Statistics and learning

Multivariate statistics

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Thursday 17$^{\text{th}}$ January 2013

First: univariate statistics

- Any data set to be analysed need to be explored first !

First: univariate statistics

- Any data set to be analysed need to be explored first !
- Tools might look simplistic but robust in interpretations.

First: univariate statistics

- Any data set to be analysed need to be explored first !
- Tools might look simplistic but robust in interpretations.
- Way to get familiar with data set at hand: missing obs., erroneous/atypic points, (exp.) bias, rare modalities, variable distribution, (non-)linear relationships . . .

First: univariate statistics

- Any data set to be analysed need to be explored first !
- Tools might look simplistic but robust in interpretations.
- Way to get familiar with data set at hand: missing obs., erroneous/atypic points, (exp.) bias, rare modalities, variable distribution, (non-)linear relationships ...
- Allow analyst for pre-processing: data transformation, class recoding ...

First: univariate statistics

- Any data set to be analysed need to be explored first !
- Tools might look simplistic but robust in interpretations.
- Way to get familiar with data set at hand: missing obs., erroneous/atypic points, (exp.) bias, rare modalities, variable distribution, (non-)linear relationships . . .
- Allow analyst for pre-processing: data transformation, class recoding . . .

Quantitative variables

- From collected data to statistical table (frequency table)
- a prelude to graphical representation: 'stem-and-leaf' presentation
- Bar and cumulative diagrams; histograms & (Kernel) density est.
- Quantiles and box(-and-whisker) plot
- Numerical features (centrality, dispersion...)
- Minor differences for continuous and discrete quantitative variables

Univariate statistics (con'd)

Qualitative variable

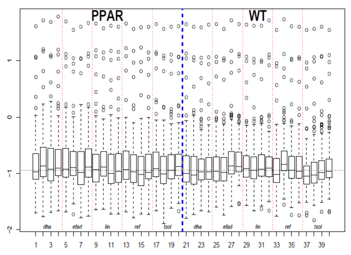
- Nominal vs. ordinal variables.
- No numerical summary from data itself $\rightarrow$ **tables** (frequency or percentages) and **graphics** (bar or pie charts).

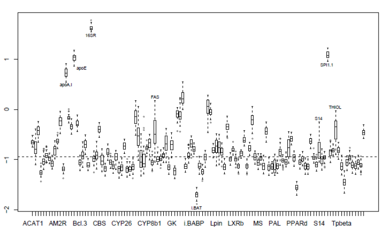
Univariate statistics (con'd)

Qualitative variable

- Nominal vs. ordinal variables.
- No numerical summary from data itself $\rightarrow$ **tables** (frequency or percentages) and **graphics** (bar or pie charts).

Genomic data





Descriptive bivariate statistics

before it's difficult to represent it

We now consider the simultaneous study of 2 variables X and Y.

The main objective is to highlight a **relationship** between these variables.
Sometimes it can be interpreted as a cause.

Descriptive bivariate statistics

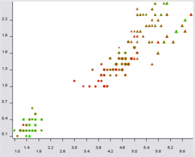
before it's difficult to represent it

We now consider the simultaneous study of 2 variables X and Y.

The main objective is to highlight a **relationship** between these variables. Sometimes it can be interpreted as a cause.

Two quantitative variables

- Scatter plot (may need to scale variables).



- Give a relationship index. *E.g.* covariance and correlation: $\operatorname{cov}(X,Y) = \sum_i (x_i - \bar{x})(y_i - \bar{y})$ and $\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$. And interpret.

Descriptive bivariate statistics (cont'd)

A quantitative variable X and a qualitative variable Y

- Parallel boxplots.
- Partial mean and sd on subpop. for all level of Y. $\rightarrow$ decomposition $\sigma_X^2 = \sigma_E^2 + \sigma_R^2$, where σ_E^2: variance explained by the partition of Y and σ_R^2: residual (between groups) variance. The ratio σ_E^2/σ_X^2 is an link index between X and Y.

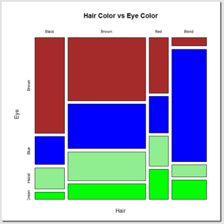
Descriptive bivariate statistics (cont'd)

A quantitative variable X and a qualitative variable Y

- Parallel boxplots.
- Partial mean and sd on subpop. for all level of Y. $\rightarrow$ decomposition $\sigma_X^2 = \sigma_E^2 + \sigma_R^2$, where σ_E^2: variance explained by the partition of Y and σ_R^2: residual (between groups) variance. The ratio σ_E^2/σ_X^2 is an link index between X and Y.

Two qualitative variables

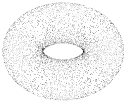
- Contingency table
- Mosaic plots with areas $\propto$ frequencies.
- Relationship index: $\chi^2 = \sum\sum \frac{(n_{kl}-s_{kl})^2}{s_{kl}}$

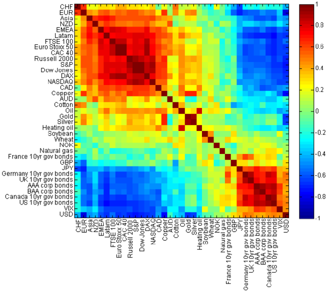


Towards multidimensional statistics

Adapting/generalising what's been seen previously:

- Matrix of correlations (symetric, positive-definite)
- Point of clouds (3D) / scatter plot matrix





Principal Component Analysis (PCA)

an introduction

- The bivariate study raised the obvious question of representing $p > 2$ variable data sets.
- Mathematically speaking, it's only a change of basis (from canonical to factor-driven). It is optimal in some sense.

Toy example

	Math.	Phys.	Engl.	Fren.
Mike	32	31	25	26
Helen	41	38	39	42
Alan	30	36	55	49
Dona	74	73	79	74
Peter	71	71	59	62
Brigit	54	51	28	35
John	26	34	70	58
William	65	62	43	47
Pam	46	48	62	61

Toy (mark) example

Toy example: data description

Elementary univariate statistics

Variable	mean	stand. dev.	min.	max
Math.	48.8	18.2	26	74
Phys.	49.3	16.1	31	73
Engl.	51.1	18.6	25	79
Fren.	50.4	14.9	26	74

Toy (mark) example

Toy example: data description

Elementary univariate statistics

Variable	mean	stand. dev.	min.	max
Math.	48.8	18.2	26	74
Phys.	49.3	16.1	31	73
Engl.	51.1	18.6	25	79
Fren.	50.4	14.9	26	74

Correlation matrix

	Math.	Phys.	Engl.	Fren.
Math.	1	0.9796	0.2316	0.4687
Phys.	0.9796	1	0.3972	0.6104
Engl.	0.2316	0.3972	1	0.9596
Fren.	0.4687	0.6104	0.9596	1

Toy (mark) example

Spectral decomposition of the covariance matrix

(Variance-)covariance matrix

	Math.	Phys.	Engl.	Fren.
Math.	330.19	286.46	78.15	126.99
Phys.	286.46	259.00	118.71	146.46
Engl.	78.15	118.71	344.86	265.69
Fren.	126.99	146.46	265.69	222.28

Toy (mark) example

Spectral decomposition of the covariance matrix

(Variance-)covariance matrix

	Math.	Phys.	Engl.	Fren.
Math.	330.19	286.46	78.15	126.99
Phys.	286.46	259.00	118.71	146.46
Engl.	78.15	118.71	344.86	265.69
Fren.	126.99	146.46	265.69	222.28

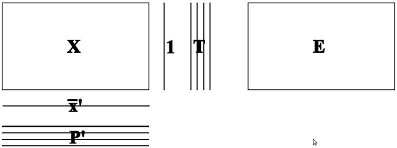
Eigen values of the covariance matrix

Factor	Eig. values	Variance percentage
F1	801.1	69.3 %
F2	351.4	30.4 %
F3	2.6	0.2 %
F4	1.2	0.1 %

PCA

- Statistical interpretation: PCA = iterative search for orthogonal linear combinations of initial variables with greatest variance.
- Geometrical interpretation: PCA = search for the best projection subspace which provides the most faithful individual/variable representation.

PCA model:
$X = {}^{\top}\bar{x} + T\,{}^{\top}P + E$

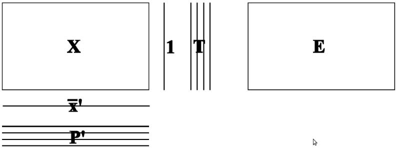


-

PCA

- Statistical interpretation: PCA = iterative search for orthogonal linear combinations of initial variables with greatest variance.
- Geometrical interpretation: PCA = search for the best projection subspace which provides the most faithful individual/variable representation.

PCA model:
$X = {}^{\top}\bar{x} + T\,{}^{\top}P + E$

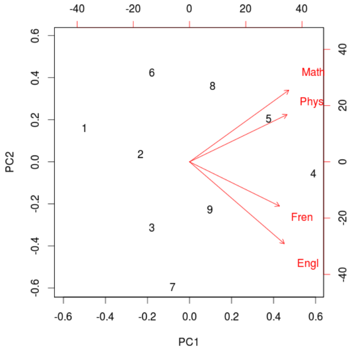


-

At the end of the day, PCA is used to (see next slide):

- Reduce the dimension of a data set
- Exhibits patterns/dependencies in high-dimensional data sets
- Represent high-dimensional data
- Bonus: detect outliers.

Studying variables and/or individuals



What's next ?

Practical session !