 35  
## 2 116 494 145 27  
## 3 72 151 583 87  
## 4 43 34 106 331

va.tabf <- xtabs(Freq ~ right+left, data = VisualAcuity, subset=gender=="female")  
  
# or, subset after  
va.tab <- xtabs(Freq ~ ., data = VisualAcuity)  
str(va.tab)

## xtabs [1:4, 1:4, 1:2] 821 116 72 43 112 494 151 34 85 145 ...  
## - attr(\*, "dimnames")=List of 3  
## ..$ right : chr [1:4] "1" "2" "3" "4"  
## ..$ left : chr [1:4] "1" "2" "3" "4"  
## ..$ gender: chr [1:2] "male" "female"  
## - attr(\*, "class")= chr [1:2] "xtabs" "table"  
## - attr(\*, "call")= language xtabs(formula = Freq ~ ., data = VisualAcuity)

va.tabm <- va.tab[,,"male"]  
va.tabf <- va.tab[,,"female"]

1. Use structable() to create a nicely organized tabular display.

structable(right ~ left + gender, data = va.tab)

## right 1 2 3 4  
## left gender   
## 1 male 821 116 72 43  
## female 1520 234 117 36  
## 2 male 112 494 151 34  
## female 266 1512 362 82  
## 3 male 85 145 583 106  
## female 124 432 1772 179  
## 4 male 35 27 87 331  
## female 66 78 205 492

1. Use xtable() to create a LaTeX or HTML table.

library(xtable)  
va.xtab <- xtable(va.tabm)  
print(va.xtab, type="html")

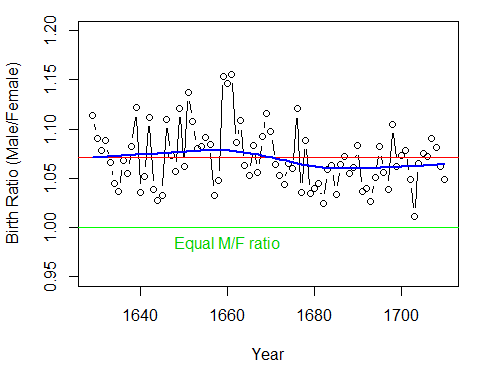
## Chapter 3: Fitting and Graphing Discrete Distributions

### Exercise 3.1

The Arbuthnot data in HistData (Example 3.1) also contains the variable Ratio, giving the ratio of male to female births.

1. Make a plot of Ratio over Year, similar to Figure 3.1. What features stand out? Which plot do you prefer to display the tendency for more male births?

library(HistData)  
data(Arbuthnot, package ="HistData")  
  
# plot of Ratio by Year  
par(mar=c(5,4,1,1)+.1)  
with(Arbuthnot, {  
 plot(Year, Ratio, type='b', ylim=c(.95, 1.2),   
 ylab="Birth Ratio (Male/Female)")  
 abline(h=1, col="green", lwd=1)  
 abline(h=mean(Ratio), col="red")  
 text(x=1660, y=1, "Equal M/F ratio", pos=1, col="green3")   
 Arb.smooth <- loess.smooth(Year,Ratio)  
 lines(Arb.smooth$x, Arb.smooth$y, col="blue", lwd=2)  
})



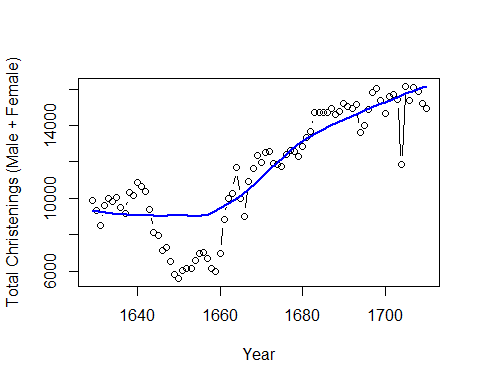
summary(Arbuthnot$Ratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.01 1.05 1.06 1.07 1.09 1.16

The plot is similar to Figure 3.1 in the text. If it is easier to think in terms of probability of a male birth, plotting that directly may be preferable.

1. Plot the total number of christenings, Males + Females or Total (in 000s) over time. What unusual features do you see?

# total number of Christenings  
with(Arbuthnot, {  
 Total= Males + Females  
 plot(Year, Total, type='b', ylab="Total Christenings (Male + Female)")  
 Arb.smooth <- loess.smooth(Year,Total)  
 lines(Arb.smooth$x, Arb.smooth$y, col="blue", lwd=2)  
})

  
 There was a large decline in births between 1640--1660, corresponding to years of plague in England.

### Exercise 3.2

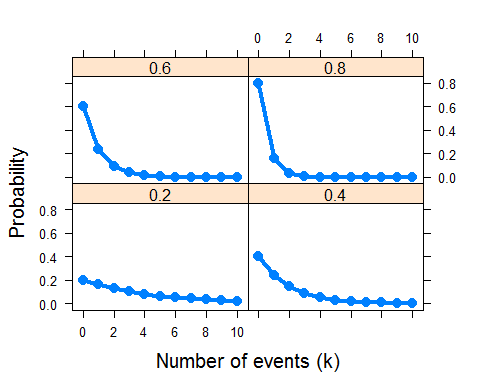
Use the graphical methods illustrated in Section 3.2 to plot a collection of geometric distributions for = 0.2, 0.4, 0.6, 0.8, over a range of values of = 0, 1, ... 10.

1. With xyplot(), try the different plot formats using points connected with lines, as in Figure 3., or using points and lines down to the origin, as in the panels of Figure 3.10.

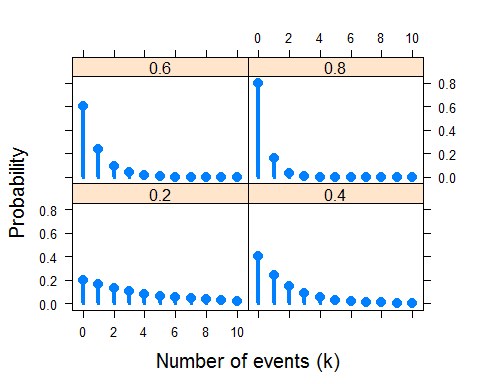
KL <- expand.grid(k = 0 : 10, p = c(0.2, 0.4, 0.6, 0.8))  
geom\_df <- data.frame(KL, prob = dgeom(KL$k, KL$p))  
geom\_df$p = factor(geom\_df$p)  
str(geom\_df)

## 'data.frame': 44 obs. of 3 variables:  
## $ k : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ p : Factor w/ 4 levels "0.2","0.4","0.6",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ prob: num 0.2 0.16 0.128 0.1024 0.0819 ...

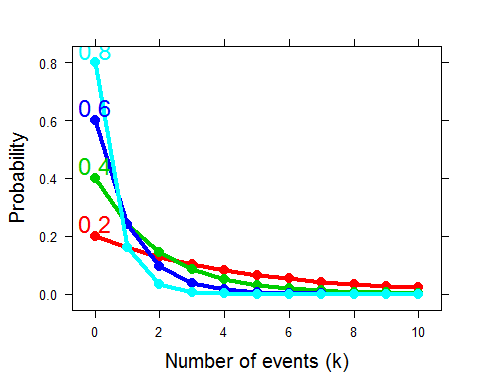
library(lattice)  
mycol<-palette()[2:5]  
xyplot(prob ~ k | p , data = geom\_df, type = c("b"),   
 pch = 16, lwd = 4, cex = 1.25,   
 xlab = list("Number of events (k)", cex = 1.25), layout = c(2,2),   
 ylab = list("Probability", cex = 1.25))



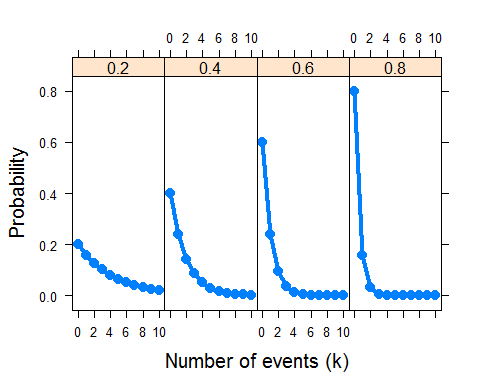
xyplot(prob ~ k | p , data = geom\_df, type = c("h", "p"),   
 pch = 16, lwd = 4, cex = 1.25,   
 xlab = list("Number of events (k)", cex = 1.25), layout = c(2,2),   
 ylab = list("Probability", cex = 1.25))

  
 (b) Also with xyplot(), produce one version of a multi-line plot in a single panel that you think shows well how these distributions change with the probability of success.

geomplt<-xyplot(prob ~ k , data = geom\_df, groups = p,   
 type = c("b"), pch = 16, lwd = 4, cex = 1.25, col = mycol,   
 xlab = list("Number of events (k)", cex = 1.25),   
 ylab = list("Probability", cex = 1.25))  
library(directlabels)  
direct.label(geomplt, list("top.points", cex = 1.5, dl.trans(y = y + 0.1)))

  
 (c) Do the same in a multi-panel version, conditional on

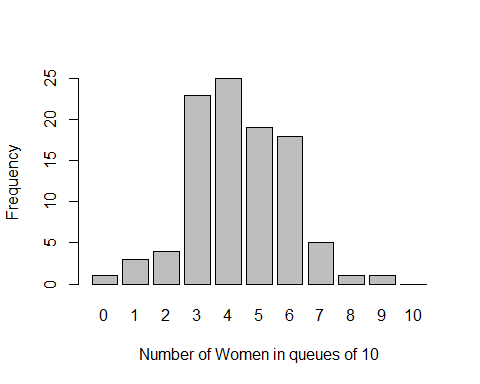
xyplot(prob ~ k | p , data = geom\_df, type = c("b"),   
 pch = 16, lwd = 4, cex = 1.25,   
 xlab = list("Number of events (k)", cex = 1.25), layout = c(4,1),   
 ylab = list("Probability", cex = 1.25))

  
 ### Exercise 3.3

Use the data set WomenQueue to:

1. Produce plots analogous to those shown in Section 3.1 (some sort of bar graph of frequencies)

data("WomenQueue", package = "vcd")  
barplot(WomenQueue,xlab="Number of Women in queues of 10",ylab= "Frequency")

  
 (b) Check for goodness-of-fit to the binomial distribution using the goodfit() methods described in Section 3.3.2.

Note that with goodfit(), you should specify for the binomial distribution as the size parameter.

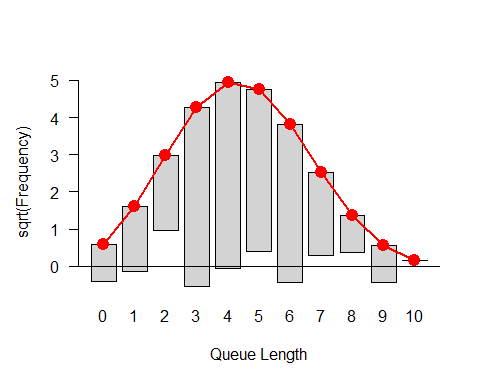
library(vcd)  
gf.women <- goodfit(WomenQueue, type = "binomial", par=list(size=10))  
summary(gf.women)

##   
## Goodness-of-fit test for binomial distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 8.651 8 0.3726

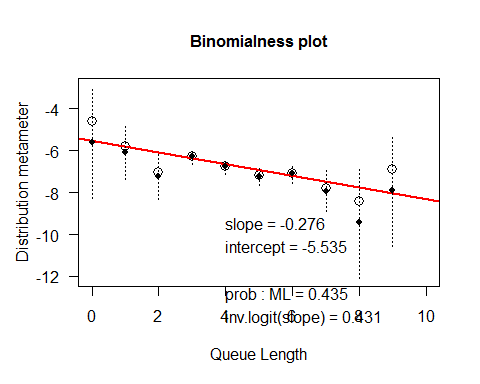
1. Make a reasonable plot showing departure from the binomial distribution.

The simplest plot is the hanging rootogram. An alternative plot is a "binomialness" plot produced by distplot().

plot(gf.women, xlab = "Queue Length")



distplot(WomenQueue, type = "binomial", size=10, xlab = "Queue Length")

  
 (d) Suggest some reasons why the number of women in queues of length 10 might depart from a binomial distribution, .

* Perhaps women (or men) are more prevalent in these queues, so .
* People often join lines in groups, so the observations are unlikely to be independent.

### Exercise 3.4

Continue Example 3.13 on the distribution of male children in families in Saxony by fitting a binomial distribution, , specifying equal probability for boys and girls.

1. Carry out the GOF test for this fixed binomial distribution. What is the ratio of Chi-sqrare/df? What do you conclude?

Note that you need to specify both and as fixed parameters here.

Saxony\_gf <-goodfit(Saxony, type = "binomial", par=list(size=12, prob=.5))  
ss <-summary(Saxony\_gf)

##   
## Goodness-of-fit test for binomial distribution  
##   
## X^2 df P(> X^2)  
## Pearson 249.2 12 2.013e-46  
## Likelihood Ratio 205.4 12 2.494e-37

# The ratio of Chi-square/df  
ss[,"X^2"] / ss[,"df"]

## Pearson Likelihood Ratio   
## 20.77 17.12

The binomial model fits very badly.

1. Test the additional lack of fit for the model Bin(n = 12; p = 1/2 ) compared to the model where is estimated from the data.

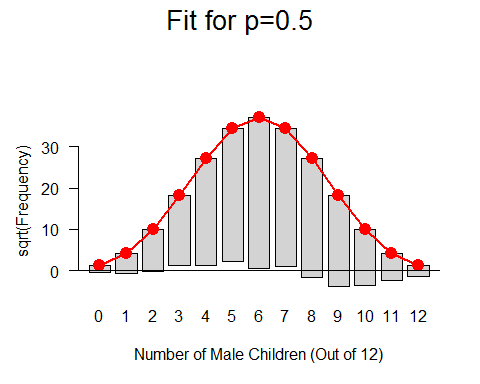
Saxony\_gf2 <- goodfit(Saxony, type = "binomial", par=list(size=12))  
summary(Saxony\_gf2)

##   
## Goodness-of-fit test for binomial distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 97.01 11 6.978e-16

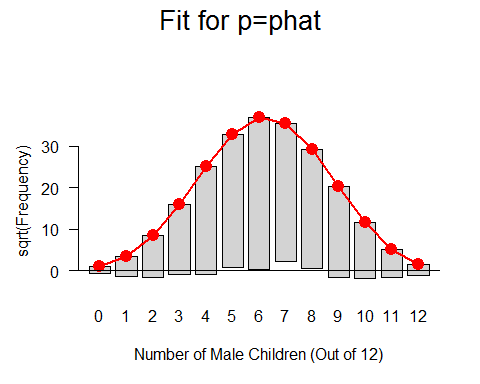
This fits much better, but still not a good fit.

1. Use the plot.goodfit() method to visualize these two models.

plot(Saxony\_gf, main = "Fit for p=0.5", xlab = "Number of Male Children (Out of 12)")



plot(Saxony\_gf2, main = "Fit for p=phat", xlab = "Number of Male Children (Out of 12)")

  
 ### Exercise 3.5

For the Federalist data, the examples in Section 3.3.1 and Section 3.3.2 showed the negative binomial to provide an acceptable fit. Compare this with the simpler special case of geometric distribution, corresponding to .

1. Use goodfit() to fit the geometric distribution. [Hint: use type="nbinomial", but specify size=1 as a parameter.]

fdfit1 <- goodfit(Federalist, type = "binomial", par = list(size=6))  
fdfit1

##   
## Observed and fitted values for binomial distribution  
## with parameters estimated by `ML'   
##   
## count observed fitted pearson residual  
## 0 156 1.307e+02 2.2107  
## 1 63 9.636e+01 -3.3986  
## 2 29 2.960e+01 -0.1097  
## 3 8 4.848e+00 1.4314  
## 4 4 4.467e-01 5.3162  
## 5 1 2.195e-02 6.6009  
## 6 1 4.495e-04 47.1440

fdfit2 <- goodfit(Federalist, type = "nbinomial", par = list(size=1))  
fdfit2

##   
## Observed and fitted values for nbinomial distribution  
## with parameters estimated by `ML with size fixed'   
##   
## count observed fitted pearson residual  
## 0 156 158.1659 -0.17222  
## 1 63 62.6833 0.04001  
## 2 29 24.8422 0.83419  
## 3 8 9.8453 -0.58810  
## 4 4 3.9018 0.04970  
## 5 1 1.5463 -0.43935  
## 6 1 0.6128 -0.01504

1. Compare the negative binomial and the geometric models statistically, by a likelihood-ratio test of the difference between these two models.

summary(fdfit1)

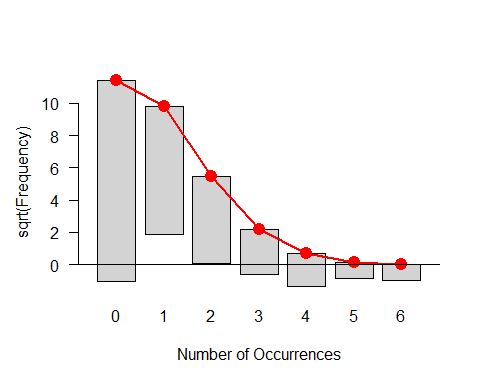
##   
## Goodness-of-fit test for binomial distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 49.03 5 2.193e-09

summary(fdfit2)

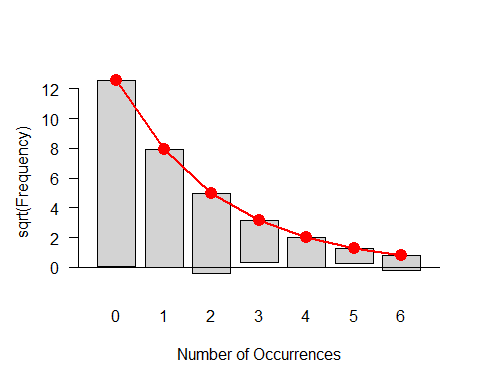
##   
## Goodness-of-fit test for nbinomial distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 2.294 5 0.8071

1. Compare the negative binomial and the geometric models visually by hanging rootograms or other methods.

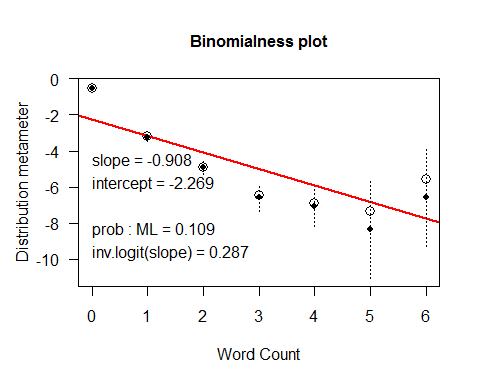
plot(fdfit1)



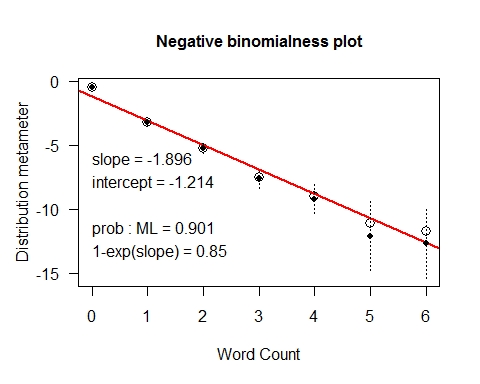
plot(fdfit2)



distplot(Federalist, type = "binomial", size=6, xlab = "Word Count")



distplot(Federalist, type = "nbinomial", size=6, xlab = "Word Count")

  
 ### Exercise 3.6

Mosteller and Wallace (1963, Table 2.4) give the frequencies, , of counts of other selected marker words in 247 blocks of text known to have been written by Alexander Hamilton. The data below show the occurrences of the word *upon*, that Hamilton used much more than did James Madison.

1. Read these data into R and construct a one-way table of frequencies of counts or a matrix or data frame with frequencies in the first column and the corresponding counts in the second column, suitable for use with goodfit().

goodfit() requires its first argument to be either a one-way table (from xtabs()), or a data.frame with frequencies in the *first* column and the corresponding counts in the second column. Both of the following forms will work.

count <- 0:5  
Freq <- c(129, 83, 20, 9, 5, 1)  
sum(Freq) # check N

## [1] 247

(Upon <- data.frame(Freq, count)) # as a data.frame

## Freq count  
## 1 129 0  
## 2 83 1  
## 3 20 2  
## 4 9 3  
## 5 5 4  
## 6 1 5

(Upon.tab <- xtabs(Freq ~ count, data=Upon)) # one-way table

## count  
## 0 1 2 3 4 5   
## 129 83 20 9 5 1

1. Fit and plot the Poisson model for these frequencies.

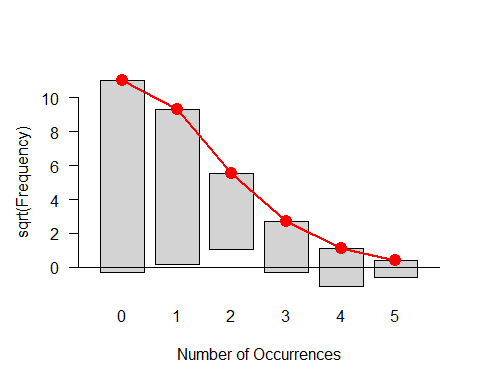
(up0 <- goodfit(Upon, type="poisson"))

##   
## Observed and fitted values for poisson distribution  
## with parameters estimated by `ML'   
##   
## count observed fitted pearson residual  
## 0 129 121.6182 0.6694  
## 1 83 86.1667 -0.3411  
## 2 20 30.5246 -1.9049  
## 3 9 7.2089 0.6671  
## 4 5 1.2769 3.2948  
## 5 1 0.1809 1.7580

summary(up0)

##   
## Goodness-of-fit test for poisson distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 13.14 4 0.01062

plot(up0)

  
 (c) Fit and plot the negative binomial model for these frequencies.

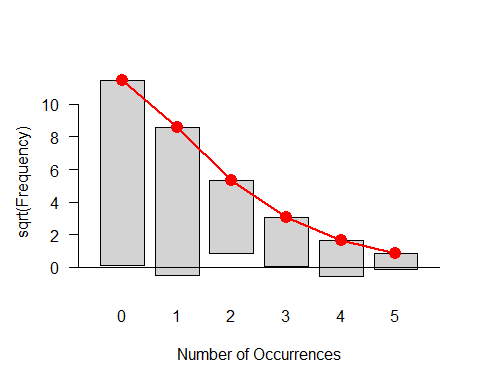
(up1 <- goodfit(Upon, type="nbinomial"))

##   
## Observed and fitted values for nbinomial distribution  
## with parameters estimated by `ML'   
##   
## count observed fitted pearson residual  
## 0 129 131.6594 -0.23177  
## 1 83 73.8942 1.05928  
## 2 20 28.4155 -1.57870  
## 3 9 9.2532 -0.08323  
## 4 5 2.7407 1.36474  
## 5 1 0.7633 -0.03643

summary(up1)

##   
## Goodness-of-fit test for nbinomial distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 6.031 3 0.1101

plot(up1)

  
 (d) What do you conclude?

The negative binomial model fits better than the Poisson.

### Exercise 3.7

The data frame Geissler in the vcdExtra package contains the complete data from Geissler's (1889) tabulation of family sex composition in Saxony. The table below gives the number of boys in families of size 11.

boys 0 1 2 3 4 5 6 7 8 9 10 11   
Freq 8 72 275 837 1,540 2,161 2,310 1,801 1,077 492 93 24

1. Read these data into R.

See Exercise 2.6, which calculates sax11 in the form of a data frame.

1. Following Example 3.13, use goodfit() to fit the binomial model and plot the results. Is there an indication that the binomial does not fit these data?

sax11.tab <- xtabs(Freq ~ boys, data=sax11)  
goodfit(sax11.tab, type="binomial", par=list(size=11))

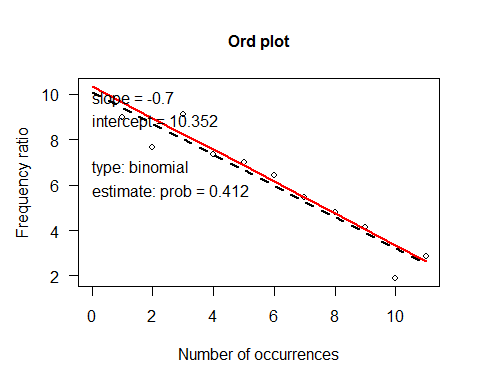
##   
## Observed and fitted values for binomial distribution  
## with parameters estimated by `ML'   
##   
## count observed fitted pearson residual  
## 0 8 3.562 2.352  
## 1 72 41.948 4.640  
## 2 275 224.572 3.365  
## 3 837 721.363 4.305  
## 4 1540 1544.756 -0.121  
## 5 2161 2315.602 -3.213  
## 6 2310 2479.363 -3.401  
## 7 1801 1896.217 -2.187  
## 8 1077 1015.159 1.941  
## 9 492 362.317 6.813  
## 10 93 77.588 1.750  
## 11 24 7.552 5.985

summary(goodfit(sax11.tab, type="binomial", par=list(size=11)))

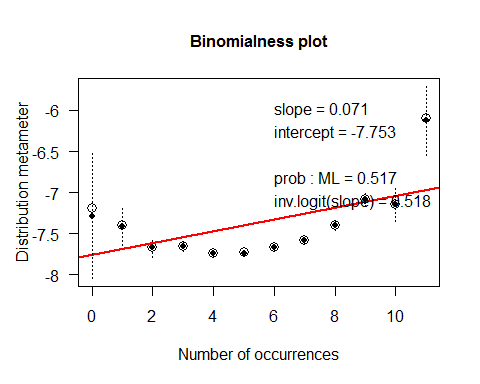
##   
## Goodness-of-fit test for binomial distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 148.1 10 9.213e-27

1. Diagnose the form of the distribution using the methods described in Section 3.4.

Ord\_plot(sax11.tab)



distplot(sax11.tab, type="binomial", size=11)



1. Try fitting the negative binomial distribution, and use distplot() to diagnose whether the negative binomial is a reasonable fit.

goodfit(sax11.tab, type="nbinomial", par=list(size=11))

##   
## Observed and fitted values for nbinomial distribution  
## with parameters estimated by `ML with size fixed'   
##   
## count observed fitted pearson residual  
## 0 8 109.1 -9.680  
## 1 72 409.1 -16.667  
## 2 275 836.6 -19.417  
## 3 837 1235.7 -11.341  
## 4 1540 1474.1 1.717  
## 5 2161 1507.3 16.839  
## 6 2310 1369.9 25.398  
## 7 1801 1134.0 19.808  
## 8 1077 869.6 7.032  
## 9 492 625.7 -5.346  
## 10 93 426.5 -16.150  
## 11 24 277.5 -25.400

### Exercise 3.8

The data frame Bundesliga gives a similar data set to that for UK soccer scores (UKSoccer) examined in Example 3.9, but over a wide range of years. The following lines calculate a two-way table, BL1995, of home-team and away-team goals for the 306 games in the year 1995.

data("Bundesliga", package = "vcd")   
BL1995 <-xtabs(~ HomeGoals + AwayGoals, data = Bundesliga,   
 subset = (Year == 1995))  
BL1995

## AwayGoals  
## HomeGoals 0 1 2 3 4 5 6  
## 0 26 16 13 5 0 1 0  
## 1 19 58 20 5 4 0 1  
## 2 27 23 20 5 1 1 1  
## 3 14 11 10 4 2 0 0  
## 4 3 5 3 0 0 0 0  
## 5 4 1 0 1 0 0 0  
## 6 1 0 0 1 0 0 0

1. As in Example 3.9, find the one-way distributions of HomeGoals, AwayGoals, and TotalGoals = HomeGoals + AwayGoals.

BL.df <- as.data.frame(BL1995, stringsASFactors=FALSE)  
BL.df <- within(BL.df, {  
 HomeGoals <- as.numeric(HomeGoals)  
 AwayGoals <- as.numeric(AwayGoals)  
 TotalGoals <- HomeGoals + AwayGoals  
 })  
# marginal distributions  
(BL.home <- xtabs(Freq ~ HomeGoals, data=BL.df))

## HomeGoals  
## 1 2 3 4 5 6 7   
## 61 107 78 41 11 6 2

(BL.away <- xtabs(Freq ~ AwayGoals, data=BL.df))

## AwayGoals  
## 1 2 3 4 5 6 7   
## 94 114 66 21 7 2 2

(BL.total <- xtabs(Freq ~ TotalGoals, data=BL.df))

## TotalGoals  
## 2 3 4 5 6 7 8 9 10 11 12 13 14   
## 26 35 98 62 39 29 10 4 2 1 0 0 0

1. Use goodfit() to fit and plot the Poisson distribution to each of these. Does the Poisson seem to provide a reasonable fit?
2. Use distplot() to assess fit of the Poisson distribution.
3. What circumstances of scoring goals in soccer might cause these distributions to deviate from Poisson distributions?

### Exercise 3.9

Repeat the exercise above, this time using the data for all years in which there was the standard number (306) of games, that is for Year>1965, tabulated as shown below.

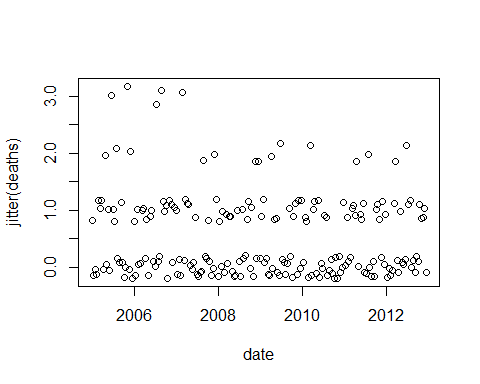
BL <-xtabs(~ HomeGoals + AwayGoals, data = Bundesliga,   
 subset = (Year > 1965))

### Exercise 3.10

Using the data CyclingDeaths introduced in Example 3.6 and the one-way frequency table CyclingDeaths.tab <- table(CyclingDeaths$deaths),

1. Make a sensible plot of the number of deaths over time. For extra credit, add a smoothed curve (e.g., using lines(lowess(...))).

data("CyclingDeaths", package="vcdExtra")  
CyclingDeaths.tab <- table(CyclingDeaths$deaths)  
plot(jitter(deaths) ~ date, data=CyclingDeaths)



1. Test the goodness of fit of the table CyclingDeaths.tab to a Poisson distribution statistically using goodfit().
2. Continue this analysis using a rootogram() and distplot().
3. Write a one-paragraph summary of the results of these analyses and your conclusions.

### Exercise 3.11

The one-way table, Depends, in vcdExtra [not] shown below gives the frequency distribution of the number of dependencies declared in 4,983 R packages maintained on the CRAN distribution network on January 17, 2014. That is, there were 986 packages that had no dependencies, 1,347 packages that depended on one other package, ... up to 2 packages that depended on 14 other packages.

1. Make a bar plot of this distribution.
2. Use Ord\_plot() to see if this method can diagnose the form of the distribution.
3. Try to fit a reasonable distribution to describe dependencies among R packages.

### Exercise 3.12

How many years does it take to get into the baseball Hall of Fame? The Lahman (Friendly, 2014b) package provides a complete record of historical baseball statistics from 1871 to the present. One table, HallOfFame, records the history of players nominated to the Baseball Hall of Fame, and those eventually inducted.

The table [not shown] below, calculated in help(HallOfFame, package="Lahman"), records the distribution of the number of years taken (from first nomination) for the 109 players in the Hall of Fame to be inducted (1936--present). Note that years==0 does not, and cannot, occur in this table, so the distribution is restricted to positive counts.

Such distributions are called ***zero-truncated distributions***. Such distributions are like the ordinary ones, but with the probability of zero being zero. Thus the other probabilities are scaled up (i.e., divided by ) so they sum to 1.

1. For the Poisson distribution, show that the zero-truncated probability function can be expressed in the form
2. Show that the mean is .
3. Enter these data into R as a one-way table, and use goodfit() to fit the standard Poisson distribution, as if you hadn't encountered the problem of zero truncation.

## Chapter 4: Two-Way Contingency Tables

### Exercise 4.1

The data set fat, created below, gives a table recording the level of cholesterol in diet and the presence of symptoms of heart disease for a sample of 23 people.

fat <- matrix(c(6, 4, 2, 11), 2, 2)   
dimnames(fat) <- list(diet = c("LoChol", "HiChol"), disease = c("No", "Yes"))

1. Use chisq.test(fat) to test for association between diet and disease. Is there any indication that this test may not be appropriate here?
2. Use a fourfold display to test this association visually. Experiment with the different options for standardizing the margins, using the margin argument to fourfold. What evidence is shown in different displays regarding whether the odds ratio differs significantly from 1?
3. oddsratio(fat, log = FALSE) will give you a numerical answer. How does this compare to your visual impression from fourfold displays?
4. With such a small sample, Fisher’s exact test may be more reliable for statistical inference. Use fisher.test(fat), and compare these results to what you have observed before.
5. Write a one-paragraph summary of your findings and conclusions for this data set.

### Exercise 4.2

The data set Abortion in vcdExtra gives a table of opinions regarding abortion in relation to sex and status of the respondent. This table has the following structure:

data("Abortion", package = "vcdExtra")   
str(Abortion)

## table [1:2, 1:2, 1:2] 171 152 138 167 79 148 112 133  
## - attr(\*, "dimnames")=List of 3  
## ..$ Sex : chr [1:2] "Female" "Male"  
## ..$ Status : chr [1:2] "Lo" "Hi"  
## ..$ Support\_Abortion: chr [1:2] "Yes" "No"

1. Taking support for abortion as the outcome variable, produce fourfold displays showing the association with sex, stratified by status.
2. Do the same for the association of support for abortion with status, stratified by sex.
3. For each of the problems above, use oddsratio to calculate the numerical values of the odds ratio, as stratified in the question.
4. Write a brief summary of how support for abortion depends on sex and status.

### Exercise 4.3

The JobSat table on income and job satisfaction created in Example 2.5 is contained in the vcdExtra package.

1. Carry out a standard test for association between income and job satisfaction. Is there any indication that this test might not be appropriate? Repeat this test using simulate.p.value = TRUE to obtain a Monte Carlo test that does not depend on large sample size. Does this change your conclusion?
2. Both variables are ordinal, so CMH tests may be more powerful here. Carry out that analysis. What do you conclude?

### Exercise 4.4

The Hospital data in vcd gives a table relating the length of stay (in years) of 132 long-term schizophrenic patients in two London mental hospitals with the frequency of visits by family and friends.

1. Carry out a test for association between the two variables.
2. Use assocstats() to compute association statistics. How would you describe the strength of association here?
3. Produce an association plot for these data, with visit frequency as the vertical variable. Describe the pattern of the relation you see here.
4. Both variables can be considered ordinal, so CMH tests may be useful here. Carry out that analysis. Do any of the tests lead to different conclusions?

### Exercise 4.5

Continuing with the Hospital data:

1. Try one or more of the following other functions for visualizing two-way contingency tables with this data: plot(), tile(), mosaic(), and spineplot(). [For all except spineplot, it is useful to include the argument shade=TRUE].
2. Comment on the differences among these displays for understanding the relation between visits and length of stay.

### Exercise 4.6

The two-way table Mammograms in vcdExtra gives ratings on the severity of diagnosis of 110 mammograms by two raters.

1. Assess the strength of agreement between the raters using Cohen’s , both unweighted and weighted.
2. Use agreementplot() for a graphical display of agreement here.
3. Compare the Kappa measures with the results from assocstats(). What is a reasonable interpretation of each of these measures?

### Exercise 4.7

Agresti and Winner (1997) gave the data in Table 4.1 on the ratings of 160 movies by the reviewers Gene Siskel and Roger Ebert for the period from April 1995 through September 1996. The rating categories were Con (“thumbs down”), Mixed, and Pro (“thumbs up”).

Ebert  
 Con Mixed Pro Total   
Con 24 8 13 45   
Mixed 8 13 11 32   
Pro 10 9 64 83   
Total 42 30 88 160

1. Assess the strength of agreement between the raters using Cohen’s , both unweighted and weighted.
2. Use agreementplot() for a graphical display of agreement here.
3. Assess the hypothesis that the ratings are *symmetric* around the main diagonal, using an appropriate test. *Hint*: Symmetry for a square table $\mat{T}$ means that for . The expected frequencies under the hypothesis of symmetry are the average of the off-diagonal cells, $\mat{E} = (\mat{T} + \mat{T}\trans) / 2$.
4. Compare the results with the output of mcnemar.test().

### Exercise 4.8

For the VisualAcuity data set:

1. Use the code shown in the text to create the table form, VA.tab.
2. Perform the CMH tests for this table.
3. Use the woolf\_test() described in sec:twoway-homog to test whether the association between left and right eye acuity can be considered the same for men and women.

### Exercise 4.9

The graph in Figure 4.23 may be misleading, in that it doesn’t take into account of the differing capacities of the 18 life boats on the *Titanic*, given in the variable cap in the Lifeboats data.

1. Calculate a new variable, pctloaded, as the percentage loaded relative to the boat capacity.
2. Produce a plot similar to Figure 4.23, showing the changes over time in this measure.

## Chapter 5: Mosaic Displays for n-Way Tables

### Exercise 5.1

The data set criminal in the package logmult gives the table below of the number of men aged 15–19 charged with a criminal case for whom charges were dropped in Denmark from 1955–1958.

data("criminal", package = "logmult")  
criminal

## Age  
## Year 15 16 17 18 19  
## 1955 141 285 320 441 427  
## 1956 144 292 342 441 396  
## 1957 196 380 424 462 427  
## 1958 212 424 399 442 430

1. Use loglm() to test whether there is an association between Year and Age. Is there evidence that dropping of charges in relation to age changed over the years recorded here?

b.Use mosaic() with the option shade=TRUE to display the pattern of signs and magnitudes of the residuals. Compare this with the result of mosaic() using “Friendly shading,” from the option gp=shading\_Friendly. Describe verbally what you see in each regarding the pattern of association in this table.

### Exercise 5.2

[lab:mosaic-crash] The data set AirCrash in vcdExtra gives a database of all crashes of commercial airplanes between 1993–2015, classified by Phase of the flight and Cause of the crash. How can you best show is the nature of the association between these variables in a mosaic plot? Start by making a frequency table, aircrash.tab:

data("AirCrash", package = "vcdExtra")   
aircrash.tab <- xtabs(~ Phase + Cause, data = AirCrash)

1. Make a default mosaic display of the data with shade=TRUE and interpret the pattern of the high-frequency cells.
2. The default plot has overlapping labels due to the uneven marginal frequencies relative to the lengths of the category labels. Experiment with some of the labeling\_args options (abbreviate, rot\_labels, etc.) to see if you can make the plot more readable. *Hint*: a variety of these are illustrated in §4.1 of vignette("strucplot").
3. The levels of Phase and Cause are ordered alphabetically (because they are factors). Experiment with other orderings of the rows/columns to make interpretation clearer, e.g., ordering Phase temporally or ordering both factors by their marginal frequency.

### Exercise 5.3

The Lahman package contains comprehensive data on baseball statistics for Major League Baseball from 1871 through 2012. For all players, the Master table records the handedness of players, in terms of throwing (L, R) and batting (B, L, R), where B indicates “both.” The table below was generated using the following code:

library(Lahman)   
data("Master", package = "Lahman")   
basehands <- with(Master, table(throws, bats))

1. Use the code above, or else enter these data into a frequency table in R.
2. Construct mosaic displays showing the relation of batting and throwing handedness, split first by batting and then by throwing.
3. From these displays, what can be said about players who throw with their left or right hands in terms of their batting handedness?

### Exercise 5.4

A related analysis concerns differences in throwing handedness among baseball players according to the fielding position they play. The following code calculates such a frequency table.

MasterFielding <- data.frame(merge(Master, Fielding, by = "playerID"))   
throwPOS <- with(MasterFielding, table(POS, throws))

1. Make a mosaic display of throwing hand vs. fielding position.
2. Calculate the percentage of players throwing left-handed by position. Make a sensible graph of this data.
3. Re-do the mosaic display with the positions sorted by percentage of left-handers.
4. Is there anything you can say about positions that have very few left-handed players?

### Exercise 5.5

For the Bartlett data described in ex:bartlett, fit the model of no three-way association, in 5.2.

1. Summarize the goodness of fit for this model, and compare to simpler models that omit one or more of the two-way terms.
2. Use a mosaic-like display to show the lack of fit for this model.

### Exercise 5.6

Red core disease, caused by a fungus, is not something you want if you are a strawberry. The data set jansen.strawberry from the agridat gives a frequency data frame of counts of damage from this fungus from a field experiment reported by . See the help file for details. The following lines create a table of crossings of 3 male parents with 4 (different) female parents, recording the number of plants in four blocks of 9 or 10 plants each showing red core disease in three ordered categories, C1, C2, or C3.

data("jansen.strawberry", package = "agridat")  
  
dat <- jansen.strawberry   
dat <- transform(dat,   
 category = ordered(category,   
 levels = c('C1','C2','C3')))  
levels(dat$male) <- paste0("M", 1:3)  
levels(dat$female) <- paste0("F", 1:4)  
  
jansen.tab <- xtabs(count ~ male + female + category, data = dat)  
names(dimnames(jansen.tab)) <- c("Male parent", "Female parent", "Disease category")   
ftable(jansen.tab)

## Disease category C1 C2 C3  
## Male parent Female parent   
## M1 F1 6 13 20  
## F2 8 15 17  
## F3 13 10 16  
## F4 8 21 11  
## M2 F1 5 13 21  
## F2 9 16 14  
## F3 16 9 15  
## F4 12 13 14  
## M3 F1 5 10 24  
## F2 13 12 15  
## F3 3 14 22  
## F4 18 9 13

1. Use pairs(jansen.tab, shade=TRUE) to display the pairwise associations among the three variables. Describe how disease category appears to vary with male and female parent. Why is there no apparent association between male and female parent?
2. As illustrated in fig:HE-fill, use mosaic to prepare a 3-way mosaic plot with the tiles colored in increasing shades of some color according to disease category. Describe the pattern of category C3 in relation to male and female parent. (Hint: the highlighting arguments are useful here.)
3. With category as the response variable, the minimal model for association is MF,C, or ~ 1\*2 + 3. Fit this model using loglm() and display the residuals from this model with mosaic. Describe the pattern of lack of fit of this model.

### Exercise 5.7

The data set caith in MASS gives another classic table tabulating hair color and eye color, this for people in Caithness, Scotland, originally from . The data is stored as a data frame of cell frequencies, whose rows are eye colors and whose columns are hair colors.

data("caith", package = "MASS")   
caith

## fair red medium dark black  
## blue 326 38 241 110 3  
## light 688 116 584 188 4  
## medium 343 84 909 412 26  
## dark 98 48 403 681 85

1. The loglm and mosaic functions don’t understand data in this format, so use Caith <- as.matrix(caith) to convert to array form. Examine the result, and use names(dimnames(Caith))<-c() to assign appropriate names to the row and column dimensions.
2. Fit the model of independence to the resulting matrix using loglm().
3. Calculate and display the residuals for this model.
4. Create a mosaic display for this data.

### Exercise 5.8

The HairEyePlace data in vcdExtra gives similar data on hair color and eye color, for both Caithness and Aberdeen as a table.

1. Prepare separate mosaic displays, one for each of Caithness and Aberdeen. Comment on any difference in the pattern of residuals.
2. Construct conditional mosaic plots, using the formula ~ Hair + Eye | Place and both mosaic and cotabplot. It is probably more useful here to suppress the legend in these plots. Comment on the difference in what is shown in the two displays.

### Exercise 5.9

[lab:mosaic-accident] used a 4-way table of frequencies of traffic accident victims in France in 1958 to illustrate his scheme for classifying data sets by numerous variables, each of which could have various types and could be assigned to various visual attributes. His data are contained in Accident in vcdExtra, a frequency data frame representing his table of the variables age, result (died or injured), mode of transportation, and gender.

data("Accident", package = "vcdExtra")   
str(Accident, vec.len=2)

## 'data.frame': 80 obs. of 5 variables:  
## $ age : Ord.factor w/ 5 levels "0-9"<"10-19"<..: 5 5 5 5 5 ...  
## $ result: Factor w/ 2 levels "Died","Injured": 1 1 1 1 1 ...  
## $ mode : Factor w/ 4 levels "4-Wheeled","Bicycle",..: 4 4 2 2 3 ...  
## $ gender: Factor w/ 2 levels "Female","Male": 2 1 2 1 2 ...  
## $ Freq : int 704 378 396 56 742 ...

1. Use loglm to fit the model of mutual independence, Freq ~ age+mode+gender+result to this data set.
2. Use mosaic to produce an interpretable mosaic plot of the associations among all variables under the model of mutual independence. Try different orders of the variables in the mosaic. (*Hint*: the abbreviate component of the labeling\_args argument to mosaic will be useful to avoid some overlap of the category labels.)
3. Treat result ("Died" vs. "Injured") as the response variable, and fit the model Freq ~ age\*mode\*gender + result that asserts independence of result from all others jointly.
4. Construct a mosaic display for the residual associations in this model. Which combinations of the predictor factors are more likely to result in death?

### Exercise 5.10

[lab:mosaic-vietnam]The data set Vietnam in vcdExtra gives a contingency table in frequency form reflecting a survey of student opinion on the Vietnam War at the University of North Carolina in May 1967. The table variables are sex, year in school, and response, which has categories: (A) Defeat North Vietnam by widespread bombing and land invasion; (B) Maintain the present policy; (C) De-escalate military activity, stop bombing and begin negotiations; (D) Withdraw military forces immediately. How does the chosen response vary with sex and year?

data("Vietnam", package = "vcdExtra")   
str(Vietnam)

## 'data.frame': 40 obs. of 4 variables:  
## $ sex : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...  
## $ year : int 1 1 1 1 2 2 2 2 3 3 ...  
## $ response: Factor w/ 4 levels "A","B","C","D": 1 2 3 4 1 2 3 4 1 2 ...  
## $ Freq : int 13 19 40 5 5 9 33 3 22 29 ...

1. With response (R) as the outcome variable and year (Y) and sex (S) as predictors, the minimal baseline model is the model of joint independence, R,YS. Fit this model, and display it in a mosaic plot.
2. Construct conditional mosaic plots of the response versus year separately for males and females. Describe the associations seen here.
3. Follow the methods shown in ex:employ to fit separate models of independence for the levels of sex, and the model of conditional independence, $R \perp Y \given S$. Verify that the decomposition of $\GSQ$ in eq:partial1 holds for these models.
4. Construct a useful 3-way mosaic plot of the data for the model of conditional independence.

### Exercise 5.11

Consider the models for 4-way tables shown in 5.3.

1. For each model, give an independence interpretation. For example, the model of mutual independence corresponds to .
2. Use the functions shown in the table together with loglin2formula() to print the corresponding model formulas for each.

### Exercise 5.12

[lab:mosaic-titanic] The dataset Titanic classifies the 2,201 pasengers and crew of the *Titanic* by Class (1st, 2nd, 3rd, Crew), Sex, Age, and Survived. Treating Survived as the response variable,

1. Fit and display a mosaic plot for the baseline model of joint independence, CGA,S. Describe the remaining pattern of associations.
2. Do the same for a “main effects” model that allows two-way associations between each of C, G, and A with S.
3. What three-way association term should be added to this model to allow for greater survival among women and children? Does this give an acceptable fit?
4. Test and display models that allow additional three-way associations until you obtain a reasonable fit.

## Chapter 6: Correspondence Analysis

These solutions use an updated version of the ca package, v. 0.64 or greater.

### Exercise 6.1

The JobSat data in vcdExtra gives a table recording job satisfaction in relation to income.

1. Carry out a simple correspondence analysis on this table. How much of the inertia is accounted for by a one-dimensional solution? How much by a two-dimensional solution?

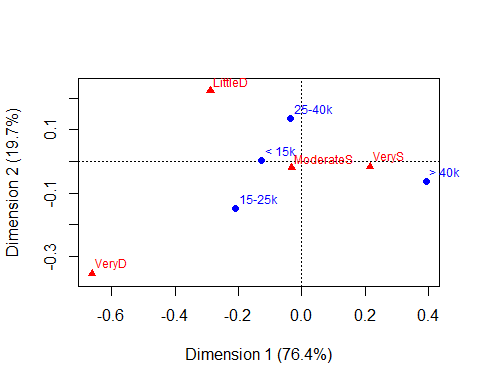
data("JobSat", package="vcdExtra")  
library(ca)  
jobsat.ca <- ca(JobSat)  
 # just show the scree plot  
summary(jobsat.ca, rows=FALSE, columns=FALSE)

##   
## Principal inertias (eigenvalues):  
##   
## dim value % cum% scree plot   
## 1 0.047496 76.4 76.4 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## 2 0.012248 19.7 96.1 \*\*\*\*\*   
## 3 0.002397 3.9 100.0 \*   
## -------- -----   
## Total: 0.062141 100.0

The 1D solution accounts for 76.4%, the 2D solution accounts for 96.1%.

1. Plot the 2D CA solution. To what extent can you consider the association between job satisfaction and income 'explained' by the ordinal nature of these variables?

plot(jobsat.ca)

  
 Job satisfaction is ordered as expected by its ordinal levels along Dimension 1. The levels of income in this plot do not appear to be ordered according to the quantitative levels they represent.

### Exercise 6.2

Refer to Exercise 5.1 in Chapter 5. Carry out a simple correspondence analysis on the table criminal from the logmult package.

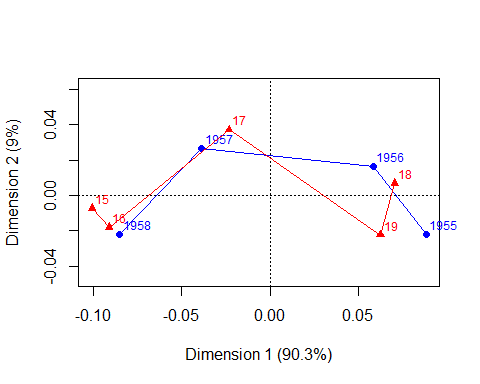
1. What percentages of the Pearson for association are explained by the various dimensions?

data("criminal", package = "logmult")  
criminal.ca <- ca(criminal)  
 # just show the scree plot  
summary(criminal.ca, rows=FALSE, columns=FALSE)

##   
## Principal inertias (eigenvalues):  
##   
## dim value % cum% scree plot   
## 1 0.004939 90.3 90.3 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## 2 0.000491 9.0 99.3 \*\*   
## 3 3.8e-050 0.7 100.0   
## -------- -----   
## Total: 0.005468 100.0

1. Plot the 2D correspondence analysis solution. Describe the pattern of association between year and age.

plot(criminal.ca, lines=TRUE)

  
 The category points for both year and age vary systematically over Dimension 1. There were more younger men in later years, and more older in earlier years.

### Exercise 6.3

Refer to Exercise 5.2 for a description of the AirCrash data from the vcdExtra package. Carry out a simple correspondence analysis on the table of Phase of the flight and Cause of the crash.

1. What percentages of the Pearson for association are explained by the various dimensions?

aircrash.tab was calculated in Exercise 5.2.

aircrash.tab

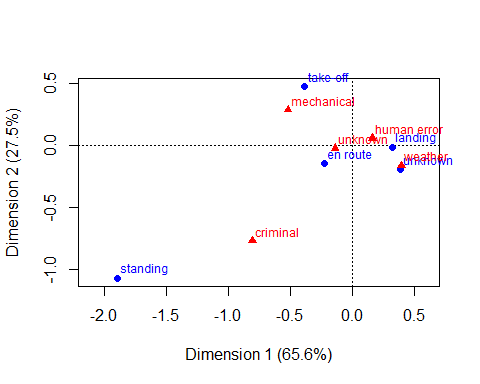
## Cause  
## Phase criminal human error mechanical unknown weather  
## en route 16 63 29 25 24  
## landing 4 114 19 18 55  
## standing 2 0 2 0 0  
## take-off 1 29 24 8 3  
## unknown 0 1 0 1 1

aircrash.ca <- ca(aircrash.tab)  
 # just show the scree plot  
summary(aircrash.ca, rows=FALSE, columns=FALSE)

##   
## Principal inertias (eigenvalues):  
##   
## dim value % cum% scree plot   
## 1 0.123002 65.6 65.6 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## 2 0.051548 27.5 93.1 \*\*\*\*\*\*\*   
## 3 0.012340 6.6 99.7 \*\*   
## 4 0.000562 0.3 100.0   
## -------- -----   
## Total: 0.187452 100.0

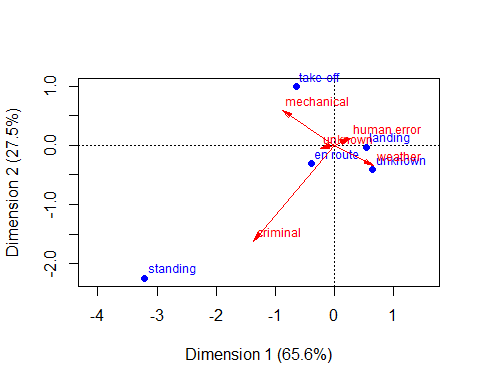
1. Plot the 2D correspondence analysis solution. Describe the pattern of association between phase and cause. How would you interpret the dimensions?

plot(aircrash.ca)

  
 Dimension 1 largely distinguishes standing from landing, where the main causes are human error and weather problems. Dimension 2 largely distinguished standing from take-off, where mechanical problems are more prevalent.

1. The default plot method uses map="symmetric" with points for both rows and columns. Try using map="symbiplot" with vectors (arrows=) for either rows or columns. (Read help(plot.ca) for a description of these options.)

plot(aircrash.ca, map="symbiplot", arrows=c(FALSE, TRUE))

  
 In the biplot scaling, the relationship of causes to phases can be seen from the projections of the points on the vectors.

### Exercise 6.4

The data set caith in MASS gives a classic table tabulating hair color and eye color of people in Caithness, Scotland, originally from Fisher (1940).

1. Carry out a simple correspondence analysis on this table. How many dimensions seem necessary to account for most of the association in the table?

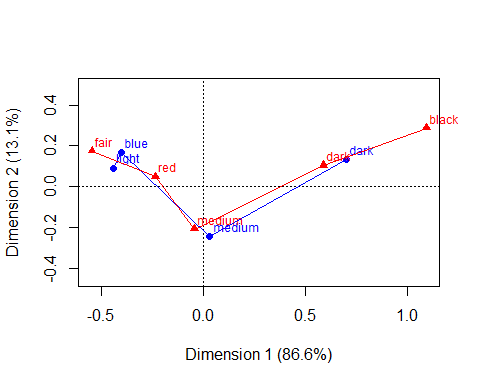
data("caith", package="MASS")  
caith.ca <- ca(caith)  
summary(caith.ca, rows=FALSE, columns=FALSE)

##   
## Principal inertias (eigenvalues):  
##   
## dim value % cum% scree plot   
## 1 0.199245 86.6 86.6 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## 2 0.030087 13.1 99.6 \*\*\*   
## 3 0.000859 0.4 100.0   
## -------- -----   
## Total: 0.230191 100.0

One could argue that a 1D solution is adequate here. The 2D solution is essentially complete.

1. Plot the 2D solution. The interpretation of the first dimension should be obvious; is there any interpretation for the second dimension?

plot(caith.ca, lines=TRUE)

  
 Dimension 1 is the obvious light--dark dimension for hair color and eye color (but not that blue and light eyes have quite similar positions). Dimension 2 could be called "extremes vs. middle", but in CA results with largely 1D association, this effect is common, often called the "horseshoe" or "arch" effect.

### Exercise 6.5

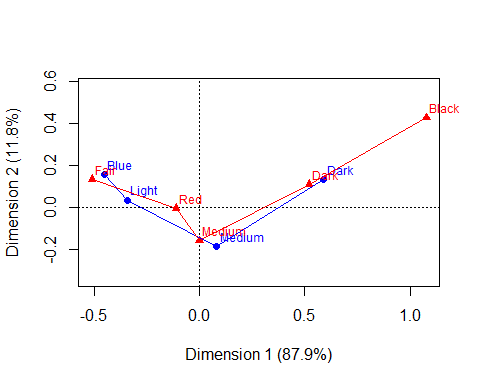
The same data, plus a similar table for Aberdeen, are given as a three-way table as HairEyePlace in vcdExtra.

1. Carry out a similar correspondence analysis to the last exercise for the data from Aberdeen. Comment on any differences in the placement of the category points.

data("HairEyePlace", package="vcdExtra")  
aberdeen <- HairEyePlace[,,2]  
aberdeen.ca <- ca(aberdeen)  
summary(aberdeen.ca, rows=FALSE, columns=FALSE)

##   
## Principal inertias (eigenvalues):  
##   
## dim value % cum% scree plot   
## 1 0.144853 87.9 87.9 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## 2 0.019405 11.8 99.7 \*\*\*   
## 3 0.000488 0.3 100.0   
## -------- -----   
## Total: 0.164746 100.0

plot(aberdeen.ca, lines=TRUE)

  
Quite similar, except in the Aberdeen sample, there is a larger difference between blue and light eyes.

1. Analyze the three-way table, stacked to code hair color and place interactively, i.e., for the loglinear model . What does this show?

### Exercise 6.6

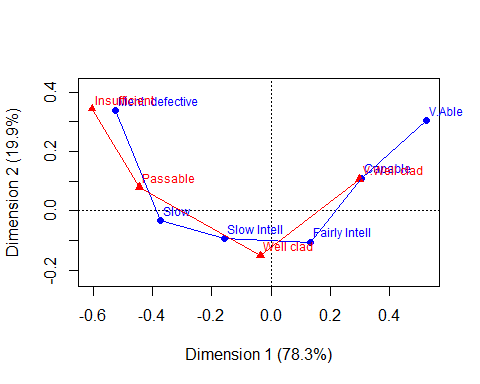
The data set Gilby in vcdExtra gives a classic (but now politically incorrect) table of English schoolboys classified according to their clothing and their teacher's rating of "dullness"" (lack of intelligence).

1. Compute and plot a correspondence analysis for this data. Write a brief description and interpretation of these results.

data("Gilby", package="vcdExtra")  
gilby.ca <- ca(Gilby)  
summary(gilby.ca, rows=FALSE, columns=FALSE)

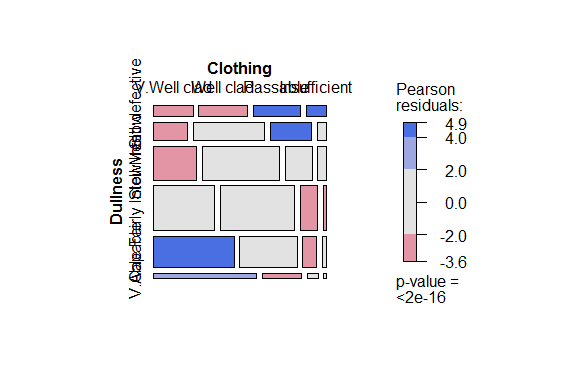
##   
## Principal inertias (eigenvalues):  
##   
## dim value % cum% scree plot   
## 1 0.079346 78.3 78.3 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## 2 0.020118 19.9 98.1 \*\*\*\*\*   
## 3 0.001881 1.9 100.0   
## -------- -----   
## Total: 0.101346 100.0

plot(gilby.ca, lines=TRUE)



1. Make an analogous mosaic plot of this table. Interpret this in relation to the correspondence analysis plot.

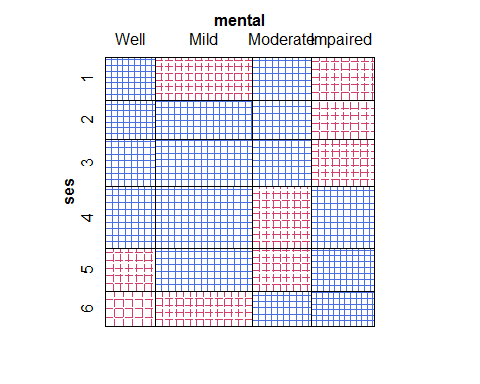
library(vcd)  
mosaic(Gilby, shade=TRUE)



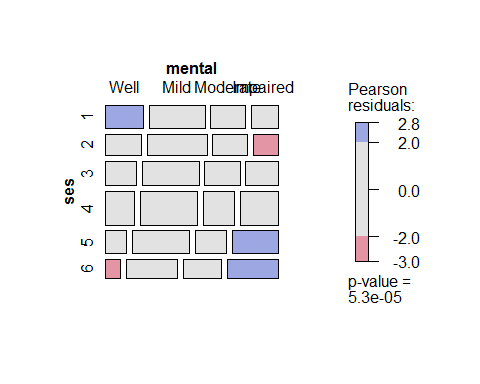
### Exercise 6.7

For the mental health data analyzed in Example 6.2, construct a shaded sieve diagram and mosaic plot. Compare these with the correspondence analysis plot shown in Figure 6.2. What features of the data and the association between SES and mental health status are shown in each?

data("Mental", package="vcdExtra")  
mental.tab <- xtabs(Freq ~ ses + mental, data=Mental)  
sieve(mental.tab, shade=TRUE)



mosaic(mental.tab, shade=TRUE)

  
The association between ses and mental in Figure 6.2 appears to be nearly entirely one dimensional. The sieve plot does have an opposite-corner pattern, but the results are less clear in terms of a unidimensional association. The mosaic display is clearer, and highlights one cell for ses=="2" and mental=="impaired" that deviates from this pattern.

### Exercise 6.8

Simulated data are often useful to help understand the connections between data, analysis methods, and associated graphic displays. Section 6.3.1 illustrated interactive coding in R, using a simulated 4-way table of counts of pets, classified by age, color, and sex, but with no associations because the counts had a constant Poisson mean, .

1. Re-do this example, but in the call to rpois(), specify a non-negative vector of Poisson means to create some associations among the table factors.
2. Use CA methods to determine if and how the structure you created in the data appears in the results.

### Exercise 6.9

The TV data was analyzed using CA in Example 6.4, ignoring the variable Time. Carry out analyses of the 3-way table, reducing the number of levels of Timeto three hourly intervals as shown below.

data("TV", package="vcdExtra")   
# reduce number of levels of Time   
TV.df <-as.data.frame.table(TV)   
levels(TV.df$Time) <-rep(c("8", "9", "10"), c(4, 4, 3))   
TV3 <-xtabs(Freq ~ Day + Time + Network, TV.df)   
structable(Day ~ Network + Time, TV3)

## Day Monday Tuesday Wednesday Thursday Friday  
## Network Time   
## ABC 8 536 861 744 735 1119  
## 9 1401 1205 1022 682 907  
## 10 910 1044 668 349 711  
## CBS 8 1167 646 550 680 509  
## 9 967 959 409 385 544  
## 10 789 798 324 270 426  
## NBC 8 858 1090 512 1927 823  
## 9 946 890 831 1858 590  
## 10 825 588 869 2101 585

1. Use the stacking approach (Section 6.3) to perform a CA of the table with Network and Time coded interactively. You can create this using the as.matrix() method for a "structable" object.

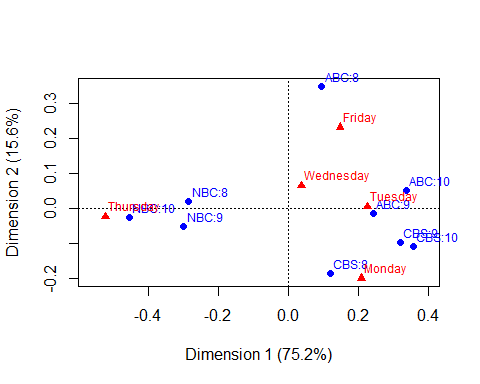
TV3S <-as.matrix(structable(Day ~ Network + Time, TV3), sep=":")

1. What loglinear model is analyzed by this approach?

The model is , asserting that the combinations of Network and Time are independent of Day.

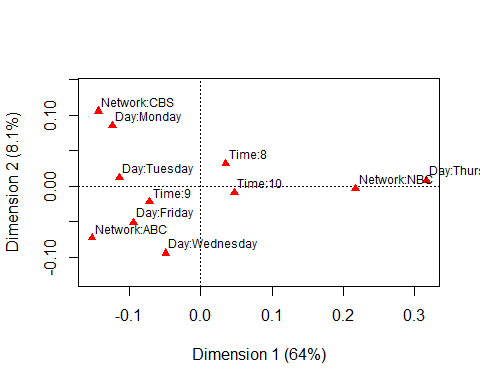
1. Plot the 2D solution. Compare this to the CA plot of the two-way table in Figure 6.4.

TV3S.ca <- ca(TV3S)  
plot(TV3S.ca)



1. Carry out an MCA analysis using mjca() of the three-way table TV3. Plot the 2D solution, and compare this with both the CA plot and the solution for the stacked three-way table.

TV3.mca <- mjca(TV3)  
plot(TV3.mca)

  
 Dimension 1 distinguishes NBC from the other two networks. Dimension 2 distinguishes ABC from CBS.

### Exercise 6.10

Refer to the MCA analysis of the PreSex data in Example 6.8. Use the stacking approach to analyze the stacked table with the combinations of premarital and extramarital sex in the rows and the combinations of gender and marital status in the columns. As suggested in the exercise above, you can use as.matrix(structable()) to create the stacked table.

1. What loglinear model is analyzed by this approach? Which associations are included and which are excluded in this analysis?
2. Plot the 2D CA solution for this analysis. You might want to draw lines connecting some of the row points or column points to aid in interpretation.
3. How does this analysis differ from the MCA analysis shown in Figure 6.10?

### Exercise 6.11

Refer to Exercise 5.10 for a description of the Vietnam data set in vcdExtra.

1. Using the stacking approach, carry out a correspondence analysis corresponding to the loglinear model [R][YS], which asserts that the response is independent of the combinations of year and sex.
2. Construct an informative 2D plot of the solution, and interpret in terms of how the response varies with year for males and females.
3. Use mjca() to carry out an MCA on the three-way table. Make a useful plot of the solution and interpret in terms of the relationship of the response to year and sex.

### Exercise 6.12

Refer to Exercise 5.9 for a description of the Accident data set in vcdExtra. The data set is in the form of a frequency data frame, so first convert to table form.

data("Accident", package="vcdExtra")  
accident.tab <-xtabs(Freq ~ age + result + mode + gender, data=Accident)

1. Use mjca() to carry out an MCA on the four-way table accident.tab.

accident.mca <- mjca(accident.tab)  
summary(accident.mca, columns=FALSE)

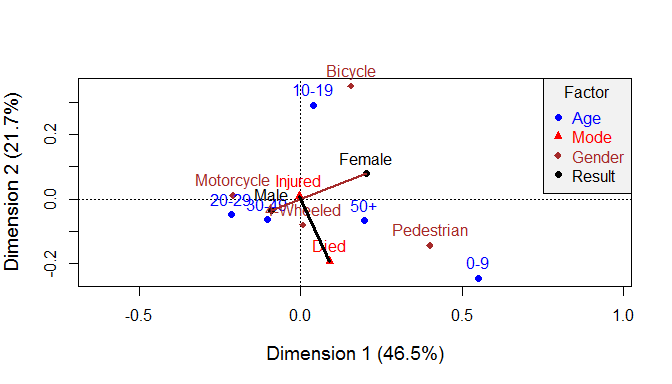
##   
## Principal inertias (eigenvalues):  
##   
## dim value % cum% scree plot   
## 1 0.025429 46.5 46.5 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## 2 0.011848 21.7 68.1 \*\*\*\*\*\*\*   
## 3 0.001889 3.5 71.6 \*   
## 4 0.000491 0.9 72.5   
## -------- -----   
## Total: 0.054700

There seem to be two substantial dimensions to the association among these variables.

1. Construct an informative 2D plot of the solution, and interpret in terms of how the variable result varies in relation to the other factors.

The default plot for "mjca" objects is hard to read. Here, we use the method described in the text to create a customized plot.

res <- plot(accident.mca, labels=0, pch='.', cex.lab=1.2)  
coords <- data.frame(res$cols, accident.mca$factors)   
cols <- c("blue", "red", "brown", "black")  
nlev <- accident.mca$levels.n  
points(coords[,1:2], pch=rep(16:19, nlev), col=rep(cols, nlev))  
text(coords[,1:2], label=coords$level, col=rep(cols, nlev), pos=3, xpd=TRUE)  
lines(Dim2 ~ Dim1, data=coords, subset=factor=="gender", lty=1, lwd=2, col="brown")  
lines(Dim2 ~ Dim1, data=coords, subset=factor=="result", lty=1, lwd=3, col="black")  
legend("topright", legend=c("Age", "Mode", "Gender", "Result"),  
 title="Factor", title.col="black",  
 col=cols, text.col=cols, pch=16:19,  
 bg="gray95")



### Exercise 6.13

The UCBAdmissions data was featured in numerous examples in Chapter 4 (e.g., Example 4.11, Example 4.15) and Chapter 5 (e.g., Example 5.14, Example 5.18).

1. Use mjca() to carry out an MCA on the three-way table UCBAdmissions.
2. Plot the 2D MCA solution in a style similar to that shown in Figure 6.10 and Figure 6.11
3. Interpret the plot. Is there some interpretation for the first dimension? What does the plot show about the relation of admission to the other factors?