

3. Características amostrais. Medidas de localização e dispersão

3.1- Introdução

No módulo de Estatística foram apresentadas as medidas ou estatísticas que se utilizam para resumir a informação contida nos dados. Destas medidas, destacam-se as medidas de localização, nomeadamente as que localizam o centro da amostra, e as medidas de dispersão, que medem a variabilidade dos dados.

Neste capítulo não nos debruçaremos sobre as propriedades destas medidas, já apresentadas no módulo referido anteriormente, abordando sobretudo a forma de as calcular, utilizando o Excel. Convém desde já adiantar que este é um trabalho grandemente facilitado pelo facto de existirem funções no Excel que nos dão directamente estas medidas.

Para facilidade de exposição vamos representar a amostra de dimensão n por

$$x_1, x_2, \dots, x_n$$

onde x_1, x_2, \dots, x_n representam, respectivamente, os resultados da 1ª observação, da 2ª observação, da n -ésima observação, a serem recolhidas, não pressupondo qualquer ordenação.

3.2 – Medidas de localização

Como medidas de localização, vamos apresentar a média, mediana e quartis.

3.2.1 – Média

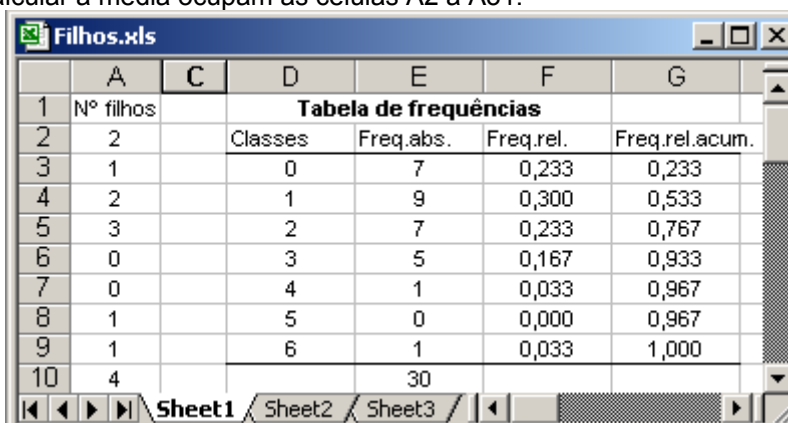
A *média* é uma medida de localização do centro da distribuição dos dados. Dada a amostra x_1, x_2, \dots, x_n , a média representa-se por \bar{x} e obtém-se adicionando todos os elementos e dividindo o resultado por n . Em Excel, determina-se a média através da função *AVERAGE* (), que retorna a média aritmética dos seus argumentos, que podem ser números ou endereços de células.

Exemplo 3.2.1 – Retomemos a amostra do exemplo 2.3.2, constituída pelo número de filhos de 30 deputados:

2, 1, 2, 3, 0, 0, 1, 1, 4, 1, 2, 1, 0, 0, 0, 2, 3, 1, 1, 6, 3, 1, 3, 2, 0, 1, 2, 0, 2, 3

Calcule a média da amostra.

Considerámos o ficheiro Filhos.xls, constituído no exemplo 2.3.2, em que os elementos de que se pretende calcular a média ocupam as células A2 a A31:



	A	C	D	E	F	G
1	Nº filhos		Tabela de frequências			
2	2		Classes	Freq.abs.	Freq.rel.	Freq.rel.acum.
3	1		0	7	0,233	0,233
4	2		1	9	0,300	0,533
5	3		2	7	0,233	0,767
6	0		3	5	0,167	0,933
7	0		4	1	0,033	0,967
8	1		5	0	0,000	0,967
9	1		6	1	0,033	1,000
10	4			30		

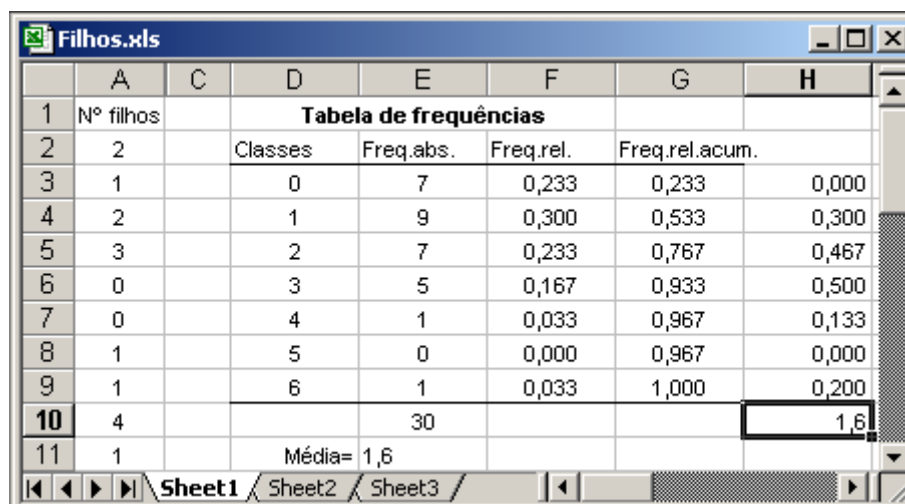
Para calcular a média pretendida, assim como para qualquer outro conjunto de dados de tipo discreto, podemos proceder de dois modos, quer considerando os dados originais, quer agrupados.

1- Cálculo da média, a partir dos dados originais, utilizando a função *AVERAGE()*:

Colocar o cursor na célula onde se pretende colocar a média, por exemplo a célula E11, e inserir a função *AVERAGE(A2:A31)* – os argumentos desta função são os endereços onde estão os elementos da amostra. Como resultado obtém-se o valor 1,6, que se apresenta na figura seguinte.

2- Cálculo da média, a partir dos dados agrupados:

Adicionar à tabela de frequências uma nova coluna com o produto dos valores que constituem as classes, pelas respectivas frequências relativas (Células H3 a H9) e somar os valores obtidos (Célula H10):



	A	C	D	E	F	G	H
1	Nº filhos		Tabela de frequências				
2	2		Classes	Freq.abs.	Freq.rel.	Freq.rel.acum.	
3	1		0	7	0,233	0,233	0,000
4	2		1	9	0,300	0,533	0,300
5	3		2	7	0,233	0,767	0,467
6	0		3	5	0,167	0,933	0,500
7	0		4	1	0,033	0,967	0,133
8	1		5	0	0,000	0,967	0,000
9	1		6	1	0,033	1,000	0,200
10	4			30			1,6
11	1		Média= 1,6				

No caso de dados discretos, como é o caso anterior, o valor da média é o mesmo, quer seja calculada utilizando os dados originais, quer os dados agrupados (utilizando as frequências relativas), em que as classes do agrupamento são os diferentes valores que surgem na amostra. O mesmo não acontece no caso de dados contínuos, como exemplificamos a seguir.

Exemplo 3.2.2 – Calcule a média das idades dos deputados do ficheiro *Deputados.xls*.

Para obter a média das idades procede-se como no primeiro caso do exemplo anterior, a partir dos dados originais. Estes dados encontram-se nas células C2 a C231 do ficheiro *Idade.xls*, inserindo a função *AVERAGE(C2:C231)* na célula L13, obtemos o valor de 48,66 anos.

Admitindo que não dispúnhamos dos dados originais, mas apenas de uma tabela de frequências com os dados agrupados, vejamos como obter um valor aproximado para a média.

Reportando-nos ainda ao ficheiro *Idade.xls*, consideremos a tabela de frequências que serviu para agrupar os dados. Para obter um valor aproximado para a média, procedemos da seguinte forma:

- i) Adicionar à tabela de frequências uma nova coluna com os pontos médios dos intervalos de classe, que se obtêm fazendo a semi-soma dos limites dos intervalos (células S4 a S11);
- ii) Adicionar à tabela uma nova coluna com os produtos dos pontos médios dos intervalos de classe, pelas frequências relativas respectivas (células T4 a T11);
- iii) Somar os resultados das células T4 a T11 (célula T12):

	C	M	N	O	P	Q	R	S	T
1	Idade								
2	53	Classes							
3	32	Limite inferior	Limite superior		Classes	Freq.Abs.	Freq.Rel.	Ponto médio	
4	61	28	33,7	c1	[28,0; 33,7[19	0,083	30,85	2,548
5	51	33,7	39,4	c2	[33,7; 39,4[29	0,126	36,55	4,608
6	48	39,4	45,1	c3	[39,4; 45,1[39	0,170	42,25	7,164
7	56	45,1	50,8	c4	[45,1; 50,8[38	0,165	47,95	7,922
8	50	50,8	56,5	c5	[50,8; 56,5[50	0,217	53,65	11,66
9	53	56,5	62,2	c6	[56,5; 62,2[38	0,165	59,35	9,806
10	44	62,2	67,9	c7	[62,2; 67,9[10	0,043	65,05	2,828
11	39	67,9	73,6	c8	[67,9; 73,6[7	0,030	70,75	2,153
12	37					230	1		48,69

Repare-se que o valor obtido de 48,69 para a média, é muito próximo do verdadeiro valor obtido com os dados originais.

3.2.2 – Mediana

Outra medida de localização do centro dos dados é a *mediana*. Ordenados os elementos da amostra, a mediana, m , é o valor (pertencente ou não à amostra) que a divide ao meio, isto é, 50% dos elementos da amostra são menores ou iguais a m e os restantes 50% são maiores ou iguais a m . Em Excel, determina-se a mediana através da função *MEDIAN()*, que retorna a mediana dos seus argumentos, que podem ser números ou endereços de células.

Exemplo 3.2.3 – Calcule a mediana das idades dos deputados. Compare com o valor obtido para a média e diga o que poderia concluir da forma como os dados se distribuem.

Voltando ao ficheiro *Idade.xls*, utilizado no exemplo anterior, insira na célula R15 a função *Median(C2:C231)* e obterá como retorno, o valor 50, como se verifica na figura seguinte.

O valor obtido para a mediana é ligeiramente superior ao da média, pelo que podemos admitir que a distribuição é aproximadamente simétrica, com um ligeiro enviesamento para a esquerda.

Se os dados se apresentarem agrupados, já vimos na secção 3.2.2 do capítulo 2, um processo de obter a mediana através da função cumulativa. No entanto, não é necessário construir esta função para obter um valor aproximado para a mediana, pois este pode ser obtido a partir da tabela de frequências, utilizando ainda o processo de interpolação.

Exemplo 3.2.4 – A partir do agrupamento considerado, no exemplo 2.3.3, para a variável idade, calcule um valor aproximado para a mediana.

Adicionando à tabela de frequências uma nova coluna com as frequências relativas acumuladas, verificamos que a mediana se encontra na classe $[45,1; 50,8[$, pois a frequência acumulada de 50% é atingida nesta classe:

	C	D	O	P	Q	R	U	V
1	Idade							
2	53							
3	32			Classes	Freq.Abs.	Freq.Rel.	Freq.rel.acum.	
4	61	c1	[28,0; 33,7[19	0,083	0,083	
5	51	c2	[33,7; 39,4[29	0,126	0,209	
6	48	c3	[39,4; 45,1[39	0,170	0,378	
7	56	c4	[45,1; 50,8[38	0,165	0,543	
8	50	c5	[50,8; 56,5[50	0,217	0,761	
9	53	c6	[56,5; 62,2[38	0,165	0,926	
10	44	c7	[62,2; 67,9[10	0,043	0,970	
11	39	c8	[67,9; 73,6[7	0,030	1,000	
12	37				230	1		
13	37							
14	41				Média=	48,7		
15	40				Mediana=	50		

Admitindo que a frequência se distribui uniformemente sobre a amplitude de classe, isto é, a frequência 0,165 se distribui uniformemente sobre o intervalo de amplitude 5,7, resolvendo a equação de proporcionalidade

$$\frac{0,165}{0,122} = \frac{5,7}{x} \quad x = \frac{0,122 \times 5,7}{0,165} = 4,2$$

onde $0,122=0,5-0,378$, obtemos para a mediana o valor aproximado $45,1 + 4,2 = 49,3$.

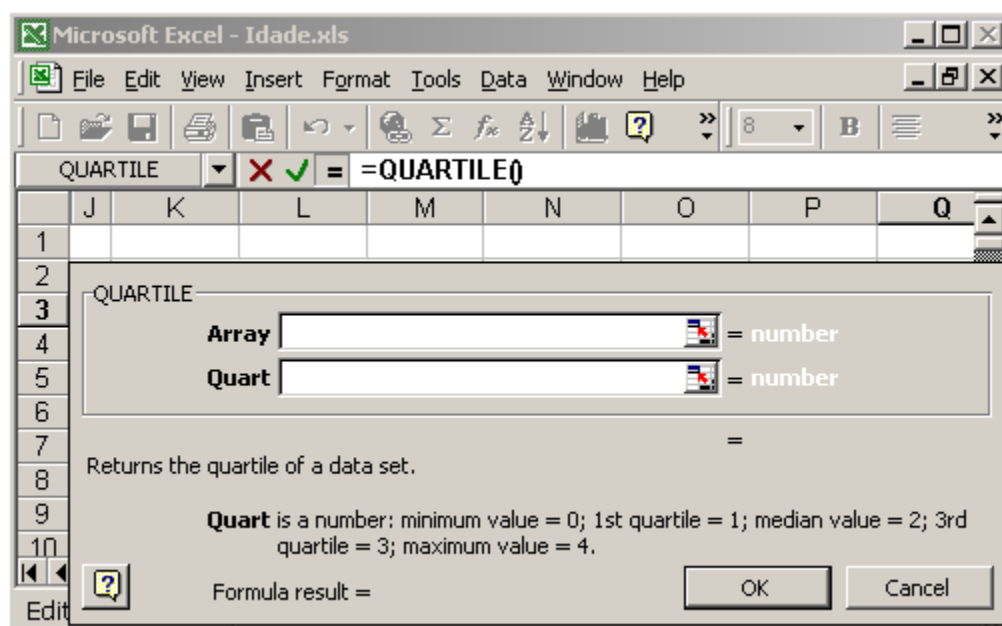
Chamamos a atenção para o seguinte facto: o valor (aproximado) que se obtém para a mediana, depende do agrupamento que se fizer para os dados, pelo que agrupamentos diferentes darão origem a valores diferentes, embora não difiram muito uns dos outros (Lembramos que o valor da mediana apresentado na figura anterior foi obtido a partir dos dados não agrupados).

3.2.3 – Quartis

Os quartis, 1.º e 3.º, definem-se de forma idêntica à mediana, mas considerando em vez da percentagem de 50%, respectivamente 25% para o 1º quartil, Q1, e 75% para o 3.º quartil, Q3.

Há vários processos para a determinação dos quartis, nem sempre conduzindo aos mesmos resultados. Este facto não é preocupante, pois de um modo geral nas situações que têm interesse em estatística, as amostras têm dimensão suficientemente elevada de forma que os diferentes processos conduzem a valores próximos.

Em Excel a determinação dos quartis faz-se utilizando a função *QUARTILE(array;quart)*:



Repare que a função $Quartile(array; quart)$ tem dois argumentos, em que o primeiro argumento é o endereço das células de que queremos calcular o quartil e o segundo argumento pode tomar vários valores, conforme a medida de localização, de entre as seguintes, que nos interesse calcular:

- 0 – mínimo
- 1 – 1º quartil
- 2 – mediana
- 3 – 3º quartil
- 4 – máximo

Assim, esta função, além do 1.º e 3.º quartis, a que estão associadas as percentagens 25% e 75%, respectivamente, ainda calcula a mediana, a que está associada a percentagem de 50% e o mínimo e máximo com percentagens associadas de 0% e 100%.

Exemplo 3.2.5 – Escolha os primeiros 15 elementos da variável Idade, do ficheiro *Idade.xls*. Obtenha o 1º e 3º quartis.

Os primeiros 15 elementos são os seguintes:

53 32 61 51 48 56 50 53 44 39 37 37 41 40 40

Utilizando a função $QUARTILE(C2:C16;1)$ e $QUARTILE(C2:C16;3)$, obtemos $Q_1=39,5$ e $Q_3=52$.

Se utilizar o processo que aprendeu no módulo de Estatística, nomeadamente considerando o 1.º quartil como a mediana da primeira parte da amostra, quando esta é dividida pela mediana, depois de ordenar a amostra e tendo em conta que a mediana é 44, temos para 1.º quartil o

32 37 37 39 40 40 41 **44** 48 50 51 53 53 56 61

valor 39, se não considerarmos a mediana como pertencente a nenhuma das partes, ou 39,5 se considerarmos a mediana pertencente às duas partes. Para o 3º quartil obteremos, respectivamente o valor 53 ou 52, utilizando a mesma metodologia.

Exemplo 3.2.5 (cont) – Repita o exemplo anterior, considerando amostras de dimensão 12 e 13.

Considere agora só os primeiros 12 elementos. Como a mediana é 49, o 1º quartil – mediana da 1ª parte da amostra, será $(37+39)/2=38$, enquanto que o 3º quartil será $(53+53)/2=53$.

32 37 37 39 44 48 50 51 53 53 56 61

Utilizando o Excel, os valores que se obtêm são $Q_1=38,5$ e $Q_3=53$.

Considere agora os primeiros 13 elementos. Como a mediana é 48, o 1º quartil – mediana da 1ª parte da amostra, será $(37+39)/2=38$, enquanto que o 3º quartil será $(53+53)/2=53$, não considerando a mediana como pertencente a nenhuma das partes. Caso contrário, teremos $Q_1=39$ e $Q_3=53$.

32 37 37 39 41 44 **48** 50 51 53 53 56 61

Utilizando o Excel, os valores que se obtêm são $Q_1=39$ e $Q_3=53$.

Observação: Repare que os valores que se obtêm para os quartis, recorrendo ao excel não são iguais aos que se obtiveram sem utilizar o Excel. Efectivamente não existe uniformidade na forma de calcular os quartis, como já havíamos referido anteriormente, embora os resultados obtidos satisfaçam a definição de quartis. Exemplificando com a mediana, repare que pela definição de mediana, quando o número de elementos da amostra é par, podemos considerar para mediana qualquer valor compreendido entre os dois elementos médios da amostra ordenada! Não é costume deixar esta opção ao critério de cada um e considera-se a semi-soma desses elementos médios.

Voltando aos quartis, pode verificar que, no Excel, o 1.º quartil corresponde à observação de ordem $(n+3)/4$, procedendo-se a uma interpolação, quando necessário (Sugestão – Tente descobrir como é calculado o 3º quartil no Excel).

3.3 – Medidas de dispersão

Continuando na mesma linha de apresentação das medidas de localização, também agora não nos vamos preocupar com as propriedades das medidas de dispersão, pois admitimos que estas já foram estudadas no módulo de Estatística. Debruçar-nos-emos sobre o seu cálculo, utilizando o Excel.

A seguir apresentaremos o cálculo da variância, desvio padrão e amplitude inter-quartil.

3.3.1 – Variância e desvio-padrão

A variância de um conjunto de dados obtém-se fazendo a média dos quadrados dos desvios dos dados, relativamente à média.

O Excel, tal como as máquinas de calcular, dispõe de duas funções para calcular a variância, conforme estejamos a calcular a variância populacional (parâmetro) ou a variância amostral (estatística).

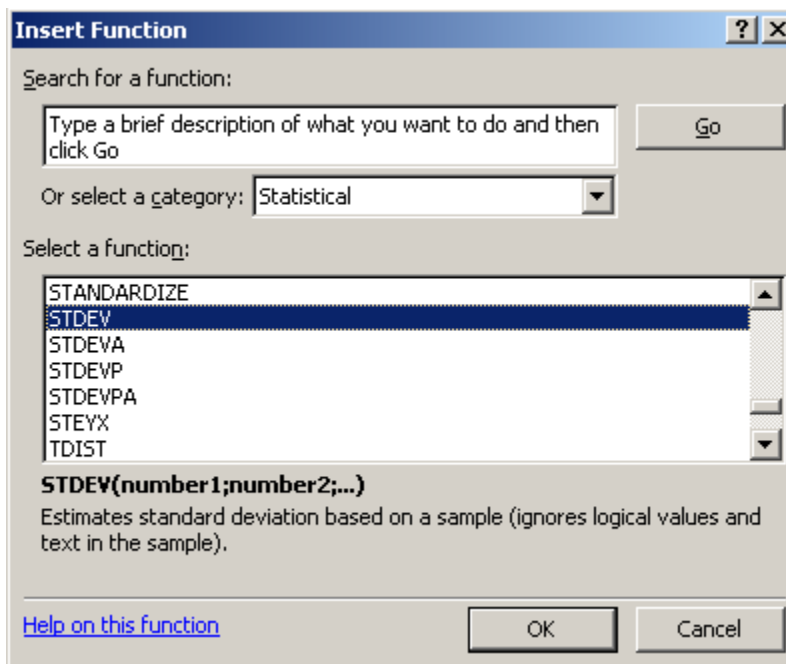
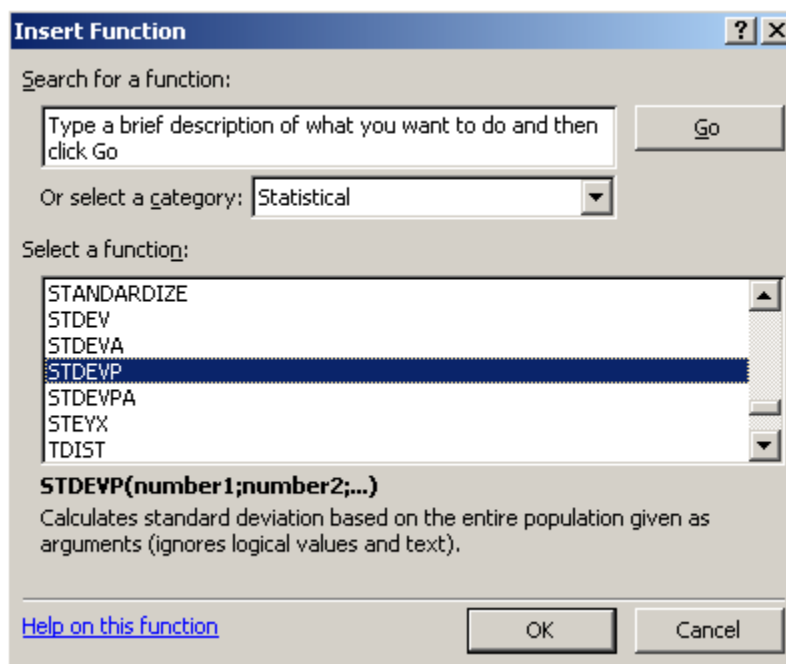
Resumimos no quadro seguinte a situação de estarmos a calcular parâmetros ou estatísticas.

População de N elementos X_1, X_2, \dots, X_N	Amostra de n elementos X_1, X_2, \dots, X_n
Valor médio $\mu = \frac{X_1 + X_2 + \dots + X_N}{N}$	Média $\bar{x} = \frac{X_1 + X_2 + \dots + X_n}{n}$
Variância populacional $\sigma^2 = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_N - \mu)^2}{N}$	Variância amostral $s^2 = \frac{(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + \dots + (X_n - \bar{x})^2}{n - 1}$
Desvio padrão populacional σ	Desvio padrão amostral s

Em Excel as funções utilizadas para calcular a variância populacional e amostral, são respectivamente *VARP()* e *VAR()*. Como argumento utiliza-se a sequência de números de que se quer calcular a variância, ou o endereço das células que os contêm.

Por exemplo, no caso da população dos deputados, que temos vindo a estudar, temos informação completa sobre a variável Idade, pelo que a fórmula que deve ser utilizada para obter a variância é a *VARP*, isto é, esta fórmula dá-nos a variância populacional. Se só dispuséssemos da idade de alguns deputados, isto é, uma amostra da população em estudo, então a fórmula a utilizar seria a *VAR*, que dá a variância amostral. A maneira de calcular as duas variâncias é idêntica, diferindo unicamente no seguinte ponto: enquanto que no caso da variância populacional se divide a soma dos quadrados dos desvios pelo número de parcelas, no caso da variância amostral divide-se a soma dos quadrados dos desvios pelo número de parcelas menos uma.

O desvio padrão obtém-se fazendo a raiz quadrada da variância ou utilizando uma função própria. Como é evidente, existem também duas fórmulas para o calcular, obtendo-se o desvio padrão populacional ou amostral, conforme a fórmula utilizada:



Repare-se que quando se selecciona a função que se quer utilizar, aparece a descrição do que é que a função faz.

Exemplo 3.3.1 – A partir do ficheiro *Idade.xls*, seleccione uma amostra aleatória simples de dimensão 40. Calcule a variância e o desvio padrão da amostra obtida. Calcule de seguida a

variância da população constituída pelas idades dos 230 deputados e compare com a variância da amostra obtida anteriormente.

Utilizando o processo descrito em 1.3.1.2, seleccionámos uma amostra de 40 elementos que posteriormente colocámos nas células A2 a D11, de uma nova folha de Excel. Colocando agora o cursor na célula onde pretendemos colocar a variância, por exemplo na célula F4, inserimos a função *VAR (A2:D11)* e a função retorna um valor aproximadamente igual a 112, para a variância da amostra.

Para calcular a variância da população das idades, inserimos na célula F5 a função *VARP(Sheet1!C2:C231)*, obtendo-se um valor aproximadamente igual a 101:

	A	B	C	D	E	F
1		Amostra				
2	66	59	34	45		
3	42	35	50	49		
4	62	33	39	55	Var. amostral=	112,20
5	57	56	54	46		
6	59	59	37	37		
7	40	50	38	42		
8	48	58	64	40		
9	41	33	31	49		
10	59	69	28	55		
11	42	46	51	55		
12						
13					Var. populacional=	100,73

Comparando as variâncias, vemos que não são iguais, o que já seria de esperar, uma vez que a variância amostral foi obtida a partir de 40 dos 230 dados e é uma estimativa da variância populacional. Se recolhermos outra amostra, também de 40 elementos, não esperamos obter o mesmo valor para a estimativa. Esperamos sim, obter valores aproximados.

Para calcular o desvio padrão, ou se calcula a raiz quadrada (positiva) do valor da variância, ou se utilizam as funções *STDEV()* ou *STDEVP()*, conforme se pretenda o desvio padrão amostral ou populacional. No nosso caso os desvios padrões amostral e populacional vêm, respectivamente, aproximadamente iguais a 10,6 e 10,0.

3.3.2 – Amplitude e amplitude interquartis

A amplitude da amostra (não confundir com dimensão da amostra), *R*, é a medida mais simples para medir a variabilidade, mas tem a grande desvantagem de ser muito sensível à existência na amostra, de uma observação muito pequena ou muito grande. Não existe, no Excel, uma função específica para a calcular, recorrendo-se às funções *MAX()* e *MIN()*. Já tivemos, aliás,

oportunidade de utilizar estas funções quando necessitamos de calcular a amplitude de um conjunto de dados, para iniciar a construção de um histograma, com classes de igual amplitude.

Uma medida mais resistente do que a anterior, é a amplitude interquartis que, como o nome indica, se define como a diferença entre os 1.º e 3.º quartis.

Exemplo 3.3.2 – Calcule a amplitude e a amplitude interquartis da amostra obtida no exemplo anterior.

Como os elementos da amostra se encontram nas células A2 a D11, temos:

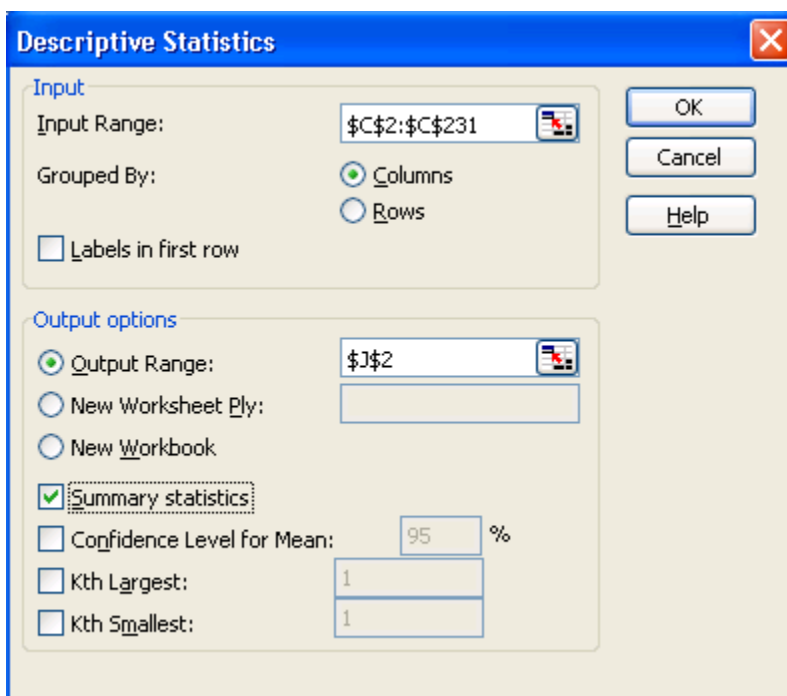
$$R = \text{MAX}(A2:D11) - \text{MIN}(A2:D11) = 69-28 = 41$$

Recorrendo à terminologia usada quando definimos os quartis, temos:

$$\text{Amplitude interquartis} = \text{QUARTILE}(A2:D11;3) - \text{QUARTILE}(A2:D11;1) = 56,25-39,75=16,5.$$

3.4 – Função Descriptive Statistics

O Excel dispõe de uma função a que se acede seleccionando Tools →Data Aalysis→Descriptive Statistics →OK



e cujo resultado é o que se apresenta a seguir:

	C	I	J	K
2	53		<i>Column1</i>	
3	32			
4	61		Mean	48,66
5	51		Standard Error	0,66
6	48		Median	50,00
7	56		Mode	50,00
8	50		Standard Deviation	10,06
9	53		Sample Variance	101,17
10	44		Kurtosis	-0,72
11	39		Skewness	-0,05
12	37		Range	45
13	37		Minimum	28
14	41		Maximum	73
15	40		Sum	11191
16	40		Count	230
17	33			

Algumas das funções já são conhecidas das secções anteriores. Chamamos a atenção para o facto de a variância das 230 idades não coincidir com o valor obtido na secção 3.3.1, uma vez que quando se considera um conjunto de dados e se pedem as Estatísticas descritivas, subentende-se que se está perante uma amostra e não da população toda! Por esta razão, a fórmula utilizada para o cálculo da variância é a da variância amostral.

As funções Standard Error, Kurtosis e Skewness saem fora do âmbito destas folhas, pelo que não entraremos em detalhe.