Humboldt-Universität zu Berlin

Ladislaus von Bortkiewicz Chair for Statistics

Data analysis I

Sigbert Klinke

May 10, 2016

- Quick-R: `http://www.statmethods.net`
- UCLA Institute for digital research and education

Choose R, SPSS or any other statistical package to fulfill the following tasks. If you do not know the software good enough then use the help of your software or search for help in the internet.

1. Read in the `BOSTONH` data set (Boston Housing data). The data contain the Housing data for 506 census tracts of Boston from the 1970 census; see Harrison, D. and Rubinfeld, D.L. (1978), *Hedonic prices and the demand for clean air*, Journal of Environmental Economics and Management, 5, 81–102.

`crim`	per capita crime rate by town
`zn`	proportion of residential land zoned for lots over 25,000 sq.ft
`indus`	proportion of non-retail business acres per town
`chas`	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
`nox`	nitric oxides concentration (parts per 10 million)
`rm`	average number of rooms per dwelling
`age`	proportion of owner-occupied units built prior to 1940
`dis`	weighted distances to five Boston employment centres
`rad`	index of accessibility to radial highways
`tax`	full-value property-tax rate per USD 10,000
`ptratio`	pupil-teacher ratio by town
`b`	$1000(B-0.63)^2$ where B is the proportion of blacks by town
`lstat`	percentage of lower status of the population
`medv`	median value of owner-occupied homes in USD 1000's

   Hint for R users: The Boston Housing data are available in R directly:

```
library("MASS")
Boston
```

2. With the Boston Housing (population!) compute varous statistics and compare the results with bootstrapped statistics.
   (a) Compute for `medv` the mean and median. What mean and median is this?
   (b) Draw a sample of size $n = 100$ with replacement and compute the same statistics as before. Explain what you expect from a theoretical point of view of the confidence interval for mean and median. The $1-\alpha\%$ confidence interval for the median is given by two observations $[x_{(l)}; x_{(u)}]$ with $l = \lfloor n/2 - z_{1-\alpha/2}\sqrt{n/4} \rfloor$ and $u = \lceil n/2 + z_{1-\alpha/2}\sqrt{n/4} \rceil$.
   (c) Draw further $B = 50, 100, 150, \ldots$ samples ($n = 100$) from the Boston Housing data and compute for each sample the mean, median. Draw a histogram of the means and medians; what kind of distribution do you see? How often does your theoretical derived confidence intervals contain the population parameter? How often it should contain it?

In 95% der Fälle sollte der Konfidenzintervall den wahren Wert enthalten

(d) Draw $B = 50, 100, 150, \ldots$ samples ($n = 100$) from the sample in exercise (b) and compute for each bootstrapped sample the mean and median. Show the distribution and derive a "bootstrapped" confidence interval. What do you observe if you compare it with your asymptotical confidence interval from exercise (b)?

3. Run `example_chisq.R` from the Seafile-server. Is it necessary to run the Monte Carlo version of χ^2-independence test?

4. Read in the `ALLBUS2014` data set (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften).

`V7`	ERHEBUNGSGEBIET <WOHNGEBIET>: WEST - OST
`V10`	WIRTSCHAFTSLAGE, BEFR. HEUTE
`V11`	WIRTSCHAFTSLAGE, BEFR. IN 1 JAHR
`V417`	BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE
`V418`	BEFR.: NETTOEINKOMMEN, LISTENABFRAGE
`V419`	BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>
`V868`	BUNDESLAND, IN DEM BEFRAGTE<R> WOHNT

5. Go to the GESIS website and read about the sampling process of the ALLBUS data. If you make a data analysis with the data set, what should you keep in mind?
Es werden Leute in die Stichprobe aufgenommen (Ostdeutsche) als in der Grundgesamtheit ???

6. For data validity: compare the values of the variable ERHEBUNGSGEBIET (`V7`) with the variable BUNDESLAND (`V868`)

7. (a) Run a χ^2 independence test between the variables BUNDESLAND and BEFRAGTE NETTOEINKOMMEN. What do you observe? Could you improve your result?

   (b) Compare the variables WIRTSCHAFTSLAGE, BEFR. HEUTE (`V10`) and WIRTSCHAFTSLAGE, BEFR . IN 1 JAHR (`V11`). Are they independent?

   (c) Redo the last exercise with a random sample of 50% and 10% of the observations. What do you notice?

   (d) For each of the last two exercises compute Cohen's w.