

Bus 273: Statistical Analysis for Business

Fall 2014

PROBLEM SHEET # 3

Problem 1: File `bank_customers.xls` contains (simulated, but realistic) data concerning 5868 customers of a bank. Each row corresponds to one customer. The variables and their values are:

| symbol | description | values |
|-----------|--|---|
| deficit | Credit limit exceeded within last 6 months? | 0: no 1: yes |
| age | age in completed years | positive integer number |
| m.status | marital status | 0: single or not specified 1: married |
| edu | highest level of education achieved | 1: no graduation or general-education graduation 2: vocational school, apprenticeship 3: university of applied sciences, university |
| econ.act | economic activity | 0: unemployed or inactive 1: employee or public servant 2: self-employed (or family member) |
| urban | Is place of residence in urban area? | 0: no 1: yes |
| stability | Is residential area stable? (Many moves during recent period indicate “no”.) | 0: no 1: yes |
| cellphone | number of cellular phone contracts | 0: no contract 1: one contract 2: two or more contracts |

- Determine the scaling of each variable.
- Find the percentage of customers with `deficit = 1`.
- Compute the average age (here: the arithmetic mean) of customers with `deficit = 0`.
- Compute the average age (here: the arithmetic mean) of customers with `deficit = 1`.
- Compute the percentage of customers with `deficit = 1` among those with no (one, two or more) cellular phone contract(s).

Problem 2: Obtaining a sound data set is often a laborious process. File `insurance_data_w_errors.xls` contains (simulated, but realistic) data on 1000 insurants: date of birth, the date they entered the insurance contract, gender, the number of damages they incurred in 2008, and the corresponding total damage amount. — Unfortunately, some typical errors have crept into this data set: Data of nine cases need to be fixed before the data set can be analyzed.

- Find the errors and fix them in a plausible way.
- How many people with at least one damage in 2008 are there?
- Compute the average damage amount (the arithmetic mean) among those who incurred a damage.

Problem 3: The distribution of new passenger vehicles registered in Germany in August of 2009 by colour is:

| colour | number | share in % |
|--------|--------|------------|
| grey | 86973 | 31.54 |
| black | 73782 | 26.76 |
| blue | 37890 | 13.74 |
| white | 27661 | 10.03 |
| red | 27540 | 9.99 |
| yellow | 5887 | 2.14 |
| green | 5047 | 1.83 |
| others | 10934 | 3.97 |
| total | 275714 | 100.00 |

(Source: www.kba.de. This distribution is also given in file `registration_new_cars_by_colour.xls`.)

- a) Draw a pie chart of this distribution. Make sure the colours in your pie chart match the colour names.
- b) The R help text for the command `pie`, which draws a pie chart, states: “Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas.” Make a plot of the distribution which incorporates this criticism.