Practice 1: Computational Statistical Inference

## R software possibilities:

* Normality Test: shapiro.test(). Library lmtest in R contains most used normality tests.
* Independence Test of Durbin-Watson: dwtest(*formula*). Use acf() for a more graphic tool.
* Test for mean of a given numeric variable:
  + t.test( var, dataframe, alternative, conf.level)
* Test for variance of a given numeric variable:
  + varTest( var, dataframe, alternative, conf.level, sigma.squared) in library(EnvStats)
* Clàssic T-TEST (dicotòmic factor):
  + t.test(*formula, dataframe, var.equal=c(TRUE,FALSE),alternative*)
  + Non parametric version: *wilcox.test(formula, dataframe)*
* Parametric contrast for the equal mean hypothesis in groups defined by the level of 1 factor: ONEWAY – Analysis of Variance for 1 factor: aov(*formula, dataframe*) o oneway.test(*formula,dataframe,var.equal=c(TRUE,FALSE)). Ex: oneway.test(*Y ~ A)
* Non Parametric contrast for the equal mean hypothesis in groups defined by the level of 1 factor: ONEWAY – Analysis of Variance for 1 factor: kruskal.test(*formula,dataframe,var.equal=c(TRUE,FALSE)). Ex:* kruskal*.test(*Y ~ A)
* Correlation test for 2 numeric variables is given in R by:
  + Parametric version for normal-like variables: cor(var1, var2,method=”Pearson”) (default option in R)
  + Non-parametric version for general variables: cor(var1, var2,method=”Spearman”)
* Parametric contrasts (assuming normal distribution of Y) for equal dispersion (variance) in groups defined by levels of the studied factor (Y ~ A is the formula parameter):
  + Dichotomic Case: var.test(*formula,dataframe)*
  + Polytomic Case: bartlett.test(*formula,dataframe).*
  + Breusch Pagan Test: bptest(prestige~type) # popular in econometrics
* Non Parametric contrasts (normal distribution of Y not required) for equal dispersion (variance) in groups defined by levels of the studied factor (Y ~ A is the formula parameter):
  + fligner.test(*formula,dataframe).*
* Comparison between individual group means: Provided that F test shows a difference between groups, the question arises of wherein the difference lies.
  + Parametric version: pairwise.t.test( Y, A ) .
  + Non-Parametric version: pairwise.wilcox.test(Y, A ) .
* Feature Selection: Let Y be a response numeric variable that has to be described in terms of the rest of variables in data set, either numerical or factors. Which of the variables are associated with response Y?
* Profiling: Going a little further levels of the factors show mean group values in Y significativelly different to the gross mean? This is the descriptive analysis included in newspapers as conclusions to surveys.

We will use package FactoMineR in R. It covers Feature Selection and Profiling for target either continuous (condes()) or factors (catdes()). Warning: no missing data should be included as response variable.

R features related to Computational Statistical Inference

* Formula equation: Y ~ A+B or Y ~ A\*B, where Y is the numeric response variable and A and B are factor (qualitative variables).
* plot.factor( formula, dataframe ) and plot.design(.) are descriptive tools for graphically assessing how a numeric response variable distributes for each level of considered factors (either dichotomy or polytomic).
* Be careful with the default order of factor levels :.
  + Reorder to simplify interpretation: factor(variable, levels=c(nivell1, …, nivellsk))
  + If factor levels are not meaningful include labels for factor levels: factor(variable, levels=c(nivell1, …, nivellsk),labels=c(nom1,…,nomk)).
* Perfect collinear dummy variables appear in design matrices for general linear models and reparametrization is mandatory, being baseline for the first level the default set in R:

Base-line: options(contrasts=c("contr.treatment","contr.treatment"))

## Exercise 1: Prestige Data on prestige of professions

### Description

The Prestige data frame has 102 rows and 6 columns. The observations are occupations..

### Usage

Prestige

### Format

This data frame contains the following columns:

|  |  |
| --- | --- |
| ***education*** | Average education of occupational incumbents, years, in 1971. |
| **income** | Average income of incumbents, dollars, in 1971. |
| **women** | Percentage of incumbents who are women. |
| **prestige** | Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s. |
| **census** | Canadian Census occupational code. |
| **type** | Type of occupation. A factor with levels (note: out of order): bc, Blue Collar;prof, Professional, Managerial, and technical; wc, White Collar. |

### Source

Canada (1971) *Census of Canada*. Vol. 3, Part 6. Statistics Canada [pp. 19-1–19-21].

Personal communication from B. Blishen, W. Carroll, and C. Moore, Departments of Sociology, York University and University of Victoria.

Exercise:

Response to study: prestige of occupation.

Firstly, create a new factor, dichotomy, indicating if an occupation refers to professional and managerial duties or not. Afterwards, answer the following questions, using graphical and inferential tools.

* Does prestige seem to be normally distributed?
* Calculate a 95% CI for prestige mean.
* Calculate a 95% CI for prestige variance.
* Does prestige depend on the type of occupation? Or on having professional/managerial profile? Formulate and quantify pvalues for testing group means.
* Does the dispersion of prestige depend on the type of occupation? Or on having professional/managerial profile? Formulate and quantify pvalues for testing group variances.
* How would you test whether the prestige between blue and white collar occupations significantly differ or not?
* Use condes() in FactoMineR package to check your previous conclusions.

## Exercise 2: Swiss Labor Market Participation Data

### Description

Cross-section data originating from the health survey SOMIPOPS for Switzerland in 1981. The concern is about female labor force participation for a sample of 872 women from Switzerland. The response variable is participation which is regressed on all further variables plus age square, i.e., on income, education, age, numbers of younger and older children and the factor foreign which indicates citizenship.

### Usage

data("SwissLabor")

### Format

A data frame containing 872 observations on 7 variables.

|  |  |
| --- | --- |
| participation | Factor. Did the individual participate in the labor force? |
| income | Logarithm of nonlabor income. |
| age | Age in decades (years divided by 10). |
| education | Years of formal education. |
| youngkids | Number of young children (under 7 years of age). |
| oldkids | Number of older children (over 7 years of age). |
| foreign | Factor. Is the individual a foreigner (i.e., not Swiss)? |

### Source

Journal of Applied Econometrics Data Archive. http://qed.econ.queensu.ca/jae/1996-v11.3/gerfin/

Exercise:

Response to study: income.

* Does the nonlabor income depend on foreigner status or not?
* Does the participation in labor force depend on the family income?

Answer the following questions, using graphical and inferential tools. Follow the previous pattern.

## Exercise 3: Wage

### Description

Cross-section data originating from the May 1985 Current Population Survey by the US Census Bureau (random sample drawn for Berndt 1991).

### Usage

data("CPS1985")

### Format

A data frame containing 534 observations on 11 variables.

|  |  |
| --- | --- |
| wage | Wage (in dollars per hour). |
| education | Number of years of education. |
| experience | Number of years of potential work experience (age - education - 6). |
| age | Age in years. |
| ethnicity | Factor with levels "cauc", "hispanic", "other". |
| region | Factor. Does the individual live in the South? |
| gender | Factor indicating gender. |
| occupation | Factor with levels "worker" (tradesperson or assembly line worker), "technical" (technical or professional worker),"services" (service worker), "office" (office and clerical worker), "sales" (sales worker), "management"(management and administration). |
| sector | Factor with levels "manufacturing" (manufacturing or mining), "construction", "other". |
| union | Factor. Does the individual work on a union job? |
|  |  |

### Source

StatLib.http://lib.stat.cmu.edu/datasets/CPS\_85\_Wages

### References

Berndt, E.R. (1991). The Practice of Econometrics. New York: Addison-Wesley.

Exercise:

Response to study: wage as numeric target and sector as categorical target.

* Does wage seem to be normally distributed?
* Does wage depend on the type of occupation? Or on having professional/managerial profile? Formulate and quantify pvalues for testing group means.
* Does the dispersion of wage depend on the type of occupation? Or on having professional/managerial profile? Formulate and quantify pvalues for testing group variances.
* How would you test whether the wage between “worker” and “technical” occupations significantly differ or not?
* Use condes() in FactoMineR package to check your previous conclusions.
* Use catdes() in FactoMineR package to address feature selection and profiling for a linear model on the categorical target sector.