Humboldt-Universität zu Berlin

Ladislaus von Bortkiewicz Chair for Statistics

Data analysis I

Sigbert Klinke May 3, 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. Explain why you agree or disagree with each of the following statements:

   (a) It's better to include a small number of subjects in a study than a large number.

   (b) All samples from the same population give the same results.

   (c) How much the mean varies from sample to sample depends on both the size of the sample and the variability of the population.

   (d) Both variables and statistics have distributions.

   (e) A sample random variable captures all uncertainty which may occur.

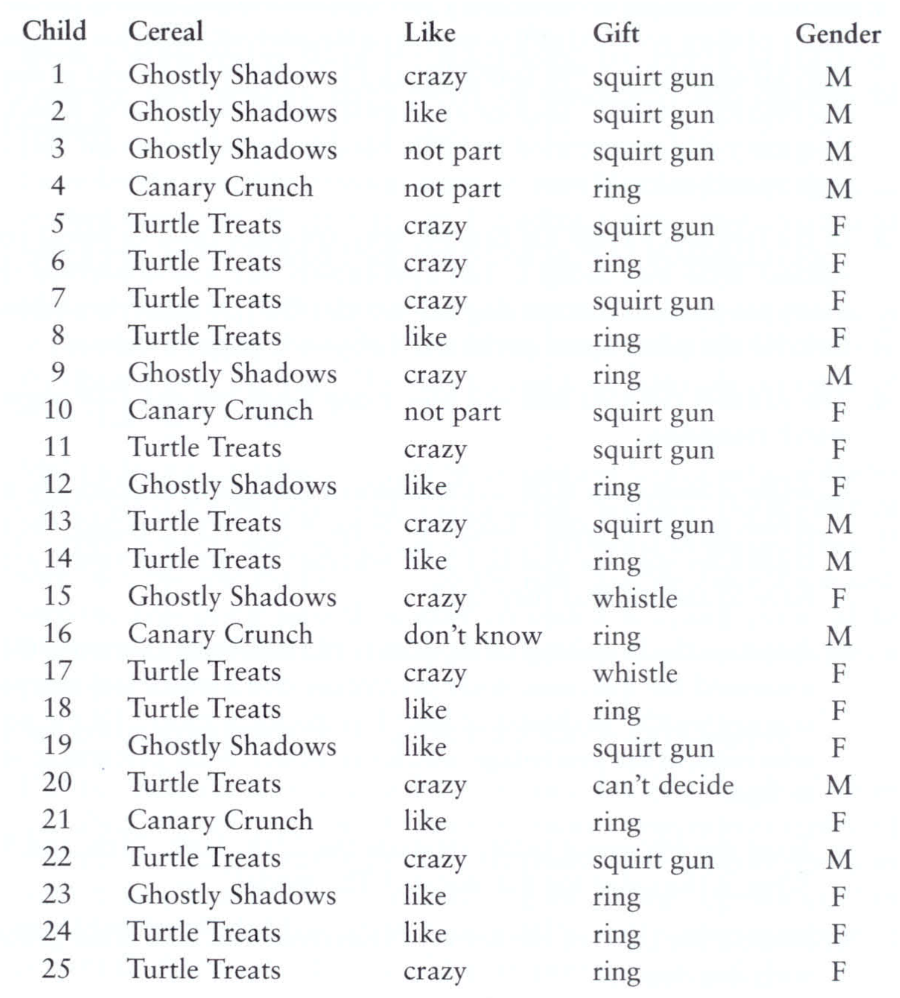
2. Twentyfive children were asked to sample each of three brands of cereals. There were asked to choose their favorite one and indicate how much they liked it. They were also asked to select which of four "gifts" they'd like to find in it: a marble, a squirt gun, a whistle, or a magic ring.

   (a) Enter the data.

   (b) What percentage of the sample are males and what percentage are females?

   (c) Which cereal was preferred by most children?

   (d) Based on the sample, which gift would you include in the cereal boxed? Explain the basis of your choice.



Child	Cereal	Like	Gift	Gender
1	Ghostly Shadows	crazy	squirt gun	M
2	Ghostly Shadows	like	squirt gun	M
3	Ghostly Shadows	not part	squirt gun	M
4	Canary Crunch	not part	ring	M
5	Turtle Treats	crazy	squirt gun	F
6	Turtle Treats	crazy	ring	F
7	Turtle Treats	crazy	squirt gun	F
8	Turtle Treats	like	ring	F
9	Ghostly Shadows	crazy	ring	M
10	Canary Crunch	not part	squirt gun	F
11	Turtle Treats	crazy	squirt gun	F
12	Ghostly Shadows	like	ring	F
13	Turtle Treats	crazy	squirt gun	M
14	Turtle Treats	like	ring	M
15	Ghostly Shadows	crazy	whistle	F
16	Canary Crunch	don't know	ring	M
17	Turtle Treats	crazy	whistle	F
18	Turtle Treats	like	ring	F
19	Ghostly Shadows	like	squirt gun	F
20	Turtle Treats	crazy	can't decide	M
21	Canary Crunch	like	ring	F
22	Turtle Treats	crazy	squirt gun	M
23	Ghostly Shadows	like	ring	F
24	Turtle Treats	like	ring	F
25	Turtle Treats	crazy	ring	F

3. 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
```

4. With the Boston Housing (population!) compute varous statistics and compare the results with bootstrapped statistics.

   (a) Compute for `lstat` and `medv` the mean, median and variance, for `chas` the percentage.

   (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 over the confidence interval for mean, median and standard deviation.

   (c) Draw further $B = 50, 100, 150, \ldots$ samples ($n = 100$) from the Boston Housing data and compute for each sample the mean, median and variance. How often does your theoretical derived confidence intervals contain the population parameter?

   (d) Derive a "bootstrapped" confidence interval and compare it with your asymptotical confidence interval from exercise b)?

   Hint: Confidence interval formulas and R codes for the median and variance you will find in the internet.