UNIDADE 2 – DADOS AGRUPADOS
MÓDULO 1 – ANÁLISE DESCRITIVA - DADOS AGRUPADOS
01

1 - DETERMINAÇÃO DE DADOS AGRUPADOS

Os dados de uma pesquisa estatística são apresentados, de maneira geral, dispersos. No máximo são apresentados de maneira ordenada como no exemplo a seguir:

Dados amostrais de altura de alunos de uma universidade: 1,60; 1,60; 1,61; 1,61; 1,61; 1,62; 1,62; 1,63; 1,64; 1,64; 1,65; 1,67; 1,67; 1,68; 1,68; 1,69; 1,69; 1,70; 1,71; 1,73; 1,73; 1,73; 1,73; 1,75; 1,76; 1,77; 1,77; 1,77; 1,78; 1,78; 1,80; 1,81; 1,83; 1,85; 1,87; 1,87.

Assim, fica complicado ao leitor entender como é a distribuição dos dados de altura dos alunos, com a grande variedade de dados apresentados. Uma saída seria então se trabalhar com as medidas de posição já estudadas.

Entretanto existe outra maneira de apresentar os dados: a **análise descritiva para dados agrupados**.

O procedimento para a determinação, ou ainda apresentação desses dados de forma agrupada é o seguinte:

a) Inicialmente temos que saber em quantos grupos, denominados estatisticamente de classes vamos dividir nossos dados. Existem na literatura diversas formulações matemáticas para isso. Uma dessas fórmulas, que talvez seja a mais consagrada, é a seguinte:

$$K = 1 + (\log n / \log 2)$$

onde:

log = operação matemática logaritmo
K = quantidade de classes;
n = número de dados (no caso amostral)

Logaritmos

Os logaritmos foram criados por John Napier (1550-1617) e desenvolvidos por Henry Briggs (1531-1630). Foram introduzidos no intuito de facilitar cálculos mais complexos. Através de suas definições podemos transformar multiplicações em adições, divisões em subtrações, potenciações em multiplicações e radiciações em divisões.

Dados dois números reais positivos a e b, onde a ?1, a>1 e b>0, existe somente um número real x, tal que

logab=x ou ax=b

O número a é chamado de **base do logaritmo**, b é o **logaritmando** e x o **logaritmo**. Quando não indicamos, significa que o valor de a é 10:

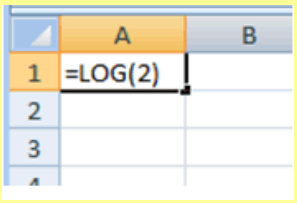
log b é o mesmo que **Log10 b**

O logaritmo de b na base a é o expoente que devemos atribuir ao número a para obter b:
a)Log2 8 = x, logo 2x = 8, portanto x=3 visto que 23=4
b)Log 100 = x, logo 10x = 100, portanto x=2 visto que 102 = 100

Podemos ver que algumas vezes o valor do algoritmo é fácil de calcular, isso ocorre quando o valor de b for múltiplo de a. Entretanto, muitas vezes é difícil fazer esse cálculo, por exemplo:

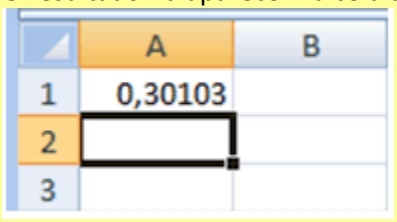
Log 2 = x, x = ?

Sabemos que 10x = 2. Mas quanto deve ser esse valor de x? Podemos fazer aproximações, mas hoje é mais comum utilizar a calculadora ou mesmo o Excel para calcular o logaritmo. No Excel usamos a função LOG10() ou LOG(). Para calcularmos Log 2 no Excel, por exemplo, basta clicar na célula e escrever:



e clicar ENTER.

O resultado irá aparecer na célula:



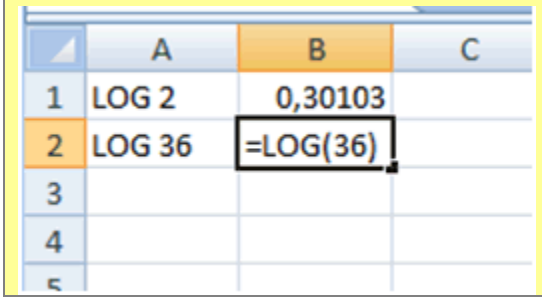
Vamos refazer o cálculo do número de classes que está no nosso conteúdo. A fórmula para achar o número de classes é:

$$K = 1 + \left(\frac{Log\ n}{Log\ 2}\right)$$

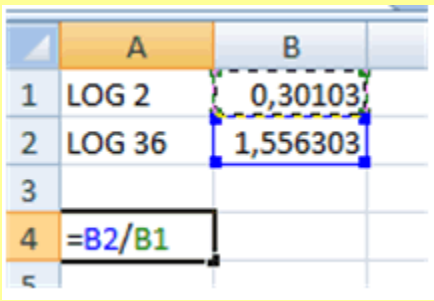
Onde n é a quantidade de dados disponíveis. Verifica-se no exemplo que a quantidade de dados disponíveis é 36, logo n=36, portanto teremos que calcular:

$$K = 1 + \left(\frac{Log\ 36}{Log\ 2}\right)$$

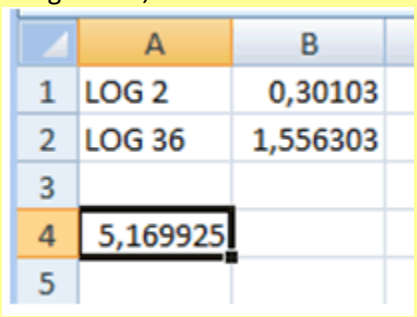
Vamos fazer inicialmente no Excel os cálculos separados para cada logaritmo da fórmula:



Fazendo a divisão:



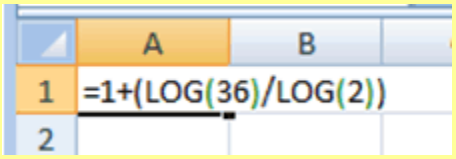
Chegando a,



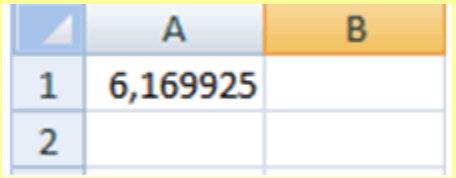
No conteúdo esse resultado foi arredondando para 5,2. Substituindo na fórmula:

$$K=1+(5,2)=6,2$$

Poderíamos calcular o valor de K de uma só vez no Excel, como segue:



Teremos como resultado:



Que seria arredondado para 6,2.

02

b) Com a quantidade de classes definidas, devemos passar então ao dimensionamento da amplitude de cada uma das classes. Por convenção estatística, todas as classes devem ter a mesma amplitude. Assim, o procedimento matemático para a determinação da amplitude das classes será dividir a amplitude total dos dados (que será calculada pela subtração do maior dado pelo menor dado) pela quantidade de classes (K), assim:

$$A_c = \frac{A_T}{K}$$

c) Com os valores da quantidade de classe e da amplitude das mesmas, passaremos a dimensionar cada uma das classes. Para esse procedimento é usual a montagem de uma tabela. O primeiro valor da primeira classe será igual ao valor do menor dado da amostra (ou da população). O último valor da classe será igual ao primeiro somado com a amplitude de classe. A segunda classe, por sua vez, deverá se iniciar no valor do final da primeira classe e findar no valor somado com a amplitude de classe. Para as demais classes deve-se seguir o mesmo procedimento até que se tenha a quantidade "K" de classes.

d) Com todas as classes dimensionadas, ou seja, sabendo-se os limites superior e inferior de cada classe, podemos passar ao cálculo de quantos dados pertencerão a cada classe. Assim, simplesmente contamos quantos dos nossos dados originais são maiores que o primeiro elemento de nossa classe e menores que o último. Esse procedimento deverá ser repetido para todas as classes até que todos os dados originais sejam distribuídos nas mesmas. Contudo, uma questão deverá ser analisada: como o valor do limite superior de uma classe será exatamente igual ao limite inferior da seguinte, um dado com esse valor deverá entrar em qual das duas classes? Para isso excluímos a possibilidade de que esse dado entre na primeira classe (a do limite superior) e fique exclusivamente na classe posterior.

e) Com a relação entre a quantidade de elementos que pertencem a cada classe e a quantidade total de elementos em análise poderemos, ainda, calcular a frequência relativa de cada classe.

03

Passemos então ao exemplo prático com os dados já apresentados anteriormente relativos à altura de alunos de uma universidade: 1,60; 1,60; 1,61; 1,61; 1,61; 1,62; 1,62; 1,63; 1,64; 1,64; 1,65; 1,67; 1,67; 1,68; 1,68; 1,69; 1,69; 1,70; 1,71; 1,73; 1,73; 1,73; 1,73; 1,75; 1,76; 1,77; 1,77; 1,77; 1,78; 1,78; 1,80; 1,81; 1,83; 1,85; 1,87; 1,87.

Inicialmente iremos calcular o valor de "K", assim, sabendo que n=36, teremos:

K = 1+(log n / log 2) → K = 1+(log 36 / log 2) → K = 1+(1,56 / 0,30)
K = 1 + 5,20 K = 6,20

Contudo, não existe 0,20 classe. O valor de "K" deverá ser inteiro. Nesse caso, se arredondarmos o valor 6,20 para 6, conforme o critério matemático, ficaremos com menos classes que o necessário e, com certeza, alguns dados ficarão sem classe. Assim, devemos sempre "arredondar para cima" o valor de K e, nesse caso, será 7.

Sabendo, então, que teremos sete classes, passaremos ao dimensionamento da amplitude das classes. Para isso inicialmente teremos que calcular a amplitude total dos dados. Assim:

At = maior dado - menor dado → At = 1,87 - 1,60 → At = 0,27

$$A_C = \frac{A_t}{k} \rightarrow A_C = \frac{0,27}{7} \rightarrow A_C = 0,039 \rightarrow A_C = 0,04$$

Com os dados da quantidade de classes e da amplitude das mesmas poderemos montar nossa tabela:

Classe	nº de dados
[1,60 - 1,64[	08
[1,64 - 1,68[	05
[1,68 - 1,72[	06
[1,72 - 1,76[	05
[1,76 - 1,80[	06
[1,80 - 1,84[	03
[1,84 - 1,88[	03
Total de dados	**36**

Repare que ora o colchete aparece virado para fora "[" ora aparece virado para dentro. A notação "[" significa justamente o impedimento desse último dado de entrar nessa classe. Enquanto que o "[" denota a permissão de que o dado ao seu lado entre na classe.

04

Poderemos ainda calcular a frequência relativa de cada classe simplesmente dividindo-se o número de dados pertencentes a cada classe pela quantidade total de dados. Para isso o procedimento usual é ampliar a tabela com mais uma coluna para o dimensionamento desses valores. Assim, ficamos com a seguinte tabela:

Classe	nº de dados	frequencia relativa
[1,60 - 1,64[	08	0,22
[1,64 - 1,68[	05	0,14
[1,68 - 1,72[	06	0,17
[1,72 - 1,76[	05	0,14
[1,76 - 1,80[	06	0,17
[1,80 - 1,84[	03	0,08
[1,84 - 1,88[	03	0,08
Total de dados	**36**	**1,00**

Conseguimos representar nossa base de dados com 36 elementos através de uma distribuição de classes, ou seja, em dados agrupados. Com esses dados conseguimos perceber de maneira mais clara como anda a distribuição de nossos dados. Podemos dizer, por exemplo, que 17% dos alunos têm estatura ente 1,76 e 1,80. Ou ainda que 36% dos alunos tem altura compreendida entre 1,60 e 1,68, acumulando-se as duas primeiras classes.

Quando se tem a distribuição original dos dados, qualquer análise estatística tem que partir dos mesmos. Em alguns casos, quando não possuímos os dados originais, mas somente os dados em distribuição de classes, poderemos calcular as medidas estatísticas a partir desse conjunto de dados agrupados. Veremos a seguir os procedimentos para tal análise.

05

2 - Média, Heterogeneidade e Discrepância

Anteriormente foi apresentada a análise descritiva para dados não agrupados e chamamos sua atenção para o fato de que as ideias vinculadas a cada uma das medidas continuariam válidas com dados agrupados. O que vai mudar de forma substancial é a forma inicial de apresentação dos dados e alguns ajustes que se farão necessários para obtenção das medidas descritivas.

Consideremos então o seguinte exemplo:

Uma empresa varejista idealiza desenvolver uma campanha de fidelização e, para isso, levanta um conjunto de dados de seus arquivos referentes às vendas dos últimos três meses, para que possa fazer um "mapeamento" dos valores gastos pelos clientes. O resultado do levantamento está apresentado, resumidamente, a seguir.

Valores Gastos (R$)	Quantidade de Clientes
[10,00 ; 20,00[	350
[20,00 ; 30,00[	725
[30,00 ; 40,00[	1290
[40,00 ; 50,00[	875
[50,00 ; 60,00[	235
[60,00 ; 70,00[	450
[70,00 ; 80,00[	280
[80,00 ; 90,00[	340
[90,00 ; 100,00[	655
[100,00 ; 110,00[	530
[110,00 ; 120,00[	370
[120,00 ; 130,00[	120
TOTAL	6220

Deve ter chamado sua atenção a existência de colchetes voltados ora para dentro, ora para fora dos intervalos numéricos. A leitura do primeiro intervalo deve ser: despesas a partir de R$ 10,00, inclusive (fechado no limite inferior), até R$ 20,00, exclusive (aberto no limite superior), ou R$19,99. Para os demais a leitura é análoga, com exceção do último que inclui os dois extremos, tanto R$ 120,00 como R$ 130,00. Deve também chamar atenção que um grupo relativamente grande de dados (6220 valores) foi apresentado de forma sintética, pois eles não estão apresentados individualmente, mas agrupados em intervalos.

06

Vamos admitir que o "mapeamento" desejado passe inicialmente por uma análise descritiva, como uma primeira abordagem exploratória do conjunto de dados. Convém lembrar que, muito possivelmente, o "dono" dos dados teria dificuldade de propor um caminho para sua análise, que posteriormente estará sustentando algum plano de ação voltado para fidelização de clientes.

E aí começam as perguntas, afinal se os dados não são efetivamente conhecidos, como proceder para calcular as medidas descritivas que, como visto anteriormente, pressupõe seu conhecimento?

07

É certo, então, que surge a necessidade de adaptações na forma de calculá-las. Se 350 clientes gastaram de R$ 10,00 a R$ 19,99, quanto teria gastado cada um? A mesma pergunta vale para as outras classes.

A resposta é que não há como saber (salvo se uma base de dados completa for disponível e optar-se pela estratégia de dados não agrupados, o que talvez fosse realmente a melhor alternativa, pois com os recursos computacionais hoje disponíveis, a agregação de dados tem função muito mais "estética"). Considerando que, por alguma razão, a base original não esteja acessível, o caminho passará, necessariamente, pela adoção de algumas premissas e aproximações.

08

A primeira premissa diz respeito aos valores das despesas: assume-se, por hipótese, que os dados de cada classe estejam uniformemente distribuídos ao longo dos intervalos, o que significa dizer que, supostamente, os 350 clientes do primeiro intervalo estariam distribuídos de forma "equilibrada" entre os diversos valores entre 10 e 20 reais, o mesmo valendo para os 725 do segundo intervalo e assim sucessivamente. Sendo assim, a despesa média para cada intervalo passa a ser o próprio ponto médio do respectivo intervalo, sendo esse o ponto que mais se aproximaria (provocaria menor erro) do conjunto das despesas inseridas em cada intervalo.

É claro que a questão da largura do intervalo passa a ter uma expressiva relevância, pois quanto maior sua amplitude, maior nossa "ignorância" a respeito dos dados ali inseridos e, consequentemente, menor a precisão das medidas que se pretende calcular.

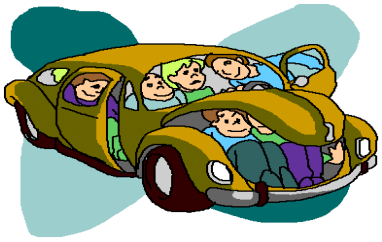
À medida que a largura dos intervalos aumenta, se por um lado a precisão diminui, a síntese/o resumo na apresentação dos dados aumenta. E, como não é possível dissociar um fato do outro, o que se busca é um equilíbrio entre perda/ganho de precisão e ganho/perda de síntese. Não há uma formulação "mágica" que assegure configurações ótimas nesse sentido (embora alguns textos apresentem fórmulas, como a de Sturges, que buscam esse objetivo, ao estabelecer um número de classes a partir de uma estrutura que utiliza a função logarítmica e o total de indivíduos/objetos estudados).

Recomenda-se bom senso, tendo-se em mente que um bom sinalizador de que o equilíbrio ainda não foi alcançado é quando uma (ou mais classes) apresentam frequências julgadas muito superiores ou muito inferiores à maioria das demais.

Por exemplo, se uma classe tem 2500 indivíduos enquanto todas as demais estão entre 50 e 450, há um indício de que se está perdendo muita informação naquela classe, sendo ela merecedora de uma quebra em duas, três ou quatro novas classes (desde que as observações individuais estejam disponíveis em um arquivo ou um banco de dados).

09

Por outro lado, se a grande maioria das classes está com frequência oscilando entre 200 e 580 e uma ou duas classes estão com 15 ou 20 observações, pode ser interessante, em busca de maior síntese quando da apresentação dos dados, agrupá-la(s) à(s) classe(s) vizinha(s) - inferior ou superior, conforme for mais adequado para uma leitura mais eficiente (não ignorando que quaisquer agregações diminuem a precisão das medidas que vierem a ser calculadas tomando por base os dados agregados, o que pode não ser conveniente).



Para a pergunta: "e se os dados não tiverem, na verdade, um comportamento semelhante a esse que se assume por hipótese?" Só há uma resposta: todo o desenvolvimento subsequente de análise descritiva pode ficar comprometido.

A apresentação dos dados, tal como aquela no início deste módulo, é uma distribuição de frequências, constando de um conjunto de classes, preferencialmente com a mesma amplitude. No caso em questão, essa amplitude padrão é de R$ 10,00 (que é a diferença entre o limite superior e o limite inferior de cada intervalo).

Amplitude
Diferença entre o limite superior e o limite inferior de uma classe em uma distribuição de frequências.

Distribuição de frequências

Uma distribuição de frequência é um método de agrupamento de dados em classes, ou intervalos, de tal forma que a cada classe fique vinculada a respectiva frequência absoluta ou porcentagem (frequência relativa). Pode ser útil se precisarmos lidar com grande quantidade de dados, em particular, quando se tratar de apresentação dos dados. Com os recursos computacionais hoje disponíveis e disseminados a utilização prática de dados agrupados (para efeito de análise da base de dados) tem sido usada com menor intensidade, mesmo quando se trata de grandes bases de dados.

10

O primeiro passo subsequente é determinar o ponto médio de cada intervalo, considerando, pelos argumentos já apresentados, que esse será o ponto que melhor representará (mais se aproximará) o conjunto de dados do respectivo intervalo.

Assim:

Valores Gastos (R$)	Pontos Médios	Quantidade de Clientes
[10,00 ; 20,00[	15,00	350
[20,00 ; 30,00[	25,00	725
[30,00 ; 40,00[	35,00	1290
[40,00 ; 50,00[	45,00	875
[50,00 ; 60,00[	55,00	235
[60,00 ; 70,00[	65,00	450
[70,00 ; 80,00[	75,00	280
[80,00 ; 90,00[	85,00	340
[90,00 ; 100,00[	95,00	655
[100,00 ; 110,00[	105,00	530
[110,00 ; 120,00[	115,00	370
[120,00 ; 130,00]	125,00	120
TOTAL	-------	6220

O fato de os limites inferiores e/ou superiores estarem incluídos ou excluídos não compromete a determinação do ponto médio, tolerando-se uma aproximação como a adotada aqui.

11

Como o ponto médio representa, nesse contexto, a média para cada intervalo, o cálculo da despesa média será feito de forma similar àquela apresentada anteriormente, ou seja:

$$\textbf{Despesa Média} = \frac{\sum_i PM_i \text{ x } f_i}{n}$$

Onde:

- PMi é o ponto médio de cada classe i.
- fi é a freqüência de cada classe i.
- n é a quantidade total de observações.

Assim:

$$\textbf{Despesa Média} = \frac{(15{,}00 \text{ X } 350) + (25{,}00 \text{ X } 725) + \ldots + (125{,}00 \text{ X } 120)}{6220} = \frac{375.400}{6220} = 60{,}35$$

E, mais uma vez, surge a pergunta: qual o significado real e prático desse valor médio?

E, mais uma vez, não é possível respondê-la sem conhecermos a variabilidade do conjunto de dados (com respeito à variável despesa realizada no ponto de comércio em questão).

12

Também como já visto anteriormente, deve-se buscar o coeficiente de variação e para chegar a ele é necessário calcular o desvio-padrão, que por sua vez é a raiz quadrada da variância.

$$CV = \frac{\text{Desvio Padrão}}{\text{Média}} = \frac{\sqrt{\text{Variância}}}{\text{Média}}$$

A variância também será calculada a partir dos pontos médios dos intervalos:

$$\textbf{Variância} = \frac{\sum_i [(PM_i - \text{Média})^2 \text{ x } f_i]}{n}$$

E substituindo pelos valores numéricos:

$$\textbf{Variância} = \frac{[(15{,}00 - 60{,}35)^2 \times 350] + \ldots + [(125{,}00 - 60{,}35)^2 \times 120]}{6220} = \frac{6.393.722}{6220} = 1027{,}93$$

O que dará origem ao seguinte coeficiente de variação:

$$CV = \frac{\sqrt{1027{,}93}}{60{,}35} = \frac{32{,}06}{60{,}35} = 0{,}5312 = 53{,}12\%$$

Agora somos sabedores que o grupo sob análise é consideravelmente heterogêneo com respeito às despesas naquele estabelecimento comercial. Logo, a média está longe de ser uma boa medida representativa do mesmo, o que demandará outras medidas descritivas.

Para diagnóstico de pontos discrepantes, faz-se:

Limite Superior = Média + (3 x Desvio-Padrão) = 60,35 + (3 x 32,06) = 156,54
Limite Inferior = Média - (3 x Desvio-Padrão) = 60,35 - (3 x 32,06) = - 35,83

A partir desse critério só seriam passíveis de serem considerados pontos atípicos valores inferiores a -35,83 ou superiores a 156,54. Como tais valores não são encontrados na base de dados (que está compreendida entre R$ 10,00 e R$ 130,00), não há nenhum ponto discrepante no presente exemplo.

13

RESUMO

Neste módulo trabalhamos com análise descritiva de dados agrupados, cuja característica maior é uma apresentação sintética dos dados observados que, por sua vez, impede que esses dados sejam vistos com exatidão. Resumindo: ganha-se síntese e perde-se precisão. Caso esse seja um recurso apenas para apresentação de dados, isso não é tão grave, porém o cálculo de medidas descritivas sem conhecimento específico dos dados traz um conjunto de aproximações, sem as quais se pode não ter medida alguma, sendo a mais relevante a hipótese de distribuição uniforme dos dados em cada intervalo.

O primeiro passo consiste em determinar o ponto médio de cada classe (sendo que, desejavelmente, mas não obrigatoriamente, todas as classes devem ter a mesma amplitude). Em seguida, calculam-se média e coeficiente de variação a partir desses pontos médios com a formulação geral. O coeficiente de variação deve ser dado pelo desvio-padrão dividido pela média e expressa o grau de variabilidade da base de dados. O próximo passo é a verificação quanto à existência de pontos discrepantes, o que é feito a partir do estabelecimento de limites superior e inferior, com o acréscimo e decréscimo de 3 desvios à média.

Diante dos recursos computacionais hoje disponíveis, o agrupamento de dados é muito mais uma forma "elegante" de apresentação dos dados, sendo recomendável que o cálculo das medidas, sempre que os dados individuais forem disponíveis, seja feito a partir dos procedimentos para dados não agrupados.

UNIDADE 2 – DADOS AGRUPADOS

MÓDULO 2 – ANÁLISE DESCRITIVA QUANDO HÁ VARIABILIDADE ALTA

01

1- DETERMINAÇÃO DA MEDIANA DE DADOS AGRUPADOS

Para determinação da **mediana**, o primeiro passo é a sua localização, isso é, qual a posição central. Rigorosamente, haveria duas posições centrais, pelo fato de haver um número par de dados, porém, quando o conjunto de dados pode ser considerado suficientemente grande, é comum adotar, por aproximação, uma única posição central, dada pelo número de dados dividido por dois. Nesse caso:

$$\text{Posição central} = \frac{6220}{2} = 3110$$

O próximo passo é determinar o valor que ocupa essa posição quando os dados estão ordenados (crescente ou decrescentemente). A tabela original já apresenta os dados/as classes em ordem crescente e para saber em qual classe essa posição está localizada é necessário construir uma nova coluna na tabela, na qual serão apresentadas frequências acumuladas.

Valores Gastos (R$)	Pontos Médios	Quantidade de Clientes	Quantidade Acumulada
[10,00 ; 20,00[	15,00	350	350
[20,00 ; 30,00[	25,00	725	1075
[30,00 ; 40,00[	35,00	1290	2365
[40,00 ; 50,00[	45,00	875	3240
[50,00 ; 60,00[	55,00	235	3475
[60,00 ; 70,00[	65,00	450	3925
[70,00 ; 80,00[	75,00	280	4205
[80,00 ; 90,00[	85,00	340	4545
[90,00 ; 100,00[	95,00	655	5200
[100,00 ; 110,00[	105,00	530	5730
[110,00 ; 120,00[	115,00	370	6100
[120,00 ; 130,00]	125,00	120	6220
TOTAL	-------	6220	-------

Pode-se constatar que 350 clientes gastam menos de R$ 20,00, 1075 clientes gastam menos de R$ 30,00, 2365 clientes gastam menos de R$ 40,00 e 3240 clientes gastam menos de R$ 50,00. Assim, a posição correspondente ao "cliente de número 3110" está na classe [40,00;50,00[. Mas qual seria o valor exato entre 40 e 50 reais que deveria estar associado a essa posição?

02

Esquematicamente, temos:

Assim, a determinação do valor da mediana deve partir do limite inferior do intervalo no qual já se sabe que ela está, no caso R$ 40,00, mais "um pedaço" dentro daquele intervalo, que deve ser utilizado para se chegar até a mediana (correspondente ao percentual calculado como sinalizado acima).

No exemplo, deve-se tomar 85,14% da amplitude do intervalo de R$ 10,00, no qual está a mediana.

Consequentemente, o valor da mediana no exemplo será:

Mediana = 40 + (0,8514 x 10) = 40 + 8,51 = 48,51

Significando que 50% das despesas estão entre R$ 10,00 e R$ 48,51 e 50% a entre R$ 48,51 e R$ 130,00 (evidentemente com as ressalvas já estabelecidas).

03

Repassando o desenvolvimento, temos:

- determinou-se a posição da mediana, fazendo 50% (ou 0,50) vezes o total de observações;
- localizou-se (com o apoio das frequências acumuladas) qual a classe na qual a mediana estava;
- determinou-se o número de observações necessárias dentro da classe da mediana para chegar-se efetivamente à posição da mediana (fazendo-se posição da mediana - frequência acumulada até a classe anterior);
- "transformou-se" o valor anterior em um percentual, ou seja, quantos por cento do intervalo no qual a mediana está deve ser tomado para se chegar a ela (com a pressuposição da distribuição uniforme);
- multiplicou-se este percentual pela amplitude do intervalo, pois este é o valor que deve ser somado ao limite inferior do intervalo no qual a mediana está;
- somou-se ao limite inferior do intervalo o "incremento" calculado no passo anterior.

Genericamente

MEDIANA

$$\text{Limite inferior do intervalo} + (1) \times \text{amplitude do intervalo} = \text{Limite inferior do intervalo} + (2) \times \text{amplitude do intervalo}$$

$$(1)\quad \frac{\text{posição da mediana - frequência acumulada até classe inferior}}{\text{frequência do intervalo}}$$

$$(2)\quad \frac{\text{(0,5 *quantidade de obs) - frequência acumulada até classe anterior}}{\text{frequência do intervalo}}$$

$$(1)\quad \frac{\text{posição da mediana - frequência acumulada até classe anterior}}{\text{frequência do intervalo}}$$

$$(2)\quad \frac{\text{(0,5 * quantidade de obs) - frequência acumulada até classe anterior}}{\text{frequência do intervalo}}$$

04

2 - Determinando o Primeiro e o Terceiro quartil

Quanto ao primeiro e terceiro quartis, o procedimento é análogo ao da mediana, pois esta é o segundo quartil (todos são obtidos por interpolação). A ideia/o raciocínio é exatamente o mesmo, com as devidas adequações, considerando que o primeiro quartil é um valor numérico correspondente à posição que deixa 25% do total de dados entre o valor mínimo e ele e o terceiro quartil é um valor numérico correspondente à posição que deixa 75% do total de dados entre o valor mínimo e ele. Relembrando: os três quartis e os valores mínimo e máximo delimitam quatro subgrupos de dados de igual tamanho.

• Determinando o primeiro quartil no exemplo sob análise.
• Determinando o terceiro quartil no exemplo sob análise.

Genericamente:

$$\text{QUARTIL} = \text{limite inferior do intervalo} + \left(\frac{\text{posição do quartil} - \text{frequência acumulada até classe anterior}}{\text{frequência do intervalo}} \right) \times \text{amplitude do intervalo}$$

E posição do quartil = 0,25 x quantidade de observações (para o primeiro quadril) ou
0,5 x quantidade de observações (para o segundo quartil/mediana) ou
0,5 x quantidade de observações (para o primeiro quartil)

Esquematicamente:

Fica evidenciado que há subgrupos cuja heterogeneidade parece muito superior a de outros, pois há 25% de clientes entre R$ 10,00 e R$ 33,72, 25% entre R$ 33,72 e R$ 48,51, 25% entre R$ 48,51 e R$ 91,83 e outros 25% entre R$ 91,83 e R$ 130,00 (uma mesma quantidade de clientes em intervalos com amplitudes bem distintas). Isso sugere que, muito possivelmente ainda não tenhamos alcançado o objetivo de subgrupos suficientemente homogêneos, sendo necessário "quebrar"/dividir a base de dados em um número maior de partes, formando subgrupos com quantidades menores de observações.

Interpolação
Processo de obtenção de valores dentro de um intervalo, segundo um conjunto de critérios matemáticos.

Determinando o primeiro quartil, temos:

1- posição do primeiro quartil é de 0,25 x 6220 = 1555;
2- essa posição está localizada na terceira classe, ou seja, entre 30 e 40 reais (até R$ 30,00 há um total de 1075 clientes e até R$ 40,00 esse total é de 2365 clientes);
3- até a classe anterior a essa, encontram-se 1075 observações, assim são necessárias mais 480 observações (= 1555 – 1075) dentro do intervalo no qual está o primeiro quartil (a partir de R$ 30,00);

4- essas 480 observações correspondem a 37,21% do intervalo de 30 a 40 reais, pois 480 / 1290 (frequência do intervalo) = 0,3721;
5- multiplicando-se esse percentual pela amplitude do intervalo (de 10 reais) chega-se a 0,3721 x 10 = 3,72;
6- somando-se esse valor ao limite inferior do intervalo chega-se ao primeiro quartil, que é 30 + 3,72 = 33,72.

Determinando o terceiro quartil, temos:

1- posição do terceiro quartil é 0,75 x 6220 = 4665;
2- essa posição está localizada na nona classe, ou seja, entre 90 e 100 reais (até R$ 90,00 há um total de 4545 clientes e até R$ 100,00 esse total é de 5200 clientes);
3- como até a classe anterior a essa encontram-se 4545 observações, são necessárias mais 120 observações (= 4665 – 4545) dentro do intervalo no qual está o terceiro quartil (a partir do R$ 90,00);
4- essas 120 observações correspondem a 18,32% do intervalo de 90 a 100 reais, pois 120 / 655 (frequência do intervalo) = 0,1832;
5- multiplicando-se esse percentual pela amplitude do intervalo (de 10 reais) chega-se a 0,1832 x 10 = 1,83;
6- somando-se esse valor ao limite inferior do intervalo chega-se ao terceiro quartil, que é 90 + 1,83 = 91,83.

05

3 - Determinando a Moda de dados agrupados

Quanto à moda, ou valor mais frequente, ela não fica bem caracterizada de forma específica, sendo muito mais adequado falar em uma classe modal, ou seja, uma classe cuja frequência é superior às demais.

No exemplo, essa classe de valores estaria de R$ 30,00 a R$ 40,00, onde estão 1290 clientes (ou 20,74% do total). Uma nova coluna mostrando as frequências percentuais pode ser útil.

Valores Gastos (R$)	Pontos Médios	Quantidade de Clientes	Frequência Relativa (%)	Quantidade Acumulada
[10,00 ; 20,00[	15,00	350	5,63	350
[20,00 ; 30,00[	25,00	725	11,66	1075
[30,00 ; 40,00[	35,00	1290	20,74	2365
[40,00 ; 50,00[	45,00	875	14,07	3240
[50,00 ; 60,00[	55,00	235	3,78	3475
[60,00 ; 70,00[	65,00	450	7,23	3925
[70,00 ; 80,00[	75,00	280	4,50	4205
[80,00 ; 90,00[	85,00	340	5,47	4545
[90,00 ; 100,00[	95,00	655	10,53	5200
[100,00 ; 110,00[	105,00	530	8,52	5730
[110,00 ; 120,00[	115,00	370	5,95	6100

[120,00 ; 130,00]	125,00	120	1,93	6220
TOTAL	-------	6220	100,00	-------

06

A literatura sobre o tema fala das fórmulas de *Czuber* e de *King* como formas aceitas para determinação de um único valor como moda em uma distribuição de frequências, mesmo que isso possa não ter nenhum compromisso com a realidade. Isso porque se houver um único valor passível de ser caracterizado como moda, ele pode não estar na classe modal (e quem disse que, nessas situações, há apenas uma moda? E se não houver moda?).

Fórmula de Czuber

$$\text{Moda} = \text{limite inferior do intervalo} + \left[\frac{(f_i - f_{i-1})}{2f_i - (f_{i+1} + f_{i-1})} \times \text{amplitude do intervalo} \right]$$

onde f_i é a frequência da classe modal,

f_{i-1} é a frequência da classe anterior à modal,

f_{i+1} é a frequência da classe posterior à modal.

Fórmula de King

$$\text{Moda} = \text{limite inferior do intervalo} + \left[\frac{f_{i+1}}{(f_{i+1} + f_{i-1})} \times \text{amplitude do intervalo} \right]$$

onde f_{i-1} é a frequência da classe anterior à modal,

f_{i+1} é a frequência da classe posterior à modal.

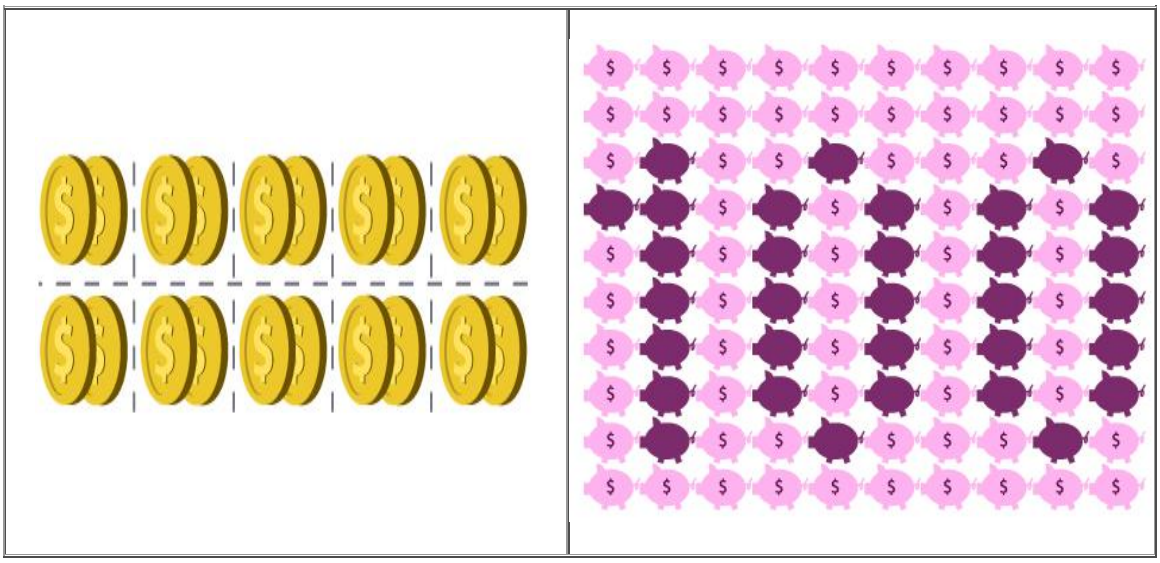
07

4 - Divisão da Base de Dados: Decis e Percentis

Em uma base de dados considerada suficientemente grande, à medida que a heterogeneidade aumenta, pode ser necessário dividi-la em mais de quatro grupos (o que é possível com utilização dos três quartis e dos valores mínimo e máximo do grupo), pois o recurso de gerar grupos menores em muito contribui para melhor visualização da distribuição do conjunto de dados, uma vez que esses subgrupos, em geral, quanto menores, mais homogêneos.

A exemplo dos quartis, também são clássicos os **decis** e os **percentis**, que, como seus próprios nomes sinalizam, dividem a base de dados em dez e em cem subgrupos, respectivamente. Os três quartis nada mais são do que o 25º, 50º e 75º percentis. Os nove decis também são percentis, correspondendo ao 10º, 20º, 30º, 40º, 50º, 60º, 70º, 80º e 90º percentis.

Decil	Percentil



O procedimento para determinação dessas medidas não tem nada de excepcional, sendo análogo àquele adotado para os quartis (que, como dito no parágrafo anterior, são percentis).

O conjunto completo desses valores que dividem/separam a base de dados forma o que se chama conjunto de **separatrizes**.

08

Exemplificando o cálculo de um dos decis (o sexto) e de um dos percentis (o nonagésimo quinto), a partir dos dados apresentados no início deste módulo:

$$\text{Posição do sexto decil} = 0{,}6 \times 6220 = 3732 \Rightarrow$$

$$\text{Sexto decil} = 60 + \left(\frac{3732 - 3475}{450}\right) \times 10 = 60 + 5{,}71 = 65{,}71$$

$$\text{Posição do nonagésimo quinto percentil} = 0{,}95 \times 6220 = 5909 \Rightarrow$$

$$\text{Nonagésimo quinto percentil} = 110 + \left(\frac{5909 - 5730}{370}\right) \times 10 = 110 + 4{,}84 = 114{,}84$$

Genericamente, qualquer separatriz pode ser determinada a partir da seguinte formulação:

$$\text{SEPARATRIZ} = \text{limite inferior da classe} + \left(\frac{\text{posição da separatriz} - \text{frequência acumulada até classe anterior}}{\text{frequência da classe da separatriz}}\right) \times \text{amplitude da classe da separatriz}$$

Sendo a posição da separatriz dada por:

Posição da separatriz = percentual correspondente à frequência até a separatriz
x
frequência total do conjunto que está sendo analisado.

09

RESUMO

Neste módulo trabalhamos com análise descritiva de dados agrupados heterogêneos. Para analisarmos esse tipo de dados, parte-se para a inclusão de novas medidas como moda e quartis e/ou decis e/ou percentis (ou seja, separatrizes). Faz-se uma ressalva quanto à moda, pois quando os dados são agrupados é mais apropriado falar-se em classe modal, sendo essa a classe com maior frequência.

Quanto às separatrizes, todas elas são determinadas por procedimentos de interpolação, partindo-se de sua posição na base de dados, a localização da classe na qual estão inseridas e incrementando-se ao limite inferior de cada classe uma parcela correspondente ao quanto é necessário percorrer dentro daquele intervalo para "encontrar" o valor da separatriz desejada.

Quanto mais heterogêneo for o grupo (com relação à variável que se está estudando) maior o número de subgrupos/segmentos que deve ser gerado para permitir melhor visualização de sua distribuição.

Novamente salientamos que diante dos recursos computacionais hoje disponíveis, o agrupamento de dados é muito mais uma forma "elegante" de apresentação dos dados, sendo recomendável que o cálculo das medidas, sempre que os dados individuais forem disponíveis, seja feito a partir dos procedimentos para dados não agrupados.

UNIDADE 2 – DADOS AGRUPADOS
MÓDULO 3 – PRÁTICA COM EXCEL - DADOS AGRUPADOS

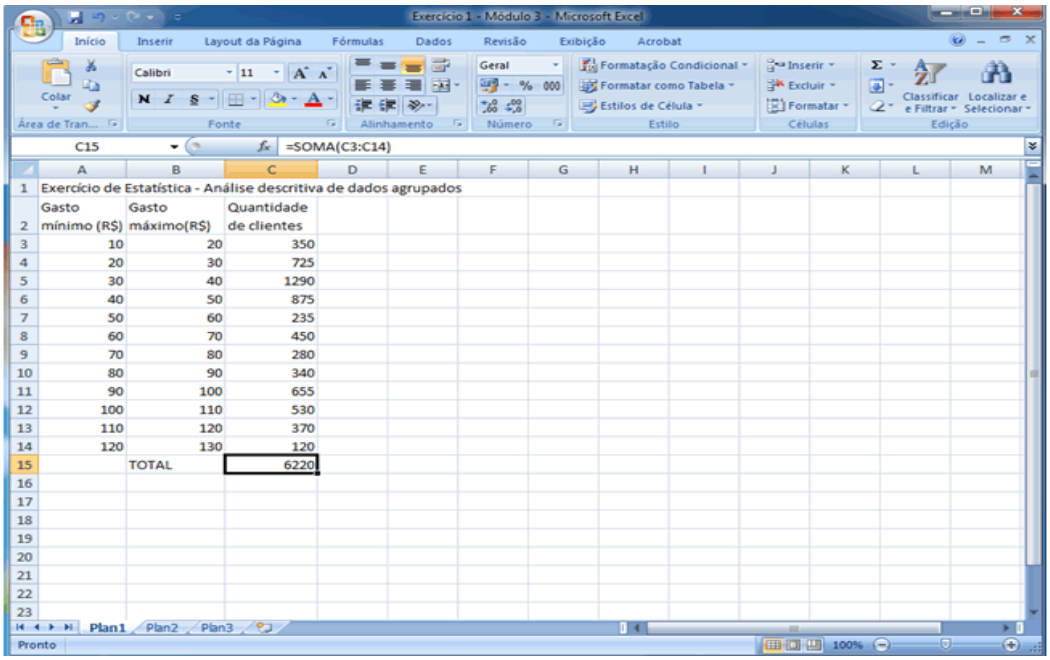
01

1 - DETERMINAÇÃO DE MEDIDAS DESCRITIVAS

A planilha eletrônica de cálculos **Excel** será utilizada para facilitar o processo de determinação de medidas descritivas, sendo que aqui o ponto de partida será um conjunto de dados agrupados em uma distribuição de frequência. O exemplo de referência será o de uma empresa varejista e os gastos realizados por seus clientes nos últimos três meses.

O primeiro passo é abrir uma planilha Excel e digitar os dados, sendo que dessa vez deve-se utilizar uma coluna para os limites inferiores dos intervalos, uma coluna para os limites superiores e uma terceira coluna para as respectivas frequências.

Assim:



Observe que na célula C15 foi digitada uma fórmula para a totalização da quantidade de clientes que formam nossa base de dados.

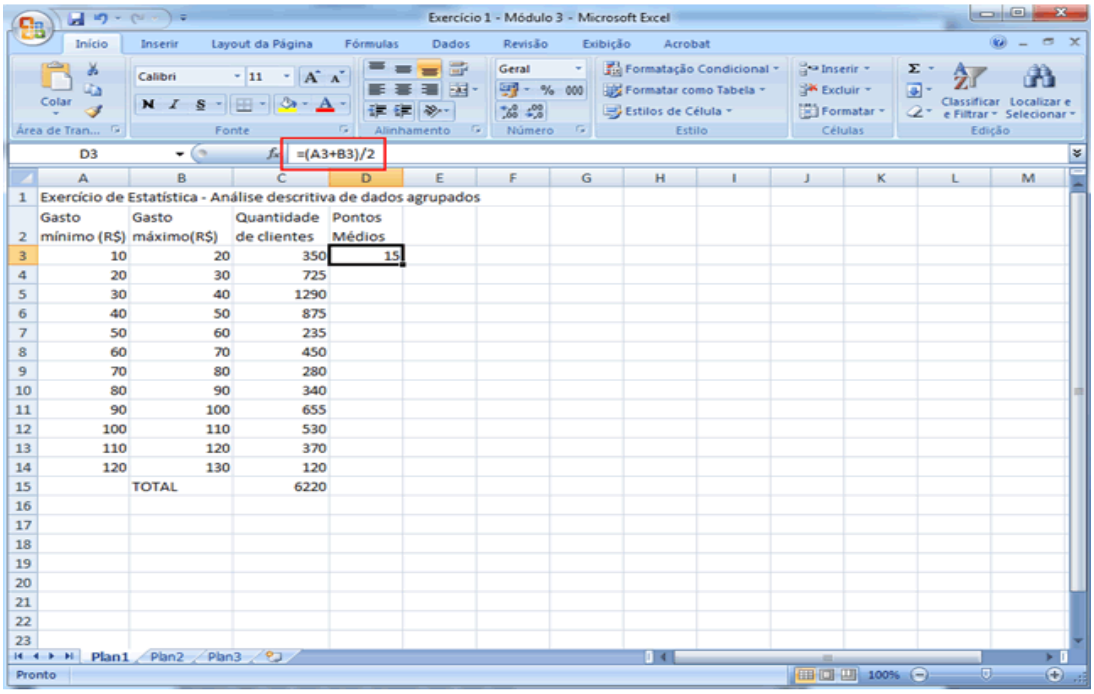
Exemplo

Uma empresa varejista idealiza desenvolver uma campanha de fidelização e para isso levanta um conjunto de dados de seus arquivos referentes às vendas dos últimos três meses, para que possa fazer um "mapeamento" dos valores gastos pelos clientes. O resultado do levantamento está apresentado, resumidamente, abaixo.

Valores Gastos (R$)	Quantidade de Clientes
[10,00 ; 20,00[	350
[20,00 ; 30,00[	725
[30,00 ; 40,00[	1290
[40,00 ; 50,00[	875
[50,00 ; 60,00[	235
[60,00 ; 70,00[	450
[70,00 ; 80,00[	280
[80,00 ; 90,00[	340
[90,00 ; 100,00[	655
[100,00 ; 110,00[	530
[110,00 ; 120,00[	370
[120,00 ; 130,00]	120
TOTAL	6220

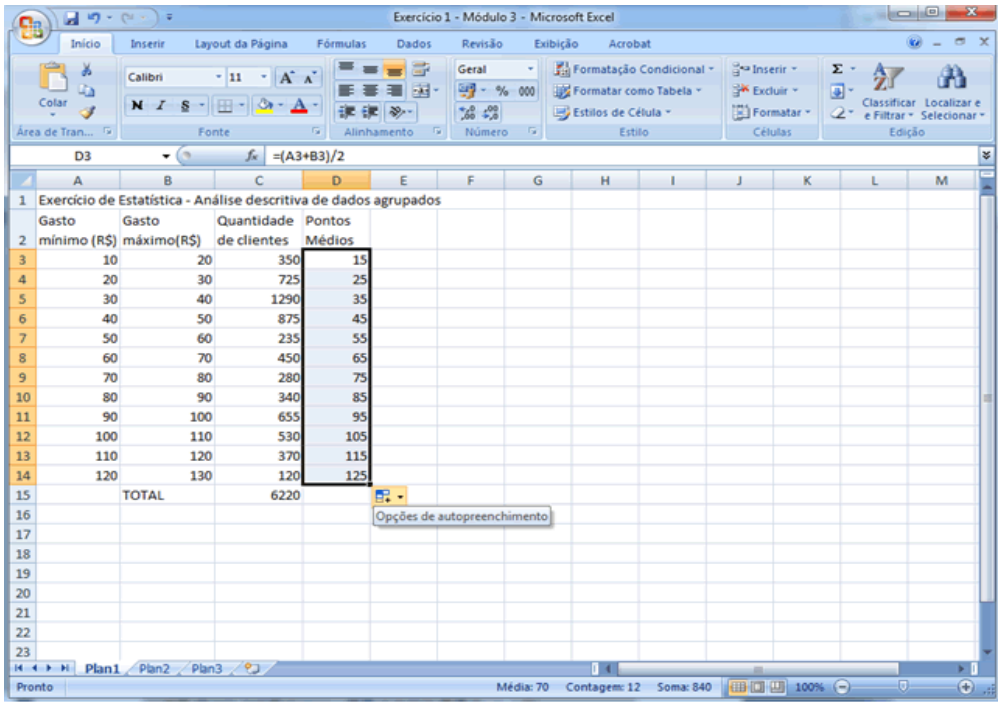
02

Como não é possível o cálculo direto de medidas descritivas, uma vez que os dados individuais não são conhecidos, deve-se, em seguida, inserir uma coluna com os pontos médios dos intervalos, o que será feito na coluna D, a partir da operação de soma dos limites mínimo e máximo e sua divisão por 2.



03

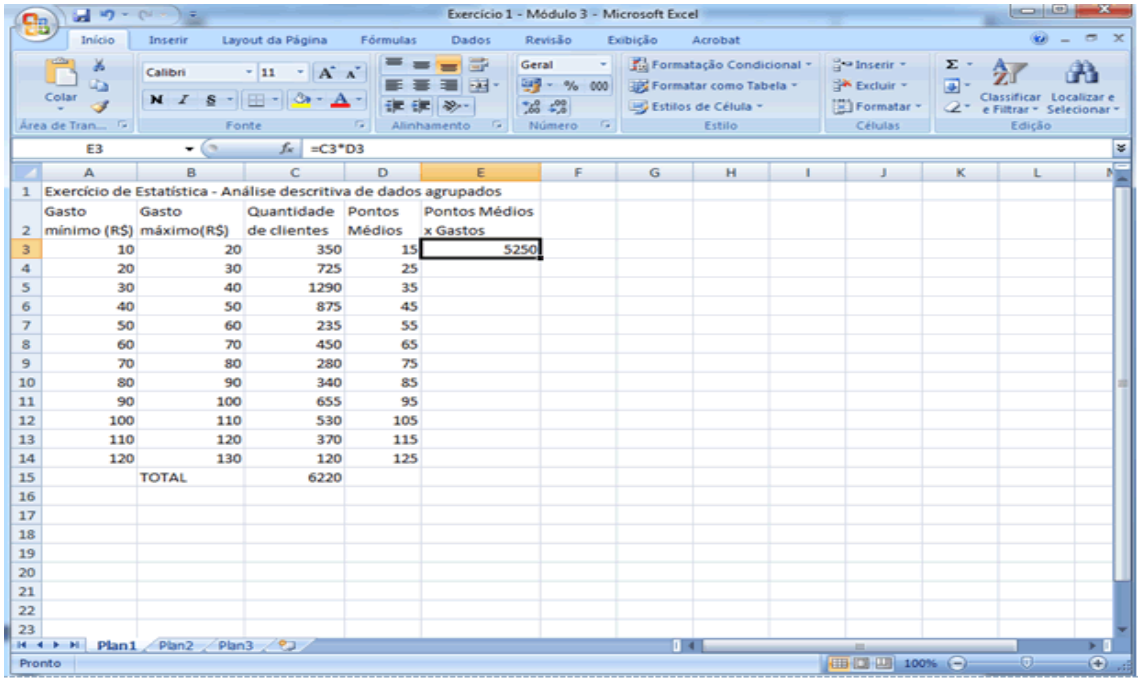
Em seguida, a célula D3 será replicada até D14:



04

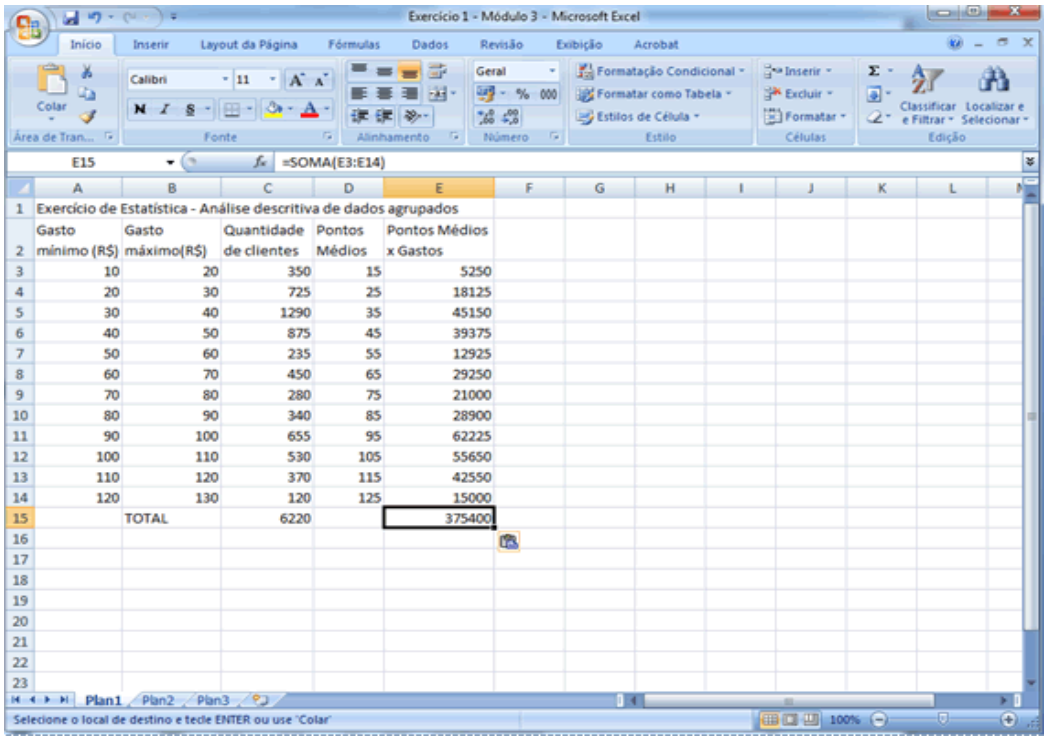
2 - CALCULANDO A MÉDIA

Como a primeira medida será a média, o próximo passo deve ser a multiplicação dos pontos médios dos intervalos de gastos pelas respectivas frequências, o que aparece na coluna E:



05

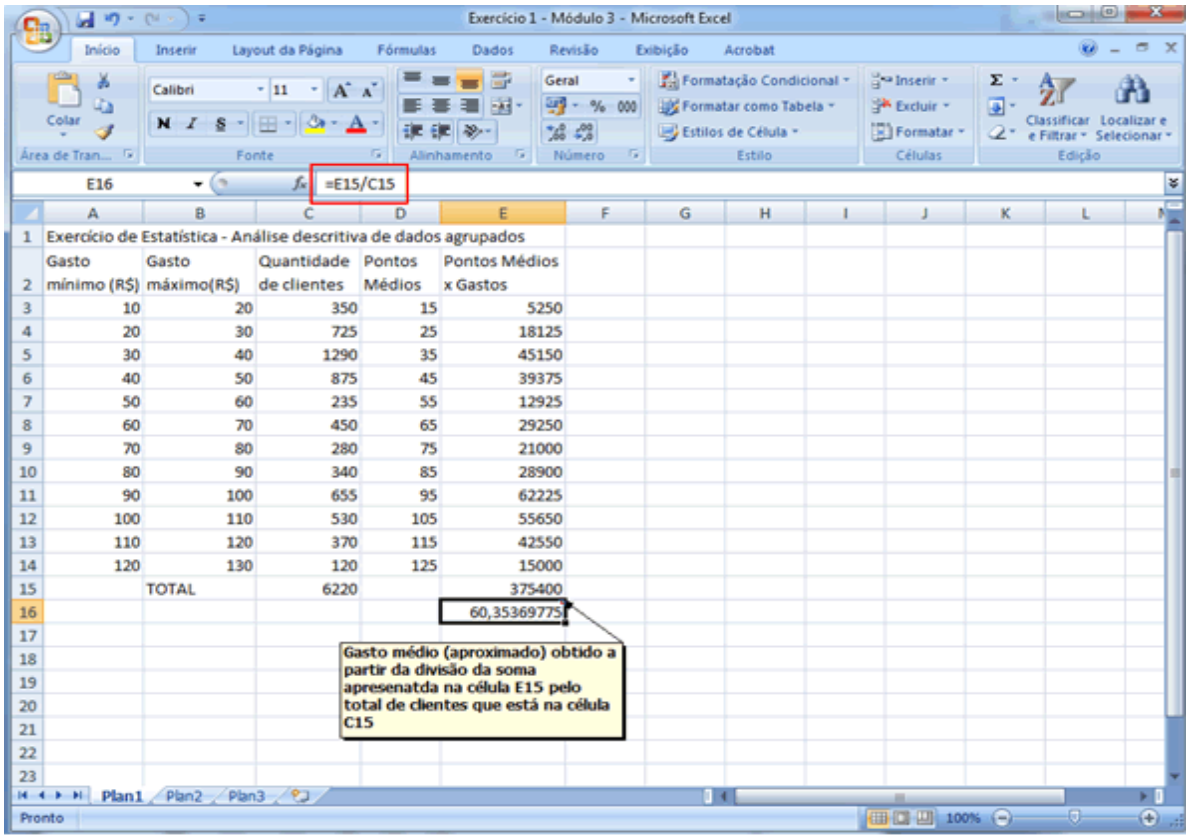
A célula E3 será replicada e na célula E15 será inserida a soma das parcelas, o que é necessário para o cálculo da média.



	A	B	C	D	E
1	Exercício de Estatística - Análise descritiva de dados agrupados				
2	Gasto mínimo (R$)	Gasto máximo(R$)	Quantidade de clientes	Pontos Médios	Pontos Médios x Gastos
3	10	20	350	15	5250
4	20	30	725	25	18125
5	30	40	1290	35	45150
6	40	50	875	45	39375
7	50	60	235	55	12925
8	60	70	450	65	29250
9	70	80	280	75	21000
10	80	90	340	85	28900
11	90	100	655	95	62225
12	100	110	530	105	55650
13	110	120	370	115	42550
14	120	130	120	125	15000
15		TOTAL	6220		375400

06

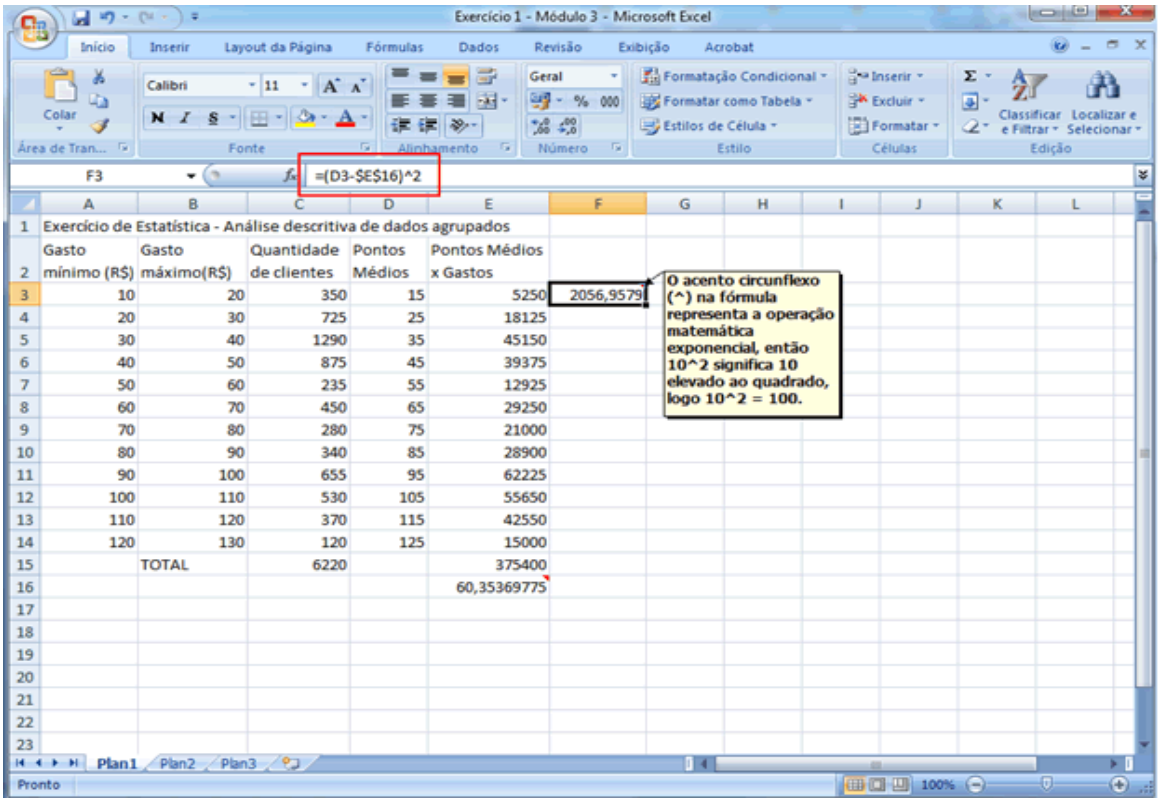
Na posição E16 será apresentada a média dos gastos disponíveis (o comentário inserido esclarece o procedimento adotado).



07

3 - Calculando o Coeficiente de Variação

Para o cálculo do coeficiente de variação, necessário para validação dessa média como medida representativa do conjunto de dados (caso haja homogeneidade), faz-se indispensável o cálculo do desvio-padrão. Assim, na coluna F iniciaremos a preparação para esse cálculo com a subtração da média de cada ponto médio e sua elevação ao quadrado (todo cálculo do coeficiente pode ser feito em uma única célula, a partir da introdução do algoritmo/fórmula adequado para ele e visto no módulo anterior). A forma aqui apresentada tem por objetivo tornar o caminho mais compreensível e didático.

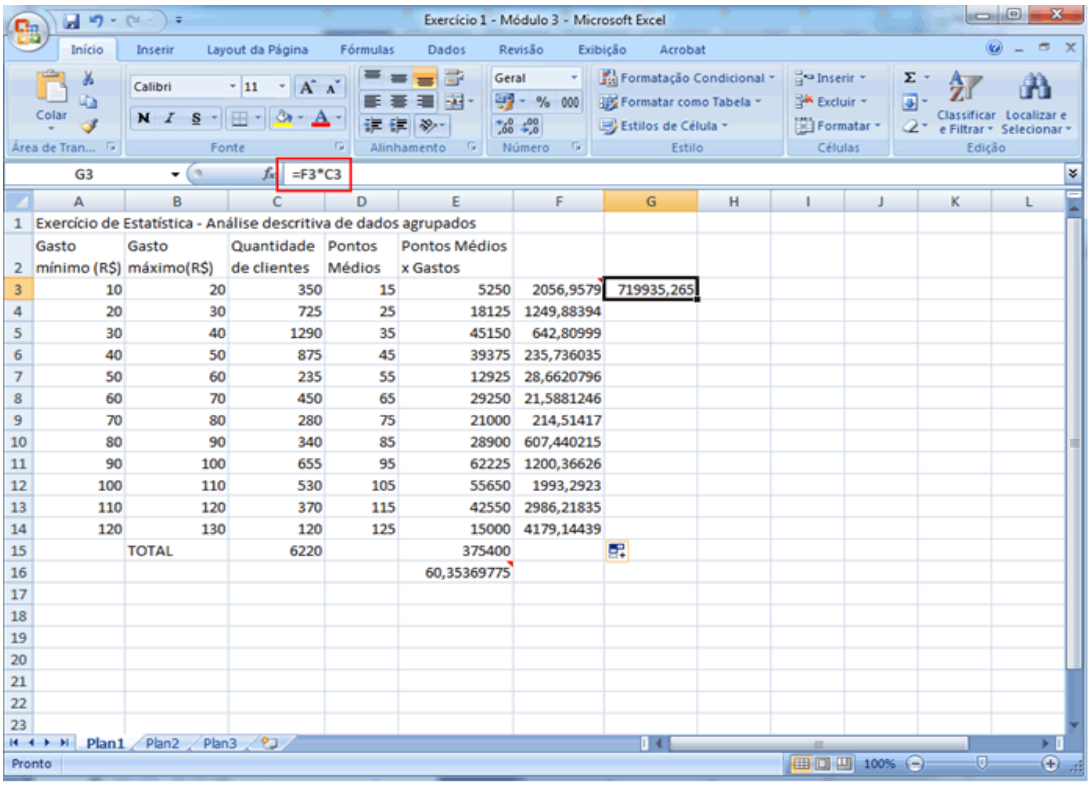


	A	B	C	D	E	F
1	Exercício de Estatística - Análise descritiva de dados agrupados					
2	Gasto mínimo (R$)	Gasto máximo(R$)	Quantidade de clientes	Pontos Médios	Pontos Médios x Gastos	
3	10	20	350	15	5250	2056,9579
4	20	30	725	25	18125	
5	30	40	1290	35	45150	
6	40	50	875	45	39375	
7	50	60	235	55	12925	
8	60	70	450	65	29250	
9	70	80	280	75	21000	
10	80	90	340	85	28900	
11	90	100	655	95	62225	
12	100	110	530	105	55650	
13	110	120	370	115	42550	
14	120	130	120	125	15000	
15		TOTAL	6220		375400	
16					60,35369775	

Observe que os dois símbolos $ foram introduzidos para que, quando a célula F3 for replicada, a posição/
referência correspondente ao valor da média permaneça fixa.

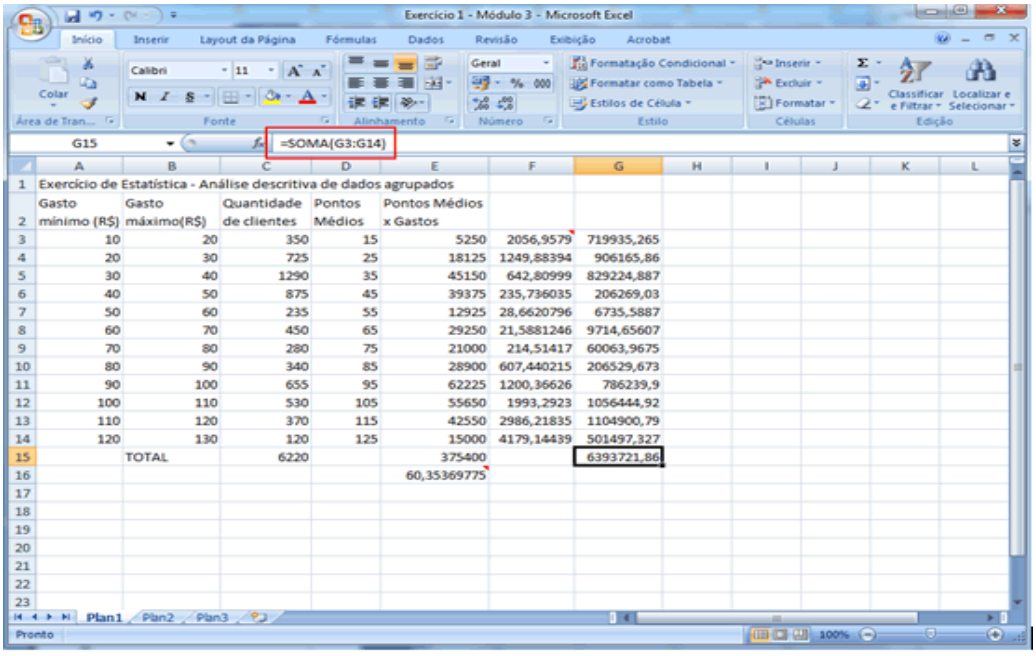
08

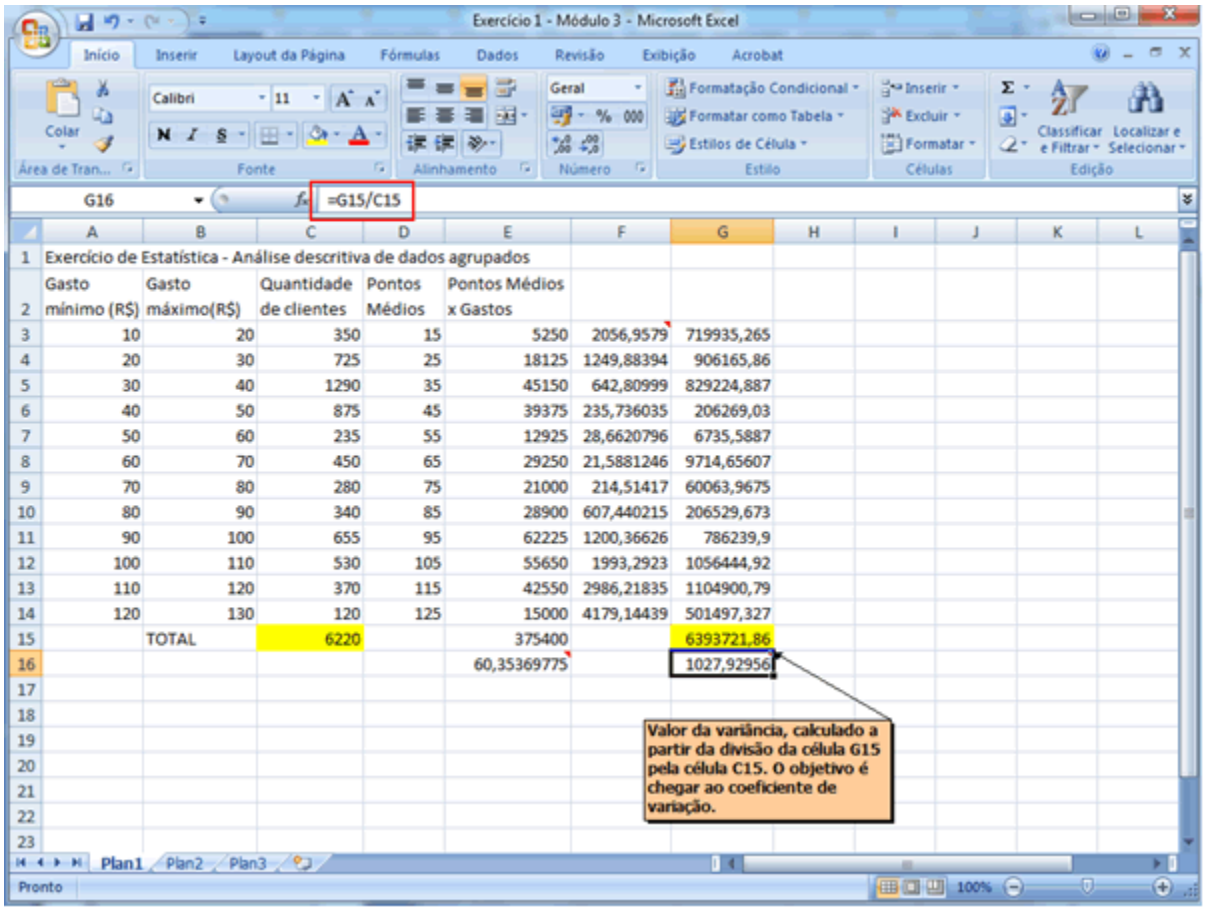
Na coluna G, serão calculados os produtos da coluna F pelas respectivas frequências (constantes na coluna C).



09

A célula G3 será replicada e na célula G15 será calculada a soma das parcelas dessa coluna, para que posteriormente esse valor possa ser dividido pela frequência total (o que gerará a variância, na célula G16).

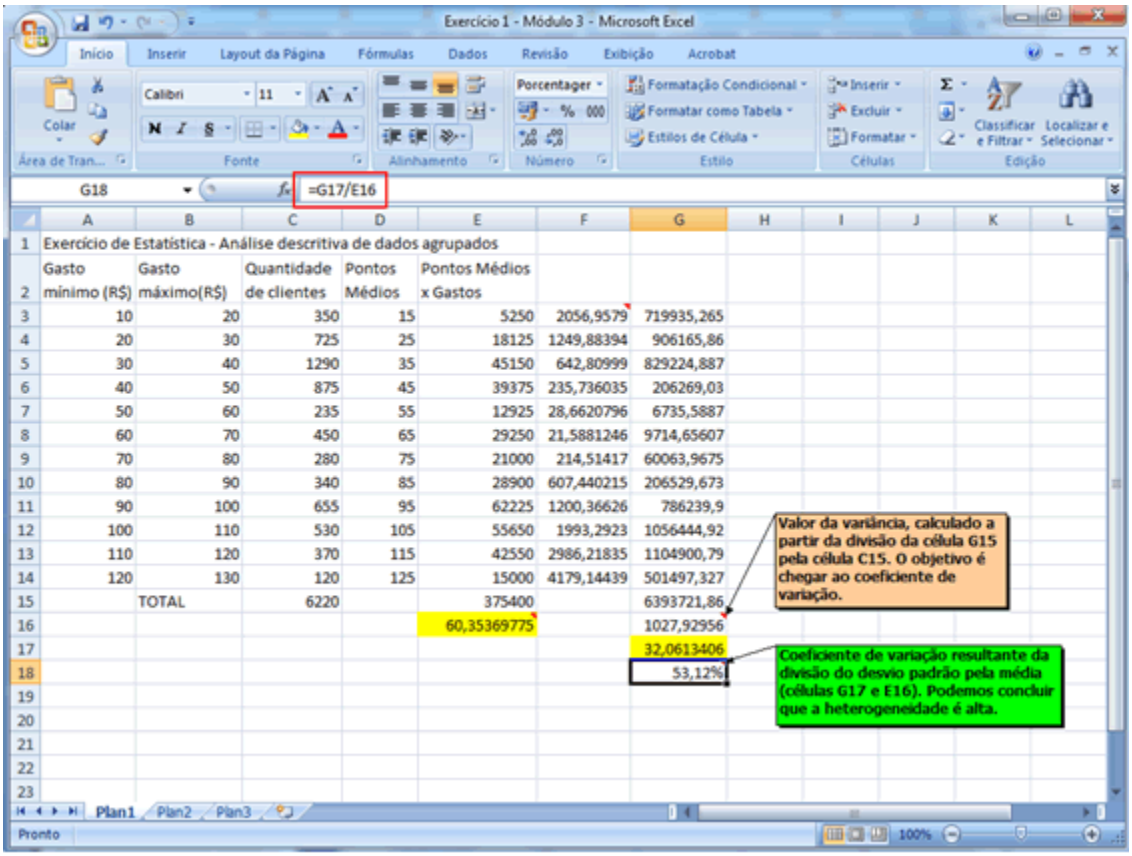




	A	B	C	D	E	F	G
1	Exercício de Estatística - Análise descritiva de dados agrupados						
2	Gasto mínimo (R$)	Gasto máximo(R$)	Quantidade de clientes	Pontos Médios	Pontos Médios x Gastos		
3	10	20	350	15	5250	2056,9579	719935,265
4	20	30	725	25	18125	1249,88394	906165,86
5	30	40	1290	35	45150	642,80999	829224,887
6	40	50	875	45	39375	235,736035	206269,03
7	50	60	235	55	12925	28,6620796	6735,5887
8	60	70	450	65	29250	21,5881246	9714,65607
9	70	80	280	75	21000	214,51417	60063,9675
10	80	90	340	85	28900	607,440215	206529,673
11	90	100	655	95	62225	1200,36626	786239,9
12	100	110	530	105	55650	1993,2923	1056444,92
13	110	120	370	115	42550	2986,21835	1104900,79
14	120	130	120	125	15000	4179,14439	501497,327
15		TOTAL	6220		375400		6393721,86
16					60,35369775		1027,92956

10

Parte-se agora para o coeficiente de variação, sendo primeiramente calculado o desvio-padrão (raiz quadrada da variância). O desvio padrão é calculado na célula G17 utilizando a função RAIZ() do Excel (escrevemos na célula G17: =RAIZ(G16)). E, em seguida, fazendo a divisão do desvio-padrão pela média, obtemos o coeficiente de variação (célula G18).

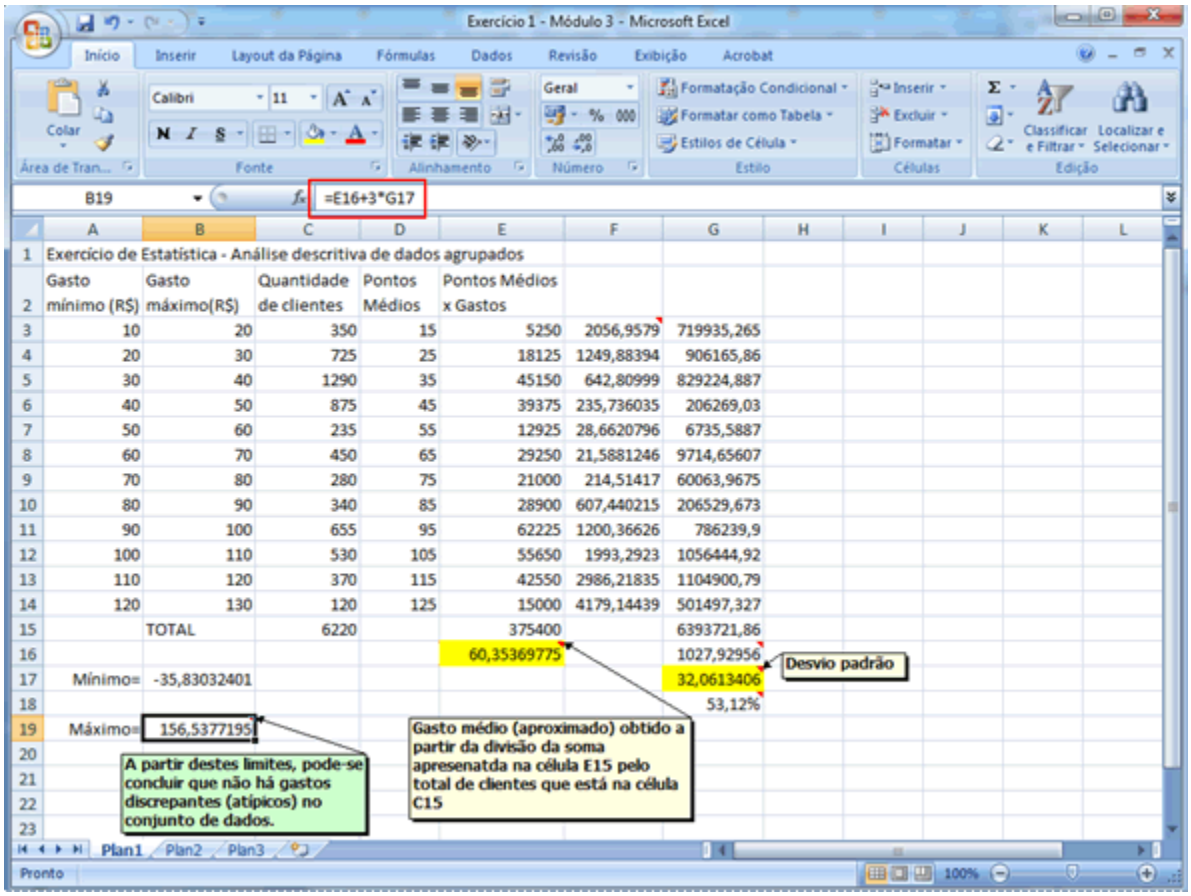


	A	B	C	D	E	F	G
1	Exercício de Estatística - Análise descritiva de dados agrupados						
2	Gasto mínimo (R$)	Gasto máximo(R$)	Quantidade de clientes	Pontos Médios	Pontos Médios x Gastos		
3	10	20	350	15	5250	2056,9579	719935,265
4	20	30	725	25	18125	1249,88394	906165,86
5	30	40	1290	35	45150	642,80999	829224,887
6	40	50	875	45	39375	235,736035	206269,03
7	50	60	235	55	12925	28,6620796	6735,5887
8	60	70	450	65	29250	21,5881246	9714,65607
9	70	80	280	75	21000	214,51417	60063,9675
10	80	90	340	85	28900	607,440215	206529,673
11	90	100	655	95	62225	1200,36626	786239,9
12	100	110	530	105	55650	1993,2923	1056444,92
13	110	120	370	115	42550	2986,21835	1104900,79
14	120	130	120	125	15000	4179,14439	501497,327
15		TOTAL	6220		375400		6393721,86
16					60,35369775		1027,92956
17							32,0613406
18							53,12%

11

4 - DIAGNOSTICANDO PONTOS DISCREPANTES

A determinação dos limites máximo e mínimo para verificação da existência de pontos discrepantes será feita a partir do cálculo do triplo do desvio-padrão e posterior soma e subtração à média, inseridos nas células B17 e B19 da planilha de trabalho (digitando-se naquelas células, respectivamente, =E16-3*G17 e =E16+3*G17).



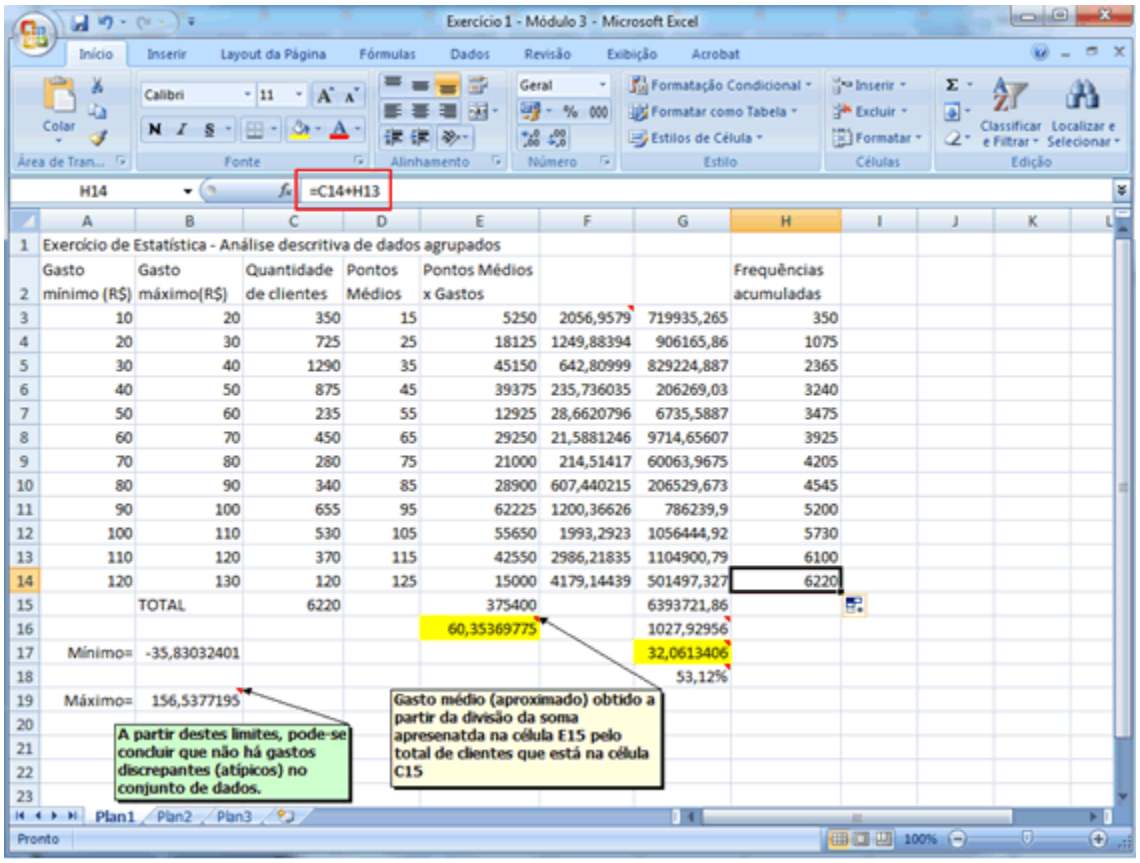
12

5 - Calculando as Separatrizes

Em continuidade à determinação das medidas descritivas, em particular porque o conjunto de dados é consideravelmente heterogêneo, será usada a coluna H para apresentação das frequências acumuladas, passo intermediário necessário para se chegar às separatrizes.

Deve-se lembrar de que o cálculo da moda não é adequado (do ponto de vista prático) quando os dados estão agrupados, sendo muito mais conveniente falar em classe modal (cuja identificação não demanda nenhum cálculo).

Com o cursor na célula H3 digita-se =C3, célula correspondente à frequência da primeira classe. Na célula H4 digita-se =H3+C4 (frequência da primeira classe somada à da segunda), na célula H5 digita-se =H4+C5 (frequência das duas primeiras classes somada à da terceira) e assim sucessivamente. Observe que o total em H14 deve "bater" com aquele já calculado anteriormente e apresentado na célula C15.



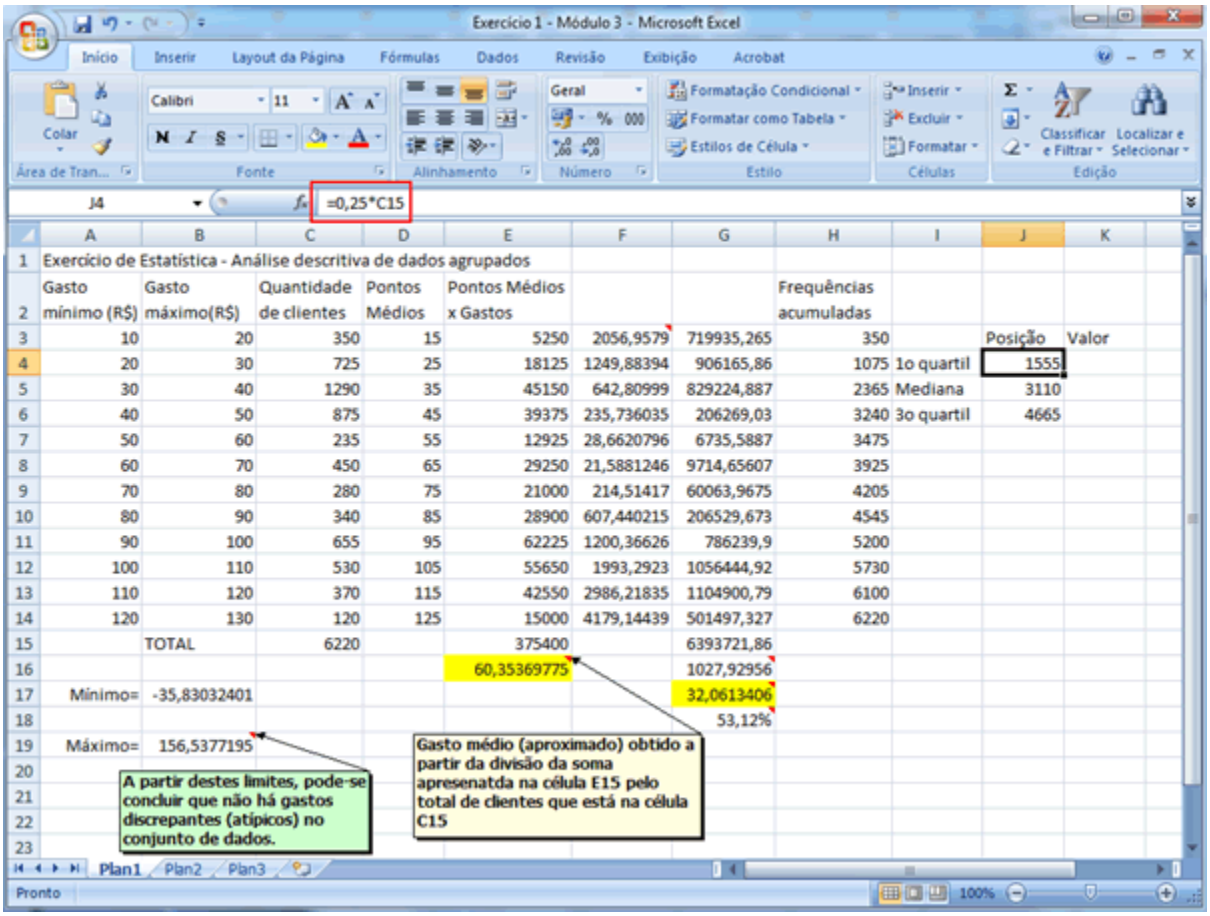
13

O cálculo dos quartis (incluindo a mediana) parte da expressão, já apresentada:

$$Quartil = \text{limite inferior do intervalo} + \left(\frac{\text{posição do quartil} - \text{frequencia acumulada até classe anterior}}{\text{frequencia do intervalo}}\right) \text{x amplitude do intervalo}$$

E posição do quartil = 0,25 x quantidade de observações (para o primeiro quartil) ou
0,5 x quantidade de observações (para o segundo quartil/mediana) ou
0,5 x quantidade de observações (para o primeiro quartil) ou

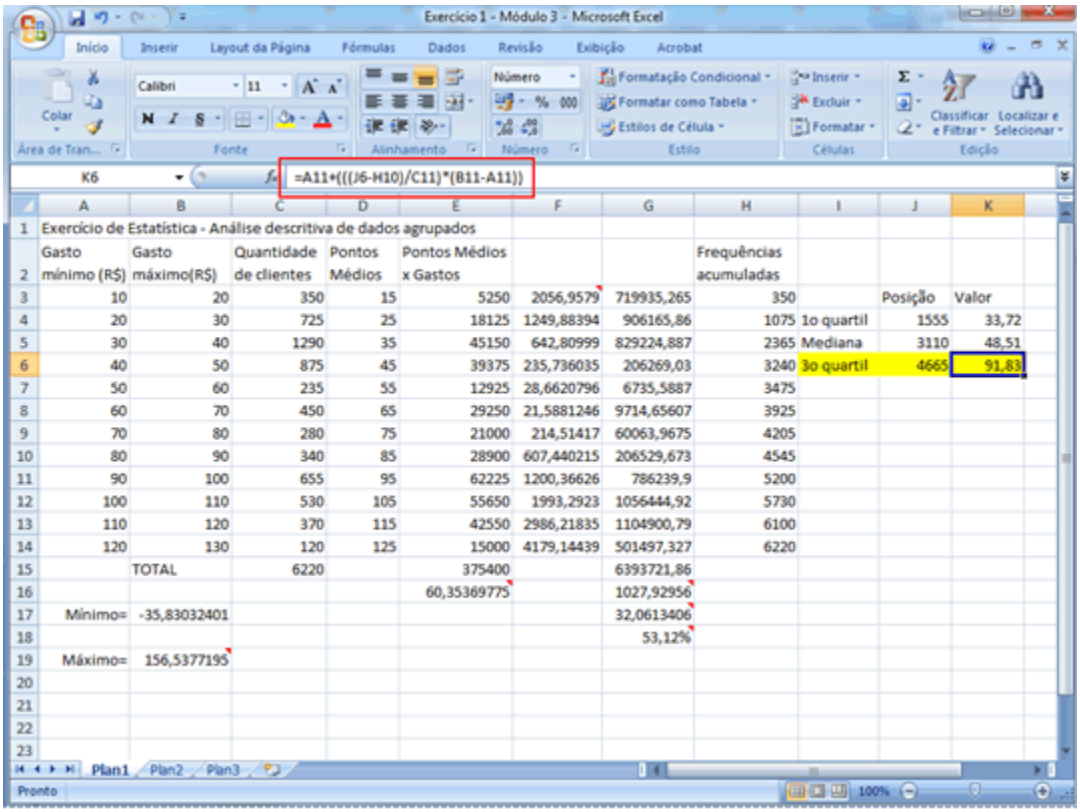
Primeiramente serão determinadas as posições dos quartis nas células J4, J5 e J6, digitando =0,25*C15 (para o primeiro quartil), =0,5*C15 (para o segundo quartil, que é a mediana) e =0,75*C15 (para o terceiro quartil).



14

O segundo passo será determinar os valores efetivos dos quartis a partir da formulação já discutida anteriormente e reapresentada no passo anterior. Na célula K4, para o primeiro quartil, deve-se digitar =A5+(((J4-H4)/C5)*(B5-A5)), pois na célula A5 está o limite inferior da classe à qual pertence o quartil, em J4 está a posição do quartil, em H4 a frequência acumulada até a classe anterior a do quartil, em C5 está a frequência da classe na qual está o quartil, e a operação B5-A5 gera a amplitude da classe na qual está o quartil.

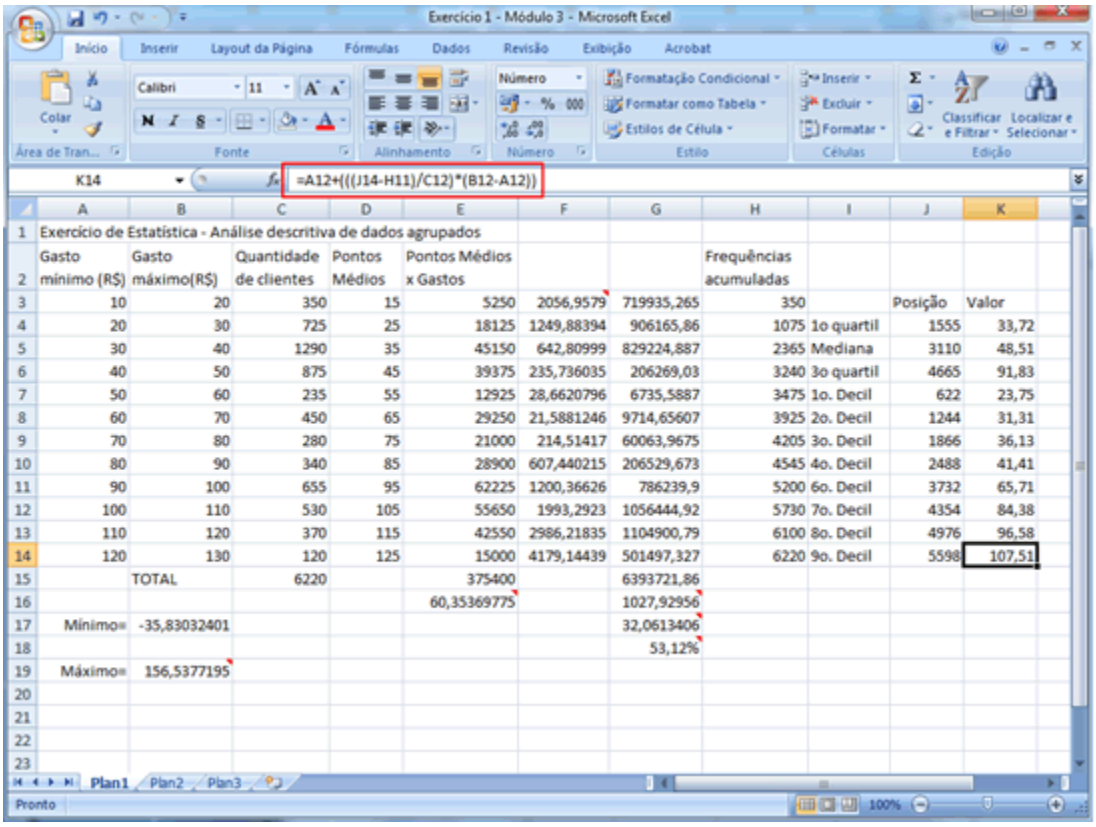
O procedimento para a mediana e terceiro quartis, apresentados nas células K5 e K6, respectivamente, é análogo.



	A	B	C	D	E	F	G	H	I	J	K
1	Exercício de Estatística - Análise descritiva de dados agrupados										
2	Gasto mínimo (R$)	Gasto máximo(R$)	Quantidade de clientes	Pontos Médios	Pontos Médios x Gastos			Frequências acumuladas			
3	10	20	350	15	5250	2056,9579	719935,265	350		Posição	Valor
4	20	30	725	25	18125	1249,88394	906165,86	1075	1o quartil	1555	33,72
5	30	40	1290	35	45150	642,80999	829224,887	2365	Mediana	3110	48,51
6	40	50	875	45	39375	235,736035	206269,03	3240	3o quartil	4665	91,83
7	50	60	235	55	12925	28,6620796	6735,5887	3475			
8	60	70	450	65	29250	21,5881246	9714,65607	3925			
9	70	80	280	75	21000	214,51417	60063,9675	4205			
10	80	90	340	85	28900	607,440215	206529,673	4545			
11	90	100	655	95	62225	1200,36626	786239,9	5200			
12	100	110	530	105	55650	1993,2923	1056444,92	5730			
13	110	120	370	115	42550	2986,21835	1104900,79	6100			
14	120	130	120	125	15000	4179,14439	501497,327	6220			
15		TOTAL	6220		375400		6393721,86				
16					60,35369775		1027,92956				
17	Mínimo=	-35,83032401					32,0613406				
18							53,12%				
19	Máximo=	156,5377195									
20											
21											
22											
23											

15

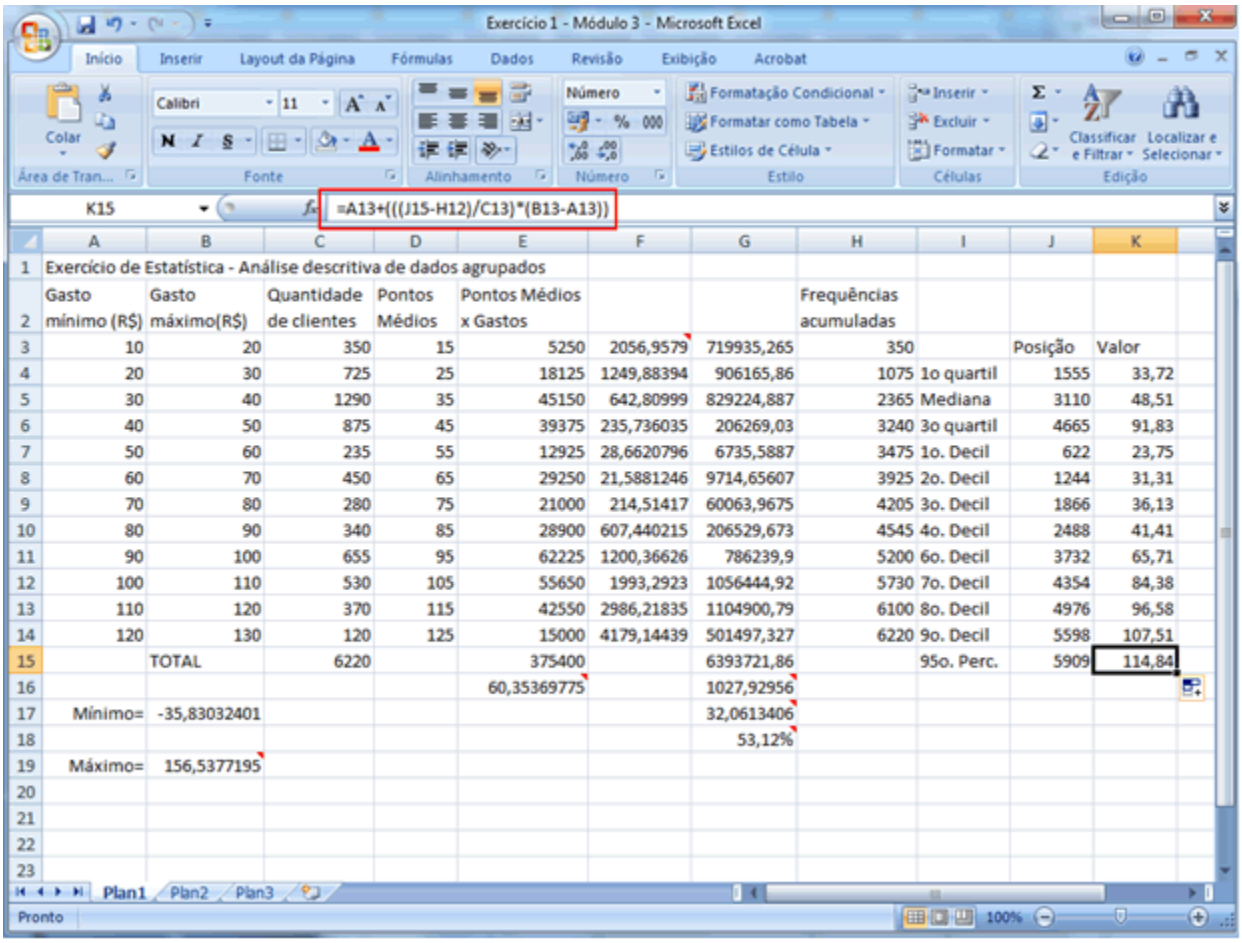
A determinação dos decis obedece a mesma sequência dos quartis (lembrando que o quinto decil, sendo a própria mediana, já está determinado). Assim, nas células de J7 a J15 estarão as posições dessas separatrizes (=0,1*C15, = 0,2*C15, =0,3*C15 e assim sucessivamente) e, em seguida, nas células de K7 a K15 os valores correspondentes aos oito decis ainda não conhecidos (lembrando a formulação especificada já demonstrada).



	A	B	C	D	E	F	G	H	I	J	K
1	Exercício de Estatística - Análise descritiva de dados agrupados										
2	Gasto mínimo (R$)	Gasto máximo(R$)	Quantidade de clientes	Pontos Médios	Pontos Médios x Gastos			Frequências acumuladas			
3	10	20	350	15	5250	2056,9579	719935,265	350		Posição	Valor
4	20	30	725	25	18125	1249,88394	906165,86	1075	1o quartil	1555	33,72
5	30	40	1290	35	45150	642,80999	829224,887	2365	Mediana	3110	48,51
6	40	50	875	45	39375	235,736035	206269,03	3240	3o quartil	4665	91,83
7	50	60	235	55	12925	28,6620796	6735,5887	3475	1o. Decil	622	23,75
8	60	70	450	65	29250	21,5881246	9714,65607	3925	2o. Decil	1244	31,31
9	70	80	280	75	21000	214,51417	60063,9675	4205	3o. Decil	1866	36,13
10	80	90	340	85	28900	607,440215	206529,673	4545	4o. Decil	2488	41,41
11	90	100	655	95	62225	1200,36626	786239,9	5200	6o. Decil	3732	65,71
12	100	110	530	105	55650	1993,2923	1056444,92	5730	7o. Decil	4354	84,38
13	110	120	370	115	42550	2986,21835	1104900,79	6100	8o. Decil	4976	96,58
14	120	130	120	125	15000	4179,14439	501497,327	6220	9o. Decil	5598	107,51
15		TOTAL	6220		375400		6393721,86				
16					60,35369775		1027,92956				
17	Mínimo=	-35,83032401					32,0613406				
18							53,12%				
19	Máximo=	156,5377195									
20											
21											
22											
23											

16

Para qualquer percentil não há novidade, devendo o procedimento ser análogo ao que foi feito para quartis e decis (lembrando que todas essas medidas já calculadas são percentis). Apenas para concluir, será apresentado o cálculo do 95º percentil nas células J15 e K15.



17

RESUMO

Neste módulo ilustrou-se o cálculo de várias medidas descritivas com o apoio da planilha eletrônica Microsoft Excel. A sequência de passos necessária para viabilizar é (lembrando que para dados agrupados não há soluções "automáticas", sendo necessário preparar a planilha da forma mais adequada e utilizar o conjunto de fórmulas trabalhadas naquele módulo):

(a) abrir uma planilha e digitar a "tabela" de dados que será trabalhada, cuidando para que os limites inferiores de cada classe ocupem uma coluna, os superiores outra e as frequências uma terceira coluna;

(b) criar uma quarta coluna na qual devem ser calculados os pontos médios de cada classe;

(c) em uma quinta coluna calcular o produto dos pontos médios pelas respectivas frequências, já disponíveis em outras duas colunas (sempre iniciando a digitação por =);

(d) selecionar uma célula na qual se deseja inserir a média e digitar a fórmula adequada iniciada pelo sinal =;

(e) para o coeficiente de variação, inicialmente deve-se chegar ao desvio-padrão, o que se faz com o auxílio de mais duas colunas nas quais se desdobra a fórmula necessária para determinação da variância, após o que se calcula a raiz quadrada e divide-se pela média (em duas células selecionadas para isso);

(f) para as separatrizes, incluindo quartis, decis e percentis, mais uma vez há utilização de duas colunas, uma para determinação das posições das separatrizes desejadas e outra para determinação dos seus valores.

É conveniente inserir comentários para cada uma das medidas calculadas, o que pode ser feito com a sequência Seleção da Célula -->Inserir -->Comentário --> Digitação do comentário ou então com a digitação direta do comentário em célula próxima àquela na qual está o valor da medida.

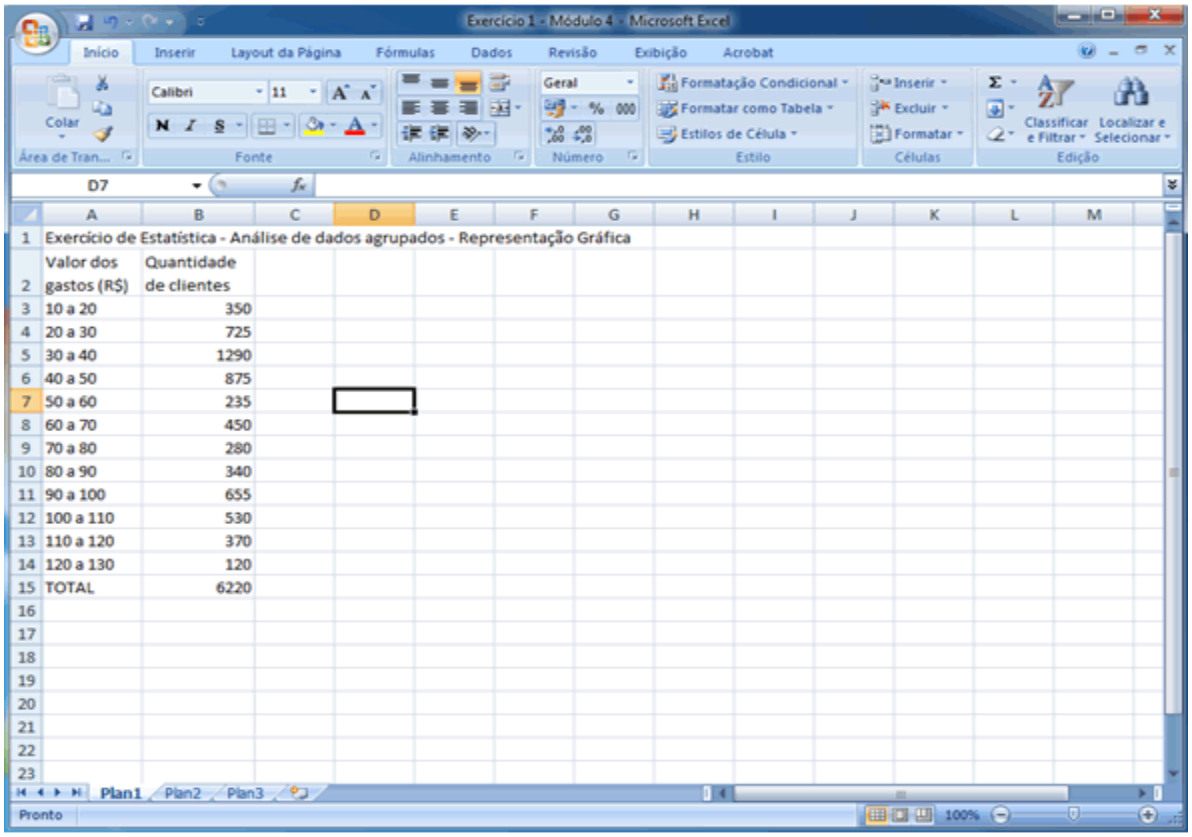
UNIDADE 2 – DADOS AGRUPADOS

MÓDULO 4 – DESCRIÇÃO GRÁFICA DE DADOS AGRUPADOS

01

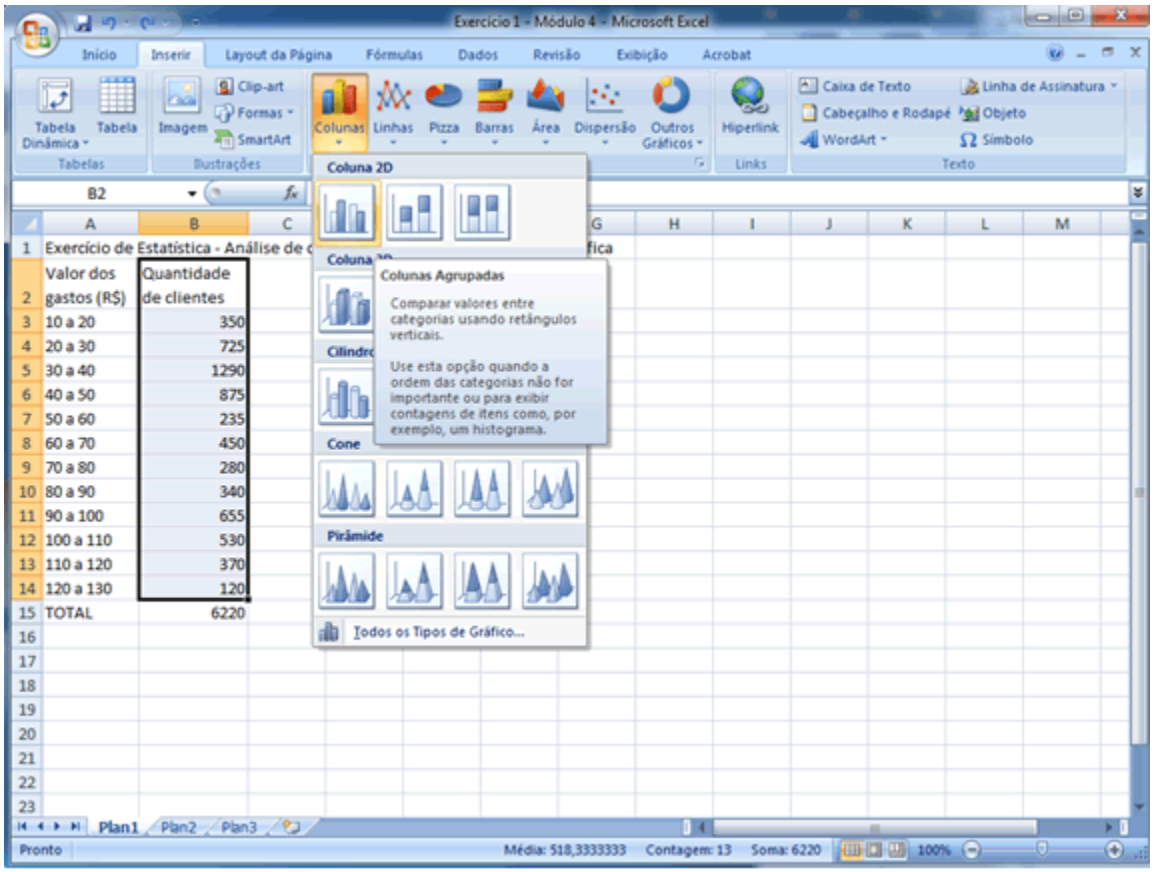
1 - REPRESENTAÇÃO GRÁFICA EM COLUNAS

Iniciaremos este módulo buscando representar graficamente o conjunto de dados agrupados.
Isso será feito com o auxílio do Excel.



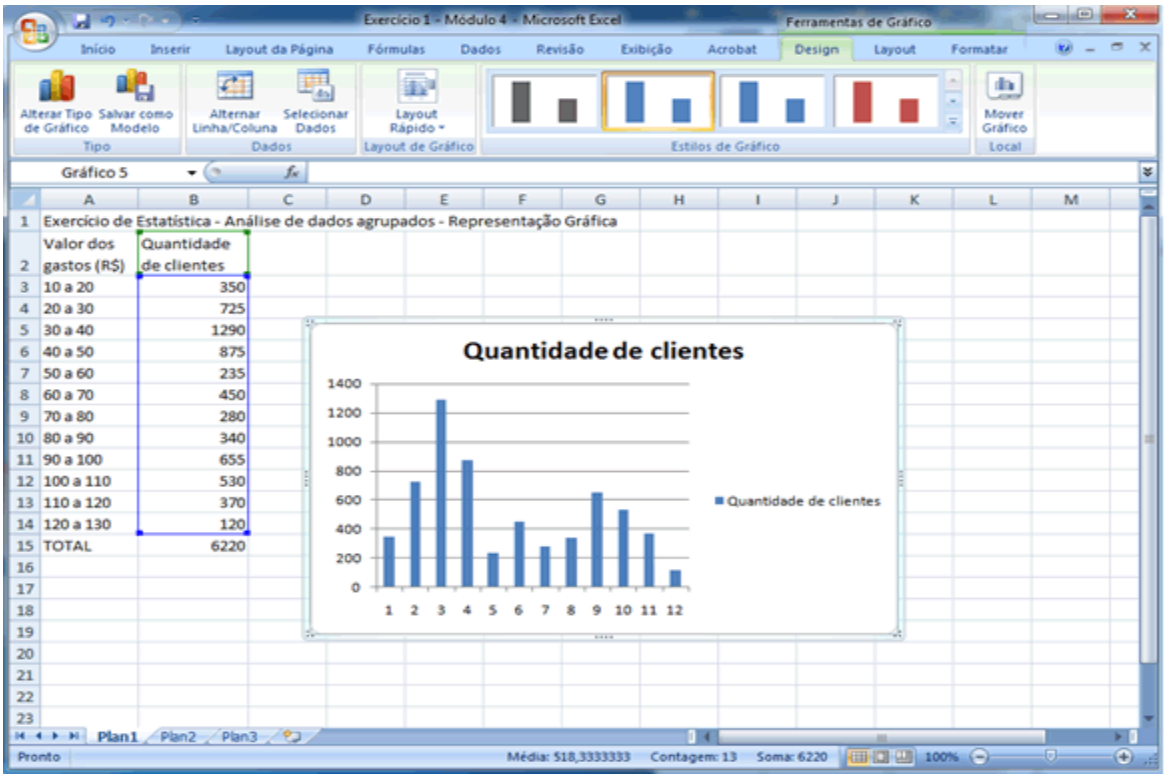
02

Vamos então selecionar o título e os dados da coluna "Quantidade de clientes" (B2 a B14). Em seguida, clicaremos na guia Inserir e escolheremos o tipo de gráfico a ser criado. No nosso caso iremos escolher um simples gráfico de colunas:

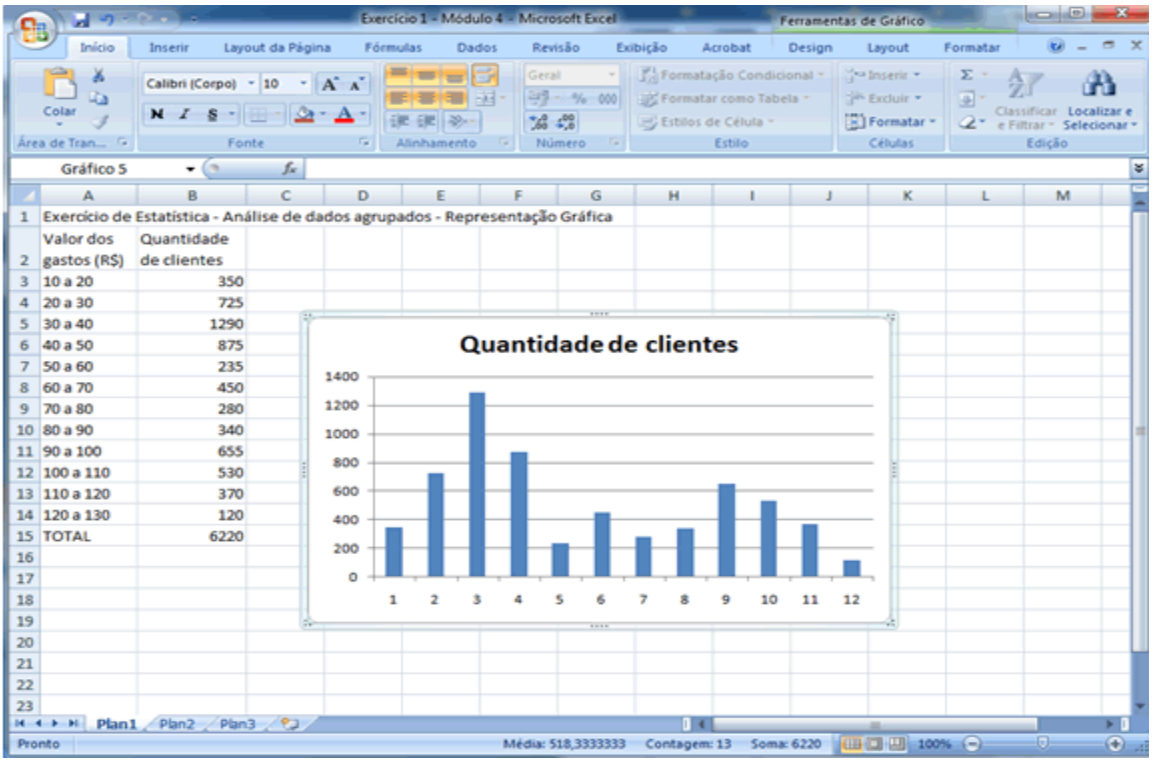


03

Aparecerá então o gráfico já com título e a série de dados que selecionamos. Observe que o gráfico ficou pequeno por causa da legenda.

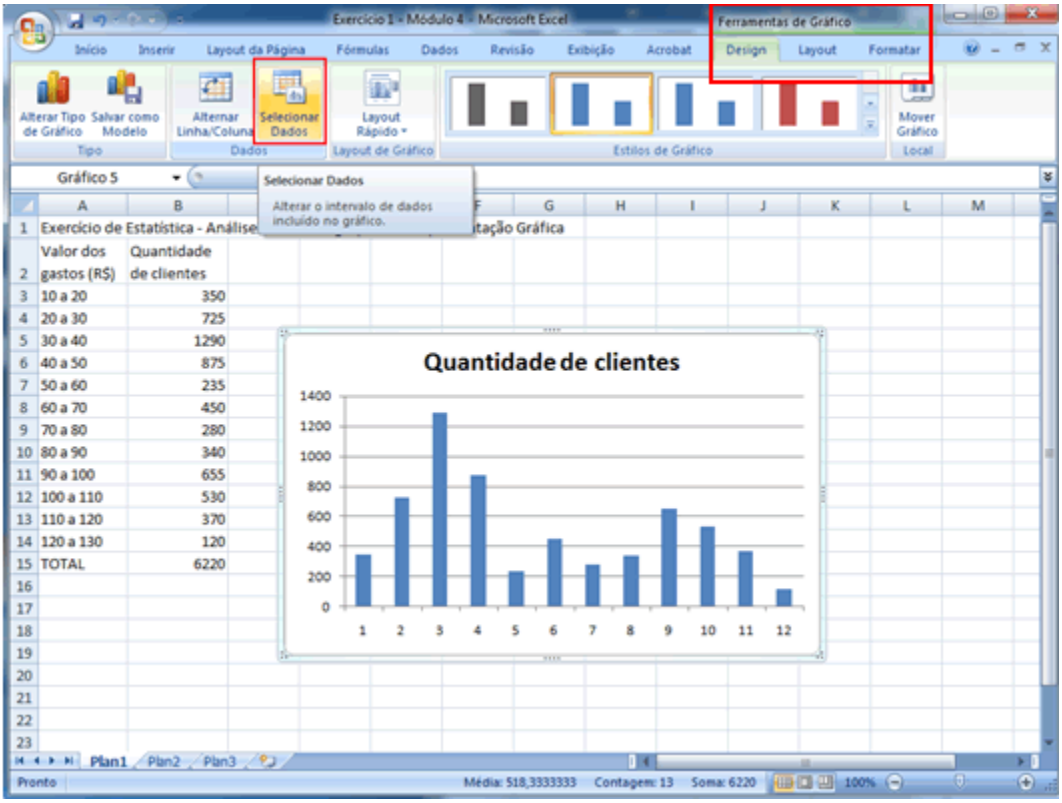


Para resolver isso, vamos então clicar na legenda e em seguida, no botão "DEL" para apagá-la. Verifique que o gráfico ficou maior.



04

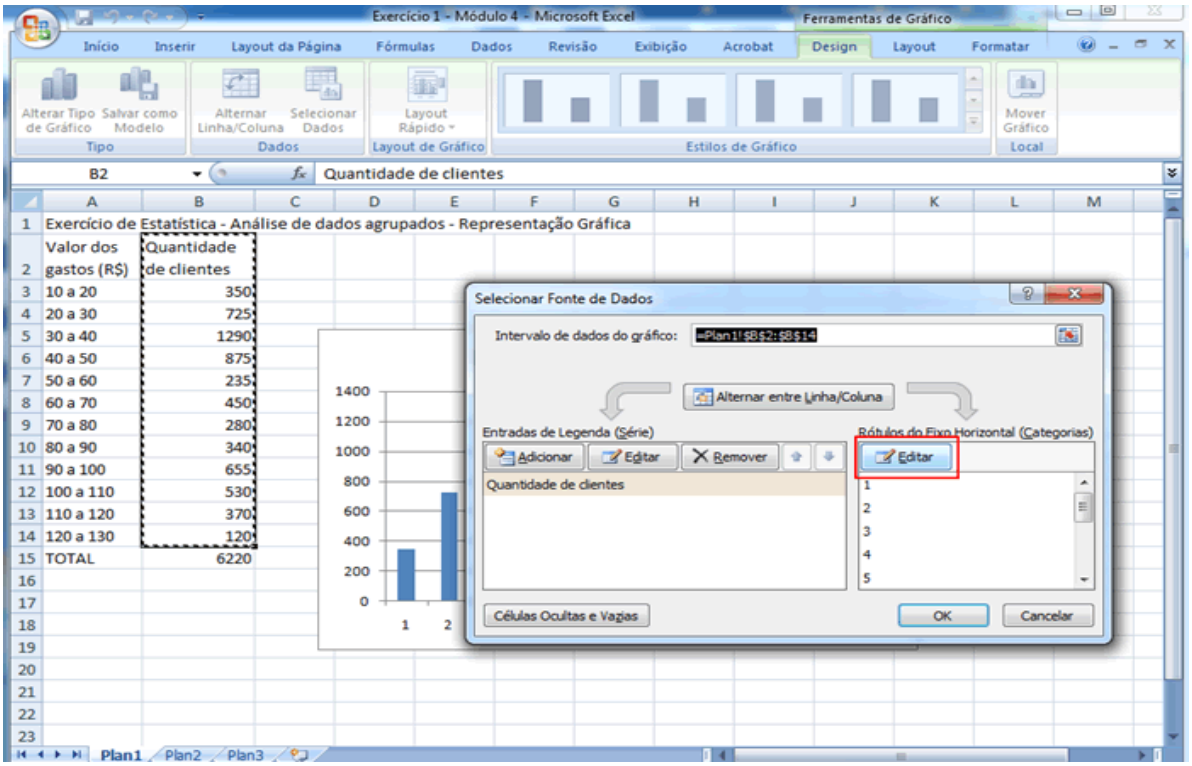
Vamos agora alterar os dados que aparecem no eixo x, pois queremos que apareça o valor dos gastos. Para isso iremos selecionar a guia "Design" e clicar no botão selecionar dados. (Importante: as guias que fazem parte das ferramentas de Gráfico: Design, Layout e Formatar só aparecem se o gráfico estiver selecionado!)



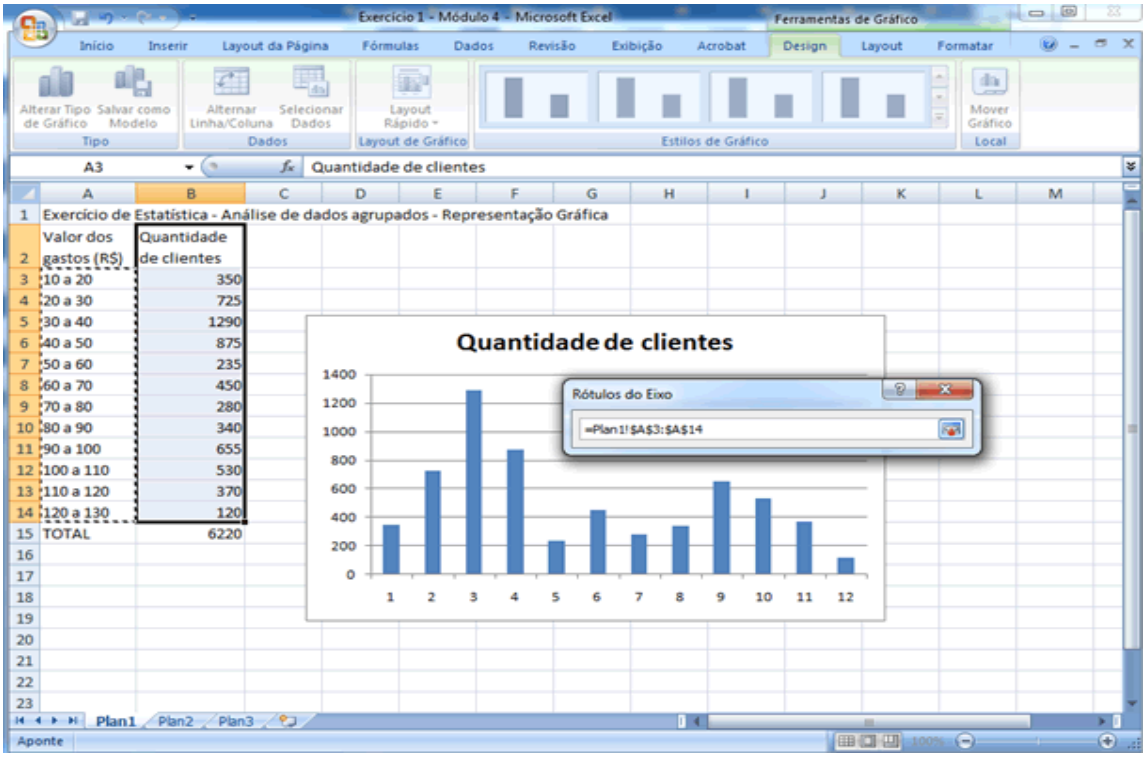
	A	B
1	Exercício de Estatística - Análise	tação Gráfica
2	Valor dos gastos (R$)	Quantidade de clientes
3	10 a 20	350
4	20 a 30	725
5	30 a 40	1290
6	40 a 50	875
7	50 a 60	235
8	60 a 70	450
9	70 a 80	280
10	80 a 90	340
11	90 a 100	655
12	100 a 110	530
13	110 a 120	370
14	120 a 130	120
15	TOTAL	6220

05

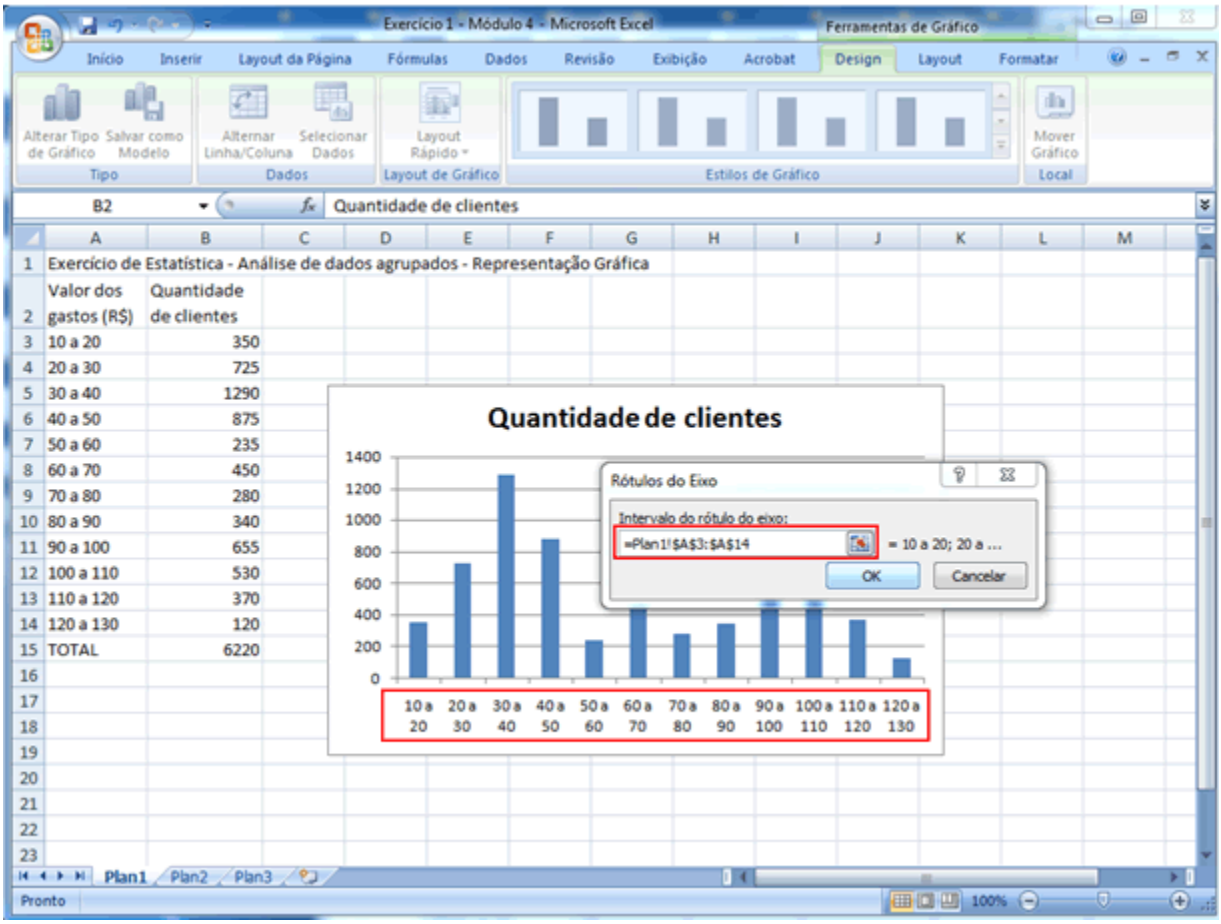
Aparecerá a caixa de diálogo que permite selecionar as fontes de dados e alterar o chamado rótulo do eixo horizontal:



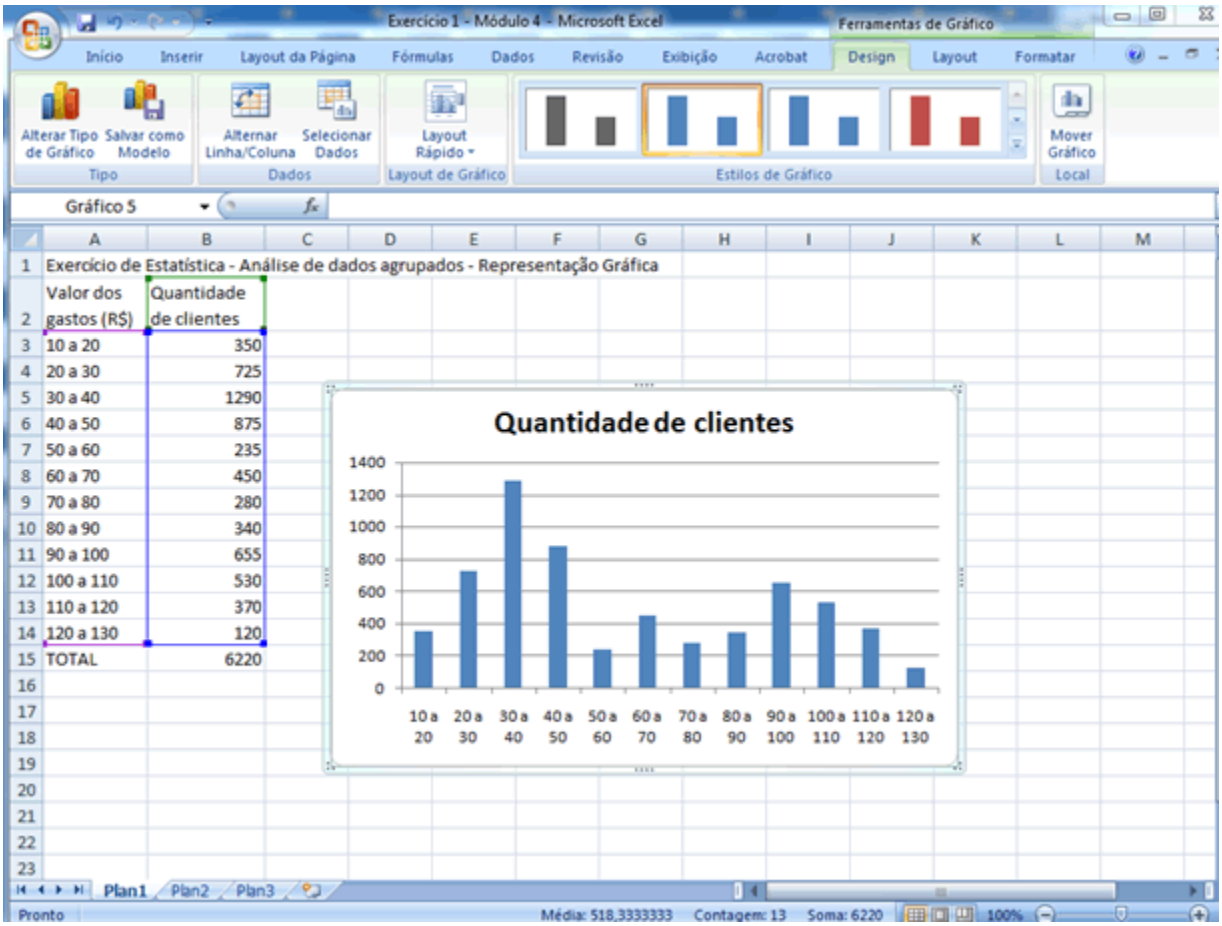
Clique em "Editar", logo depois em  e selecione os dados da coluna "valor dos gastos". Em seguida clique ENTER.



Podemos perceber que já é apresentada uma pré-visualização dos rótulos.



Basta clicar em OK e novamente em OK para confirmar a alteração.



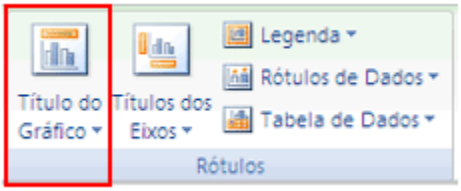
Claro que podemos melhorar a aparência do gráfico que construímos. Para isso vamos utilizar as guias que fazem parte das Ferramentas de Gráfico que falamos anteriormente. Vamos conhecer um pouco mais essas ferramentas.

07

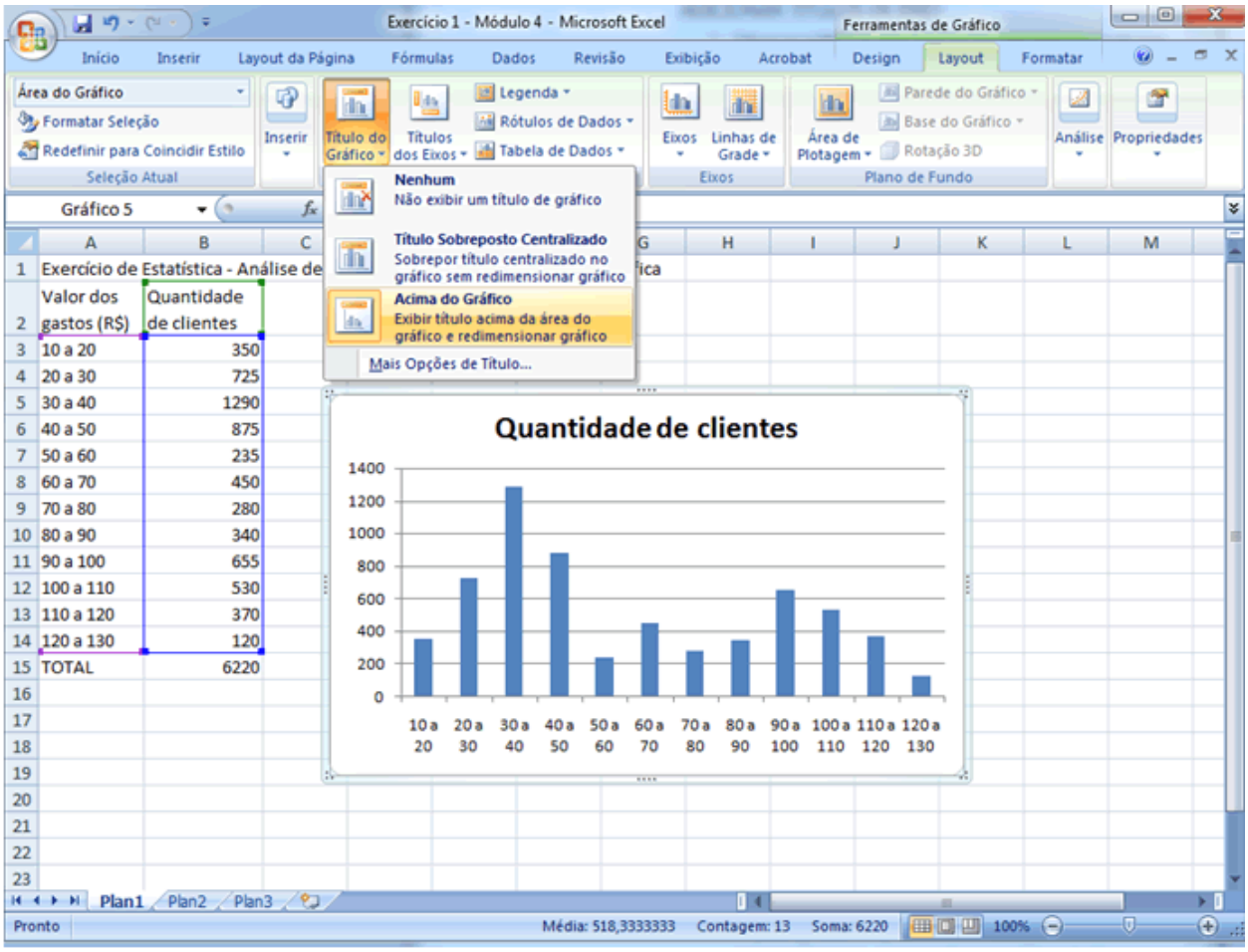
Adicionando o título ao Gráfico

Para adicionar o título iremos clicar no gráfico para que sejam exibidas as Ferramentas de Gráfico, adicionando as guias Design, Layout e Formatar.

Na guia Layout, no grupo Rótulos, clique em Título do Gráfico.



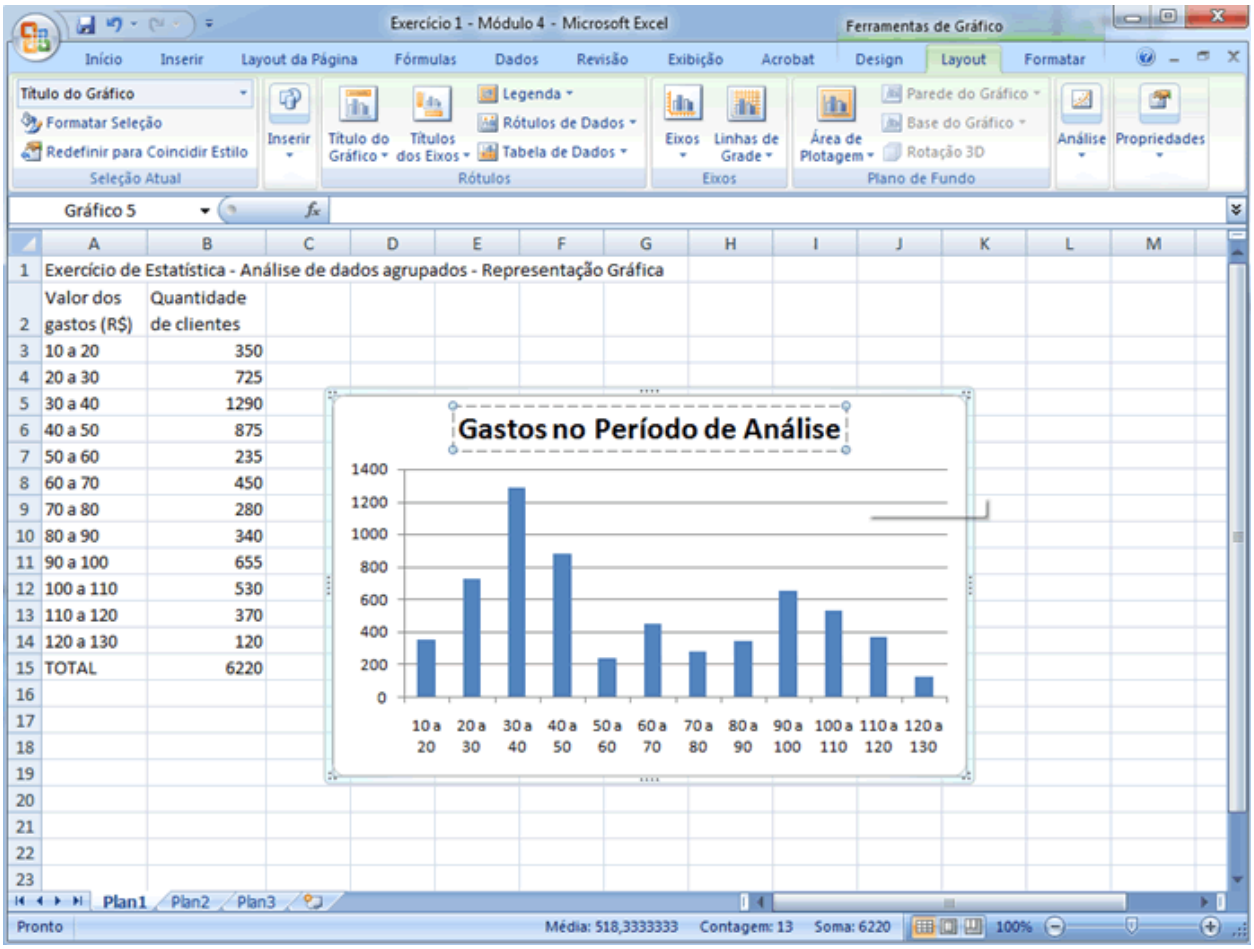
Clique em "Acima do Gráfico":



08

Na caixa de texto “Título do Gráfico” exibida no gráfico, digite o título "Gastos no Período de Análise". Para formatar o texto, selecione-o e clique nas opções de formatação desejadas na Minibarra de ferramentas.

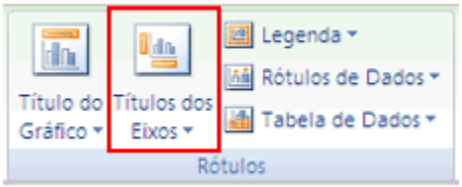
Dica: você também pode usar os botões de formatação da Faixa de Opções (guia Início, grupo Fonte). Para formatar o título inteiro, clique nele com o botão direito, clique em Formatar Título de Gráfico e selecione as opções de formatação desejadas.



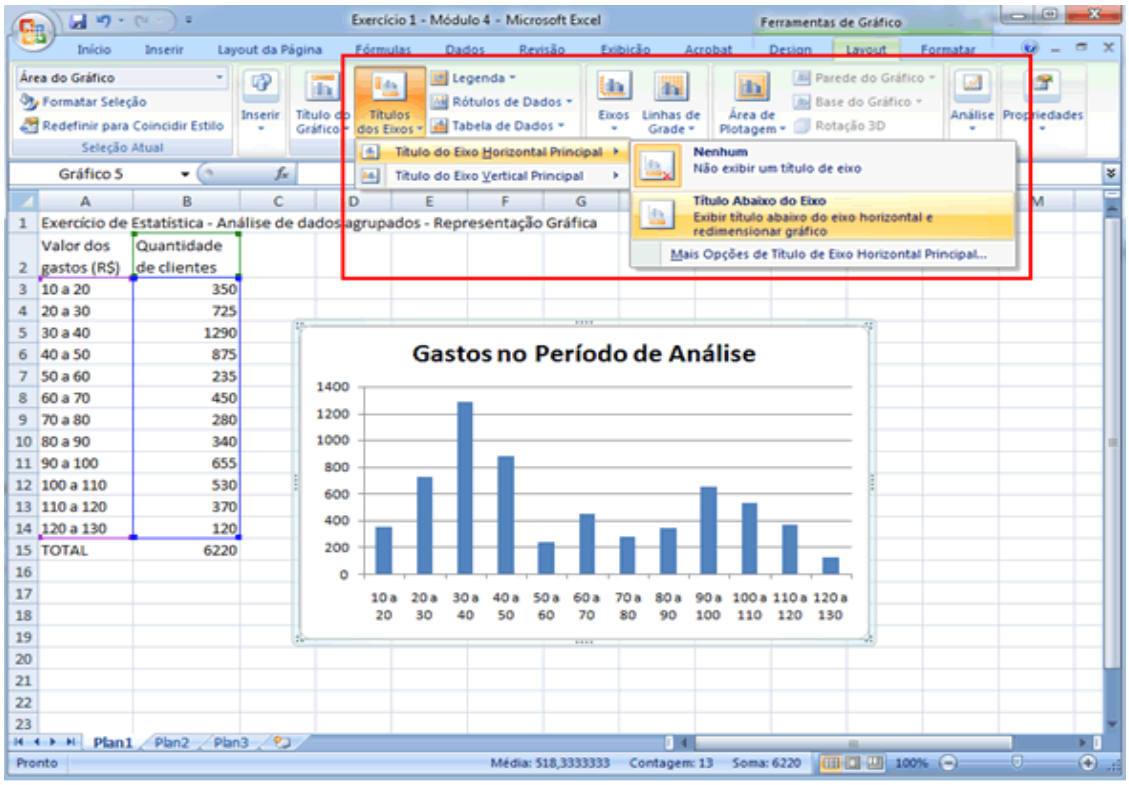
09

Adicionando os títulos nos eixos

Vamos agora adicionar os títulos dos eixos. Na guia Layout, no grupo Rótulos, clique em Títulos dos Eixos.

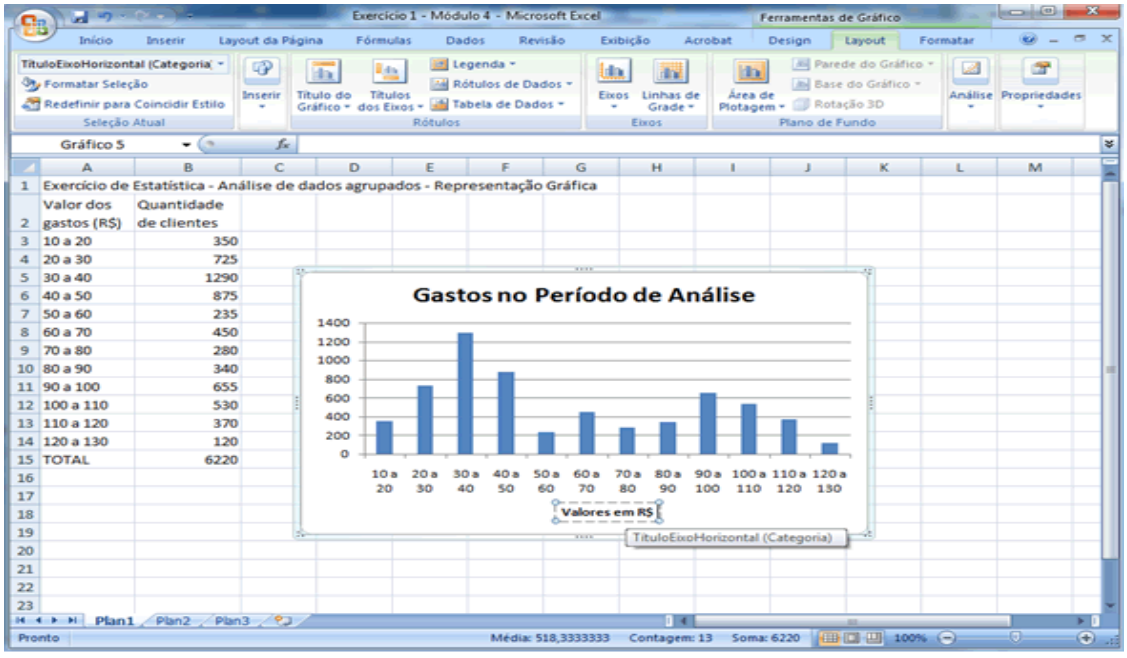


Primeiramente vamos adicionar um título ao eixo horizontal, clique em Título do Eixo Horizontal Principal e selecione a opção "Título Abaixo do Eixo".

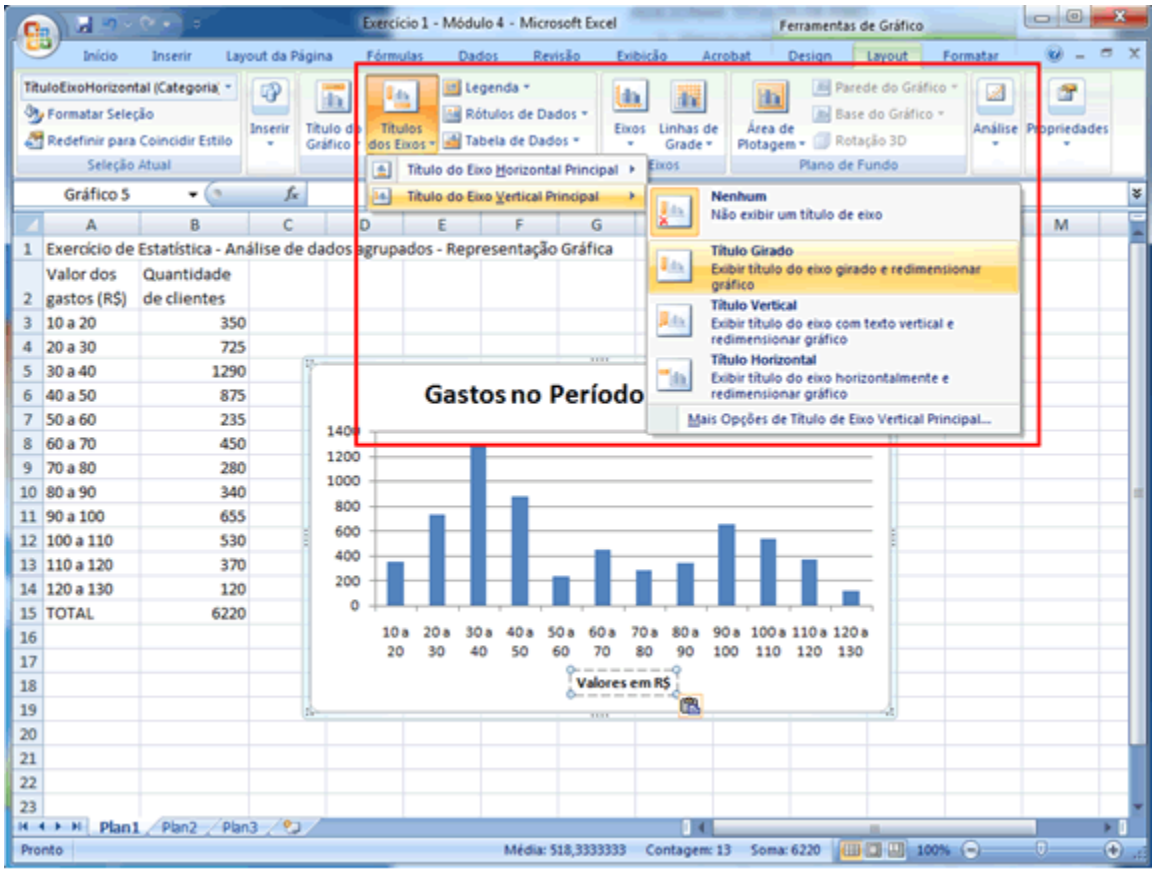


10

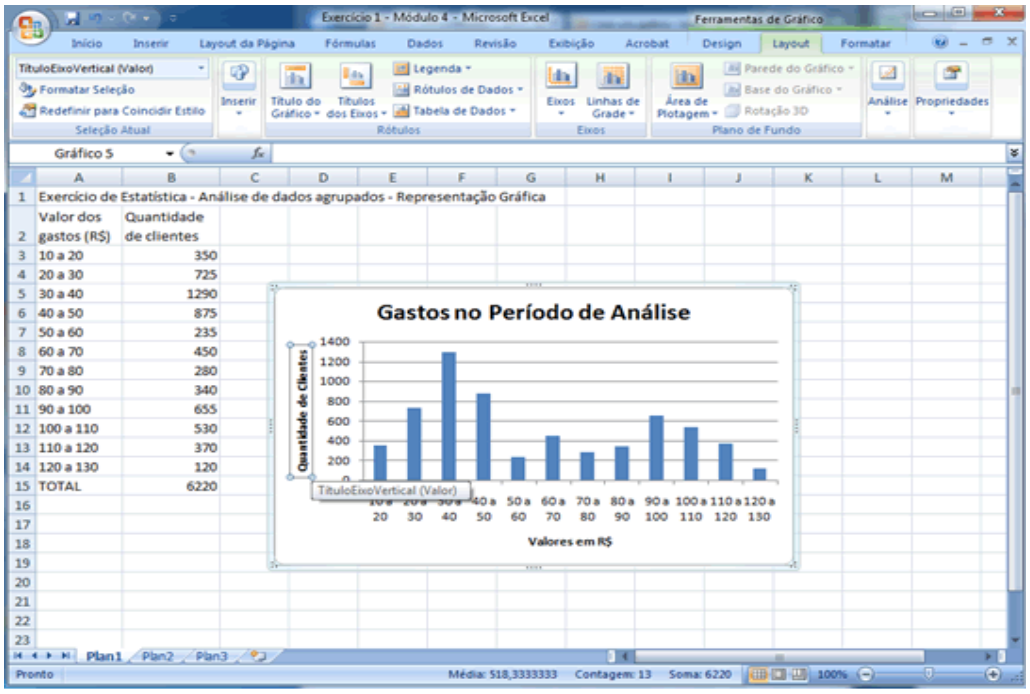
Dica: se o gráfico tiver um eixo horizontal secundário, você também poderá clicar em Título do Eixo Horizontal Secundário. Na caixa de texto Título do Eixo exibida no gráfico, digite o texto "Valores em R$".



Vamos agora repetir o procedimento para o eixo vertical, clicando no botão "Títulos dos Eixos", em "Título do Eixo Vertical Principal" e na opção "Título Girado":

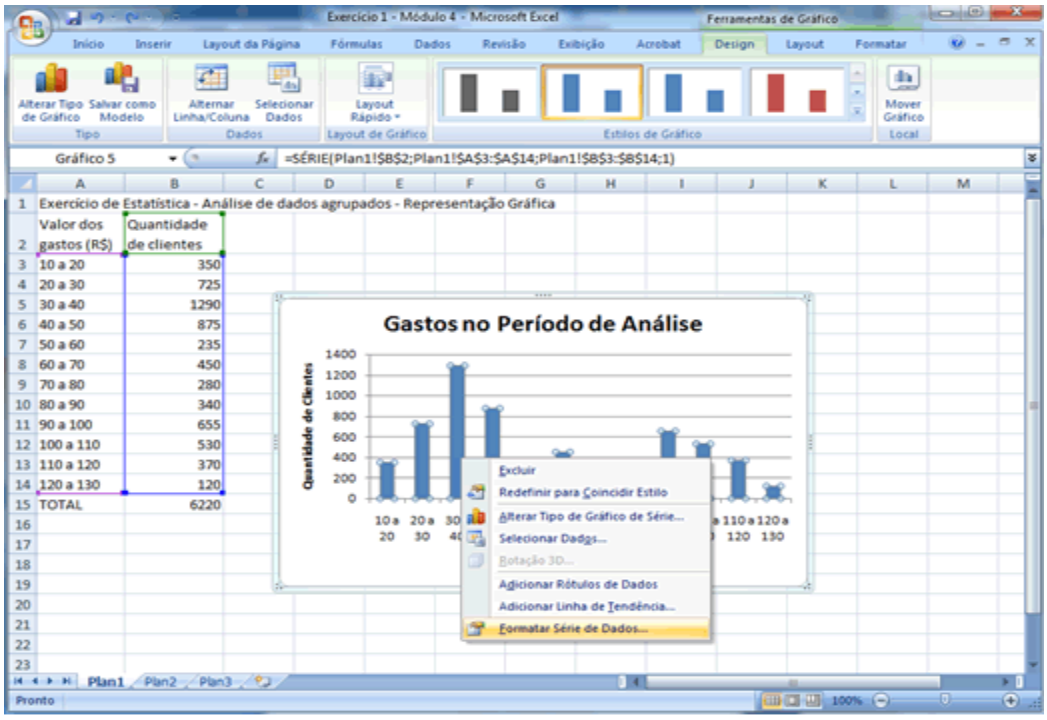


Edite o texto na caixa de título que aparecer escrevendo "Quantidade de clientes".

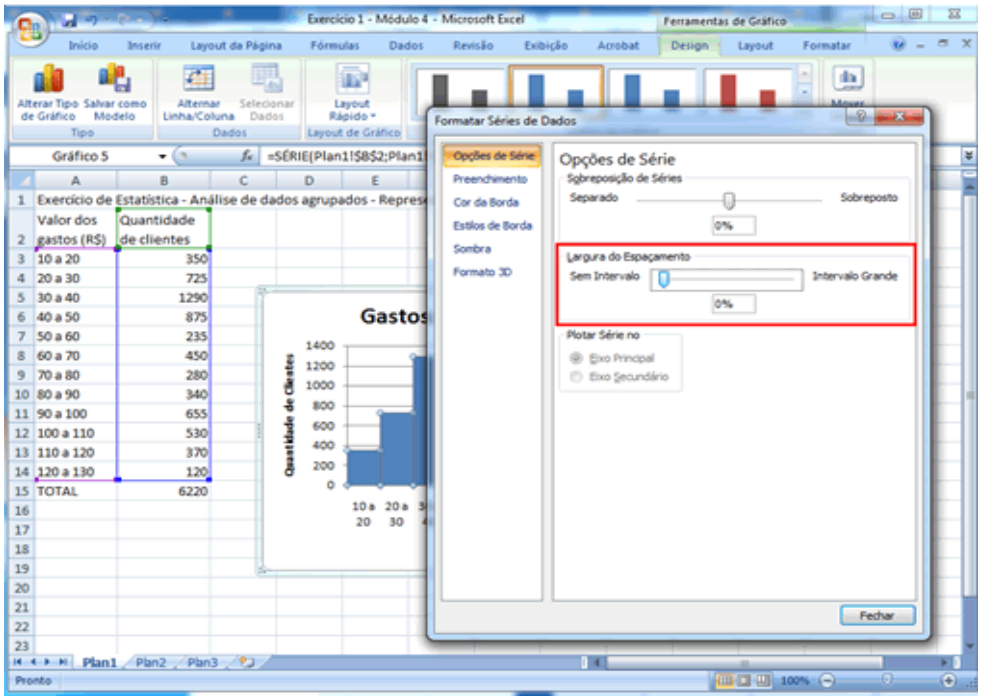


12

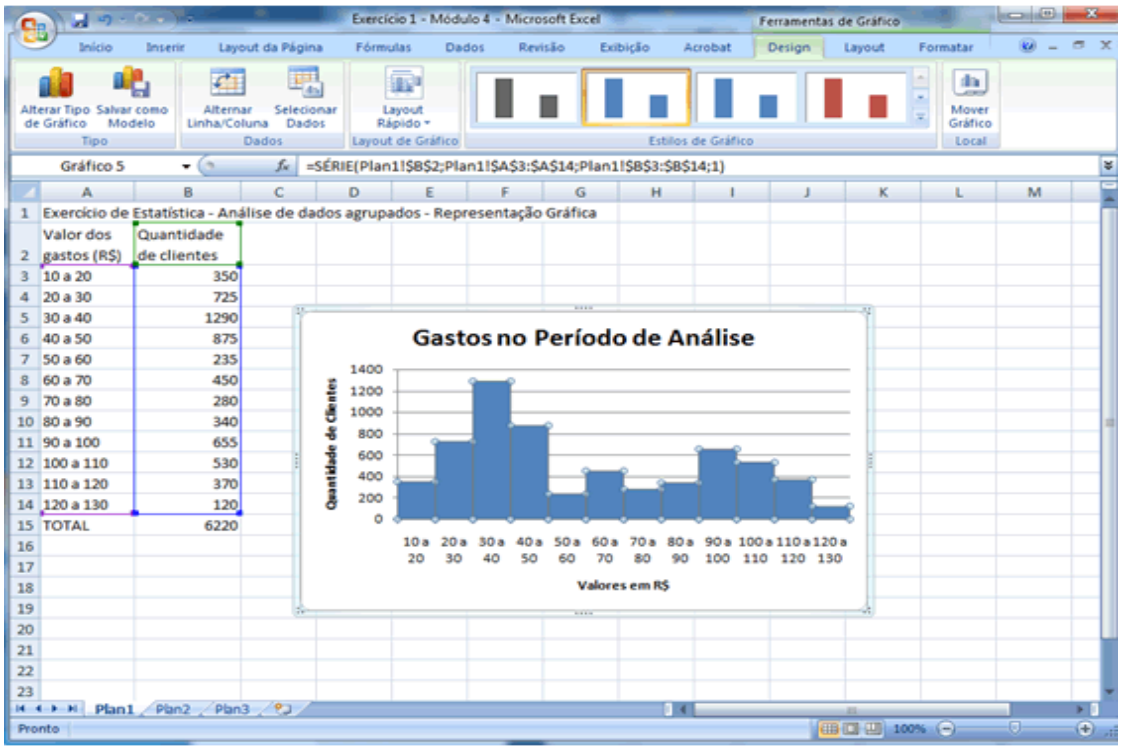
Como há um senso de continuidade para a variável que se está estudando (Despesas/gastos realizados), em particular na transição de uma classe de gastos para outra, pode-se alargar as colunas de forma que elas fiquem "coladas" umas às outras. Isso será feito clicando uma vez sobre uma das barras do gráfico, e com o botão direito do mouse escolhendo a opção "Formatar Série de Dados".



Na caixa de diálogo basta ajustar a largura do espaçamento para 0%.



Em seguida, clique em Fechar.



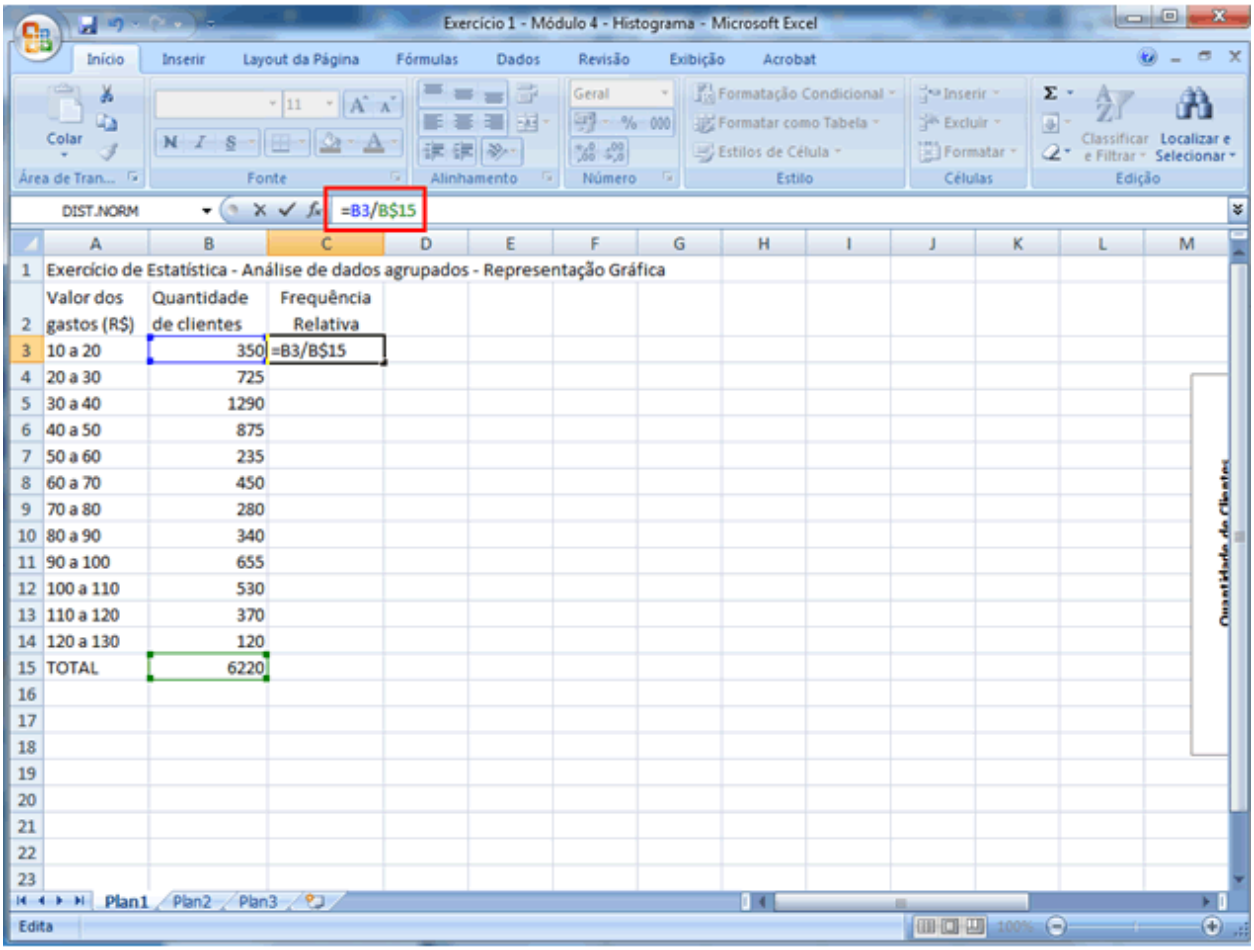
Se você já está familiarizado com os recursos gráficos do Excel sabe que outros ajustes podem ser feitos, particularmente no que diz respeito a cores e formatos. Caso contrário, o que foi feito até aqui é adequado e expressa convenientemente a situação dada.

14

2 - DESENHANDO UM HISTOGRAMA E UM POLÍGONO DE FREQUÊNCIAS

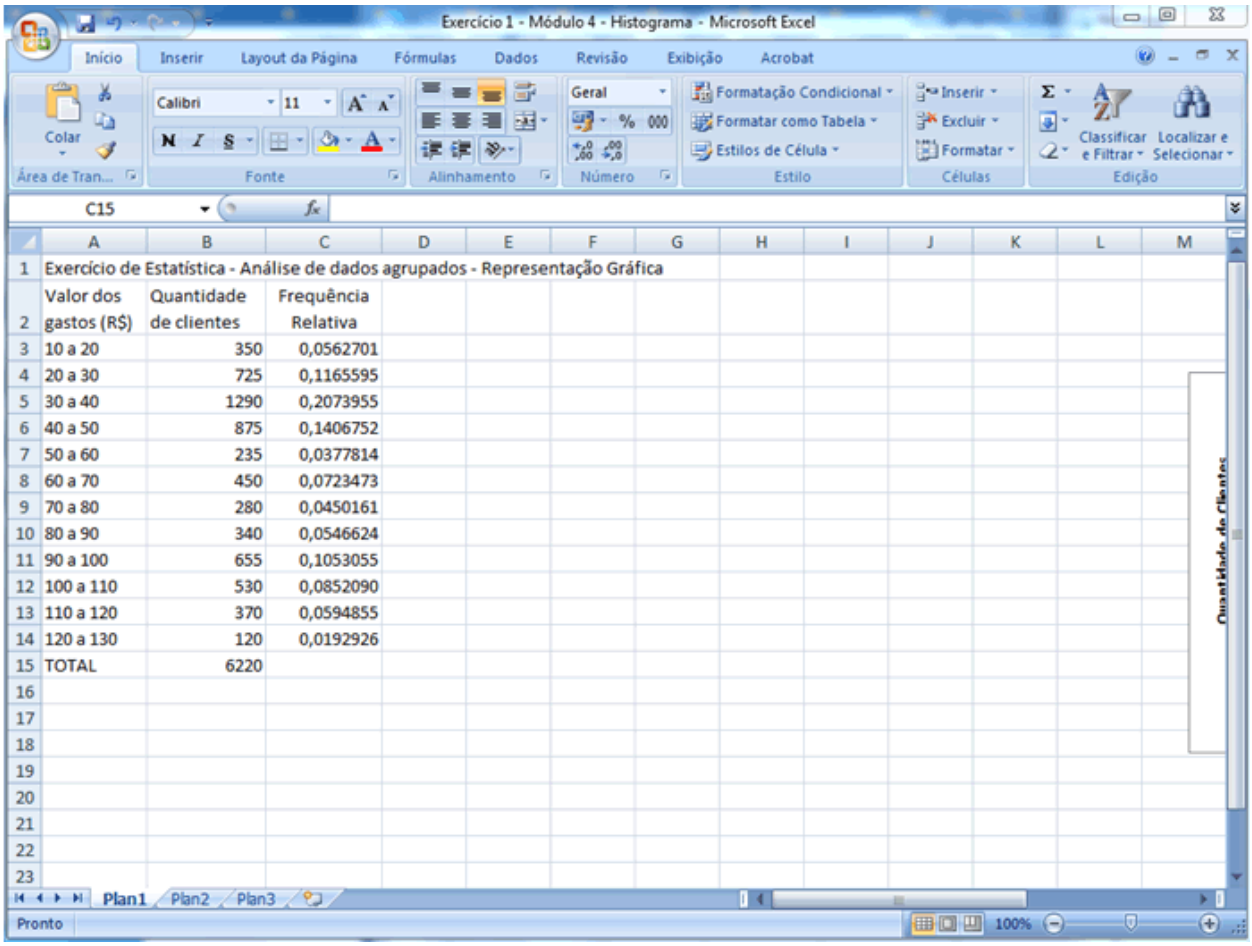
Uma adaptação interessante no gráfico apresentado nessa última tela seria fazer com que a área total das colunas correspondesse à frequência total, ou seja, 100% = 1. Assim a área de cada coluna corresponderia à frequência relativa daquela classe.

Para tanto vamos criar uma coluna chamada de "Frequência Relativa". A frequência relativa é dada pela quantidade de clientes na classe/dividido pela quantidade total de clientes. No caso da primeira classe seria B3/B$15 (Lembre-se de que o símbolo $ em uma fórmula serve para “travar” a célula B15, de forma que se arrastarmos a fórmula para baixo ela continue dividindo as células por B15).



15

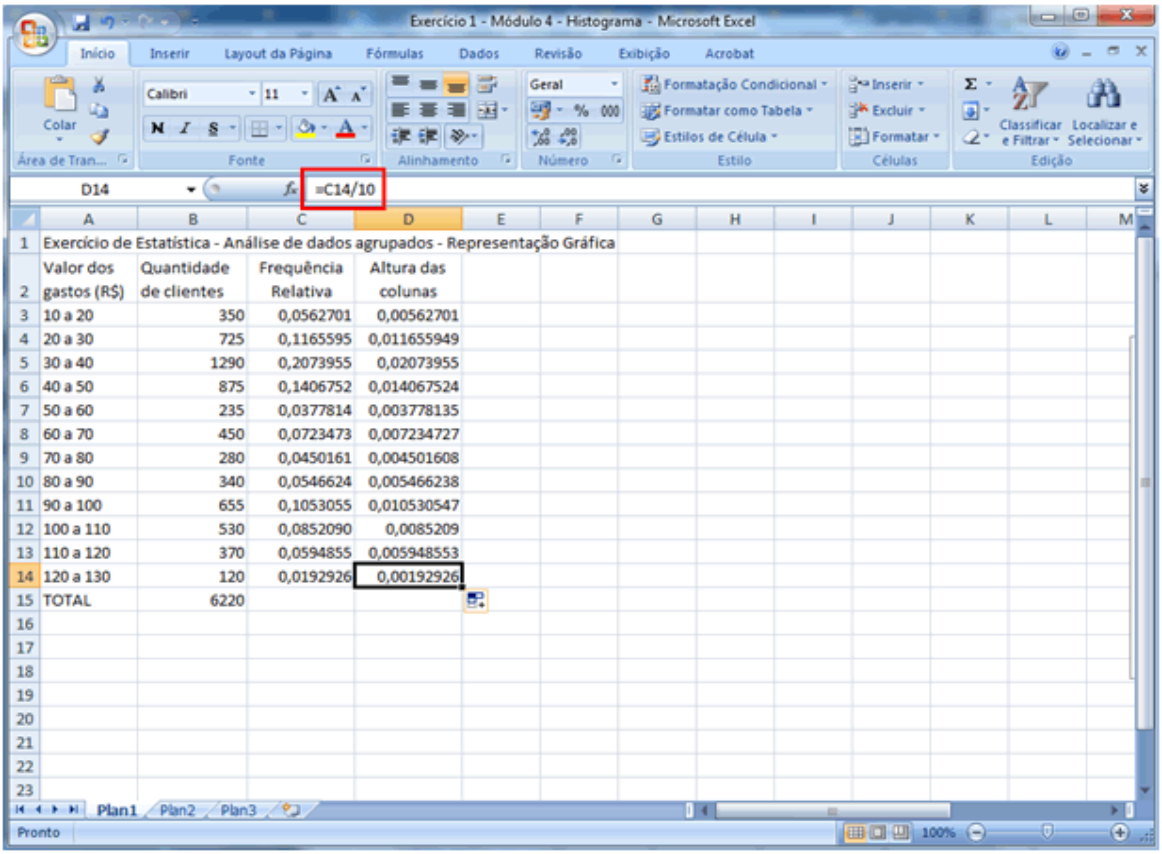
Vamos arrastar a fórmula para calcular a frequência relativa das demais classes.



	A	B	C
1	Exercício de Estatística - Análise de dados agrupados - Representação Gráfica		
2	Valor dos gastos (R$)	Quantidade de clientes	Frequência Relativa
3	10 a 20	350	0,0562701
4	20 a 30	725	0,1165595
5	30 a 40	1290	0,2073955
6	40 a 50	875	0,1406752
7	50 a 60	235	0,0377814
8	60 a 70	450	0,0723473
9	70 a 80	280	0,0450161
10	80 a 90	340	0,0546624
11	90 a 100	655	0,1053055
12	100 a 110	530	0,0852090
13	110 a 120	370	0,0594855
14	120 a 130	120	0,0192926
15	TOTAL	6220	

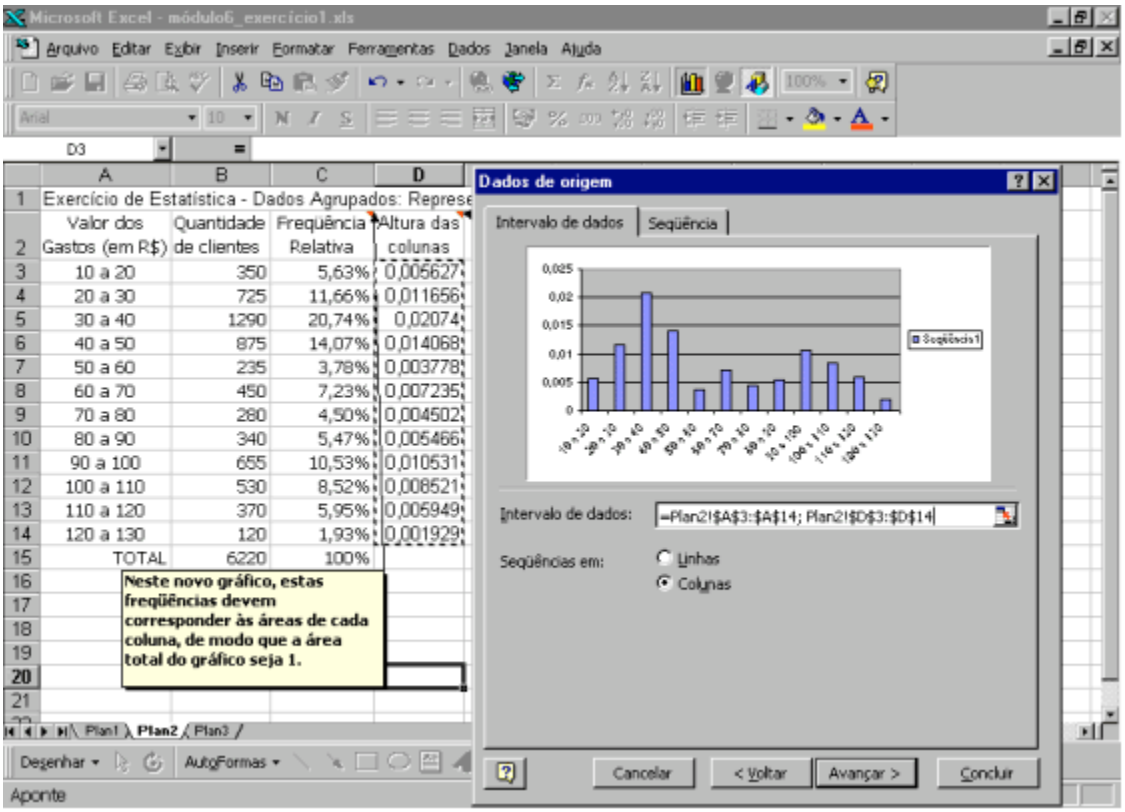
16

Vamos agora calcular a altura das colunas. Para que a área total sob o gráfico seja 1, a altura das colunas será dada pela frequência relativa dividida pela amplitude da classe. Sabe-se que para todas as classes a amplitude é 10 (isso é, todas as bases são iguais a 10). Incluiremos então uma nova coluna chamada "Altura das colunas", como segue:

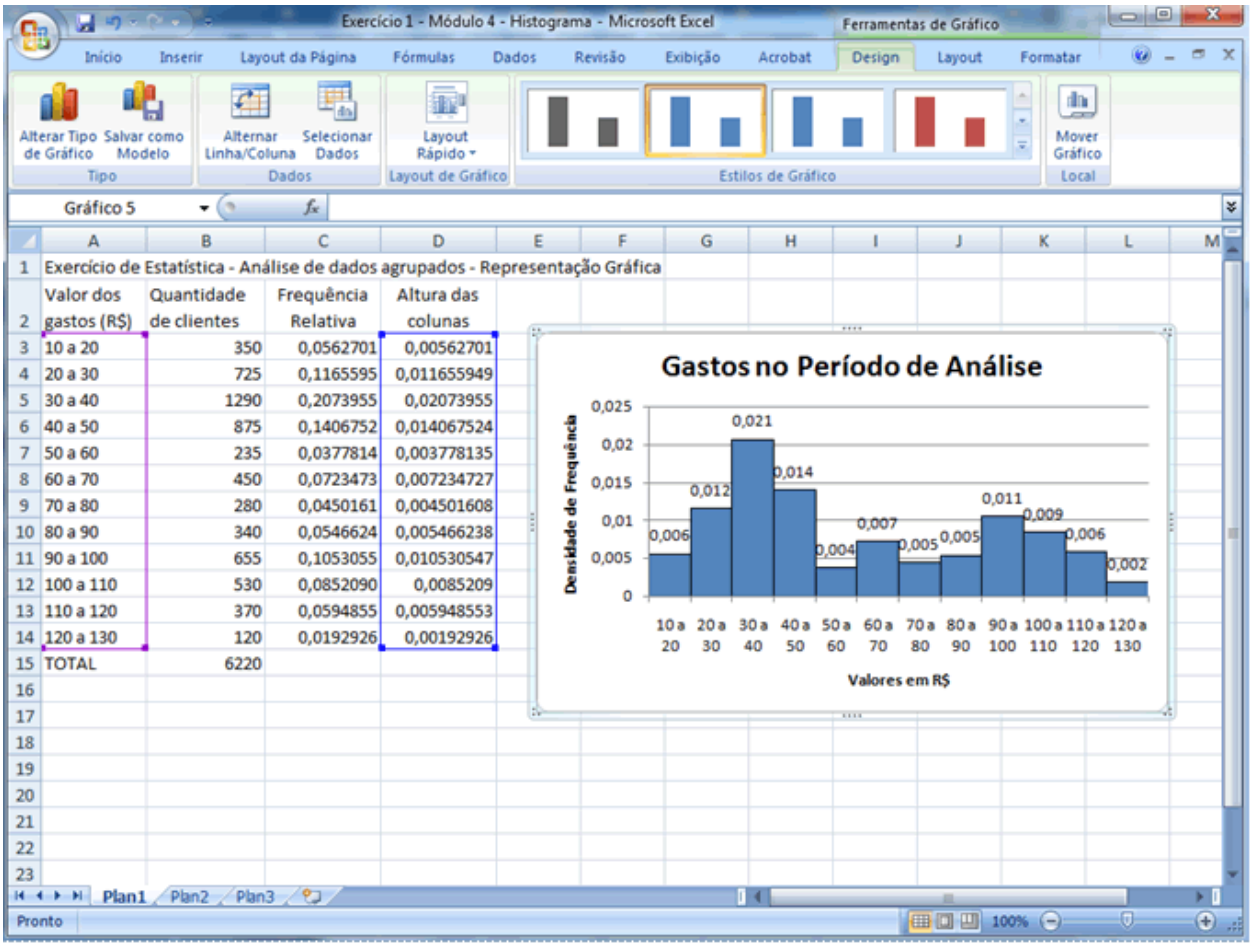


17

Repetindo-se agora todos os passos anteriores para a construção de um gráfico de barras, com o cuidado de inserir a área com os dados corretamente (coluna A com os rótulos e coluna D com os dados para o eixo Y), tal como mostrado a seguir:



O gráfico ficará com o seguinte aspecto:



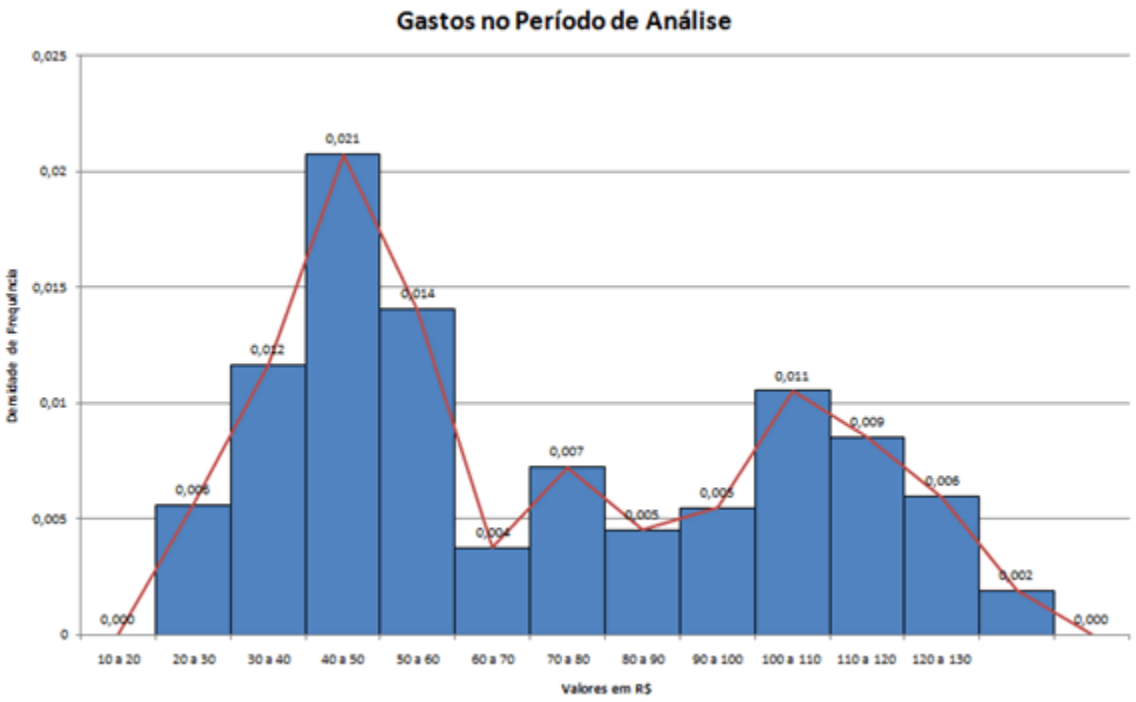
Os valores correspondentes a cada classe (altura de cada coluna) foram inseridos acima das respectivas colunas por meio da opção "Rótulo de dados" da guia "Layout".

O gráfico dessa última tela recebe a denominação de histograma, embora alguns autores também atribuam essa denominação àquele inicialmente obtido com as frequências reais no eixo Y.

18

Polígono de Frequências

Outra forma de representar a densidade de distribuição é por meio do polígono de frequências, que é representado pela linha em vermelho:



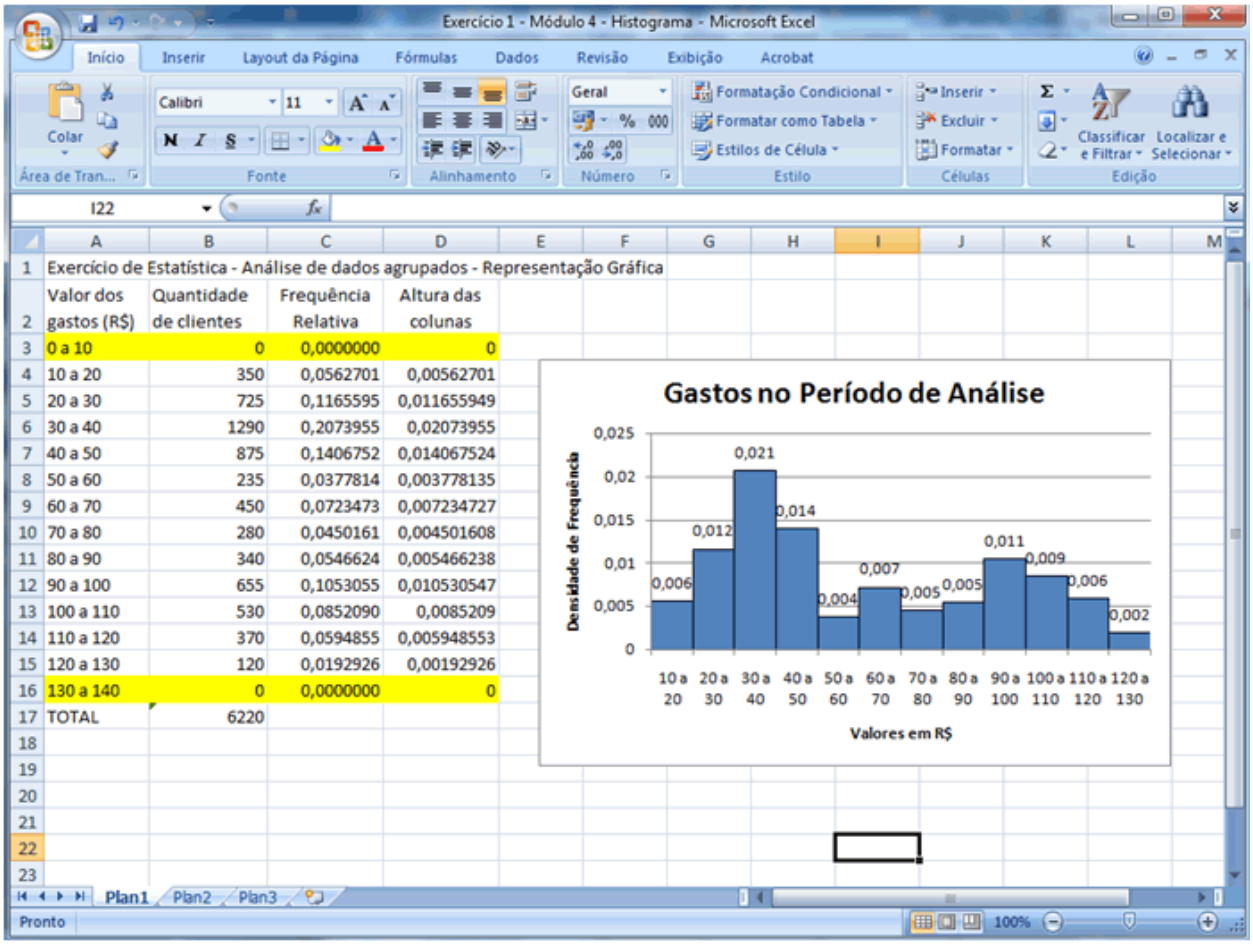
O polígono de frequências é traçado tomando-se os pontos médios/centrais de cada coluna do histograma (em sua parte superior) e unindo-os com segmentos de reta: o da primeira classe une-se ao da segunda, que une-se ao da terceira e assim, sucessivamente.

Para que esse polígono toque o eixo horizontal, arbitram-se duas classes hipotéticas (uma à esquerda da menor e outra à direita da maior) com a mesma amplitude encontrada nas classes existentes (lembre-se do que já foi comentado anteriormente: é desejável que as classes tenham a mesma amplitude). Caso as amplitudes sejam diferentes, devem ser tomadas as amplitudes da primeira e última classes. Nessas novas duas classes hipotéticas, marcam-se os pontos médios, os quais devem ser unidos por segmentos de reta ao gráfico já traçado (em seus pontos inicial e final).

19

Vamos elaborar na prática para que isso se torne mais claro. Utilizaremos como base a planilha anterior, mas inseriremos uma classe à esquerda da menor classe e uma à direita da maior classe.

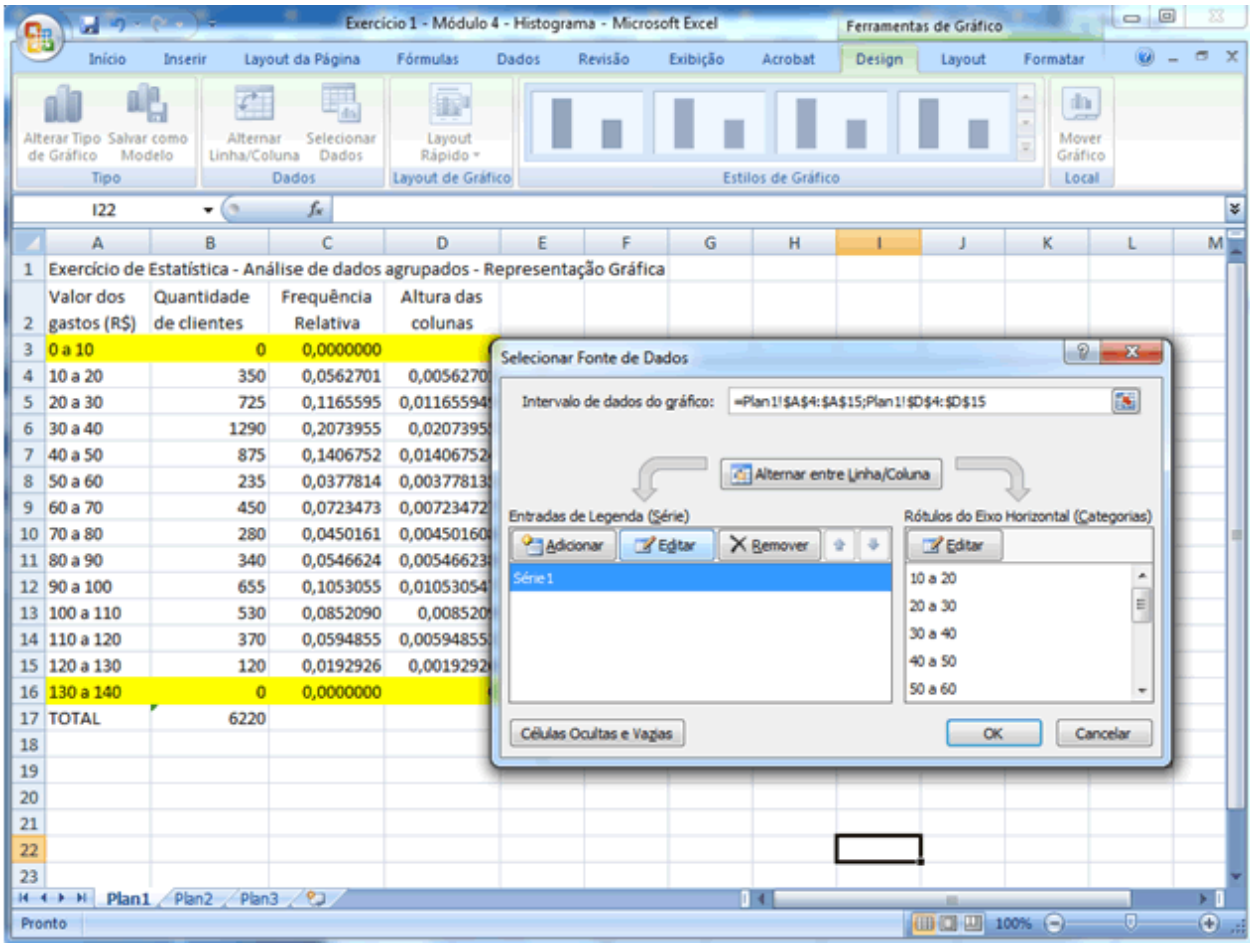
Como a amplitude de todas as classes do exemplo é 10, então iremos criar uma nova classe de gastos de 0 a 10 com quantidade de clientes 0 e outra de 130 a 140 com a quantidade de clientes zero. Para isso, insira as duas novas linhas na planilha e inclua os valores indicados:



	A	B	C	D
1	Exercício de Estatística - Análise de dados agrupados - Representação Gráfica			
2	Valor dos gastos (R$)	Quantidade de clientes	Frequência Relativa	Altura das colunas
3	0 a 10	0	0,0000000	0
4	10 a 20	350	0,0562701	0,00562701
5	20 a 30	725	0,1165595	0,011655949
6	30 a 40	1290	0,2073955	0,02073955
7	40 a 50	875	0,1406752	0,014067524
8	50 a 60	235	0,0377814	0,003778135
9	60 a 70	450	0,0723473	0,007234727
10	70 a 80	280	0,0450161	0,004501608
11	80 a 90	340	0,0546624	0,005466238
12	90 a 100	655	0,1053055	0,010530547
13	100 a 110	530	0,0852090	0,0085209
14	110 a 120	370	0,0594855	0,005948553
15	120 a 130	120	0,0192926	0,00192926
16	130 a 140	0	0,0000000	0
17	TOTAL	6220		

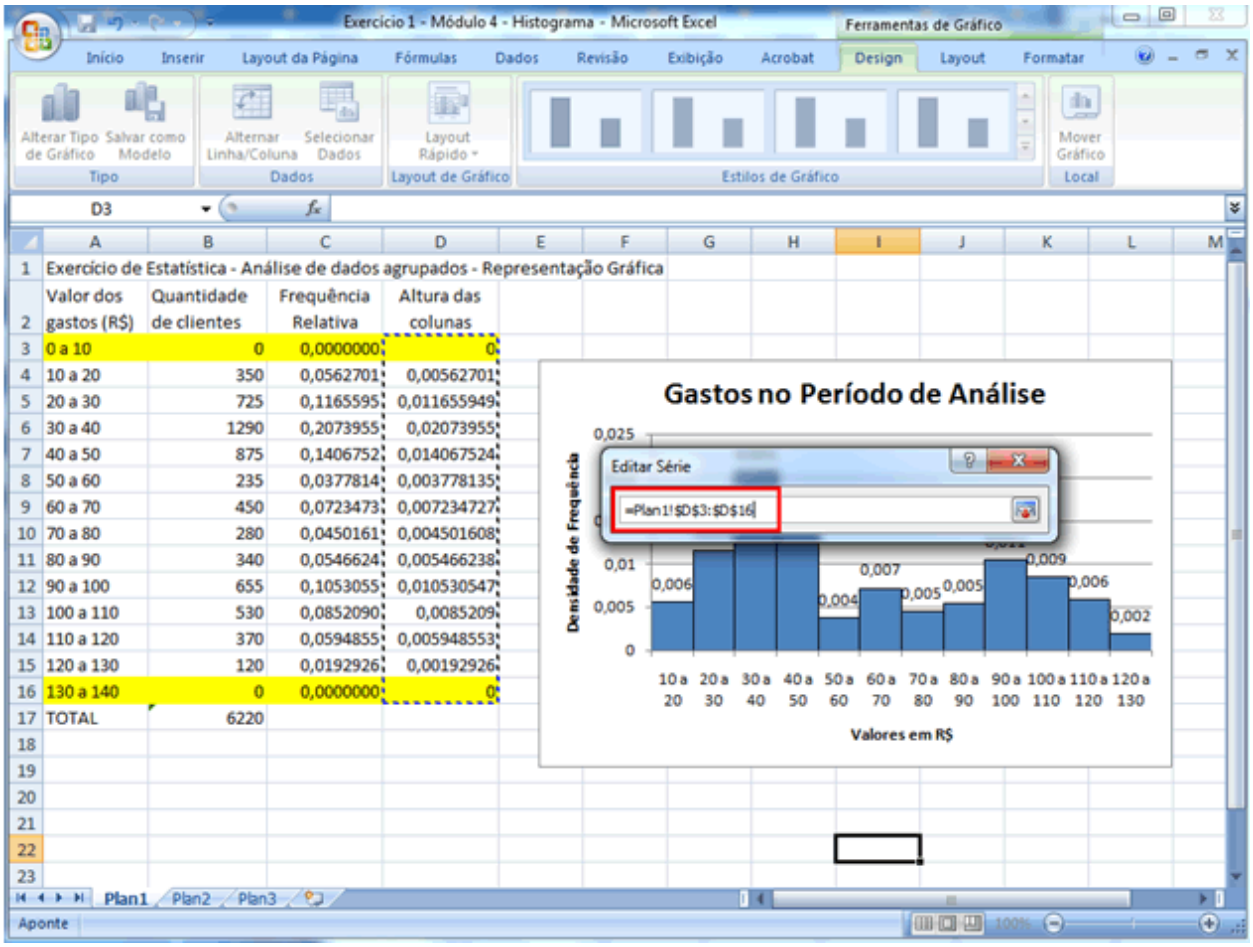
20

Vamos ajustar o gráfico para considerar essas duas novas classes. Iremos clicar no gráfico, logo em seguida, na guia Design, iremos clicar em "selecionar dados". Na caixa de diálogo selecionar "Série1" e clique em Editar, como mostra a figura a seguir.



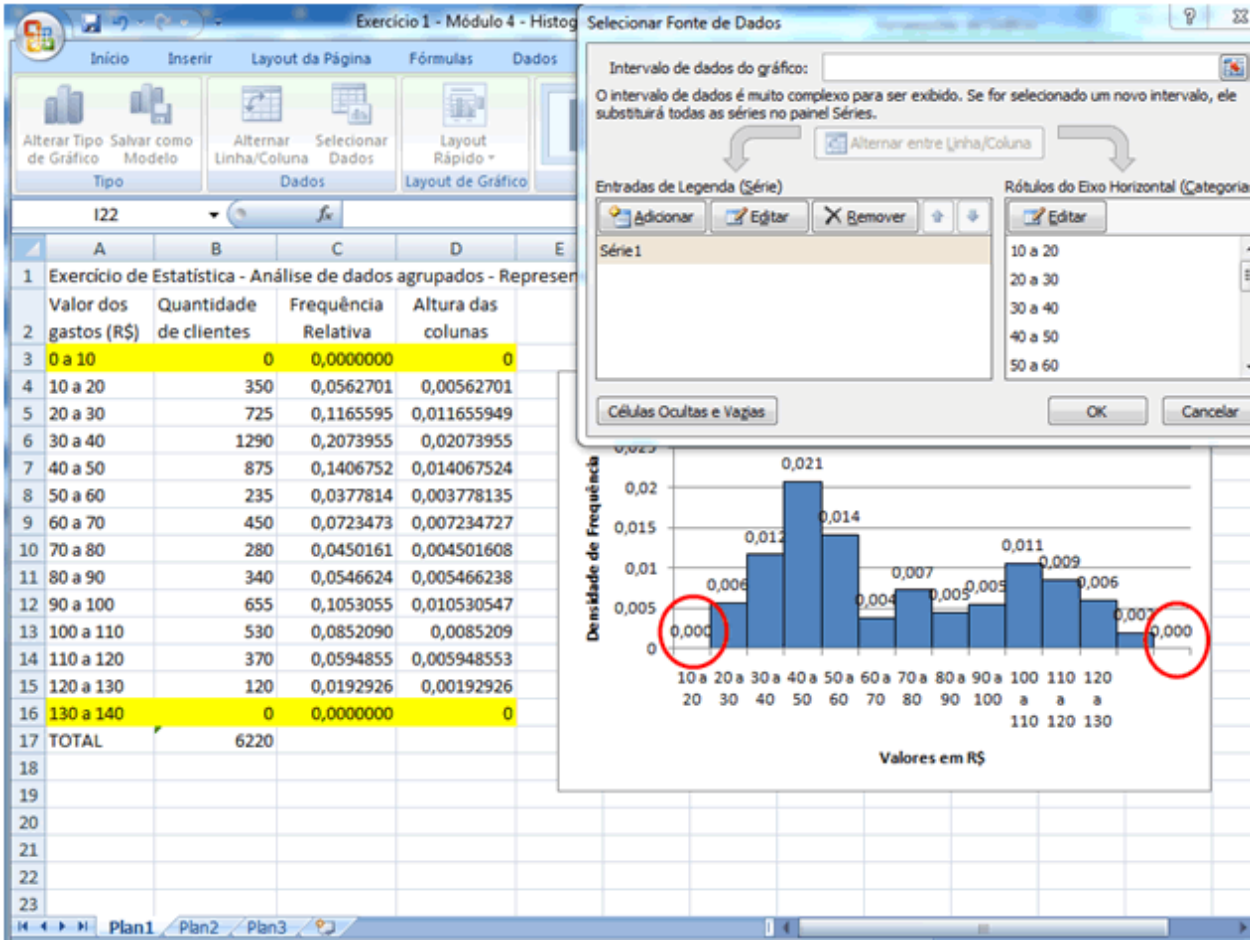
21

Altere os valores da série para considerar também as duas novas alturas de coluna (B3 a B16) e clique ENTER e OK.



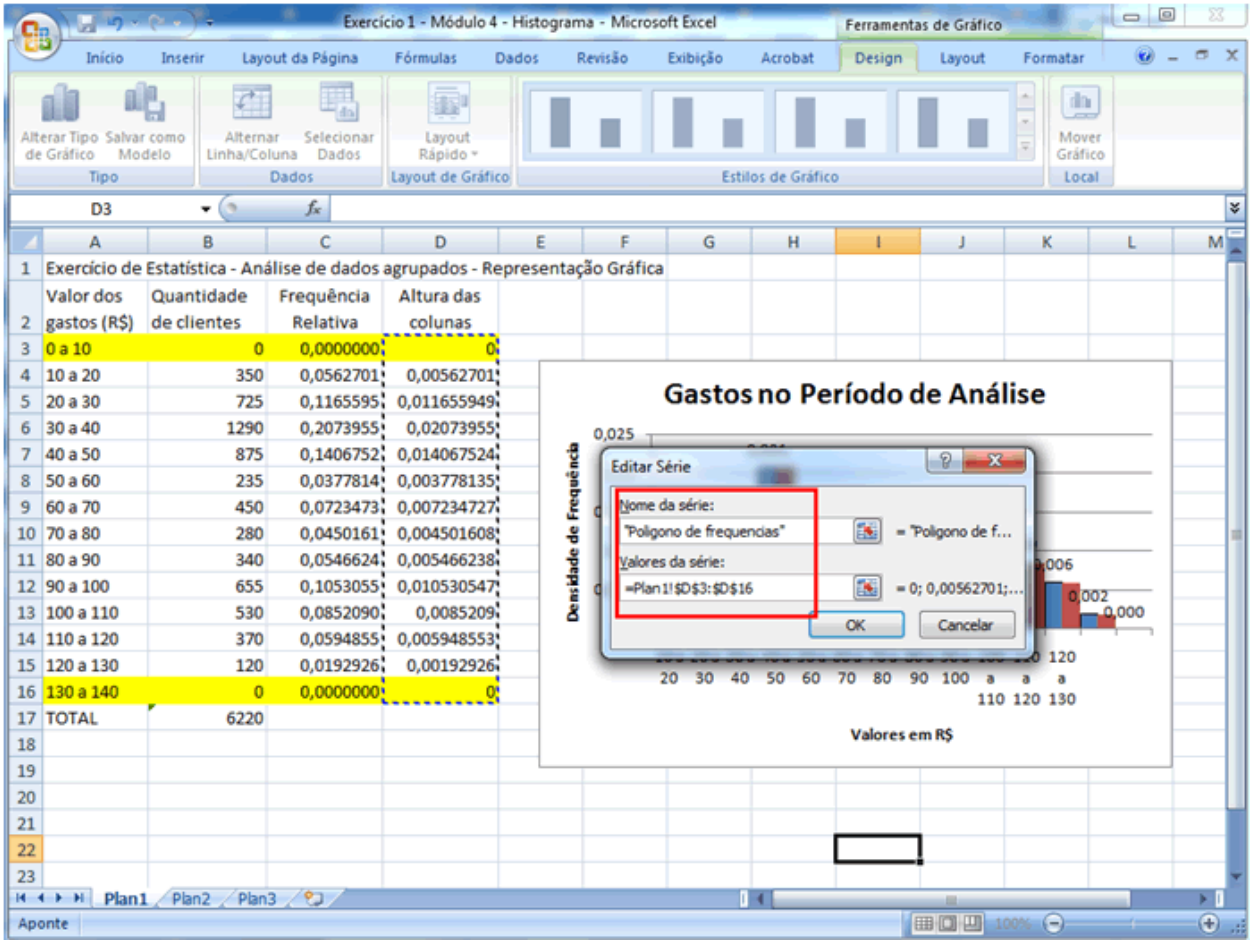
22

Veja que o gráfico já está alterado com as novas classes:



23

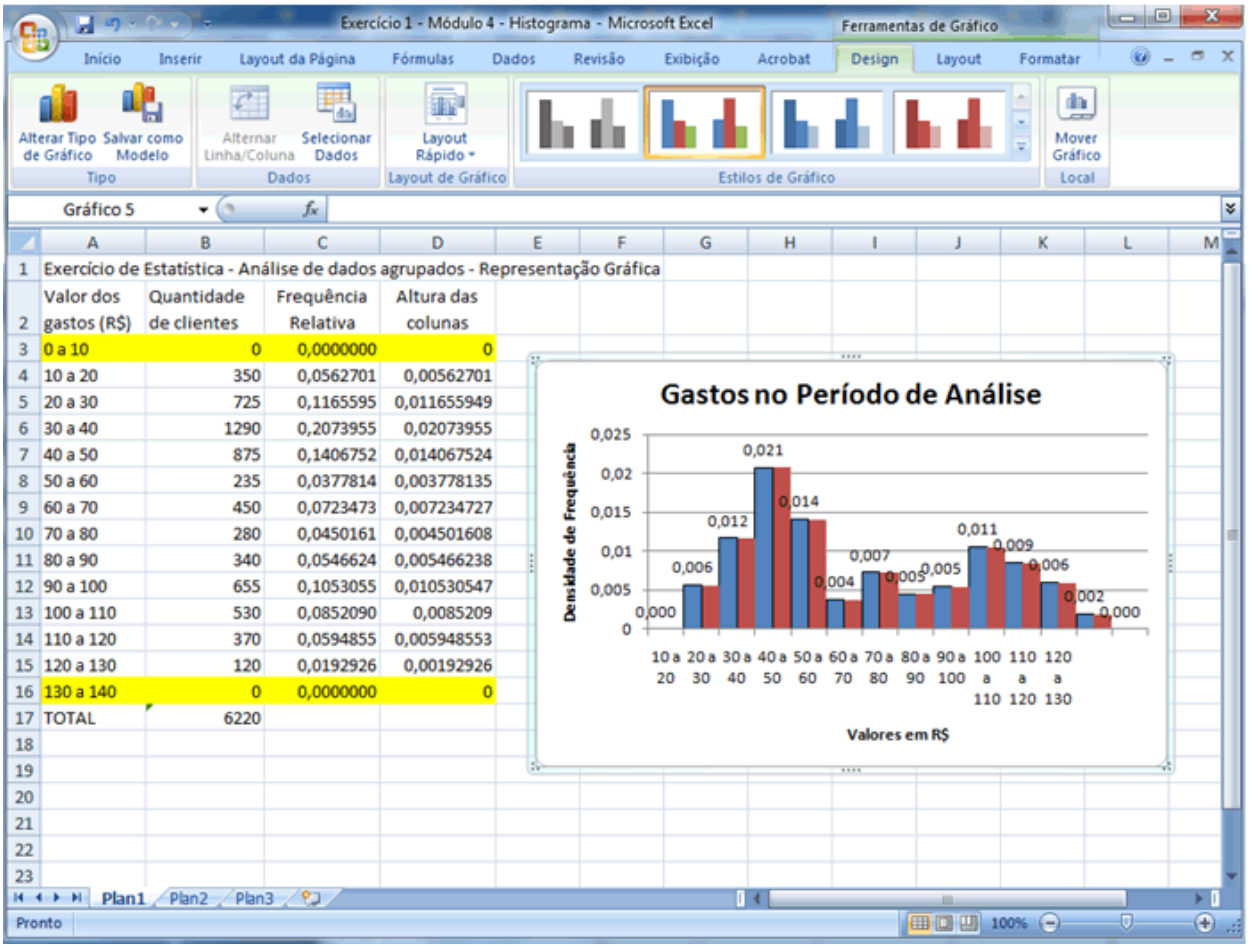
Podemos, então, incluir o polígono de frequências de forma bastante fácil. Vamos clicar no botão "Adicionar", para incluir uma nova série. Escreva no nome da série "Polígono de frequências" e, nos valores da série, selecione a coluna "Altura das colunas", como mostrado a seguir:



24

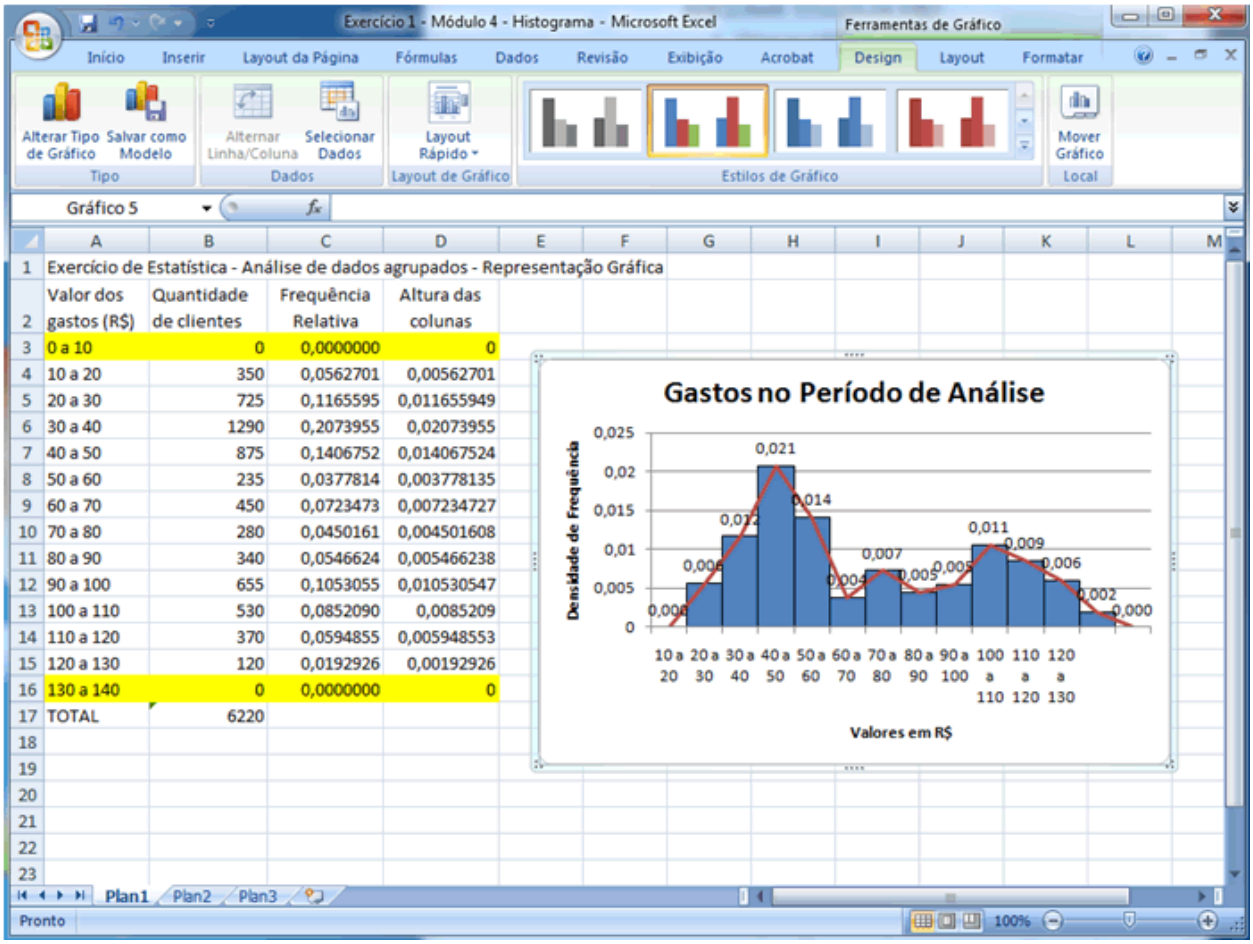
Confirme as modificações clicando em OK duas vezes, chegando ao gráfico.

As colunas em azul correspondem ao histograma e a linha em vermelho ao polígono de frequências.



25

O último passo para obter o polígono de frequências é clicar em uma barra vermelha e, clicando com o botão direito do mouse, selecionar a opção "Alterar tipo de Gráfico da Série". Escolha um gráfico do tipo linha, para obter o resultado final.



26

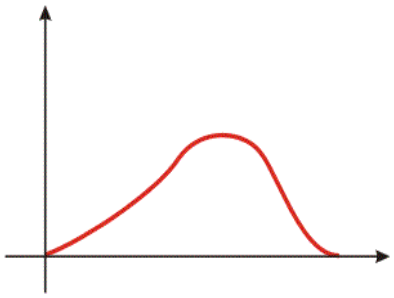
3 - Assimetria e Curtose

Com relação à sua forma, as distribuições de frequência são classificadas, de forma mais genérica, como simétricas ou assimétricas (podendo ser assimétrica positiva ou assimétrica negativa).

Uma distribuição é denominada simétrica quando a metade esquerda (ou direita) do histograma construído (a partir da base de dados) é aproximadamente um "espelho" da metade direita (ou esquerda), ou seja, um eixo vertical de simetria funciona como um espelho no qual a imagem de um dos dois lados reflete-se no outro.

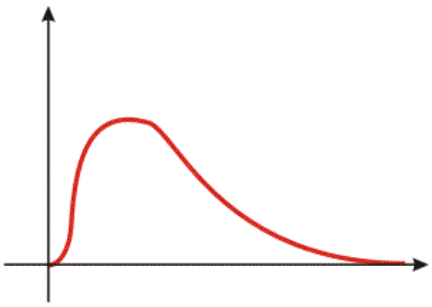
Os formatos (também genéricos) dessas possibilidades de curva são:

Curva assimétrica negativa (ou assimétrica para a esquerda)



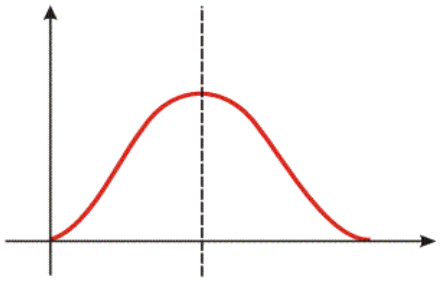
A média é um valor menor que a mediana, que é menor que a moda.

Curva assimétrica positiva (ou assimétrica para a direita)



A média é um valor maior que a mediana, que é maior que a moda.

Curva simétrica (ou curva com assimetria zero)



A média, moda e mediana são valores idênticos.

27

Devem-se a Pearson duas possibilidades de cálculo para o valor da assimetria, dadas por dois coeficientes de assimetria:

$$As_1 = \frac{\text{Média - Moda}}{\text{desvio padrão}} \qquad \text{e} \qquad As_2 = \frac{3 \times (\text{Média - Mediana})}{\text{desvio padrão}}$$

A curtose avalia o grau de achatamento da distribuição e numericamente pode ser obtida por:

$$k = \frac{\left(3^{\circ}\,Quartil - 1^{\circ}\,Quartil\right)/2}{9^{\circ}\,Decil - 1^{\circ}\,Decil}$$

A tendência é que à medida que haja maior homogeneidade, menor será o achatamento da curva, enquanto se houver maior heterogeneidade (maior dispersão), mais achatada será essa curva.

Tanto as medidas de assimetria como aquela de curtose não são de muito valor prático (em um contexto empresarial, por exemplo), contribuindo para uma melhor visualização da distribuição dos pontos sem, no entanto, agregar valor mais objetivo à análise descritiva da base de dados (em especial no escopo do presente curso).

28

RESUMO

Neste módulo buscou-se dar um tratamento gráfico à base de dados com o apoio da planilha Microsoft Excel.

Diante de diferentes possibilidades que o Excel oferece para a apresentação gráfica, optou-se pela utilização de gráfico de colunas (ou barras verticais) com a sequência de passos bastante acessível e praticamente autoexplicativa.

Em seguida, com aproximação das barras e uma adaptação na escala do eixo vertical, para que a área de cada coluna (um retângulo, geometricamente falando) passasse a expressar a frequência relativa (ou percentual), constrói-se o histograma.

A partir da união dos pontos médios superiores de cada coluna do histograma com segmentos de reta, chegou-se ao polígono de frequências.

Por fim, foram trabalhados os conceitos de simetria/assimetria e curtose, sendo o primeiro relativo à existência ao não de um lado mais alongado (direito ou esquerdo) da distribuição de dados e o segundo relativo ao achatamento da curva (em geral decorrente da maior ou menor variabilidade dos dados). Tanto um como outro podem ser quantificados a partir de algoritmos desenvolvidos com esse fim, muito embora a aplicação prática de tal medida seja bastante limitada no âmbito deste curso.