INQUÉRITO ALIMENTAR NACIONAL
E DE ATIVIDADE FÍSICA

WEIGHTING COMPLEX SAMPLES

IAN-AF Databases

Tutorial using software SPSS e R

Table of Contents

References

[1] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R- roject.org/.

[2] T. Lumley (2017) "survey: analysis of complex survey samples". R package version 3.32.

[3] T. Lumley (2004) Analysis of complex survey samples. Journal of Statistical Software. 9(1): 1-19

Introductory Notes

In the National Food and Physical Activity Survey, IAN-AF 2015-2016, participants were randomly selected from the National Register of Users of the National Health Service, based on a two-stage complex sampling process. The sampling process proceeded as follows:

i. Primary Health Units were randomly selected in each Territorial Unit for Statistical Purposes (NUTS II). In each region, the sampling was weighted taking into account the number of individuals. The number of Primary Health Units selected was 21 in the North region, Centre and Metropolitan Area of Lisbon, 12 in the Algarve and Alentejo regions and six in the Autonomous Regions of Madeira and the Azores.
ii. Individuals registered in each Primary Health Units were randomly selected, with a fixed number of elements by sex and age group.

To estimate the results according to the IAN-AF 2015-2016 complex sample design, at national and regional level (NUTS II), the results are weighted according to a created variable. The sample weights represent how many individuals of the Portuguese population (in number) each individual of the sample represents. The calculation of sample weights included the following criteria:

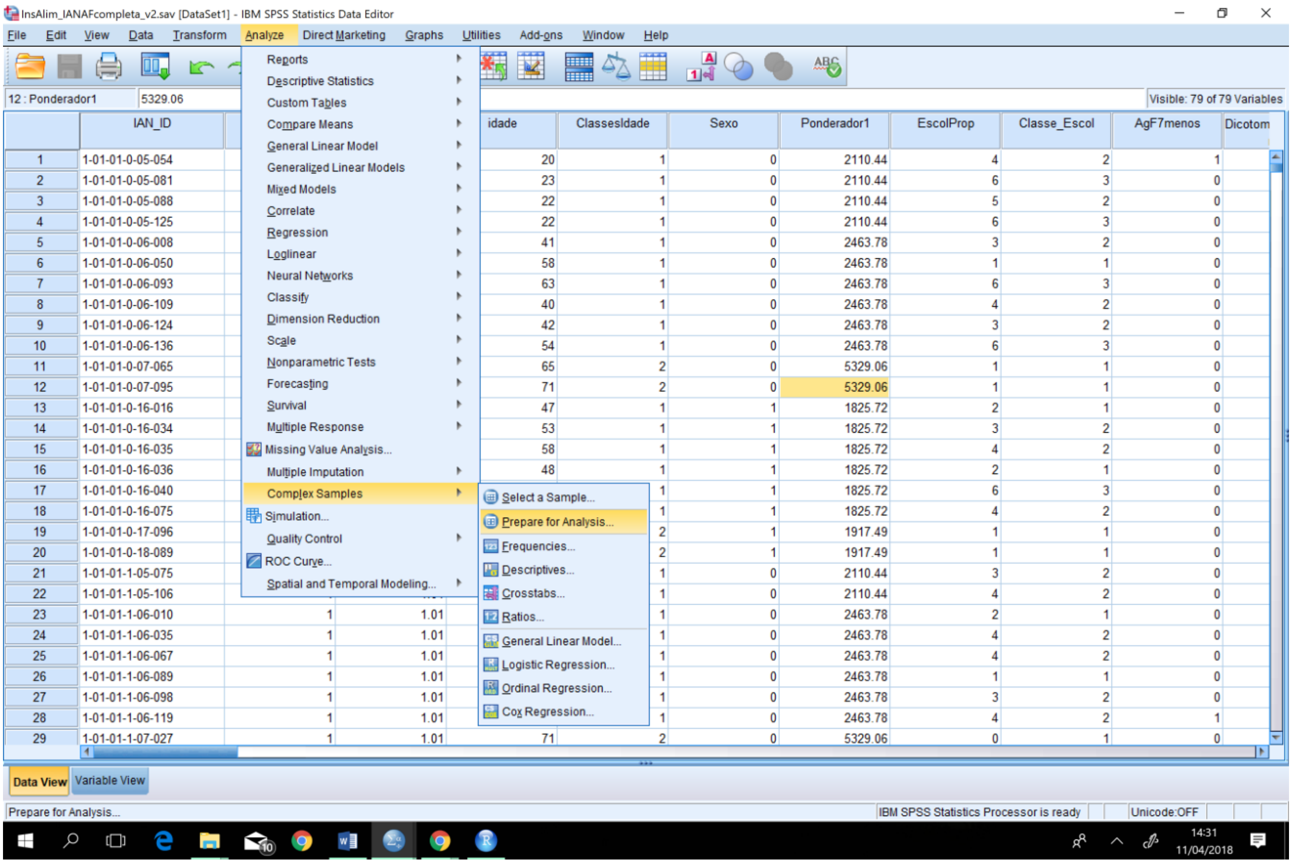
i. initial weighting to compensate for the different probabilities of selection of each Primary Health Units;
ii. weighting to compensate for the different probabilities of selection of each individual in each Primary Health Units, by sex and age group (considering the individuals in each Primary Health Units, in the closest recruitment wave);
iii. correction of the initial weights for the non-response bias.

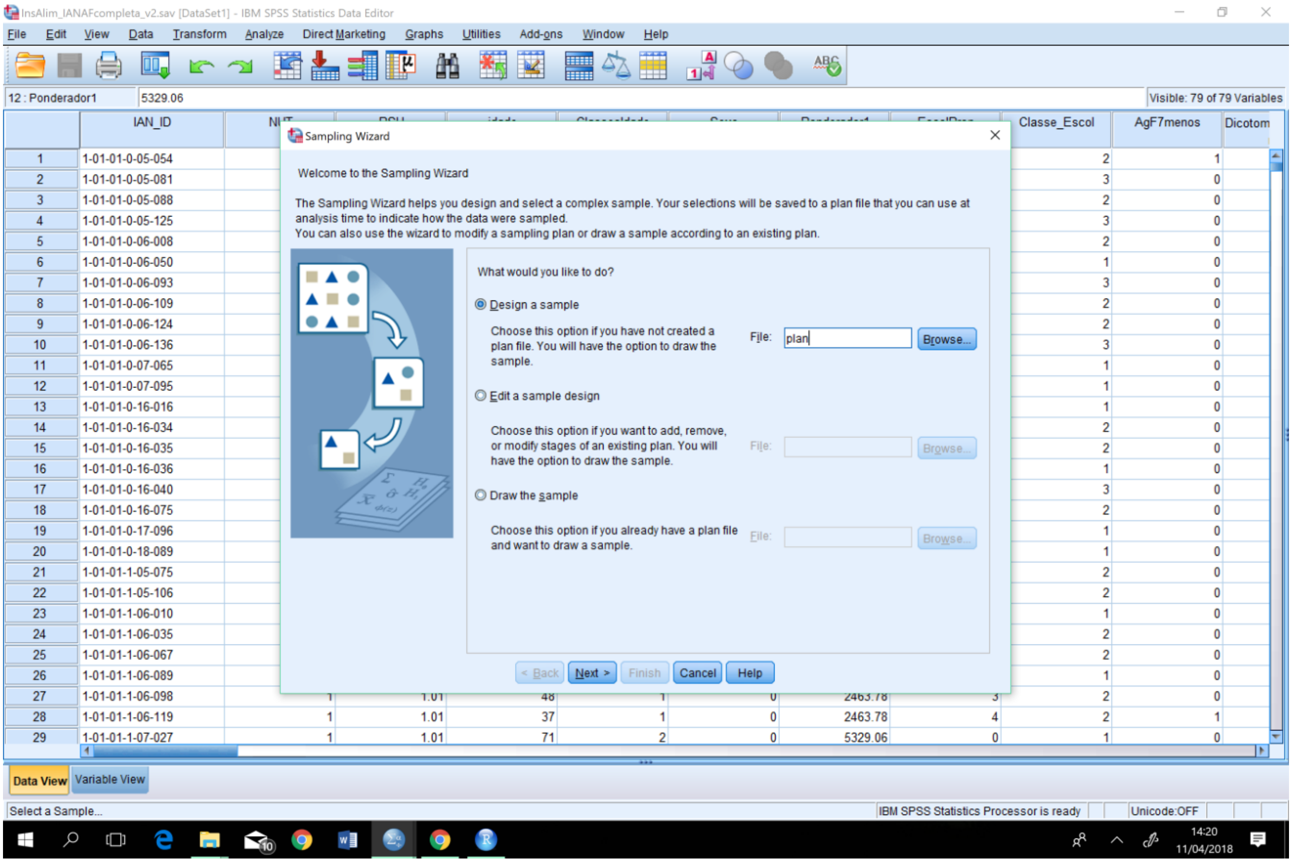
At the end, in order to correct data for non-response bias of both first and second interview, two weight variables were created. The first, ***Ponderador1***, is used for data collected in the first interview, and the second weight variable, ***Ponderador2***, is used for data collected in the second interview. **Thus, all estimates referring to the domains Physical Activity and Nutritional Status must use the weight variable *Ponderador1*, while the domain Food must use the weight variable *Ponderador2*.**

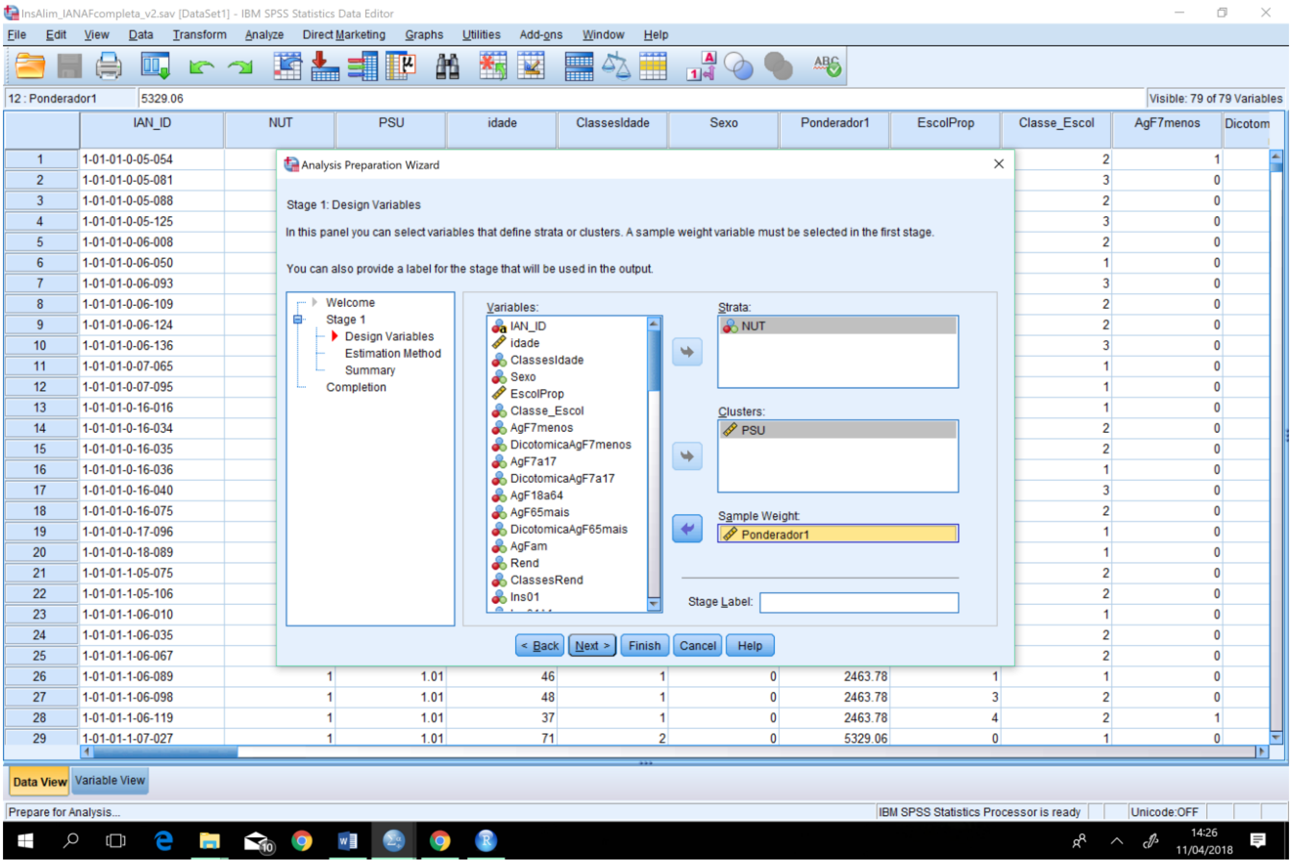
Next, we present a brief tutorial on how to use the SPSS and R [1] software in order to obtain weighted estimates according to the complex sampling design of the IAN-AF 2015-2016, using the SPSS and R software [1].

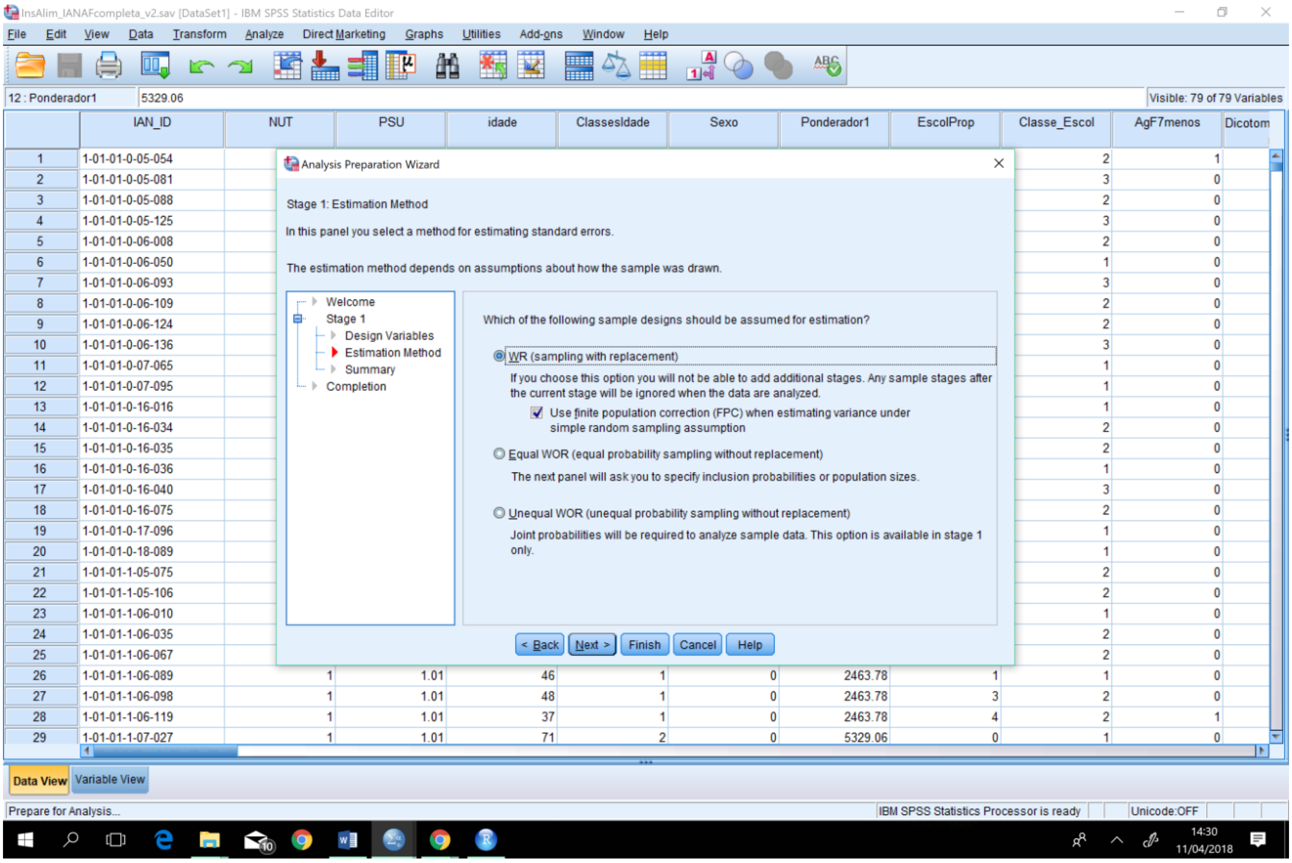
1.
Software SPSS

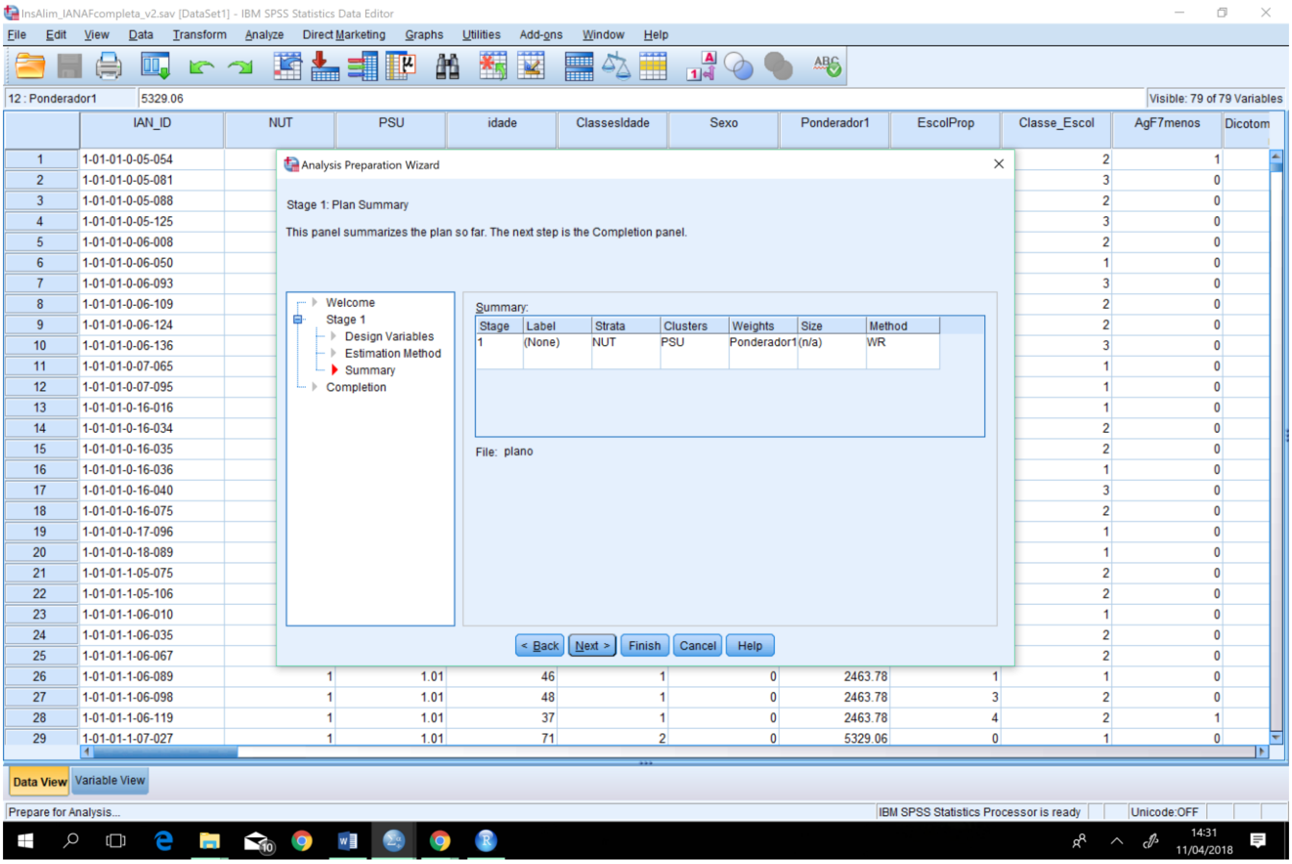
In order to obtain weighted estimates according to the IAN-AF 2015-2016 complex sampling design in SPSS, first it is necessary to create a file that indicates the complex sampling design used. To do it so, it is mandatory to have the variables "PSU", "NUT" and the respective weighting variable, which can be found in the sociodemographic database. Thus, it is always necessary to merge the sociodemographic database with the database containing the variables under study.

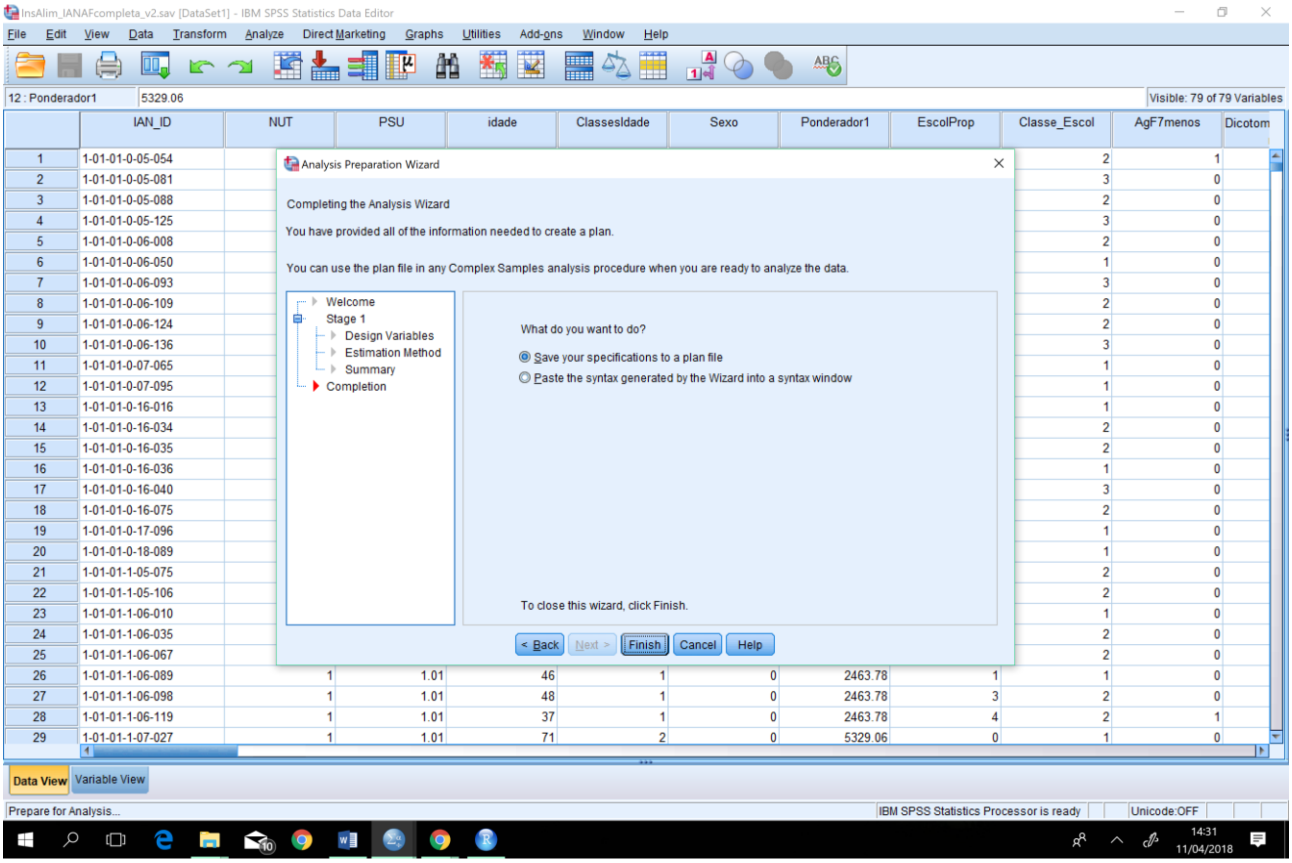



InsAlim_IANAFcompleta_v2.sav [DataSet1] - IBM SPSS Statistics Data Editor
Sampling Wizard
Welcome to the Sampling Wizard
The Sampling Wizard helps you design and select a complex sample. Your selections will be saved to a plan file that you can use at analysis time to indicate how the data were sampled.
You can also use the wizard to modify a sampling plan or draw a sample according to an existing plan.
What would you like to do?
Design a sample
Choose this option if you have not created a plan file. You will have the option to draw the sample.
File: plan
Browse...
Edit a sample design
Choose this option if you want to add, remove, or modify stages of an existing plan. You will have the option to draw the sample.
Draw the sample
Choose this option if you already have a plan file and want to draw a sample.
< Back
Next >
Finish
Cancel
Help
Data View
Variable View
Select a Sample...
IBM SPSS Statistics Processor is ready
Unicode:OFF

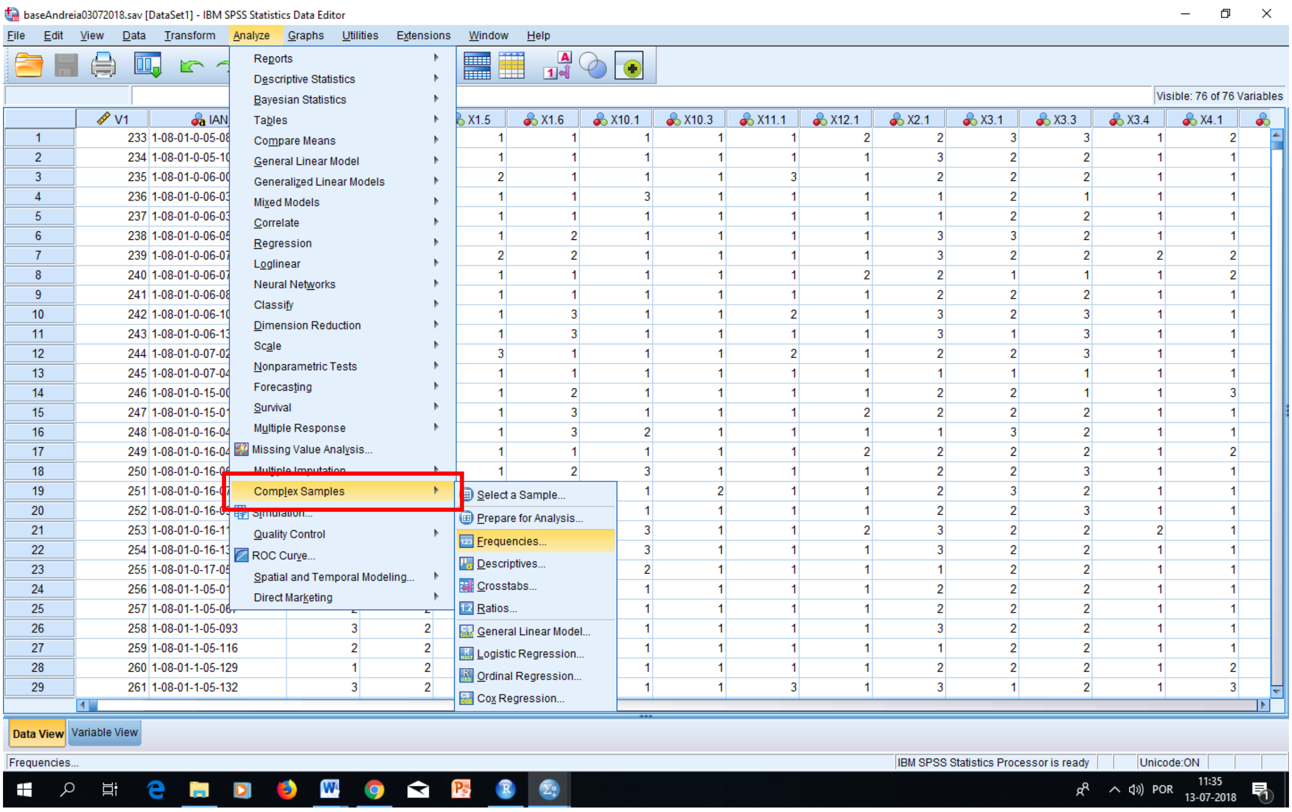

InsAlim_IANAFcompleta_v2.sav [DataSet1] - IBM SPSS Statistics Data Editor
Analysis Preparation Wizard
Stage 1: Design Variables
In this panel you can select variables that define strata or clusters. A sample weight variable must be selected in the first stage.
You can also provide a label for the stage that will be used in the output.
Welcome
Stage 1
Design Variables
Estimation Method
Summary
Completion
Variables:
IAN_ID
idade
ClassesIdade
Sexo
EscolProp
Classe_Escol
AgF7menos
DicotomicaAgF7menos
AgF7a17
DicotomicaAgF7a17
AgF18a64
AgF65mais
DicotomicaAgF65mais
AgFam
Rend
ClassesRend
Ins01
Strata:
NUT
Clusters:
PSU
Sample Weight:
Ponderador1
Stage Label:
< Back
Next >
Finish
Cancel
Help
Data View
Variable View
Prepare for Analysis...
IBM SPSS Statistics Processor is ready
Unicode:OFF


Analysis Preparation Wizard
Stage 1: Estimation Method
In this panel you select a method for estimating standard errors.
The estimation method depends on assumptions about how the sample was drawn.
Welcome
Stage 1
Design Variables
Estimation Method
Summary
Completion
Which of the following sample designs should be assumed for estimation?
WR (sampling with replacement)
If you choose this option you will not be able to add additional stages. Any sample stages after the current stage will be ignored when the data are analyzed.
Use finite population correction (FPC) when estimating variance under simple random sampling assumption
Equal WOR (equal probability sampling without replacement)
The next panel will ask you to specify inclusion probabilities or population sizes.
Unequal WOR (unequal probability sampling without replacement)
Joint probabilities will be required to analyze sample data. This option is available in stage 1 only.
< Back
Next >
Finish
Cancel
Help


Analysis Preparation Wizard
Stage 1: Plan Summary
This panel summarizes the plan so far. The next step is the Completion panel.
Welcome
Stage 1
Design Variables
Estimation Method
Summary
Completion
Summary:
Stage | Label | Strata | Clusters | Weights | Size | Method
1 | (None) | NUT | PSU | Ponderador1 | (n/a) | WR
File: plano
< Back
Next >
Finish
Cancel
Help

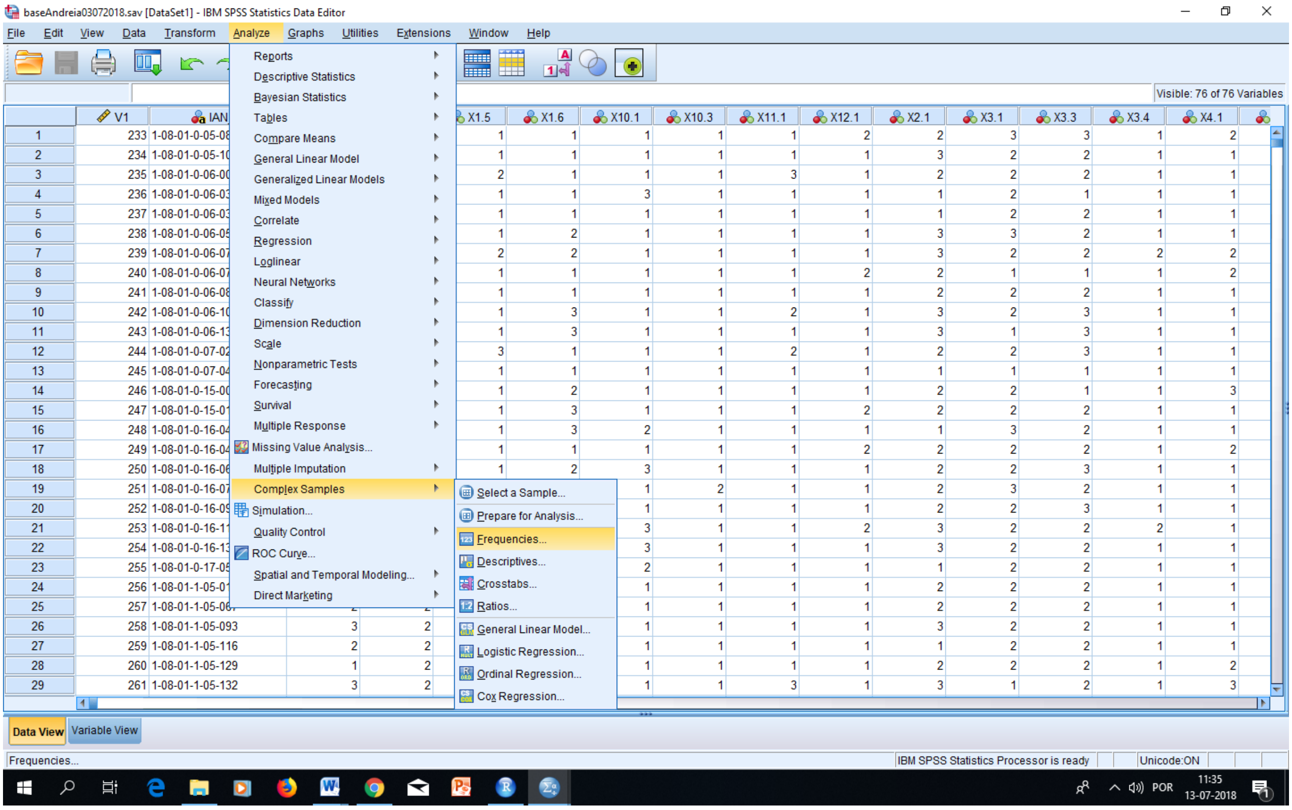


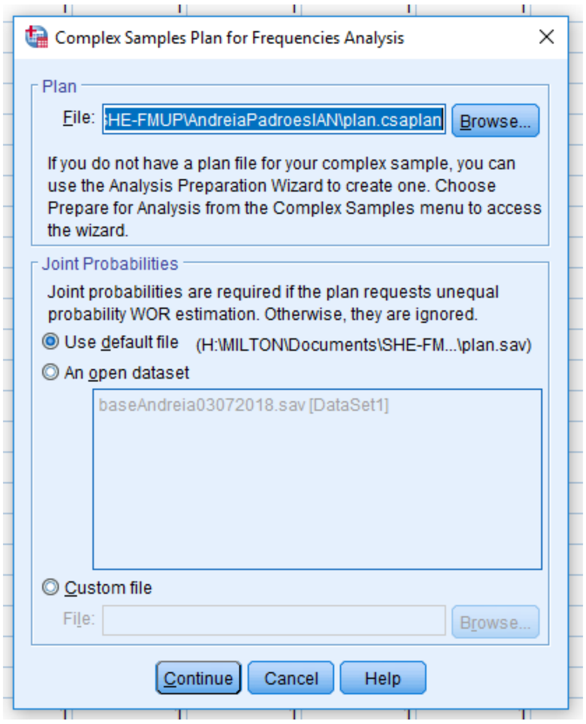
This file will be used to perform all weighted statistical analyses, which must be uniquely made in the **Analyze >> Complex Samples** menu.



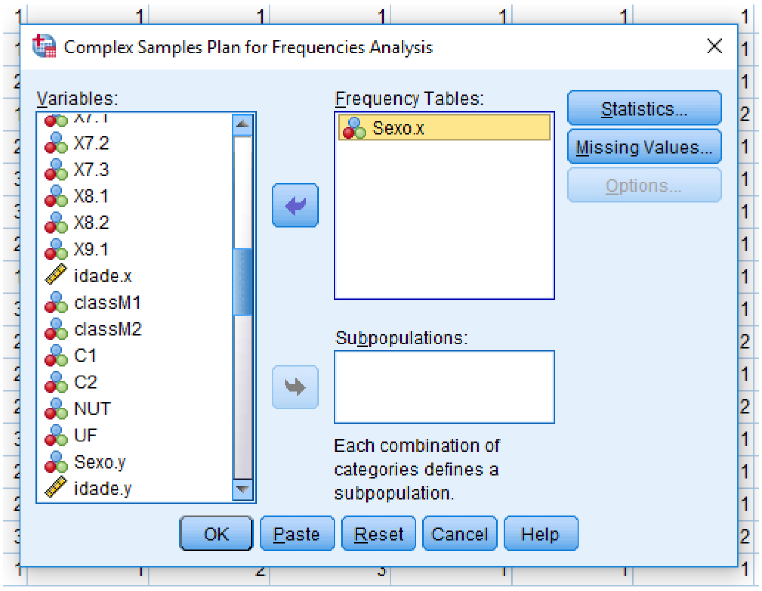
1.1. Weighted frequencies

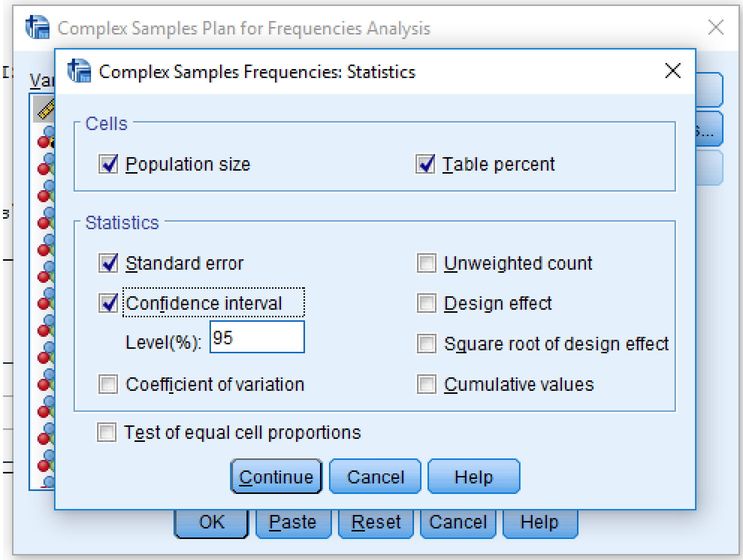
In order to estimate weighted frequencies, one should go to **Analyze >> Complex Samples >> Frequencies** and select the previously created file.





Next, one should select the variable under study and the associated statistics.





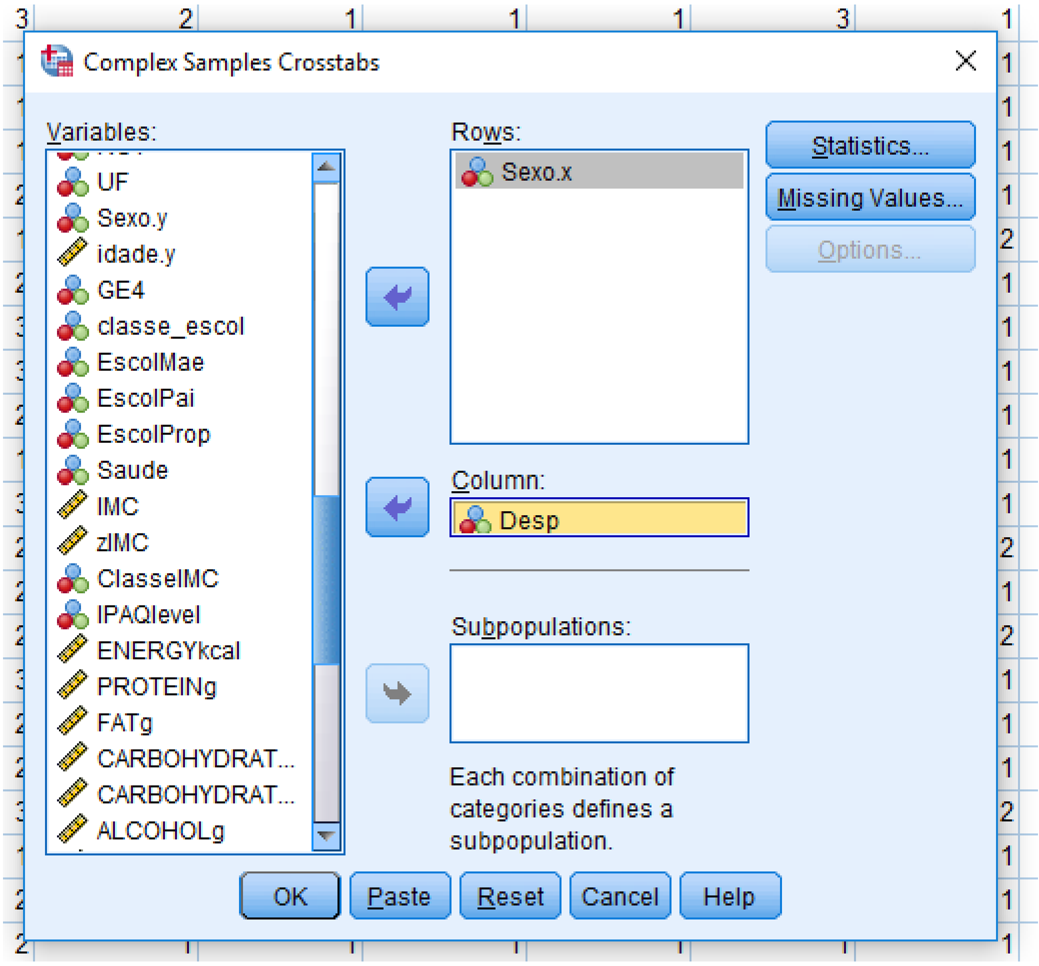
Result:

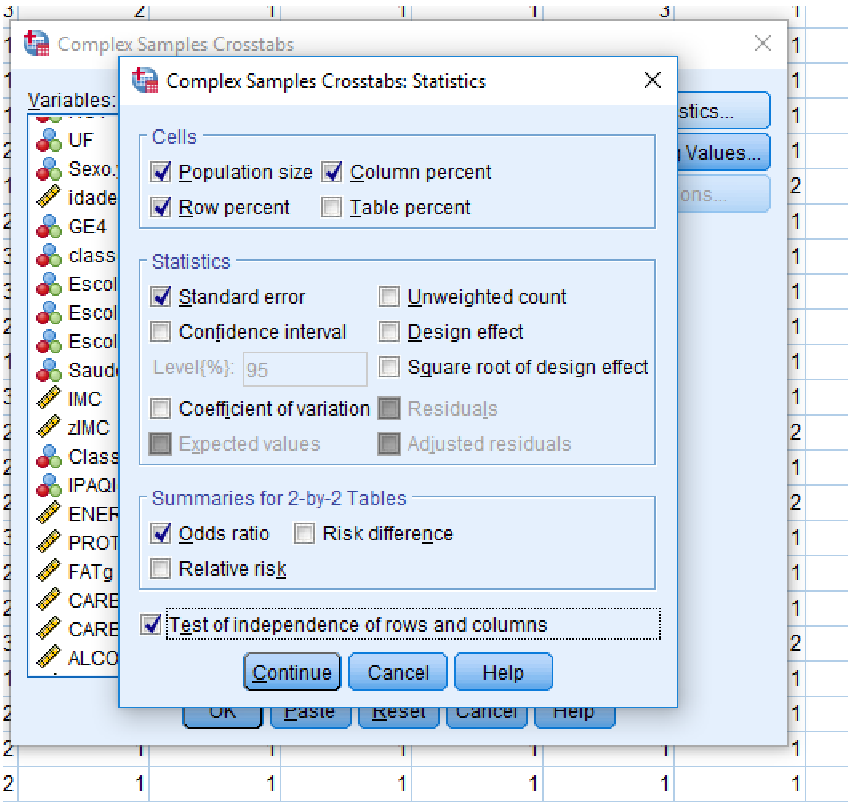
Sexo.x					
		Estimate	Standard Error	95% Confidence Interval	
				Lower	Upper
Population Size	0	4739432,770	145329,479	4450795,879	5028069,661
	1	4449227,520	126039,458	4198902,276	4699552,764
	Total	9188660,290	239273,706	8713442,056	9663878,524
% of Total	0	51,6%	0,7%	50,2%	53,0%
	1	48,4%	0,7%	47,0%	49,8%
	Total	100,0%	0,0%	100,0%	100,0%

1.2. Test independence/association between 2 categorical variables

In order to test the independence/association between two categorical variables, one should access to the **Analyze >> Complex Samples >> Crosstabs** menu and select the previously created file.

Then, select the variables under hypothesis and the respective statistics.





Result:

Sexo.x * Desp

Sexo.x			Desp 0	Desp 1	Total
0	Population Size	Estimate	2916200,750	1689662,870	4605863,620
		Standard Error	119981,932	104059,923	143375,307
	% within Sexo.x	Estimate	63,3%	36,7%	100,0%
		Standard Error	1,9%	1,9%	0,0%
	% within Desp	Estimate	53,4%	47,1%	50,9%
		Standard Error	1,3%	1,7%	0,7%
1	Population Size	Estimate	2547897,160	1899139,430	4447036,590
		Standard Error	109990,959	108317,206	126295,420
	% within Sexo.x	Estimate	57,3%	42,7%	100,0%
		Standard Error	2,0%	2,0%	0,0%
	% within Desp	Estimate	46,6%	52,9%	49,1%
		Standard Error	1,3%	1,7%	0,7%
Total	Population Size	Estimate	5464097,910	3588802,300	9052900,210
		Standard Error	183758,461	173125,807	234706,467
	% within Sexo.x	Estimate	60,4%	39,6%	100,0%
		Standard Error	1,5%	1,5%	0,0%
	% within Desp	Estimate	100,0%	100,0%	100,0%
		Standard Error	0,0%	0,0%	0,0%

Tests of Independence

		Chi-Square	Adjusted F	df1	df2	Sig.
Sexo.x * Desp	Pearson	14,388	6,020	1	92	,016
	Likelihood Ratio	14,394	6,022	1	92	,016

The adjusted F is a variant of the second-order Rao-Scott adjusted chi-square statistic. Significance is based on the adjusted F and its degrees of freedom.

Measures of Association

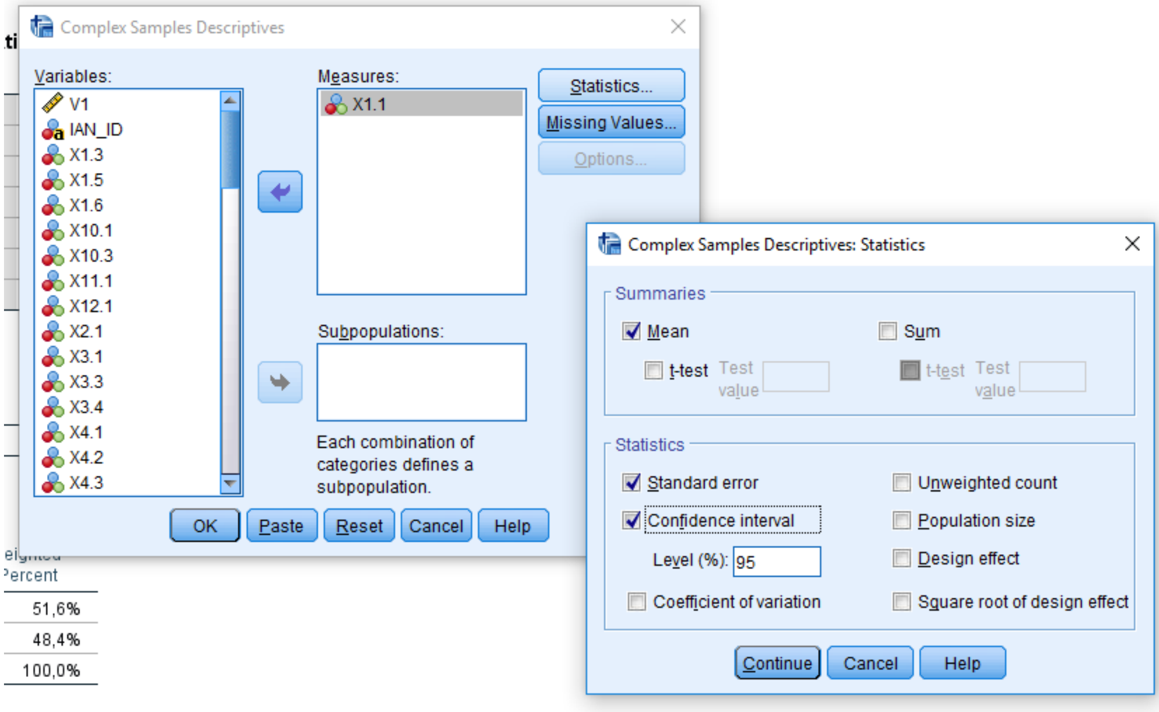
		Estimate
Sexo.x * Desp	Odds Ratio	1,286

Statistics are computed only for 2-by-2 tables with all cells observed.

1.3. Weighted mean

In order to estimate the weighted mean and the respective confidence interval of a continuous variable, one should access to the **Analyze >> Complex Samples >> Descriptives** menu and select the previously created file.

Then, select the continuous variable under study and the respective statistics.



Result:

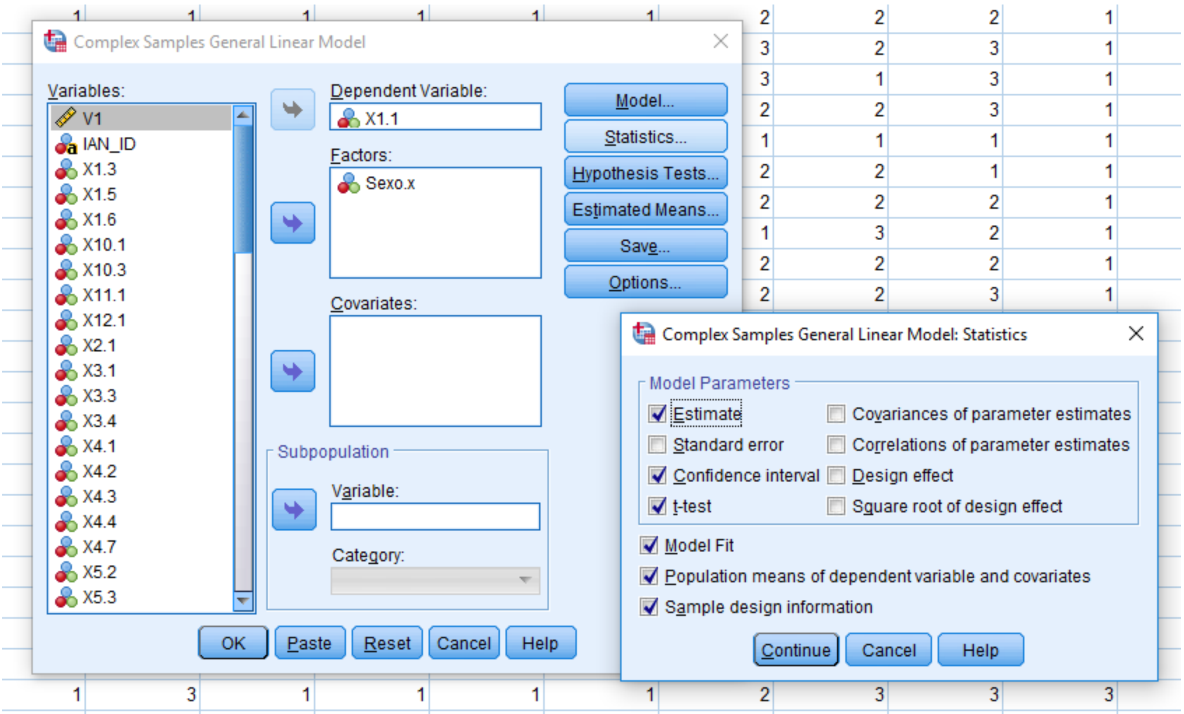
Univariate Statistics

		Estimate	Standard Error	95% Confidence Interval	
				Lower	Upper
Mean	X1.1	2,14	,027	2,09	2,19

1.4. Linear Regression

In order to compare weighted mean values or a linear regression for weighted data, one should access to the **Analyze >> Complex Samples >> General Linear Model** menu and select the previously created file.

Then, select the dependent variable and the independent variables, as well as the respective statistics. If a variable is of type categorical, then the variable must be added in "Factors". Otherwise, if a variable is of type continuous, then the variable must be added in "Covariates".



Resultado:

Parameter Estimates[a]

Parameter	Estimate	95% Confidence Interval		Hypothesis Test		
		Lower	Upper	t	df	Sig.
(Intercept)	2,129	2,056	2,203	57,592	92,000	,000
[Sexo.x=0]	,020	-,068	,108	,456	92,000	,649
[Sexo.x=1]	,000[b]	.	.	.	.	.

a. Model: X1.1 = (Intercept) + Sexo.x

b. Set to zero because this parameter is redundant.

2.

Software R

In order to obtain weighted estimates according to the IAN-AF 2015-2016 complex sampling design in R, the library “survey” is used [2,3].

```
> install.packages(“survey”)
> library(survey)
```

When creating the database, it is mandatory to include the variables "PSU", "NUT" and the respective weighting variable, which can be found in the sociodemographic database. Thus, it is always necessary to join the sociodemographic database with the variables under study.

```
# mudar nome das tabelas de acordo com os nomes dos ficheiros exportados
# mudar variável ponderador de acordo com as variáveis a analisar

> base = read.csv2("Tabela_Ponderador_Sociodem.csv", stringsAsFactors = F)
> atvfis = read.csv2("Tabela_AFisica.csv", stringsAsFactors = F)
> b = merge(base, atvfis)

> svdx<-svydesign(id = ~PSU, strata = ~NUT, weights = ~Ponderador1, data = b)
> summary(svdx)
```

Next, some statistical analysis using the indicated library are exemplified. More information about the implemented functions in this library is available in the respective documentation.

2.1. Weighted frequency and mean values of categorical and continuous variables, respectively

The "svymean" function calculates the weighted mean of a variable according to the complex sampling design previously established. If the variable under study is of type "factor", then this function calculates the weighted proportion of each category of the variable.

```
> svymean(~idade, svdx)
        mean     SE
idade 42.686 0.3652

> svymean(~factor(Sexo), svdx)
                 mean     SE
factor(Sexo)0 0.51217 0.0064
factor(Sexo)1 0.48783 0.0064
```

2.2. Statistics on subsets

In order to estimate statistics on subsets defined by a factor, use the "svyby" function.

```
> svyby(~idade, ~Sexo, svdx, svymean)
  Sexo    idade        se
0    0 42.22272 0.4738476
1    1 42.11595 0.4994525
```

It is also possible to define separately a subset, and proceed as usual.

```
> subsvdx = subset(svdx, Sexo==1)
> svymean(~idade, subsvdx)
        mean    SE
idade 42.116 0.475
```

2.3. Hypothesis tests

t-test for comparison of mean values:

```
> svyttest(Idade~factor(Sexo), svdx)

        Design-based t-test

data:  Idade ~ factor(Sexo)
t = -2.1346, df = 91, p-value = 0.03548
alternative hypothesis: true difference in mean is not equal to 0 sample estimates:
difference in mean
         -1.153271
```

$\chi 2$-test to comparison of proportions:

```
> svychisq(~GE4+Sexo, svdx)

        Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~GE4 + Sexo, svdx)
F = 4.4883, ndf = 1.9053, ddf = 175.2800, p-value = 0.01385
```

2.4. Regression models

Linear regression model:

```
> m1=svyglm(IMC ~ Sexo + Idade + factor(EscolClass_Prop) , family=gaussian(), svdx)
> summary(m1)

Call:
svyglm(formula = IMC ~ Sexo + Idade + factor(EscolClass_Prop),
    family = gaussian(), subsvdx)

Survey design:
svdx

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              24.445613   0.472124  51.778  < 2e-16 ***
Sexo                     -0.332601   0.241667  -1.376    0.172
Idade                     0.084928   0.007141  11.894  < 2e-16 ***
factor(EscolClass_Prop)2 -1.399916   0.272237  -5.142 1.63e-06 ***
factor(EscolClass_Prop)3 -2.057181   0.269839  -7.624 2.70e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 20.84462)

Number of Fisher Scoring iterations: 2

> cbind(coef(m1),confint(m1))
                                            2.5 %      97.5 %
(Intercept)              24.44561278 23.52026639 25.37095917
Sexo                     -0.33260125 -0.80626059  0.14105808
Idade                     0.08492765  0.07093221  0.09892308
factor(EscolClass_Prop)2 -1.39991563 -1.93349039 -0.86634087
factor(EscolClass_Prop)3 -2.05718129 -2.58605546 -1.52830711
```

Logistic regression model:

```
> m1=svyglm(factor(Desp) ~ factor(GrupoEtario), family=binomial(link = 'logit'), svdx)
> summary(m1)

Call:
svyglm(formula = factor(Desp) ~ factor(GrupoEtario), family = binomial(link = "logit"),
    subsvdx)

Survey design:
svdx

Coefficients:
                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)            0.44697    0.14980   2.984   0.00367 **
factor(GrupoEtario)2 -0.08235    0.18099  -0.455   0.65023
factor(GrupoEtario)3 -0.83873    0.15511  -5.407  5.32e-07 ***
factor(GrupoEtario)4 -1.15278    0.18788  -6.136  2.30e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000187)

Number of Fisher Scoring iterations: 4

> cbind(exp(coef(m1)),exp(confint(m1)))
                                    2.5 %    97.5 %
(Intercept)          1.5601185 1.1636513 2.0916658
factor(GrupoEtario)2 0.9240598 0.6467305 1.3203127
factor(GrupoEtario)3 0.4309102 0.3187190 0.5825935
factor(GrupoEtario)4 0.3164551 0.2187010 0.4579029
```

ian·af

INQUÉRITO ALIMENTAR NACIONAL
E DE ATIVIDADE FÍSICA