N.º 25 – HISTOGRAMA

Por: Maria Eugénia Graça Martins
Departamento de Estatística e Investigação Operacional da FCUL
memartins@fc.ul.pt

Emília Oliveira
Escola Secundária de Tomaz Pelayo
ecmo.estp@gmail.com

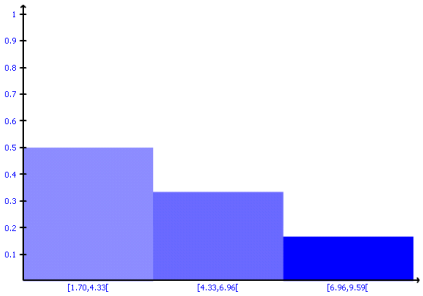
O HISTOGRAMA é um gráfico, formado por uma sucessão de rectângulos adjacentes, tendo cada um por base um intervalo de classe e uma área igual (ou proporcional) à frequência relativa (ou absoluta) dessa classe.

Ao contrário do gráfico de barras, em que estas estão separadas e em que a altura de cada barra é o mais relevante, no histograma as barras (rectângulos) estão juntas e o que é importante é a área de cada uma.

Contudo, para se proceder à sua construção, é necessário começar por organizar os dados na forma de uma tabela de frequências.

Assim, nesta ficha são dadas inicialmente algumas orientações para a organização dos dados em classes, número de classes a considerar, bem como o seu processo de construção, procedimentos prévios à construção da tabela de frequências. De seguida, e utilizando exemplos, explica-se a construção do histograma.

Esta ActivALEA é acompanhada de uma aplicação interactiva[1] que permite a construção de histogramas com classes de igual amplitude ou com amplitudes diferentes. A apresentação da aplicação e das suas funcionalidades é feita em documento anexo.



Para construir uma tabela de frequências ou um histograma na folha de cálculo Excel, sugerimos uma consulta ao capítulo 2 do dossiê XVII Estatística Descritiva com Excel – Complementos, disponível em:
https://www.alea.pt/images/dossies_pdf/dossie13a.pdf

[1] Da autoria de Manuel Ramos (mjlr.estp@gmail.com).

1. Introdução

O **histograma** é a representação gráfica mais conhecida quando se pretende representar dados contínuos[2]. Contudo, para se proceder à sua construção, é necessário começar por organizar os dados na forma de uma tabela de frequências.

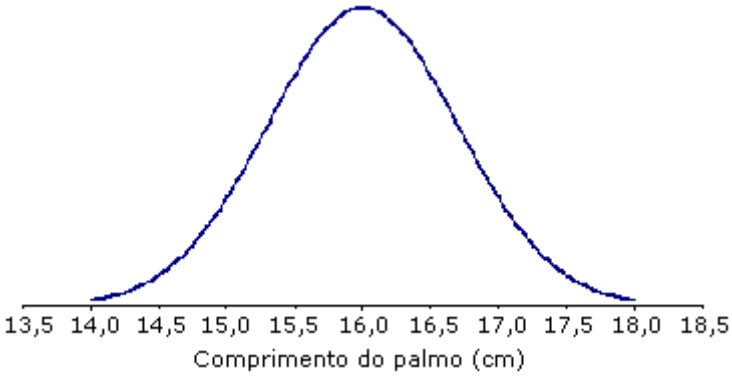
Dados contínuos são todos os que resultam de medições, ou seja, a variável em estudo pode ser medida com algum instrumento (régua, balança, relógio, termómetro, barómetro, etc.) e os dados são constituídos pelos valores resultantes das medições efectuadas.

Estas **variáveis** chamam-se **contínuas**, pois qualquer valor, dentro de um certo intervalo, pode ser obtido como resultado da medição.

Assim, uma característica comum a um conjunto de dados de natureza contínua é o facto de registarem poucos valores repetidos. A ocorrência de valores iguais com maior frequência do que a que se esperaria para dados contínuos deriva do facto de o instrumento de medida não ter uma grande precisão. Por exemplo, os valores resultantes da medição das variáveis contínuas como o *tempo que demora de casa à escola* e *comprimento do palmo* apresentam-se "discretizados" por uma limitação do instrumento que se utilizou para as medir. Outro exemplo de uma variável contínua que se apresenta "discretizada" é a *idade*. Quando se diz que um jovem tem 9 anos, significa que já fez os 9 anos, mas ainda não fez os 10, pelo que o 9 representa um intervalo de valores que se pode exprimir da seguinte forma: $9 \leq \text{idade} < 10$.

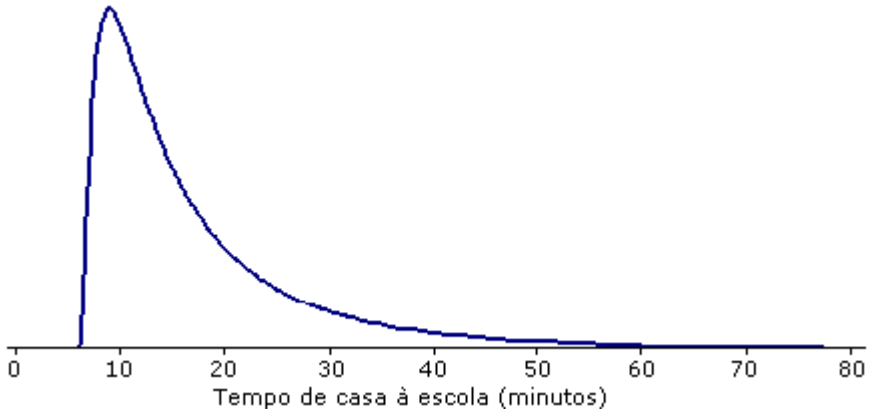
Mesmo existindo alguns valores iguais, o número de valores distintos pode ser tão grande que a metodologia utilizada para construir as tabelas de frequências de dados quantitativos discretos, em que se consideram para classes os valores distintos nos dados, não pode ser aqui utilizada. Correríamos o risco de a frequência observada para cada valor distinto ser 1! Então, a alternativa é considerar classes na forma de intervalos.

Ao organizar os dados na forma de intervalos, o nosso objectivo é visualizar o padrão subjacente a esses dados. Por exemplo, é natural esperar que uma forma usual para a distribuição da variável *comprimento do palmo*, medida num conjunto de alunos do 3º ciclo, tenha um aspecto simétrico, como o que se apresenta na figura seguinte, com uma concentração de valores em volta dos 16 cm e cada vez menos valores à medida que o comprimento para o palmo diminui ou aumenta.



Já para a variável *tempo de casa à escola* em que, de um modo geral, predominam os tempos mais pequenos, em detrimento dos tempos maiores, esperamos uma distribuição com uma forma enviesada, como a que se apresenta a seguir.

[2] Nesta introdução segue-se de perto as Brochuras: Graça Martins, M.E. et al – Análise de Dados, DGIDC, Ministério da Educação, 2007 e Graça Martins, M.E. et al – Organização e tratamento de dados, DGIDC, Ministério da Educação, 2010.



2. Como organizar os dados em classes?

Perante um conjunto de dados quantitativos contínuos, ao agrupá-los e ao representá-los graficamente, temos como objectivo que essa representação nos ajude a compreender os dados, fazendo sobressair algum padrão subjacente. Algumas questões que procuramos responder são, por exemplo:

- A distribuição é simétrica ou enviesada?
- Qual o centro da distribuição dos dados?
- Tem pequena ou grande variabilidade?

O primeiro passo no processo de agrupamento dos dados é saber em quantas classes vamos agrupar os dados. Muitas vezes, o tipo da variável que se está a estudar pode dar indicação do número de classes e de como construir essas classes.

Exemplo 1 – Altura e peso dos alunos de uma escola do 1.º ciclo. Pretendemos estudar as variáveis *altura* e *peso* dos alunos de uma escola do 1.º ciclo. Para isso, recolhemos a altura e o peso de 50 alunos dessa escola, obtendo os valores (em cm) para a *altura* e os valores (em kg) para o *peso*, que se apresentam na seguinte tabela:

Altura	Peso	Altura	Peso	Altura	Peso	Altura	Peso	Altura	Peso
132	26	135	29	146	40	142	32	143	35
145	39	145	35	141	33	143	34	147	40
150	45	136	30	144	35	146	40	147	40
149	45	143	32	159	57	151	46	135	29
130	26	137	30	157	49	135	30	132	28
135	30	141	30	158	58	143	38	140	30
145	40	135	29	134	30	140	31	138	30
130	28	141	32	146	40	146	43	154	47
148	40	145	35	145	34	156	45	150	45
150	47	136	30	148	43	133	29	130	28

Ao percorrer, na tabela, os dados referentes à variável *altura*, verificamos que o valor mínimo é 130 cm e o valor máximo 159 cm. Assim, é natural considerar como classes, para organizar os dados, as seguintes:

$$130 \leq \text{altura} < 135$$
$$135 \leq \text{altura} < 140$$
$$140 \leq \text{altura} < 145$$
$$145 \leq \text{altura} < 150$$
$$150 \leq \text{altura} < 155$$
$$155 \leq \text{altura} < 160$$

As classes são todas disjuntas e a sua união contém todos os elementos da amostra, isto é, cada elemento da amostra só pode pertencer a uma das classes, mas pertence necessariamente a uma dessas classes.

Considerando agora os dados referentes à variável *peso*, verificamos que os valores máximos e mínimos são, respectivamente, 26 kg e 58 kg, pelo que uma escolha possível para as classes é:

$$25 \leq \text{peso} < 30$$
$$30 \leq \text{peso} < 35$$
$$35 \leq \text{peso} < 40$$
$$40 \leq \text{peso} < 45$$
$$45 \leq \text{peso} < 50$$
$$50 \leq \text{peso} < 55$$
$$55 \leq \text{peso} < 60$$

Do mesmo modo que para a variável *altura*, também as classes anteriores foram construídas sem ambiguidade, na medida em que cada elemento da amostra pertence a alguma das classes e só a uma das classes.

2.1. Quantas classes se consideram?

Nos exemplos apresentados anteriormente, a formação de classes foi fácil de fazer de forma intuitiva. No entanto, isso nem sempre acontece. Nestes casos, podemos usar a chamada ***regra de Sturges***, que nos sugere o número de classes a usar para agrupar os dados:

- ***Regra de Sturges*** – para organizar uma amostra de dados contínuos de dimensão n, pode considerar-se para número de classes o valor k, onde k é o menor inteiro, tal que $2^k > n$.

Assim, se o número de elementos da amostra for 50, como nos exemplos apresentados anteriormente, o número aconselhado de classes é 6, já que $2^5 < 50$ e $2^6 > 50$. Note-se que esta regra não tem que ser seguida "à letra" e deve ser entendida como uma ajuda, quando não se tem qualquer ideia de quantas classes construir para proceder ao agrupamento dos dados. É apresentada sobretudo como informação para o professor.

2.2. Como se constroem as classes?

Para a formação das classes, na forma de intervalos *com a mesma amplitude*, considera-se a seguinte metodologia:

- Passo 1 – Toma-se como amplitude h, de cada intervalo, um valor arredondado por excesso, do quociente que se obtém dividindo a amplitude da amostra (máximo – mínimo) pelo número de classes, k.
- Passo 2 – Formam-se as classes como intervalos fechados à esquerda e abertos à direita, ou abertos à esquerda e fechados à direita, sendo o extremo esquerdo do primeiro intervalo o mínimo da amostra ou o extremo direito do k-ésimo intervalo o máximo da amostra, respectivamente.

Exemplo 2 – Recolheu-se a informação sobre o tempo (em minutos) que 24 alunos demoravam a chegar de casa à escola. Os valores observados são, depois de ordenados:

5 6 6 7 7 8 9 10 10 11 12 12 13 13 14 15 15 15 16 17 18 19 20 21

Como a dimensão da nossa amostra é n=24, o menor inteiro k que satisfaz a condição $2^k > 24$ é k=5. Para obter a amplitude de classe h, vamos dividir a amplitude da amostra (16 = 21 – 5) por 5. Este quociente vem igual a 3,2, pelo que um valor aproximado **por excesso** é, por exemplo, 3,25.

Para a construção das classes, vamos convencionar que todos os intervalos são fechados à esquerda e abertos à direita, isto é, da forma $[a, b[$, onde a pertence ao intervalo, mas b já não pertence. Utilizando esta metodologia, temos os seguintes intervalos, para as classes:

1ª classe: [5; 5+3,25[→ [5; 8,25[
2ª classe: [8,25; 8,25+3,25[→ [8,25; 11,50[
3ª classe: [11,50; 11,50+3,25[→ [11,50; 14,75[
4ª classe: [14,75; 14,75+3,25[→ [14,75; 18,00[
5ª classe: [18,00; 18,00+3,25[→ [18,00; 21,25[

O valor de 3,25 que utilizámos para a amplitude de classe, como aproximação por excesso do valor 3,2, é pouco natural. Mas o mesmo não acontece com 3 minutos e meio, pelo que outra alternativa possível para a amplitude de classe será h=3,5. Se se considerar este valor, o número de classes a usar é ainda 5, como se pode ver facilmente, já que as classes assim obtidas

[5; 8,5[, [8,5; 12,0[, [12,0; 15,5[, [15,5; 19,0[e [19,0; 22,5[

contêm todos os elementos da amostra.

Se se pretender construir intervalos em que os limites sejam números inteiros, podemos considerar como amplitude de classe 3 minutos ou 4 minutos, obtendo-se, respectivamente, as seguintes classes:

Amplitude de classe igual a 3 minutos	Amplitude de classe igual a 4 minutos
[5; 8[	[5; 9[
[8; 11[	[9; 13[
[11; 14[	[13; 17[
[14; 17[	[17; 21[
[17; 20[	[21; 25[
[20; 23[	

Repare-se que, quando se considerou como amplitude de classe 3 minutos, foi necessário construir 6 classes, de modo a cobrirem a totalidade dos dados; por outro lado, quando se considerou como amplitude de classe o valor 4 minutos, consideraram-se 5 classes, mas a última classe só tem um elemento. Note-se que não é correcto considerar a quarta classe na forma [17; 21], com o objectivo de evitar mais uma classe. A metodologia na construção dos intervalos de classe deve ser sempre a mesma: fechados à esquerda e abertos à direita, ou vice-versa.

Como vemos, existe uma grande flexibilidade na construção dos intervalos de classe. Em muitas situações, a regra básica a seguir é utilizar a informação disponível sobre a variável a estudar e o "bom senso" para a definição dos limites das classes.

A regra de *Sturges* pode ser usada como um primeiro passo na indicação de um número apropriado de classes. Na verdade, o que nós procuramos é um agrupamento dos dados em classes, para depois construirmos o histograma que, como veremos, deve evidenciar a estrutura subjacente aos dados. Assim, se se construírem muitas classes, essa representação apresentará muita da variabilidade presente nos dados, não conseguindo fazer sobressair o padrão que procuramos. Também um número muito pequeno de classes esconderá esse padrão.

2.3. Construção da tabela de frequências

Uma vez formadas as classes, procede-se à construção da tabela de frequências:

- Os dados contínuos são organizados na forma de uma **tabela de frequências**, com três ou mais colunas. Na primeira coluna, *coluna das classes*, consideram-se os intervalos (classes) escolhidos para agrupar os dados; na coluna seguinte, *coluna das frequências absolutas* $\mathbf{n_i}$, regista-se o total de elementos da amostra que pertencem a cada classe. Numa terceira coluna, *coluna das frequências relativas* (ou percentagens) $\mathbf{f_i}$, regista-se, para cada classe, o valor que se obtém dividindo a frequência absoluta pela dimensão da amostra.

Vamos agora construir a tabela de frequências para os dados observados para a variável *altura* de um aluno da escola do 1.º ciclo, considerados na secção anterior.

Considerámos as 6 classes aí definidas, com intervalos de amplitude 5 cm, fechados à esquerda e abertos à direita:

Classes	Freq. Abs. n_i	Freq. Rel. f_i
[130, 135[	7	0,14
[135, 140[	9	0,18
[140, 145[	11	0,22
[145, 150[	14	0,28
[150, 155[	5	0,10
[155, 160[	4	0,08
Total	50	1,00

A frequência absoluta da classe [130, 135[é 7, porque existem nos dados 7 valores maiores ou iguais a 130 e menores que 135. Para as outras classes, a metodologia é idêntica.

A soma das frequências absolutas é igual a 50, que é o número de dados, enquanto a soma das frequências relativas é igual a 1. Por vezes, esta soma não dá exactamente 1, sendo esta situação devida ao facto de os valores das frequências relativas serem arredondados.

Como se verifica a partir da tabela, predominam as alturas das classes centrais, havendo uma diminuição das frequências para as classes inferiores e superiores.

3. Construção do histograma

Agrupados os dados numa tabela de frequências, estamos aptos a construir o histograma, que é a representação gráfica mais utilizada para os dados quantitativos contínuos.

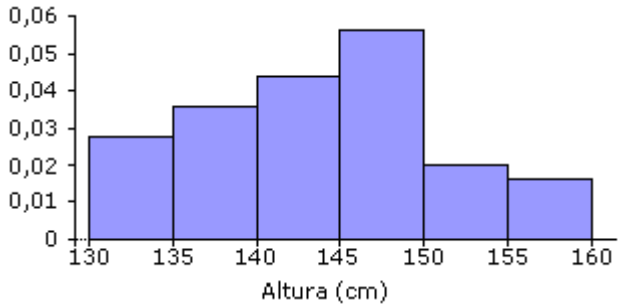
- O histograma é um gráfico, formado por uma sucessão de rectângulos adjacentes, tendo cada um por base um intervalo de classe e uma área igual (ou proporcional) à frequência relativa (ou absoluta) dessa classe.

Ao contrário do gráfico de barras, em que estas estão separadas e em que a altura de cada barra é o mais relevante, no histograma as barras (rectângulos) estão juntas e o que é importante é a área de cada uma.

Considerando, então, para áreas das barras as frequências relativas, vemos que a área total ocupada pelo histograma é igual a 1 ou 100%.

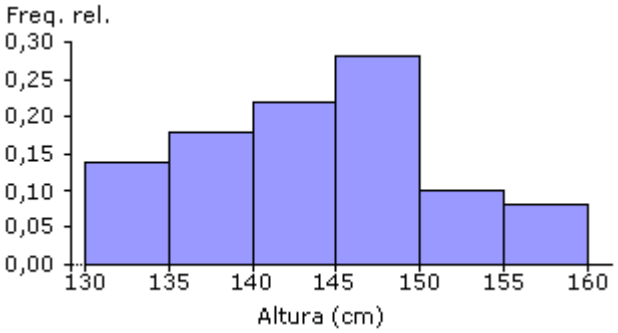
Tendo em conta a definição de histograma, para a sua construção é conveniente acrescentar uma nova coluna à tabela de frequências, com as frequências relativas a dividir pela amplitude de classe. Os valores desta coluna serão as alturas dos rectângulos com base nas classes respectivas:

Classes	Freq. Abs. n_i	Freq. Rel. f_i	Altura rectângulo classe i=f_i/h
[130, 135[	7	0,14	0,028
[135, 140[	9	0,18	0,036
[140, 145[	11	0,22	0,044
[145, 150[	14	0,28	0,056
[150, 155[	5	0,10	0,020
[155, 160[	4	0,08	0,016
Total	50	1,00	



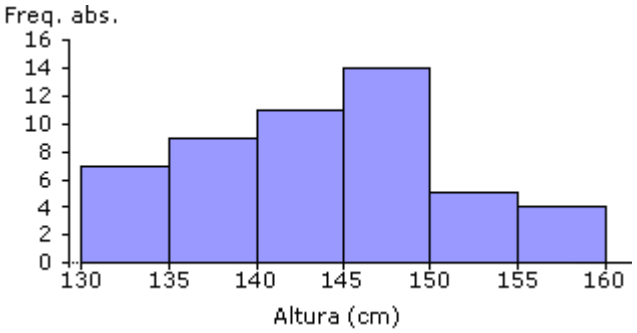
No histograma ao lado, a área do rectângulo mais à esquerda é igual a 5×0,028=0,14; a área do rectângulo seguinte é 5×0,036=0,18 e assim sucessivamente, donde a **área total do histograma é igual a 1** (soma das frequências relativas).

Suponhamos que, em vez de construirmos o histograma como anteriormente, tínhamos considerado para alturas dos rectângulos as frequências relativas. Então, neste caso, as áreas dos rectângulos já não seriam iguais às frequências relativas, mas sim proporcionais, e a área total ocupada pelo histograma seria igual a 5, em que 5 é a amplitude de classe:



No histograma ao lado, a área do rectângulo mais à esquerda é igual a 5×0,14; a área do rectângulo seguinte é 5×0,18 e assim sucessivamente, donde a **área total do histograma é igual a 5** (=5**×**1 onde 1 é a soma das frequências relativas).

Suponhamos ainda que agora se considerava para altura dos rectângulos as frequências absolutas. O resultado seria o seguinte:



No histograma ao lado, a área do rectângulo mais à esquerda é igual a 5×7; a área do rectângulo seguinte é 5×9 e assim sucessivamente, donde a **área total do histograma é igual a 250** (=5×50, onde 50 é a soma das frequências absolutas)**.**

Como se verifica, a imagem transmitida tem sempre o mesmo aspecto, já que as áreas dos rectângulos ou são iguais às frequências relativas, como é o caso do primeiro dos 3 histogramas anteriores, ou são proporcionais, com a mesma constante de proporcionalidade, que é igual à amplitude de classe no caso do segundo histograma ou à amplitude de classe vezes o número de dados, no caso do terceiro histograma. Assim, o eixo vertical só serve como auxílio para a construção dos rectângulos, não transmitindo, no caso do histograma, qualquer informação relevante:

- Não devemos perder de vista que o histograma representa os dados através das áreas das barras e não das alturas, o que constitui uma grande diferença relativamente ao gráfico de barras.
- Outra grande diferença é que no histograma as barras estão juntas, para transmitir a ideia de continuidade da variável em estudo, enquanto no gráfico de barras estas são separadas.

De um modo geral, se tivermos **n** dados e estes tiverem sido organizados em **k** classes, todas com a mesma amplitude **h**, e representarmos por n_i e f_i, respectivamente, as frequências absoluta e relativa da classe i, com i=1,...,k, a área total ocupada pelo histograma será igual a:

a) **1**, se se considerar f_i/h para altura do rectângulo correspondente à classe i, com i=1,...,k.
b) **h**, se se considerar f_i para altura do rectângulo correspondente à classe i, com i=1,...,k.
c) **h×n**, se se considerar n_i para altura do rectângulo correspondente à classe i, com i=1,...,k.

Qualquer das formas anteriores pode ser utilizada para construir o histograma, excepto nas seguintes situações:

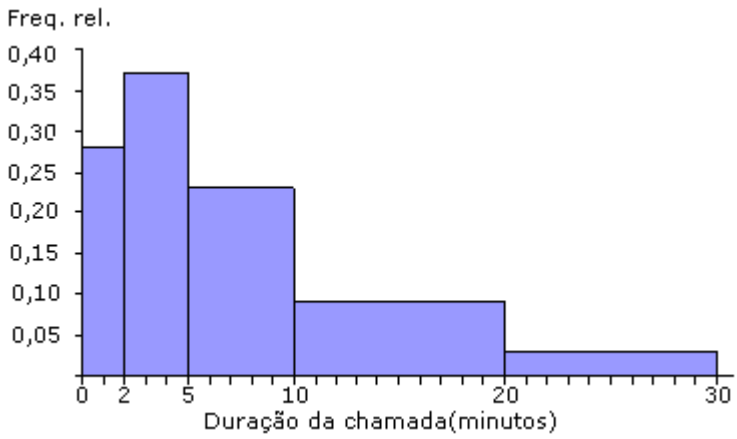
1) As classes têm amplitudes diferentes, sendo necessário utilizar, neste caso, o primeiro procedimento;
2) Pretende-se comparar histogramas de amostras com dimensão diferente, sendo também necessário utilizar o primeiro procedimento, para compararmos figuras com a mesma área (igual a 1).

Exemplo 3 – Duração das chamadas telefónicas. Uma empresa, preocupada com os gastos em telefone, decidiu fazer um estudo sobre a *duração* (em minutos) das chamadas telefónicas. Assim, o departamento de controlo de qualidade recolheu uma amostra de dimensão 100, tendo construído a seguinte tabela de frequências com os dados recolhidos:

Duração da chamada (em minutos)

Classes	Freq. absoluta	Freq. relativa
[0, 2[	28	0,28
[2, 5[	37	0,37
[5, 10[	23	0,23
[10, 20[	9	0,09
[20, 30[	3	0,03
Total	100	1,00

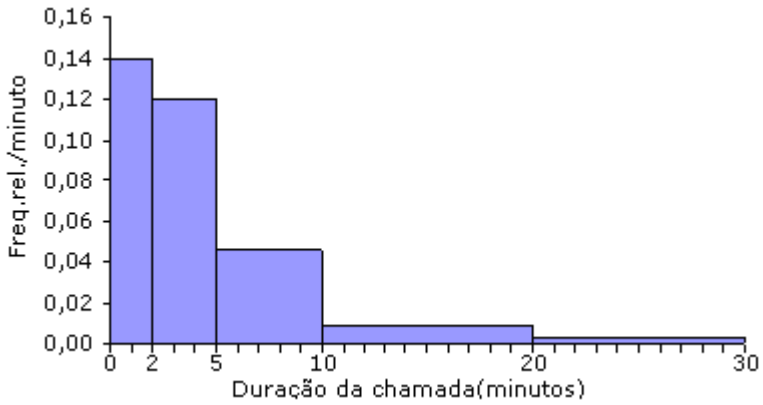
Elaborou depois o seguinte histograma, que apresentou à gerência:



Um dos gerentes, que sabia o que era um histograma, manifestou-se bastante preocupado com a percentagem de chamadas razoavelmente longas, já que a percentagem de chamadas com duração entre 5 e 10 minutos era um pouco superior às de duração entre 2 e 5 minutos e só um pouco inferior às de duração de 10 a 20 minutos, como se depreende pelas áreas dos rectângulos correspondentes às classes respectivas. Pediu para consultar a tabela de frequências e concluiu que aquela representação gráfica não estava correcta, pois as áreas dos rectângulos não eram proporcionais às frequências, induzindo em erro. Ele próprio acrescentou mais uma coluna à tabela de frequências, com as alturas correctas dos rectângulos, e construiu o histograma correspondente:

Duração da chamada (em minutos)

Classes	Freq. absoluta	Freq. relativa	Freq. relativa/amplitude classe
[0, 2[	28	0,28	0,140
[2, 5[	37	0,37	0,123
[5, 10[	23	0,23	0,046
[10, 20[	9	0,09	0,009
[20, 30[	3	0,03	0,003
Total	100	1,00	



Repare-se que as duas representações são completamente diferentes. Agora, podemos concluir que predominam as chamadas com duração entre 2 e 5 minutos e que as chamadas com duração superior a 10 minutos são pouco frequentes.