

# Assignment Data Analysis

Beat Hulliger

The case is adapted from Sharpe, N.R., Ali, A., Potter, M.E. (2001): A Casebook for Business Statistics, Wiley, New York, p 13-20. The case (Sharpe-et-al-NPD-Case-A-1-3.pdf), the questionnaire of the survey (NPD-Questionnaire.pdf) and the data of the case (npd.sav) are on moodle. The data is stored as an SPSS-file. Read the data using the data import of RStudio and carry out the tasks below using R.

**Give a short comment on each result!**

You will need the following packages of R:

```
library(DescTools) # for CramerV() function
library(haven) # to read SPSS files
library(labelled) # to deal with variable and value labels
library(knitr) # to get the kable() function for nice tables
library(gplots) # to get the plotmeans() function
```

## 1 Data Preparation

### 1.1 Read the data

### 1.2 Check the the dimension and the variables

What is the dimension of the data set? What are the variable names? Are there labels?

### 1.3 Response (unit non-response)

The gross sample for the survey consisted of 592 companies. Not all of them answered the questionnaire. Check the number of observations (companies) on the data file and give the (unit) response rate.

### 1.4 Measurement levels and missingness (item non-response)

The variables of Sections C and D of the questionnaire are described in Table ???. Check the measurement level of these variables and whether they have missing values. Add variable labels where useful.

### 1.5 Missingness of D9IncYes1

Is the size of the company (D4) related to item non-response of the question on incubator participation, i.e. is the distribution of the size the same for those companies that respond D9IncYes1 as for those that do not respond D9IncYes1?

Create a contingency table between `is.na(D4)` and `is.na(D9IncYes1)` first.

Then use a grouped boxplot to compare the distributions of D4 for the responders and non-responders of D9IncYes1.

## 2 Analysis

### 2.1 Proportion of Programme participation

Analyse the participation at incubator programmes (D9IncYes1): What proportion of survey participants did participate in an incubator programme? Calculate a 95% confidence interval for the proportion of incubator participants.

### 2.2 Satisfaction with product development process

Analyse the satisfaction with the product development process (C5) using a barchart!

### 2.3 Association of incubator participation and satisfaction with the product development process

Does the satisfaction with the product development process (C5) differ between participants and non-participants in incubator programs? Create the corresponding contingency table, a mosaicplot and test the independence of satisfaction with the product development process (C5)! Also calculate the association measure Cramer's V!

### 2.4 Number of R+D personnel

Investigate the distribution of the number of R+D personnel (D5) using a histogram and a boxplot! Calculate the quartiles and Bowley-Skewness<sup>1</sup>. Repeat the graphics with the log transformed D5 (Actually use  $\log(D5 + 1)$  to avoid problems with  $D5 = 0$ ).

### 2.5 R+D personnel vs participation

Is the number of R+D personnel (D5) the same for incubator participants and non-participants? Check the distribution of D5 for participants and non-participants with boxplots and normal plots (use `aggregate()` with function `qqnorm()` to plot by group). Also compare the boxplots of log-transformed numbers to gain a better picture. Carry out a two-sample Wilcoxon-Mann-Whitney U-test to compare D5 for participants and non-participants and compare with the t-test.

---

<sup>1</sup>Bowley-Skewness is  $[(Q_3 - Q_2) - (Q_2 - Q_1)] / (Q_3 - Q_1)$