Dependency Analysis in Detail

The SPSS dataset 'employee_survey.sav' is based on an employee survey conducted in firms in Berlin 2010. Respondents were randomly selected. Based on this dataset several dependency measures should be calculated discussed. It cannot be assumed that motivation is measured based on an interval scale.

The dataset and the idea of the exercise is based on Eckstein (2015), p. 73.

1. Find the file on your hard disc. Copy the folder name in the clipboard.
2. Set the working directory to the folder. So the script can easily be used on another computer. Just by modifying this command!
3. R Studio offers a broad range of options and generates the code at the same time. In R Studio …
   a. go to 'Files \ Import Dataset \ From SPSS'
   b. Using browse select the file 'employee_survey.sav' and click open
   c. In the preview you can see the result.
   d. Do not click import. Instead copy the script commands (!) and paste them in a script file!

      ```
      library(haven)
      X005b_employee_survey <- read_sav("C:/Folder name/005b_employee_survey.sav")
      View(X005b_employee_survey)
      ```

   e. Modify the commands!
      Remove the folder name! That's because the file is in the working directory.
      Modify the variable name, e.g., to 'employee.survey'.
4. Convert 'employee.survey' to a data frame.
5. Show the first couple of rows and determine the sample size.
6. Measuring dependencies depend on the scale of measurement of the variables assessed. Therefore, determine the scale of measurement of all variables in the dataset. Add your note here …

gender	motivation	age	salary

7. Before we bin variables we want to analyse dependencies using graphs. Add the correct axes annotations!
   a. Create a graph to assess the dependency between gender and salary.
   b. Create another graph to assess age vs. gender.

8. To get a better overview of the variables we bin the age as follows
   variable 'age.category'
   N.B. The upper boundaries belong to the category!

   young up to 27
   middle between 28 - 40
   older between 41 - 65

   To bin age use the following commands. Explain the commands in detail!

```
age.thresholds<-c(0,27,40,65)
age.labels<-c('young','middle','older')

# determine categories
employee.survey$age.categories<-cut(age,breaks=age.thresholds,labels=age.labels,right=T)

# show data
View(employee.survey)

# assess result
class(age.categories)
unclass(age.categories)


# show original variables and category side by side
cbind(age,age.categories)
```

9. Now bin the variable ‘salary’ as follows. Calculate a new factor ‘salary.category’.
   The variable must be a new column in the data frame employee.survey.
   N.B. the upper boundaries belong not to the category!

low	below 1500
middle	1500 to below 2500
high	2500 to below 4000

10. The dependencies between variables should be examined. The following table provides an overview of the different measures. For more details see, e.g., IBM Knowledge Center (2017) and Michigan State University (2017)

	Nominal	**Ordinal**	**Metrical**
Nominal	Phi and Cramer's V		
Ordinal	Phi and Cramer's V Mann-Whitney test or Wilcoxon test (only two groups or "levels" of X); Kruskal Wallis test (X can have more than two groups)	Spearmen’s Rho Gamma Kendall's tau-b take ties into account Kendall's tau-c ignores ties	
Metrical	Eta T-test (only two groups or ‘variable levels’ ONEWAY-ANOVA (variable can have more than two groups)		Pearson correlation coefficient

11. Determine the dependencies between

dependency of interest	statistical measure of dependency	significance	conclusion
age.category and gender			
salary.category and gender			

12. Now we want to use the original variables instead of the binned. Found out if there is a significant difference in the means of the salary depending on gender.

dependency of interest	statistical measure of dependency	significance	conclusion
gender vs. age			
gender vs. salary			
gender vs. motivation			
Why we cannot apply the t-test for age vs. motivation?			

REMARKS: Using a t-test we have to distinguish
Independent 2-group t-test / Two sample assuming equal variances
Example: Comparing the performance of students of intake this year with the students of the previous intake. The number of students is not equal but we need to compare the means. Is there any significant difference in the means of the performance?
Command:
t.test(y~x,paired=FALSE,var.equal=TRUE) # where y is numeric and x is a binary factor
t.test(y1,y2,paired=FALSE,var.equal=TRUE) # where y1 and y2 are numeric

Independent 2-group t-test / Two sample assuming unequal variances
Example: On area B plant fertiliser is being used. Now we compare harvest on area A in comparison with harvest on area B. Same plants – same volatility but on a different level?
Command:
t.test(y~x,paired=FALSE,var.equal=FALSE) # where y is numeric and x is a binary factor
t.test(y1,y2,paired=FALSE,var.equal=FALSE) # where y1 and y2 are numeric

Paired t-test
Example: Assess the performance of students before and after power napping. It is the same group, so there are paired means of performance!
Command: t.test(y1,y2,paired=TRUE) # where y1 & y2 are numeric