Lucas Faijdherbe(2594812) & Ruben van der Ham(2592271) – CS40

**Assignment 1**

**Exercise 1.1**

10. Sampling method: Voluntary response. This is a flawed sampling method, since respondents can decide themselves if they want to respond, which can lead to a ‘biased sample.

12. Sampling method: Randomized sample. This can be considered a sound approach, since everybody has an equal chance of getting picked. The group has a substantial size of over a 1000 respondents, which also helps in making sure there will really be different kinds of people responding.

26. Everybody can respond to this poll. This means that not only students that follow a college major can respond, but also people that already have a job, etc. Now, the number says absolutely nothing. It might be that all college major students responded that their major prepared them for their chosen careers (so, 41%), but that the other 59% of the respondents did not even follow a college major, leading to a flawed conclusion. It would be better to only poll students.

**Exercise 1.2**

22. The level of measurement of the depth is ratio. The measurements can be ordered, and also have significance; the difference in depths can be useful for earthquake research.

32. This is a nominal level of measurement. There is nothing meaningful to the numbers; they just ‘represent’ the player. Because the numbers have no meaning, when ordered the mean of these numbers doesn’t mean (pun not intended) anything.

**Exercise 1.3**

***// Explanation?***

6. The study described corresponds to an experiment. The subjects were given a treatment and therefore they are modified.

12. Type of sampling used: systematic sampling

18. Type of sampling used: cluster sampling

**Exercise 1.4**

a. The Y-axis starts at 50 instead of 0. It seems like CDU has 10, maybe even 20 times more seats than the CSU, but in reality, they only have 4x more seats. The representation exaggerates the difference between the parties.

b. Suppose that you are preparing the annual report of a big social network company. One of

your datasets contain the average numbers of daily public posts for each registered user.

Which of the following graphs would be best for describing the distribution of the average

number of posts: histogram; bar chart; Pareto chart; pie chart?

Which graph is the best to compare the means of all average numbers of posts in the

subgroups “male, single”, ”female, single”, ”male, married”, ”female, married”?

R-exercises

Hints concerning R:

• For the exercises below you can use, for instance, the R-functions hist, boxplot, mean,

median, sd, min, max, and summary. If necessary, experiment with the different options

these functions have.

• The R-function quantile(x,α) gives the α-quantile of the values in the vector x. For

example, quantile(x,0.25) gives the first quartile of x. Instead of one single value, also

a vector (α 1 , α 2 , . . . , α k ) can be inserted for the parameter α in quantile. Check which

output this function gives when the parameter α is not specified.

**Exercise 1.5**

**Exercise 1.6**

**Exercise 1.7**