



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

Nota introdutória .....	3
1. Software SPSS .....	4
2. Software R .....	16

## 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-project.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 considerando o processo de amostragem complexa bietápica, a nível nacional e regional, a análise estatística utiliza 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, de forma a obter dados corrigidos para o viés de não-resposta quer da primeira quer da segunda entrevista, criaram-se dois ponderadores, sendo que o primeiro ponderador, **Ponderador1**, utiliza-se para dados recolhidos na primeira entrevista e o segundo, **Ponderador2**, para dados recolhidos na segunda entrevista. **Assim, todas estimativas referentes aos domínios Atividade Física e Estado Nutricional devem ser feitas recorrendo ao Ponderador1, enquanto que o domínio Alimentação deve utilizar o Ponderador2.**

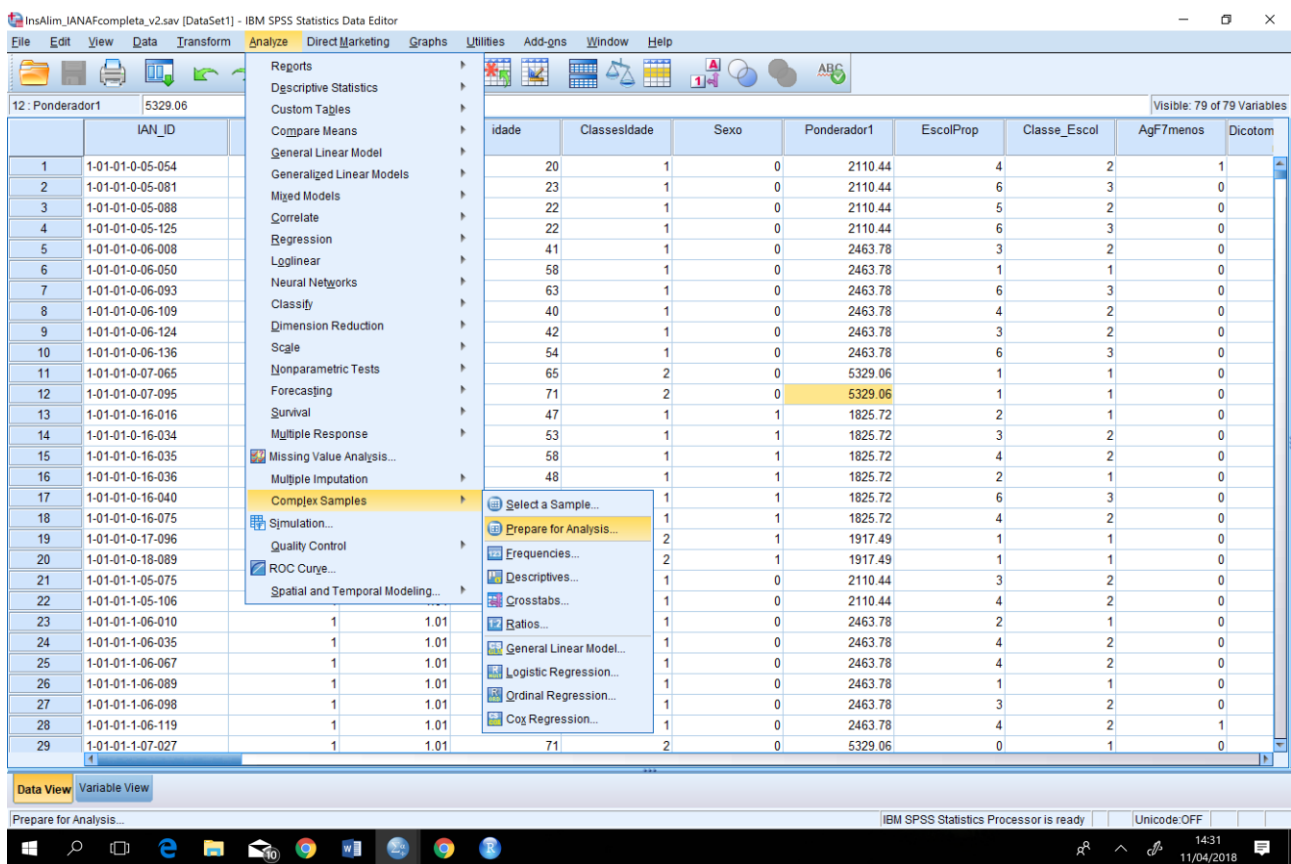
Neste tutorial, exemplifica-se as etapas a seguir de forma a obter estimativas ponderadas de acordo com o desenho de amostragem complexo do IAN-AF 2015-2016, utilizando os softwares SPSS e R [1].

**1.**

# Software

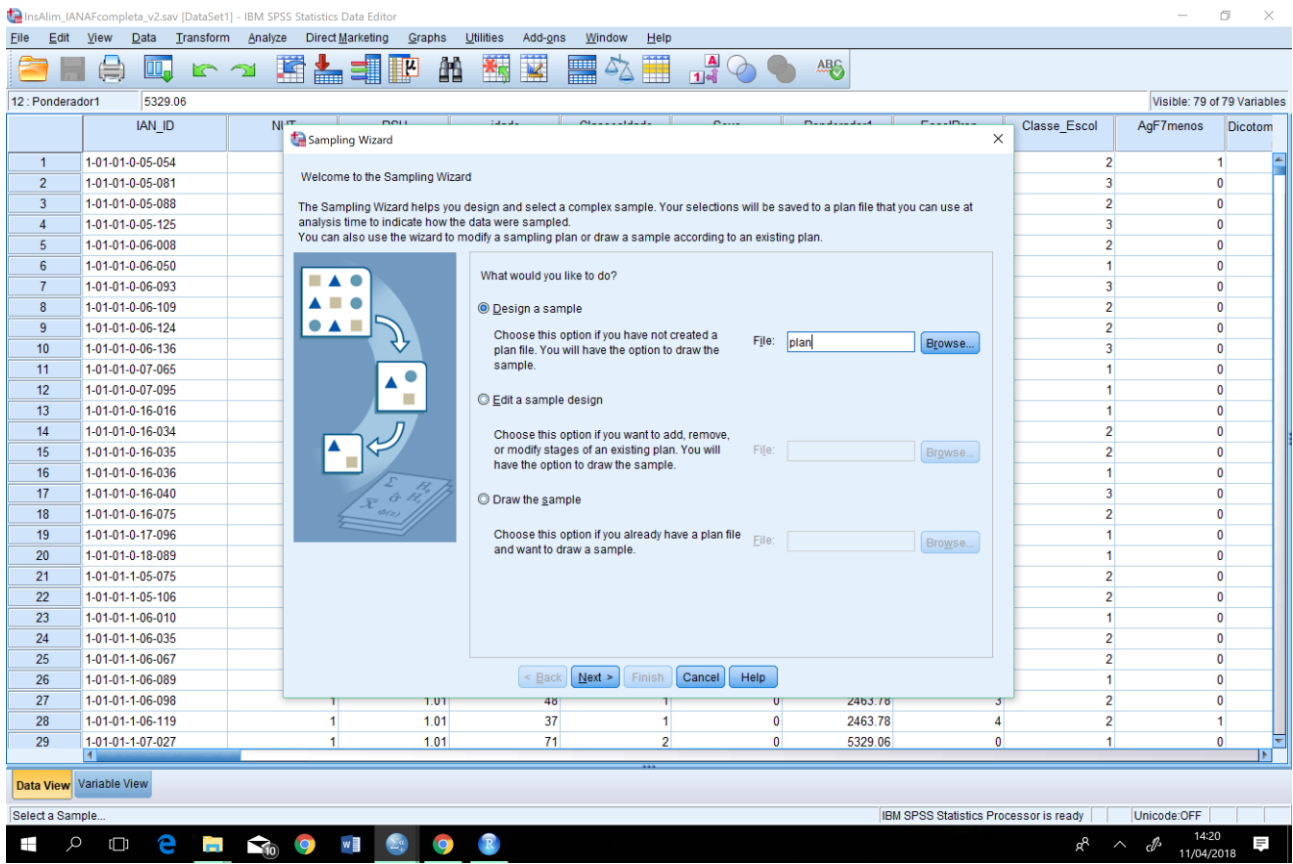
## SPSS

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. Para tal, 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.



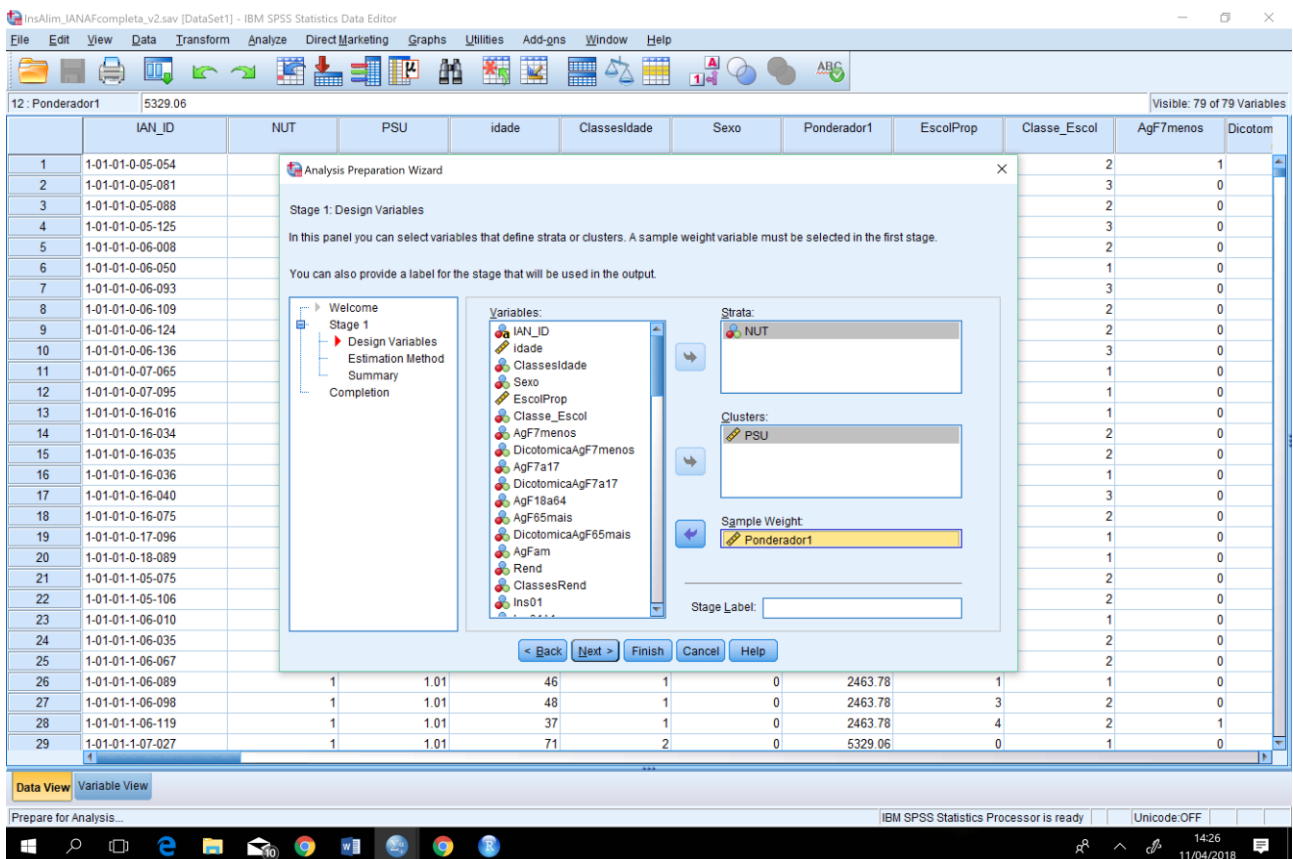
The screenshot shows the IBM SPSS Statistics Data Editor interface. The main window displays a data table with the following columns: idade, Classesidade, Sexo, Ponderador1, EscolProp, Classe\_Escol, AgF7menos, and Dicotom. The 'Analyze' menu is open, and the 'Complex Samples' option is selected. The 'Prepare for Analysis...' sub-option is also visible. The status bar at the bottom indicates 'Prepare for Analysis...' and 'IBM SPSS Statistics Processor is ready'.

idade	Classesidade	Sexo	Ponderador1	EscolProp	Classe_Escol	AgF7menos	Dicotom
20	1	0	2110.44	4	2	1	
23	1	0	2110.44	6	3	0	
22	1	0	2110.44	5	2	0	
22	1	0	2110.44	6	3	0	
41	1	0	2463.78	3	2	0	
58	1	0	2463.78	1	1	0	
63	1	0	2463.78	6	3	0	
40	1	0	2463.78	4	2	0	
42	1	0	2463.78	3	2	0	
54	1	0	2463.78	6	3	0	
65	2	0	5329.06	1	1	0	
71	2	0	5329.06	1	1	0	
47	1	1	1825.72	2	1	0	
53	1	1	1825.72	3	2	0	
58	1	1	1825.72	4	2	0	
48	1	1	1825.72	2	1	0	
1	1	1	1825.72	6	3	0	
1	1	1	1825.72	4	2	0	
2	1	1	1917.49	1	1	0	
2	1	1	1917.49	1	1	0	
1	0	2110.44	3	2	0		
1	0	2110.44	4	2	0		
1	0	2463.78	2	1	0		
1	0	2463.78	4	2	0		
1	0	2463.78	1	1	0		
1	0	2463.78	3	2	0		
1	0	2463.78	4	2	1		
71	2	0	5329.06	0	1	0	



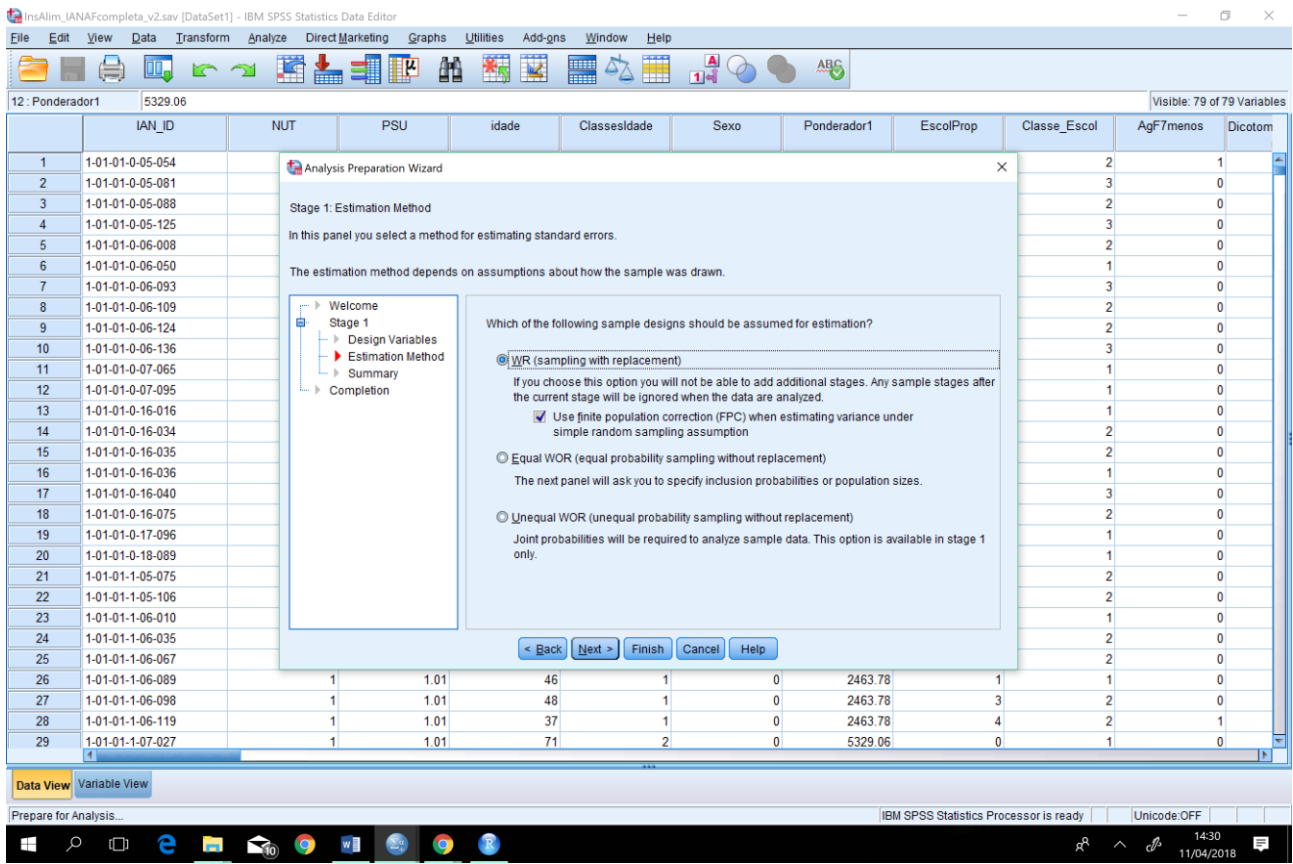
The screenshot shows the SPSS Sampling Wizard dialog box. The 'Design a sample' option is selected. The 'File' field contains 'plan'. The background shows a data table with columns: IAN\_ID, NUT, PSU, idade, Classesidade, Sexo, Ponderador1, EscolProp, Classe\_Escol, AgF7menos, and Dicotom.

IAN_ID	NUT	PSU	idade	Classesidade	Sexo	Ponderador1	EscolProp	Classe_Escol	AgF7menos	Dicotom	
1	1-01-01-0-05-054							2	1		
2	1-01-01-0-05-081							3	0		
3	1-01-01-0-05-088							2	0		
4	1-01-01-0-05-125							3	0		
5	1-01-01-0-06-008							2	0		
6	1-01-01-0-06-050							1	0		
7	1-01-01-0-06-093							3	0		
8	1-01-01-0-06-109							2	0		
9	1-01-01-0-06-124							2	0		
10	1-01-01-0-06-136							3	0		
11	1-01-01-0-07-065							1	0		
12	1-01-01-0-07-095							1	0		
13	1-01-01-0-16-016							1	0		
14	1-01-01-0-16-034							2	0		
15	1-01-01-0-16-035							2	0		
16	1-01-01-0-16-036							1	0		
17	1-01-01-0-16-040							3	0		
18	1-01-01-0-16-075							2	0		
19	1-01-01-0-17-096							1	0		
20	1-01-01-0-18-089							1	0		
21	1-01-01-1-05-075							2	0		
22	1-01-01-1-05-106							2	0		
23	1-01-01-1-06-010							1	0		
24	1-01-01-1-06-035							2	0		
25	1-01-01-1-06-067							2	0		
26	1-01-01-1-06-089							1	0		
27	1-01-01-1-06-098		1	1.01	48	1	0	2463.78	3	2	0
28	1-01-01-1-06-119		1	1.01	37	1	0	2463.78	4	2	1
29	1-01-01-1-07-027		1	1.01	71	2	0	5329.06	0	1	0

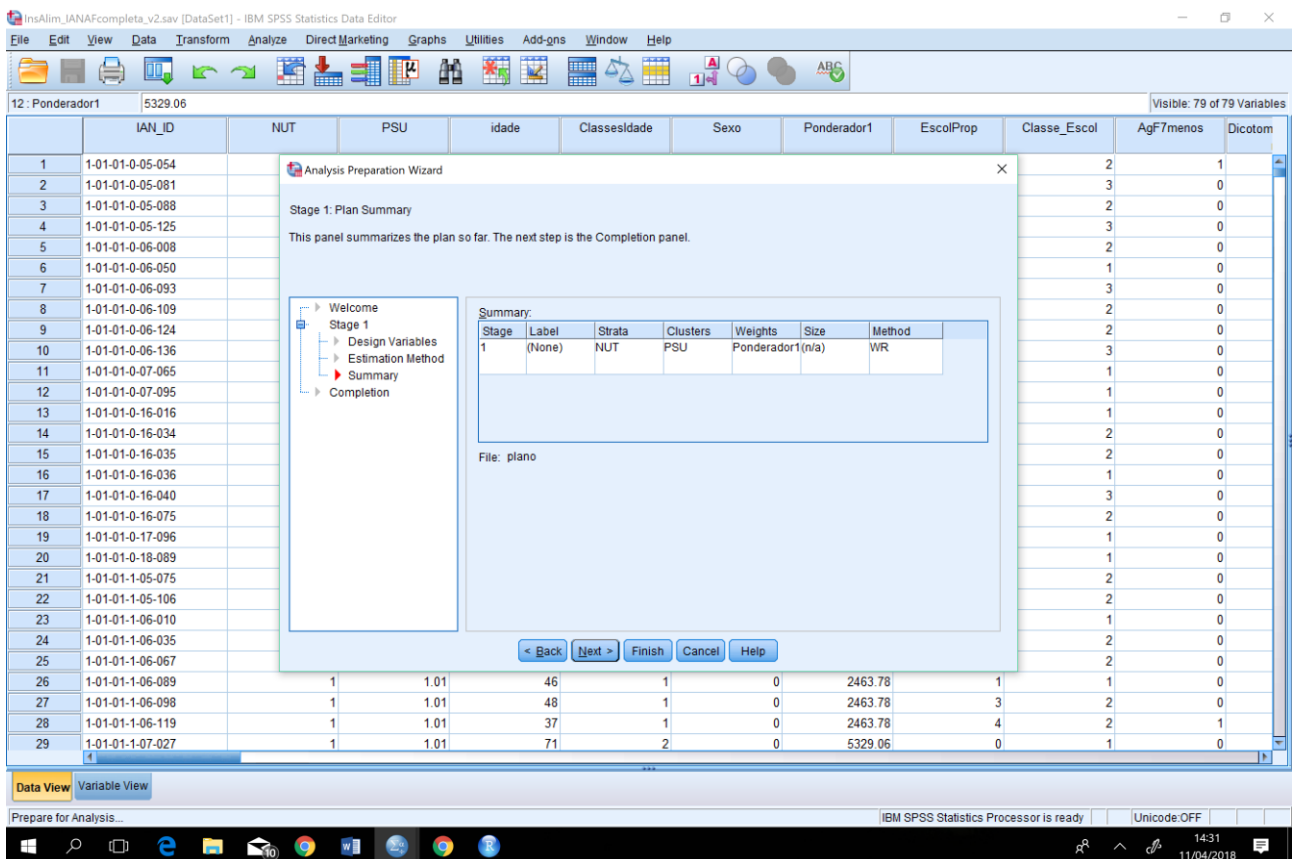


The screenshot shows the SPSS Analysis Preparation Wizard dialog box, Stage 1: Design Variables. The 'Strata' field contains 'NUT', 'Clusters' contains 'PSU', and 'Sample Weight' contains 'Ponderador1'. The background shows the same data table as the previous screenshot.

IAN_ID	NUT	PSU	idade	Classesidade	Sexo	Ponderador1	EscolProp	Classe_Escol	AgF7menos	Dicotom	
1	1-01-01-0-05-054							2	1		
2	1-01-01-0-05-081							3	0		
3	1-01-01-0-05-088							2	0		
4	1-01-01-0-05-125							3	0		
5	1-01-01-0-06-008							2	0		
6	1-01-01-0-06-050							1	0		
7	1-01-01-0-06-093							3	0		
8	1-01-01-0-06-109							2	0		
9	1-01-01-0-06-124							2	0		
10	1-01-01-0-06-136							3	0		
11	1-01-01-0-07-065							1	0		
12	1-01-01-0-07-095							1	0		
13	1-01-01-0-16-016							1	0		
14	1-01-01-0-16-034							2	0		
15	1-01-01-0-16-035							2	0		
16	1-01-01-0-16-036							1	0		
17	1-01-01-0-16-040							3	0		
18	1-01-01-0-16-075							2	0		
19	1-01-01-0-17-096							1	0		
20	1-01-01-0-18-089							1	0		
21	1-01-01-1-05-075							2	0		
22	1-01-01-1-05-106							2	0		
23	1-01-01-1-06-010							1	0		
24	1-01-01-1-06-035							2	0		
25	1-01-01-1-06-067							2	0		
26	1-01-01-1-06-089		1	1.01	46	1	0	2463.78	1	1	0
27	1-01-01-1-06-098		1	1.01	48	1	0	2463.78	3	2	0
28	1-01-01-1-06-119		1	1.01	37	1	0	2463.78	4	2	1
29	1-01-01-1-07-027		1	1.01	71	2	0	5329.06	0	1	0



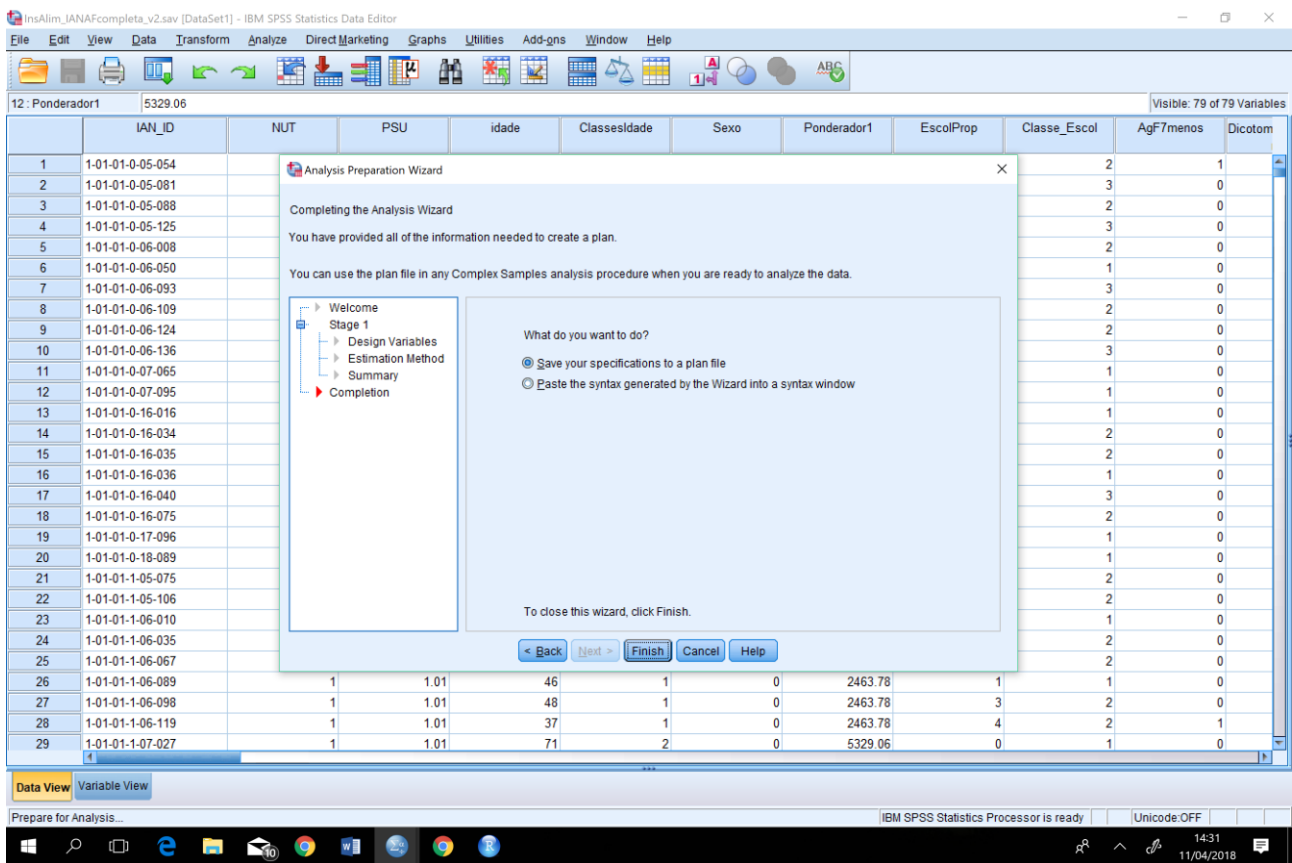
The screenshot shows the IBM SPSS Statistics Data Editor with the 'Analysis Preparation Wizard' dialog box open. The wizard is at 'Stage 1: Estimation Method'. The background data table has columns: IAN\_ID, NUT, PSU, idade, Classesidade, Sexo, Ponderador1, EscolProp, Classe\_Escol, AgF7menos, and Dicotom. The wizard dialog has a tree view on the left with 'Estimation Method' selected. The main text asks 'Which of the following sample designs should be assumed for estimation?' and lists three options: 'WR (sampling with replacement)', 'Equal WOR (equal probability sampling without replacement)', and 'Unequal WOR (unequal probability sampling without replacement)'. The 'WR' option is selected. Below the options, there are checkboxes for 'Use finite population correction (FPC) when estimating variance under simple random sampling assumption' and 'Joint probabilities will be required to analyze sample data. This option is available in stage 1 only.' Buttons for '< Back', 'Next >', 'Finish', 'Cancel', and 'Help' are at the bottom.



The screenshot shows the IBM SPSS Statistics Data Editor with the 'Analysis Preparation Wizard' dialog box open. The wizard is at 'Stage 1: Plan Summary'. The background data table is the same as in the previous screenshot. The wizard dialog has a tree view on the left with 'Summary' selected. The main text says 'This panel summarizes the plan so far. The next step is the Completion panel.' Below this is a 'Summary' table:

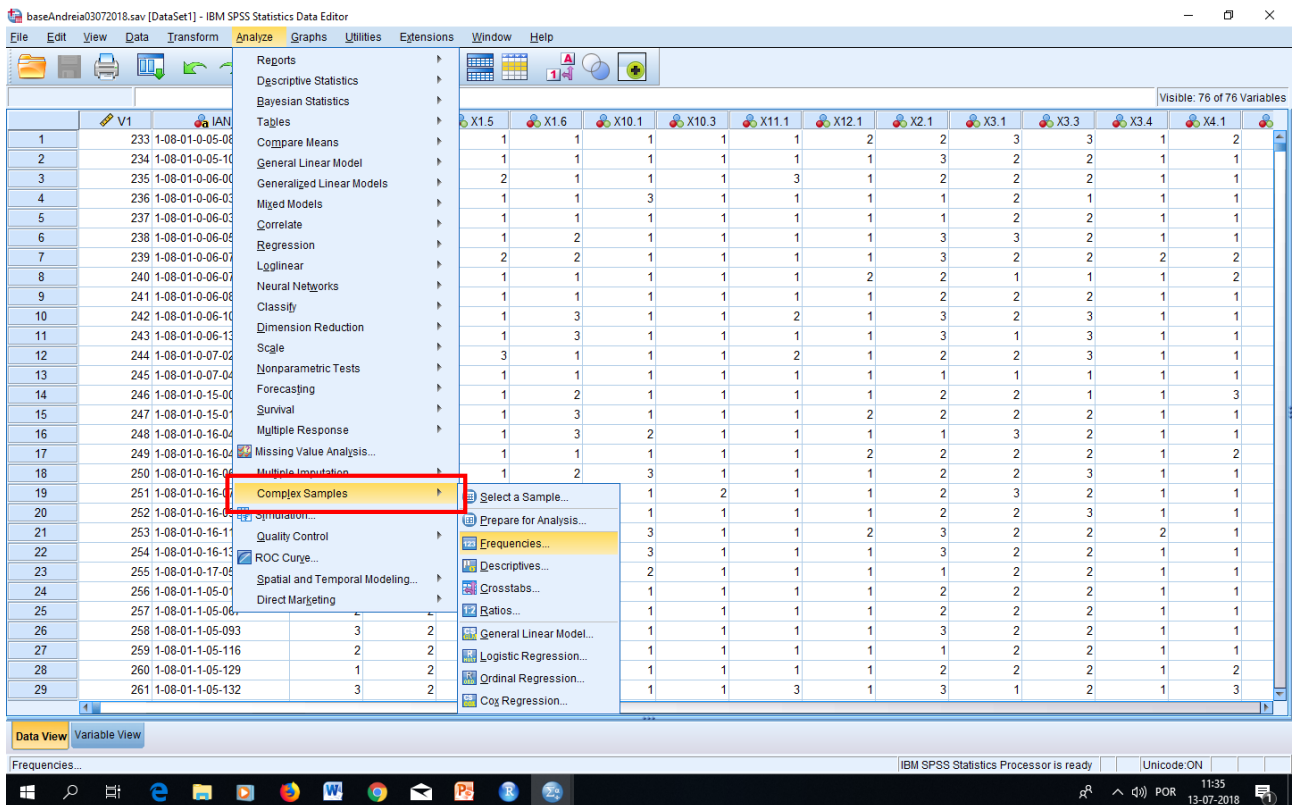
Stage	Label	Strata	Clusters	Weights	Size	Method
1	(None)	NUT	PSU	Ponderador1(n/a)		WR

Below the table, it says 'File: plano'. Buttons for '< Back', 'Next >', 'Finish', 'Cancel', and 'Help' are at the bottom.



The screenshot shows the IBM SPSS Statistics Data Editor with a data table and an open 'Analysis Preparation Wizard' dialog box. The data table has columns: IAN\_ID, NUT, PSU, idade, Classesidade, Sexo, Ponderador1, EscolProp, Classe\_Escol, AgF7menos, and Dicotom. The wizard is at the 'Completion' stage, asking 'What do you want to do?' with two options: 'Save your specifications to a plan file' (selected) and 'Paste the syntax generated by the Wizard into a syntax window'. The 'Finish' button is highlighted.

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

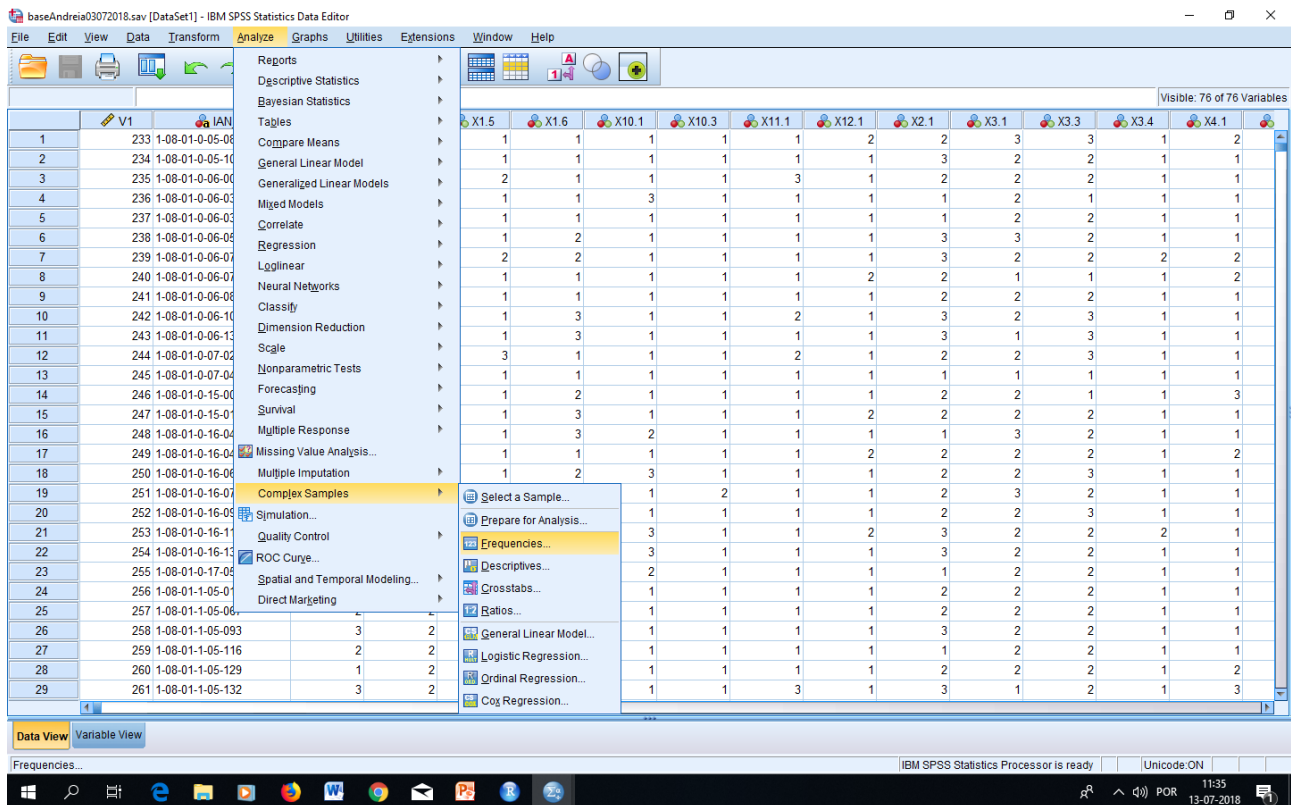


The screenshot shows the IBM SPSS Statistics Data Editor with the 'Analyze' menu open. The 'Complex Samples' option is highlighted with a red box. The data table in the background has columns: V1, IAN, X1.5, X1.6, X1.0, X1.0.3, X1.1, X1.2, X2.1, X3.1, X3.3, X3.4, X4.1. The 'Complex Samples' menu is open, showing options like 'Select a Sample...', 'Prepare for Analysis...', 'Erequencies...', 'Descriptives...', 'Crosstabs...', 'Rabios...', 'General Linear Model...', 'Logistic Regression...', 'Ordinal Regression...', and 'Cog Regression...'.

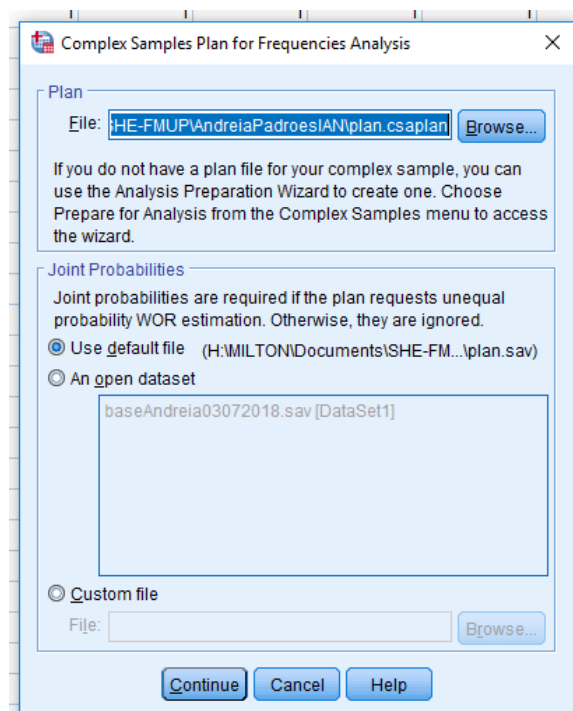


## 1.1. Estimar frequências ponderadas

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

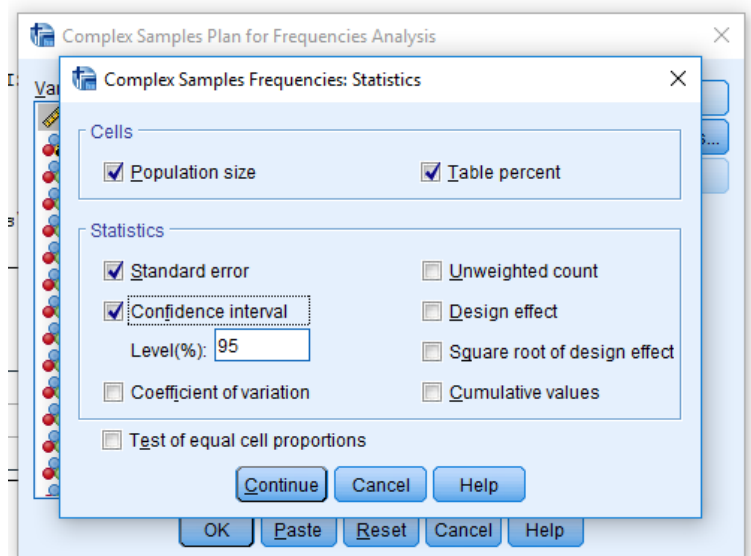
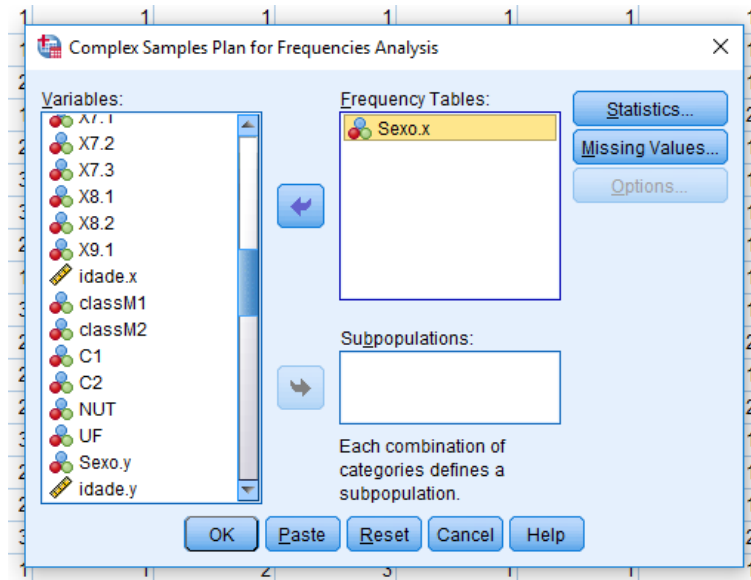


The screenshot shows the IBM SPSS Statistics Data Editor interface. The 'Analyze' menu is open, and the path 'Complex Samples' > 'Frequencies...' is highlighted. The background shows a list of variables (V1 to V29) and a data table with columns X1.5 to X4.1.



The dialog box 'Complex Samples Plan for Frequencies Analysis' is shown. It has a 'Plan' section with a 'File:' field containing 'H:\FMUP\AndreiaPadroes\IAN\plan.csaplan' and a 'Browse...' button. Below this is a text box explaining that if no plan file is present, the Analysis Preparation Wizard should be used. The 'Joint Probabilities' section has two radio buttons: 'Use default file (H:\MILTON\Documents\SHE-FM...\plan.sav)' which is selected, and 'An open dataset' with a text box containing 'baseAndreia03072018.sav [DataSet1]'. At the bottom, there are 'Continue', 'Cancel', and 'Help' buttons.

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



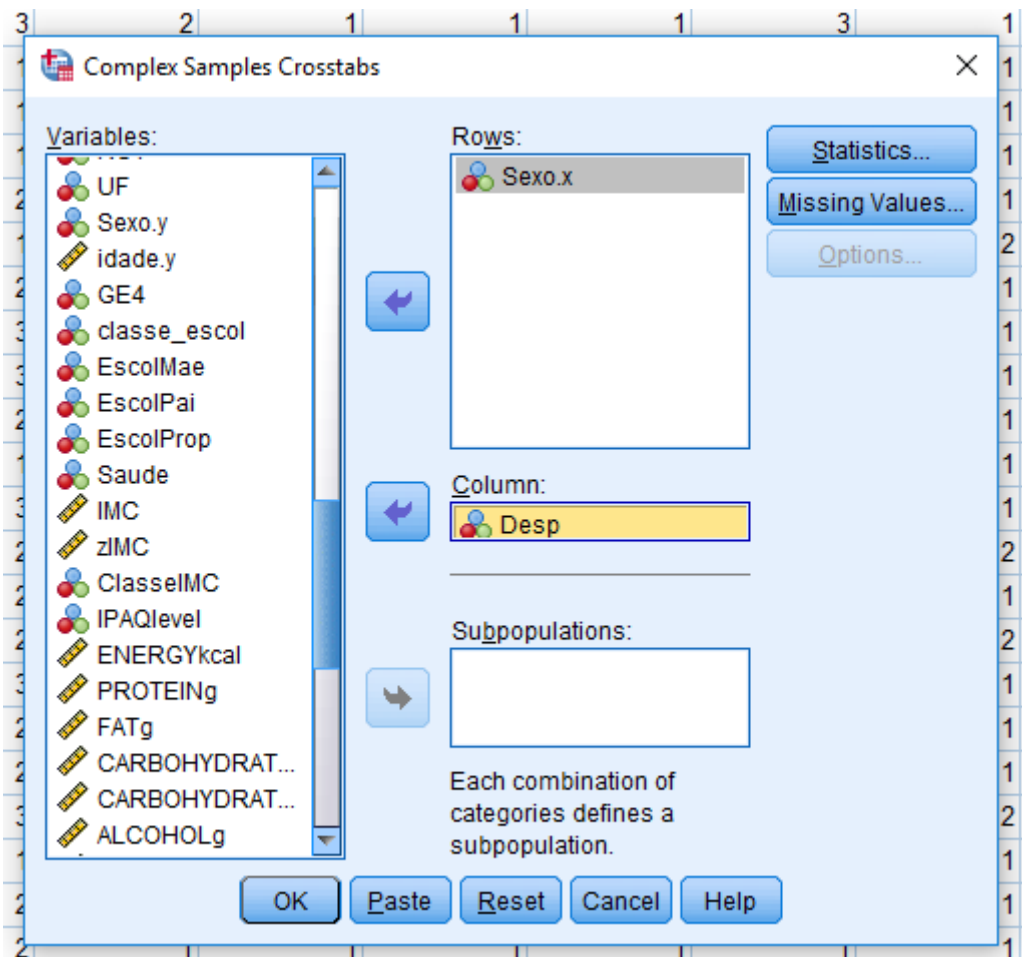
Resultado:

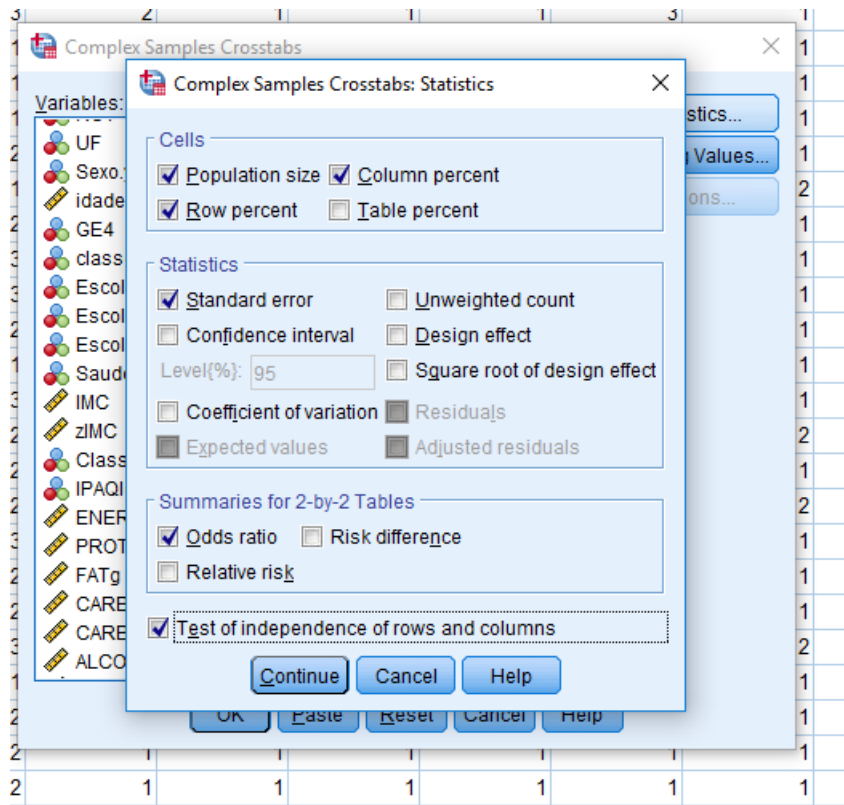
Sexo.x					
		Estimate	Standard Error	95% Confidence Interval	
				Lower	Upper
Population Size	0	4739432,770	145329,479	4450795,879	5028069,661
	1	4449227,520	126039,458	4198902,276	4699552,764
	Total	9188660,290	239273,706	8713442,056	9663878,524
% of Total	0	51,6%	0,7%	50,2%	53,0%
	1	48,4%	0,7%	47,0%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





Resultado:

Sexo.x \* Desp

Sexo.x		Desp			
		0	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

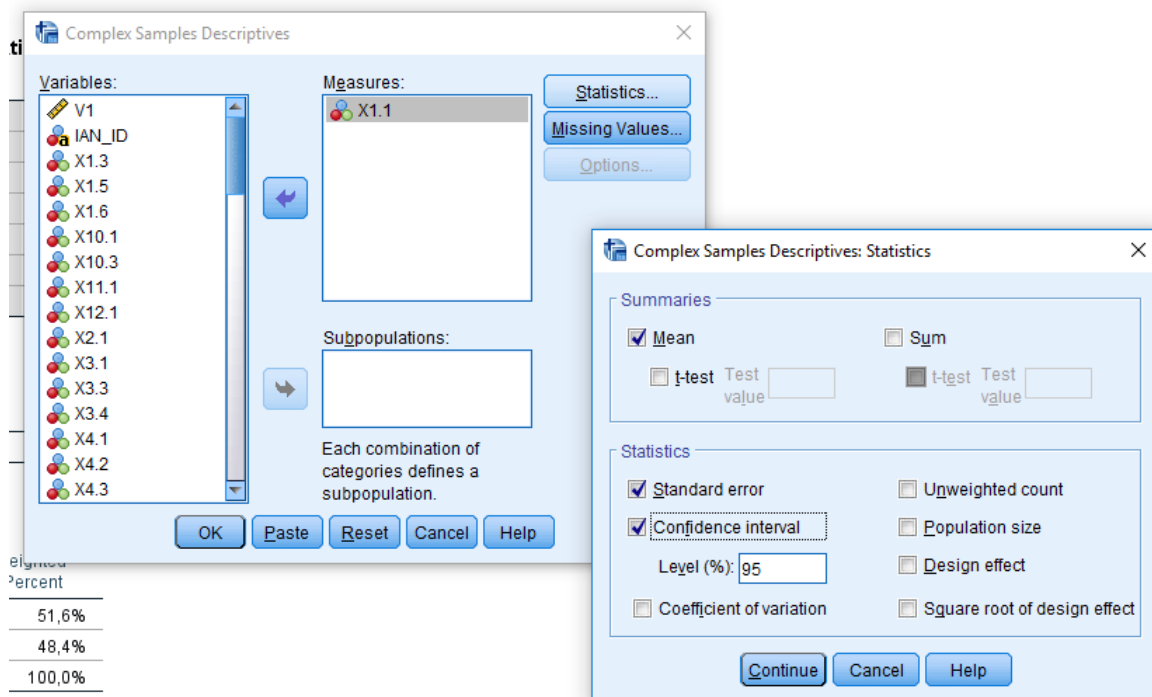
		Estimate
Sexo.x * Desp	Odds Ratio	1,286

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

### 1.3. Estimar média ponderada

Para estimar a média ponderada e o respetivo intervalo de confiança de uma variável contínua, 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.



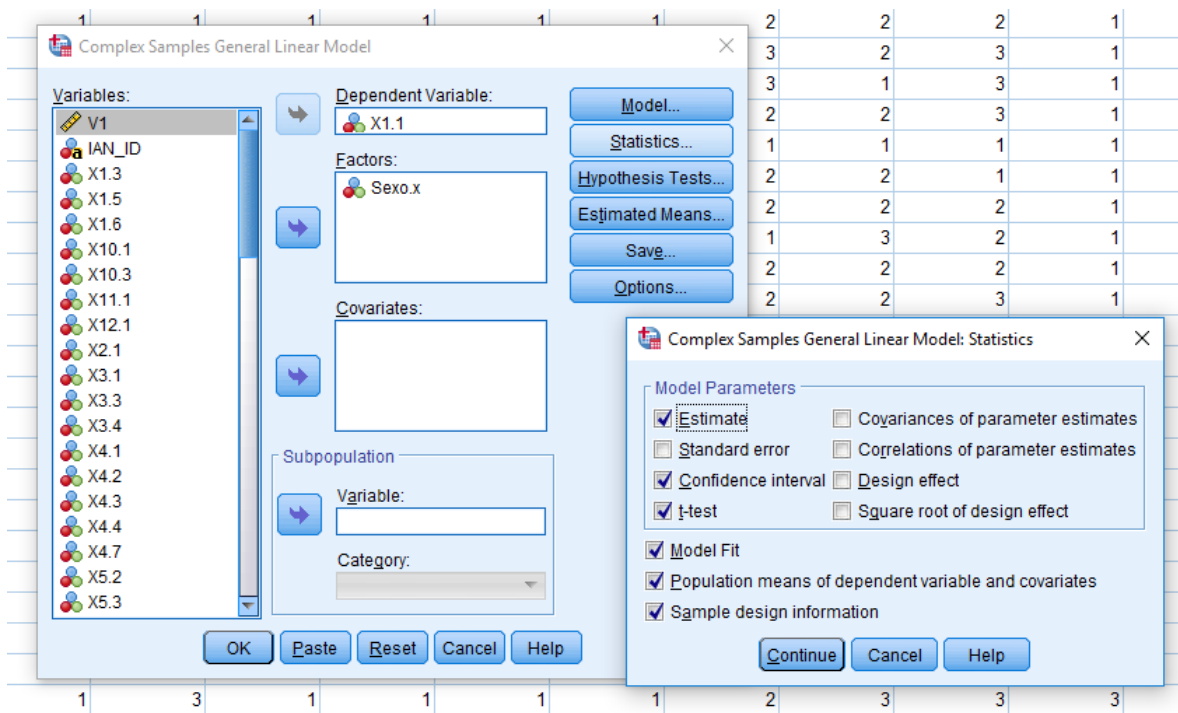
Resultado:

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. Se a variável selecionada for do tipo categórica, deve ser adicionada em Factors, caso contrário, se for do tipo contínua, deve ser adicionada em Covariates.



Resultado:

Parameter Estimates<sup>a</sup>

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 <sup>b</sup>	.	.	.	.	.

a. Model:  $X1.1 = (\text{Intercept}) + \text{Sexo.x}$

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.

```
# 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)
```

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, svdx, 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:

```
> svytest(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  $\chi^2$ : 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:

```
> 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

Modelo de regressão logística:

```
> 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
```



INQUÉRITO ALIMENTAR NACIONAL  
E DE ATIVIDADE FÍSICA

