# Lab05: Bivariate plots and descriptive statistics

**Handout date:** Wednesday, October 2, 2019

**Due date:** Wednesday, October 9, 2019 as hardcopy

*This lab counts 4 % toward your total grade*

**Objectives:** In this lab   
[a] you will practice exploring bivariate relationships variables on different measurement scales and decide on the dependency status of the variables,   
[b] explore the Simpson paradox,  
[c] and evaluate the correlation among a set of metric variables.

**Format of answer:** Your answers (graphs and verbal description) should be handed in as ***hard-copy*** in ***one*** document. Add a running title into the header of the document with the following information: ***your name***, ***Lab05*** and ***page numbers***. Label each answer properly starting with its task number. Maintain the sequence of questions. Format any code and computer output properly before inserting it into the document with your answer.  code and text output need to be in a ***monospaced*** font (i.e., fixed-pitch font) such as Courier New so proper spacing and alignments are preserved. Excessive, but irrelevant, output will lead to a deduction of your accumulated points.

# 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  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)

Task 2: Show the spinogram for the three data-frames **nn1**, **nn2** and **nn3**. (0.15 points)

Task 3: Decide for the three data-frames **nn1**, **nn2** and **nn3** whether their variables are independent, dependent or whether it is difficult to decide their dependency status. ***Justify your answer***. (0.3 points)

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)

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

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)

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)

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

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)

# 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  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]](#footnote-1) associated with this dataset. The  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.

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

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***?

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

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

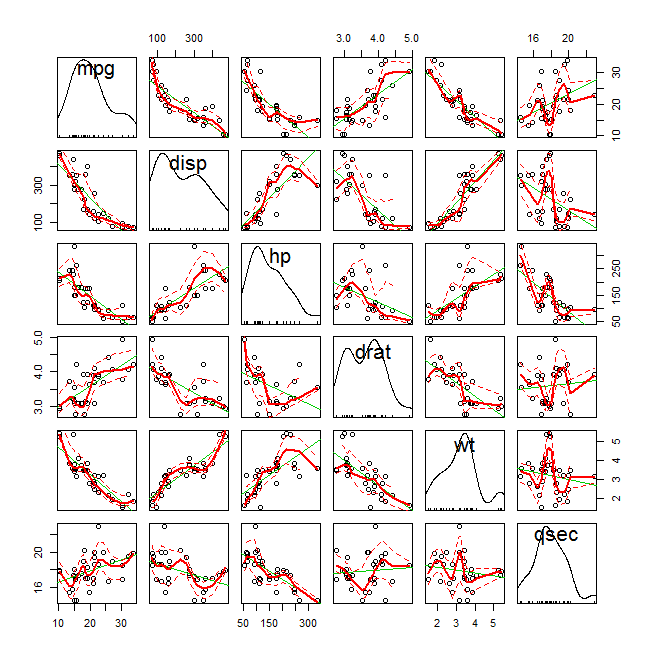
# Part 3: Association Pattern Among Car Characteristics (1 point)

Link the **mtcars** data-frame in the package **datasets** to your  session with the command **data(mtcars, package= "datasets")** and study its description in the online help.

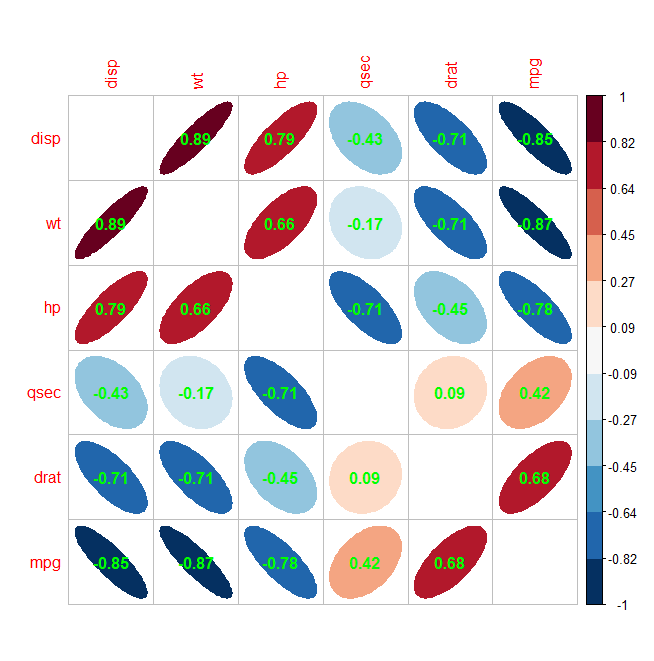
Extract the metric variables **mpg**, **disp**, **hp**, **drat**, **wt** and **qsec** from this data-set and use these variable for your subsequent correlation analyses.

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)

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



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



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)

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. [↑](#footnote-ref-1)