INQUÉRITO ALIMENTAR NACIONAL
E DE ATIVIDADE FÍSICA

AMOSTRAGEM COMPLEXA

Bases de Dados IAN-AF

Tutorial para análise ponderada recorrendo aos softwares SPSS e R

Conteúdo

Referências

[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

Nota Introdutória

No Inquérito Alimentar Nacional e de Atividade Física, IAN-AF 2015-2016, os participantes foram selecionados aleatoriamente por um processo de amostragem complexa bietápica, a partir do Registo Nacional de Utentes do Serviço Nacional de Saúde. O processo de amostragem desenvolveu-se da seguinte forma:

i. Selecionou-se aleatoriamente Unidades Funcionais de Saúde (UFS) em cada Unidade Territorial para Fins Estatísticos (NUTS II), ponderada para o número de inscritos; o número de USF selecionadas foi 21 nas regiões do Norte, Centro e Área Metropolitana de Lisboa, 12 nas regiões do Algarve a Alentejo e seis nas Regiões Autónomas da Madeira e Açores.
ii. Selecionou-se aleatoriamente indivíduos registados em cada Unidade Funcional de Saúde, com um número fixo de elementos por sexo e grupo etário.

Para calcular as estimativas do IAN-AF 2015-2016, a nível nacional e regional, a análise estatística assume uma ponderação dos dados amostrais. O peso amostral representa quantos indivíduos (em número) da população Portuguesa representa cada indivíduo da amostra em estudo. O cálculo dos pesos amostrais incluiu os seguintes critérios:

i. ponderação inicial para compensar as diferentes probabilidades de seleção de cada Unidade Funcional de Saúde;
ii. ponderação para compensar as diferentes probabilidades de seleção de cada indivíduo em cada Unidade de Saúde, por sexo e grupo etário (considerando os indivíduos inscritos no RNU na onda de recrutamento mais próxima)
iii. correção dos pesos iniciais para o viés de não-resposta.

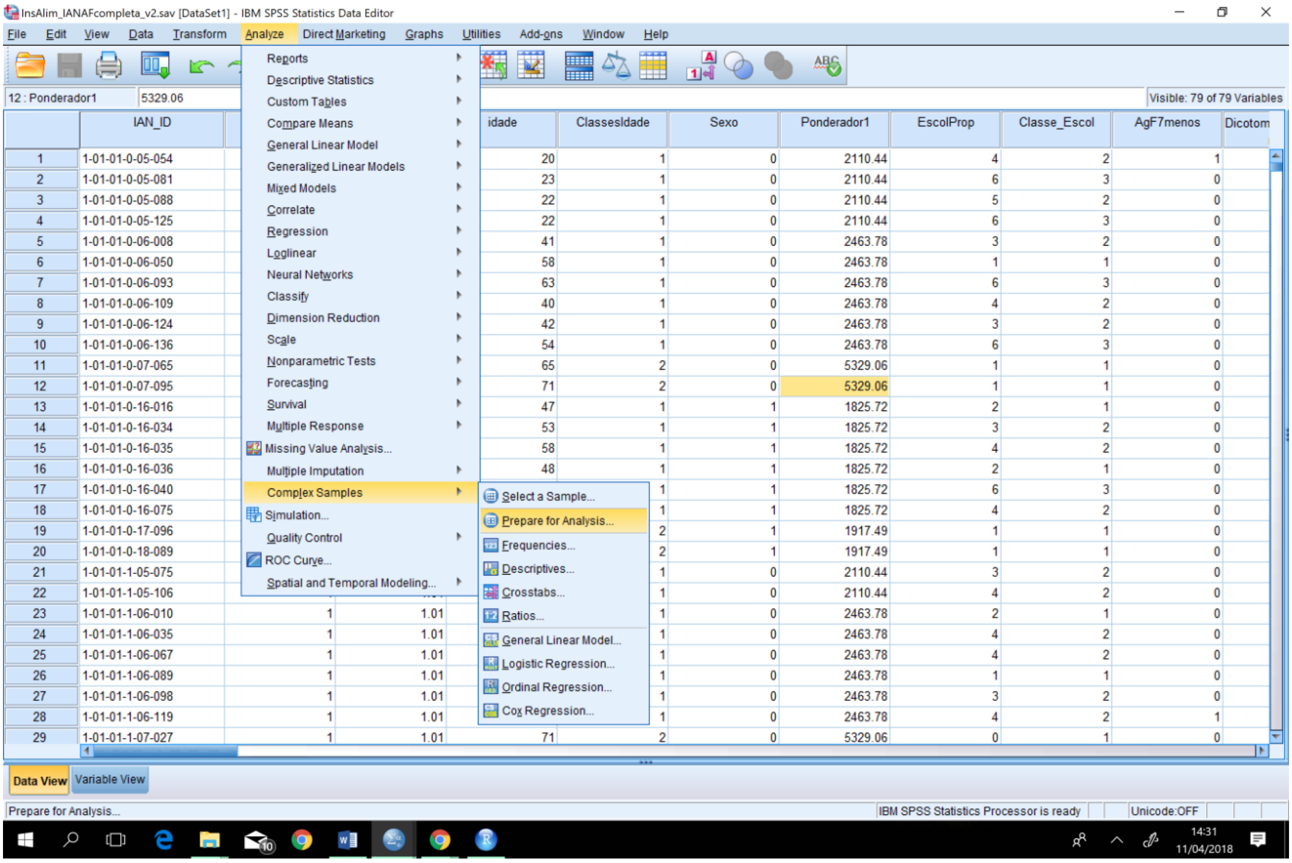
No final, criaram-se dois ponderadores de forma a cobrir as diferentes dimensões analisadas, sendo que o primeiro ponderador, ***Ponderador1***, corresponde às dimensões analisadas na primeira entrevista e o segundo, ***Ponderador2***, corresponde às dimensões analisadas na segunda entrevista. **Assim, todas estimativas referentes às dimensões Atividade Física e Estado Nutricional devem ser feitas recorrendo ao Ponderador1, enquanto que a dimensão Alimentação deve utilizar o Ponderador2.**

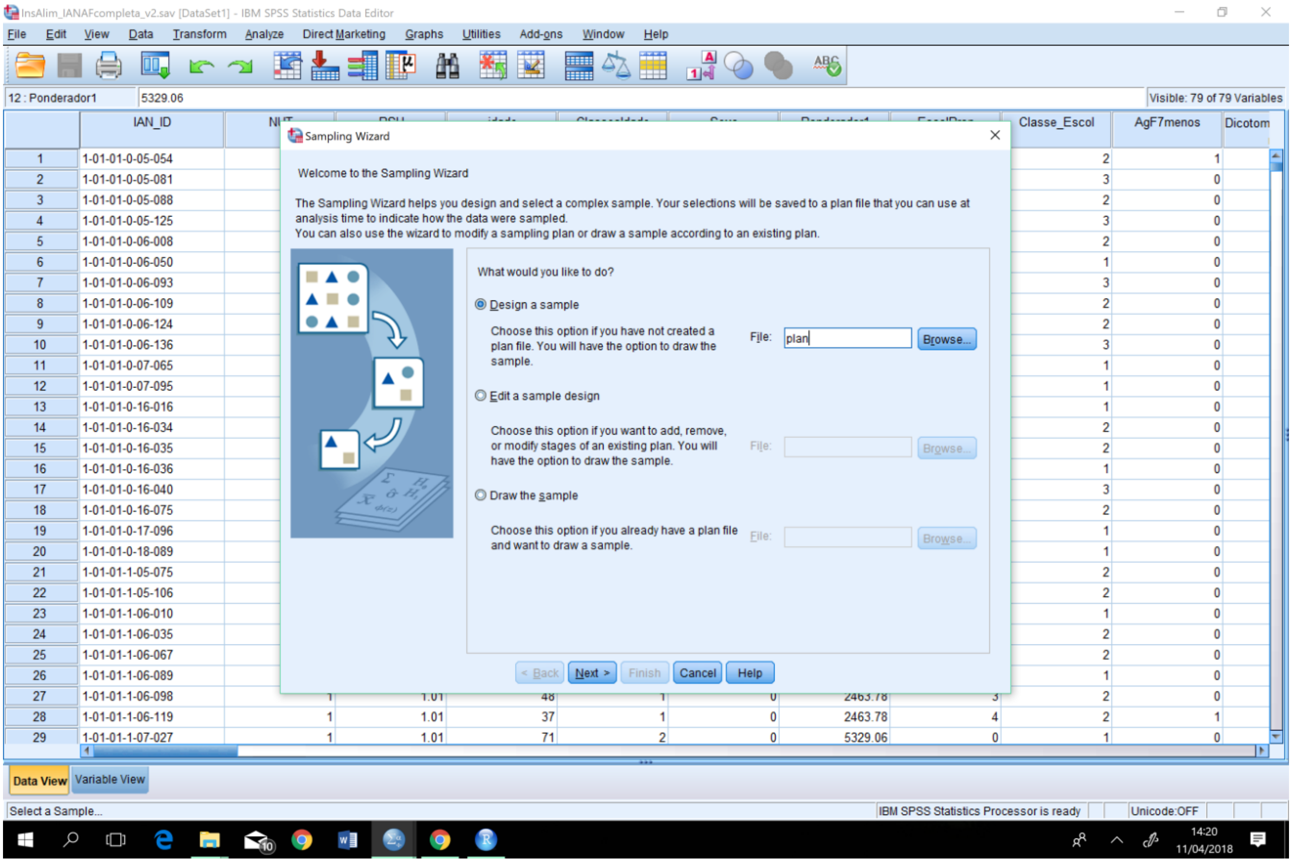
Neste tutorial, exemplifica-se as etapas a seguir de forma a obter estimativas ponderadas de acordo com o desenho de amostragem complexo do IAN-AF 2015-2016, utilizando os softwares SPSS e R [1].

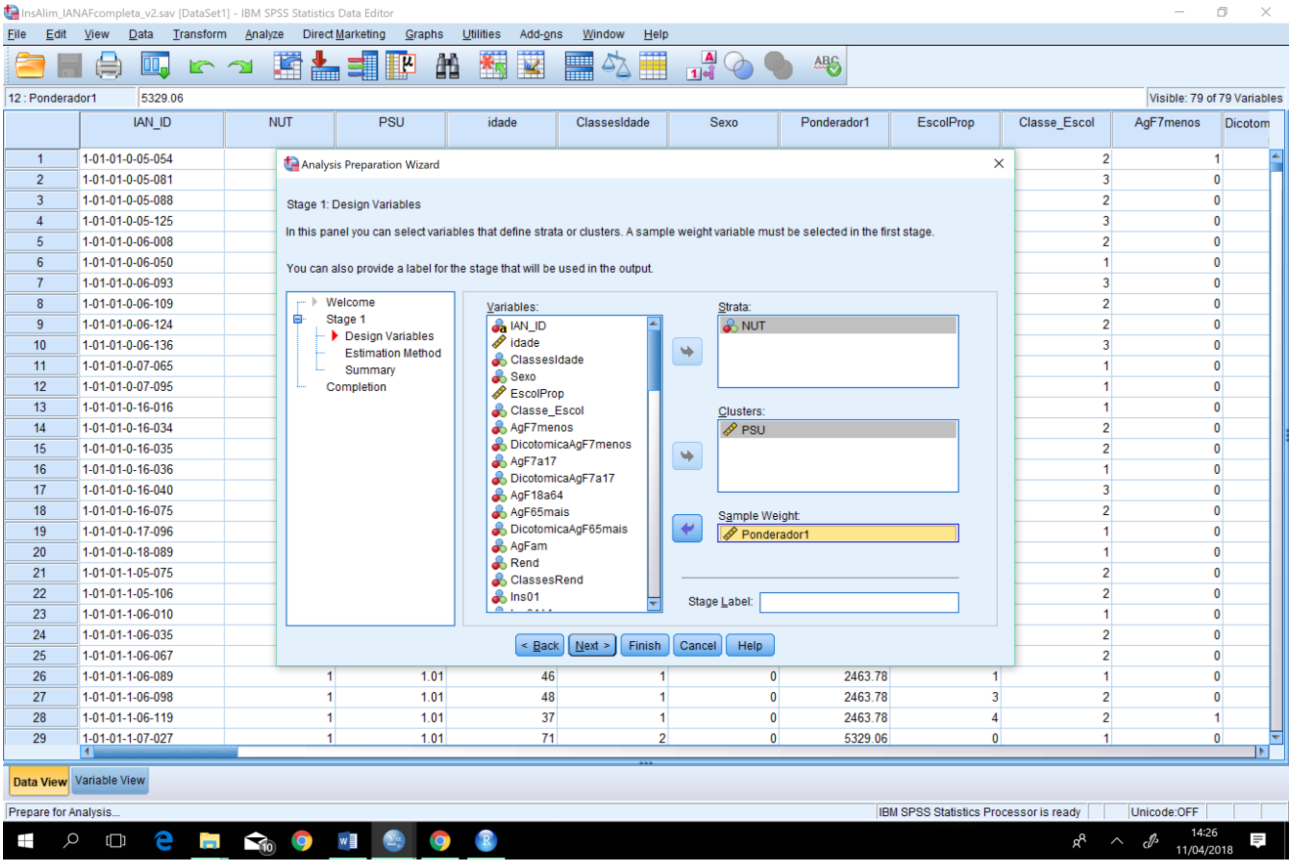
1.

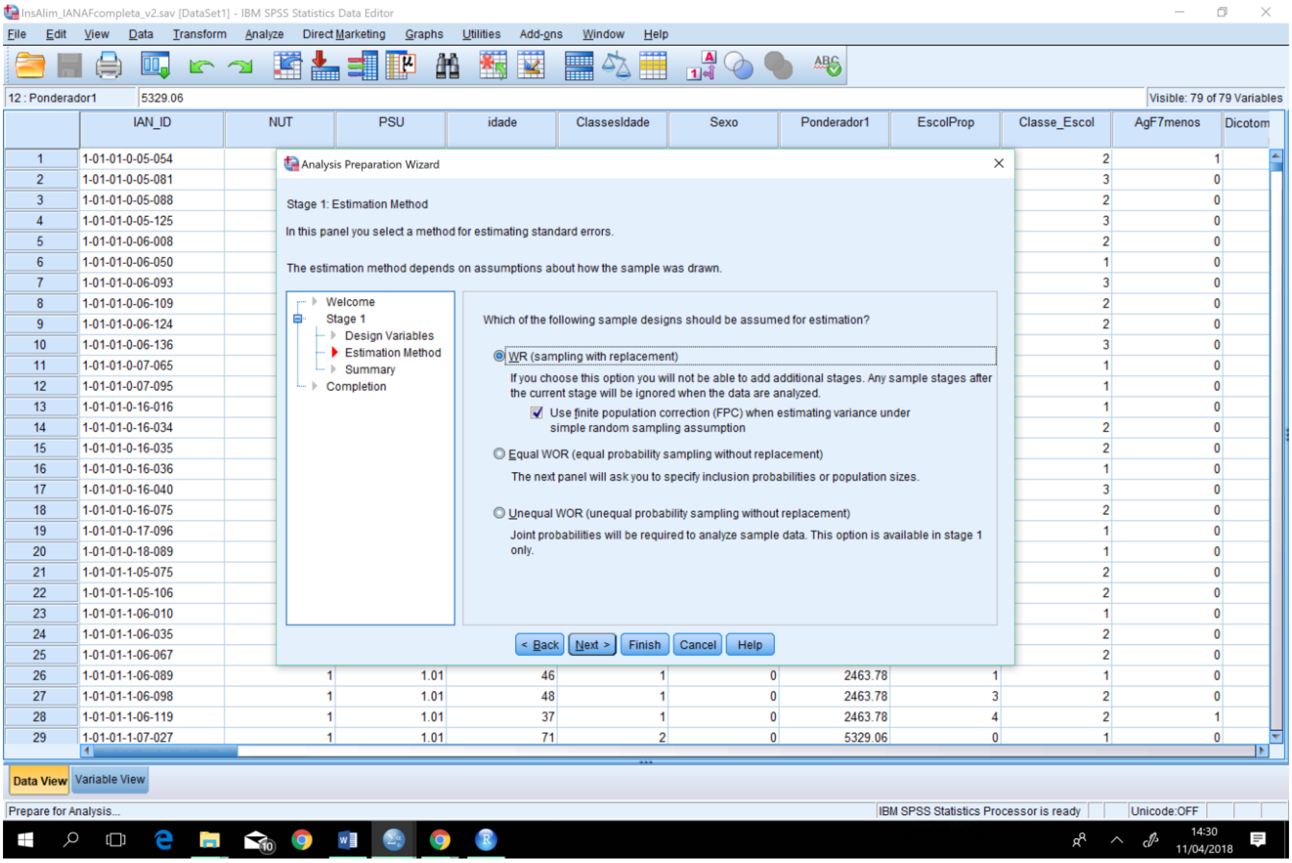
Software SPSS

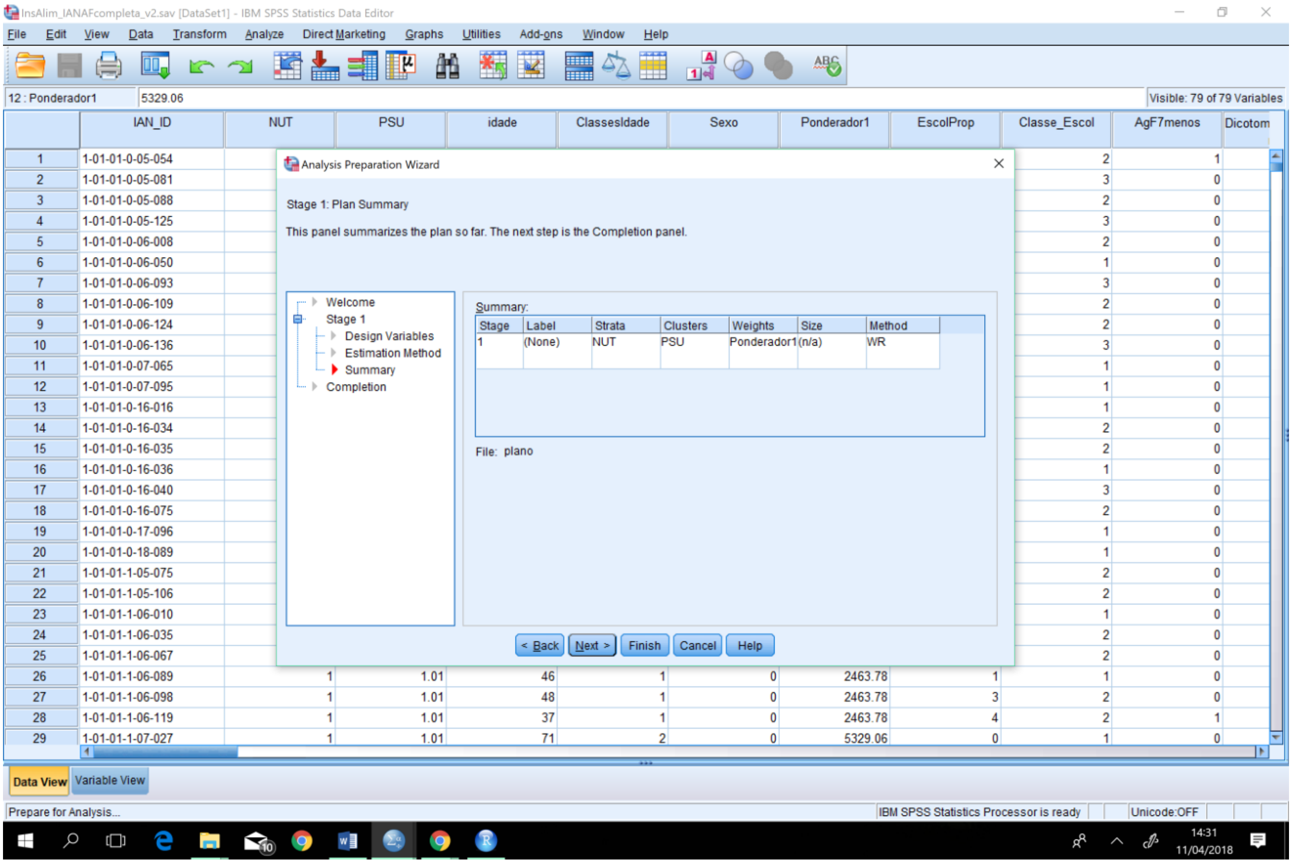
De forma a obter estimativas ponderadas em SPSS de acordo com o desenho de amostragem complexo IAN-AF 2015-2016, é necessário, numa primeira fase, construir um ficheiro indicador do desenho da amostragem complexa.

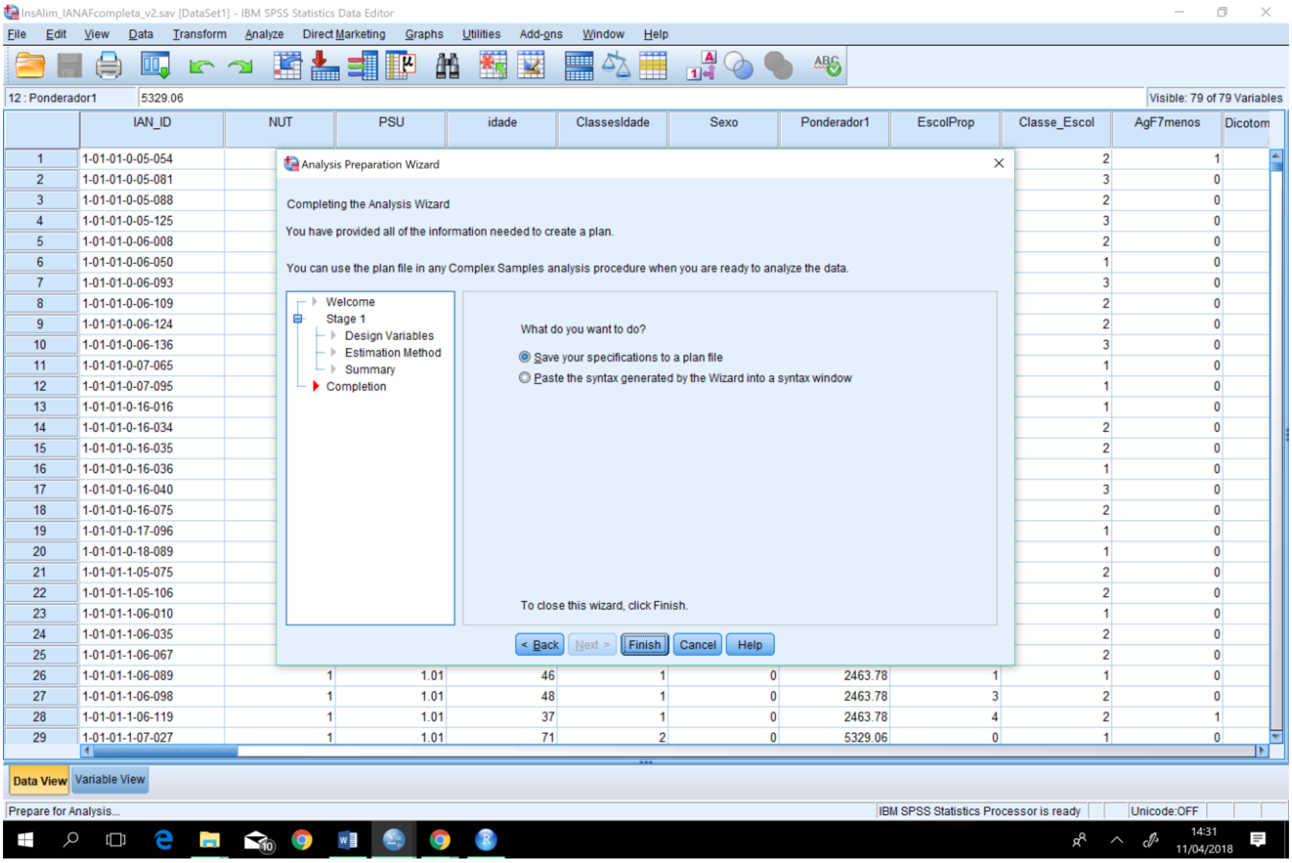



Sampling Wizard
Welcome to the Sampling Wizard
The Sampling Wizard helps you design and select a complex sample. Your selections will be saved to a plan file that you can use at analysis time to indicate how the data were sampled.
You can also use the wizard to modify a sampling plan or draw a sample 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

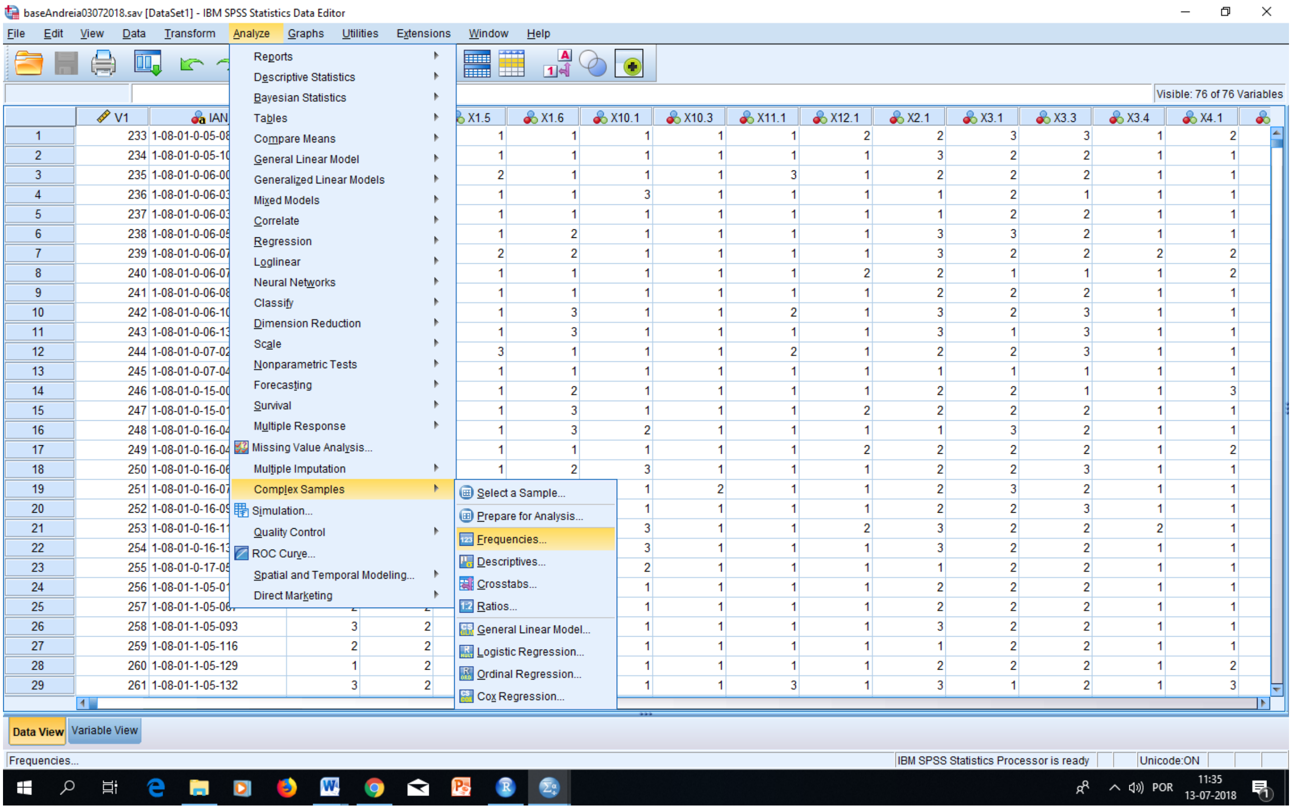

Analysis Preparation Wizard
Stage 1: Design Variables
In this panel you can select variables that define strata or clusters. A sample weight variable must be selected in the first stage.
You can also provide a label for the stage that will be used in the output.
Welcome
Stage 1
Design Variables
Estimation Method
Summary
Completion
Variables:
Strata:
NUT
Clusters:
PSU
Sample Weight:
Ponderador1
Stage Label:
< Back
Next >
Finish
Cancel
Help


InsAlim_IANAFcompleta_v2.sav [DataSet1] - IBM SPSS Statistics Data Editor
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


InsAlim_IANAFcompleta_v2.sav [DataSet1] - IBM SPSS Statistics Data Editor
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

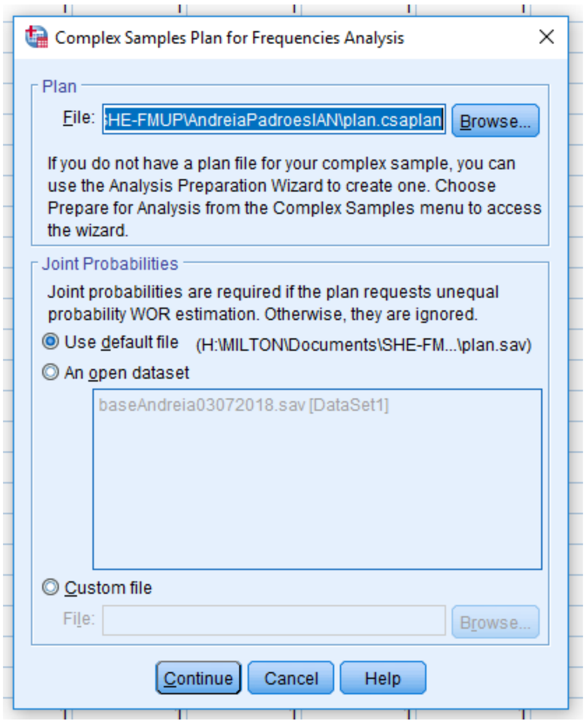


Este ficheiro será usado para todas as análises estatísticas que terão de ser realizadas obrigatoriamente no menu Analyze >> Complex Samples.

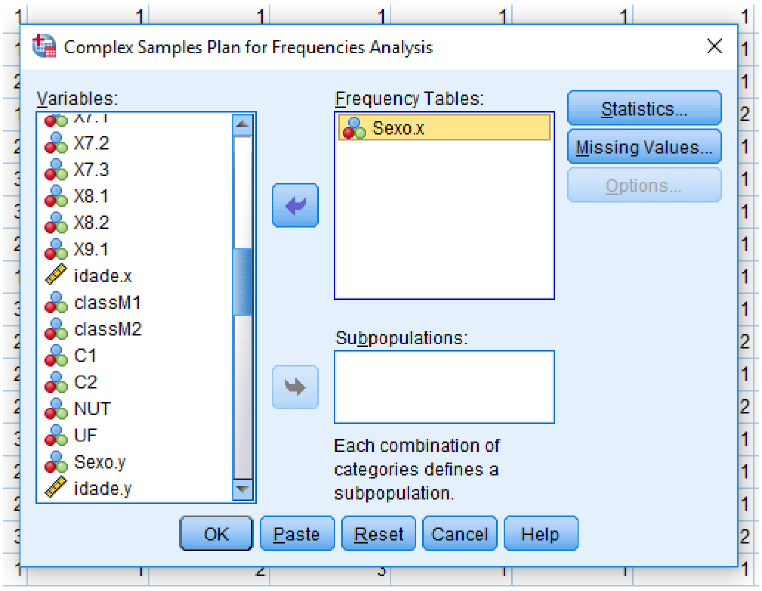


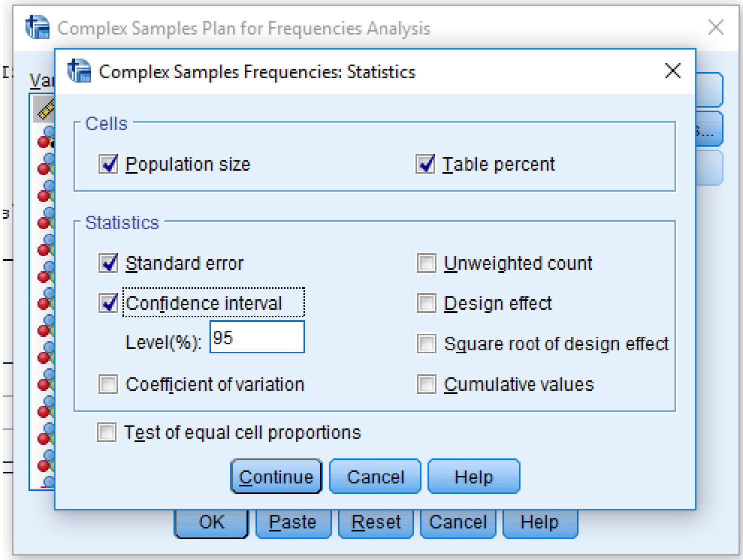
1.1. Estimar frequências ponderadas

Para estimar frequências ponderadas, deve-se aceder a Analyze >> Complex Samples >> Frequencies e selecionar o ficheiro anteriormente construído.



De seguida, seleciona-se a variável para a qual queremos estimar as frequências ponderadas e as respetivas estatísticas associadas.





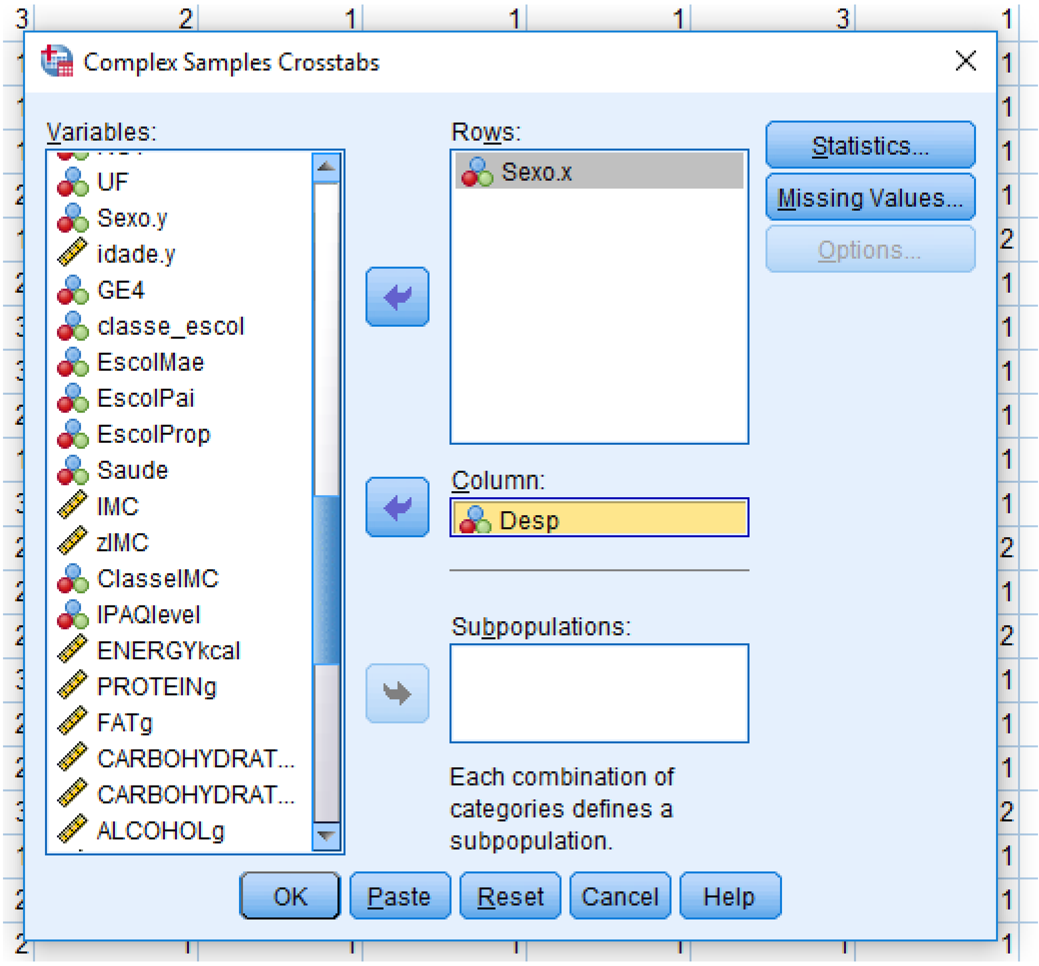
Resultado:

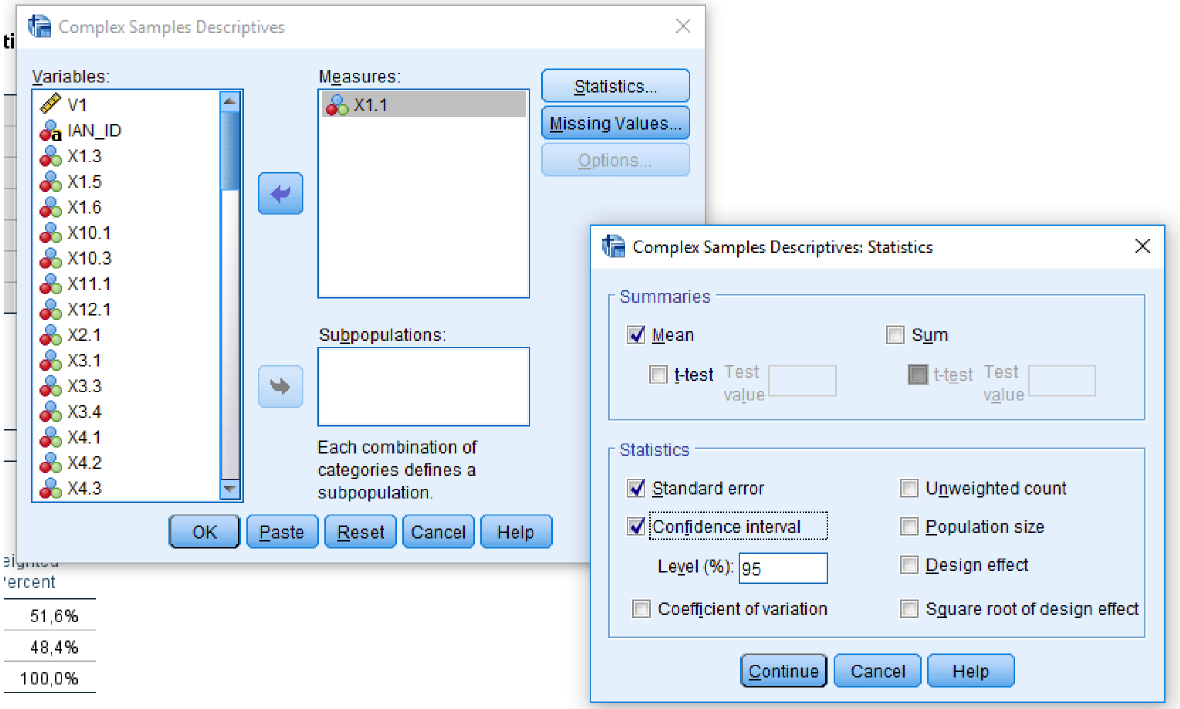
Sexo.x						
		Estimate	Standard Error	95% Confidence Interval	
				Lower	Upper
Population Size	0	4739432,770	145329,479	4450795,879	5028069,661
	1	4449227,520	126039,458	4198902,276	4699552,764
	Total	9188660,290	239273,706	8713442,056	9663878,524
% of Total	0	51,6%	0,7%	50,2%	53,0%
	1	48,4%	0,7%	47,0%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

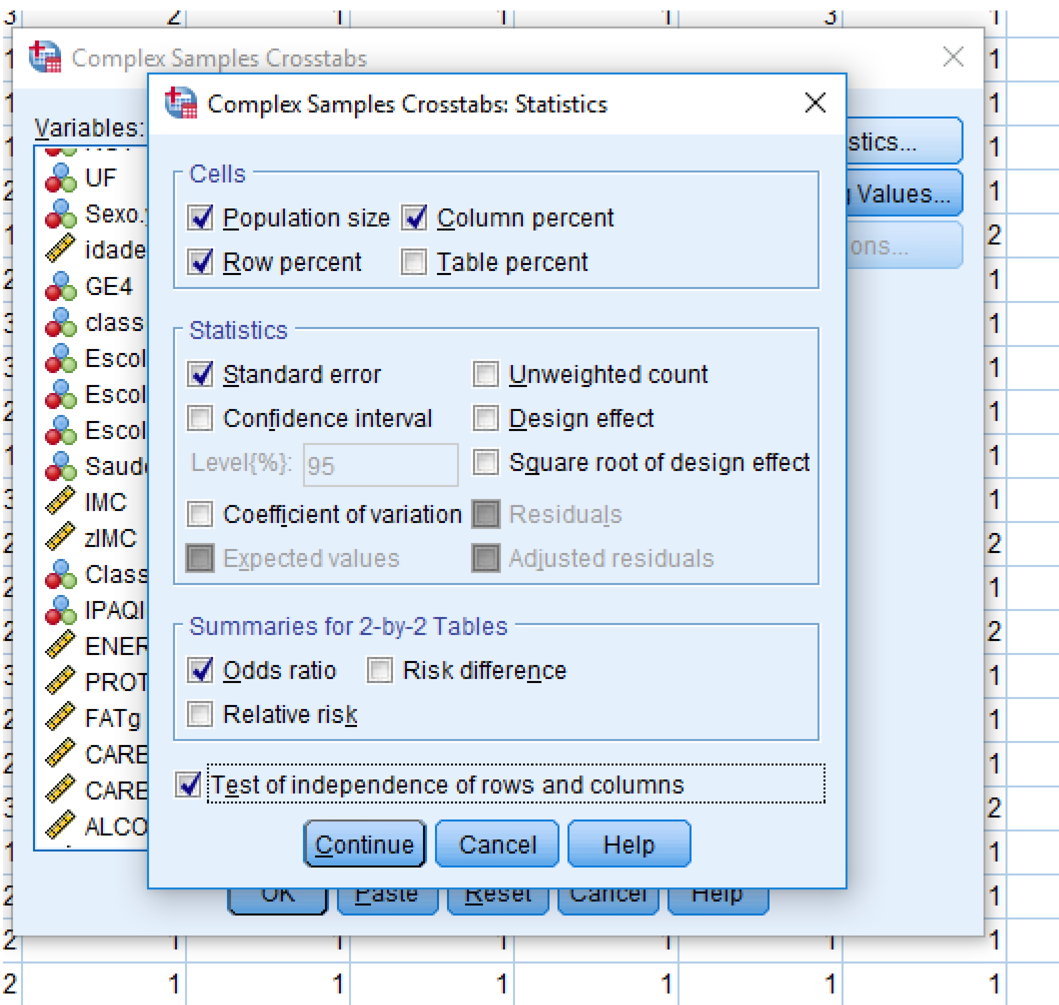
1.2. Testar a independência/associação entre 2 variáveis categóricas

Para testar a independência/associação entre duas variáveis categóricas, deve-se aceder a Analyze >> Complex Samples >> Crosstabs e selecionar o ficheiro anteriormente construído.

De seguida, selecionam-se as variáveis a testar e as estatísticas desejadas.




Complex Samples Descriptives
Variables:
V1
IAN_ID
X1.3
X1.5
X1.6
X10.1
X10.3
X11.1
X12.1
X2.1
X3.1
X3.3
X3.4
X4.1
X4.2
X4.3
Measures:
X1.1
Subpopulations:
Each combination of categories defines a subpopulation.
Statistics...
Missing Values...
Options...
OK
Paste
Reset
Cancel
Help
Percent
51,6%
48,4%
100,0%
Complex Samples Descriptives: Statistics
Summaries
Mean
Sum
t-test
Test value
Statistics
Standard error
Unweighted count
Confidence interval
Population size
Level (%): 95
Design effect
Coefficient of variation
Square root of design effect
Continue
Cancel
Help


Complex Samples Crosstabs
Variables:
UF
Sexo
idade
GE4
class
IMC
zIMC
IPAQI
FATg
Complex Samples Crosstabs: Statistics
Cells
Population size
Column percent
Row percent
Table percent
Statistics
Standard error
Unweighted count
Confidence interval
Design effect
Level(%): 95
Square root of design effect
Coefficient of variation
Residuals
Expected values
Adjusted residuals
Summaries for 2-by-2 Tables
Odds ratio
Risk difference
Relative risk
Test of independence of rows and columns
Continue
Cancel
Help
OK
Paste
Reset
Cancel
Help

Resultado:

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

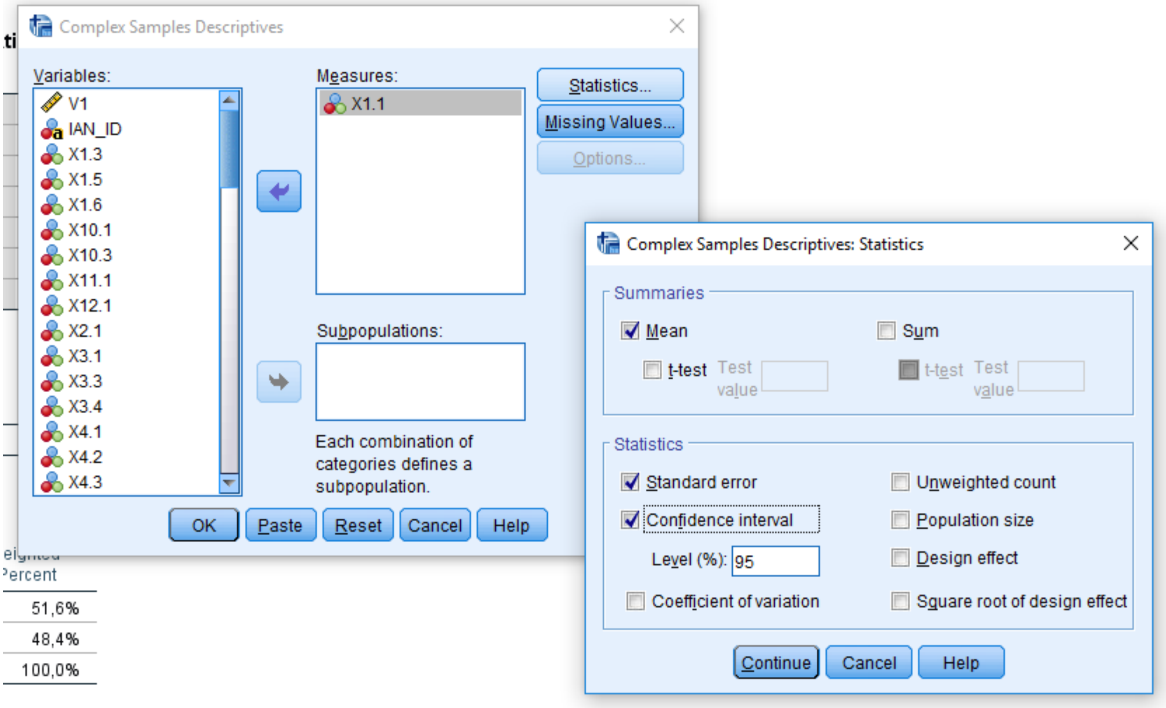
		Estimate
Sexo.x * Desp	Odds Ratio	1,286

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

1.3. Estimar média ponderada

Para testar a independência/associação entre duas variáveis categóricas, deve-se aceder a Analyze >> Complex Samples >> Descriptives e selecionar o ficheiro anteriormente construído.

De seguida, selecionam-se a variáveis cuja média se deseja estimar e as estatísticas desejadas.



Resultado:

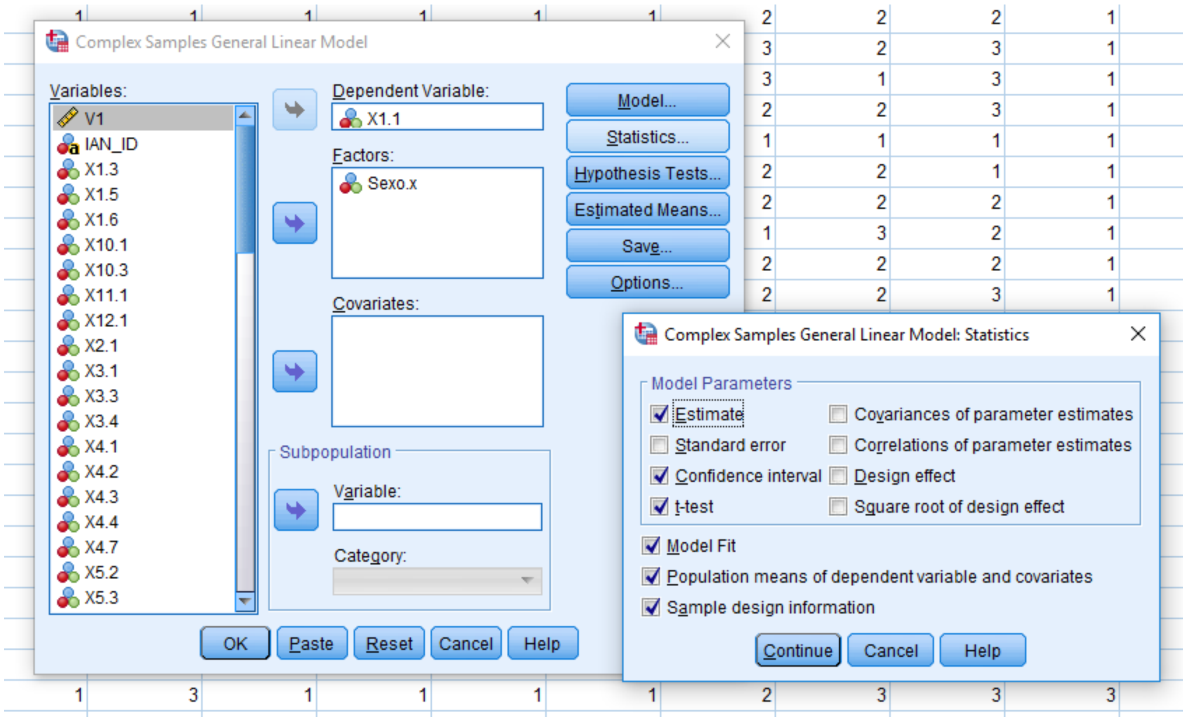
Univariate Statistics

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

1.4. Regressão Linear

Para fazer comparação de médias ponderadas ou regressão linear para os dados ponderados, deve-se aceder a Analyze >> Complex Samples >> General Linear Model e selecionar o ficheiro anteriormente construído.

De seguida, selecionam-se a variáveis dependente e as independentes assim como as estatísticas desejadas.



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

Para obter estimativas ponderadas em R de acordo com o desenho de amostragem complexo IAN-AF 2015-2016, recorre-se à biblioteca “survey” [2,3].

```
> install.packages("survey")
> library(survey)
```

Ao criar a base de dados a usar para realizar estimativas ponderadas é obrigatório ter presente as variáveis “PSU”, “NUT” e a respetiva variável de ponderação, que se encontram na tabela de dados sociodemográficos. Assim, é sempre necessário juntar a base de dados sociodemográficos à base com as variáveis em estudo.

```
> 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)
```

De seguida, exemplifica-se algumas análises possíveis recorrendo a este package. Mais informações sobre funções implementadas nesta biblioteca encontram-se disponíveis na respetiva documentação.

2.1. Frequência de variáveis categóricas e média de variáveis contínuas

O comando “svymean” calcula a média ponderada de uma variável de acordo com o desenho de amostragem complexo. Se a variável em questão for do tipo “factor”, então esta função calcula a proporção ponderada de cada categoria da variável.

```
> 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. Estatísticas em subconjuntos

Para estimar estatísticas em subconjuntos definidos por um fator, usa-se o comando "svyby".

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

É ainda possível definir isoladamente um subconjunto para posterior análise.

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

2.3. Testes de hipóteses

Teste t para comparação de médias:

```
> 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
```

Teste $\chi 2$ para comparação de proporções:

```
> 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. Modelos de regressão

Modelo de regressão linear:

```
> subsvdx = subset(svdx, IMC<700 & EscolClass_Prop!=9)
> m1=svyglm(IMC ~ Sexo + Idade + factor(EscolClass_Prop) , family=gaussian(), subsvdx)
> summary(m1)

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

Survey design:
subset(svdx, IMC < 700 & EscolClass_Prop != 9)

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
```

Modelo de regressão logística:

```
> subsvdx = subset(svdx, Desp!=9)
> m1=svyglm(factor(Desp) ~ factor(GE4), family=binomial(link = 'logit'), subsvdx)
> summary(m1)

Call:
svyglm(formula = factor(Desp) ~ factor(GE4), family = binomial(link = "logit"),
    subsvdx)

Survey design:
subset(svdx, Desp != 9)

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   0.44697    0.14980   2.984   0.00367 **
factor(GE4)2 -0.08235    0.18099  -0.455   0.65023
factor(GE4)3 -0.83873    0.15511  -5.407  5.32e-07 ***
factor(GE4)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
```

ian·af

INQUÉRITO ALIMENTAR NACIONAL
E DE ATIVIDADE FÍSICA