Afinal o que são e como se calculam os quartis?

Susana Fernandes Mónica Pinto

Universidade do Algarve
Departamento de Matemática

Introdução

Imaginemos que queremos calcular os quartis de um conjunto de dados. Se consultarmos vários manuais de matemática do ensino básico e do ensino secundário e sebentas de estatística do ensino superior encontramos diferentes definições e fórmulas para o cálculo dos quartis, que algumas vezes conduzem a resultados diferentes. Se recorrermos a meios tecnológicos para determinar o valor dos quartis do mesmo conjunto de dados, usando uma máquina de calcular, uma folha de cálculo ou um programa estatístico, pode acontecer ficarmos com mais algumas respostas diferentes. Afinal o que são e como se calculam os quartis? Neste texto abordaremos esta questão, considerando apenas dados discretos não agrupados. Salientamos que no ensino superior se ensinam formas de cálculo dos quartis distintas das apresentadas nos ensinos básico e secundário, o que confunde os alunos e gera neles uma certa desconfiança relativamente à estatística. Argumentamos que o método introduzido no ensino superior é preferível e apresentamos uma proposta de uniformização do cálculo dos quartis em todos os níveis de ensino.

Quartis - definição simples

Uma forma simples, mas pouco rigorosa, de definir os quartis é

Definição 1: Quartis são os valores que dividem um conjunto de dados em quatro partes iguais.

Uma vez ordenado o conjunto de dados, o segundo quartil (Q_2 - também conhecido como mediana) é o valor que fica a meio dos valores dos elementos do conjunto de dados, isto é, o valor que divide o conjunto de dados em duas partes iguais (metades). Depois o primeiro quartil (Q_1) será o valor que fica a meio da primeira metade do conjunto de dados e o terceiro quartil (Q_3) será, analogamente, o valor que fica a meio da segunda metade do conjunto de dados. Consideremos por exemplo o conjunto de dados (já ordenados)

$$\{x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 10, x_5 = 14, x_6 = 18, x_7 = 21, x_8 = 25, x_9 = 29, x_{10} = 32\}.$$

Ao todo temos 10 elementos logo o meio dos valores estará entre os quinto e sexto elementos, isto é, entre $x_5 = 14$ e $x_6 = 18$. O ponto médio entre 14 e 18 é 16 logo $Q_2 = 16$. Valor este que divide o conjunto de dados em duas metades: $\{x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 10, x_5 = 14\}$ e $\{x_6 = 18, x_7 = 21, x_8 = 25, x_9 = 29, x_{10} = 32\}$. Agora o primeiro quartil será o valor que fica a meio do primeiro subconjunto. Este subconjunto tem um número ímpar de elementos logo o valor que fica no meio é um elemento do subconjunto - o elemento $x_3 = 6$, logo $Q_1 = 6$. Da mesma forma o terceiro quartil será o valor que fica a meio do subconjunto $\{x_6 = 18, x_7 = 21, x_8 = 25, x_9 = 29, x_{10} = 32\}$ e mais uma vez, como o número de elementos deste subconjunto é ímpar, o valor que fica no meio é $x_8 = 25$, logo $Q_3 = 25$. Consideremos agora o conjunto de dados $\{x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 10, x_5 = 14, x_6 = 18, x_7 = 21, x_8 = 25, x_9 = 29\}$ com 9 elementos (usamos a letra n para representar o número de elementos do conjunto, neste caso $n = 9$). Como o número de elementos é ímpar o segundo quartil será o elemento central, isto é $Q_2 = x_5 = 14$. Até aqui tudo bem. Mas agora como dividimos o conjunto de dados em duas metades? O elemento x_5 deve ser incluído em ambas as metades, em nenhuma das duas ou apenas numa delas? Esta incerteza origina diferentes métodos de cálculo dos quartis. Vejamos dois deles.

Método inclusivo: quando o conjunto de dados tem um número ímpar de elementos, o elemento correspondente ao Q_2 é incluído em ambas as metades do conjunto de dados para cálculo dos Q_1 e Q_3.

Usando este método as duas metades do conjunto de dados serão $\{x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 10, x_5 = 14\}$ e $\{x_5 = 14, x_6 = 18, x_7 = 21, x_8 = 25, x_9 = 29\}$ e então teremos $Q_1 = x_3 = 6$ e $Q_3 = x_7 = 21$. Repare-se que tendo o conjunto de dados n elementos, desta forma cada uma das metades terá $(n + 1)/2$ elementos.

Método exclusivo: quando o conjunto de dados tem um número ímpar de elementos, o elemento correspondente ao Q_2 não é incluído em nenhuma das metades do conjunto de dados para cálculo dos Q_1 e Q_3.

Usando este método as duas metades do conjunto de dados serão $\{x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 10\}$ e $\{x_6 = 18, x_7 = 21, x_8 = 25, x_9 = 29\}$ e então teremos $Q_1 = (x_2 + x_3)/2 = 4.5$ e $Q_3 = (x_7 + x_8)/2 = 23$. Repare-se que tendo o conjunto de dados n elementos, desta forma cada uma das metades terá $(n - 1)/2$ elementos.

O processo de determinação de cada quartil inclui dois passos: primeiro determinar a posição do quartil no conjunto de dados; segundo calcular o valor do quartil. Quando o quartil coincide com um elemento do conjunto de dados dizemos que a sua posição é um valor inteiro k, e neste caso o valor do quartil é imediato. Por exemplo, para o conjunto de dados inicial com 10 elementos a posição de Q_1 é $k = 3$ logo o seu valor é $Q_1 = x_3 = 6$. Quando o quartil fica entre dois elementos dizemos que a sua posição é um valor não inteiro. Ainda neste conjunto de dados com $n = 10$ a posição de Q_2 é entre os elementos x_5 e x_6 pelo que dizemos que a sua posição é $k = 5.5$ e neste caso é necessário calcular o seu valor fazendo a semi-soma dos valores dos elementos nas posições 5 e 6, isto é, $Q_2 = (x_5 + x_6)/2 = (14 + 18)/2 = 16$.

A posição dos quartis pode também ser calculada matematicamente. Todos os métodos determinam a posição da mediana (segundo quartil) da mesma forma. Considere-se um conjunto de dados ordenado com n elementos $x_1 \leq x_2 \leq x_3 \leq .... \leq x_n$.

A posição de Q_2 num conjunto de dados com n elementos é $(n+1)/2$.

Já a forma de determinar a posição dos Q_1 e Q_3 no conjunto de dados varia consoante o método utilizado. A Tabela 1 apresenta as fórmulas para cálculo das posições dos primeiro e terceiro quartis correspondentes aos métodos inclusivo e exclusivo.

n par		n ímpar		
Q_1	Q_3	Q_1	Q_3	
$k = \frac{n+2}{4}$	$k = \frac{3n+2}{4}$	$k = \frac{n+3}{4}$	$k = \frac{3n+1}{4}$	Método Inclusivo
		$k = \frac{n+1}{4}$	$k = \frac{3n+3}{4}$	Método Exclusivo

Tabela 1: Fórmulas para cálculo das posições de Q_1 e Q_3

Uma vez conhecida a posição dos quartis, a forma de cálculo do seu valor é igual em ambos os métodos. A tabela 2 apresenta a forma de calcular um quartil Q_p, dada a sua posição k no conjunto de dados.

k inteiro	$Q_p = x_k$
k não inteiro $i < k < i + 1$	$Q_p = \frac{x_i + x_{i+1}}{2}$

Tabela 2: Fórmulas para cálculo do valor de Q_p dada a sua posição k

Todos os manuais do ensino básico consultados ensinam a determinar os quartis usando o método exclusivo, de acordo com as indicações do Ministério da Educação e Ciência [1]. Começam por definir a mediana como o centro do conjunto de dados ordenado e depois definem o 1^{o} quartil como sendo a mediana dos dados que ficam à esquerda da mediana do conjunto de dados e o 3^{o} quartil como sendo a mediana dos dados que ficam à direita da mediana do conjunto de dados. Os manuais do ensino secundário consultados ensinam também o método exclusivo (embora a brochura de estatística do 10^{o} ano faça referência aos métodos inclusivo e exclusivo [4]), apresentando as fórmulas para determinação das posições dos quartis.

Existem muitos métodos para o cálculo dos quartis, diferindo quer no primeiro passo do processo (determinação da posição dos quartis) quer no segundo passo do processo (cálculo do valor dos quartis). Quando a posição dos quartis no conjunto de dados não é um inteiro, os métodos inclusivo e exclusivo calculam a semi-soma dos valores dos elementos do conjunto de dados mais próximos da posição do quartil. Métodos há que arredondam a posição do quartil, sendo este sempre igual a um elemento do conjunto de dados, e existem métodos com formas de arredondamentos distintas. Outros métodos optam por fazer interpolação dos valores dos elementos mais próximos da posição do quartil. Por exemplo, imaginemos que a posição do quartil é 2.85, fazendo interpolação o valor do quartil seria dado por $0.85 \times x_2 + 0.15 \times x_3$. O leitor interessado encontrará no artigo [3] a descrição de 15 métodos para o cálculo de quartis, assim como as respectivas referências.

Mas então qual será o melhor método para calcular os quartis? Depende de qual o uso que deles queremos fazer.

Quartis vistos como estimadores

No ensino superior interessa-nos olhar para os dados como valores observados de uma dada população e, neste sentido, o conjunto de dados é uma amostra de observações e as medidas calculadas com base na amostra são vistas como estimativas de parâmetros da população subjacente (assume-se que a amostra é aleatória e representativa da população em estudo - neste texto não abordaremos o assunto da recolha/construção de uma amostra válida). Ora a forma de calcular os quartis nos ensino básico e secundário, embora muito intuitiva, não fornece um bom estimador para o parâmetro da população correspondente (ver por exemplo [2] pag. 87). Para perceber de que parâmetros falamos interessa aqui introduzir o conceito de percentil (ou quantil). Considere-se X a variável aleatória discreta que representa a característica da população em estudo. Percentil populacional de proporção p (ou de percentagem $100p\%$) é o valor P_p tal que $P(X \leq P_p) \geq p$ e $P(X \geq P_p) \geq 1 - p$. O primeiro quartil é pois o percentil de proporção 0.25, isto é, é o valor $P_{0.25}$ tal que a probabilidade da variável X tomar um valor não superior a $P_{0.25}$ é pelo menos 0.25 e simultaneamente a probabilidade de a variável X tomar um valor não inferior a $P_{0.25}$ é pelo menos 0.75. Analogamente, o terceiro quartil é o percentil de proporção 0.75 (75%) e o segundo quartil (mediana) é o percentil de proporção 0.5 (50%).

Quem nunca ouviu a mãe de uma criança pequena comentar "O meu filho está no percentil 95 da altura"? O que isto significa é que, considerando todas as crianças portuguesas com a mesma idade do filho da senhora, pelo menos 95% dessas crianças terão uma altura não superior à altura do filho da senhora e simultaneamente pelo menos 5% dessas crianças terão uma altura não inferior à altura do filho da senhora. Dada uma amostra representativa da população descrita pela variável aleatória X estimamos a probabilidade da variável assumir um valor não superior a determinado valor x pela proporção de valores não superiores a x na amostra, isto é, estimamos a função de distribuição de X pela função de distribuição cumulativa dos valores da amostra. Assim uma definição mais rigorosa e geral para todos os percentis amostrais será:

Definição 2: Percentil amostral de proporção p - P_p - é o valor tal que a proporção de valores da amostra não superiores a P_p é pelo menos p e simultaneamente a proporção de valores da amostra não inferiores a P_p é pelo menos $1 - p$.

Desta definição surge um novo método para o cálculo dos percentis de uma amostra que muitas vezes se designa por método *CDF*, do inglês, *"cumulative distribution function"*. O leitor interessado encontrará no artigo [3] a demonstração de que o método *CDF* produz sempre um percentil de acordo com esta definição.

Método CDF: dada uma amostra com n observações, se np é um valor inteiro então $P_p = (x_{np} + x_{np+1})/2$; se np não é um valor inteiro, seja k a parte inteira de np, então $P_p = x_{k+1}$.

Este método produz um bom estimador para o percentil populacional mas tem a desvantagem de não ser nada intuitivo. Os métodos inclusivo e exclusivo referidos anteriormente, embora intuitivos, além de não produzirem bons estimadores, têm a desvantagem de não serem generalizáveis ao cálculo de outros percentis.

Proposta para uniformização do cálculo dos quartis em todos os ciclos de ensino

Como dissemos na introdução deste texto, o facto de ensinarmos ao longo dos vários ciclos de ensino formas diferentes de calcular os quartis que em certas situações conduzem a valores distintos, confunde e é desconfortável para os alunos, o que em nada contribui para a aquisição do conceito e dá origem a uma certa desconfiança relativamente à estatística. Argumentamos que tal é evitável e propomos uma forma mais intuitiva do método *CDF*, passível de ser ensinada nos ciclos básico e secundário do ensino português. Voltemos à noção intuitiva de que os quartis dividem uma amostra com n observações em 4 partes iguais. Ora ao fazer a divisão inteira de n por 4 ou o resto dá 0, ou dá 1, ou dá 2 ou dá 3. Isto é, dado n o número de observações da amostra, existe um m inteiro não negativo tal que ou $n = 4m$, ou $n = 4m+1$, ou $n = 4m+2$, ou $n = 4m+3$. Podemos então determinar as fórmulas para a posição dos primeiro e terceiro quartis segundo o método *CDF* de acordo com o resto da divisão inteira de n por 4 da forma explicitada na Tabela 3:

	$n = 4m$	$n = 4m + 1$	$n = 4m + 2$	$n = 4m + 3$
Q_1	$k = \frac{n+2}{4}$	$k = \frac{n+3}{4}$	$k = \frac{n+2}{4}$	$k = \frac{n+1}{4}$
Q_3	$k = \frac{3n+2}{4}$	$k = \frac{3n+1}{4}$	$k = \frac{3n+2}{4}$	$k = \frac{3n+3}{4}$

Tabela 3: Fórmulas para cálculo das posições de Q_1 e Q_3 com o método *CDF*

Note-se que as fórmulas são iguais nos casos em que o resto da divisão inteira de n por 4 é 0 ou 2, isto é, quando o n é par. Alternativamente podemos organizar a tabela para n par e ímpar, considerando dois casos para n ímpar - um quando o resto é 1 e outro quando o resto é 3, como mostra a Tabela 4:

	n par	n ímpar	
		$n = 4m + 1$	$n = 4m + 3$
Q_1	$k = \frac{n+2}{4}$	$k = \frac{n+3}{4}$	$k = \frac{n+1}{4}$
Q_3	$k = \frac{3n+2}{4}$	$k = \frac{3n+1}{4}$	$k = \frac{3n+3}{4}$

Tabela 4: Fórmulas para cálculo das posições de Q_1 e Q_3 com o método *CDF* - apresentação alternativa

Note-se que a posição da mediana (segundo quartil) continua a ser dada por $(n+1)/2 = (2n+2)/4$, para todos os valores de n.

Uma vez conhecida a posição dos quartis, a forma de cálculo do seu valor é igual ao apresentado para os métodos inclusivo e exclusivo na Tabela 2. Quando a posição do quartil é um número inteiro o seu valor é igual ao da observação nessa posição na amostra; no caso de a posição do quartil não dar um número inteiro o valor do quartil será dado pela semi-soma das observações mais próximas dessa posição não inteira.

Repare-se que no caso em que $n = 4m+1$ as fórmulas para o cálculo da posição dos quartis com o método *CDF* coincidem com as fórmulas do método inclusivo e, no caso em que $n = 4m+3$, as fórmulas do método *CDF* coincidem com as fórmulas do método exclusivo. Assim uma forma intuitiva de apresentar o método *CDF* ao alunos do ensino básico será o de, definindo os quartis como as medianas das metades inferior e superior da amostra, apresentar a problemática que surge no caso de n ser ímpar: incluir ou não a mediana da amostra nas duas metades a considerar para cálculo dos quartis? e indicar o uso do método inclusivo quando n a dividir por 4 dá resto 1 e o uso do método exclusivo quando n a dividir por 4 dá resto 3.

No artigo [3] o autor sugere uma forma intuitiva de aplicar o método *CDF* sem ser necessário recorrer a fórmulas para determinar a posição dos quartis. O autor sugere que definindo os quartis como as medianas das metades inferior e superior da amostra e apresentada a problemática de incluir ou não a mediana da amostra nas duas metades no caso de n ser ímpar, se apresente a solução de, quando n ímpar, incluir ou não a mediana da amostra nas metades inferior e superior da amostra por forma a que o número de elementos das metades também seja ímpar.

Assim por exemplo para uma amostra com $n = 5$ observações (note-se que 5 a dividir por 4 dá resto 1), por exemplo $\{x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 10, x_5 = 14\}$, a mediana será $Q_2 = x_3 = 6$ e as metades a considerar irão incluir a mediana sendo $\{x_1 = 1, x_2 = 3, x_3 = 6\}$ e $\{x_3 = 6, x_4 = 10, x_5 = 14\}$, o que conduzirá a $Q_1 = x_2 = 3$ e $Q_3 = x_4 = 10$. Já para uma amostra com por exemplo 7 observações (note-se que 7 a dividir por 4 dá resto 3), por exemplo $\{x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 10, x_5 = 14, x_6 = 18, x_7 = 21\}$ a mediana será $Q_2 = x_4 = 10$ e as metades a considerar não incluirão a mediana sendo $\{x_1 = 1, x_2 = 3, x_3 = 6\}$ e $\{x_5 = 14, x_6 = 18, x_7 = 21\}$, o que conduzirá a $Q_1 = x_2 = 3$ e $Q_3 = x_6 = 18$.

Considerações finais

Ainda que o método *CDF*, que quanto a nós é o método mais indicado para calcular quartis de amostras (com poucos elementos repetidos), venha a ser o método de cálculo dos quartis adoptado em todos os ciclos de ensino, as máquinas de calcular, as folhas de cálculo e os programas de estatística continuam a calcular os quartis por outros métodos, que muitas vezes produzirão valores diferentes. É pois por isso importante que o professor esteja ciente deste facto quando decidir usar algum destes recursos. É também importante que desde muito cedo se apresente aos alunos a noção de que ao estudar amostras de observações de uma população com o objectivo de inferir informação sobre toda a população estamos inevitavelmente a correr o risco de errar, risco esse que pretendemos quantificar (e minimizar). E como existem diferentes formas de quantificar o erro existem também métodos diferentes de abordar as questões (por forma a minimizar o erro quantificado).

Referências

[1] A. Bivar, C. Grosso, F. Oliveira, M.C. Timóteo, *Metas Curriculares do Ensino Básico - Matemática*, Direção-Geral da Educação, Ministério da Educação e Ciência, 2013.
portal da DCE www.dgidc.min-edu.pt/index.php?s=noticias&noticia=396

[2] A. Bivar, C. Grosso, F. Oliveira, M.C. Timóteo, *Metas Curriculares do Ensino Básico - Matemática, Caderno de Apoio 3 º Ciclo*, Direção-Geral da Educação, Ministério da Educação e Ciência, 2013.
portal da DCE www.dgidc.min-edu.pt/index.php?s=noticias&noticia=396

[3] E. Langford, "Quartiles in Elementary Statistics", *Journal of Statistics Education*, Vol. 14, No. 3, (2006). www.amstat.org/publications/jse/v14n3/langford.html

[4] M.E.G. Martins, C. Monteiro, J.P. Viana, M. A. A. Turkman, *Brochuras de Matemática para o Secundário - ESTATÍSTICA 10 º Ano*, Ministério da Educação, Departamento do Ensino Secundário, 1997.
portal da DCE www.dgidc.min-edu.pt/outrosprojetos/index.php?s=directorio&pid=148