DDAR: Solutions and Hints for Exercises

2015-11-30

## Chapter 1

These exercises are all conceptual. There are no hints or solutions.

## Chapter 2

### Exercise 2.2

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.
2. For each department, find the total number of applicants.
3. For each department, find the overall proportion of applicants who were admitted.
4. 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 2.3

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 2.4

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.
2. Find the marginal total of the number of goals scored by each of the home and away teams.
3. Express each of the marginal totals as proportions.
4. 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?

### Exercise 2.5

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.
2. Select the columns for boys and Freq.
3. Use xtabs with a formula, Freq ~ boys, to create the one-way table.
4. Do the same steps again to create a one-way table, Saxony11, containing similar frequencies for families of size==11.

### Exercise 2.6

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

### Exercise 2.7

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.
2. Use structable() to create a nicely organized tabular display.
3. Use xtable() to create a LaTeX or HTML table.

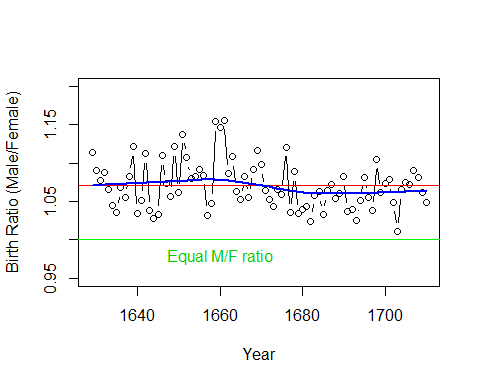
## Chapter 3

### 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  
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)  
})



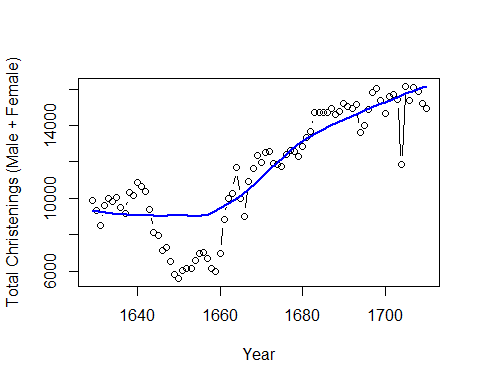
summary(Arbuthnot$Ratio)

## 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 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)  
})

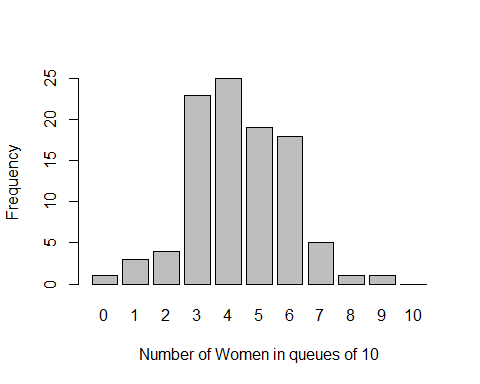


There was a large decline in births between 1640--1660, corresponding to years of plague in England.

### Exercise 3.3

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")



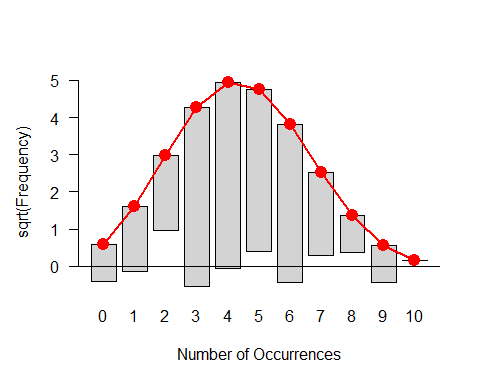
1. Check for goodness-of-fit to the binomial distribution using the goodfit() methods described in Section 3.3.2.

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.

plot(gf.women)



1. Suggest some reasons why the number of women in queues of length 10 might depart from a binomial distribution, .
2. Perhaps women (or men) are more prevalent in these queues, so .
3. People often join lines in groups, so the observations are unliekly 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.

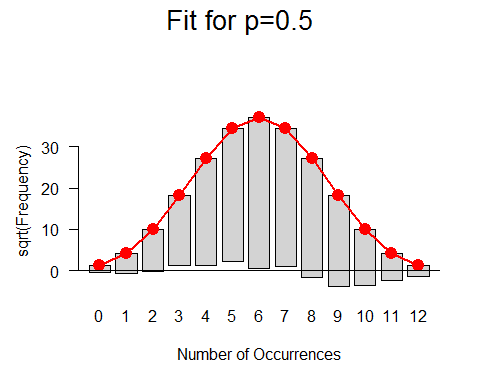
Saxony\_gf2 <- goodfit(Saxony, type = "binomial", par=list(size=12))  
summary(Saxony\_gf2)

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

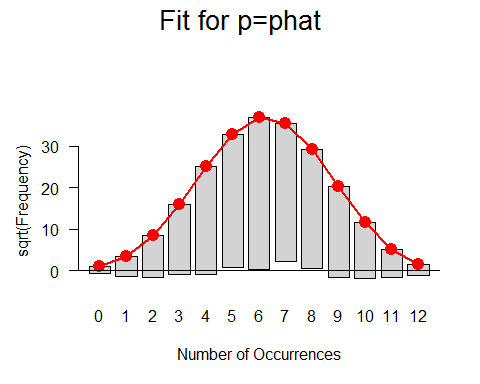
This fits much better, but still not a good fit.

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

plot(Saxony\_gf, main = "Fit for p=0.5")



plot(Saxony\_gf2, main= "Fit for p=phat")



### Exercise 3.5

### 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.

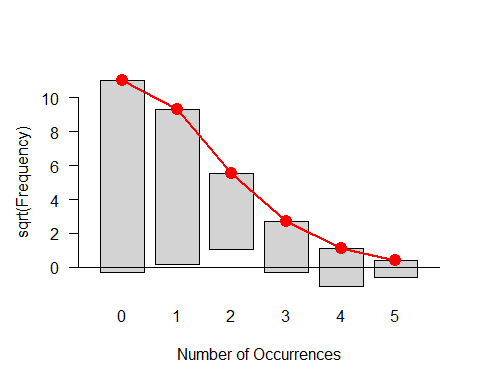
(up0 <- goodfit(Upon, type="poisson"))

##   
## 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)



1. Fit and plot the negative binomial model for these frequencies.

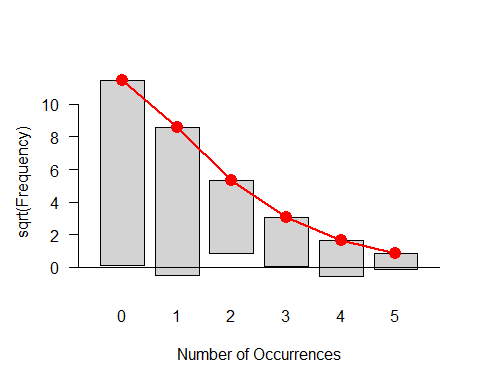
(up1 <- goodfit(Upon, type="nbinomial"))

##   
## 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)



1. What do you conclude?

### Exercise 3.7

### Exercise 3.8

### Exercise 3.9

### Exercise 3.10

### Exercise 3.11

### Exercise 3.12

## Chapter 4

### 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 ex:jobsat1 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

gave the data in tab:siskel-ebert 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”). ch04/tab/siskel-ebert

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 fig:lifeboats2 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 fig:lifeboats2, showing the changes over time in this measure.