Sample Answer Lab05: Bivariate plots and descriptive statistics

Part I: Analyzing bivariate relationships of variables on different measurement scales (2 points)

The workspace **`Part1Data.RData`** holds 3 groups of 3 associated data-frames with two variables. Import the workspace into your R session's environment with the function **`load( )`**.

For each group of data-frames you will visualize the relationship between their two variables with an appropriate statistical graphs, calculate requested information and decide whether the two variables in each data-frame are: [a] statistically independent, [b] statistically dependent or [c] whether it is difficult to decide on their dependency status because the data may be affected to some degree by sampling variations.

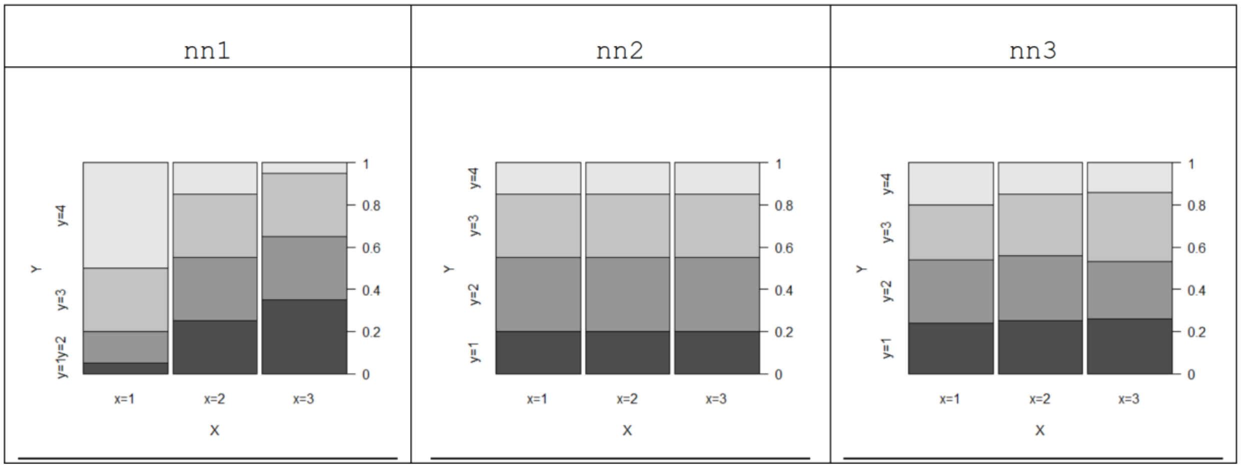
For each task show your code along with the requested output. Since you explore an undirected relationship between two variables their order does not matter.

Task 1: Show the cross-tabulated variables for the data-frames **nn1**, **nn2** and **nn3**. Pad the cross-tabulation at their edges with the row and column sums (see the sample code or Kabacoff section 7.2). (0.3 points)

```
load(file = 'Part1Data.RData')

mytable1 <- xtabs(freq~X+Y, data=rbind(nn1,nn2,nn3))

addmargins(mytable1)

     Y

X      y=1 y=2 y=3 y=4 Sum

  x=1   49  80  86  85 300

  x=2   70  96  89  45 300

  x=3   81  92  93  34 300

  Sum  200 268 268 164 900
```

Task 2: Show the spinogram for the three data-frames **nn1**, **nn2** and **nn3**. See Kabacoff p 122, sub-section 6.1.5 (0.15 points)

```
mytable1 <- xtabs(freq~X+Y, data=nn1)

vcd::spine(mytable1)
```



Task 3: Decide for the tree data-frames nn**1**, nn**2** and nn**3** whether their variables are independent, dependent or whether it is difficult to decide their dependency status. ***Justify your answer***. (0.3 points)

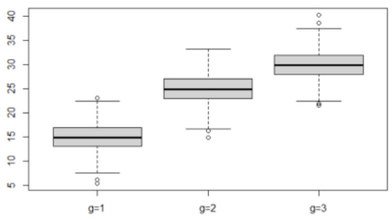
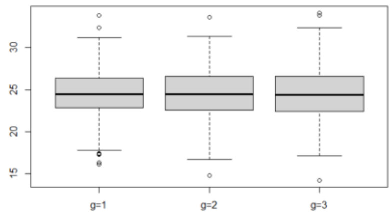
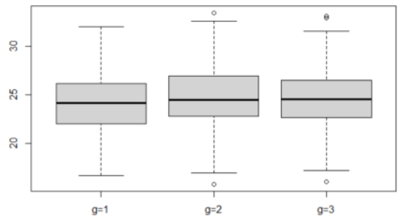
Comment: variables in **nn1** variables are statistically dependent because proportions vary substantially, based on the individual levels of either X or Y. Variables of **nn2** are statistically independent because the proportions are not affected by the individual levels of the variables. However, it is difficult to decide the dependency status of **nn3** due to the slight change of proportions, which may be due to minor random variations from the independence property

Task 4: Calculate the group means of the variable **X** in the data-frames **nc1**, **nc2** and **nc3**. Report your results in a table (see Kabacoff sub-section 7.1.3). (0.15 points)

```
tapply(nc1$X,nc1$G,mean)

g=1 g=2 g=3

 15  25  30

tapply(nc2$X,nc2$G,mean)

g=1  g=2  g=3

24.5 24.5 24.5

tapply(nc3$X,nc3$G,mean)

g=1  g=2  g=3

24.2 24.8 24.5
```

Task 5: Generate a parallel boxplot (see Kabacoff section 6.5) for each data-frame **nc1**, **nc2** and **nc3**. (0.15 points)

```
boxplot(X~G,data = nc1)
boxplot(X~G,data = nc2)

boxplot(X~G,data = nc3)
```

nc1	nc2	nc3
		

Task 6: Decide for the three data-frames **nc1**, **nc2** and **nc3** whether their variables are independent, dependent or whether it is difficult to decide their dependency status. ***Justify your answer***. (0.3 points)

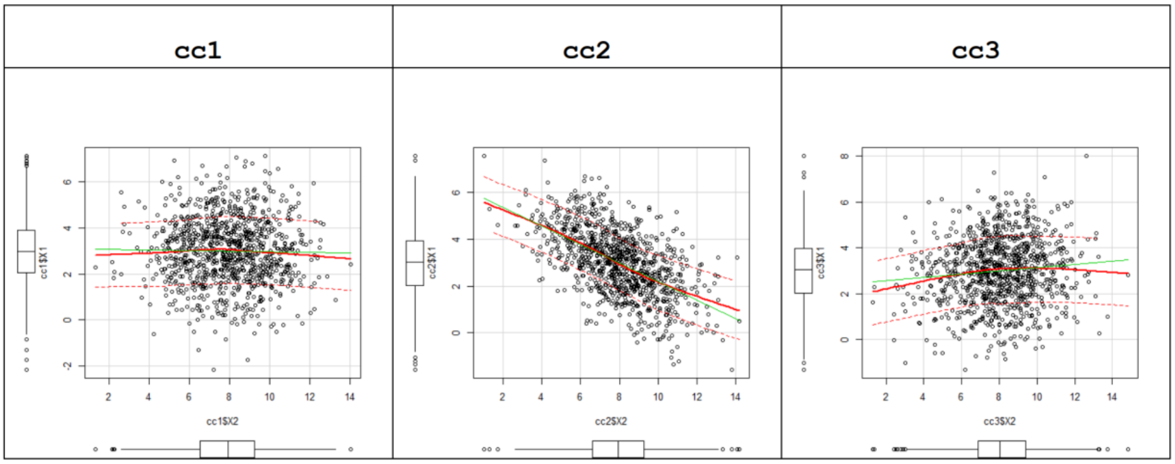
Comment: Variables of **nc1** are dependent because means differ from group by group. Variables of **nc2** are independent due to the same median/median across all three groups. It is difficult to decide the dependency status of **nc3** due to the slight drifts of means/medians which most like are the result of minor random variations from independence.

Task 7: Calculate the pairwise correlation between the variables in the data-frames **cc1**, **cc2** and **cc3**. Report your results in a table. (0.15 points)

```
cc_lst<- list(cc1,cc2,cc3)

df <- as.data.frame(lapply(1:3, function(x) cor(cc_lst[[x]]$X1,
cc_lst[[x]]$X2)))

colnames(df) <- c("CC1","CC2","CC3")

df

          CC1        CC2        CC3

1 -0.01831999 -0.5663649 0.09831319
```

Task 8: Generate a scatterplot of the variables in the three data-frames **cc1**, **cc2** and **cc3**. You may use the R function **car::scatterplot( )**. (0.15 points)

```
car::scatterplot(X1~X2,data = cc1,col =
"black",regLine=list(col="green"),smooth=list(col.smooth="red",col.var
="red"))
car::scatterplot(X1~X2,data = cc2,col =
"black",regLine=list(col="green"),smooth=list(col.smooth="red",col.var
="red"))
car::scatterplot(X1~X2,data = cc3,col =
"black",regLine=list(col="green"),smooth=list(col.smooth="red",col.var
="red"))
```



Task 9: Decide for the three data-frames **cc1**, **cc2** and **cc3** whether their variables are independent, dependent or whether it is difficult to decide their dependency status. ***Justify your answer***. (0.35 points)

Comment: variables of **cc1** are virtually statistically independent because points are randomly distributed around the horizontal regression line. Variables of **cc2** are dependent due to the strong linear correlation. Variables of **cc3** are slightly linearly dependent due to the small correlation coefficient 0.09831319; only a statistical test can tell here whether both variables are dependent of independent.

Part 2: Spurious Relationships (1 point)

Data from the famous Berkeley gender bias study are available in the table **UCBAdmissions**. To link it to your R session use the command **data(UCBAdmissions, package = "datasets")**.

Study the online description associated with this dataset; departments A and B are considered hard departments. The website http://vudlab.com/simpsons/ discusses in full length the Simpson Paradox[1] associated with this dataset. The R script to generate the tables in Tasks 10 and 11 can be found in the file **BerkleyTables.r**. You can use this script as starting point for generating your graphs.

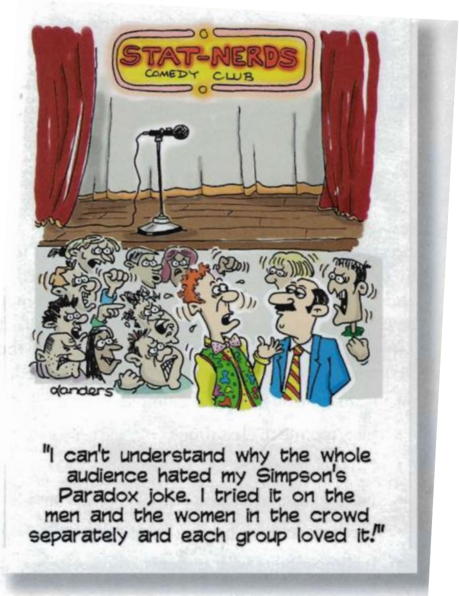
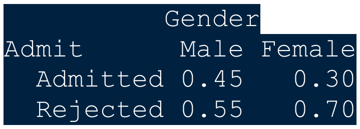

STAT-NERDS
COMEDY CLUB
olanders
"I can't understand why the whole audience hated my Simpson's Paradox joke. I tried it on the men and the women in the crowd separately and each group loved it!"

Figure 1b. A cartoon from the CAUSEweb collection. www.causeweb.org/cause/resources/fun/cartoons/Simpsons-Comic

Task 10: Generate a spinogram of the admission and rejection rates by gender. Show your code. (0.2 points)

[1] Visit also http://blog.revolutionanalytics.com/2013/07/a-great-example-of-simpsons-paradox.html, which discusses the Simpson Paradox for overall salary increases and increases by educational group.

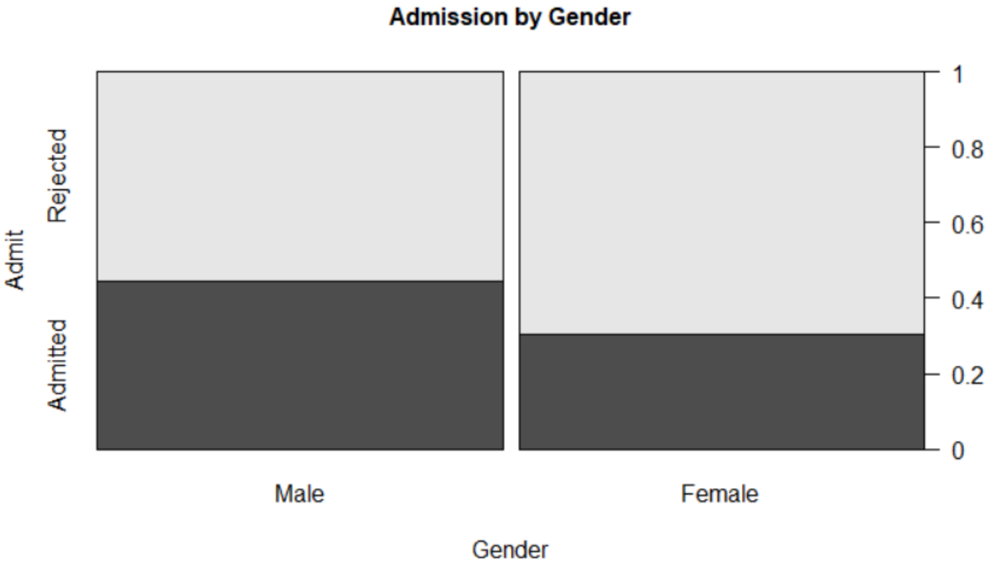
Your chart should visualize the table below. Make sure that your chart has a proper title.



```
                Gender
Admit         Male Female
   Admitted 0.45   0.30
   Rejected 0.55   0.70
```

Interpret the table. Discuss whether there is an ***apparent gender bias***?

```
source('BerkleyTables.R')

library(vcd)

spine(t(AdmissByGenPropMat), main="Admission by Gender")
```

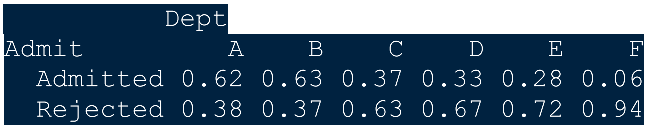


From this graph, the accepted rate of male is 50% high than female, so we could assume bias exists based on this analysis.

Task 11: Repeat the analysis separately for each gender and show both spinograms for the cross-tabulation the admission and rejection rates by department. Show your code. (0.4 points)

For comparison purposes you should place both charts side-by-side. Make sure that each chart has a proper title.

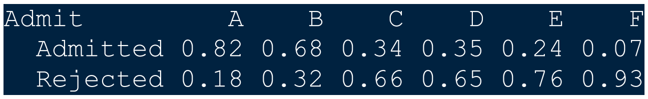
Male Applicants:



```
                Dept
Admit            A    B    C    D    E    F
   Admitted 0.62 0.63 0.37 0.33 0.28 0.06
   Rejected 0.38 0.37 0.63 0.67 0.72 0.94
```

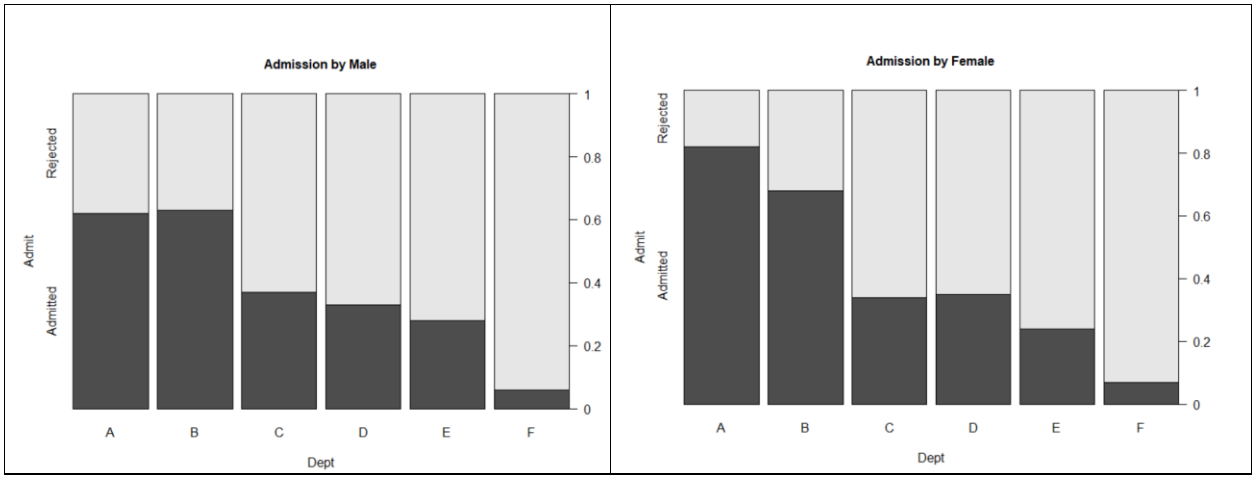
Female Applicants:

```
                Dept
```



```
Admit            A    B    C    D    E    F
   Admitted 0.82 0.68 0.34 0.35 0.24 0.07
   Rejected 0.18 0.32 0.66 0.65 0.76 0.93
```

```
AdmissByDeptMale <- round(prop.table(AdmissByDeptMale,2), 2)

spine(t(AdmissByDeptMale), main="Admission by Male")

AdmissByDeptFem <- round(prop.table(AdmissByDeptFem,2),2)

spine(t(AdmissByDeptFem), main="Admission by Female")
```



Task 12: ***Interpret both tables and their associated spinograms***. Does your conclusion about the gender admission bias change? (0.4 points)

Comment: In almost all departments the admittance rate is higher for females (exceptions are the departments C and E). So, after controlling for the individual departments, the effect of the ***gender bias seems to point slightly into the opposite direction, i.e., in favor of more female admissions***.
It is also noticeable that some departments have an overall high rejection rate. These are popular departments that can be more selective in their admission policy. These departments are also conceived by some applicant as offering "easy" programs (such as the Arts History department), because they do not have a math requirement. Unfortunately, compared to their male counterparts, less female students satisfy the math requirement, therefore, their choices are limited.
Visit the website http://vudlab.com/simpsons/ for further details.

Part 3: Association Pattern Among Car Characteristics (1 point)

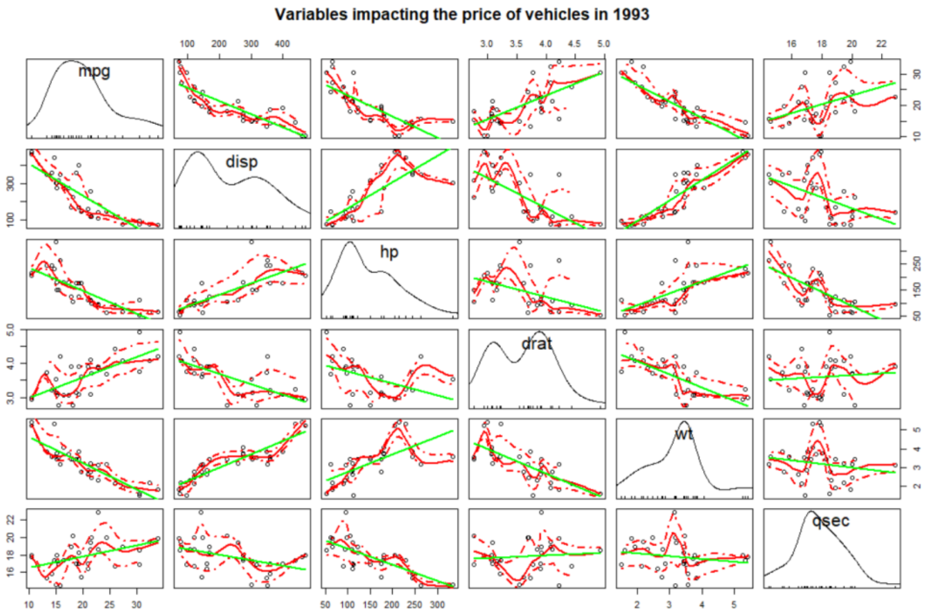
Link the **mtcars** data-frame in the package **datasets** to your R session with the command **data(mtcars, package= "datasets")** and study its description in the online help. Extract the metric variables **mpg**, **disp**, **hp**, **drat**, **wt** and **qsec** from this data-set and use these variable for your subsequent correlation analyses.

Task 13: Generate and show a Pearson ***correlation matrix*** with the **`mpg`** as first row and column variable. Round the values in the correlation matrix to 2 digits after the decimal point. Discuss and explain how the miles per gallon relate to the remaining variables (0.25 points)

```
data(mtcars, package="datasets")

car <- mtcars[ ,c("mpg","disp","hp","drat","wt","qsec")]

corMat <- round(cor(car, method= "pearson"),2)

corMat

       mpg  disp    hp  drat    wt  qsec

mpg   1.00 -0.85 -0.78  0.68 -0.87  0.42

disp -0.85  1.00  0.79 -0.71  0.89 -0.43

hp   -0.78  0.79  1.00 -0.45  0.66 -0.71

drat  0.68 -0.71 -0.45  1.00 -0.71  0.09

wt   -0.87  0.89  0.66 -0.71  1.00 -0.17

qsec  0.42 -0.43 -0.71  0.09 -0.17  1.00
```

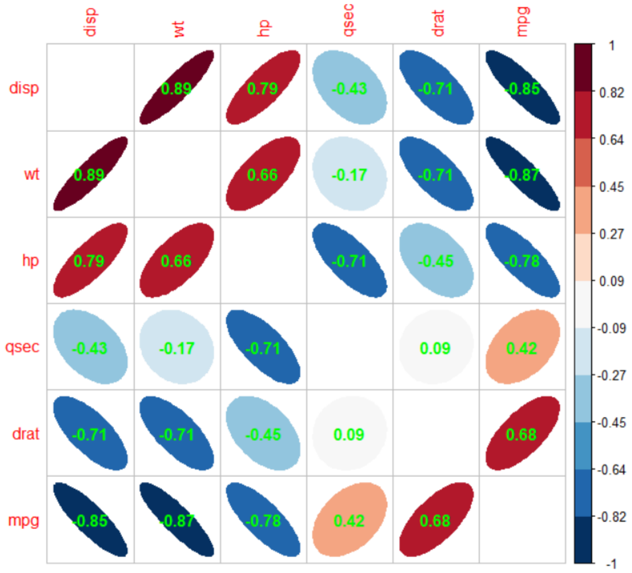
Task 14: Read the online help for the **`car::scatterplotMatrix`** function. Generate the following ***scatterplot matrix*** and show your code. (0.25 points)

```
car::scatterplotMatrix(~mpg+disp+hp+drat+wt+qsec, data=car,main="Variables
impacting the price of vehicles in 1993",col = "black", smooth=list(span =
0.35,lty.smooth=1,col.smooth="red",col.var="red"),regLine=list(col="green"))
```



Task 15: Generate the ***correlation ellipse plot*** shown below and show your code. (0.25 points)

```
colors <- RColorBrewer::brewer.pal(11,"RdBu")

colors <- rev(colors)

car1 <- mtcars[ ,c("disp","wt","hp","qsec","drat","mpg")]

corMat <- round(cor(car1, method= "pearson"),2)

corrplot::corrplot(corMat,col=colors,method = "ellipse",diag=FALSE,
addCoef.col="green")
```



Task 16: ***Discuss*** which of the two visualization methods in tasks 14 and 15 is the most informative with regards to identifying ***pairwise not necessarily linear relationships*** among the cars' characteristics. (0.25 points)

Comment: The correlation ellipsoid plot is based on the underlying matrix of correlation coefficients. The correlation coefficients can only identify linear relationships. Therefore, the correlation ellipsoid plot is a nice way to visualize the correlation matrix but it cannot capture non-linear relationships and does not show individual observations. A bi-polar color theme is use to distinguish between positive and negative correlation, as do the shapes and directions of the ellipsoids.

The scatterplot matrix shows substantially more information. A kernel density estimate for each variable on the diagonal provides information about the distribution of the variables. A linear regression line is shown by the green lines and a lowess smoother estimate, depicting the potentially non-linear relationships, are show by the red lines. The dashed bands depict the bounds within which with 95% probability the lowess line smoother will fall. Since the data are not aggregated, individual outliers and influential observations are clearly identifiable in the scatterplot matrix.