Humboldt-Universität zu Berlin

Ladislaus von Bortkiewicz Chair for Statistics

Data analysis I

Sigbert Klinke

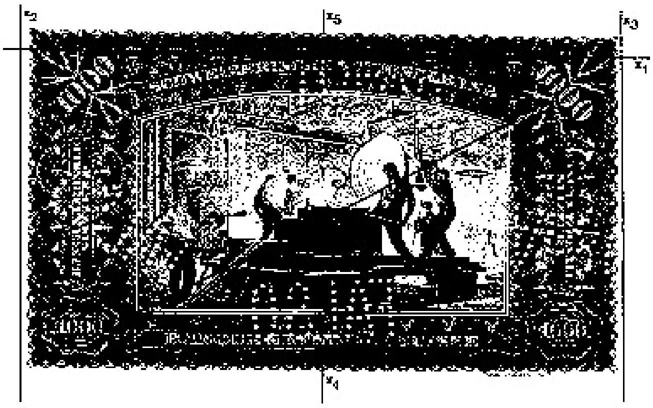
May 31, 2016

- 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 `BANK2` data set. The data are taken from the book of Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A practical approach.* London: Chapman & Hall. They are about the measurement of 200 genuine and forged old Swiss bank notes.

WIDTH (X1)	Width of the bank note
LEFT (X2)	Height of the bank note, measured on the left
RIGHT (X3)	Height of the bank note, measured on the right
LOWER (X4)	Distance of inner frame to the lower border
UPPER (X5)	Distance of inner frame to the upper border
DIAGONAL (X6)	Length of the diagonal



2. The first half of the observations consists of genuine bank notes, the second half of forged bank notes.
   (a) By numerical inspection answer: would you agree that all variables are continuous and metric? (see `unique`)
   (b) Visualize each variable. Which graphical method do you choose to show that in some variables the observation are concentrated on a few values?

3. 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

4. The first half of the observations consists of genuine bank notes, the second half of forged bank notes.

   (a) Visualize the distribution of the variable NETTOEINKOMMEN (V417). What do you observe?

   (b) Check graphically if the variable is normal distributed. If you delete outliers, do you think the data become normal distributed?

5. Read in the GSS data set (General Social Survey - US equivalent to ALLBUS).

age	Age of Respondent
sex	Respondent's Sex (1=Male, 2=Female)
educ	Highest Year of School Completed
sibs	Number of brothers and sisters
life	Is life dull (=1), routine (=2) or exciting (=3)
speduc	Highest Year of School Completed by Spouse
paeduc	Highest Year of School Completed by Father
maeduc	Highest Year of School Completed by Mother
tvhours	Hours of Television Watched
wrkstat	Labor Force Status (1=Working fulltime)
hrs1	Number of Hours Worked Last Week
rincmdol	Respondent's Income (in US$)
wifeduc	Wife: number of years of education
husbeduc	Husband: number of years of education
wifeft	Wife employed full time (0=No, 1=Yes)

6. Compute for the variable sibs and educ appropriate statistical graphics. Which location parameter describe the distribution best?

7. Consider the variable age.

   (a) Create a histogram.

   (b) Overlay the histogram with the density of an appropriate normal distribution. Does the variable look normally distributed?

   (c) Make a Q-Q-Plot.

8. Analyse the variable age with boxplots grouped after life.

   (a) Which group has the largest dispersion? Which criterion do you use for dispersion?

   (b) Is the distribution in one or more groups symmetric?

   (c) What is the median in each group?

   (d) From the graphics: Do you believe there is a relationship between age and life?