DAE 8 (Correlation) Assignment

Dr. Hendrik Purwins, Associate Professor
Dept. Architecture, Design and Media Technology, Aalborg University Copenhagen
A.C. Meyers Vænge 15, DK-2450 Copenhagen SV, Denmark

1. (Correlation) Load `anscombe.csv`. In the first row, this file contains the names of 8 variables: $x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4$. Read the data into Matlab so that the first column in the csv file (not including the top row with the variable names) is assigned to variable x_1, the second column is assigned to variable y_1, the third column to x_2 and so forth.

   (a) For each of the 4 (x, y) pairs (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4) plot and perform the correlation using `corrplot`. (*Hint*: consult `help` to find out how to pass the input variables to `corrplot`.) What is Pearson's r and associated p for all four variable pairs? Check for all pairs, whether their relation is linear, and whether there are outliers.

2. Is there a relationship between the amount of personal savings (in DKK) and happiness? 300 people are interviewed, and asked to provide the amount of their savings and how happy they are (on a 7 point scale). What is the level of measurement of 'happy'? Load the data from `cash_happy.csv`. Give the correlation and the p-value.

3. We are interested in the question how the average h per week of playing computer games relates to gender (male vs. female) among Medialogy students. Load the variables `ftimes` (average gaming times for females) , `mtimes` (averge gaming times for males) from `gametimes.mat`.

   (a) Perform a t-test on the null hypothesis 'Average weekly gaming hours are the same for male and female Medialogy students' vs 'Average weekly gaming hours are not the same for male and female Medialogy students' . Assume that the assumptions to apply a t-test are fulfilled. Perform the test in matlab and give the resulting p-value.

   (b) Calculate the effect size.

   (c) Calculate the *point-biserial correlation coefficient* on the data. The point-biserial correlation is employed when one of the two variables is a *discrete dichotomy*. A dichotomy is called discrete, if there is no underlying continuum between two categories, e.g. dead vs. alive. Create a 60-dimensional column vector x with all gaming hours per week for both female and male. Then create another 60-dimensional column vector y which contains a 0 if the corresponding measurement in x is derived from a female student and 1 otherwise. Then calculate the Pearson correlation coefficient for x and y. Give the resulting r and p. How do these values relate to the p-value and effect size previously calculated from the t-test?

(d) Perform Pearson correlation on x and y', where in y is derived from y' by inverting 0 and 1, i.e. y' is 60-dimensional column vector y which contains a 0 if the measurement is derived from a male students and 1 otherwise. What are the resulting r and p?