Quantitative Data Analysis

(version 0.7, 1/4/05)

Code: analysis-quant

Daniel K. Schneider, TECFA, University of Geneva

Menu

1. Scales and "data assumptions"

1.1 Types of quantitative measures (scales)

Types of measures	Description	Examples
nominal or category	enumeration of categories	male, female district A, district B, software widget A, widget B
ordinal	ordered scales	1st, 2nd, 3rd
interval or quantitative or "scale" (in SPSS)	measure with an interval	1, 10, 5, 6 (on a scale from 1-10) 180cm, 160cm, 170cm

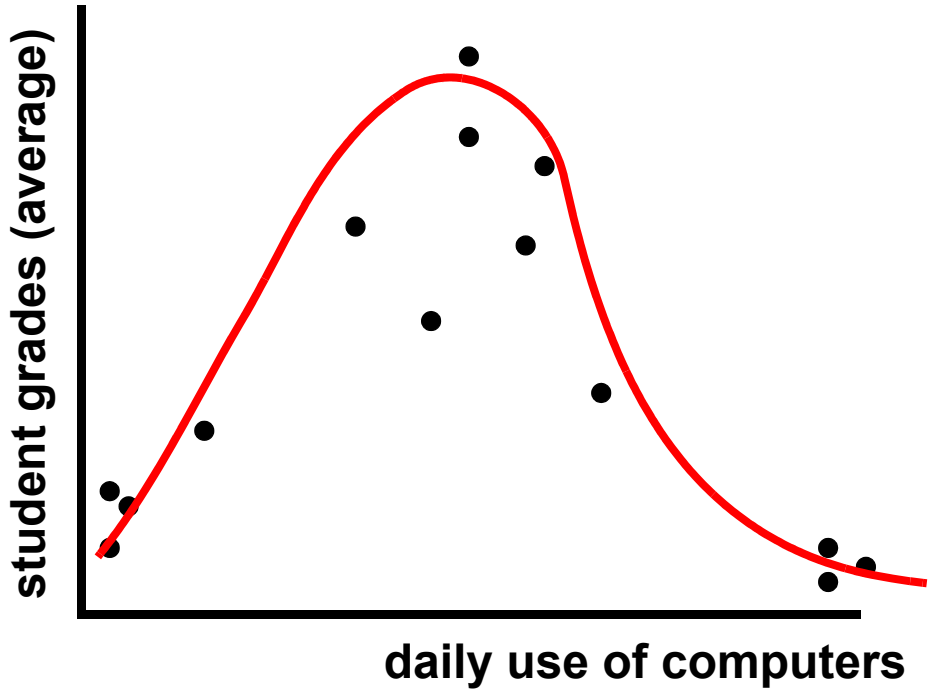
- For each type of measure or combinations of types of measure you will have to use different analysis techniques.
- For interval variables you have a bigger choice of statistical techniques.
  - Therefore scales like (1) strongly agree, (2) agree, (3) somewhat agree, etc. usually are treated as interval variables.

1.2 Data assumptions

- not only you have to adapt your analysis techniques to types of measures but they also (roughly) should respect other data assumptions.

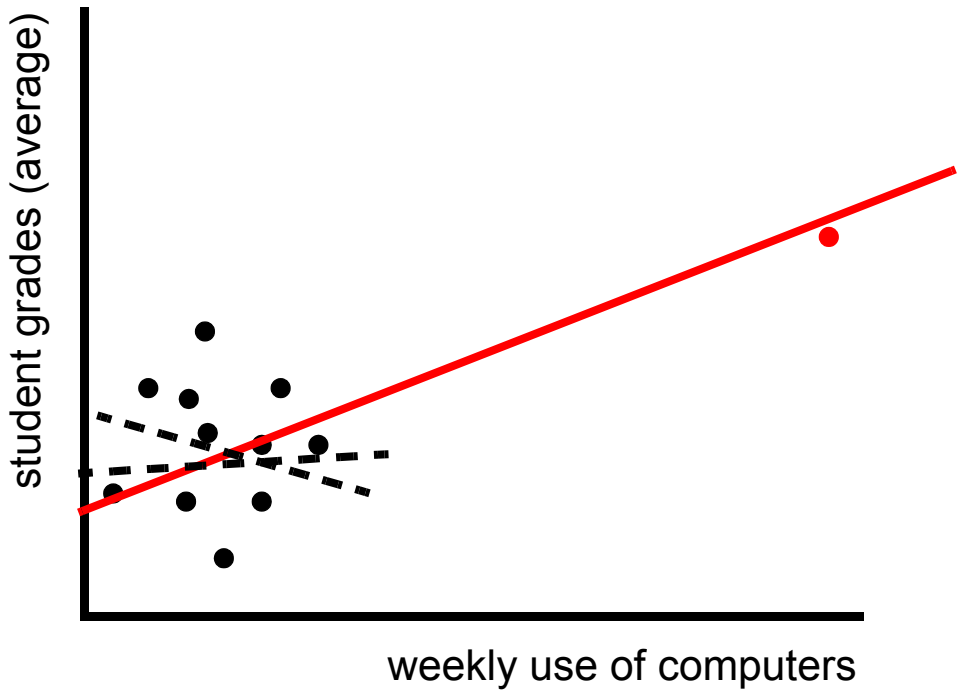
A. Linearity

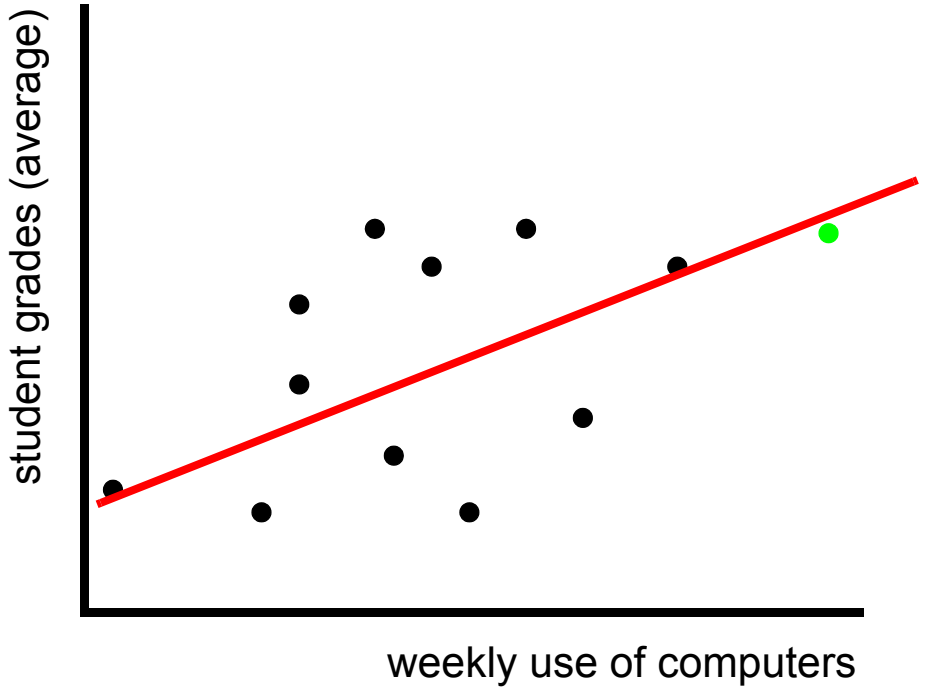
- Example: Most popular statistical methods for interval data assume ***linear relationships***:
  - In the following example the relationship is non-linear: students that show weak daily computer use have bad grades, but so do they ones that show very strong use.
  - Popular measures like the Pearson's r will "not work", i.e. you will have a very weak correlation and therefore miss this non-linear relationship



B. Normal distribution

- Most methods for interval data also require "***normal distribution***"
- If you have data with "extreme cases" and/or data that is skewed, some individuals will have much more "weight" than the others.
- Hypothetical example:
  - The "red" student who uses the computer for very long hours will determine a positive correlation and positive regression rate, whereas the "black" ones suggest an inexistent correlation. Mean use of computers does not represent "typical" usage.
  - The "green" student however, will not have a major impact on the result, since the other data are well distributed along the 2 axis. In this second case the "mean" represents a "typical" student.





2. The principle of statistical analysis

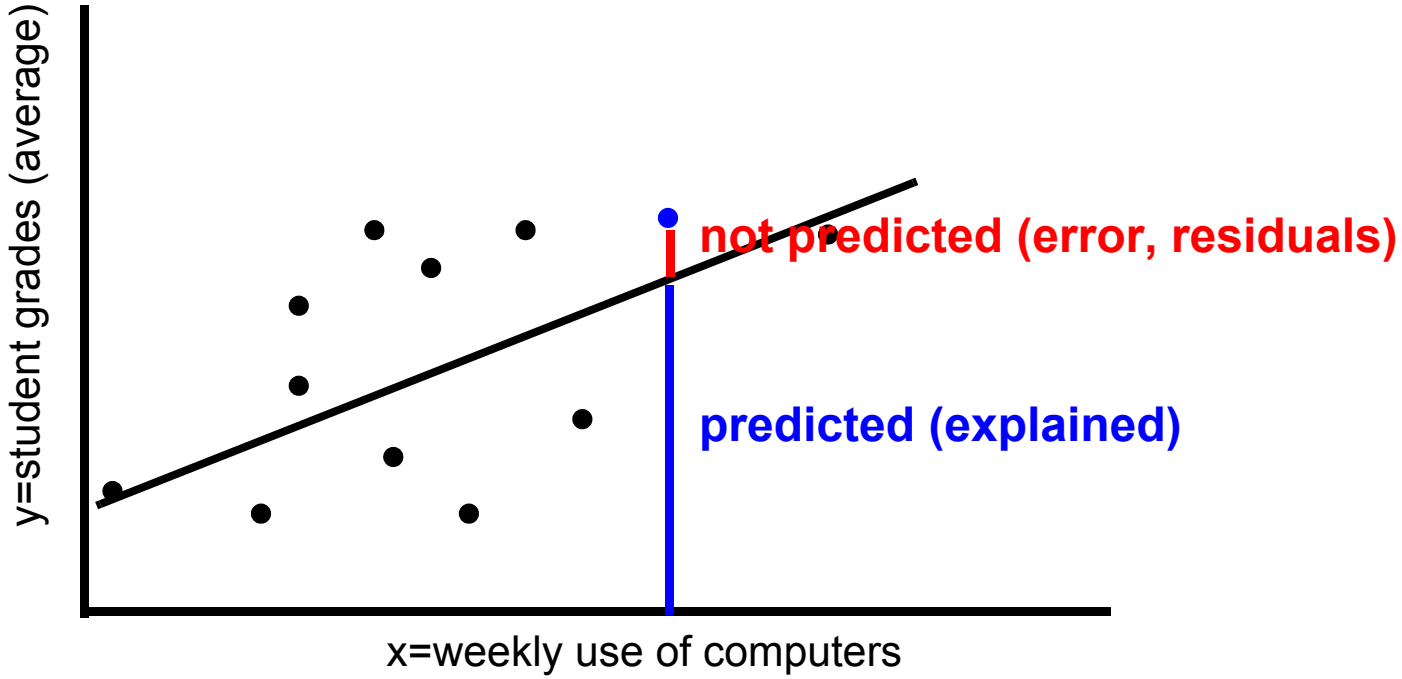
- The goal of statistical analysis is quite simple: find structure in the data

DATA = STRUCTURE + NON-STRUCTURE

DATA = EXPLAINED VARIANCE + NOT EXPLAINED VARIANCE

Example: Simple regression analysis

- DATA = **predicted** regression line + **residuals**
- in other words: regression analysis tries to find a line that will maximize prediction and minimize residuals



3. Stages of statistical analysis

Note: With statistical data analysis programs you easily can do several steps in one operation.

1. Clean your data
   - Make very sure that your data are correct (e.g. check data transcription)
   - Make very sure that missing values (e.g. not answered questions in a survey) are clearly identified as missing data
2. Gain knowledge about your data
   - Make lists of data (for small data sets only !)
   - Produce descriptive statistics, e.g. means, standard-deviations, minima, maxima for each variable
   - Produce graphics, e.g. histograms or box plot that show the distribution
3. Produce composed scales
   - E.g. create a single variable from a set of questions
4. Make graphics or tables that show relationships
   - E.g. Scatter plots for interval data (as in our previous examples) or crosstabulations
5. Calculate coefficients that measure the strength and the structure of a relation
   - Strength examples: Cramer's V for crosstabulations, or Pearson's R for interval data
   - Structure examples: regression coefficient, tables of means in analysis of variance
6. Calculate coefficients that describe the percentage of variance explained
   - E.g. R^2 in a regression analysis
7. Compute significance level, i.e. find out if you have to right to interpret the relation
   - E.g. Chi-2 for crosstabs, Fisher's F in regression analysis

4. Data preparation and composite scale making

4.1 Statistics programs and data preparation

Statistics programs

- If available, plan to use a real statistics program like SPSS or Statistica
- Good freeware: WinIDAMS (statistical analysis require the use of a command language)
  url: **http://portal.unesco.org/ci/en/ev.php-URL_ID=2070&URL_DO=DO_TOPIC&URL_SECTION=201.html**
- Freeware for advanced statistics and data visualization: R (needs good IT skills !)
  url: **http://lib.stat.cmu.edu/R/CRAN/**
- Using programs like Excel will make you loose time
  - only use such programs for simple descriptive statistics
  - ok if the main thrust of your thesis does not involve any kind of serious data analysis

Data preparation

- Enter the data
  - Assign a number to each response item (planned when you design the questionnaire)
  - Enter a clear code for missing values (no response), e.g. -1
- Make sure that your data set is complete and free of errors
  - Some simple descriptive statistics (minima, maxima, missing values, etc.) can help
- Learn how to document the data in your statistics program
  - Enter labels for variables, labels for responses items, display instructions (e.g. decimal points to show)
  - Define data-types (interval, ordinal or nominal)

4.2 Composite scales (indicators)

Basics:

- Most scales are made by simply adding the values from different items (sometimes called "Lickert scales")
- Eliminate items that have a high number of non responses
- Make sure to take into account missing values (non responses) when you add up the responses from the different items
  - A real statistics program (SPSS) does that for you
- Make sure when you create your questionnaire that all items use the same range of response item, else you will need to standardize !!

Quality of a scale:

- Again: use a published set of items to measure a variable (if available)
  - if you do, you can avoid making long justifications !
- Sensitivity: questionnaire scores discriminate
  - e.g. if exploratory research has shown higher degree of presence in one kind of learning environment than in an other one, results of presence questionnaire should demonstrate this.
- Reliability: internal consistency is high
  - Intercorrelation between items (alpha) is high
- Validity: results obtained with the questionnaire can be tied to other measures
  - e.g. were similar to results obtained by other tools (e.g. in depth interviews),
  - e.g. results are correlated with similar variables.

Exemple 4-1: The COLLES surveys

url: **http://surveylearning.moodle.com/colles/**

- The Constructivist On-Line Learning Environment Surveys include one to measure preferred (or ideal) experience in a teaching unit. It includes 24 statements measuring 6 dimensions.
- We only show the first two (4 questions concerning relevance and 4 questions concerning reflection).
- Note that in the real questionnaire you do not show labels like "Items concerning relevance" or "response codes".

Statements	Almost Never	Seldom	Some-times	Often	Almost Always
response codes	**1**	**2**	**3**	**4**	**5**
Items concerning relevance					
a. my learning focuses on issues that interest me.	O	O	O	O	O
b. what I learn is important for my prof. practice as a trainer.	O	O	O	O	O
c. I learn how to improve my professional practice as a trainer.	O	O	O	O	O
d. what I learn connects well with my prof. practice as a trainer.	O	O	O	O	O
Items concerning Reflection					
… I think critically about how I learn.	O	O	O	O	O
… I think critically about my own ideas.	O	O	O	O	O
… I think critically about other students' ideas.	O	O	O	O	O
… I think critically about ideas in the readings.	O	O	O	O	O

Algorithm to compute each scale:

for each individual add response codes and divide by number of items

or use a "means" function in your software package:

relevance = mean (a, b, c, d)

Examples:

Individual A

```
who answered a=sometimes, b=often, c=almost always, d= often gives:
(3 + 4 + 5 + 4 ) / 4 = 4
```

Missing values (again)

- Make sure that you do not add "missing values"

  Individual B

  ```
  who answered a=sometimes, b=often, c=almost always, d=missing gives:
  (3 + 4 + 5) / 3 = 4
  ```

  and certainly NOT:

  ```
  (3 + 4 + 5 + 0) / 4 or (3 + 4 + 5 -1) / 4 !!
  ```

5. Overview on statistical methods and coefficients

5.1 Descriptive statistics

- Descriptive statistics are not very interesting in most cases
  (unless they are used to compare different cases in comparative systems designs)
- Therefore, do not fill up pages of your thesis with tons of Excel diagrams !!

Some popular summary statistics for interval variables

- Mean
- Median: the data point that is in the middle of "low" and "high" values
- Standard deviation: the mean deviation from the mean, i.e. how far a typical data point is away from the mean.
- High and Low value: extremes a both end
- Quartiles: same thing as median for 1/4 intervals

5.2 Which data analysis for which data types?

Popular bi-variate analysis

		Dependant variable Y: Quantitative (interval)	Dependant variable Y: Qualitative (nominal or ordinal)
Independent (explaining) variable X	Quantitative	Correlation and Regression	Transform X into a qualitative variable and see below
	Qualitative	Analysis of variance	Crosstabulations

Popular multi-variate anaylsis

		Dependant variable Y: Quantitative (interval)	Dependant variable Y: Qualitative (nominal or ordinal)
Independent (explaining) variable X	Quantitative	Factor Analysis, multiple regresstion, SEM, Cluster Analysis,	Transform X into a qualitative variable and see below
	Qualitative	Anova	Multidimensional scaling etc.

5.3 Types of statistical coefficients:

- First of all make sure that the coefficient you use is more or less appropriate for you data

The big four:

1. Strength of a relation
   - Coefficients usually range from **-1** (total negative relationship) to **+1** (total positive relationship). **0** means no relationship.
2. Structure (tendency) of a relation
3. Percentage of variance explained
4. Signification level of your model
   - Gives that chance that you are in fact gambling
   - Typically in the social sciences a sig. level lower than 5% (0.05) is acceptable
   - Do not interpret data that is above !

These four are mathematically connected:

E.g. Signification is not just dependent on the size of your sample, but also on the strength of a relation.

6. Crosstabulation

- Crosstabulation is a popular technique to study relationships between normal (categorical) or ordinal variables

Computing the percentages (probabilities)

- See the example on the next slides

1. For each value of the explaining (independent) variable compute de percentages
   - Usually the X variable is put on top (i.e. its values show in columns). If you don't you have to compute percentages across lines !
   - Remember this: you want to know the probability (percentage) that a value of X leads to a value of Y
2. Compare (interpret) percentages across the dependant (to be explained) variable

Statistical association coefficients (there are many!)

- Phi is a chi-square based measure of association and is usually used for 2x2 tables
- The Contingency Coefficient (Pearson's C). The contingency coefficient is an adjustment to phi, intended to adapt it to tables larger than 2-by-2.
- Somers' d is a popular coefficient for ordinal measures (both X and Y). Two variants: symmetric and Y dependant on X (but less the other way round).

Statistical significance tests

- Pearson's chi-square is by far the most common. If simply "chi-square" is mentioned, it is probably Pearson's chi-square. This statistic is used to text the hypothesis of no association of columns and rows in tabular data. It can be used with nominal data.

Exemple 6-1: Crosstabulation Avez-vous reçu une formation à l'informatique ?* Créer des documents pour afficher en classe

			X= Avez-vous reçu une formation à l'informatique ?		Total
			Non	Oui	
Y= Utilisez-vous l'ordinateur pour créer des documents pour afficher en classe ?	Régulièrement	Effectif	**4**	**45**	49
		% dans X	**44.4%**	**58.4%**	57.0%
	Occasionnellement	Effectif	**4**	**21**	25
		% dans X	**44.4%**	**27.3%**	29.1%
	2 Jamais	Effectif	**1**	**11**	12
		% dans X	**11.1%**	**14.3%**	14.0%
Total		Effectif	**9**	**77**	86
		% dans X	**100.0%**	**100.0%**	100.0%

- The probability that computer training ("oui") leads to superior usage of the computer to prepare documents is very weak (you can see this by comparing the % line by line.

Statistics:

- Pearson Chi-Square = 1.15 with a signification= .562
  - This means that the likelihood of results being random is > 50% and you have to reject relationship
- Contingency coefficient = 0.115, significance = .562
  - Not only is the relationship very weak (but it can't be interpreted)

Exemple 6-2: Crosstabulation: Pour l'élève, le recours aux ressources de réseau favorise l'autonomie dans l'apprentissage * Rechercher des informations sur Internet

			X= Pour l'élève, le recours aux ressources de réseau favorise l'autonomie dans l'apprentissage				
			0 Tout à fait en désaccord	1 Plutôt en désaccord	2 Plutôt en accord	3 Tout à fait en accord	Total
Y= Rechercher des informations sur Internet	0 Régulièrement	Count	0	2	9	11	22
		% within X	.0%	18.2%	19.6%	42.3%	25.6%
	1 Occasionnellement	Count	1	7	23	11	42
		% within X	33.3%	63.6%	50.0%	42.3%	48.8%
	2 Jamais	Count	2	2	14	4	22
		% within X	66.7%	18.2%	30.4%	15.4%	25.6%
	Total	Count	3	11	46	26	86
		% within X	100.0%	100.0%	100.0%	100.0%	100.0%

- We have a weak significant relationship: the more teachers agree that students will increase learning autonomy from using Internet resources, the more they will let students do so.

Statistics: Directional Ordinal by Ordinal Measures with Somer's D

Values	Somer's D	Significance
Symmetric	-.210	.025
Y = Rechercher des informations sur Internet Dependent	-.215	.025

7. Simple analysis of variance

- Analysis of variance (and it's multi-variate variant Anova) are the favorite tools of the experimentalists.
- X is an experimental condition (therefore a nominal variable) and Y usually is an interval variable.
  - E.g. Does presence or absence of ICT usage influence grades ?
- You can show that X has an influence on Y if means achieved by different groups (e.g. ICT vs. non-ICT users) are significantly different.
- Significance improves when:
  - means of the X groups are different (the further apart the better)
  - variance inside X groups is low (certainly lower than the overall variance)

Exemple 7-1: Differences between teachers and teacher students

Population		COP1 Fréquence de différentes manières de travailler des élèves	COP2 Fréquence des activités d'exploration à l'extérieur de la classe	COP3 Fréquence des travaux individuels des élèves
1 Etudiant(e) LME	Mean	1.528	1.042	.885
	N	48	48	48
	Std. Deviation	.6258	.6260	.5765
2 Enseignant(e) du primaire	Mean	1.816	1.224	1.224
	N	38	38	38
	Std. Deviation	.3440	.4302	.5893
Total	Mean	1.655	1.122	1.035
	N	86	86	86
	Std. Deviation	.5374	.5527	.6029

- COP1, COP2, COP3 sont des indicateurs composé allant de 0 (peu) et 2 (beaucoup)
- The difference for COP2 is not significant (see next slide)
- Standard deviations within groups are rather high (in particular for students), which is a bad thing: it means that among students they are highly different.

Anova Table and measures of associations

		Sum of Squares	df	Mean Square	F	Sig.
Var_COP1 Fréquence de différentes manières de travailler des élèves * Population_bis Population	Between Groups	1.759	1	1.759	6.486	.013
	Within Groups	22.785	84	.271		
	Total	24.544	85			
Var_COP2 Fréquence des activités d'exploration à l'extérieur de la classe * Population_bis Population	Between Groups	.703	1	.703	2.336	.130
	Within Groups	25.265	84	.301		
	Total	25.968	85			
Var_COP3 Fréquence des travaux individuels des élèves * Population_bis Population	Between Groups	2.427	1	2.427	7.161	.009
	Within Groups	28.468	84	339		
	Total	30.895	85			

Measures of Association

	Eta	Eta Squared
Var_COP1 Fréquence de différentes manières de travailler des élèves * Population	.268	.072
Var_COP2 Fréquence des activités d'exploration à l'extérieur de la classe * Population	.164	.027
Var_COP3 Fréquence des travaux individuels des élèves * Population	.280	.079

- associations are week and explained variance very weak

8. Regression Analysis and Pearson Correlations

Exemple 8-1: Does teacher age explain exploratory activities outside the classroom ?

- Independant variable: AGE
- Dependent variable: Fréquence des activités d'exploration à l'extérieur de la classe

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate	Pearson Correlation	Sig. (1-tailed)	N
.316	.100	.075	.4138	.316	.027	38

Model Coefficients

	Coefficients		Stand. coeff.	t	Sig.	Correlations
	B	Std. Error	Beta			Zero-order
(Constant)	.706	.268		2.639	.012	
AGE Age	.013	.006	.316	1.999	.053	.316

Dependent Variable: Var_COP2 Fréquence des activités d'exploration à l'extérieur de la classe

All this means:

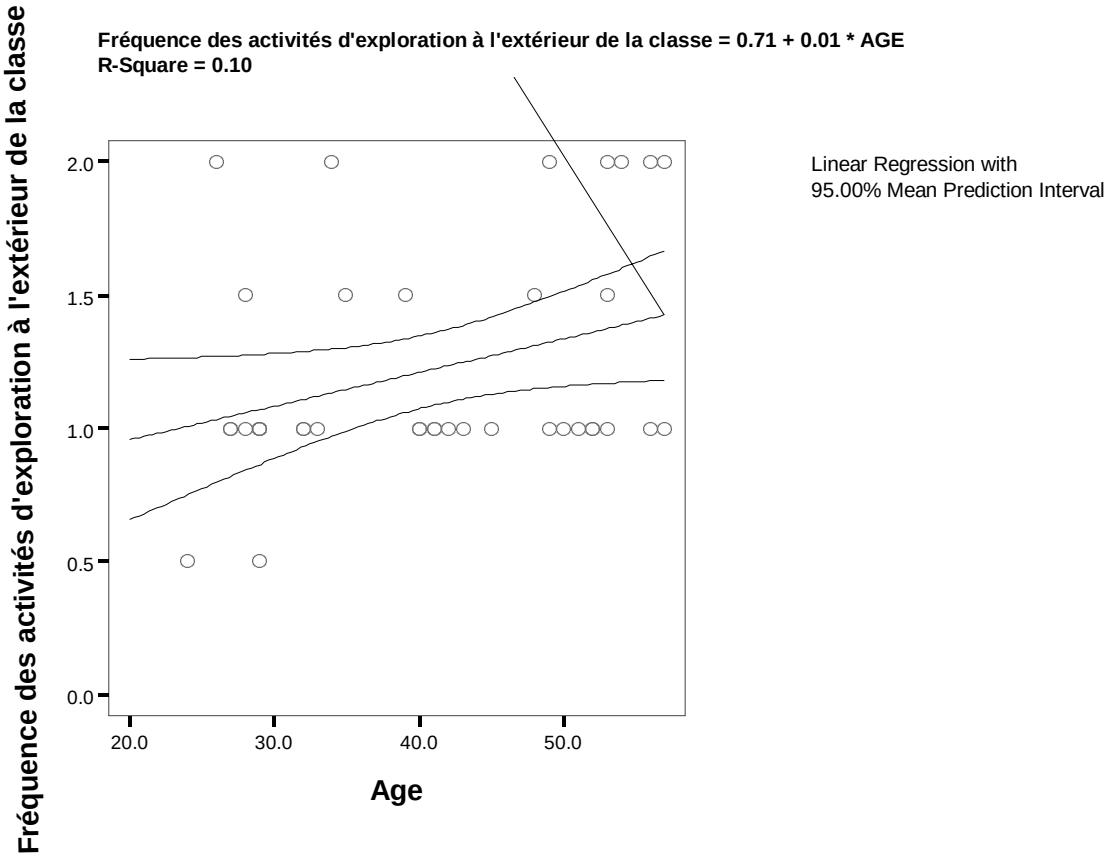
- We have a week relation (.316) between age and exploratory activities. It is significant (.027)
- Formally the relation is:

```
exploration scale = .705 + 0.013 * AGE
```

(roughly: only people over 99 are predicted a top score of 2)

Here is a scatter plot of this relation

- No need for statistical coefficients to see that the relation is rather week and why the prediction states that it takes a 100 years ... :)



9. Exploratory Multi-variate Analysis

There many techniques, here we just introduce cluster analysis, e.g. Factor Analysis (principal components) or Discriminant analysis are missing here

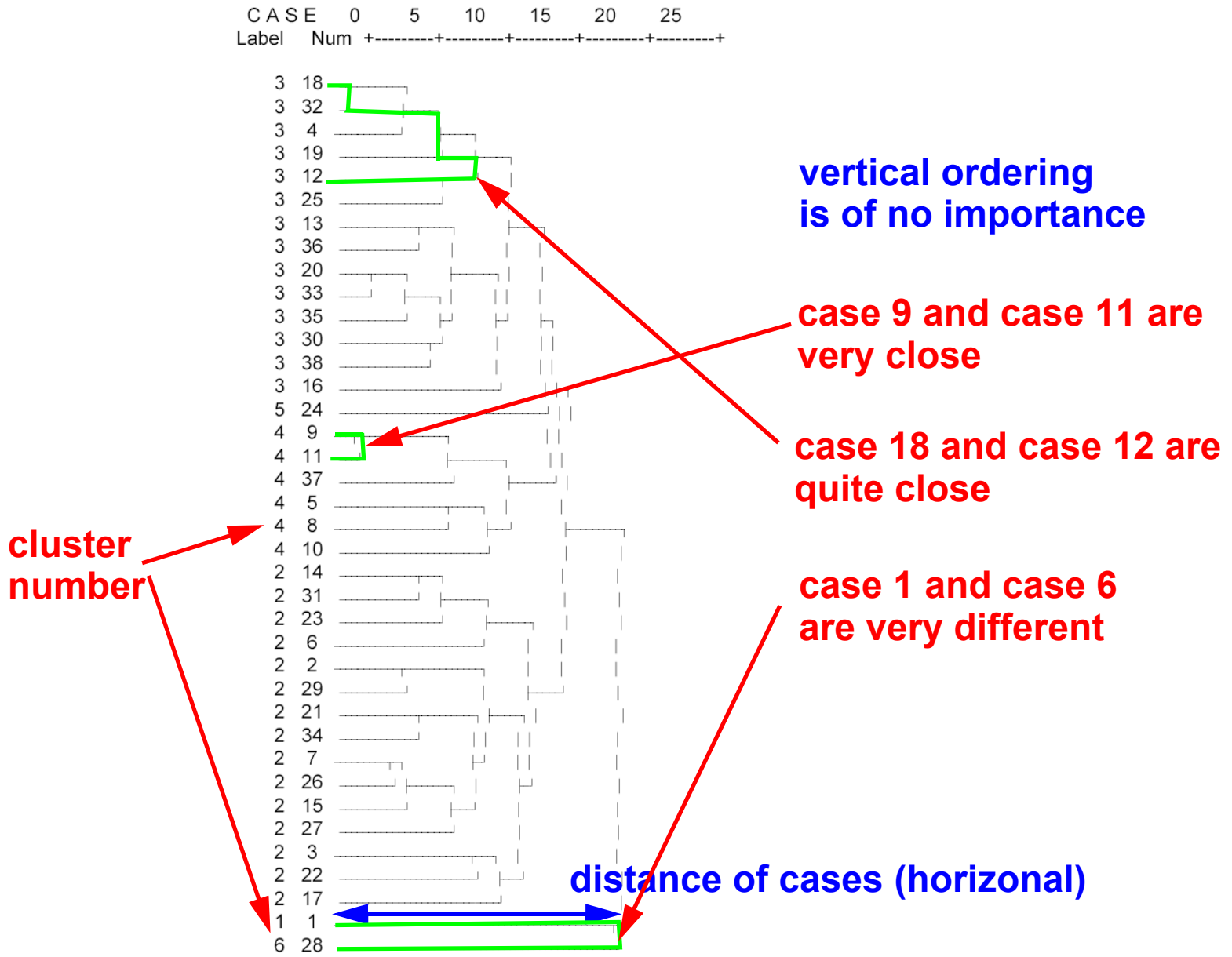
9.1 Cluster Analysis

- Cluster analysis or classification refers to a set of multivariate methods for grouping elements (subjects or variables) from some finite set into clusters of similar elements (subjects or variables).
- There 2 different kinds: hierarchical cluster analysis and K-means cluster.
- Typical examples: Classify teachers into 4 to 6 different groups regarding ICT usage

Exemple 9-1: Gonzalez classification of teachers

- A hierarchical analysis allow to identify 6 major types of teachers
- Type 1 : l'enseignant convaincu
- Type 2 : les enseignants actifs
- Type 3 : les enseignants motivés ne disposant pas d'un environnement favorable
- Type 4 : les enseignants volontaires, mais faibles dans le domaine des technologies
- Type 5 : l'enseignant techniquement fort mais peu actif en TIC
- Type 6 : l'enseignant à l'aise malgré un niveau moyen de maîtrise

Dendogram (tree diagram of the population)



Statistics of a subset of the 36 variables used for analysis:

	Types [nb d'enseignants]					
	1 [1]	2 [15]	3 [14]	4 [6]	5 [1]	6 [1]
	Moyenne	Moyenne	Moyenne	Moyenne	Moyenne	Moyenne
Degré d'importance des outils d'entraide et de collaboration pour les élèves	4.7	2.1	1.5	2.9	.0	5.0
Degré d'importance des outils de communication entre élèves	4.0	2.4	1.7	2.7	1.0	4.3
Accord sur ce qui favorise les apprentissages de type constructiviste	3.0	1.7	1.5	1.9	1.0	2.7
Accord par rapport à l'influence du milieu familial sur les apprentissages	3.0	2.7	2.6	2.1	3.5	3.5
Accord sur le sentiment de sûreté face aux élèves et de contrôle sur leurs apprentissages	3.0	3.0	2.5	2.7	2.0	2.0
Accord par rapport à la maîtrise des relations avec les élèves	4.0	3.3	2.8	2.8	5.0	3.0
Effets de l'utilisation de l'ordinateur sur la préparation et la gestion de l'enseignement	3.0	2.9	2.2	2.8	2.3	2.3
Préoccupations liées au projet et aux ressources disponibles	1.3	1.0	.5	.4	.0	2.8
Préoccupations liées aux relations avec les élèves, parents et collègues et leur statut	2.0	.8	1.0	.5	.0	1.8
Accord sur l'importance d'utiliser l'informatique dans la classe	.0	2.7	1.9	2.3	1.0	3.0
Matériel informatique "avancé" que les enseignants possèdent chez eux	.5	.8	.4	.3	1.0	.0
Niveau de maîtrise de l'utilisation d'outils TIC de communication et de documentation	2.3	2.6	2.3	1.7	3.0	1.8

- Final note: confirmatory multivariate analysis (e.g. structural equation modelling) is not even mentionnend in this document