WEIGHTING COMPLEX SAMPLES
IAN-AF Databases
Tutorial using software SPSS e R
Table of Contents

Nota introdutória ........................................................................................................3
1. Software SPSS ........................................................................................................4
2. Software R .............................................................................................................17

References


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.
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:

<table>
<thead>
<tr>
<th>Sexo.x</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Interval Lower</th>
<th>95% Confidence Interval Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population Size</td>
<td>0</td>
<td>4739432,770</td>
<td>145329,479</td>
<td>4450795,879</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4449227,520</td>
<td>126039,458</td>
<td>4198902,276</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>9188660,290</td>
<td>239273,706</td>
<td>8713442,056</td>
</tr>
<tr>
<td>% of Total</td>
<td>0</td>
<td>51,6%</td>
<td>0,7%</td>
<td>50,2%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>48,4%</td>
<td>0,7%</td>
<td>47,0%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>100,0%</td>
<td>0,0%</td>
<td>100,0%</td>
</tr>
</tbody>
</table>
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**

<table>
<thead>
<tr>
<th></th>
<th>Sexo.x</th>
<th></th>
<th>Desp</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Standard Error</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Population Size</strong></td>
<td>2916200,750</td>
<td>119981,932</td>
<td>1689662,870</td>
<td>104059,923</td>
<td>4605863,620</td>
</tr>
<tr>
<td>% within Sexo.x</td>
<td>63,3%</td>
<td>1,9%</td>
<td>36,7%</td>
<td>1,9%</td>
<td>100,0%</td>
</tr>
<tr>
<td>% within Desp</td>
<td>53,4%</td>
<td>1,3%</td>
<td>47,1%</td>
<td>1,7%</td>
<td>50,9%</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>Standard Error</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Population Size</strong></td>
<td>2547897,160</td>
<td>109990,959</td>
<td>1899139,430</td>
<td>108317,206</td>
<td>4447036,590</td>
</tr>
<tr>
<td>% within Sexo.x</td>
<td>57,3%</td>
<td>2,0%</td>
<td>42,7%</td>
<td>2,0%</td>
<td>100,0%</td>
</tr>
<tr>
<td>% within Desp</td>
<td>46,6%</td>
<td>1,3%</td>
<td>52,9%</td>
<td>1,7%</td>
<td>49,1%</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>Standard Error</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Population Size</strong></td>
<td>5464097,910</td>
<td>183758,461</td>
<td>3588802,300</td>
<td>173125,807</td>
<td>9052900,210</td>
</tr>
<tr>
<td>% within Sexo.x</td>
<td>60,4%</td>
<td>1,5%</td>
<td>39,6%</td>
<td>1,5%</td>
<td>100,0%</td>
</tr>
<tr>
<td>% within Desp</td>
<td>100,0%</td>
<td>0,0%</td>
<td>100,0%</td>
<td>0,0%</td>
<td>100,0%</td>
</tr>
<tr>
<td></td>
<td>Standard Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% within Sexo.x</td>
<td>0,0%</td>
<td>0,0%</td>
<td>0,0%</td>
<td>0,0%</td>
<td>0,0%</td>
</tr>
<tr>
<td>% within Desp</td>
<td>0,0%</td>
<td>0,0%</td>
<td>0,0%</td>
<td>0,0%</td>
<td>0,0%</td>
</tr>
</tbody>
</table>
Tests of Independence

<table>
<thead>
<tr>
<th>Sexo.x * Desp</th>
<th>Chi-Square</th>
<th>Adjusted F</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>14,388</td>
<td>6,020</td>
<td>1</td>
<td>92</td>
<td>.016</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>14,394</td>
<td>6,022</td>
<td>1</td>
<td>92</td>
<td>.016</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Sexo.x * Desp</th>
<th>Odds Ratio</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds Ratio</td>
<td>1,286</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>Univariate Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
</tr>
<tr>
<td>Mean</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>95% Confidence Interval</th>
<th>Hypothesis Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2,129</td>
<td>2,056 - 2,203</td>
<td>57,592 92,000 ,000</td>
</tr>
<tr>
<td>[Sexo.x=0]</td>
<td>0,020</td>
<td>-0,068 - 0,108</td>
<td>456 92,000 ,649</td>
</tr>
<tr>
<td>[Sexo.x=1]</td>
<td>0,000b</td>
<td>- -</td>
<td>- - - -</td>
</tr>
</tbody>
</table>

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].

```r
> 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.

```r
# 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.

```r
> 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.

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

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

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

**t-test for comparison of mean values:**

```r
> 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
```

**χ²-test to comparison of proportions:**

```r
> 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:

```r
> 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
```

```r
> 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  0.19892308
factor(EscolClass_Prop)2 -1.39991563 -1.93349039 -0.86634087
factor(EscolClass_Prop)3 -2.05718129 -2.58605546 -1.52830711
```
Logistic regression model:

```r
> 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
```

```r
> 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
```