Variáveis bidimensionais

Muitas vezes, a análise estatística tem em vista o estudo, em simultâneo, de duas características de cada indivíduo dando origem a amostras bivariadas (ou bidimensionais). Assim, estas amostras são constituídas por pares de dados; o primeiro elemento corresponde à observação de um atributo de um determinado indivíduo e o segundo elemento corresponde à observação de outro atributo do mesmo indivíduo. Tais atributos podem ser ambos quantitativos, ambos qualitativos ou um de cada tipo.

Na análise de uma amostra bivariada, para além de se poderem considerar separadamente os dados relativos a cada atributo, interessa frequentemente verificar se existe algum tipo de associação entre eles e, no caso afirmativo, caracterizar essa relação.

1 Dados qualitativos *vs* dados qualitativos

1.1 Tabelas e gráficos

Para resumir amostras bivariadas de dados qualitativos usam-se habitualmente *tabelas de informação cruzada* (também designadas *tabelas de contingência*) e *diagramas de barras*.

Ilustramos estes procedimentos recorrendo ao exemplo de uma empresa que é abastecida de um determinado produto por três firmas distribuidoras, A, B e C. A empresa classifica o grau de qualidade (GQ) das entregas da seguinte forma: grau I - entregas efectuadas em perfeitas condições, grau II - entregas efectuadas em condições aceitáveis, grau III - entregas efectuadas em más condições. As entregas feitas num determinado período foram distribuídas da seguinte forma (dados *in* Guimarães e Sarsfield Cabral, pág 35[1]):

- a firma A efectuou 21 entregas de grau I, 10 entregas de grau II e 8 entregas de grau III;
- a firma B efectuou 12 entregas de grau I, 4 entregas de grau II e 2 entregas de grau III;
- a firma C efectuou 3 entregas de grau I, 3 entregas de grau II e 2 entregas de grau III.

Esta amostra bivariada corresponde a duas variáveis qualitativas (*String*) que designamos por *Firma* e *GQ*. Dispomos de 65 pares de observações ($21+10+8+12+4+2+3+3+2 = 65$), dos quais 21 são iguais a (A,I), 10 são iguais a (A,II), 8 são iguais a (A,III), 12 são iguais a (B,I), 4 são iguais a (B,II), 2 são iguais a (B,III), 3 são iguais a (C,I), 3 são iguais a (C,II) e 2 são iguais a (C,III). Assim, o ficheiro de dados a construir para esta amostra é constituído por duas colunas correspondentes às variáveis *Firma* e *GQ*. Na primeira coluna (*Firma*) a modalidade A aparece 39 vezes ($21 + 10 + 8 = 39$), a modalidade B aparece 18 vezes ($12 + 4 + 2 = 18$) e a modalidade C aparece 8 vezes ($3 + 3 + 2 = 8$). Na segunda coluna (*GQ*) tem-se, respectivamente, 21 vezes a modalidade I, 10 vezes a modalidade II, 8 vezes a modalidade III, 12 vezes a modalidade I, 4 vezes a modalidade II, 2 vezes a modalidade III, 3 vezes a modalidade I, 3 vezes a modalidade II e 2 vezes a modalidade III. Assim, os dois registos de cada uma das 65 linhas correspondem, respectivamente, aos 65 pares acima descritos.

[1] Guimarães, R.C. e Sarsfield Cabral, J.A. (2007) Estatística (2^a edição) McGraw-Hill.

O SPSS permite obter a tabela e o gráfico que resumem esta amostra bivariada. Para tal, usamos a opção *Analyze* seguida de *Descriptive Statistics* e *Crosstabs....* Na janela que aparece a seguir, colocamos a variável *Firma* em *Row*(s) e a variável *GQ* em *Column*(s). Seleccionamos a opção *Display clustered bar charts* para obter o gráfico de barras e podemos obter a tabela mais simples em *Cells*, seleccionando *Observed* em *Counts*. Na figura 1 apresenta-se a tabela fornecida no output.

Count

		GQ			
		I	II	III	Total
Firma	A	21	10	8	39
	B	12	4	2	18
	C	3	3	2	8
Total		36	17	12	65

Figura 1: Tabela de contingência.

Os valores 39, 18, 8 e 65 que surgem na última coluna correspondem aos totais acima referidos enquanto que os valores 36, 17 e 12 que podem ser observados na última linha correspondem ao total de entregas feitas na empresa que foram efectuadas, respectivamente, em perfeitas condições (I), em condições aceitáveis (II) e em más condições (III).

Apresenta-se a seguir (figura 2) uma representação gráfica da amostra bivariada em análise.

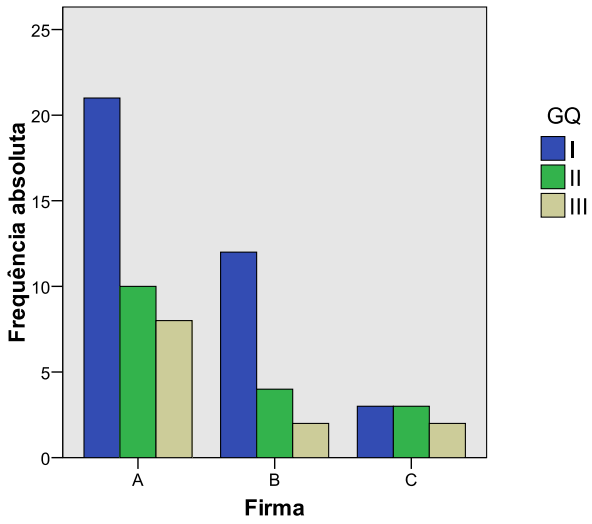

Figura 2: Gráfico de barras de uma amostra bivariada.

Neste gráfico, a altura da primeira barra (resp., segunda barra e terceira barra) em cada grupo de três, representa o número de vezes que aparece a modalidade I (resp., II e III) na correspondente modalidade da variável *Firma*.

Notamos que a ordem pela qual as modalidades aparecem no gráfico pode ser alterada. Para tal, basta clicar duas vezes no gráfico, surgindo uma janela intitulada *Chart Editor*. Clicando mais duas vezes sobre uma das barras, surge uma janela (*Properties*) onde a opção *Categories* permite fazer as referidas alterações em cada uma das variáveis.

Podemos obter uma tabela com mais informação seleccionando, em *Cells*, no quadro *Percentages*, as opções *Row*, *Column* e *Total*, ou apenas uma ou duas delas.

Na figura 3 apresenta-se a tabela que se obtém seleccionando simultaneamente as três opções acima referidas.

Firma * GQ Crosstabulation

			GQ			
			I	II	III	Total
Firma	A	Count	21	10	8	39
		% within Firma	53,8%	25,6%	20,5%	100,0%
		% within GQ	58,3%	58,8%	66,7%	**60,0%**
		% of Total	32,3%	15,4%	12,3%	60,0%
	B	Count	12	4	2	18
		% within Firma	66,7%	22,2%	11,1%	100,0%
		% within GQ	33,3%	23,5%	16,7%	**27,7%**
		% of Total	18,5%	6,2%	3,1%	27,7%
	C	Count	3	3	2	8
		% within Firma	37,5%	37,5%	25,0%	100,0%
		% within GQ	8,3%	17,6%	16,7%	**12,3%**
		% of Total	4,6%	4,6%	3,1%	12,3%
Total		Count	36	17	12	65
		% within Firma	**55,4%**	**26,2%**	**18,5%**	**100,0%**
		% within GQ	100,0%	100,0%	100,0%	**100,0%**
		% of Total	55,4%	26,2%	18,5%	100,0%

Figura 3: Tabela de contingência com percentagens.

Nesta tabela figuram, em cada célula (de cima para baixo), a frequência absoluta do par correspondente, a percentagem de observações desse par relativamente ao total de pares observados com a modalidade da linha correspondente, a percentagem de observações desse par relativamente ao total de pares observados com a modalidade da coluna correspondente e a percentagem de observações desse par relativamente ao total de pares observados. A apresentação destes valores em simultâneo permite a observação de várias características da amostra, nomeadamente

- a importância relativa das firmas distribuidoras (ver o total de cada linha),
- a distribuição do número de entregas por grau de qualidade (ver o total de cada coluna),
- a relação entre as firmas distribuidoras e a qualidade das entregas (ver, para as diferentes células correspondentes a cada firma, as percentagens do número de observações da linha).

Os gráficos podem também ser obtidos através da opção *Graphs* da barra de ferramentas. O trajecto *Graphs* $\rightarrow$ *Chart Builder* $\rightarrow$ *OK* conduz a uma janela onde podemos escolher o tipo de gráfico pretendido. Por exemplo, se seleccionarmos *Bar* em *Gallery,Choose from* e escolhermos o tipo *Stacked Bar*, obtemos o gráfico apresentado na figura 4 depois de colocarmos a variável *Firma* no $X - Axis$ (automaticamente, aparece *Count* no $Y - Axis$) e a variável *GQ* em *Stack: set color*.

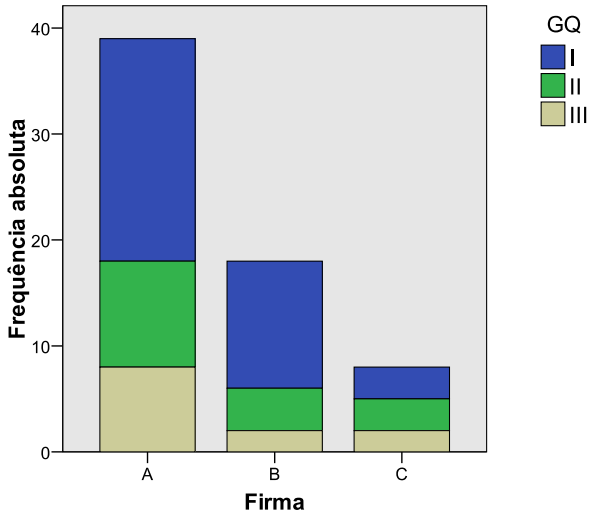

Figura 4: Gráfico de barras sobrepostas.

No gráfico de barras sobrepostas, os três rectângulos sobrepostos que correspondem à modalidade A (resp., B e C) representam o número de observações de cada uma das modalidades I, II e III associadas a A (resp., B e C). Assim, a altura total destes rectângulos sobrepostos é igual ao número de pares observados com a modalidade A (resp., B e C).

1.2 Associação entre as variáveis

Uma análise importante é, sem dúvida, a da existência de algum tipo de associação entre as duas variáveis em estudo. Podemos verificar se a amostra é constituída por valores de variáveis estatísticas independentes usando a tabela de contingência correspondente.

Consideremos por exemplo a seguinte tabela:

	B_1	B_2	B_3	Total da linha
A_1	4	10	2	**16**
A_2	6	15	3	**24**
Total da coluna	**10**	**25**	**5**	n**=40**

A independência acima referida existe se e só se a frequência de cada célula é igual ao produto da frequência total da linha correspondente pela frequência total da coluna correspondente dividido pelo total de observações. No presente exemplo temos $4 = 16 \times 10/40$, $10 = 16 \times 25/40$, $2 = 16 \times 5/40$, $6 = 24 \times 10/40$, $15 = 24 \times 25/40$ e $3 = 24 \times 5/40$. Assim, verifica-se a independência entre as duas variáveis estatísticas em análise.

Se não se verifica a independência, existem algumas medidas de associação que podemos calcular. Uma delas é o *coeficiente de contingência de Pearson* que é baseado na comparação das frequências absolutas observadas com as frequências absolutas que se teriam no caso de independência entre as variáveis. Representando por n_{ij} a frequência da célula que figura na linha i e na coluna j e por e_{ij} a correspondente frequência esperada em caso de independência,

começamos por calcular o valor χ^2 (lê-se "qui-quadrado") da seguinte forma:

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - e_{ij})^2}{e_{ij}},$$

onde p é o número de linhas e q é número de colunas. Note-se que e_{ij} é dado pelo produto do total da linha i pelo total da linha j dividido por n (total de observações).

O *coeficiente de contingência de Pearson* é então dado por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

Este coeficiente toma valores entre 0 e 1, mas nunca atinge o valor 1. O valor máximo que C pode atingir é $\sqrt{\frac{m-1}{m}}$, onde m é o mínimo entre p e q. Por outro lado, o valor 0 corresponde ao caso em que se tem independência e valores mais elevados correspondem a associação mais forte entre as variáveis.

Este coeficiente (e outras medidas de associação) obtém-se no SPSS através de *Analyze* → *Descriptive Statistics* → *Crosstabs* → *Statistics* → *Contingency coef f icient*.

2 Dados quantitativos *vs* dados quantitativos

No caso em que as duas variáveis são quantitativas, a possibilidade de efectuar cálculos permite que, além do resumo da amostra através de tabelas de contingência e de representações gráficas, seja ainda feita a descrição numérica da variável bidimensional em estudo. Esta descrição tem como principal objectivo a análise da existência de algum tipo de dependência funcional de uma variável relativamente à outra.

2.1 Representação gráfica: diagrama de dispersão

Um guia útil para indicar o tipo de função subjacente à estrutura de dependência entre duas variáveis é o *gráfico de dispersão* (também chamado *nuvem de pontos*). Este gráfico é constituído pelos pontos correspondentes aos n pares ordenados (x_i, y_i), $i = 1, \ldots, n$, que constituem a amostra bivariada. Os valores y_i correspondem às observações da *variável dependente*, também designada *variável resposta*, que representamos por Y, enquanto que os valores x_i correspondem às observações da *variável independente* ou *variável explicativa*, que representamos por X.

Para exemplificar, consideremos os dados do quadro abaixo relacionados com a taxa de oxigénio consumido por determinado tipo de animais em zonas com temperaturas ambientais diferentes.

Temperat. (^oC) - x_i	-18	-15	-10	-5	0	5	10	19
T. de ox. $(ml/g/h)$ - y_i	5.2	4.7	4.5	3.6	3.4	3.1	2.7	1.8

A figura 5 mostra a nuvem de pontos correspondente a esta amostra bivariada. Observamos que os pontos se distribuem de forma quase linear com tendência decrescente, o que significa que podemos considerar que a estrutura de dependência entre as duas variáveis em causa é bem representada por uma recta com declive negativo, i.e., por uma função da forma $y = a+bx$, com

$b < 0$. Claro que esta função só fica completamente determinada com a indicação de valores para a e b. Na realidade, o que conseguimos obter são valores aproximados que dependem da amostra recolhida, como veremos adiante.

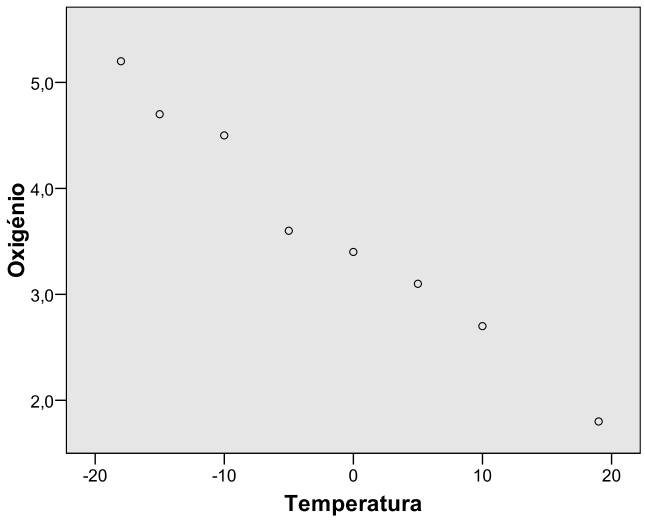

Figura 5: Gráfico de dispersão.

A obtenção deste gráfico no SPSS é conseguida de modo semelhante ao indicado acima para a construção o gráfico de barras sobrepostas, finalizando com *Scatter Dot* + *Simple Scatter*.

2.2 Descrição numérica

A *média amostral* de dados bivariados é o par $(\bar{x}, \bar{y})$, onde $\bar{x}$ (resp., $\bar{y}$) representa a média dos valores x_i (resp., y_i), $i = 1, \ldots, n$. O ponto $(\bar{x}, \bar{y})$ é interpretado como o centro de massa (baricentro) do conjunto de pontos que constituem a amostra e não faz necessariamente parte dela.

Tendo em vista a análise da dependência linear entre X e Y, impõe-se a definição de medidas que permitam avaliar a ligação existente entre aquelas variáveis. A primeira dessas medidas é a *covariância amostral* entre X e Y, que representamos por s_{XY} e é dada por

$$s_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Esta medida tem a desvantagem de depender das unidades em que os dados são expressos, desvantagem essa que é ultrapassada dividindo s_{XY} pelo desvio-padrão de X, s_X e pelo desvio-padrão de Y, s_X. Obtém-se assim o chamado *coeficiente de correlação amostral* que denotamos por r_{XY}:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

Claro que r_{XY} não se define quando $s_X s_Y = 0$, mas tal só acontece se os valores observados de (pelo menos) uma das variáveis forem todos iguais.

Note-se que r_{XY} e s_{XY} têm o mesmo sinal. Se esse sinal for positivo (resp., negativo) então dizemos que há uma correlação positiva (resp., negativa).

Além disso, verifica-se que r_{XY} toma apenas valores entre -1 e 1. Nos casos particulares $r_{XY} = -1$ e $r_{XY} = 1$, os pontos (x_i, y_i), $i = 1, \ldots, n$, estão todos sobre a mesma recta, i.e., existe uma relação perfeitamente linear entre eles (no primeiro caso a recta tem declive negativo e no segundo caso a recta tem declive positivo). Podemos então inferir que a relação entre X e Y está tão mais próxima da linear quanto mais próximo de 1 ou de -1 estiver o valor de r_{XY}. Por outro lado, se este valor está próximo de 0 é de excluir a explicação dos valores tomados por Y através de uma relação linear com X. No entanto, tal facto não exclui a possibilidade da existência de outro tipo de relação funcional entre as duas variáveis. Consideremos por exemplo a nuvem de pontos representada na figura 6.

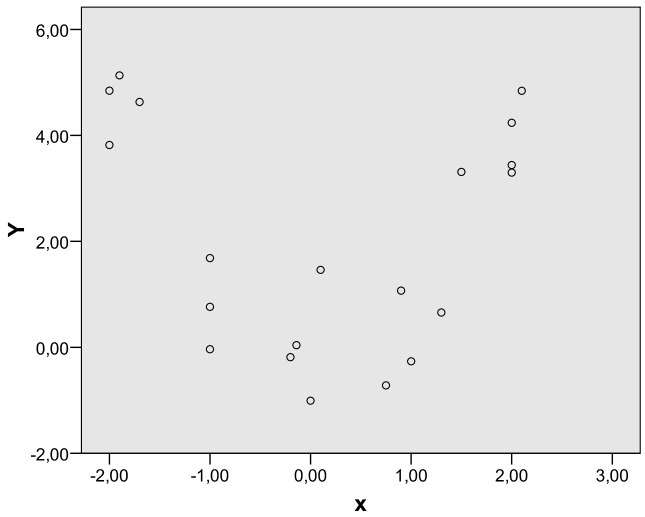

Figura 6: Gráfico de dispersão.

A amostra correspondente verifica $r_{XY} = -0.079$, indicando que não deve ser considerada uma relação linear entre X e Y. No entanto, a observação da nuvem de pontos sugere uma relação do tipo $y = a + bx + cx^2$ (parábola).

Pelo contrário, à nuvem de pontos representada na figura 5 corresponde, como seria de esperar, um valor de r_{XY} muito próximo de -1, mais precisamente, $r_{XY} = -0.99$. Este valor pode ser obtido no SPSS através, por exemplo, do seguinte percurso: *Analyse* → *Descriptive Statistics* → *Crosstabs*. Depois de colocar a variável X em *Row*(s) e a variável Y em *Column*(s), clicar em *Statistics* e seleccionar *Correlations*. A tabela com o valor de r_{XY} encontra-se na figura 7 (*Pearson's R*).

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Interval by Interval	**Pearson's R**	**-,990**	,007	-17,576	,000
Ordinal by Ordinal	Spearman Correlation	-1,000	,000		
N of Valid Cases		8			

Figura 7: Tabela com o valor do coeficiente de correlação.

2.3 Recta dos mínimos quadrados

Quando o valor de r_{XY} justifica a existência de uma relação linear entre as duas variáveis em análise, é de todo o interesse encontrar a equação da recta que melhor se ajusta (segundo um critério a definir) aos pontos do gráfico de dispersão. Usualmente, procura-se a recta tal que seja mínima a média dos quadrados das distâncias de cada ponto da nuvem ao ponto da recta que possui a mesma abcissa, i.e., procuramos os valores a e b de modo que o valor de $\frac{1}{n}\sum_{i=1}^{n}[y_i - (a + bx_i)]^2$ seja o mínimo possível. Denotando por $(\hat{a}, \hat{b})$ a solução do problema de minimização acima descrito, verifica-se que

$$\begin{cases} \hat{b} = \dfrac{s_{XY}}{s_X^2} = \dfrac{s_Y}{s_X}\dfrac{s_{XY}}{s_X s_Y} = r_{XY}\dfrac{s_Y}{s_X} \\ \\ \hat{a} = \bar{y} - \hat{b}\,\bar{x} \end{cases}.$$

A recta obtida tem então a equação $y = \hat{a} + \hat{b}\,x$ a qual, efectuando alguns cálculos, também pode ser escrita na forma

$$y = r_{XY}\frac{s_Y}{s_X}(x - \bar{x}) + \bar{y}. \tag{1}$$

Esta recta é designada por *recta dos mínimos quadrados*, ou *de regressão, de Y sobre X*.

A partir da equação (1), conclui-se de imediato que o ponto $(\bar{x}, \bar{y})$ pertence à recta de regressão.

Observemos ainda que, uma vez que os valores s_X e s_Y são positivos, o declive da recta de regressão, $\hat{b}$, tem o sinal de r_{XY}. Assim, no caso em que r_{XY} é positivo (resp., negativo) um acréscimo numa das variáveis acarreta um acréscimo (resp., diminuição) na outra.

Os valores $\hat{a}$ e $\hat{b}$ pode ser obtida no SPSS através de *Analyse* → *Regression* → *Linear...*. Na figura 8 encontra-se a tabela do output onde figuram tais coeficientes para a recta de regressão correspondente à nuvem de pontos da figura 5.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	**(Constant)**	**3,471**	,060		57,738	,000
	temperatura	**-,088**	,005	-,990	-17,576	,000

a. Dependent Variable: oxigenio

Figura 8: Tabela dos coeficientes da recta de regressão.

O valor de $\hat{a}$ é indicado como (*Constant*) e o valor de $\hat{b}$ corresponde ao nome da variável dependente. A equação da recta de regressão para o caso aqui considerado é então

$$y = 3.471 - 0.088\,x.$$

Esta equação permite-nos inferir valores de Y correspondentes a valores de X que não figuram na amostra. Por exemplo, a uma temperatura de 2^oC corresponde uma taxa de oxigénio de $3.471 - 0.088 \times 2 = 3.295\,ml/g/h$.

Na figura 9 apresenta-se a recta de regressão acima indicada sobre a correspondente nuvem de pontos.

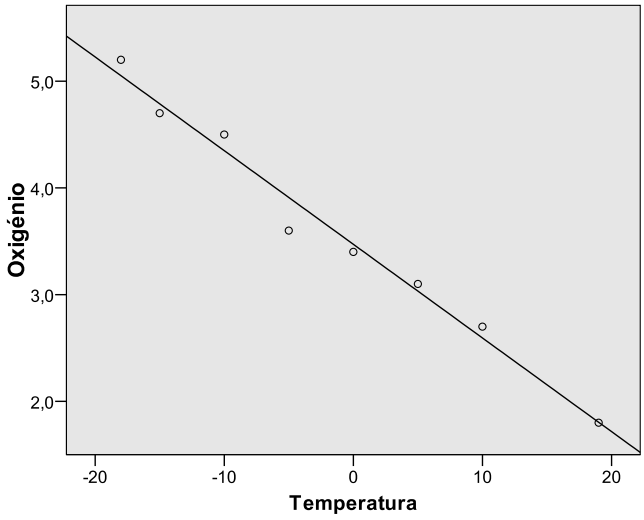

Figura 9: Nuvem de pontos com recta de regressão.

3 Dados qualitativos *vs* dados quantitativos

No caso em que uma das variáveis é qualitativa e a outra é quantitativa os gráficos e tabelas adequados são apenas os que foram referidos para os dados qualitativos. No entanto, ao escolhermos a opção *Graphs*→ *ChartBuilder* para fazer os gráficos de barras, o SPSS não aceita variáveis com o tipo *Scale* sendo portanto necessário alterar o tipo da variável quantitativa para *Ordinal*.