Análise de Dados Bivariada

3. Análise de dados bivariada

Nota: A base de dados que vamos utilizar neste capítulo é um extracto da base de dados do *European Social Survey* (*round* 1) com os resultados da aplicação do questionário em Portugal, modificada com os exercícios do capítulo 1 (ficheiro **ESS Portugal 2002 (TPAUB2).sav**)

3.1. Cruzamentos e teste de independência χ^2 (Qui-quadrado)

O teste do χ^2 de independência "serve para testar se duas ou mais populações (ou grupos) independentes diferem relativamente a uma determinada característica, i.e. se a frequência com que os elementos da amostra se repartem pelas classes de uma variável nominal categorizada é ou não idêntica"[11]. Admite como nível mínimo de mediada o nível nominal.

Exemplo: Pretende-se saber se há relação entre o sexo e o facto de ter comprado produtos por razões de ordem política, ética ou ambiental.

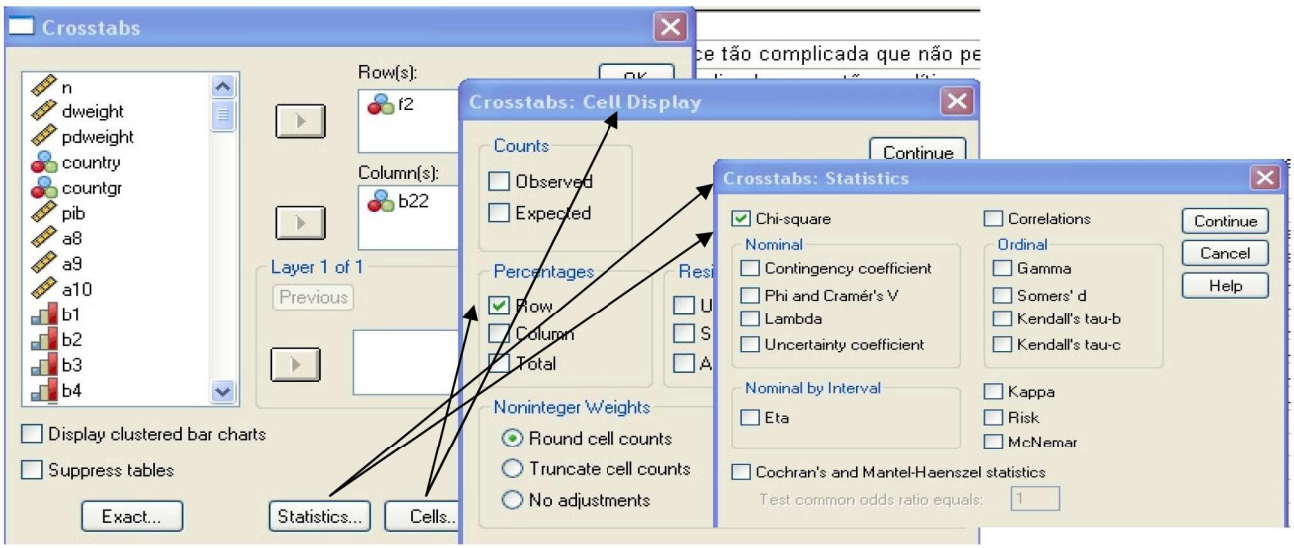
O procedimento consiste em cruzar as variáveis sexo (***f2***) e (***b22***) e solicitar o teste de independência do χ^{2*}.

Hipótese do teste (bilateral):

H_o: A compra de produtos por razões de ordem política, ética ou ambiental, é igual entre homens e mulheres

H_a: A compra de produtos por razões de ordem política, ética ou ambiental, é diferente entre homens e mulheres

3.1.1. Utilizando o comando *Crosstabs*



[11] Maroco, idem: 103.

[*] Para a selecção dos testes estatísticos, ver o Anexo 1.

O resultado é o seguinte:

Sexo * Comprou produtos por razões de ordem política, ética ou ambiental Crosstabulation

			Sim	Não	Total
Sexo	Masculino	Count	45	579	624
		Row %	7,2%	92,8%	100,0%
	Feminino	Count	67	804	871
		Row %	7,7%	92,3%	100,0%
Total		Count	112	1383	1495
		Row %	7,5%	92,5%	100,0%

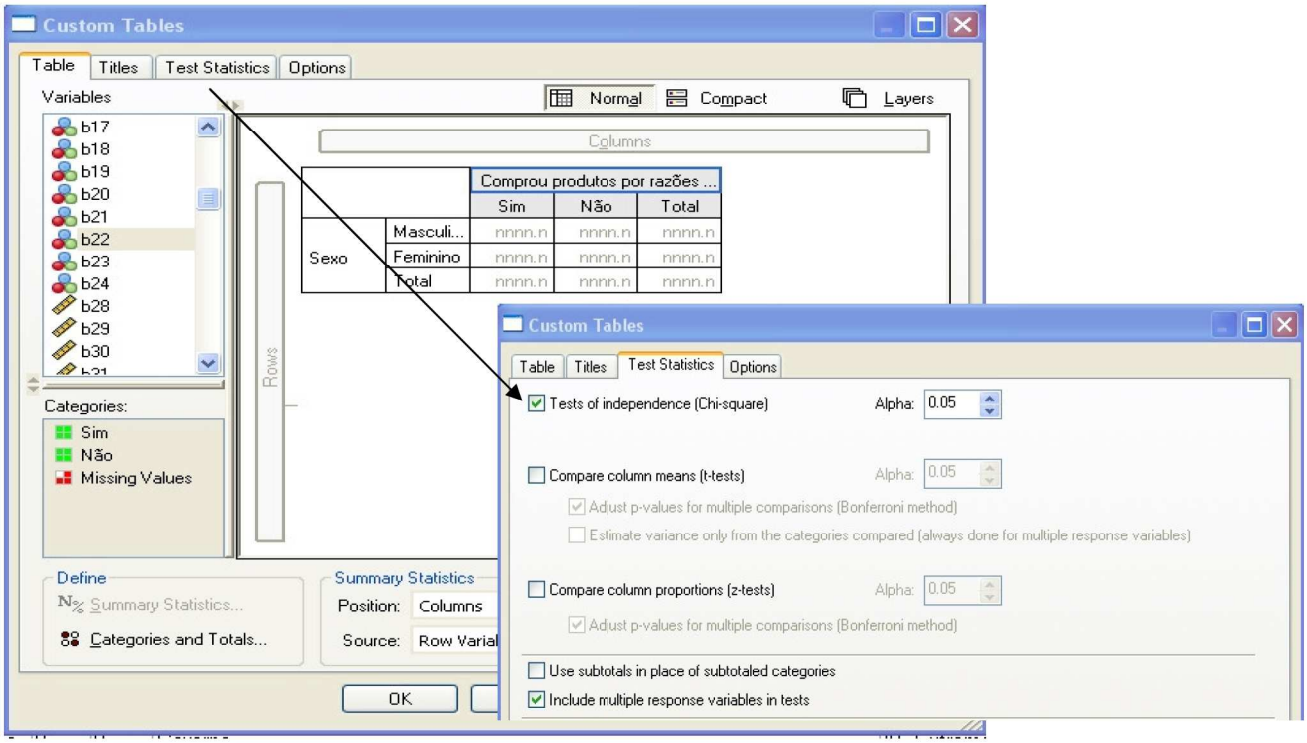
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,121[b]	1	,728		
Continuity Correctiõ	,062	1	,804		
Likelihood Ratio	,122	1	,727		
Fisher's Exact Test				,766	,404
Linear-by-Linear Association	,121	1	,728		
N of Valid Cases	1495				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expect count is 46,75.

3.1.2. Utilizando o comando *Tables*



O resultado é o seguinte:

Sexo	Comprou produtos por razões de ordem política, ética ou ambiental			
	Sim		Não	
	N	%	N	%
Masculino	45	7,2	579	92,8
Feminino	67	7,7	804	92,3
Total	**112**	**7,5**	**1383**	**92,5**

Pearson Chi-Square Tests

		Comprou produtos por razões de ordem política, ética ou ambiental
Sexo	Chi-square	,121
	df	1
	Sig.	,728

Results are based on nonempty rows and columns in each innermost subtable.

Interpretação: Não se rejeita-se a hipótese nula (χ^2 (1)=0,121; p>0,05). Embora as mulheres comprem ligeiramente mais que os homens, as diferenças entre homens e mulheres não são estatisticamente significativas.

3.2. Testes não paramétricos (procedimento *Nonparametric Tests*)

Os testes não paramétricos não estão condicionados por qualquer distribuição de probabilidades dos dados em análise, ou seja, não estão sujeitos aos condicionalismos da verificação dos pressupostos, como acontece nos testes paramétricos, e constituem uma alternativa à sua utilização, quando aqueles não se verificam e a sua violação é "grave".

3.2.1. Duas amostras independentes: Teste de *Mann-Whitney*

"O teste de Wilcoxon-Mann-Whitney ou simplesmente teste de Mann-Whitney, é o teste não-paramétrico adequado para comparar as funções de distribuição de uma variável pelo menos ordinal medida em duas amostras independentes"[12].

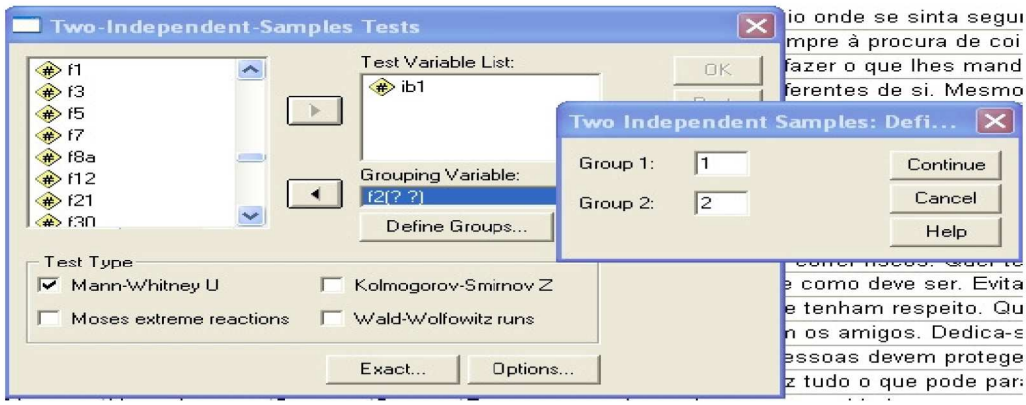
Exemplo: Pretende-se testar se há relação entre o sexo (***f2***) e o interesse pela política (***ib1***).

O procedimento consiste na realização do teste não paramétrico para 2 amostras independentes (***Mann-Whitney***)[13].

Hipótese do teste (bilateral):

H_o: Homens e mulheres têm o mesmo interesse pela política

H_a: Homens e mulheres não têm o mesmo interesse pela política



O resultado é o seguinte:

Ranks

	Sexo	N	Mean Rank	Sum of Ranks
Qual o seu interesse pela política	Masculino	628	815.82	512336.00
	Feminino	875	706.19	617920.00
	Total	1503		

Test Statistics [a]

	Qual o seu interesse pela política
Mann-Whitney U	234670.000
Wilcoxon W	617920.000
Z	-5.058
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: Sexo

Interpretação: Rejeita-se a hipótese nula. Os homens têm mais interesse pela política do que as mulheres. A média das ordenações (*Mean Rank*) é superior nos homens[14] e as diferenças são estatisticamente significativas (*M-W*=234670; *p*=0,000).

[12] Marôco, idem: 219. Este teste pode também ser utilizado como alternativa ao teste *t-Student* para amostras independentes, nomeadamente quando os pressupostos deste teste não são válidos.
[13] Consultar o Anexo 1.
[14] A escala é crescente.

3.2.2. K amostras independentes (*Kruskal-Wallis*)

"O teste de Kruskal-Walis pode ser considerado como a alternativa não-paramétrica à ANOVA *one-way* (Kruskal & Wallis, 1952). Este teste pode ser então usado para testar se duas ou mais amostras provém de uma mesma população ou se de populações diferentes ou se, de igual modo, as amostras provém de populações com a mesma distribuição"[15].

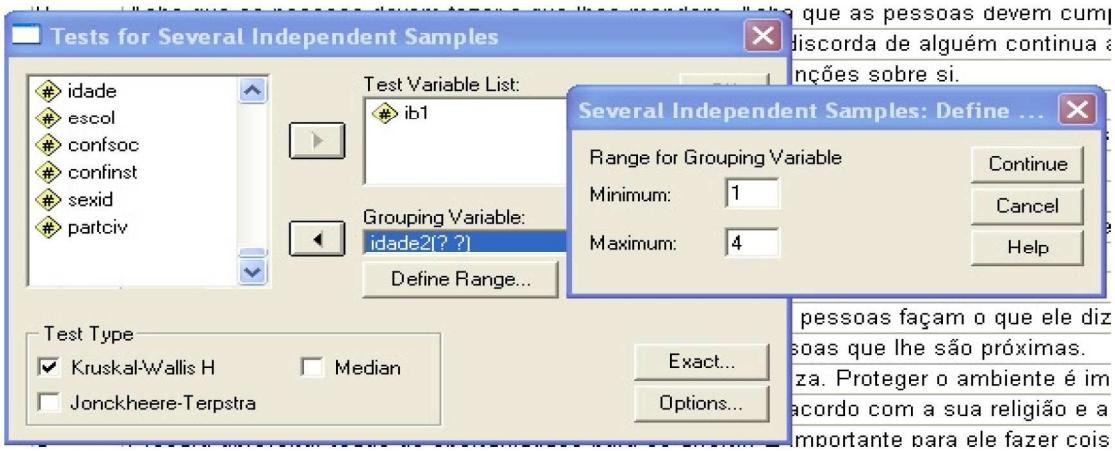
Exemplo: Pretende-se testar se há relação entre a idade (***idade2***) e o interesse pela política (***ib1***).

O procedimento consiste na realização do teste não paramétrico para k amostras independentes (***Kruskal-Wallis***)[16].

Hipótese do teste (bilateral):

H_o : O interesse pela política tem igual distribuição nos diversos escalões etários

H_a: O interesse pela política não tem igual distribuição nos diversos escalões etários



O resultado é o seguinte:

		Qual o seu interesse pela política				
		Nenhum interesse	Pouco interesse	Algum interesse	Muito interesse	Total
Idade	Até 30 anos	27.4	33.6	31.9	7.1	100.0
	31 - 50 anos	28.0	28.8	32.9	10.3	100.0
	51 - 65 anos	30.7	30.4	29.7	9.3	100.0
	> 65 anos	45.8	25.1	25.6	3.5	100.0
	Total	32.5	29.3	30.3	7.8	100.0

Ranks

	Idade	N	Mean Rank
Qual o seu interesse pela política	Até 30 anos	339	777.35
	31 - 50 anos	504	803.46
	51 - 65 anos	313	768.74
	> 65 anos	347	637.38
	Total	1503	

Test Statistics [a,b]

	Qual o seu interesse pela política
Chi-Square	36.088
df	3
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: Idade

Interpretação: Rejeita-se a hipótese nula. Os indivíduos entre 31 a 50 anos são os que revela maior interesse pela política (*mean rank*=803,046) e os mais velhos são os que revelam menos interesse (*mean rank*=637,38). As diferenças são estatisticamente significativas (*K-W* (3)=36,088; *p*=0,000).

[15] Maroco, idem: 227.
[16] Consultar o Anexo 1.

3.3. Testes paramétricos

Os Testes paramétricos são baseados em parâmetros da amostra, por exemplo média e desvio padrão, assumindo a normalidade e a homogeneidade das distribuições.

3.3.1. Duas amostras emparelhadas (*t Student* para amostras emparelhadas)

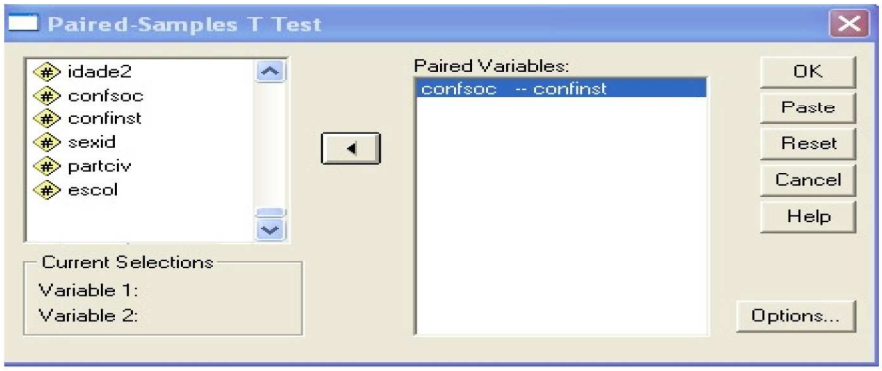
"O teste *t-Student* para amostras emparelhadas é o teste a utilizar quando se pretende comparar duas populações de onde foram extraídas duas amostras emparelhadas relativamente a uma variável dependente quantitativa"[17].

Exemplo: Pretende-se testar se a média da confiança social (***confsoc***) é idêntica, ou não, à média da confiança institucional (***confinst***).
O procedimento consiste na realização do teste paramétrico para duas amostras emparelhadas (***Paired-Samples T-Test***)[18].

Hipótese do teste (bilateral):

H_o: A média da confiança social é igual à média da confiança institucional

H_a: A média da confiança social não é igual à média da confiança institucional



O resultado é o seguinte:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Índice sintético de Confiança social	4,313	1319	1,7165	,0473
	Índice sintético de Confiança institucional	4,087	1319	1,7606	,0485

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Índice sintético de Confiança social & Índice sintético de Confiança institucional	1319	,310	,000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Índice sintético de Confiança social - Índice sintético de Confiança institucional	,2255	2,0428	,0562	,1151	,3358	4,009	1318	,000

Interpretação: Rejeita-se a hipótese nula (t (1318)=4,009; p=0,000). (A média da confiança social (4,313) é ligeiramente superior à média da confiança institucional (4,087). A correlação entre as duas variáveis é média fraca (0,310) e estatisticamente significativa (p=0,000).

[17] Maroco, idem: 270. Ao contrário do teste *t-Student* para amostras independentes, no caso de apenas duas amostras emparelhadas, não é necessário assegurar a homocedasticidade das variâncias.
[18] Consultar o Anexo 1.

3.3.2. Duas amostras independentes (*t Student* de independência)

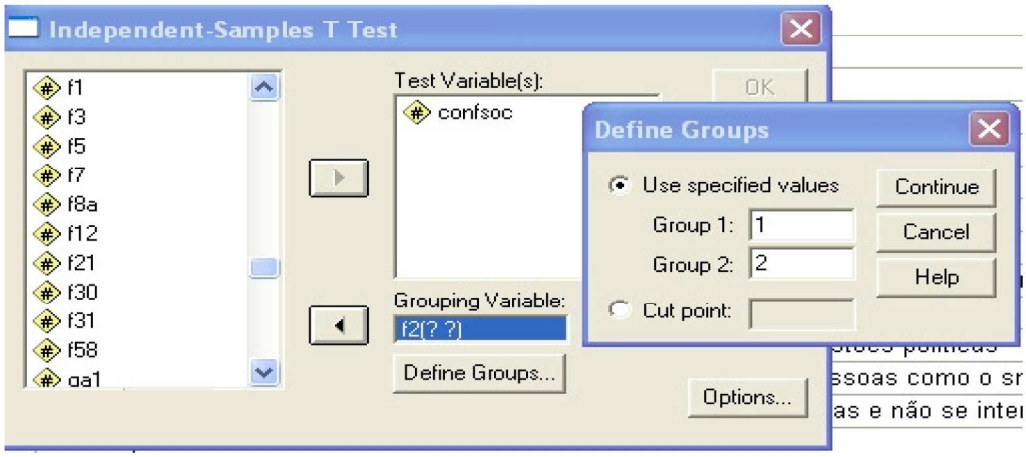
"O teste *t-Student* serve também para testar se as médias de duas populações são ou não significativamente diferentes. Este teste requer que as duas amostras tenham sido obtidas aleatoriamente de duas populações e que as variáveis dependentes possuam distribuição normal com variâncias homogéneas"[19].

Exemplo: Pretende-se testar se há relação entre o sexo (***f2***) e a confiança social (***confsoc***).
O procedimento consiste na realização do teste paramétrico para duas amostras independentes (***Independent-Samples T-Test***)[20].

Hipótese do teste (bilateral):

H_o: A média da confiança social é igual entre homens e mulheres

H_a: A média da confiança social é diferente entre homens e mulheres



O resultado é o seguinte:

Group Statistics

	Sexo	N	Mean	Std. Deviation	Std. Error Mean
Indice sintético de Confiança social	Masculino	612	4.412	1.7476	.0706
	Feminino	868	4.248	1.7442	.0592

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Indice sintético de Confiança social	Equal variances assumed	.182	.669	1.781	1478	.075	.164	.0921	-.0167	.3448
	Equal variances not assumed			1.780	1313.926	.075	.164	.0922	-.0167	.3449

Interpretação: Não se rejeita a hipótese nula. Os homens (4,412) revelam mais confiança social que as mulheres (4,248)[21], mas a diferença não é estatisticamente significativa ($t\,(1478)=1,781$; $p>0,05$).

[19] Marôco, idem: 147-148.
[20] Consultar o Anexo 1.
[21] O índice de confiança social varia entre 0=nenhuma confiança e 10=toda a confiança.

3.3.3. k amostras independentes (Análise de Variância simples paramétrica - ANOVA)

"A comparação de médias de duas ou mais populações de onde foram extraídas amostras aleatórias e independentes pode fazer-se através de uma metodologia proposta por Sir Ronald Fisher e genericamente designada por Análise de Variância (abreviadamence ANOVA do inglês *Analysis of Variance*) (Fisher, 1935) se a distribuição da variável em estudo for Normal e se as variâncias populacionais forem homogéneas"[22].

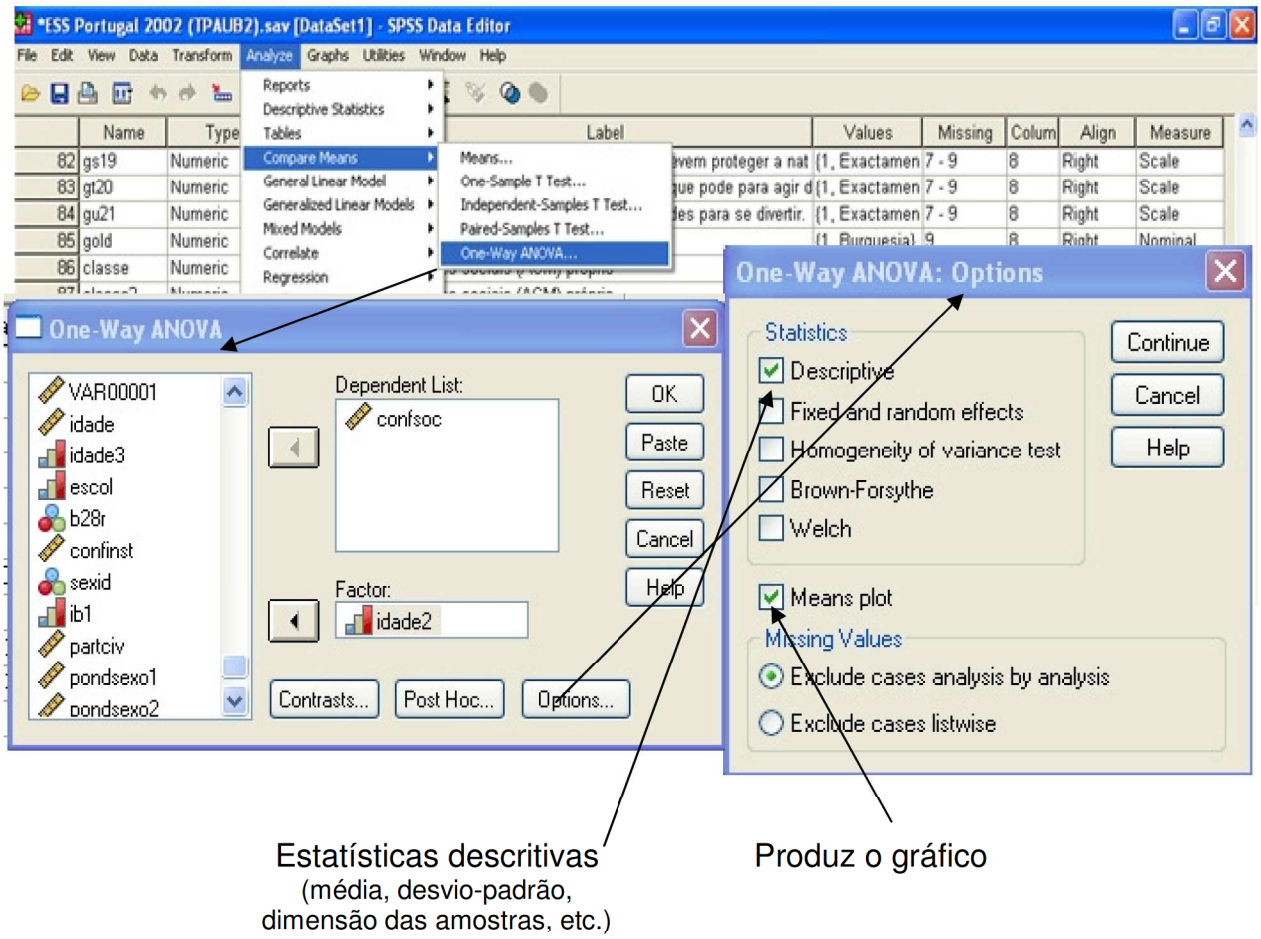
Exemplo: Pretende-se testar se há relação entre a idade (***idade2***) e a confiança social (***confsoc***).

O procedimento consiste na realização da Análise de Variância Simples Paramétrica (***One-way Anova***)[23].

Hipótese do teste (bilateral):

H_o: A média da confiança social é igual em todos os escalões etários

H_a: A média da confiança social é diferente em pelo menos um escalão etário



[22] Maroco, idem: 154.
[23] Consultar o Anexo 1.

O resultado é o seguinte:

Descriptives

Índice sintético de Confiança social

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Até 30 anos	338	4.722	1.5585	.0848	4.555	4.889	.0	9.3
31 - 50 anos	497	4.058	1.6747	.0751	3.911	4.206	.0	9.3
51 - 65 anos	312	4.068	1.8340	.1038	3.864	4.273	.0	10.0
> 65 anos	333	4.519	1.8500	.1014	4.319	4.718	.0	10.0
Total	1480	4.316	1.7469	.0454	4.226	4.405	.0	10.0

Nota: como as amostras são superiores a 30 (N), não se torna necessário testar a "normalidade"

Test of Homogeneity of Variances

Índice sintético de Confiança social

Levene Statistic	df1	df2	Sig.
4.050	3	1476	.007

Hipótese para a "Homogeneidade de variâncias":

H_0: as variâncias são homogéneas

H_1:as variâncias não são homogéneas

Tomada de decisão: Rejeita-se a hipótese nula, não se verifica o pressuposto.

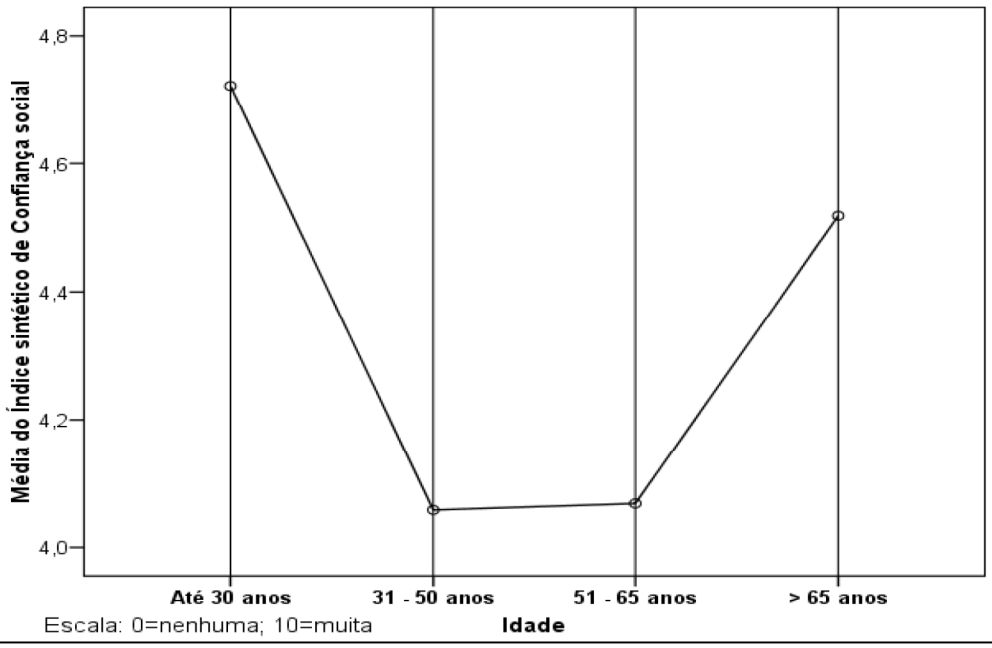
No entanto, a "violação deste pressuposto não tem consequências sérias se os grupos ou amostras tiverem dimensões razoáveis"[24]. Na prática, a Anova é robusta à violação do pressuposto se a maior das amostras não for superior ao dobro da menor.

ANOVA

Índice sintético de Confiança social

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	121.466	3	40.489	13.607	.000
Within Groups	4391.954	1476	2.976		
Total	4513.420	1479			

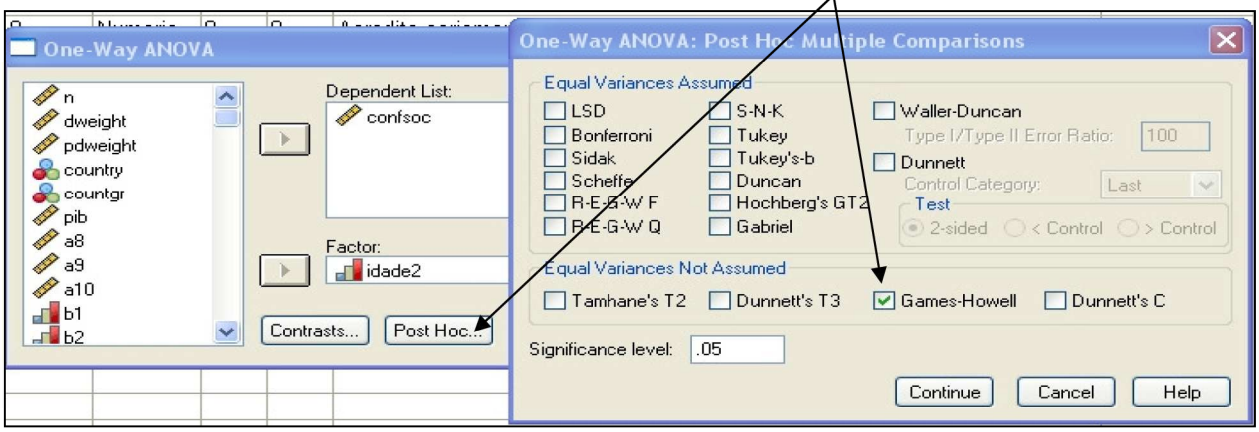
Tomada de decisão:
Rejeita-se a hipótese nula, há pelo menos um escalão etário em que a média da confiança social é significativamente diferente.



Nota: Tendo-se rejeitado a hipótese nula, conclui-se que há pelo menos um escalão etário onde a confiança social apresenta valores médios significativamente diferentes dos restantes escalões. Nesta situação importa, por conseguinte, saber quais são os escalões que diferem uns dos outros. Para o efeito realiza-se um teste de comparações múltiplas à posteriori (***Post Hoc***).

[24] Murteira, B. (1990), Probabilidades e Estatística, Lisboa McGraw-Hill, vol.II: :349.

O SPSS disponibiliza vários testes para este fim, sendo os mais utilizados, o teste de *Scheffe*[25], no caso de as variâncias serem iguais, e o teste *Games-Howell* no caso de serem diferentes. Neste caso, uma vez que se rejeita a hipótese de as variâncias serem iguais (p=0,007), vamos solicitar o teste *Games-Howell*. Par o efeito, repete-se o procedimento e selecciona-se o teste no separador "PostHoc":



O resultado é o seguinte:

Multiple Comparisons

Dependent Variable: Índice sintético de Confiança social

Games-Howell

(I) Idade	(J) Idade	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Até 30 anos	Até 30 anos					
	31 - 50 anos	,6635*	,1133	,000	,372	,955
	51 - 65 anos	,6535*	,1340	,000	,308	,999
	> 65 anos	,2034	,1321	,415	-,137	,544
31 - 50 anos	Até 30 anos	-,6635*	,1133	,000	-,955	-,372
	31 - 50 anos					
	51 - 65 anos	-,0100	,1282	1,000	-,340	,320
	> 65 anos	-,4602*	,1262	,002	-,785	-,135
51 - 65 anos	Até 30 anos	-,6535*	,1340	,000	-,999	-,308
	31 - 50 anos	,0100	,1282	1,000	-,320	,340
	51 - 65 anos					
	> 65 anos	-,4501*	,1451	,011	-,824	-,076
> 65 anos	Até 30 anos	-,2034	,1321	,415	-,544	,137
	31 - 50 anos	,4602*	,1262	,002	,135	,785
	51 - 65 anos	,4501*	,1451	,011	,076	,824
	> 65 anos					

*. The mean difference is significant at the .05 level.

Estamos agora em condições de interpretar o resultado da Análise de Variância.

Interpretação: Os mais novos (4,722), seguidos dos mais velhos (4,519) são os que mais confiam. Os escalões intermédios 31-50 anos (4,058) e 51-65 anos (4,068) confiam um pouco menos. As diferenças são estatisticamente significativas (F (3)=13,787; p=0,000). O quadro seguinte sintetiza as diferenças de médias significativas entre os quatro escalões etários:

Confiança Social: diferenças de médias entre escalões etários

	Até 30 anos	31-50 anos	51-65 anos
Até 30 anos			
31-50 anos	0,6635**		
51-65 anos	0,6535**	*n.s.*	
> 65 anos	*n.s.*	0,4602*	0,4501*

** p< 0,05; ** p<0,001;*

[25] Que é também o mais conservador,

4. Correlação

Em estatística, a *correlação*, também chamada de *coeficiente de correlação*, indica a força e a direcção do relacionamento entre duas variáveis. Refere-se à medida da relação entre duas variáveis, embora a correlação não implique causalidade. Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados.

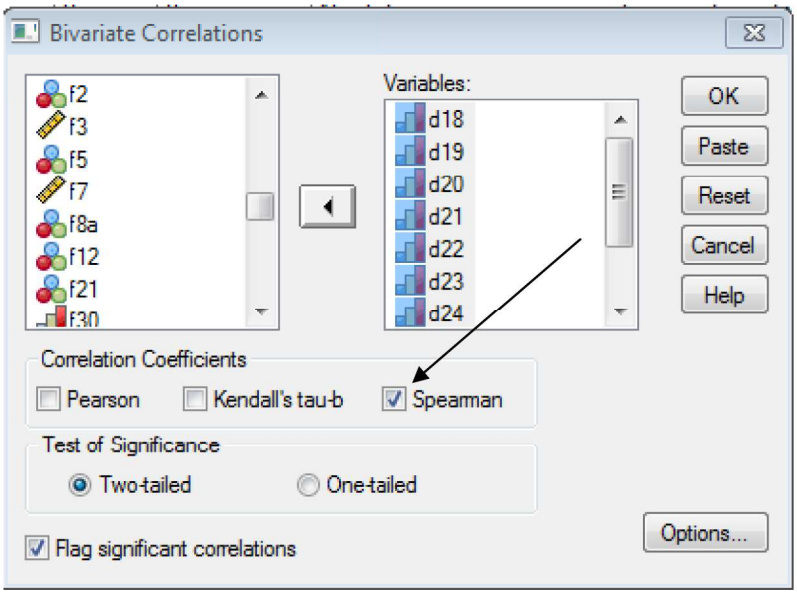
O coeficiente de correlação varia entre –1 e 1[26] e deve ser interpretado da seguinte forma:

0:	ausência de correlação
+/-]0 – 0,25]:	muito fraca
+/-]0,25 – 0,40]:	fraca
+/-]0,40 – 0,60]:	moderada
+/-]0,60 – 0,75]:	moderada forte
+/-]0,75 – 0,90]:	forte
+/-]0,90 – 1[:	muito forte
+/- 1:	correlação perfeita

4.1. Correlação não linear de Spearman (*rho* de Spearman)

"O coeficiente de correlação de Spearman mede a intensidade da relação entre variáveis ordinais. Utiliza os valores de ordem das observações em vez do seu valor observado. Deste modo, este coeficiente não é sensível a assimetrias na distribuição, nem à presença de outliers, não exigindo que os dados provenham de duas populações normais. Aplica-se igualmente em variáveis intervalo/rácio como alternativa ao R de Pearson, quando neste último se viola a normalidade"[27].

Exemplo: Correlação entre as variáveis atitudinais sobre os imigrantes (***d18*** a ***d24***). Trata-se de variáveis ordinais pelo que o coeficiente de correlação mais adequado a este tipo de variáveis é o *rho de Spearman*:



[26] O sinal – significa uma relação negativa e a ausência de sinal uma relação positiva.
[27] Pestana, M. H. e J. N. Gageiro (2000), *Análise de dados para Ciências Sociais – A Complementaridade do SPSS*, Lisboa, Sílabo. 2ª edição.

O resultado é o seguinte[28]:

Correlações (*Spearman's rho*)

		d18	d19	d20	d21	d22	d23
d18 As pessoas que vêm viver e trabalhar para cá fazem com que os salários baixem	Rho						
	Sig.	.					
	N	1406					
d19 As pessoas que vêm viver e trabalhar para cá, em regra, prejudicam mais as expectativas económicas dos pobres do que dos ricos	Rho	,569(**)					
	Sig.	,000	.				
	N	1366	1420				
d20 As pessoas que vêm viver e trabalhar para cá ajudam a preencher lugares em que há falta de trabalhadores	Rho	-,188(**)	-,204(**)				
	Sig.	,000	,000	.			
	N	1378	1391	1453			
d21 Se as pessoas que vieram viver e trabalhar para cá estiverem desempregadas por muito tempo deviam ser obrigadas a ir embora	Rho	,250(**)	,327(**)	-,134(**)			
	Sig.	,000	,000	,000	.		
	N	1353	1371	1391	1422		
d22 As pessoas que vieram viver para cá devem ter os mesmos direitos do que todas as outras pessoas	Rho	-,223(**)	-,187(**)	,318(**)	-,111(**)		
	Sig.	,000	,000	,000	,000	.	
	N	1382	1398	1428	1405	1462	
d23 As pessoas que vieram viver para cá cometerem um crime grave, devem ser obrigadas a ir embora	Rho	,226(**)	,271(**)	-,051	,307(**)	,037	
	Sig.	,000	,000	,057	,000	,161	.
	N	1378	1391	1418	1396	1429	1455
d24 As pessoas que vieram viver para cá cometerem qualquer crime, devem ser obrigadas a ir embora	Rho	,250(**)	,219(**)	-,110(**)	,341(**)	-,073(**)	,594(**)
	Sig.	,000	,000	,000	,000	,006	,000
	N	1376	1388	1413	1392	1425	1444

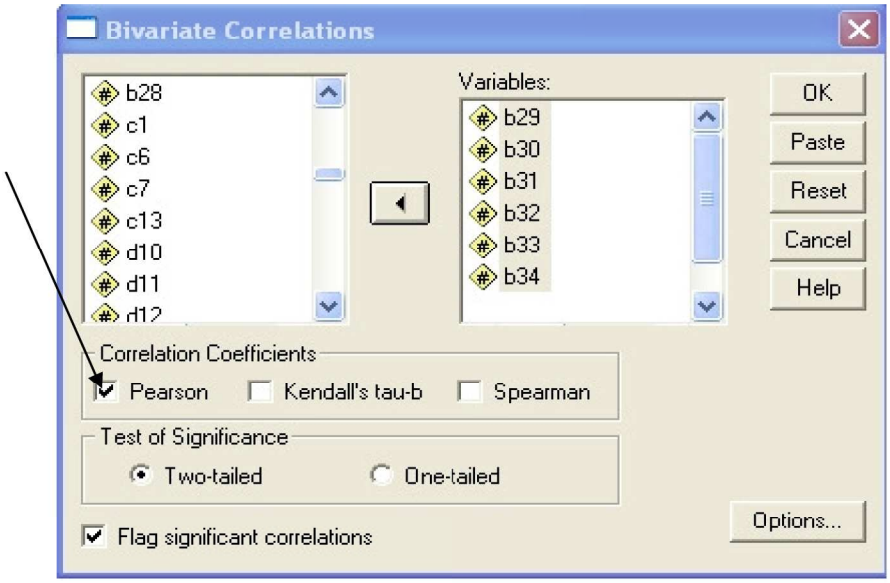
** Correlation is significant at the 0.01 level (2-tailed).

[28] O output foi melhorado no Word.

4.2. Correlação linear simples (*r* de Pearson)

A correlação linear simples permite obter uma medida (*coeficiente de correlação r de* Pearson) através da qual se determina a força ou intensidade de uma associação linear entre duas ou mais variáveis quantitativas ou tratadas como tal (escalas tipo *Likert*).

Exemplo: Correlação entre as variáveis satisfação com a vida (***b29***), com a economia (***b30***), com o Governo (***b31***), com a democracia (***b32***), com a educação (***b33***) e com os serviços de saúde (***b34***):



O resultado é o seguinte:

Correlations

		Satisfação com a vida em geral	Economia	Governo	Democracia	Educação
Satisfação com a vida em geral	Pearson Correlation					
	Sig. (2-tailed)					
	N					
Economia	Pearson Correlation	,339**				
	Sig. (2-tailed)	,000				
	N	1441				
Governo	Pearson Correlation	,280**	,578**			
	Sig. (2-tailed)	,000	,000			
	N	1413	1392			
Democracia	Pearson Correlation	,348**	,403**	,507**		
	Sig. (2-tailed)	,000	,000	,000		
	N	1371	1353	1339		
Educação	Pearson Correlation	,205**	,361**	,289**	,300**	
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	1429	1389	1367	1338	
Serviços de Saúde	Pearson Correlation	,195**	,396**	,340**	,294**	,537**
	Sig. (2-tailed)	,000	,000	,000	,000	,000
	N	1489	1440	1412	1370	1433

**. Correlation is significant at the 0.01 level (2-tailed).

Interpretação: as correlações são positivas e significativas entre todas as variáveis; (*p*=0,000), sendo a menor entre a satisfação com a vida e com a educação (0,205) e a maior entre a satisfação com o Governo e com a economia (0,578