UNIDADE 1 – NOÇÕES BÁSICAS E DADOS NÃO AGRUPADOS
MÓDULO 1 – CONTEXTUALIZAÇÃO E DEFINIÇÕES BÁSICAS

01

1 - UTILIZANDO A ESTATÍSTICA

Relatamos quatro situações que ocorrem em diferentes contextos e para as quais os recursos estatísticos podem ser utilizados de forma a dar grande contribuição, seja para: estruturação de uma pesquisa e análise dos dados levantados; acompanhamento e relacionamento de valores de natureza econômica; interpretação de medidas até certo ponto "populares" (como é o caso de renda *per capita*e expectativa de vida, que são duas médias, do ponto de vista estatístico).

1 - A área de recursos humanos de uma empresa deseja implementar uma pesquisa de clima organizacional (satisfação dos empregados) para posterior elaboração de um "plano de retenção de talentos".

Como fazer?

O que medir ou avaliar?

Quantos empregados entrevistar?

Como selecioná-los?

02

2 - Uma determinada instituição financeira tem recebido, por meio de sua ouvidoria - serviço de atendimento ao cliente - sucessivas reclamações de clientes que alegam estar esperando muito tempo antes do início do atendimento nas agências. Deve ser elaborado um diagnóstico que forneça subsídios a uma possível elaboração de um plano de ação para melhoria do atendimento.

O que deve ser medido? Como? Deve-se incluir todas as agências ou não?

Em caso negativo, quais devem ser incluídas no estudo?

Que medidas devem ser tomadas e em quais dias da semana?

Em quais horários?

Os clientes devem ser entrevistados?

Em caso afirmativo, como selecioná-los e o que perguntar.

03

3 - Considere que você esteja "tentado" a investir no mercado de ações e decida acompanhar a performance de um conjunto de empresas, na bolsa de valores, por um determinado período de tem po, antes de tomar uma decisão sobre investir ou não.

Quais as melhores alternativas para acompanhamento dos valores/variações diárias nas cotações das ações das diferentes empresas sob análise?

É possível comparar/relacionar a performance das ações de uma determinada companhia com aquela divulgada diariamente para a Bolsa como um todo?

04

4 - Acaba de ser divulgado um estudo informando o resultado da renda per capita brasileira no último ano e também da expectativa de vida de homens e mulheres em nosso país.

O que significam exatamente estes valores?

Como interpretá-los de forma adequada?

Seriam necessárias medidas estatísticas complementares para permitir um melhor entendimento da renda *per capita* e da expectativa de vida?

Quais e por quê?

05

2 - E a Estatística?

Embora os censos sejam uma "manifestação" bastante concreta e familiar da Estatística, há muitas outras possibilidades de estudos contemplados por esta área do conhecimento. Convém lembrar que, mesmo com aplicações históricas que datam de milênios, como as já apresentadas, o termo, tal como hoje é conhecido, só surgiu na literatura em meados do século XVIII.

Assim podemos entender Estatística como:

- Ramo da Matemática Aplicada dedicado à análise de dados de observação (Fischer)

- Coleção de métodos (conjunto e técnicas) desenvolvida (o) para planejar "experimentos", obter, organizar (classificar), apresentar, analisar e interpretar dados e utilizá-los para a tomada de decisão

(Mário Triola/Toledo e Ovalle)

E no plural (estatísticas):

- Qualquer coleção consistente de dados numéricos reunidos com a finalidade de fornecer informações acerca de uma atividade qualquer.

Exemplos: estatísticas econômicas - taxa de desemprego; índices de inflação;
PIB (produto interno bruto).

Como grandes áreas ou funções da Estatística, temos:

- Estatística Descritiva
- Probabilidade
- Inferência

Inferência

Refere-se a um processo de generalização a partir de dados particulares (em geral obtidos / levantados em uma amostra).Os "alicerces" da estatística inferencial foram construídos por matemáticos bastante reconhecidos, como Bernoulli, DeMoivre e Gauss, embora apenas a partir do início do século passado é que métodos e técnicas pertinentes à Inferência passaram a ser estabelecidos por outros estudiosos famosos da Estatística, como Pearson, Fisher e Gosset.

Probabilidade

Útil para analisar situações que envolvem o acaso (incerteza) quanto à ocorrência ou não de um evento futuro. A formulação matemática da teoria das probabilidades teve início na metade do século XVII, tendo por base troca de correspondências entre o filósofo Pascal e o jogador Chevalier de Mere.

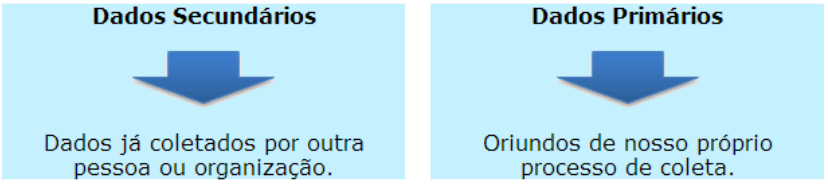
Estatística Descritiva

É a parte que utiliza números para descrever fatos. Compreende a organização, resumo e, em geral, a simplificação de informações.

06

3 - Coletando Dados

Tendo em vista que os dados constituem a matéria prima da Estatística, a fim de viabilizar estudos dos mais diferentes fenômenos e de variáveis a eles associadas, é muito importante que sua coleta seja feita da forma mais adequada, sendo que há duas formas básicas pelas quais eles podem ser obtidos:



O conhecimento das fontes de dados secundários é de grande valia para o processo de tomada de decisão. Uma decisão pode ser mais bem apontada quando subsidiada por **fatos** e **dados reais**.

Se estão disponíveis dados secundários adequados, você pode economizar a coleta dispendiosa de dados primários.

Variáveis

"Característica" que pode assumir distintos resultados para cada um dos "objetos" ou pessoas estudados.
1. QUANTITATIVA: representa uma realização numérica (números resultantes de uma contagem ou mensuração)
1.1. DISCRETA: assume apenas um número finito ou enumerável de valores (geralmente está associada a um processo de contagem).
Ex.: número de filhos, quantidade de defeitos, quantidade de clientes
1.2. CONTÍNUA: assume qualquer valor num intervalo (geralmente está associada a mensuração).
Ex.: peso ou altura de um indivíduo, tempo para execução de uma tarefa.
2. QUALITATIVA: representa uma qualidade ou atributo do "indivíduo" pesquisado.
2.1. NOMINAL: não existe nenhuma ordenação nas possíveis realizações (resultados).
Ex.: procedência ou naturalidade (cidades de origem);
Sexo - masculino ou feminino;
Estado civil - solteiro, casado, viúvo, outros.
2.2. ORDINAL: existe uma possibilidade de ordenação / hierarquização dos resultados. Ex.: classe socioeconômica - possíveis realizações: A1, A2, B1, B2, C, D, E;
Grau de instrução - possíveis realizações: analfabeto; 1o grau incompleto, 1º grau completo, 2º grau incompleto, 2º grau completo, graduação incompleto, graduação, pós-graduação.

Fenômenos

Qualquer evento (acontecimento), seja ele natural, social, econômico ou biológico, que se pretenda analisar e cujo estudo seja passível da utilização de técnicas estatísticas.

07

Em relação aos dados secundários, Mick Silver (2000) afirma que, quando se usa esses dados, as definições, a finalidade, a cobertura, a frequência (quantas vezes), o nível de desagregação (detalhes), a temporalidade (atualidade) e a exatidão (incluindo tamanho da amostra, quão representativa a amostra é, a tendenciosidade nas perguntas feitas) podem ser impróprios para seus objetivos, porque foram delineados com propósitos genéricos ou diferentes do seu.

O uso de dados secundários deve levar em conta se os mesmos são compatíveis com o objetivo pretendido.

O autor também ressalta que os títulos das publicações e as definições podem ser bastante limitados ou, então, tecnicamente corretos no sentido usado por técnicos, porém por ninguém mais. Os resultados podem omitir fatos importantes ou incluir estimativas e o método de compilação pode ter sido modificado, prejudicando comparações com resultados publicados anteriormente.

Os dados secundários podem ser classificados em três, segundo Mick Silver:

- estatísticas oficiais;
- estatísticas não oficiais;
- estatísticas obtidas dentro de empresas.

Universo ou população

"Conjunto de elementos (indivíduos, animais, objetos) que têm pelo menos uma variável comum observável." Mick Silver.

Amostra

"Subconjunto do universo ou população que se quer estudar. Deve ser obtida por procedimentos técnicos adequados de forma a validar os resultados que serão decorrentes do trabalho." Mick Silver.

Estatísticas obtidas dentro de empresas

"Nas grandes organizações, uma enorme quantidade de dados é, em geral, coletada e analisada sem

que os administradores de outros departamentos sejam inteirados; outras vezes, os dados estão em forma inadequada e são analisados usando um software restrito, que não atende às necessidades de outras pessoas. Os dados podem ser, até, bastante detalhados e em grande quantidade, como, por exemplo, as vendas de um produto específico em uma grande loja - isto é, registros de cada compra, quantidade, data e hora, preço - que podem estar relacionadas às características do comprador, por meio do cartão de crédito da loja. Também estão, em geral, disponíveis dados detalhados sobre força de trabalho, produção, ações, compras etc. Quando a natureza dos dados ou dos softwares usados é inadequada para obter a informação necessária, cria-se a necessidade de, a longo prazo, reconsiderar o sistema de informação da organização. Se o software e/ou o hardware não atendem às necessidades imediatas sem modificação cara, a solução de curto prazo, na maior parte dos casos, é gerar uma amostra representativa de dados com base no banco de dados da organização, que pode ser acessado para análise por microcomputador, usando-se um software mais adequado." Mick Silver.

Estatísticas não oficiais

"Além das estatísticas oficiais, há organizações semigovernamentais e privadas que ocupam espaços no mercado da informação. Por exemplo, se você quiser informação sobre um mercado consumidor específico, Mintel, Key Note Market Reports, e Euromonitor produzem regularmente certo número de relatórios e estudos; relatórios sobre setores industriais são produzidos por, entre outros, The Economist Group, o Financial Times Business Information Service e o National Economic Development Office (Nedo)." Mick Silver.

Estatísticas oficiais

"Esta denominação é dada às estatísticas coletadas e compiladas por órgãos do governo que, pelo menos em princípio, seguem diretrizes aceitas internacionalmente. Os diversos órgãos internacionais têm responsabilidade em manter padrões similares em áreas geográficas diferentes. Por exemplo, entre as responsabilidades do FMI (Fundo Monetário Internacional), está a obtenção de estatísticas sobre as Contas do Governo e o Balanço de Pagamentos. (...) Um órgão independente do governo (no Reino Unido, o Departamento Nacional de Estatística, resultado da fusão de outro departamento com o Departamento de Censo da População e Amostragem) é responsável por fornecer as estatísticas de cada país. A confiança do público na qualidade das informações é importante, principalmente no que se refere à independência desses departamentos. (...)
As estatísticas oficiais fornecem informações abundantes e bem preparadas que incluem: padrões de despesas; salários, emprego, suspensões e horas trabalhadas; estatísticas detalhadas sobre importação e exportação; características sociais; séries econômicas; estatísticas sobre energia; indicadores financeiros importantes; informação detalhada sobre as características dos negócios em empresas específicas - que inclui índices de preços para a atualização de valores; estimativas da população e suas características sociais e econômicas por região, úteis para marketing (Censo Demográfico); diferenças regionais e muito, muito mais." Mick Silver.

08

Dados Primários - A coleta de dados é uma etapa muito delicada da pesquisa. Deve ser feita com muito cuidado e de preferência pelo próprio pesquisador. Repassar essa atividade para terceiros pode levar a

coletas distantes do que realmente estamos pesquisando. Vejamos um exemplo de Mick Silver (2000) para pesquisa de uma população.

Podemos obter dados primários por meio dos seguintes recursos:

1. Observação
2. Experimentação
3. Entrevistas
4. Fontes de Documentação

Fontes de documentação

"Porque, embora a maioria dos dados sobre as pessoas seja confidencial, é possível obter informação para uso especial ou para completar um questionário. Por exemplo, as respostas às perguntas feitas aos funcionários podem ser comparadas às informações do banco de dados do departamento de recursos humanos." Mick Silver.

Entrevistas

"Por correio ou por outra forma de comunicação escrita, por entrevista pessoal, pelo sistema de autopreenchimento, por telefone ou em discussões com pequenos grupos." Mick Silver.

Experimentação

"Como, por exemplo, registrar a qualidade dos itens produzidos em determinado dia por uma máquina que foi ajustada (o ajuste é o tratamento). Um grupo controle (outra máquina idêntica, que não foi ajustada) deve ser testado ao mesmo tempo, mas sem o tratamento. Se a máquina em teste (que foi ajustada) fosse observada no início da tarde e a máquina-controle de manhã, a diferença entre elas poderia ser explicada pelos períodos diferentes (manhã e tarde) e não pelo efeito do ajuste. No Japão, métodos baseados na estatística, como o delineamento de experimentos e as técnicas para controle de qualidade e melhoria da produtividade, são bastante utilizados na produção. Tais métodos são também usados para medir o efeito de anúncios de promoções, de mudanças nos preços ou de liquidações. Quando são testados vários tratamentos (como, por exemplo, mudança de preço, de embalagem, da disposição dentro da loja etc.), exigem-se delineamentos mais complexos, como o quadrado latino. Em alguns casos, não conduzimos um experimento real, embora tratemos a situação como se ela tivesse sido criada por um experimento. Por exemplo, podemos considerar os empregados que permaneceram na empresa e aqueles que a deixaram, olhar os dados e identificar as características daqueles que permaneceram e daqueles que saíram. Isso é conhecido como estudo retrospectivo." Mick Silver.

Observação

"Participativa ou não, de comportamento, ou registro de uma informação, como o sexo de um consumidor de uma loja, quadros que atraem mais atenção num museu, tempo ocioso no manuseio e plantas. As câmaras escondidas podem facilitar a observação e já foram usadas para estimar o número de leitores de anúncios colocados do lado de fora dos ônibus em Londres (observando as expressões

faciais)." Mick Silver.

Exemplo

É possível examinar o desempenho acadêmico de todos os alunos de determinada universidade, matriculados entre 1997 e 1998; outras vezes, pode-se tomar uma amostra representativa devido às restrições de tempo e/ou dinheiro, ou devido à natureza destrutiva da pesquisa - é o caso dos testes de telas de televisão, nos quais se pressiona a tela até quebrar.

09

4 - O Mau Uso da Estatística

"Distorções *on-line"*

As estatísticas brasileiras devem ser lidas com cuidado. Tome-se um exemplo no campo do *e-commerce*, a venda de produtos pela internet. Segundo dados da GM, dirigida por José Carlos Pinheiro Neto, 80% dos carros Celta são vendidos on-line. Ocorre que a maior parte dessas transações é fechada nas concessionárias. Motivo: os vendedores orientam o comprador a usar a internet na loja mesmo, porque há desconto para a compra via computador." Revista Veja, edição 1728, de 28 de novembro de 2001, página 34, seção Holofote.

Este pequeno comentário contém uma séria advertência sobre a utilização absolutamente distorcida de estatísticas (como já definido anteriormente), chama nossa atenção para que sejamos criteriosos e procuremos buscar a validação da base de dados disponível e fazer a leitura correta de quaisquer indicadores que nos sejam apresentados.

10

Resumo

Foi apresentada uma breve "localização" histórica de levantamentos estatísticos (com destaque para os censos), mostrando que, a despeito de conceitos teóricos, a necessidade de dispor de dados confiáveis a respeito de determinados fenômenos, de uma população (ou **universo**) e algumas de suas características (ou variáveis), sempre se fez presente na história, desde anos anteriores ao nascimento de Cristo.

Os conceitos introdutórios de Estatística incluem variáveis (quantitativas e qualitativas), censo e amostragem, universo / população (todos os indivíduos ou objetos) e amostra (uma parte do universo). Suas grandes áreas ou funções são Estatística Descritiva, Probabilidade e Inferência.

Estatísticas dizem respeito a conjuntos numéricos (taxas, índices/indicadores) vinculados a determinados fenômenos (econômicos, sociais, demográficos).

A coleta de dados pode ser feita via dados secundários, já disponíveis como fruto da coleta de outra pessoa, grupo ou organização (que incluem estatísticas oficiais, estatísticas não oficiais e estatísticas obtidas dentro de empresas) ou primários (que são coletados pelo pesquisador/instituição que conduz o estudo e podem ser obtidos por observação, experimentação, entrevistas - com ou sem aplicação de um questionário formal, fontes de documentação disponibilizadas por algumas organizações).

Deve-se ter especial cuidado com o mau uso (inadequado ou distorcido) da Estatística, sendo necessário um estado de alerta para que ela não seja utilizada para manipulação de dados, no mau sentido, para "legitimar" premissas inconsistentes ou induzir, deliberadamente, interpretações equivocadas o que, definitivamente, não é um papel ao qual ela se preste.

1 - MEDIDAS ESTATÍSTICAS

A análise descritiva busca, com a utilização de algumas medidas estatísticas, sintetizar um conjunto de dados (sejam esses amostrais ou populacionais), permitindo uma compreensão satisfatória de seu comportamento (de sua distribuição) ou, em outras palavras, oferecendo uma efetiva descrição da base de dados disponível.

Passamos agora a descrever um conjunto de dados que não esteja agrupado, ou seja, cada valor da base é apresentado "individualmente", de forma que pode ser identificado claramente.

Exemplo:

Considere que a empresa para a qual você trabalha solicita que você acompanhe a cotação da ação de uma determinada Companhia (Companhia C) na Bolsa de Valores de São Paulo ao longo de um mês (22

dias úteis). Diligentemente você faz isto e ao final daquele mês você tem os seguintes dados (em reais, o lote de 1000 ações):

24,04; 25,70; 23,14; 22,87; 23,42; 23,96; 25,69; 26,78; 26,88; 27,05; 27,05; 26,12; 24,60; 23,56; 23,04; 21,30; 21,75; 21,85; 22,89; 23,14; 21,74; 20,32.

Ao final do período de observação você deve dar uma breve descrição do comportamento das cotações daquela Companhia, para seu supervisor, permitindo uma visualização minimamente consistente da situação.

Deixando de lado algumas verdades (tais como: o mercado acionário é bastante complexo, com muitas variáveis subjetivas envolvidas – "humor" dos investidores, suas incertezas e seus receios – ou que, diante da natureza da variável em questão, o tempo para observação pode ser insuficiente), e tendo como preocupação prestar uma informação com valor e que seja mais que a apresentação dos 22 valores numéricos acompanhada de um gráfico, o que pode ser dito?

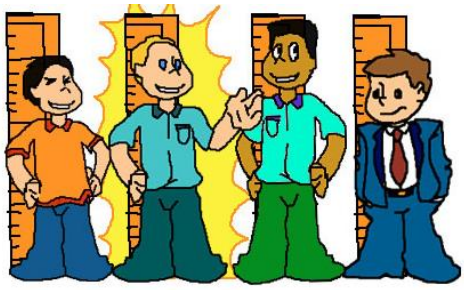
O significado das medidas que serão apresentadas em seguida valem tanto para dados não agrupados (tais como os exemplificados) como para dados agrupados.

13

2 - A MÉDIA

A média tem feito parte do nosso cotidiano com certa regularidade, afinal, todos nós já fomos confrontados ou submetidos a "notas médias", "salários médios", "pesos médios", "alturas médias" ou outros resultados do gênero.

Sua obtenção é simples, pois basta que sejam somados todos os valores correspondentes a todos os dados disponíveis e esta soma seja dividida pela quantidade de dados. Esta definição é válida para **média aritmética simples**.



14

De modo mais formal, se utilizarmos o exemplo anterior e dissermos que a variável *cotação do lote de ações da Companhia C* será denotada por V (representando a variável *valores*) e considerando ainda que essa base de dados é **amostral**, a cotação média será:

$$\overline{V} = \frac{\sum_{i=1}^{n} V_i}{n}$$

onde:

$\overline{V}$ é o valor médio das cotações do lote de ações ao longo do mês.

V_i corresponde ao valor de cada cotação observada a partir do primeiro dia do mês (i = 1) até o último (i = n = 22).

n é a quantidade de dados observados, neste caso o número de cotações diárias, ou seja 22.

Entretanto, se tivéssemos trabalhando com uma base de **dados populacionais**, o símbolo da média não seria mais o nome da variável com uma barra em cima, mas sim a letra grega **μ**. Assim, a formulação matemática para a determinação da média para dados populacionais seria dada por:

$$\mu = \frac{\sum_{i=1}^{n} V_i}{n}$$

15

Por outro lado, considerando o fato de que alguns dados podem se repetir e representando por $\mathbf{f_i}$ o número de repetições, ou frequência, de cada valor dentro da base de dados, chegaríamos, para uma base de dados amostrais, por exemplo, a uma formulação genérica como:

$$\overline{V} = \frac{\sum_{i=1}^{n} (V_i \times f_i)}{n}$$

Aplicando a primeira ou segunda formulação aos dados chega-se a:

$$\overline{V} = \frac{24{,}04 + 25{,}70 + 23{,}14 + \ldots + 21{,}74 + 20{,}32}{22} = \frac{526{,}89}{22} \cong 23{,}95$$

Isso permite dizer que a cotação média do lote de 1000 ações da Companhia C, no mês de referência, foi de **R$ 23,95**.

Mas o que significa exatamente isto? Será que o ouvinte ou o leitor que recebe este resultado é capaz de entendê-lo corretamente?

Frequência

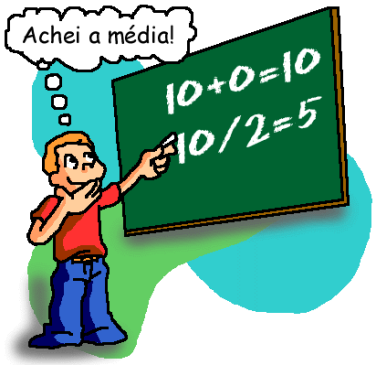
É um número de repetições correspondente ao resultado (evento).

16

O que significa dizer que um aluno concluiu seu curso com nota média de 7,5 (numa escala de zero a dez), após concluir as 20, 30 ou 40 disciplinas do currículo?

Ou ainda, o que significa dizer, dentro daquela mesma ótica, que uma turma de 45 alunos teve nota média de 6,4 (também em uma escala de zero a dez) em um exame de Estatística?

Por último, como deveríamos entender quando um noticiário da televisão divulga resultado de um recente estudo, de um órgão internacional respeitadíssimo, dizendo que a renda por habitante no ano passado em um determinado país é de US$ 8.900,00 (que nada mais é que uma renda média por habitante)?

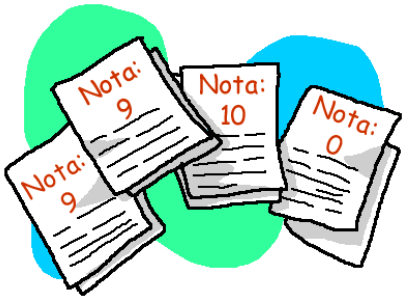


17

Seria razoável supor que todos os valores de uma base de dados (ou pelo menos a maioria deles) são suficientemente próximos do valor médio? Se, "por obra do acaso", metade de uma turma tira nota máxima (dez) em um exame na faculdade e a outra metade nota mínima (zero), a conclusão é que a nota média é cinco, o que não reflete proximidade dos valores originais com esta média. Claro que este é um caso extremo, mas útil para que despertemos para esta questão do uso indiscriminado de médias.

O exemplo anterior (nota mínima zero, nota máxima dez e nota média cinco) poderia induzir-nos a acreditar que média é o valor que está "no meio" (ou pelo menos muito próximo a ele), entre o mínimo e o máximo. Seria isto verdade?

Suponhamos que um estudante faça quatro provas com pesos iguais e tire as seguintes notas:



Observe que ele foi muito bem em 3 provas e resolve não fazer a última para dedicar-se a outras matérias. Sua média será 28 / 4 = **7**

Essa média (7) não é o "meio" entre sua nota máxima 10 e sua nota mínima 0. Esta ideia, na verdade, corresponde ao **ponto médio**, que de fato é a média aritmética entre os valores máximo e mínimo de uma base de dados (ou de uma escala numérica).

18

Podemos continuar indagando: o que significa, dentro de uma visão aplicada, dizer que a cotação média do lote de 1000 ações da Companhia analisada, no mês de referência, foi de R$ 23,95?

Vejamos a seguir mais exemplos que poderão nos ajudar a refletir sobre esses dados.

19

Imaginemos agora que todos os 30 aprovados em um concurso público, para um cargo bastante concorrido, ficaram com pontuação final entre 96,35 e 99,20 pontos (numa escala de zero a cem pontos). Poder-se-ia inferir que essas notas foram suficientemente próximas (por enquanto com base unicamente no bom senso), de modo que a performance dos 30 aprovados (ou 30 primeiros classificados) foi semelhante. Caso fôssemos informados que a nota média deste grupo foi de 98,07, será que, mesmo com as ressalvas já feitas anteriormente, seríamos capazes de afirmar que ela reflete com fidelidade a performance do grupo?

Um caso extremo: todos os 30 primeiros colocados totalizaram exatamente a mesma pontuação de 99,20 pontos. Não deve haver dúvidas de que esta também seria a pontuação média dos 30 aprovados e, consequentemente, representaria com exatidão o desempenho do grupo selecionado.

Pelo que vimos até aqui, deve ficar claro que a média será uma medida fiel, uma medida que representará a base de dados estudada, quando essa base puder ser considerada **suficientemente homogênea**, isto é, com variabilidade julgada **suficientemente pequena**. Por mais que queiramos, não é possível chegar a esta conclusão apenas com nossos olhos e uma dose de bom senso, em particular, à medida que a quantidade de dados cresce.

Surge então a necessidade de avaliar/medir a variabilidade ou dispersão (o grau de heterogeneidade ou de homogeneidade) da base de dados - objeto de análise. Essa avaliação deve ser feita da forma mais objetiva possível, para chegar a conclusões seguras.

Inferir

Generalizar um resultado obtido a partir da análise de uma amostra, para o universo.

20

3 - Dispersão dos Dados: variância, desvio-padrão e coeficiente de variação.

a) Variância e desvio-padrão

Uma primeira possibilidade, para avaliar a variabilidade ou dispersão da base de dados, seria falar na **amplitude do intervalo**, que apresenta os mesmos procedimentos de cálculo para dados amostrais e para dados populacionais.

No exemplo da cotação do lote de ações da Companhia C, teríamos:

Amplitude = valor máximo - valor mínimo = 27,05 - 20,32
Amplitude = 6,73

Esta análise não ajudaria muito, considerando que uma amplitude, mesmo quando considerada grande, nada indica quanto à distribuição dos dados, ou seja, diz muito pouco quanto à variabilidade deles (pode-se ter, por exemplo, uma grande amplitude com pequena variabilidade).

21

Pode-se então pensar em uma medida cujo ponto de partida seja a comparação de cada um dos valores da base de dados com o valor médio, buscando avaliar se a base de dados, como um todo, está suficientemente próxima da média.

Evidentemente, eventuais **pontos discrepantes** - aqueles pontos que "fogem" de um perfil, quer seja ele homogêneo ou heterogêneo, valores excessivamente grandes ou excessivamente pequenos, quando comparados com os demais - não são representados adequadamente pelo valor médio.

Chega-se então a uma medida denominada variância, que é calculada, para dados amostrais, da seguinte maneira:

$$S^2 = \frac{\sum_{i=1}^{n} (V_i - \overline{V})^2}{n-1}$$

Variância

Variância também conhecida como desvio quadrático médio é uma representação da média das distâncias ao quadrado entre o conjunto de valores observado e o valor médio.

22

Se os dados a serem analisados forem populacionais, isto é, todos os dados existentes, a variância (σ^2) será dada por:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (V_{i-\mu})^2}{n}$$

Perceba que, na formulação utilizada para a determinação da variância populacional, o denominador da expressão é dado apenas por "n", que indica o número de dados em análise. Já para uma base de dados amostrais, esse denominador será "n-1". Essa subtração de uma unidade no denominador da variância amostral está ligada à redução de um grau de liberdade (GL) que passamos a ter quando trabalhamos com esses dados.

Devemos lembrar que, na análise de dados amostrais, estamos, na verdade, tentando caracterizar informações de sua população de origem e a redução de uma unidade no denominador da expressão da variância faz com que, matematicamente, este valor fique um pouco maior, tentando garantir a explicação da dispersão total dos dados populacionais que deram origem à amostra que está sendo estudada.

Então, apenas para reforçar o conceito:

Quando trabalhamos com todos os dados estaremos trabalhando com **valores populacionais**. Quando estivermos trabalhando apenas com uma amostra, ou seja, parte dos dados, estaremos trabalhando com **valores amostrais**.

No exemplo das ações, como estamos trabalhando com apenas algumas ações então estamos trabalhando apenas com uma amostra, logo deveremos utilizar a fórmula dos dados amostrais (**n-1**). Caso a fizéssemos uma análise com todas as ações, então deveríamos utilizar as fórmulas populacionais, ok?

Outro exemplo: suponha que você queira avaliar a média de altura dos alunos da sua sala. Caso você meça todos os alunos da sala e use esses dados para medir a média, então estará usando dados populacionais. Caso você não tenha tempo, ou não tenha como medir a altura de todos os alunos e meça apenas uma parte dos alunos, então estará trabalhando com dados amostrais. Fácil, não é?

Mais uma vez, admitindo a possibilidade de repetição de alguns dos valores, pode-se reescrever a expressão da variância amostral, por exemplo, como:

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left[(V_i - \overline{V})^2 \times f_i\right]}{n}$$

Se os dados forem populacionais, teremos:

$$S^2 = \frac{\sum_{i=1}^{n}\left[(V_i - \overline{V})^2 \times f_i\right]}{n-1}$$

Vamos desenvolver o cálculo para o exemplo das ações (dados amostrais) e tentar interpretar o resultado:

$$S^2 = \frac{(24{,}04 - 23{,}95)^2 + (25{,}70 - 23{,}95)^2 + \ldots + (20{,}32 - 23{,}95)^2}{22 - 1} = \frac{86{,}24}{21} = 4{,}11$$

Variância Amostral $S^2 = 4{,}11$

Estamos vendo que, se tivermos muitos dados, teremos um cálculo bem grande pela frente. Entretanto se tivermos dados repetidos, podemos simplificar a fórmula incluindo um termo **fi**, que é na verdade a quantidade de vezes que o valor **Vi** se repete. Pode-se, então, reescrever a expressão da variância amostral, como:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left[(V_i - \overline{V})^2 \times f_i \right]}{n}$$

Se os dados forem populacionais, teremos:

$$S^2 = \frac{\sum_{i=1}^{n} \left[(V_i - \overline{V})^2 \times f_i \right]}{n - 1}$$

24

Vejamos um exemplo. Querendo saber qual é o gasto médio e a variância dos salários de uma pequena empresa. Os cargos, a quantidade de empregados em cada cargo e o salário de cada cargo são mostrados na tabela a seguir:

Cargo	Quantidade	Salário ($)
Operário	100	600
Gerente	10	10.000
Diretor	5	20.000
Presidente	1	30.000

Como estamos analisando os salários, então os valores de salário são os nossos Vi, e como a quantidade é a número de funcionários com aquele salário, então esse será o nosso fi.

Vamos primeiramente calcular a média, usando a fórmula da média que leva em conta os valores repetidos:

$$\overline{V} = \frac{\sum_{i=1}^{n} (V_i \times f_i)}{n}$$

Assim:

$$\overline{V} = \frac{(100x600)+10x10000)+(5x20000)+(1x30000)}{100+10+5+1} = 2500$$

Como estamos considerando todos os funcionários da empresa, então estamos trabalhando com dados populacionais. Logo, a variância será dada por:

$$\sigma^2 = \frac{(600-2500)^2 \text{ x}100+(10000-2500)^2 \text{ x}10+(20000-2500)^2 \text{ x}5+(30000-2500)^2 \text{ x}1)}{116}$$

$$\sigma^2 = \frac{(-1900)^2 \text{ x}100+(7500)^2 \text{ x}10+(17500)^2 \text{ x}5+(27500)^2 \text{ x}1)}{116}$$

$$\sigma^2 = \frac{3610000x100+56250000x10+306250000x5+756250000x1}{116} = 9675517,241$$

25

Entendendo melhor a variância e a sua fórmula

Para entender se o grupo varia pouco ou muito temos que definir um valor de referência e verificar qual a "distância" entre cada ponto (valor de Vi) e essa referência. Qual a nossa referência? Acertou! A média!

Legal, então se soubermos a distância entre cada ponto e a média, temos uma ideia da variação do conjunto de dados? Não, você terá ideia da variação de cada ponto do conjunto, mas não uma medida de variação do conjunto, como um todo. Para ter a ideia de variação do conjunto, somamos cada uma dessas distâncias e dividimos pela quantidade de dados:

Variância do conjunto (σ^2) = soma da distância de cada ponto (Vi) em relação à média / quant. de dados reescrevendo:

$$\sigma^2 = \frac{\Sigma \text{ distância de cada ponto Vi}}{\text{quantidade de dados}}$$

Não está parecida com a fórmula original? E olha que conclusão interessante: a variação do grupo é a média das "distâncias"! Vale a pena refletir sobre isso.

26

Para entender o que nós estamos falando, observe a nossa fórmula da variância populacional:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (V_i - \mu)^2}{n}$$

Podemos ver que $(V_i - \mu)^2$ representa justamente a distância de que estávamos falando: a distância entre o ponto Vi e a média.

Quando todas as diferenças foram elevadas ao quadrado elas "inflacionaram" a diferença real entre cada valor e o valor médio que se quer validar como representativo (do ponto de vista aplicado) para a base de dados.

24,04 -	Média =	0,0905
25,70 -	Média =	1,7505
23,14 -	Média =	-0,8095
22,87 -	Média =	-1,0795
23,42 -	Média =	0,5295
23,96 -	Média =	0,0105
25,69 -	Média =	1,7405
26,78 -	Média =	2,8305
26,88 -	Média =	2,9305
27,05 -	Média =	3,1005
27,05 -	Média =	3,1005
26,12 -	Média =	2,1705
24,60 -	Média =	0,6505
23,56 -	Média =	-0,3895
23,04 -	Média =	-0,3895
21,30 -	Média =	-2,6495
21,75 -	Média =	-2,1995
21,85 -	Média =	-2,0995
22,89 -	Média =	-1,0595
23,14 -	Média =	-0,8095
21,74 -	Média =	-2,2095
20,32 -	Média =	-3,6295
Soma	=	0,0000

Mas por que estamos trabalhando com distâncias ao quadrado? Não bastaria ter trabalhado apenas com as diferenças, sem elevá-las ao quadrado?

Infelizmente não, pois as distâncias poderiam se anular, gerando um falso valor de variância, como podemos ver na imagem.

As diferenças positivas (valores reais maiores que a média) e as diferenças negativas (valores reais menores que a média) simplesmente se anulam, o que em nada irá contribuir para que cheguemos a uma conclusão sobre a variabilidade. Elevar ao quadrado é um recurso matemático, que é utilizado para trabalhar apenas com números não negativos.

Uma observação importante: o fato de elevar ao quadrado também faz com que a unidade de medida da variância não seja a unidade de medida original dos dados.

Veja um exemplo.

Exemplo

Suponhamos que o peso médio de indivíduos fosse de 72,4 kg. Se tomarmos o peso do João que é de 78,3 kg e fazer (78,3 kg - 72,4 kg)², o resultado seria de 34,81 kg².
O valor de 34,81Kg é uma unidade de difícil interpretação e, mesmo que fosse simples, ela não corresponde à original, o que compromete a avaliação, uma vez que estamos estudando pesos em quilogramas e gramas e a variabilidade está expressa em quilogramas ao quadrado!)

27

Desvio-padrão

Vimos que foi necessário elevar as diferenças ao quadrado para resolver um problema, entretanto, geramos outro, que é o fato da variância ter uma unidade diferente daquela unidade original dos dados. A operação inversa, raiz quadrada, deve solucionar estes dois pontos simultaneamente. Veja a seguir.

Para dados amostrais:

$$\sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (V_i - \overline{V})^2}{n-1}}$$

E para dados populacionais:

$$\sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (V_i - \mu)^2}{n}}$$

Esta nova medida é denominada **desvio-padrão**, simbolizada por S no caso de dados amostrais e por σ no caso de dados populacionais.

Aplicando às ações (dados amostrais), teremos:

$$S = \sqrt{S^2} \rightarrow S = \sqrt{4{,}11} \rightarrow S = 2{,}03.$$

Coeficiente de variação

Agora que já temos uma medida de variação que tem a mesma unidade dos dados originais, o desvio-padrão, já podemos responder as perguntas abaixo:

1. um desvio-padrão de R$ 2,03 significa que os dados das 22 cotações formam um conjunto suficientemente homogêneo ou não?

2. A média é capaz de retratar com fidelidade esta base de dados ou não?

3. Resumindo: o desvio-padrão de R$ 2,03, aqui encontrado, é pequeno ou grande?

Num primeiro momento a única resposta que parece adequada é:

... o desvio-padro 2,03 só pode ser considerado pequeno ou grande quando comparado com a ordem de grandeza dos dados que estão sendo analisados, pois se pouco mais de dois reais dão uma ideia de muito pouco dinheiro, isto não parece ser verdade se os preços/valores alvo de investigação são de algum dispositivo eletrônico que oscilam de 80 centavos a 3 reais e vinte centavos.

29

Então devemos ter uma ideia relativa deste desvio-padrão, o que dá origem a uma nova medida denominada **coeficiente de variação** (simbolizada por CV apresenta a mesma formulação matemática tanto para dados amostrais, como para dados populacionais):

$$CV = \frac{Desvio - Padrão}{Média}$$

Logo, para dados amostrais, teremos:

$$CV = \frac{S}{\bar{V}}$$

Já, para dados populacionais, teremos:

$$CV = \frac{\sigma}{\mu}$$

No exemplo, tem-se que:

$$CV = \frac{2,03}{23,95} \cong 0,0848 \cong 8,48\%$$

Uma vantagem adicional do coeficiente de variação é a possibilidade de comparação da variabilidade de duas variáveis com unidades de medidas distintas.

Veja um Exemplo:

Exemplo
Se um grupo de crianças de 1 ano está sendo estudado com relação a variação de seu peso e altura durante a orientação de uma nova dieta, pode ser interessante conhecer qual das duas variáveis têm um comportamento mais homogêneo, mais parecido. Isto não é possível com o desvio-padrão, pois o de uma variável estaria em gramas e o da outra em centímetros. Com o recurso do coeficiente de variação, que é adimensional, consegue-se contornar aquela impossibilidade, pois para ambas as variáveis serão obtidos percentuais de variação, permitindo a visão de qual das duas variáveis (peso ou altura) apresentou maior variabilidade.

30

Um último ponto a indagar com relação à medida de variação dos dados é o seguinte:

- Qual o percentual de variação que caracterizaria um grupo como homogêneo ou heterogêneo?

Do visto até agora, percebe-se que um coeficiente de variação igual a zero ocorrerá quando **todos os valores forem iguais** (assim a variância é zero, o desvio-padrão é zero e, consequentemente, o coeficiente é zero por cento).

Logo, quanto mais próximo disso, maior a homogeneidade do grupo e mais consistente é a validade da média como medida representativa daqueles dados (lembrando que a média, nesse caso, poderia ser vista como o valor "capaz" de substituir todos os outros com fidelidade, devido à proximidade que estaria sendo mantida entre eles, a menos dos pontos discrepantes, que em breve seremos capazes de diagnosticar).

31

Uma regra empírica (baseada na experiência, no uso) indica que:

• quando CV for de até 15%, aproximadamente, a variabilidade/dispersão pode ser considerada pequena;
• acima disso e até 30%, aproximadamente, a variabilidade/dispersão pode ser considerada média;
• acima disso, tem-se uma variabilidade/dispersão que pode ser considerada suficientemente grande.

Deve-se ter um cuidado adicional com as seguintes circunstâncias:

• quando a base de dados for muito pequena (por exemplo, 10 ou 12 observações), pois um ou dois valores podem ser suficientes para gerar um coeficiente de variação mais alto, pois 1 valor em 10 corresponde a 10% da base de dados. Pode-se, neste caso, ser um pouco mais flexível com os limites de 15 e 30% indicados anteriormente;

• quando a análise estiver inserida num contexto de gestão ou controle de qualidade, um coeficiente de 12% pode ser considerado muito alto. Exemplos:

1. quando por exemplo se está analisando medidas do diâmetro, comprimento, espessura ou largura de um componente;
2. quando as avaliações implicarem risco à vida ou à integridade de indivíduos, pode-se ser sensivelmente mais rigoroso para admitir a homogeneidade da base de dados.

Para esses casos, existem técnicas estatísticas mais apropriadas.

32

Finalmente, é possível chegar a uma conclusão a respeito da validade da cotação média de R$ 23,95 para o lote de 1000 ações da Companhia C ao longo do mês estudado.

Como o coeficiente de variação foi de 8,48% e esse percentual pode ser considerado pequeno, então a cotação média é um bom indicador das cotações ao longo daquele mês, representando o grupo de

cotações como um todo (mais uma vez destacando que será indispensável verificar a existência de pontos discrepantes, que, se existirem, devem merecer análise específica para serem interpretados corretamente).

33

b) Pontos discrepantes

Será que existem pontos discrepantes em nosso conjunto de dados?

Existe um teorema (desigualdade de Tchebycheff, ou Tchebichev, ou Chebychev segundo a grafia de alguns textos) que permite assegurar que para qualquer distribuição amostral de dados com uma determinada média e um determinado desvio-padrão:

a) pelo menos 75% dos dados amostrais estarão compreendidos no intervalo (média - 2 desvios; média + 2 desvios);

b) pelo menos 88,89% dos dados amostrais estarão compreendidos no intervalo (média - 3 desvios; média + 3 desvios).

34

Um critério, então, para o diagnóstico acerca da presença de pontos discrepantes é o seguinte:

...todos os pontos que estiverem fora do intervalo delimitado pela média mais ou menos três desvios será passível de ser considerado discrepante (ou, em outra terminologia, um outlier)...

No exemplo da cotação da Companhia C, teríamos estes "limites" dados por:

média – 3 desvios = 23,95 – (3 x 2,03) = 23,95 – 6,09 = 17,86
média + 3 desvios = 23,95 + (3 x 2,03) = 23,95 + 6,09 = 30,04

Assim, quaisquer cotações compreendidas entre R$ 17,86 e R$ 30,04 não seriam consideradas atípicas e aquelas cotações, no período observado, que fossem inferiores a R$ 17,86 ou superiores a R$ 30,04 seriam pontos atípicos, pontos discrepantes ou *outliers*.

Conclui-se, então, que na base de dados apresentada não há nenhuma cotação nesta situação. Isto reforça a representatividade da média para aquele contexto, embora, como já tenhamos registrado anteriormente, se o objetivo de sua empresa for realmente decidir a respeito de investir ou não na Companhia C há necessidade de uma análise mais aprofundada.

Mas e quando a média não ajuda? Quando isto ocorre, ou seja, quando a média não é validada como medida descritiva da base de dados, faz-se necessário tentar descrevê-la com o apoio de outras medidas, sob pena de não se ter uma compreensão correta a respeito do comportamento/da distribuição dos dados.

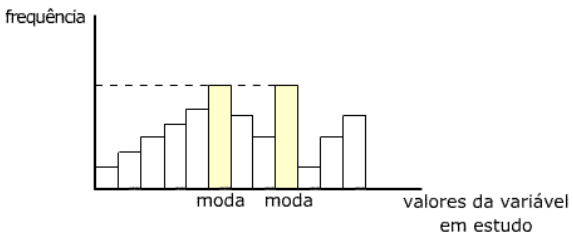
4 - A Moda

Uma primeira alternativa é a moda, que nada mais é do que o valor que aparece com a maior frequência (maior número de repetições). Uma base de dados pode ter um ou mais valores repetindo-se um mesmo número de vezes, ou até mesmo não ter nenhum valor com maior incidência que os demais. A análise da moda é a mesma tanto para dados amostrais, como para dados populacionais. A classificação fica:

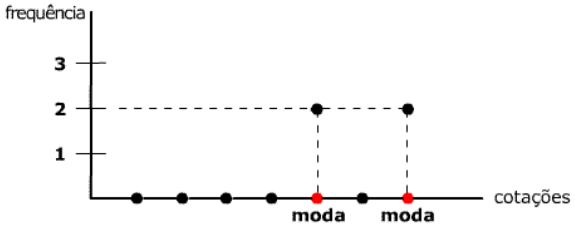
- quando há apenas uma **moda** ? distribuição unimodal
- quando há **duas modas** ? distribuição bimodal
- quando há **mais de duas modas** ? distribuição multimodal
- quando não há **nenhuma moda** ? distribuição amodal

Distribuição Bimodal

Exemplo das cotações das ações



Representação Gráfica de uma Distribuição Bimodal



36

Mais uma vez é necessária parcimônia quando da interpretação prática desta medida, pois ela pode induzir equívocos. No exemplo da Companhia C, qual seria a moda?

24,04; 25,70; 23,14; 22,87; 23,42; 23,96; 25,69; 26,78; 26,88; 27,05; 27,05; 26,12; 24,60; 23,56; 23,04; 21,30; 21,75; 21,85; 22,89; 23,14; 21,74; 20,32.

As cotações de R$ 23,14 e R$ 27,05 aparecem duas vezes, enquanto todas as demais apenas uma vez, logo temos duas modas que são estes dois valores. Dizer que a moda, por ser o valor que mais aparece, representa/descreve o grupo pode estar longe de ser verdade, como é o caso que acaba de ser mostrado, pois cada um dos valores que mais aparece representa menos de 10% da base de dados, assim o "poder" da moda como medida representativa do grupo aumenta à medida que ela corresponda a um percentual expressivo do total de dados (o mesmo valendo caso haja mais de uma moda).

37

5 - A Mediana

Outra alternativa a ser considerada é a que segue.

> Uma base de dados heterogênea, quando "quebrada" em grupos menores tende a gerar subgrupos mais homogêneos e, por isto mesmo, de compreensão/interpretação mais fácil.

Quanto maior a base de dados, maior o número de subgrupos que pode/deve ser produzido para permitir melhor visualização da distribuição dos dados.

Uma primeira possibilidade é dividir a base (ordenada de forma crescente ou decrescente) em dois subgrupos de igual tamanho. O valor que ocupa a posição central é a **mediana**.

Quando o número de observações for **ímpar** existe uma única posição central e o valor correspondente a esta posição será a mediana.

1, 2, 3, 4, 5

Porém quando o número de observações for **par** não há uma única posição central, mas sim duas, e a mediana será a **média dos dois valores** correspondentes a estas duas posições centrais. Essa análise, de determinação da mediana, também é idêntica tanto para dados amostrais, como para dados populacionais.

1, 2, 3, 4

38

Este seria o caso do exemplo da Companhia C. Os dados ordenados ficariam:

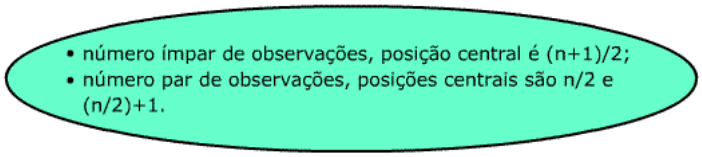
20,32; 21,30; 21,74; 21,75; 21,85; 22,87; 22,89; 23,04; 23,14; 23,14; 23,42; 23,56; 23,96; 24,04; 24,60; 25,69; 25,70; 26,12; 26,78; 26,88; 27,05; 27,05

Como são vinte duas posições, as centrais são as 11ª e 12ª, "ocupadas" pelas cotações R$ 23,42 e R$ 23,56, logo a mediana será dada por:

$$\text{Mediana} = \frac{23{,}42 + 23{,}56}{2} = 23{,}49$$

Caso fossem 21 ou 23 dados, a posição central seria única, no primeiro caso a 11ª, e no segundo a 12ª.

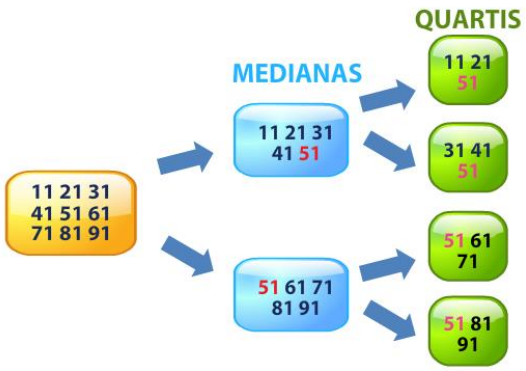
Quando a base de dados não é pequena permitindo fácil visualização deste centro, pode-se utilizar o recurso abaixo para identificá-lo:



39

Voltando à mediana calculada, a interpretação é mais ou menos óbvia, pois o valor mediano de R$ 23,49 indica que 50% das cotações foram de até R$ 23,49 enquanto 50% das cotações ficaram de R$ 23,49 para cima. Quando a base é suficientemente **homogênea**, em particular se não há pontos discrepantes, a mediana pode não agregar valor à análise, pois a média já fazia um bom papel. Quando a base for muito **heterogênea** (especialmente se for uma base grande) só a mediana não é suficiente para clarear substancialmente nossa análise descritiva.

Assim, pode-se dividir o conjunto de observações em uma quantidade maior de subgrupos. Quando se utilizam os quartis para dividir a base de dados, são gerados quatro subgrupos de igual tamanho, o primeiro deles compreendido entre o valor mínimo e o primeiro quartil, o segundo entre o primeiro e o segundo quartil (que é a própria mediana), o terceiro entre o segundo e o terceiro quartil e o quarto (e último subgrupo) entre o terceiro quartil e o valor máximo do grupo analisado.



Quartis

São três valores numéricos que dividem a base de dados (ordenada de forma crescente) em quatro subconjuntos, cada um deles contendo 25% dos dados originais.

40

Por analogia com a determinação da mediana, o primeiro passo para a determinação do 1º e 3º quartis é identificar sua posição na base de dados ordenada. Caso tivéssemos 100 observações (apenas para facilitar o entendimento), o 1º quartil estaria localizado entre a 25ª e 26ª posições, enquanto o 3º quartil estaria entre a 75ª e 76ª posições. A mediana (ou 2º quartil) estaria entre a 50ª e 51ª posições.

O procedimento, neste caso, e como já visto para a mediana, também seria o cálculo das médias entre os valores que ocupassem aquelas posições. Importante frisar, mais uma vez, que os quartis são três e que são valores numéricos específicos, permitindo a divisão de uma base de dados em quatro subgrupos.

No caso de 22 dados, o primeiro quartil seria o valor correspondente à 6ª posição, e o terceiro quartil seria o valor correspondente à 17ª posição, ou seja, R$ 22,87 e R$ 25,70, respectivamente (apenas relembrando que o segundo quartil, igual a R$ 23,49, foi dado pela média aritmética entre os valores localizados na 11ª e 12ª posições).

Para bases de dados maiores podem ser necessárias quebras/divisões das bases de dados em uma quantidade maior de subgrupos, por exemplo dez ou cem, o que é feito com a ajuda dos 9 decis e dos 99 percentis, respectivamente.

Na maioria dos textos estatísticos as medidas aqui tratadas são agrupadas em:

1. medidas de posição (ou de tendência central):média, moda, mediana, quartis, decis, percentis;
2. medidas de dispersão (ou de variabilidade): amplitude, variância, desvio-padrão, coeficiente de variação.

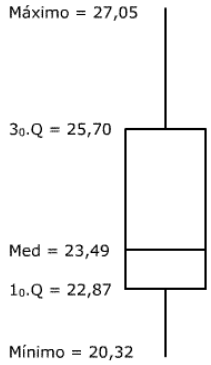
41

6 - O Box Plot

J. W. Tukey, em seu texto sobre Análise Exploratória de Dados, de 1977, sugere um conjunto de cinco medidas para representar uma base de dados numéricos: os valores extremos (mínimo e máximo) e os três quartis (também denominados juntas), que seriam medidas mais robustas/resistentes, considerando que seriam pouco ou nada afetadas caso houvesse variação em uma pequena quantidade daqueles dados, o que é especialmente conveniente quando a base não for suficientemente grande (medidas estatísticas descritivas calculadas a partir de bases de dados suficientemente grandes não são

significativamente influenciadas, em geral, por uma pequena parte dela, mesmo que sejam valores atípicos). Estas cinco medidas fazem-se acompanhar do total de dados analisados.

Pode-se construir, a partir delas, um desenho esquemático denominado *box plot* ou diagrama de caixas (disponível em alguns *softwares*estatísticos), ilustrado com resultados, já calculados, para as ações da Companhia C.



A utilização de uma planilha eletrônica, como Microsoft Excel, muito nos ajuda, simplificando o exercício de ter que lembrar todas as fórmulas e o "trabalho braçal" de efetuar todas as contas. Certamente ela em nada contribui em termos analíticos, pois se não formos capazes de interpretar os resultados gerados ficaremos como que "perdidos", sem sermos capazes de fornecer subsídios consistentes para o processo de tomada de decisão.

Exemplo de análise de dados estatísticos publicados na mídia.

O brasileiro está vivendo mais num país que lhe é perigoso

A expectativa de vida do brasileiro cresceu 2,6 anos durante a década passada, seguindo uma tendência mundial de aumento da longevidade, segundo dados do IBGE divulgados na segunda-feira 3. A mulher está vivendo em média 72 anos e o homem, 64. Essa é uma forma técnica de se ler a estatística. Há outra no âmbito social e da saúde pública. Não pessimista mas realista: a expectativa de vida no Brasil é a segunda

pior da América do Sul, e aqui só se vive um pouco mais do que na Bolívia. No mundo, os brasileiros estão em 30º lugar. E os campeões da longevidade são Japão, Suécia e Canadá (homem: 79 anos; a mulher: 80). A diferença de expectativa de vida entre os dois sexos no Brasil embute uma bomba-relógio: a mortalidade entre os homens com idades de 15 a 35 anos é 3,5 vezes maior do que entre as mulheres, motivada principalmente pelo aumento da violência - incluídas as chacinas. Outra bomba: a mulher, se está vivendo mais, nem por isso está ganhando mais atenção social: uma entre quatro é portadora de HPV, responsável pelo câncer de útero. Some-se a isso o fato de que o brasileiro está ganhando mais idade, mas num país que já aponta para cerca de 1,3 milhões casos de mal de Alzheimer, sem uma política de saúde para cuidar desse estado demencial que acomete justamente os velhos. A última bomba, segundo o próprio Ministério da Saúde, dá conta de um acelerado avanço da hepatite C, que já atinge 3,3 milhões de pessoas.

RESUMO

Neste módulo foram trabalhadas algumas medidas descritivas básicas para análise de uma base de dados não agrupados. Falamos da média, como medida mais simples para representar esta base, desde que a mesma seja suficientemente homogênea (ou seja, variabilidade suficientemente pequena).

Esta avaliação da homogeneidade/heterogeneidade será feita com base no coeficiente de variação, que é uma expressão relativa do desvio-padrão (comparativamente à média), sendo então adimensional, apresentando de forma percentual a variabilidade.

Vimos que, neste caso, foram estabelecidos intervalos que, empiricamente, têm sido utilizados e aceitos como válidos para expressar a ordem de grandeza da dispersão dos dados: até 15% (pequena); acima de 15% até 30% (média) e acima de 30% (grande), devendo-se ter o cuidado de considerar o contexto sob análise, que pode exigir um limite muito inferior a 15% para admissão de homogeneidade dos dados (como muitas situações de gestão da qualidade, por exemplo).

Surge então a necessidade, tanto com bases de dados homogêneas, como com heterogêneas, de diagnosticar a existência de pontos atípicos/discrepantes (ou *outliers*), que são aqueles que "fogem" excessivamente do perfil traçado pelo conjunto como um todo. Isto pode ser feito por duas abordagens: (1) estabelecer os limites máximos e mínimos de "tolerância" a partir da média mais ou menos três vezes o desvio-padrão, o que assegurará que este intervalo contenha no mínimo 88,89% dos dados amostrais (desigualdade de Tchebychev), sendo os restantes considerados pontos discrepantes; (2) estabelecer estes limites partindo do primeiro e terceiro quartis, fazendo então a diferença entre estes dois valores, multiplicando esta diferença por 3/2, somando este resultado ao terceiro quartil e subtraindo-o do primeiro quartil.

Feito isto e no caso de bases de dados que não sejam consideradas suficientemente homogêneas, deve-se partir para medidas complementares, uma vez que a média não representa o conjunto analisado, nestas circunstâncias. As medidas apresentadas foram moda, mediana e os próprios quartis (além de citação dos decis e percentis). A moda é o valor que mais aparece, aquele com maior incidência, o que pode agregar valor à análise desde que represente uma quantidade significativa de dados (uma moda que represente, por exemplo, dois por cento do grupo pode não caracterizar aquele valor como um destaque frente aos demais), considerando-se que a moda é única ou múltipla (ou mesmo se existe uma moda). A mediana divide o grupo, já ordenado (de forma crescente ou decrescente), deixando metade dos dados entre o valor mínimo e ela, mediana, e a outra metade entre ela e o valor máximo observado. O objetivo da utilização de valores que dividam a base de dados em grupos menores é visualizar melhor o comportamento de sua distribuição, uma vez que estes subgrupos devem ser mais homogêneos.

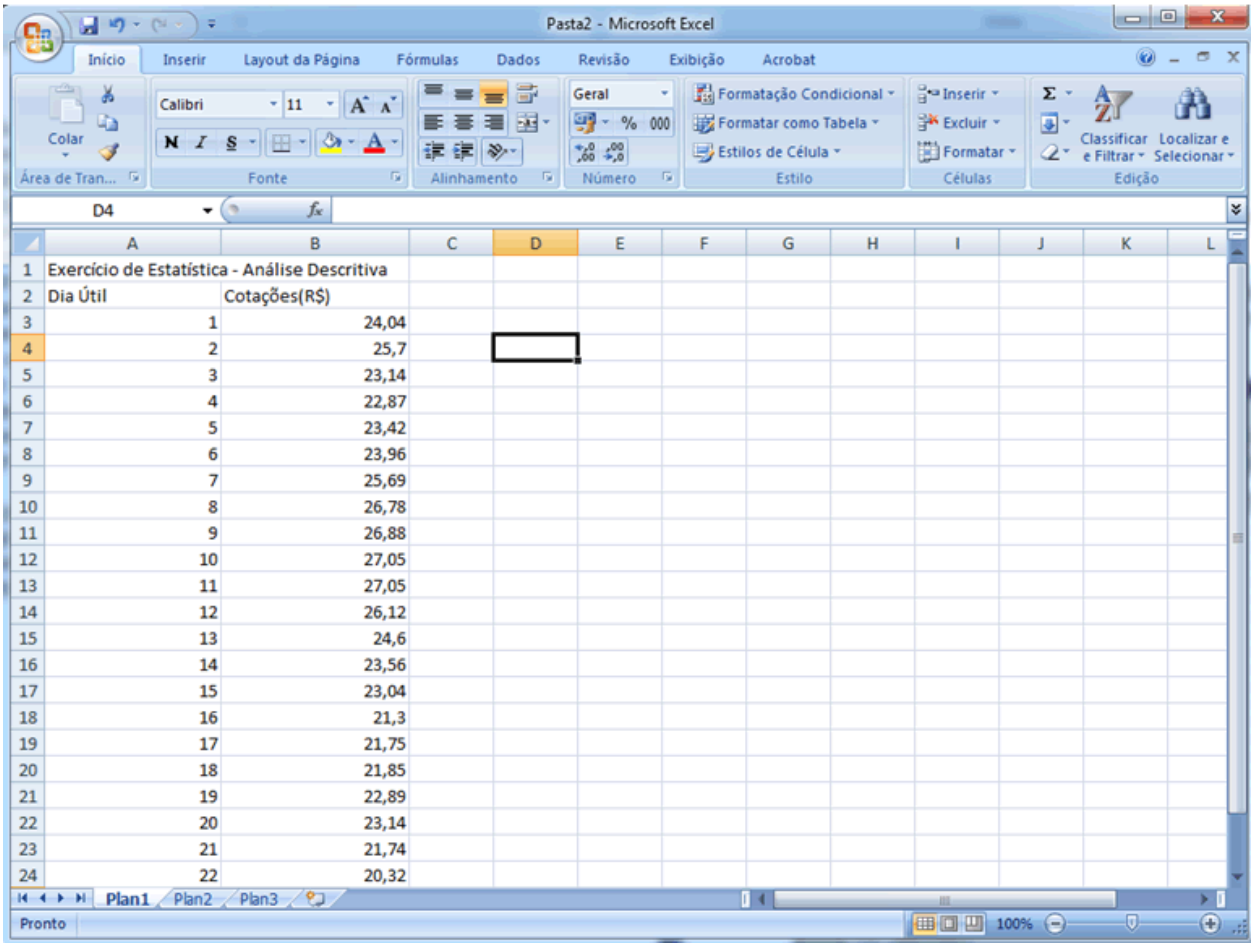
Foi também apresentada uma proposta sintética, idealizada por Tukey, e subsequente estrutura esquemática para suporte à análise descritiva. Apresentam-se simultaneamente os valores mínimos e

máximos e os três quartis. O *box-plot* permite ter uma ideia a respeito da distribuição, sendo que os pontos discrepantes também são sinalizados no desenho.

1 - DETERMINAÇÃO DE MEDIDAS DESCRITIVAS

A **Planilha de Cálculo (Excel)** irá nos auxiliar na tarefa de determinar o valor das medidas descritivas.

Será mantido inicialmente, até para permitir comparação, o exemplo das cotações das ações da Companhia C. O primeiro passo é abrir uma planilha e digitar os dados levantados:



Exemplo das cotações das ações

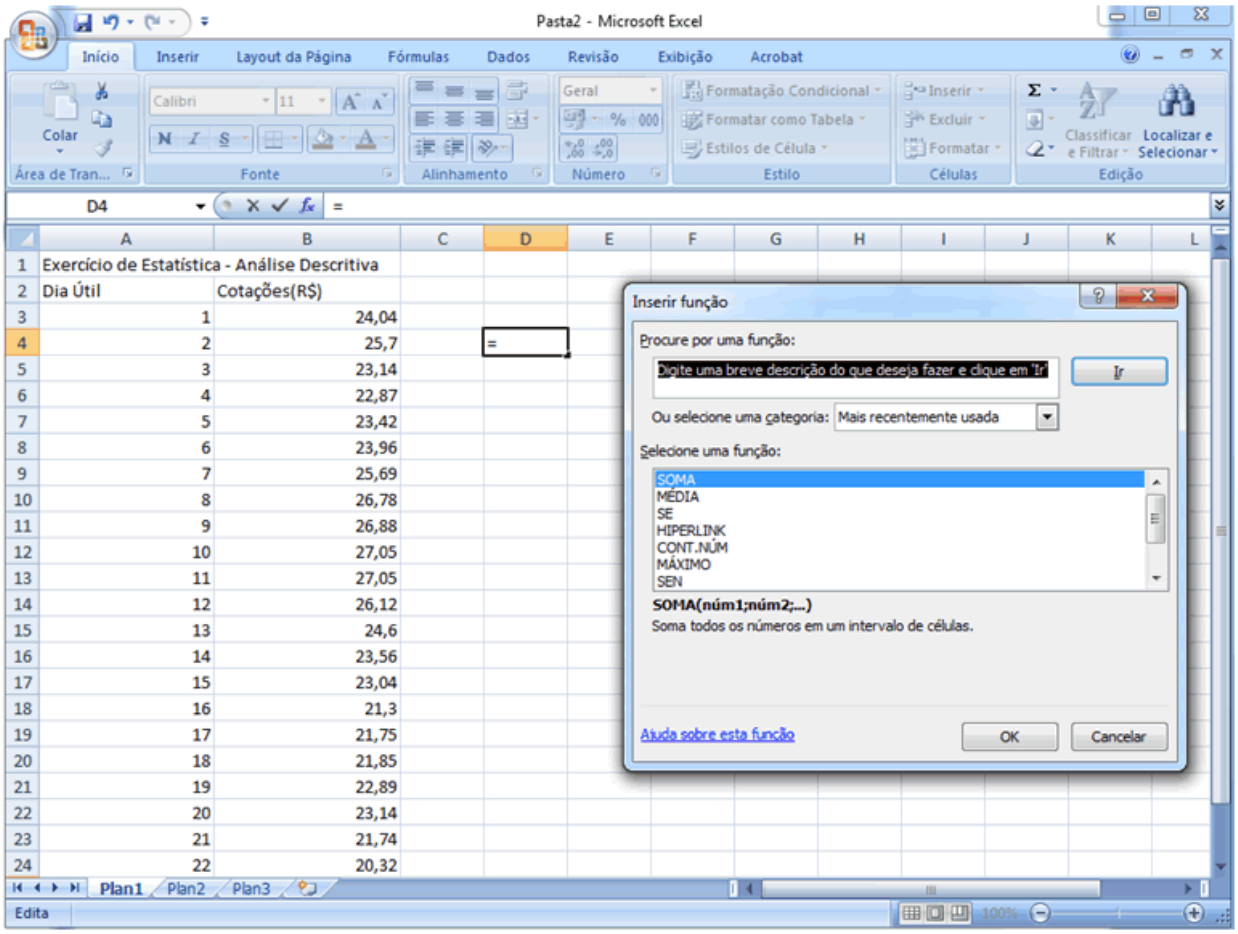
Considere que a empresa para a qual você trabalha solicita que você acompanhe a cotação da ação de uma determinada companhia (companhia C) na Bolsa de Valores de São Paulo ao longo de um mês (22

dias úteis). Diligentemente você faz isto e ao final daquele mês você tem os seguintes dados (em reais, o lote de 1000 ações):
24,04; 25,70; 23,14; 22,87; 23,42; 23,96; 25,69; 26,78; 26,88; 27,05; 27,05; 26,12; 24,60; 23,56; 23,04; 21,30; 21,75; 21,85; 22,89; 23,14; 21,74; 20,32.

44

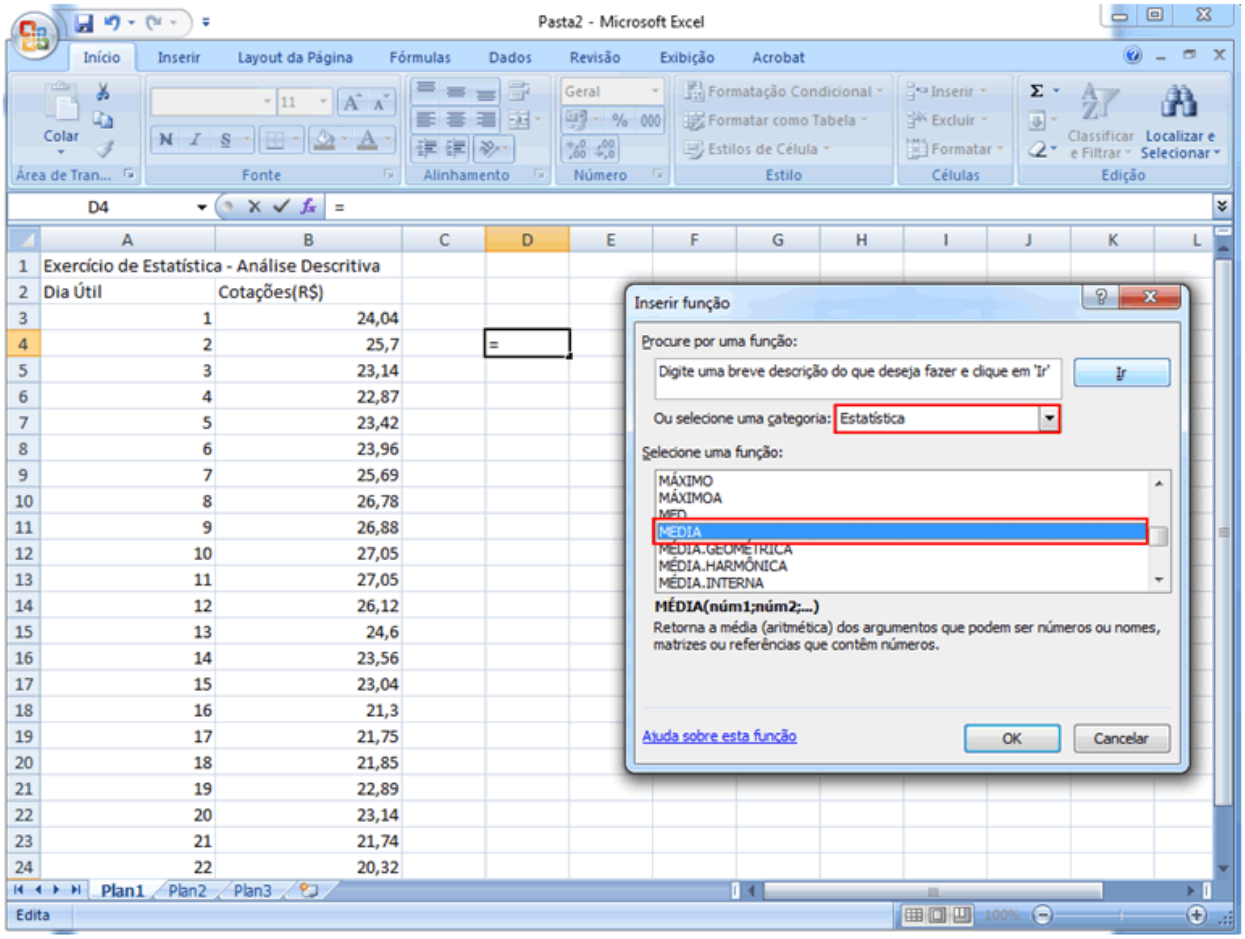
2 - Calculando a Média

Após digitar os dados do exemplo da Companhia C, selecione a célula na qual você deseja que seja inserido o valor da média e posicione o cursor ali, por exemplo, célula D4, e clique no ícone f_x (colar função) da barra de ferramentas (e sinalizado na planilha abaixo):



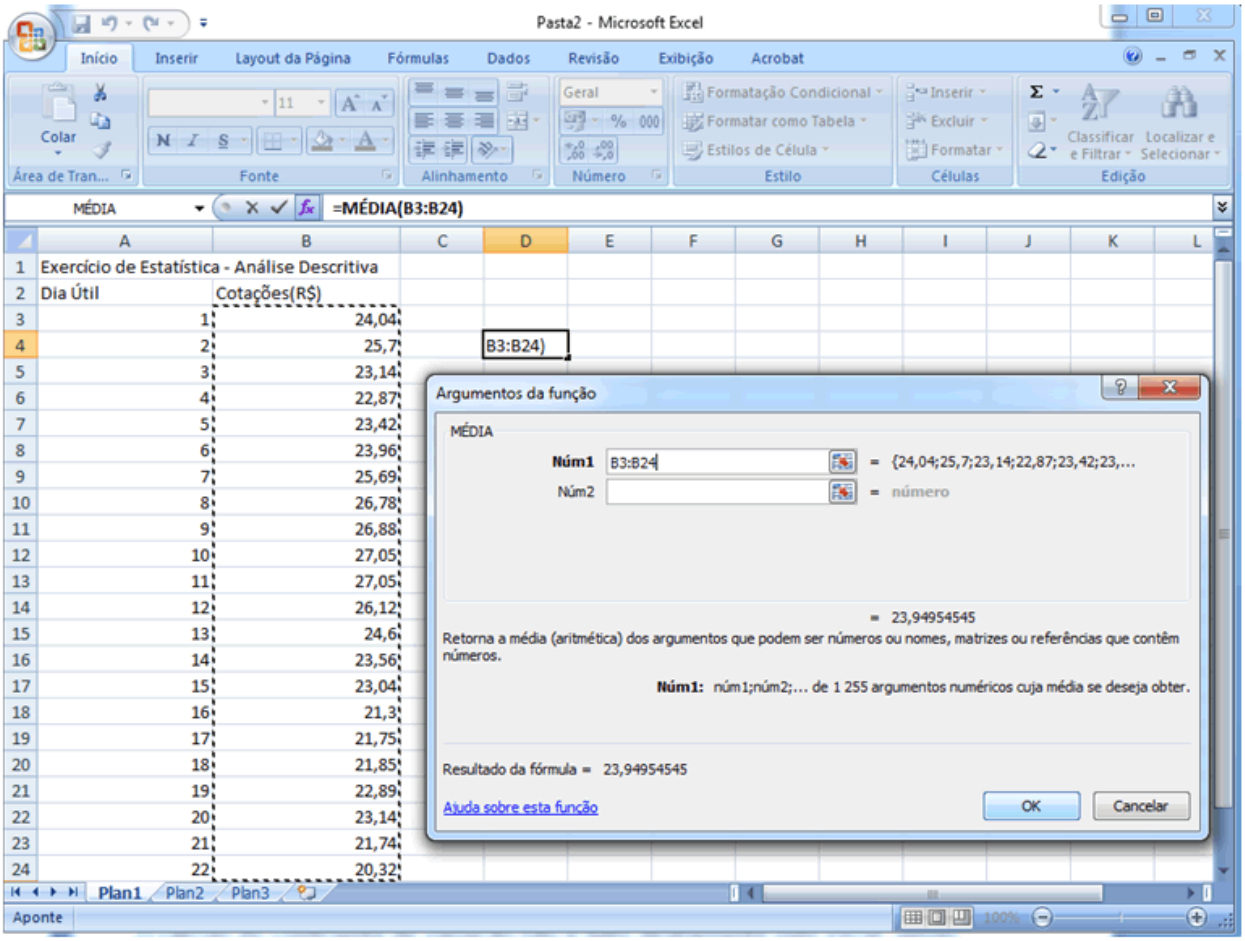
45

Selecione agora as opções Estatística e Média, tal como indicado na próxima tela:



46

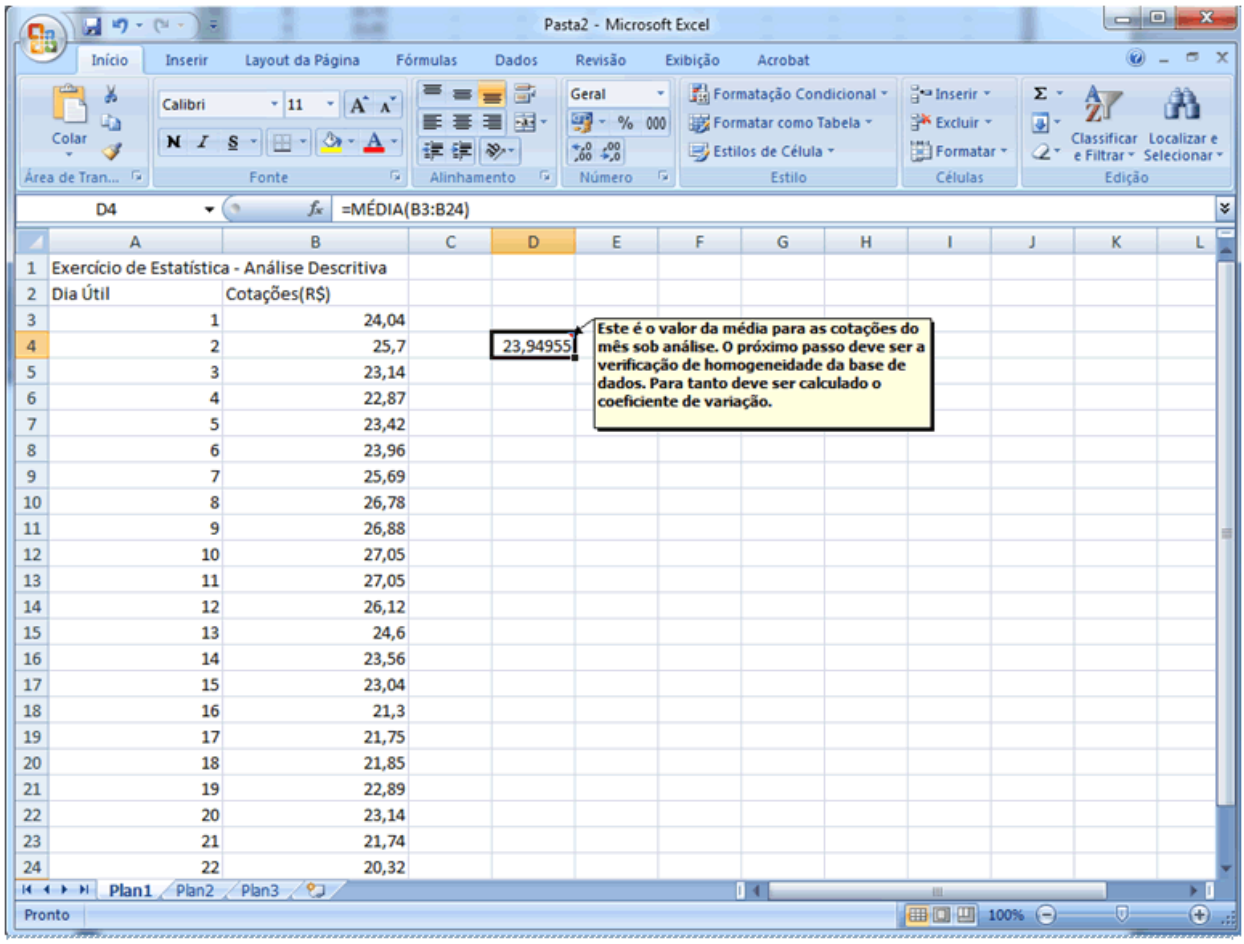
Ao clicar em OK você terá (observe que já foi feita a seleção das posições nas quais se encontram os valores, ou seja de B3 a B24, o que tanto pode ser feito com o botão esquerdo do mouse como digitando-se diretamente no campo apropriado da tela Média, iniciado por Núm1):



Já é possível ver o resultado da fórmula no canto inferior esquerdo da tela.

47

Ao clicar em OK o resultado da fórmula aparecerá na célula D4, originalmente selecionada, podendo-se inserir um comentário ou digitar na célula ao lado alguma nota de esclarecimento.

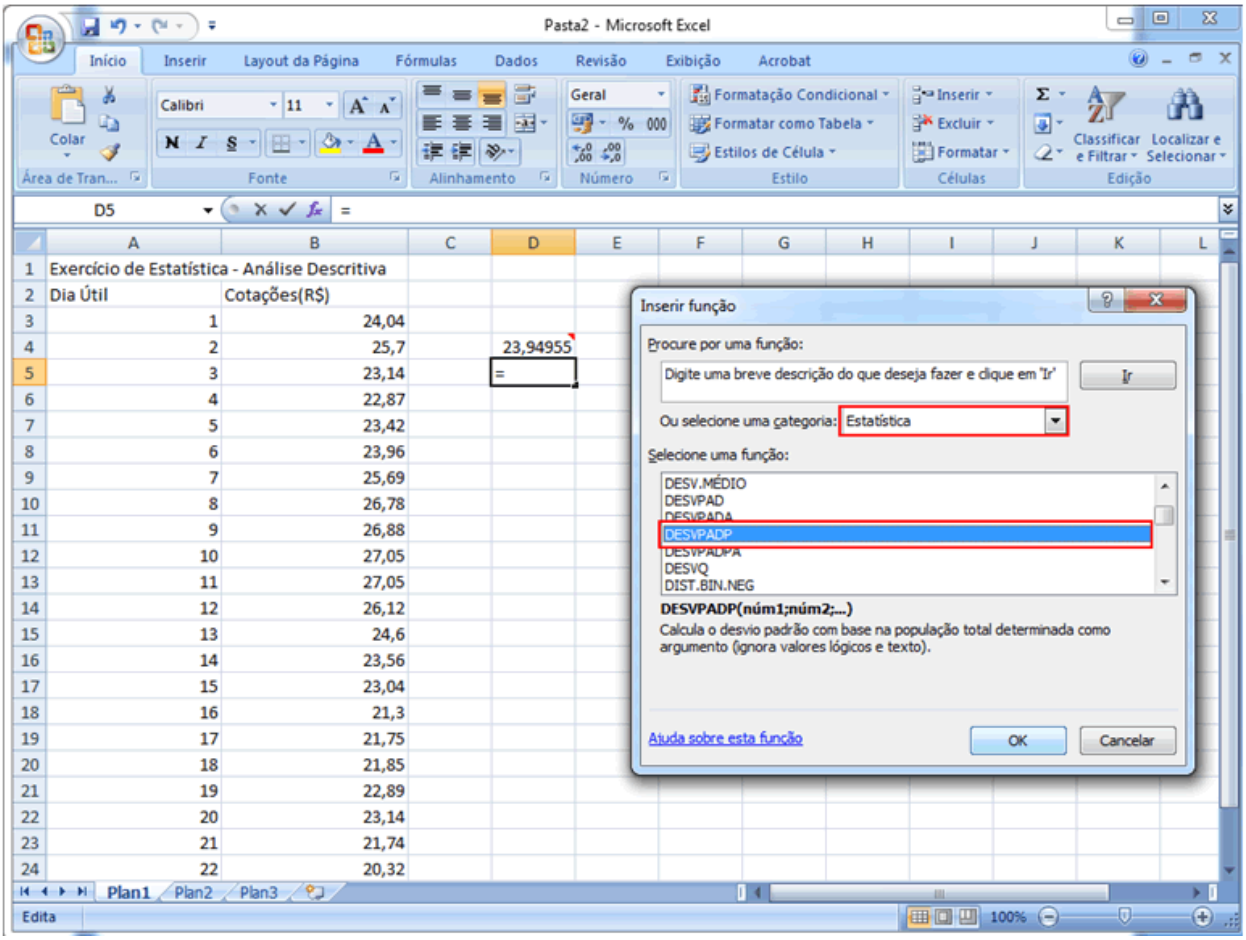


	A	B	C	D	E	F	G	H	I	J	K	L
1	Exercício de Estatística - Análise Descritiva											
2	Dia Útil	Cotações(R$)										
3	1	24,04										
4	2	25,7		23,94955								
5	3	23,14										
6	4	22,87										
7	5	23,42										
8	6	23,96										
9	7	25,69										
10	8	26,78										
11	9	26,88										
12	10	27,05										
13	11	27,05										
14	12	26,12										
15	13	24,6										
16	14	23,56										
17	15	23,04										
18	16	21,3										
19	17	21,75										
20	18	21,85										
21	19	22,89										
22	20	23,14										
23	21	21,74										
24	22	20,32										

48

3 - Calculando o Coeficiente de Variação

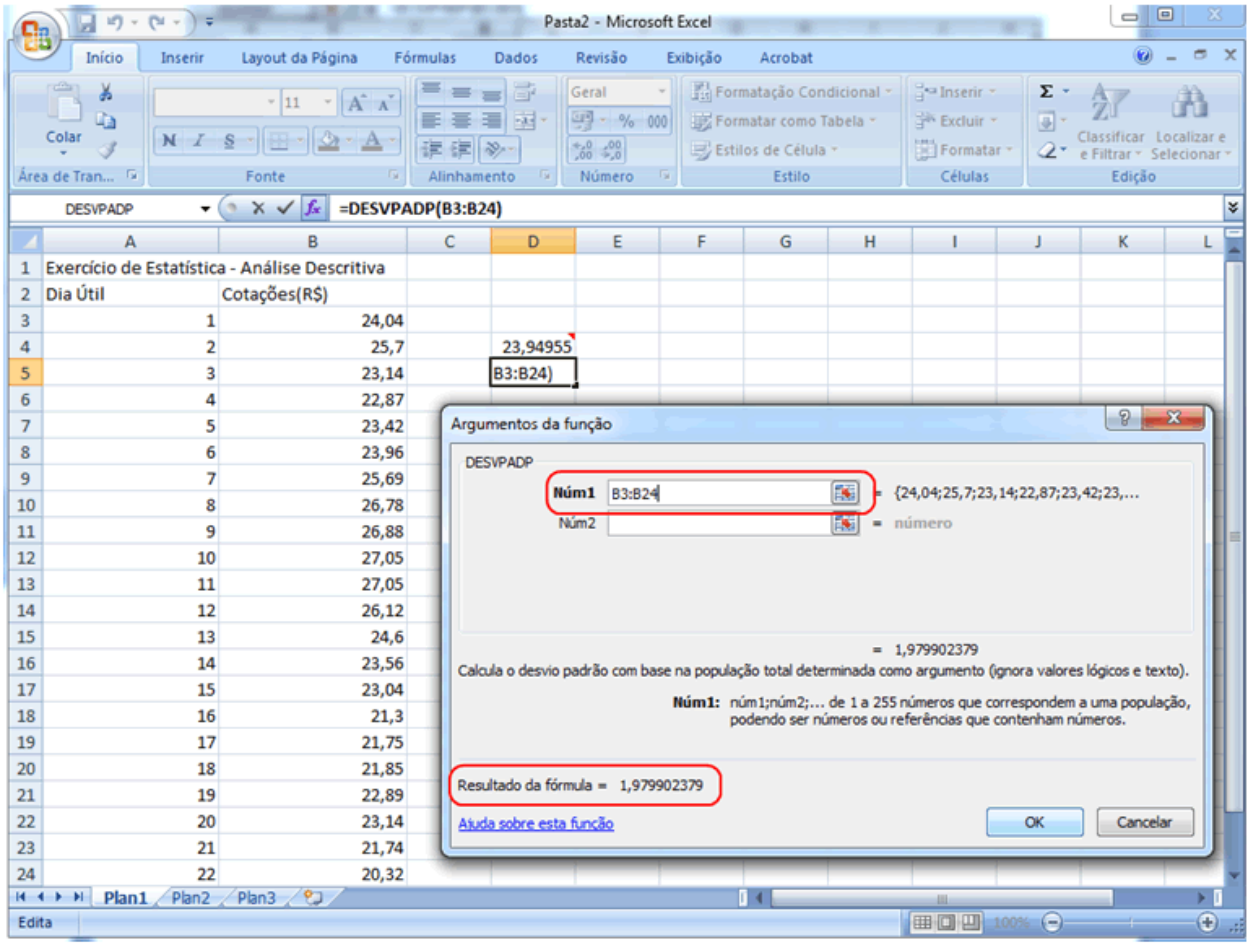
O cálculo do coeficiente de variação não é feito diretamente pelo Excel, sendo necessário calcular o desvio-padrão para depois dividi-lo pela média. O procedimento é análogo ao cálculo da média sendo que será selecionada a célula D5 e quando da seleção da opção Estatística deve-se também selecionar DESVPADP (desvio padrão populacional). Caso você faça DESVPAD (desvio padrão amostral), o Excel adotará n-1 no denominador da variância. Assim:



Clique em OK para ir para a caixa de diálogo da função selecionada.

49

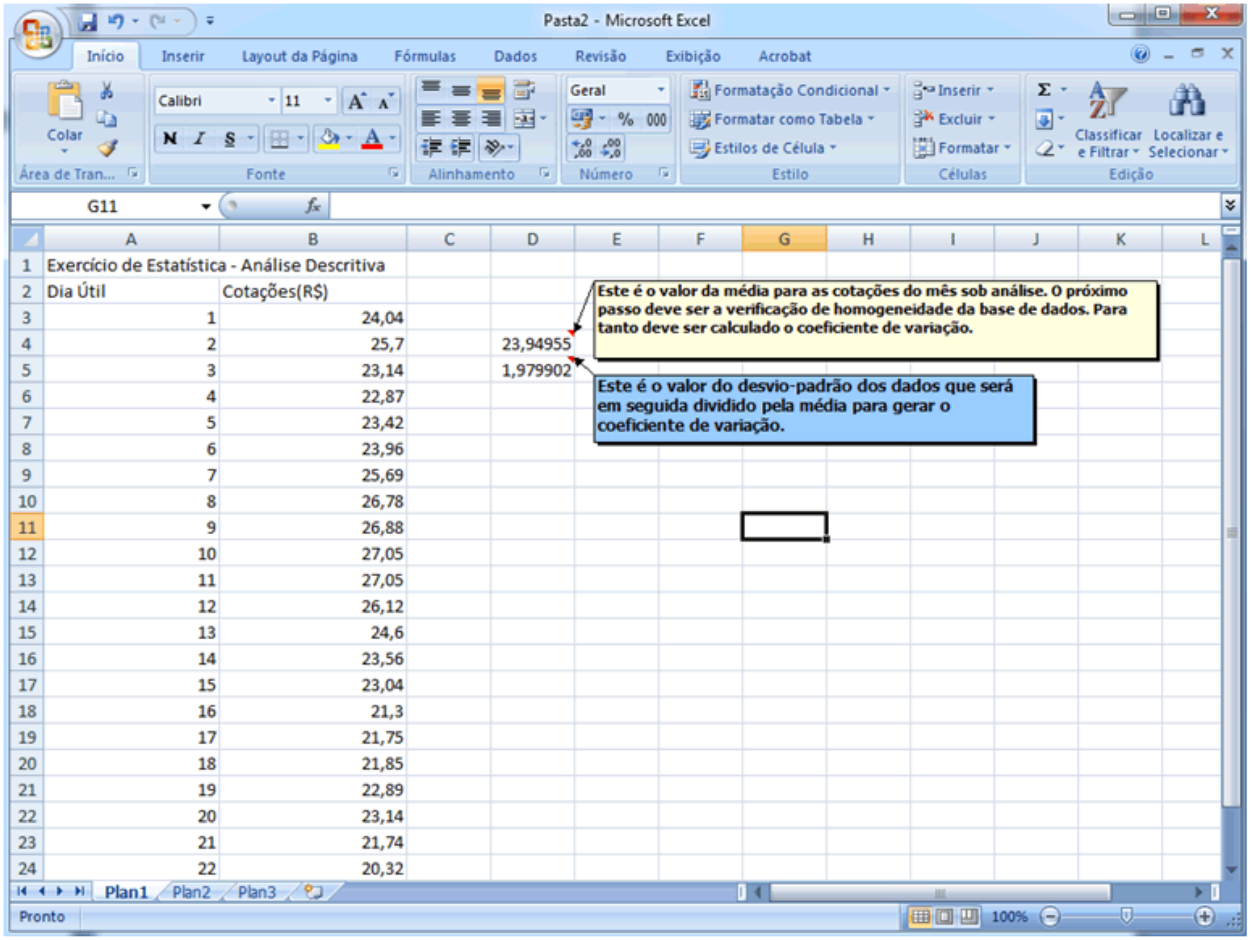
Em seguida, marcando as posições de B3 a B24, vem:



Mais uma vez é possível visualizar o resultado no canto inferior esquerdo da tela.

50

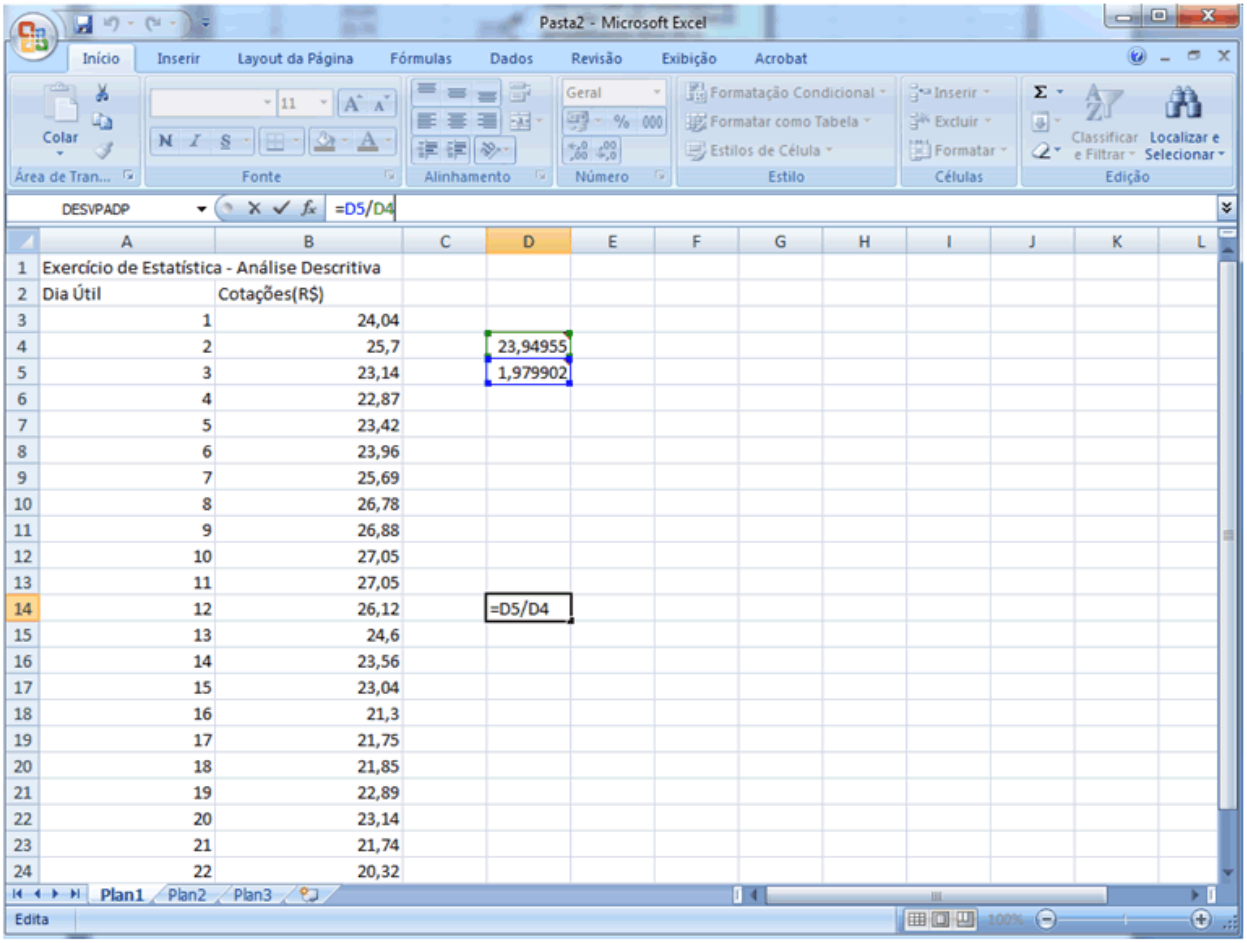
Clicando agora em OK e inserindo um comentário chega-se a:



Para que haja exibição dos comentários e não haja sobreposição, clicar em Exibir, depois em Comentários e com duplo clique na área de cada comentário é possível movê-lo de forma a permitir que todos sejam visíveis simultaneamente.

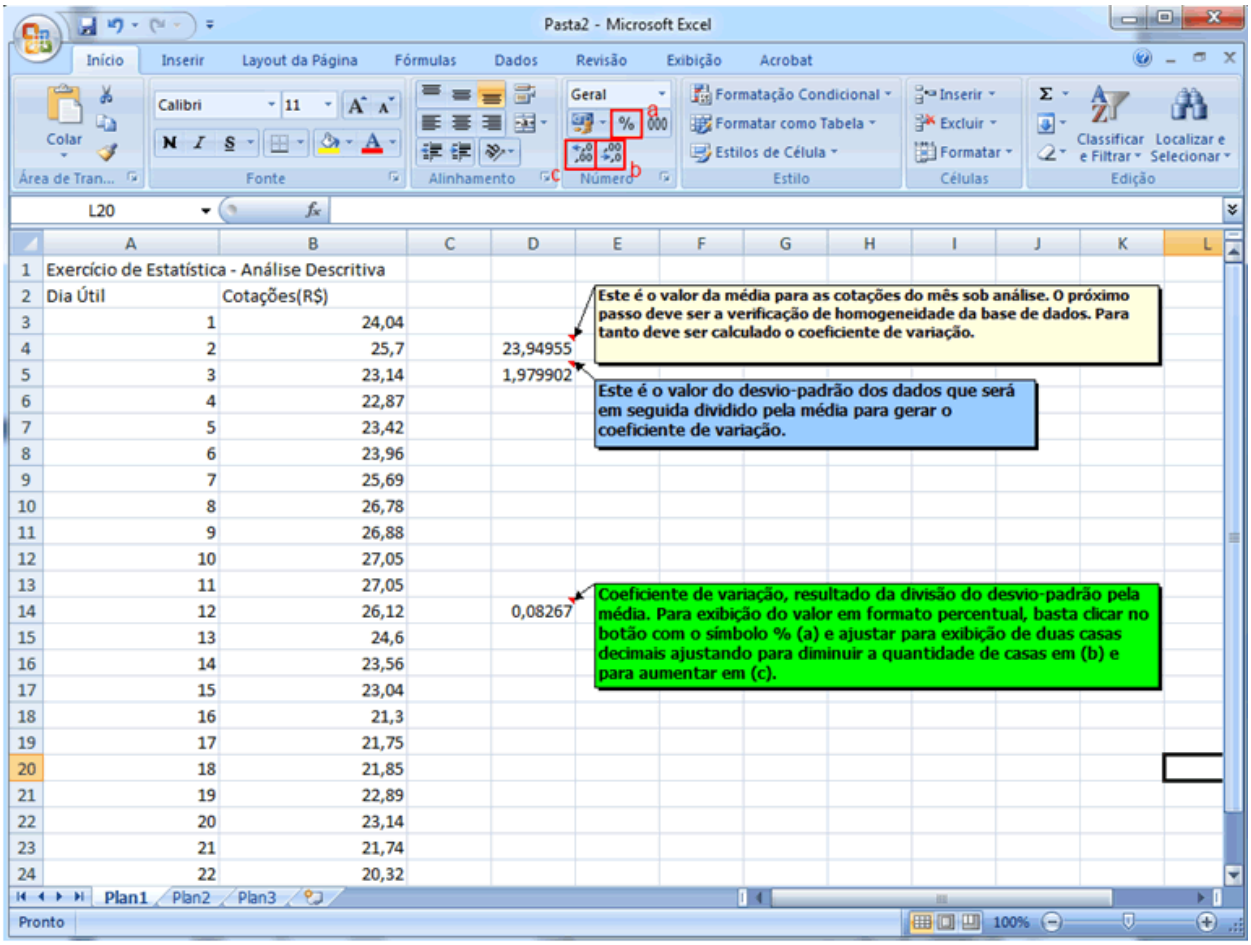
51

A determinação do coeficiente de variação deve ser feita dividindo-se o desvio-padrão pela média e isto será feito na célula D14 acompanhado de um comentário (como poderia ser em qualquer outra).



52

Deve-se lembrar que para iniciar a digitação de uma fórmula no Excel utiliza-se o sinal =, como mostrado na tela abaixo (também está sinalizado o ícone para aumento de casas decimais, mencionado no comentário):

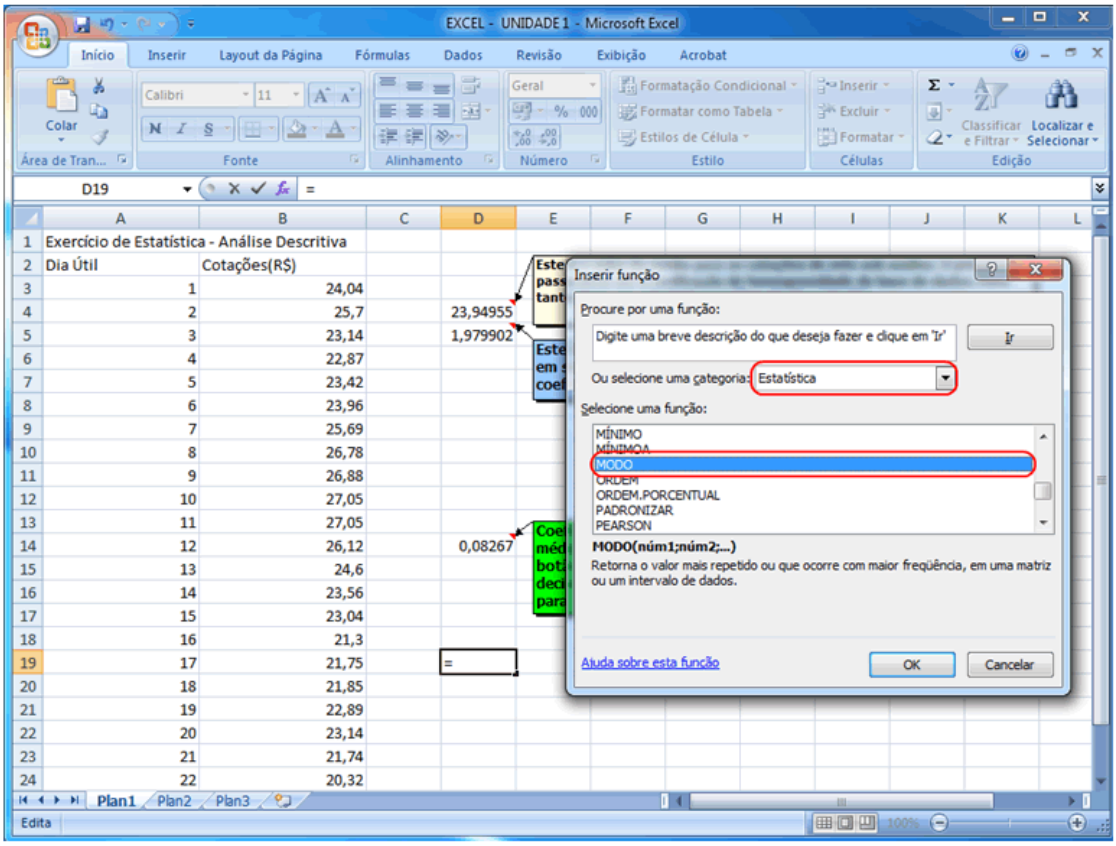


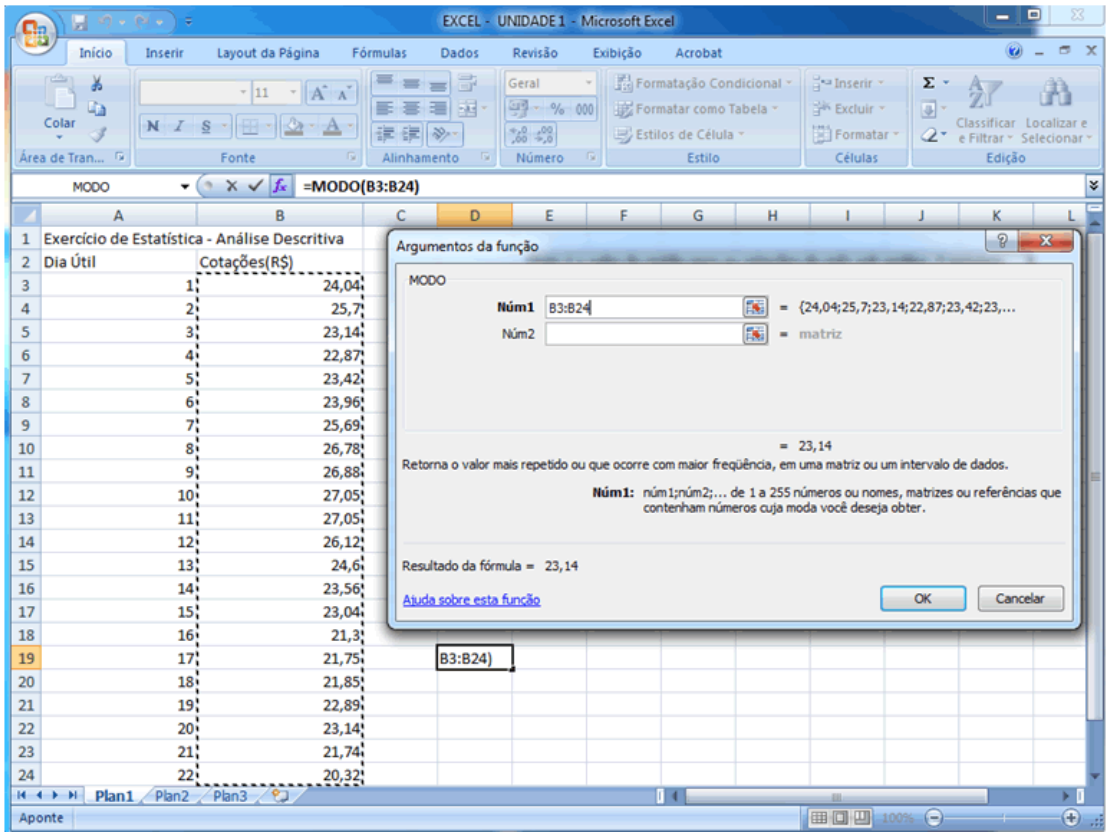
53

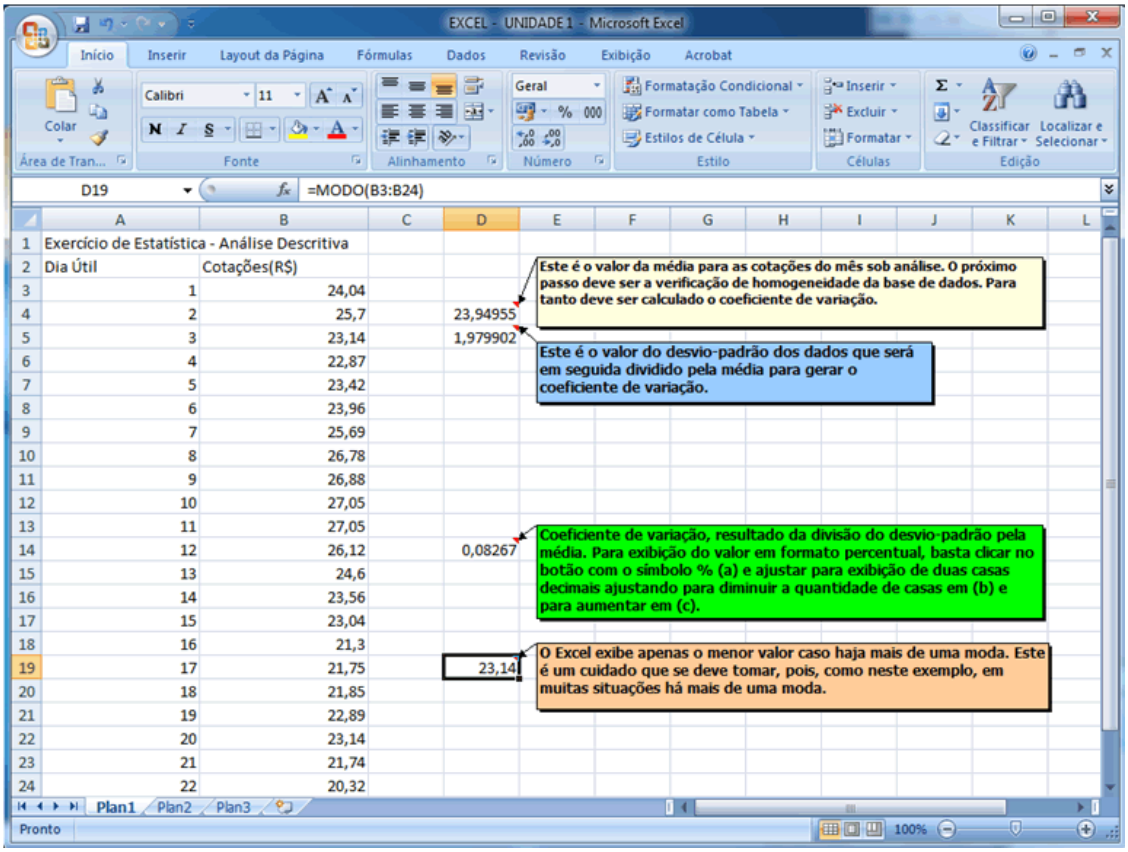
4 - Calculando a Moda e a Mediana

Partindo agora para a determinação da moda e da mediana, a sequência inicial permanece, qual seja: selecionar uma célula na qual se deseja que a medida estatística apareça e posicionar o cursor, clicar em colar função, Estatística, MODO (no caso da moda) ou MED (no caso da mediana), OK , marcar ou digitar a posição da base de dados no campo ao lado de Num1, OK. Vejamos como fica.

Para a moda (célula D19):

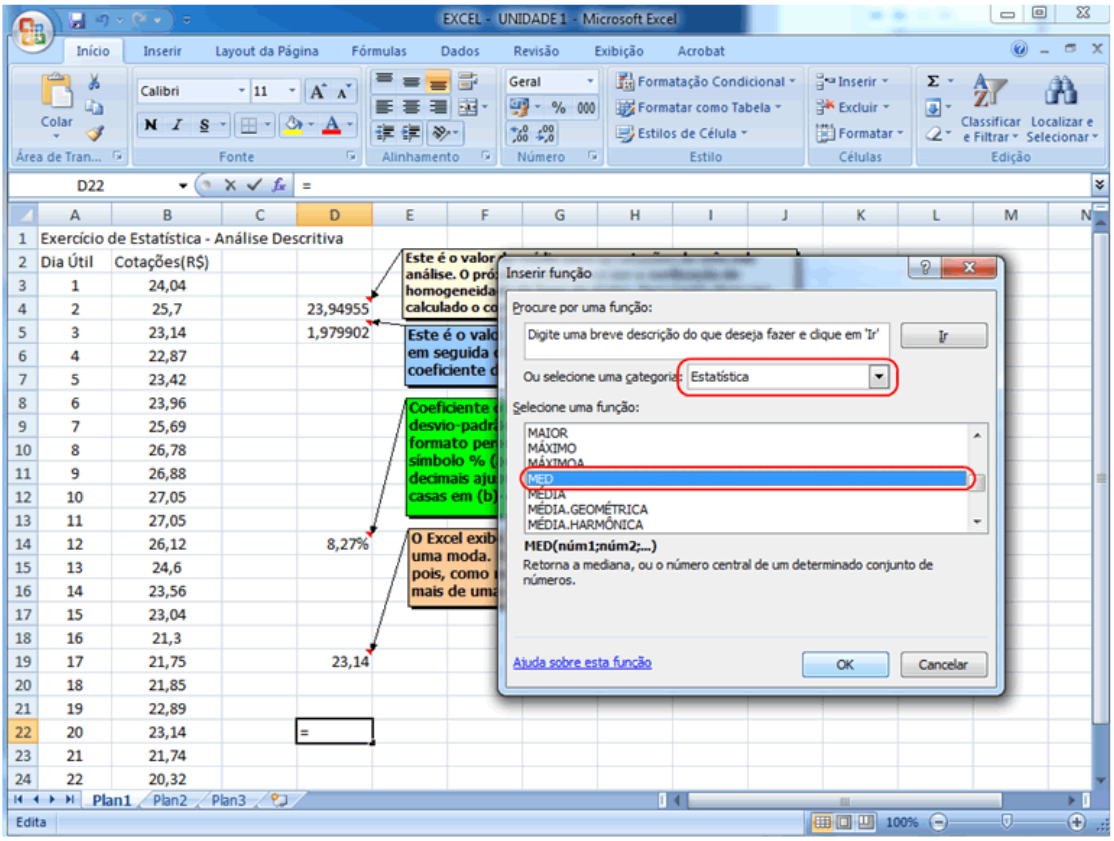


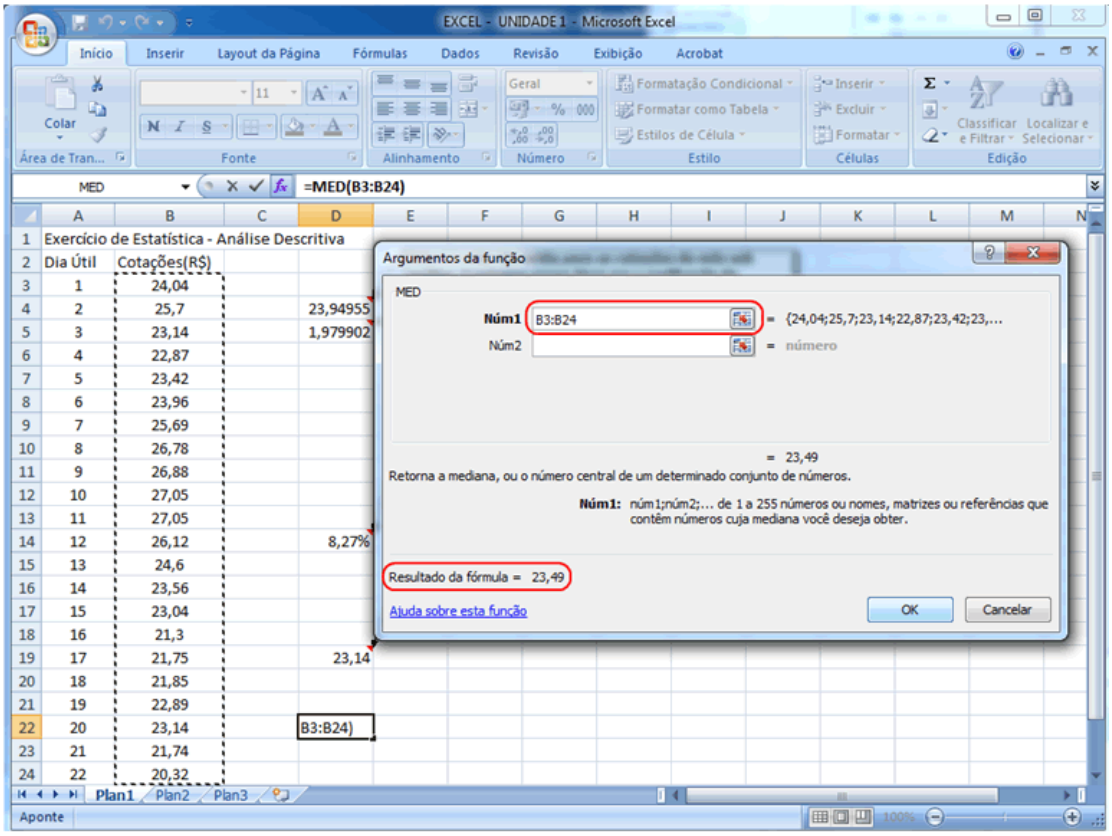


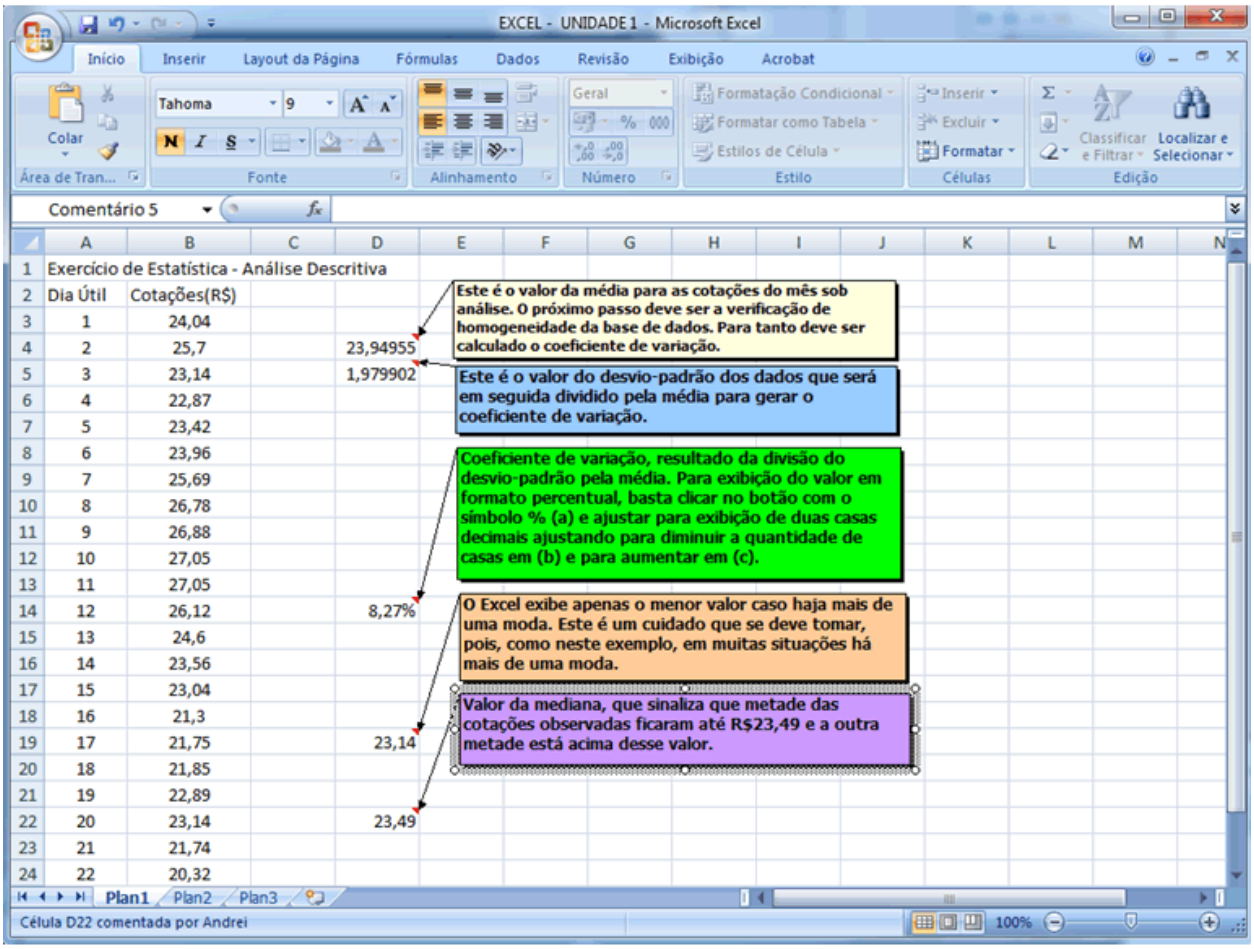


54

Para a Mediana (célula D22):





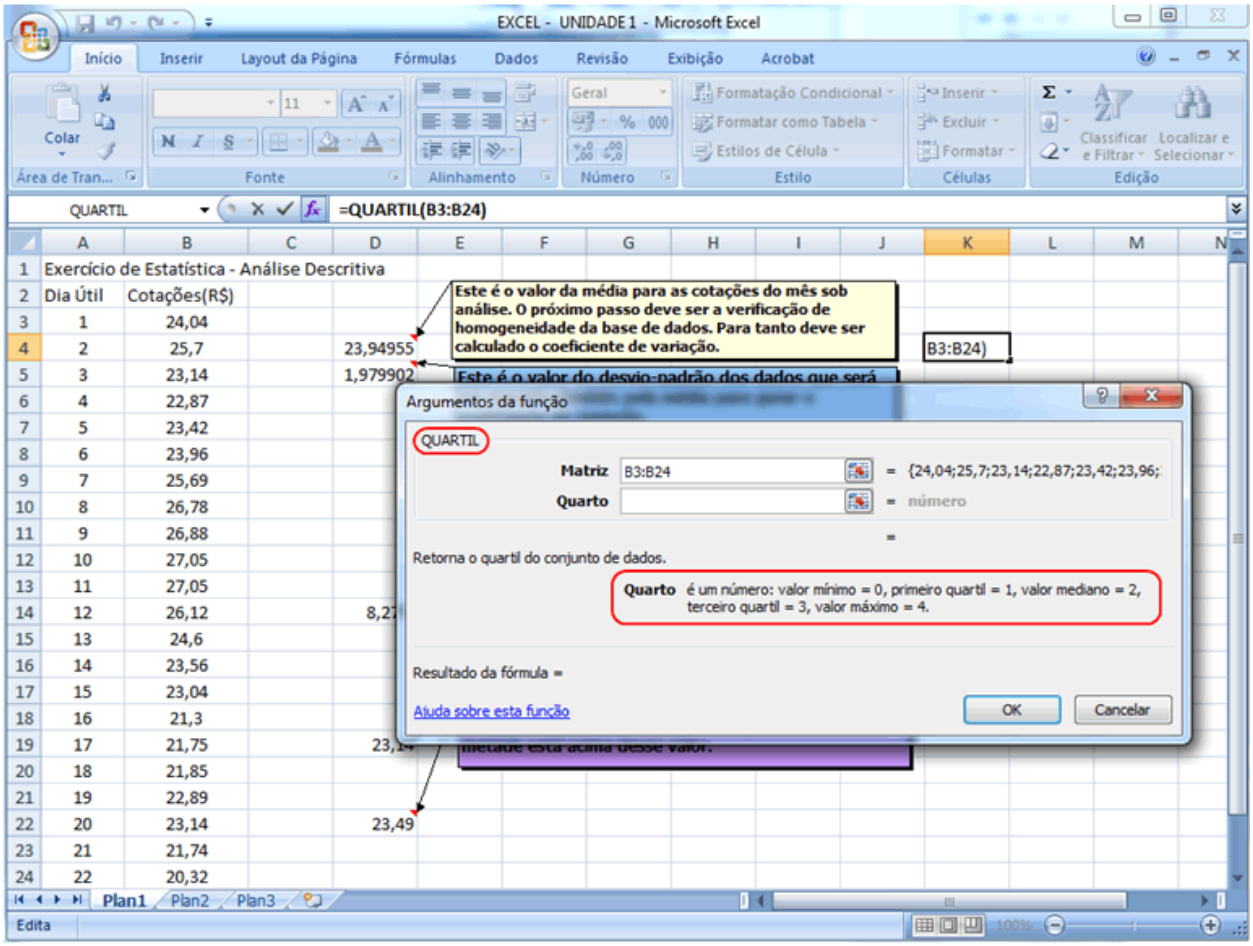


55

5 - Diagnosticando Pontos Discrepantes

Falta ainda determinar os quartis e os valores limite para diagnóstico de pontos discrepantes. A determinação dos quartis é análoga, mais uma vez, a tudo que estamos fazendo. Para **calcular os quartis**, deve-se observar a sequência apresentada adiante.

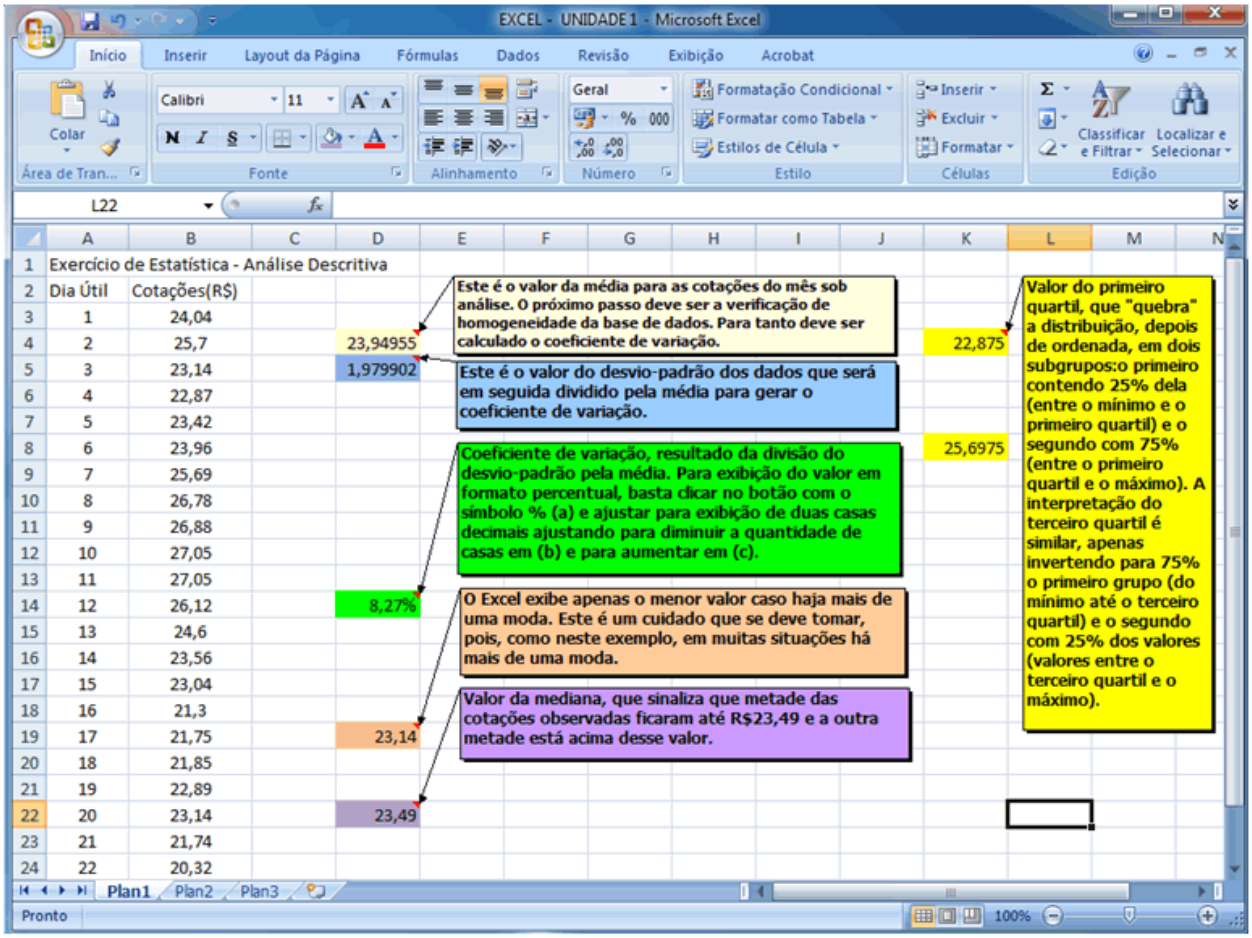
Selecionar uma célula na qual se deseja que a medida estatística apareça e posicionar o cursor, clicar em colar função, Estatística, QUARTIL, OK, e, então surgem dois campos para preenchimento - Matriz e Quarto, como mostrado a seguir.



No campo Matriz deve-se entrar (digitando ou marcando com o *mouse*) com as posições de B3 a B24, como também já foi feito para as outras medidas. No caso do campo Quarto, digita-se 1, se for o primeiro quartil e digita-se 3, se for o terceiro quartil, e clica-se em OK (foram selecionadas as células K4 para inserção do valor do primeiro quartil e K8 para a do terceiro).

56

Consequentemente teremos:

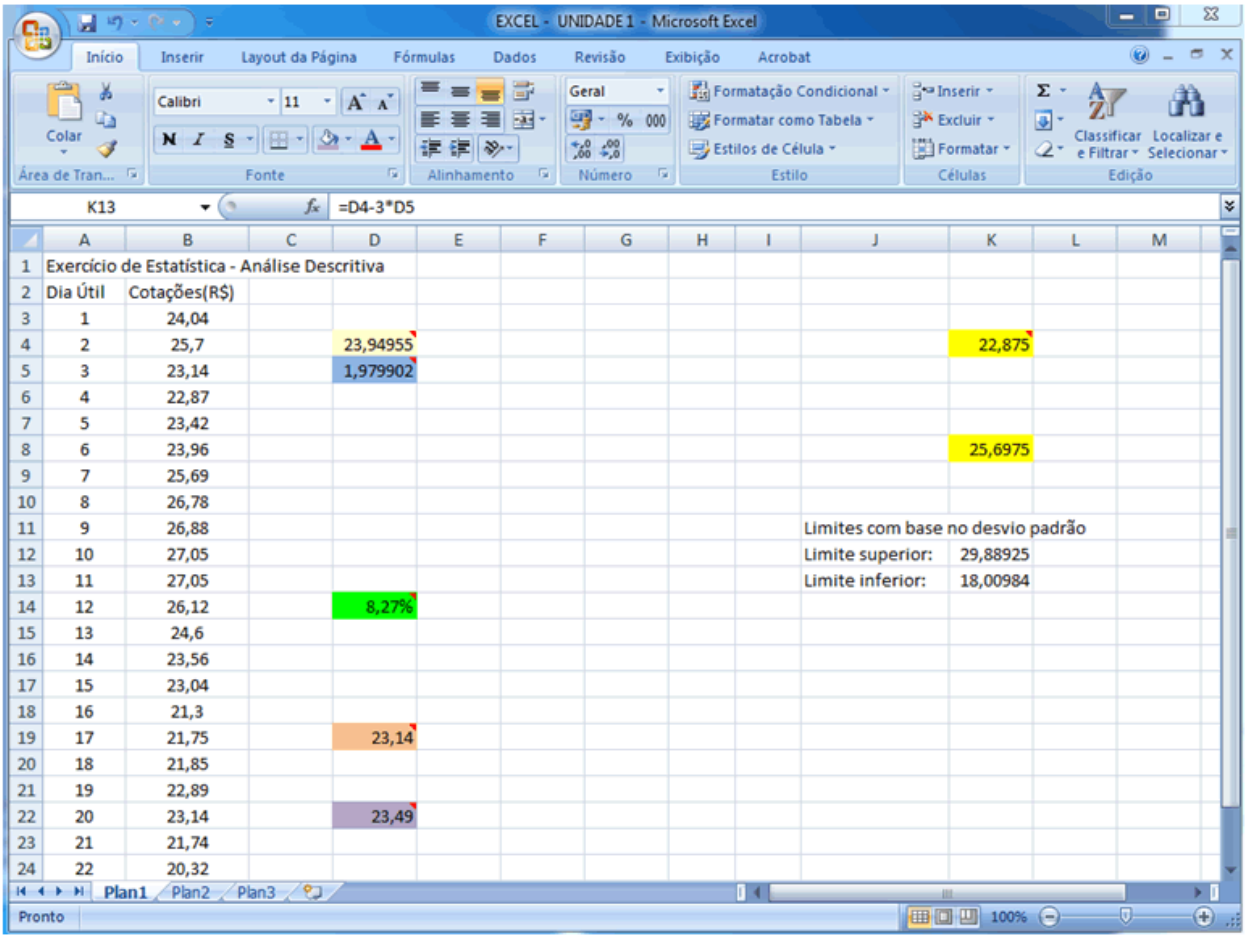


57

A determinação dos **limites máximo e mínimo** para verificação da existência de pontos discrepantes não é feita de forma "automática" pelo Excel, sendo necessário utilizar as fórmulas já indicadas anteriormente. A seguir serão apresentadas as duas estratégias de diagnóstico de pontos discrepantes.

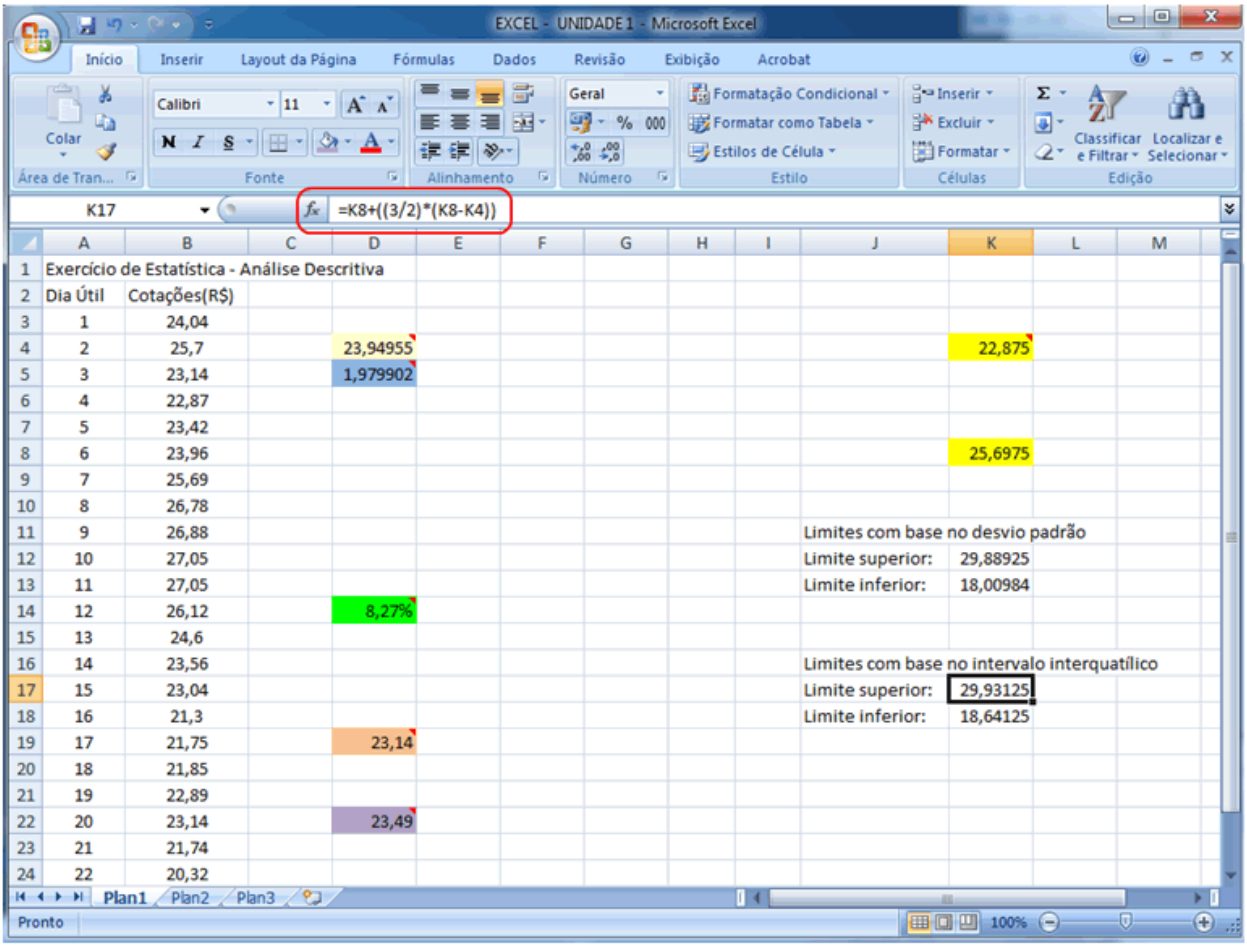
Primeiramente, considerando o triplo do desvio-padrão (células K12 e K13) e, em seguida, a abordagem de 3/2 do intervalo interquartílico(células K17 e K18).

Na primeira situação, depois de posicionar o cursor na célula selecionada, digitar =D4+3*D5 (para o limite máximo, observando que nas células D4 e D5 encontram-se a média e o desvio-padrão, respectivamente, e digitar =D4-3*D5 (para o limite mínimo).



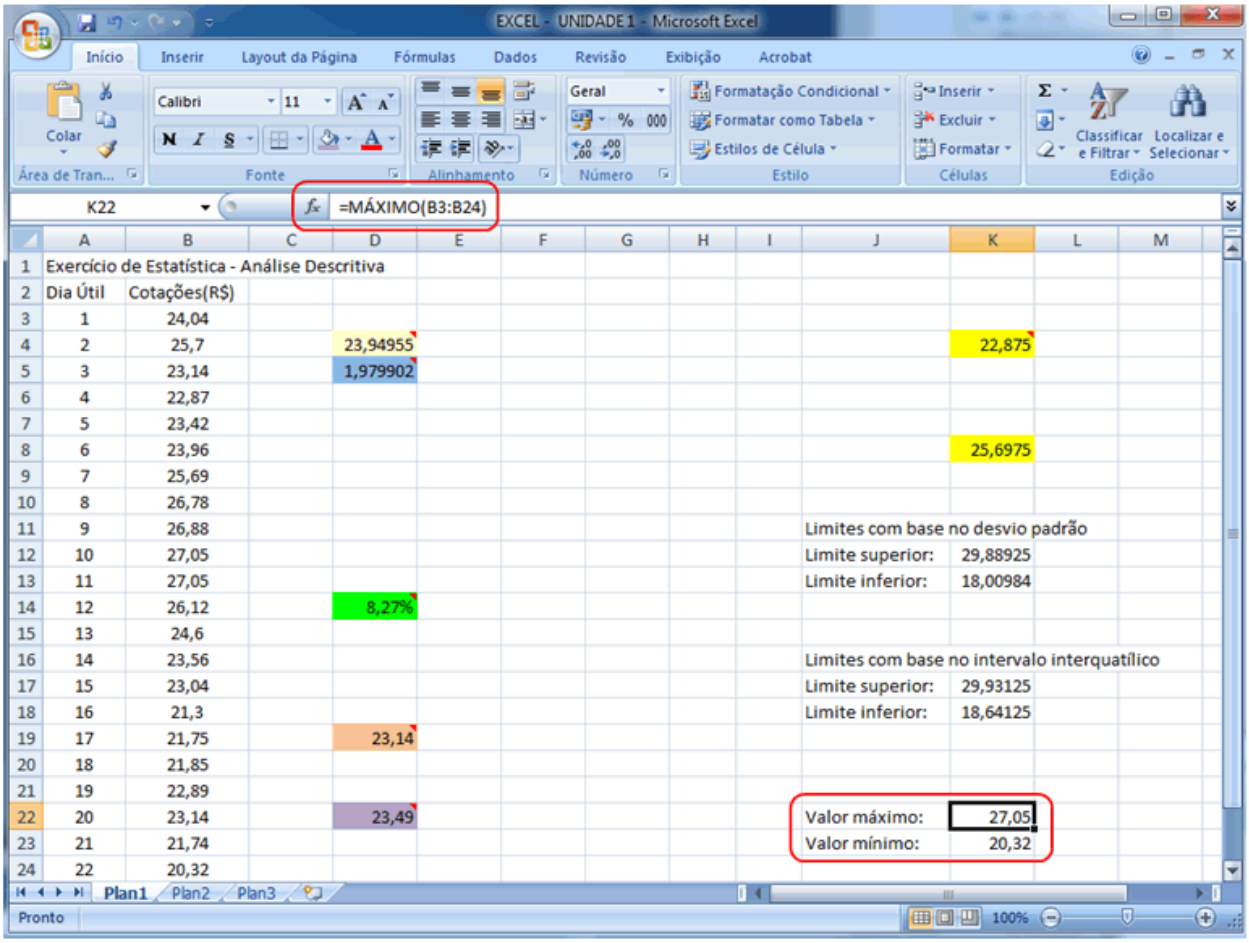
58

No segundo procedimento, digita-se na célula selecionada para o máximo (K17), =K8+((3/2)*(K8-K4)) e na célula selecionada para o mínimo (K18), =K4-((3/2)*(K8-K4)), lembrando que em K8 está o terceiro quartil e em K4 está o primeiro quartil.



59

Por último utilizaremos o Excel para determinar os valores máximo e mínimo do conjunto de observações, o que é particularmente útil quando a base de dados é grande e não é possível visualizá-los com clareza. Isto é feito de forma análoga à média e desvio-padrão, devendo-se selecionar as células, posicionar o cursor, clicar em fx, Estatística, MÁXIMO ou MÍNIMO, OK, marcar ou digitar a posição da base de dados no campo ao lado de Num1 e OK. Veja na planilha abaixo os resultados nas células K22 e K23.



59

RESUMO

Neste módulo ilustrou-se o cálculo de várias medidas descritivas apresentadas anteriormente com o apoio da planilha eletrônica Microsoft Excel. Na maioria dos casos a sequência de passos é:

(a) abrir uma planilha e digitar os dados a serem analisados;

(b) selecionar a célula na qual se quer inserir a medida e lá posicionar o cursor;

(c) clicar no ícone fx - colar função (da barra de ferramentas) e então clicar em Estatística (na coluna à esquerda, na janela que se abre);

(d) clicar em qualquer uma das opções, abaixo, na coluna à direita, na janela que está aberta:

- MÉDIA, para calcular a média;
- DESVPADP, para calcular o desvio-padrão;
- MODO, para calcular a moda;

• MED, para calcular a mediana;
• QUARTIL, para calcular os quartis (1º, 2º e 3º, sendo que o segundo é a própria mediana);
• MÁXIMO, para determinar o maior valor observado;
• MÍNIMO, para determinar o menor valor observado.

(e) clicar em OK e digitar/marcar com o mouse, na próxima janela, o conjunto de células no qual está a base de dados, no caso de cálculo de média, desvio-padrão, moda, mediana ou determinação do valor máximo e do valor mínimo. No caso de cálculo do primeiro e terceiro quartis, deve-se, depois deste passo, digitar 1 ou 3 ao lado da opção QUARTO na mesma janela (caso esta opção fosse utilizada para cálculo da mediana, bastaria digitar 2 neste campo);

(f) clicar em OK para que o resultado seja lançado na célula inicialmente selecionada.

No caso do cálculo do coeficiente de variação e dos limites para diagnóstico de pontos discrepantes, não há solução "automática" no Excel, logo, é necessário digitar as fórmulas adequadas, sempre iniciadas pelo sinal =.

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.

1 - DISTRIBUIÇÃO NORMAL – GAUSS

A distribuição normal ou de Gauss é uma distribuição estatística de dados amostrais baseada na relação dos dados em análise com sua medida de tendência central (média) e com sua respectiva medida de dispersão (desvio-padrão).

Como exemplo, podemos citar os pesos e medidas esperados para crianças recém-nascidas no primeiro ano de vida. Perceba, pelo cartão de vacinação destas crianças, que um controle destas duas variáveis é realizado, a fim de se garantir o bom desenvolvimento das crianças.

Entretanto, a lógica utilizada, através de uma função matemática, na determinação dos limites máximo e mínimo aceitáveis para os pesos e altura destas crianças, é exatamente a da determinação da distribuição normal ou de Gauss. Nessa análise leva-se em conta um peso médio e uma variação para mais ou para menos (desvio-padrão), que serve para delimitar os limites máximo e mínimo aceitáveis nas variáveis peso e altura.

Devemos lembrar que a toda função matemática corresponde um gráfico, como, por exemplo, retas, associadas às funções do primeiro grau, e parábolas, associadas às funções do segundo grau (isto para não falar nas funções exponenciais, logarítmicas e trigonométricas). Assim, esta curva normal também é fruto de uma função que pode ser expressa como:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-(x-\overline{x})^2/2\sigma^2}$$

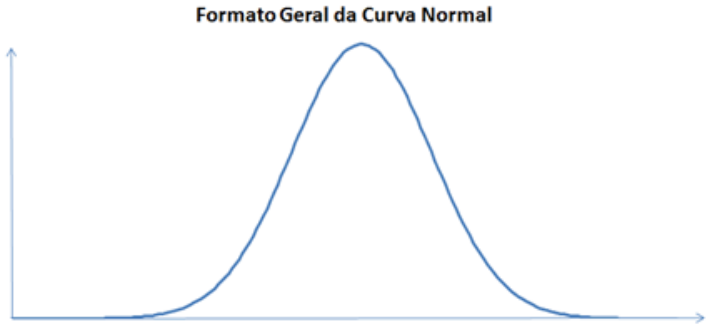
Onde:

Y = valor do eixo vertical correspondente a um dado valor da variável X, cuja média é dada por $\overline{X}$ e o desvio padrão é σ;

$\pi = 3,1416$

$e = 2,7183$

O aspecto gráfico, desta distribuição, de forma genérica, é o seguinte:



sendo que, no eixo horizontal, são representados os valores da variável que se está estudando e, no eixo vertical, valores associados às frequências relativas correspondentes aos valores específicos da variável sob análise.

A formulação apresentada, com certeza, apresenta uma forma bastante assustadora. Entretanto, felizmente não teremos necessidade de utilizá-la tal como se apresenta. O que a fórmula demonstra é que qualquer distribuição normal é determinada por dois parâmetros: a **média** e o **desvio-padrão** dos dados. Saiba +

Saiba +

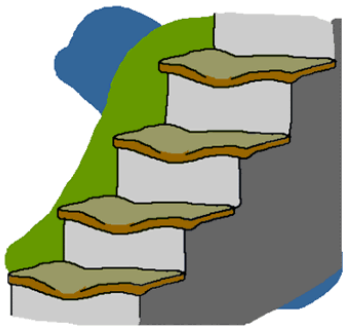
Jack Levin, em seu livro Estatística Aplicada a Ciências Humanas, considera a distribuição normal como "um modelo teórico ou ideal que resulta muito mais de uma equação matemática do que de um real delineamento de pesquisa com posterior coleta de dados. Entretanto, a utilidade da curva normal para o

pesquisador pode ser evidenciada por meio de suas aplicações a efetivas situações de pesquisa". Escreve, ainda, Jack Levin: "a curva normal pode ser usada na descrição de distribuições de escores, na interpretação do desvio padrão e em afirmações relacionadas com a noção de probabilidade" e "a curva normal constitui um ingrediente essencial para a tomada de decisões estatísticas, a partir da qual o pesquisador pode generalizar para populações as conclusões a que tenha chegado ao lidar com amostras."

63

Se atentássemos para as características físicas dos seres humanos, estatura, por exemplo, veríamos que a maioria dos adultos estaria na faixa que vai de 152 cm (aproximadamente) até 183 cm (aproximadamente), com muito pouca gente menor que 152 cm ou maior que 183 cm. O QI também seria previsível - a maioria dos QIs situando-se entre 90 e 110, havendo uma 'descida' gradual dos escores para ambas as caudas, com pouquíssimos 'gênios' que têm QI superior a 140 e, da mesma forma, pouquíssimas criaturas menos privilegiadas, cujos QIs estão abaixo de 60. Por igual raciocínio, relativamente poucos sujeitos poderiam ser considerados políticos extremistas - de direita ou de esquerda - enquanto a tendência política da maioria seria considerada moderada.

Finalmente, mesmo o desgaste dos pisos, resultante do fluxo de transeuntes, lembra a distribuição normal: a maior parte do desgaste ocorre no centro dos pisos (degraus etc.), enquanto nos lados, à medida que nos afastamos do centro, o desgaste vai-se tornando cada vez menor.



64

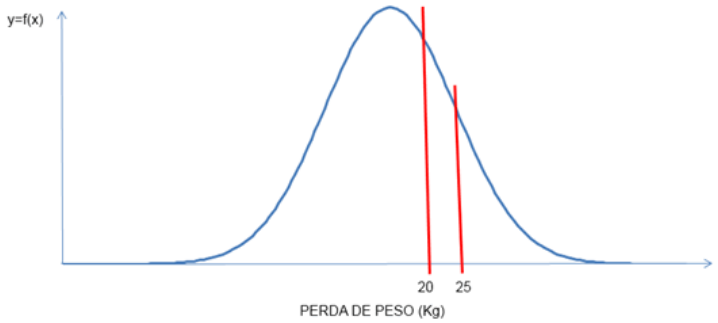
Observamos que o mundo hipotético da curva normal não difere de forma radical do mundo 'real' (que vivemos no momento). Fenômenos tais como estatura, QI, orientação política, desgaste dos pisos etc. aproximam-se, na prática, até que muito bem da distribuição normal teórica. Pelo fato de tantos fenômenos terem essa característica - isto é, pelo fato de ela ocorrer tão frequentemente na natureza (e por outras razões que logo se tornarão aparentes) - pesquisadores de diferentes campos têm feito uso extensivo da curva normal, aplicando-a aos dados que eles coletam e analisam.

Observe-se, porém, que alguns fenômenos no campo social - como em qualquer outro - simplesmente não se ajustam à noção teórica da distribuição normal. Muitas distribuições são assimétricas; outras têm mais de uma moda; outras são simétricas, mas não têm a forma de 'sino'.

Como exemplo concreto, consideremos a distribuição de riqueza no mundo. É fato bem conhecido que 'os que têm' superam de longe 'os que não têm'. Assim, a distribuição de riquezas (indicada pela renda per capita) é de extrema assimetria (pelo menos na aparência), de sorte que apenas uma pequena proporção da população mundial recebe porção significativa da renda total.

65

De forma análoga, especialistas em demografia dizem-nos que os Estados Unidos da América do Norte tornaram-se, nos últimos tempos, uma terra de jovens e velhos. Do ponto de vista econômico, essa distribuição de idades representa um fardo pesado para um grupo relativamente pequeno de trabalhadores, isto porque, sendo todos cidadãos de meia-idade, têm a seu encargo um número assustador tanto de velhos (aposentados) quanto de jovens (ainda em período escolar).



Nestas circunstâncias em que temos boas razões para esperar grandes divergências da normalidade - como, por exemplo, no caso da idade e da renda - a curva normal não pode ser usada como 'modelo' para os dados coletados. Vemos, assim, que não é possível aplicá-la com liberdade a todas as distribuições que o pesquisador obtém, e deve, ao contrário, ser usada com uma boa dose de bom senso. Felizmente os estatísticos sabem que grande quantidade de fenômenos de interesse segue o modelo normal."

Já deve ser do seu conhecimento, mesmo que apenas de ouvir falar, que vários sistemas previdenciários do mundo estão em crise ou na iminência de, se o cenário não mudar no curto e médio prazos.

66

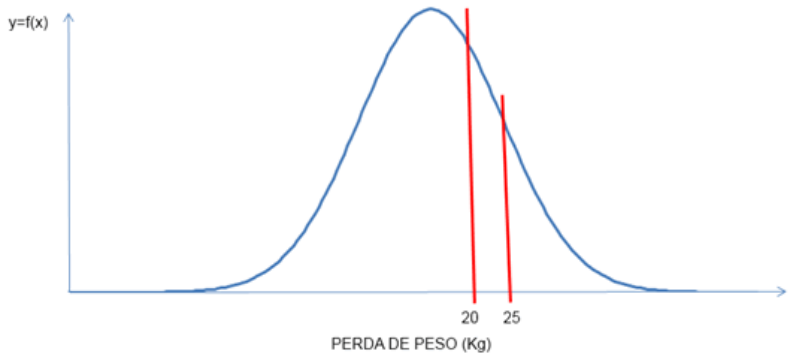
2 - Padronização da Distribuição

Os fenômenos sociais, psicológicos e físicos são exemplos de fenômenos que se apresentam de forma normalmente distribuídos. Temos uma medida padrão para estas variáveis e alguma dispersão em torno desta medida. Assim, em função desta dispersão em torno da média, podemos destacar uma propriedade para este tipo de distribuição: se a base de dados estudada tem distribuição normal, pode-se garantir que:

a) 68,26% dos dados estarão compreendidos no intervalo delimitado pela média mais ou menos um desvio-padrão;
b) 95,44% dos dados estarão compreendidos no intervalo delimitado pela média mais ou menos dois desvios;
c) (c) 99,74% dos dados estarão compreendidos no intervalo delimitado pela média mais ou menos três desvios.

A determinação do percentual de dados compreendidos entre determinados valores estipulados para a variável sob estudo pressupõe que haja uma padronização destes valores (vinculados à "quantidade" de desvios-padrão que cada um representa de distância relativa à média).

Sendo mais claro: considere que uma nova dieta está sendo testada em um SPA/clínica de emagrecimento. Depois de algum tempo de avaliação deste novo tratamento, observa-se que a distribuição da perda de peso (em kg) assemelha-se bastante a uma distribuição normal. Sabe-se que a perda média de peso foi de 20 kg no período estipulado para o tratamento, com desvio-padrão de 3,5 kg. Pode ser importante para a administração de algum órgão governamental de fiscalização ter uma ideia do percentual de pacientes que perderam acima de 25 kg. Tomando nossa curva, pode-se ver que:



e que o percentual desejado corresponde à área à direita do ponto 25, assinalado no eixo horizontal, lembrando que a área total da figura corresponde a 100% dos dados (ou seja = 1).

67

Existe um procedimento matemático formal que permite o cálculo de áreas de figuras como aquela ilustrada anteriormente, delimitada pelos valores 20 e 25. No entanto, não é uma alternativa das mais simples. De maneira alternativa, pode-se pensar em um procedimento quase "automático", aplicável a todas as situações sob análise. Isto passa pela padronização dos valores/dados que compõem a base de dados.

O que se faz é subtrair a média de todos os dados e dividir estes resultados pelo desvio-padrão, gerando-se, então, uma base de dados padronizados (também conhecidos por escores padronizados ou z).

No exemplo,

$Z_1 = \frac{25-20}{3,5} = \frac{5}{3,5} \cong 1,43$ ou seja, equivale a 1,43 desvios - padrão acima da média, ou ainda, em outras palavras, o valor original (=25) fica 1,43 desvios - padrão à direita da média.

O interesse recai sobre o percentual de pessoas cuja perda de peso superou este valor padronizado de 1,43. Este procedimento para padronização dos dados viabiliza uma leitura única, independente da situação sob análise. Com isto, foi possível a construção de uma tabela da Distribuição Normal Padrão, que é apresentada na grande maioria dos livros de Estatística Básica (senão em todos).

Fundamentalmente, o que esta tabela retrata é o percentual de dados (área da figura) entre o valor padronizado e o valor médio padronizado, que sempre será zero, afinal

$$\frac{\text{Média - Média}}{\text{Desviopadrão}} = \frac{\textit{Zero}}{\textit{desvio}} = \text{Zero}$$

68

A tabela revela que para z = 1,43 esta porcentagem é de 0,4236, o que significa dizer que 42,36% dos clientes da clínica perderam entre 20 e 25 quilos (a média e o valor de referência, respectivamente). Só que a pergunta ainda não foi respondida, pois o que se quer é a porcentagem acima de 25. Assim, como nossa curva é simétrica, a média a divide em duas áreas iguais, cada uma correspondendo a 50% do total (mesmo porque em uma curva simétrica a média e a mediana são iguais, portanto basta aplicar o conceito de mediana para isto ficar bem claro). Assim, como 50% é o percentual acima da média, e 42,36% é o percentual entre a média e 25 kg, o resultado desejado é:

50% - 42,36% = 7,64%.

É importante frisar que, quando os dados de uma determinada distribuição normal são padronizados, a "nova" base de dados continua com distribuição normal, porém com média zero e desvio-padrão igual a um. O fato da média ser zero já foi mostrado e, certamente, se você calcular o desvio-padrão deste conjunto de dados "transformados", obterá 1 (um) como resultado.
Mais um exemplo:

Consultando uma tabela, por exemplo, chega-se a:
Percentual entre 0 e 3,03 = 0,4988, logo o percentual acima de 3,03 é 0,5 - 0,4988 = 0,0012 = 0,12%;

Percentual entre 0 e 2,78 = percentual entre -2,78 e 0 (pela simetria da distribuição) = 0,4973, logo o percentual abaixo de -2,78 é 0,5 - 0,4973 = 0,0027 = 0,27%.
Concluindo, a porcentagem de substratos fora dos limites estabelecidos é de 0,0012 + 0,0027 = 0,0039, ou seja, 0,39%, o que parece indicar que o processo de fabricação está funcionando bem, com uma "perda", resultante de componentes fora da especificação, muito baixa.

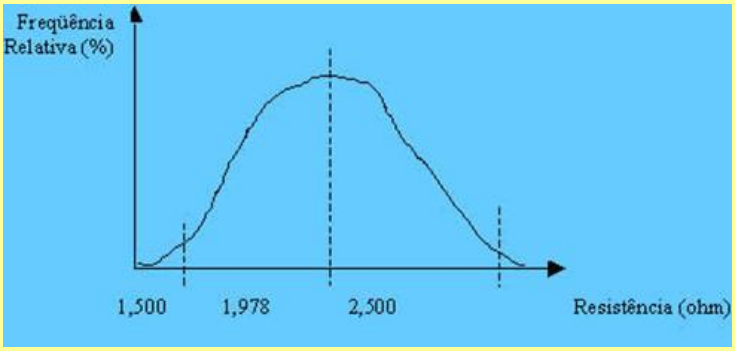
Exemplo

No livro Introdução à Estatística, de Mario Triola, em seu capítulo 5, encontramos a seguinte situação: "Um subfornecedor da IBM foi contratado para fabricar substratos de cerâmica, utilizados para transmitir sinais entre chips de silício para computador. As especificações exigem uma resistência entre 1,500 ohm e 2,500 ohms, mas a população tem resistências distribuídas normalmente com média de 1,978 ohm e desvio-padrão de 0,172 ohm. Que percentagem dos substratos de cerâmica foge às especificações do fabricante? Esse processo de fabricação parece estar funcionando bem?"
Por analogia com a situação tratada na etapa anterior, é muito importante padronizar os valores estabelecidos como referência.

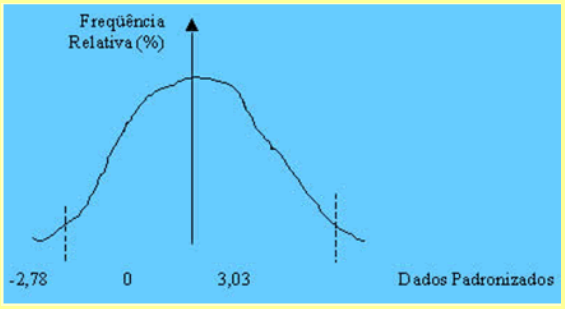
$Z_1 = \frac{1,500 - 1,978}{0,172} = \frac{-0,478}{0,172} = -2,78$

$Z_2 = \frac{2,500 - 1,978}{0,172} = \frac{-0,522}{0,172} = 3,03$

Assim,



O que após a padronização:



A porcentagem desejada é correspondente à área fora dos limites estabelecidos pelos valores padronizados -2,78 e 3,03, isto é, deseja-se saber a área à esquerda de -2,78 somada àquela à direita do 3,03.

Tabela

TABELA A-2 Distribuição Normal Padronizada (z)

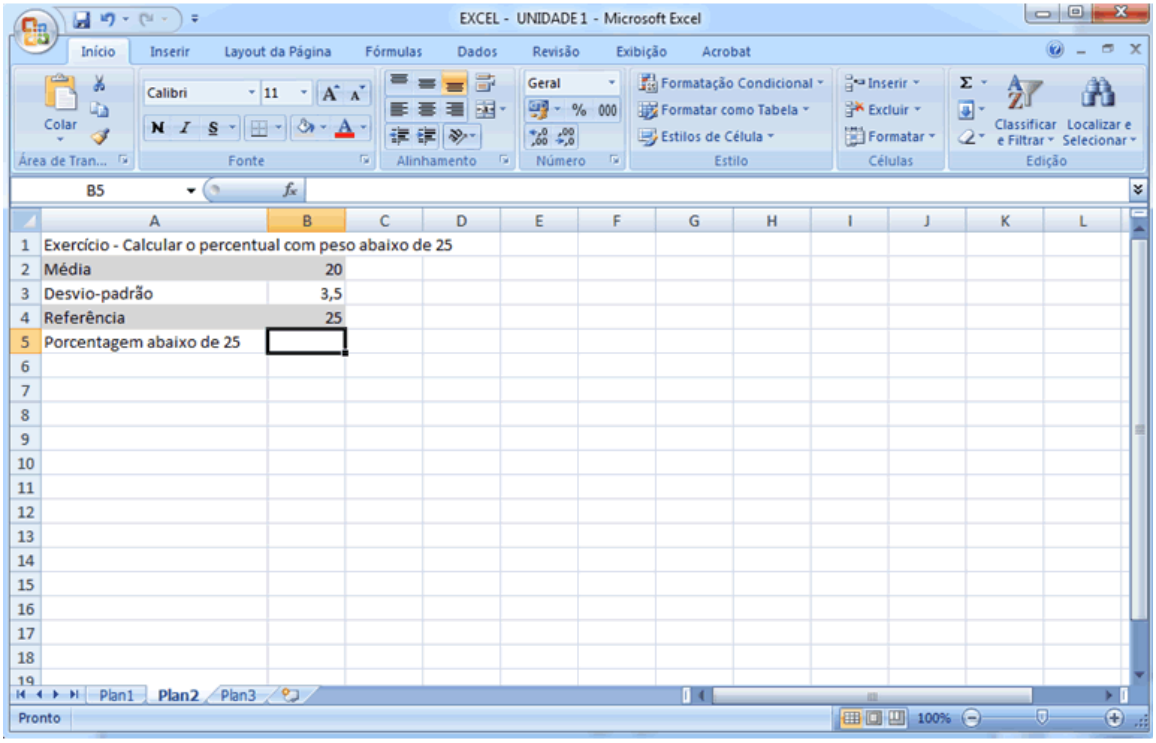
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,10 ou mais	0,4999									

69

3 - USANDO O EXCEL

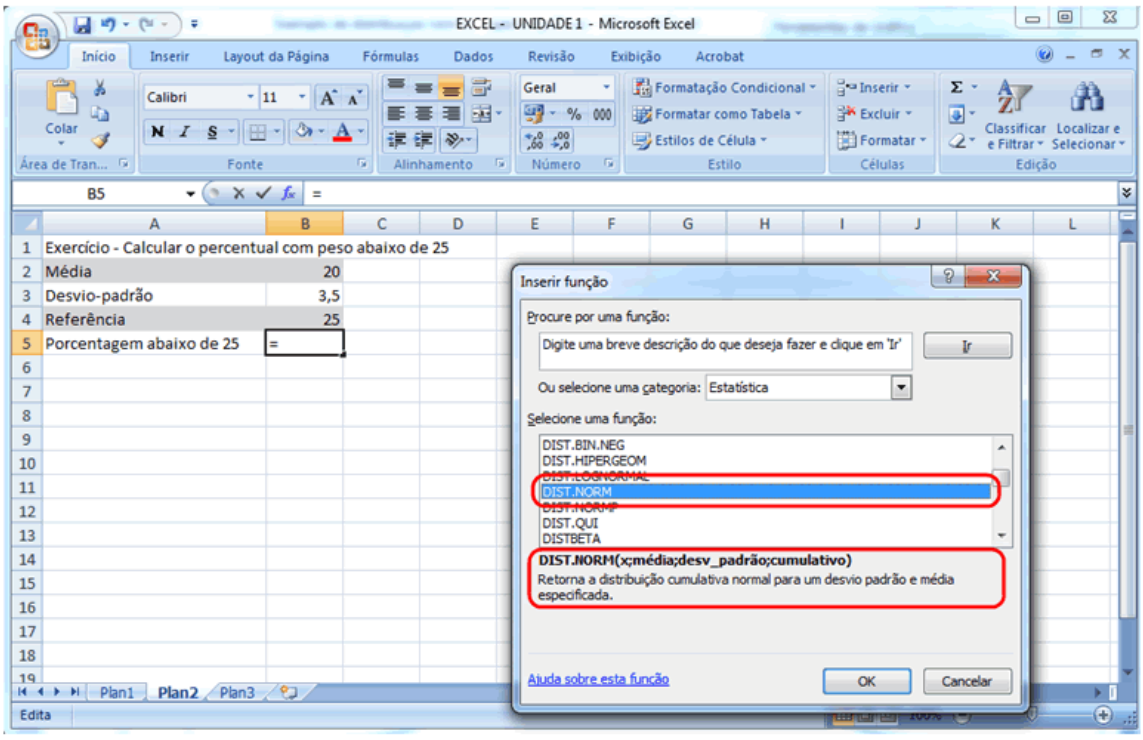
A utilização da planilha Microsoft Excel, nos dois casos trabalhados neste módulo, também permitiria obtenção dos resultados solicitados.

No caso do primeiro exercício, é interesse o percentual de casos acima de 25 kg em uma distribuição normal cuja média é 20 kg e desvio-padrão 3,5 kg. Abrindo-se uma planilha Excel, pode-se digitar estas referências para, em seguida, utilizar a ferramenta adequada.



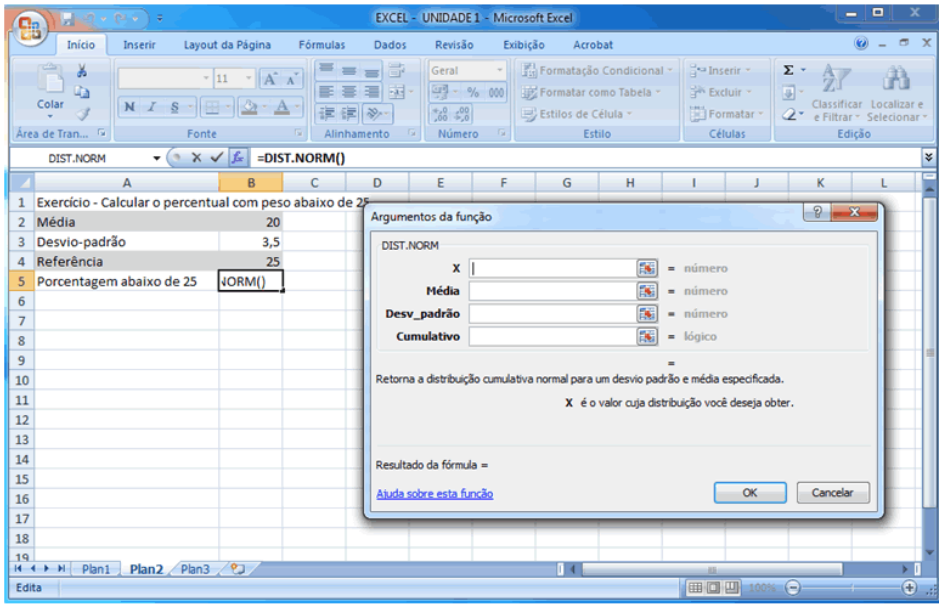
70

Posicionando o cursor na posição na qual se quer o percentual que fica abaixo do valor de referência (no caso 25 kg), clica-se no ícone fx, em seguida Estatística, em seguida DIST.NORM e OK.



71

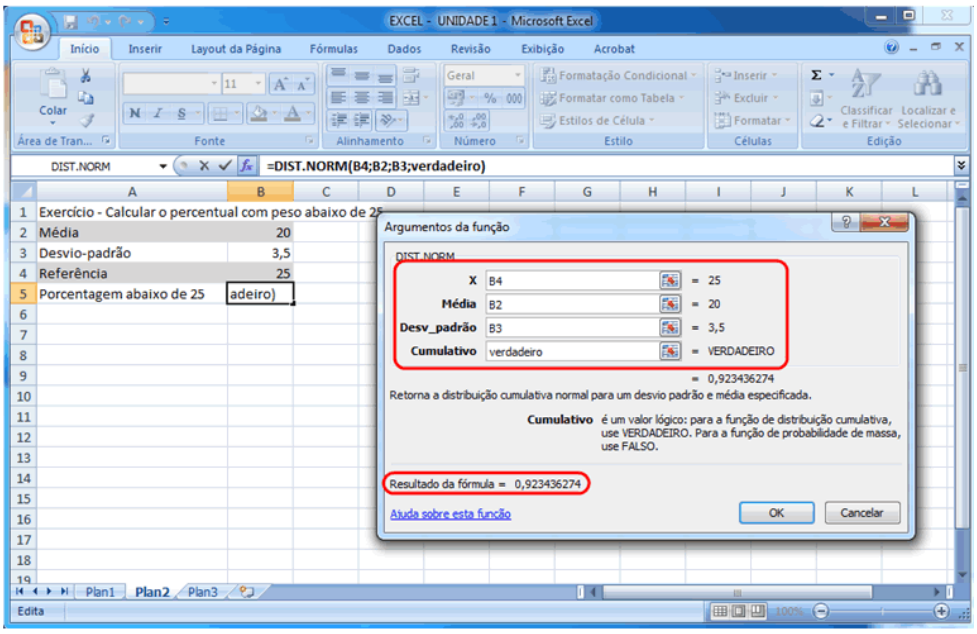
Na janela aberta, há quatro campos para que sejam inseridos os valores (ou respectivas células) para geração da proporção de valores abaixo da referência a ser digitada no campo ao lado da letra X.



72

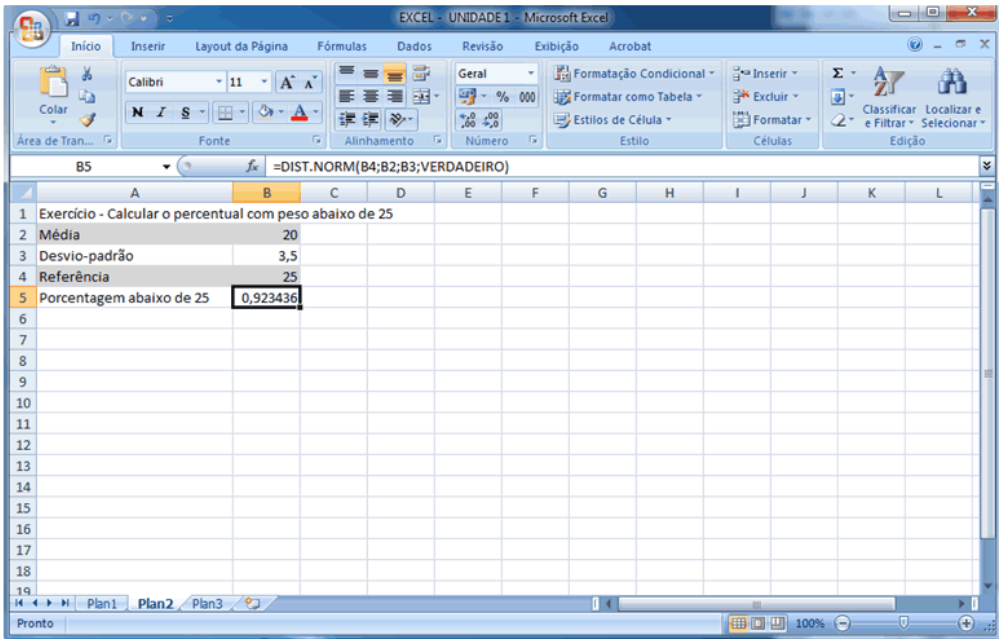
Observe que, no último campo, digitou-se VERDADEIRO, para obtenção da porcentagem de dados abaixo do valor de X.

Clica-se agora em OK.



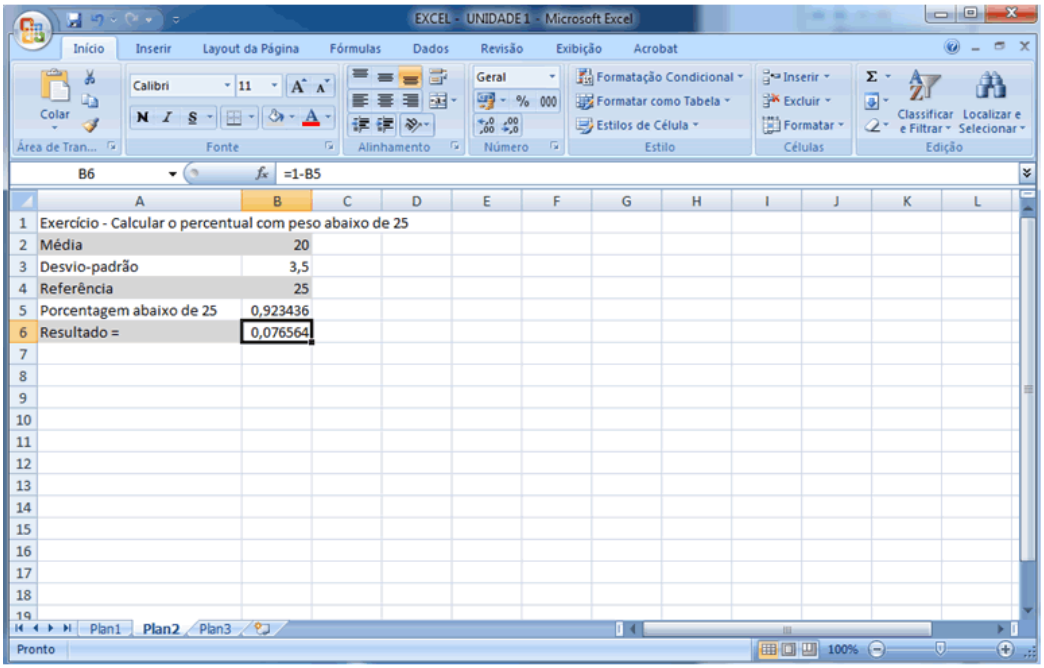
73

O valor de 92,34% não é ainda a resposta solicitada, mas, fazendo 100% - 92,34% (ou 1 - 0,9234), chegar-se-á à solução.

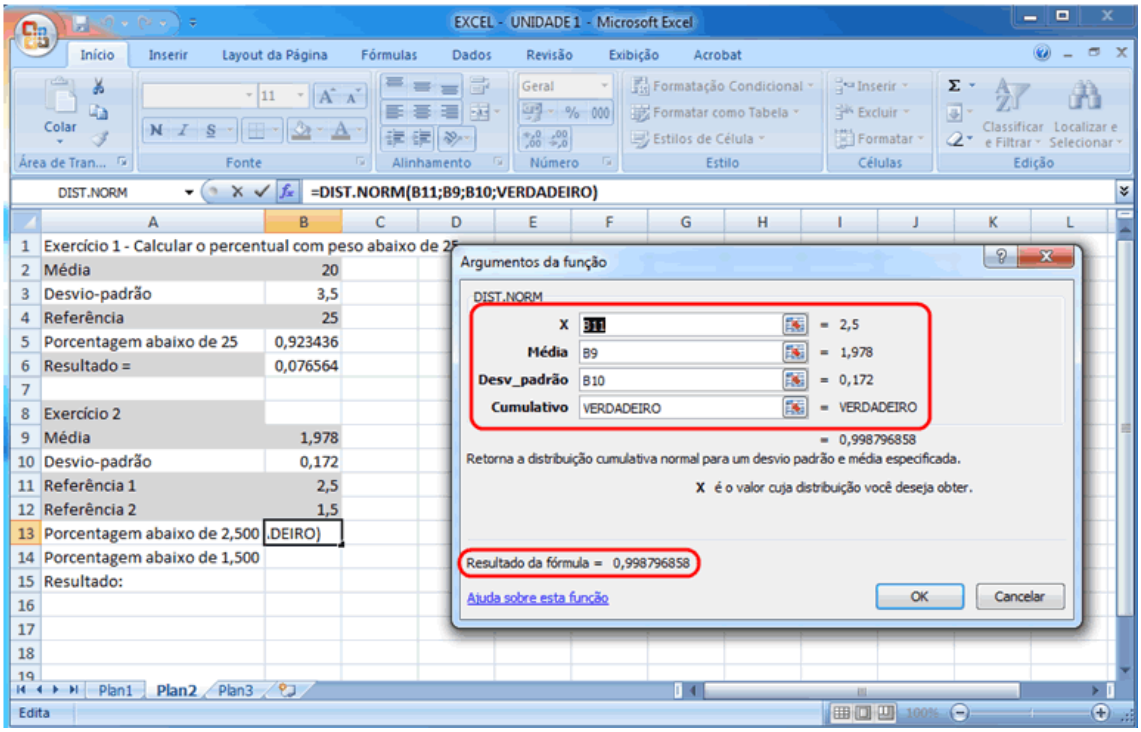


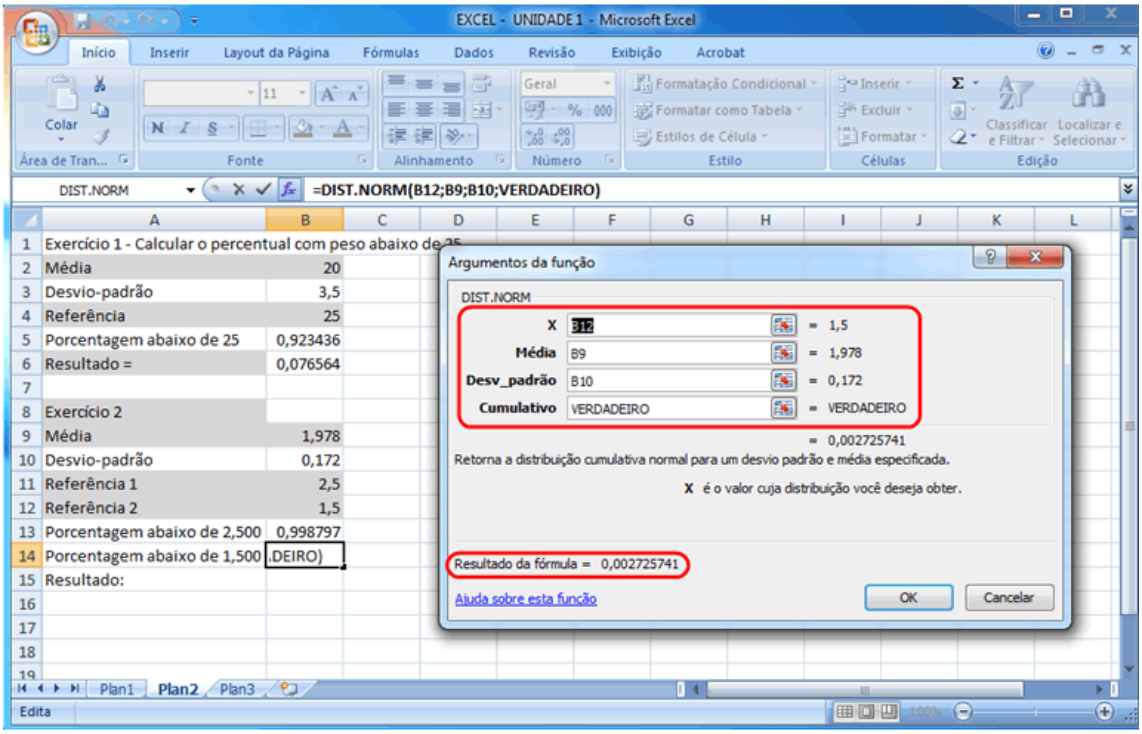
74

O resultado obtido na planilha abaixo é praticamente o mesmo daquele obtido sem a sua utilização com uma variação quase insignificante.



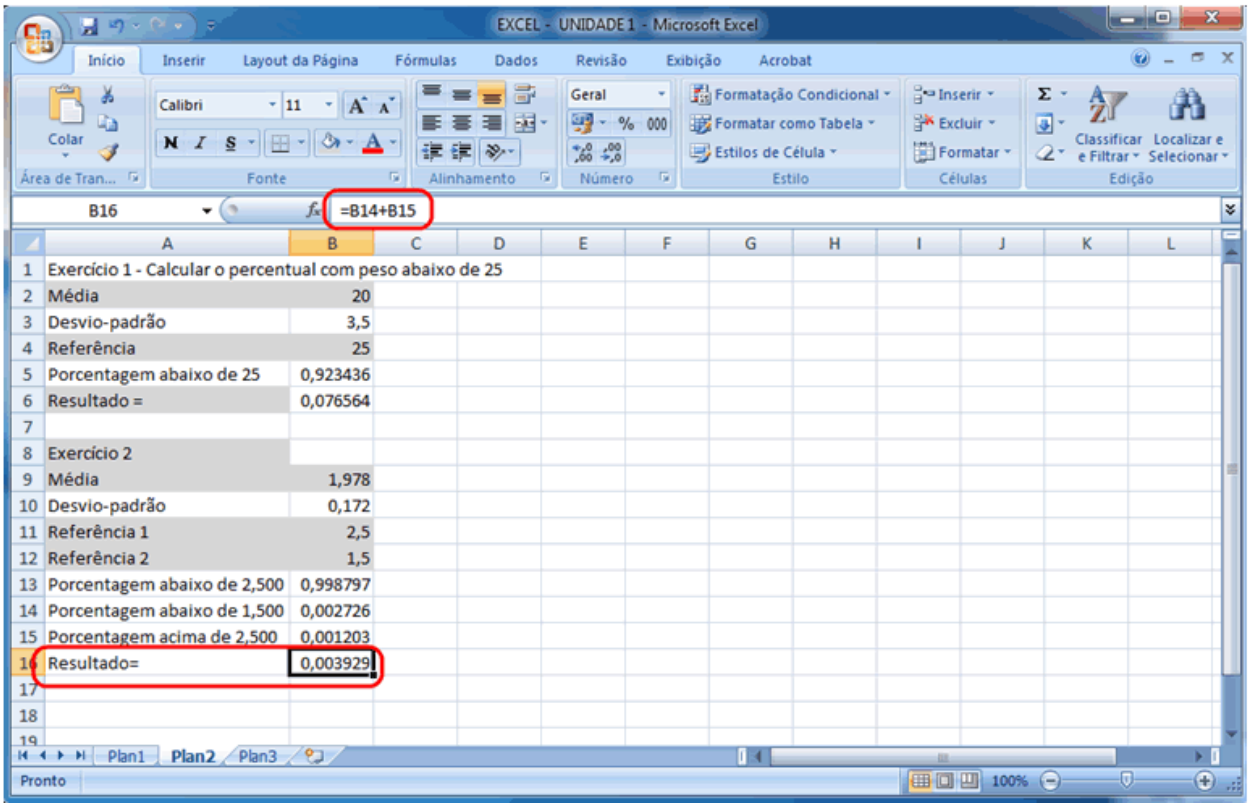
No segundo exercício, será necessário usar o recurso da DIST.NORM duas vezes:





76

Como pode ser visto na última planilha, a seguir, o resultado foi obtido de forma análoga àquela quando foi utilizada a tabela, com geração do mesmo resultado 0,39%.



	A	B
1	Exercício 1 - Calcular o percentual com peso abaixo de 25	
2	Média	20
3	Desvio-padrão	3,5
4	Referência	25
5	Porcentagem abaixo de 25	0,923436
6	Resultado =	0,076564
7		
8	Exercício 2	
9	Média	1,978
10	Desvio-padrão	0,172
11	Referência 1	2,5
12	Referência 2	1,5
13	Porcentagem abaixo de 2,500	0,998797
14	Porcentagem abaixo de 1,500	0,002726
15	Porcentagem acima de 2,500	0,001203
16	Resultado=	0,003929
17		
18		
19		

77

Resumo

A distribuição normal ou de Gauss é uma distribuição estatística de dados amostrais baseada na relação dos dados em análise com sua medida de tendência central (média) e com sua respectiva medida de dispersão (desvio-padrão).
Enfatizou-se a distribuição normal de dados pelo seu conjunto de características (principalmente simetria e percentuais de dados compreendidos entre a média e um, dois e três desvios-padrão para mais e para menos).

Há um conjunto de variáveis que, do ponto de vista prático, assume o aspecto de uma distribuição normal (fenômenos sociais, psicológicos e físicos), porém há um conjunto de outras que tipicamente apresentam outro formato (distribuição de renda, por exemplo).

O recurso de padronização dos dados, origina uma distribuição normal com média zero e desvio-padrão igual a um. Todas as bases de dados que formam uma distribuição normal são passíveis de serem padronizados, particularmente quando houver a necessidade de utilização de uma tabela para

determinação de percentuais de dados acima de um determinado valor, abaixo de outro ou entre dois valores especificados.

Caso seja utilizada a planilha Excel, para determinação de percentuais de dados, que atendam determinadas especificações, seguir-se-á os seguintes passos:

(a) abre-se uma planilha;
(b) digitam-se os dados de média, desvio- padrão e também o valor especificado como referência;
(c) clica-se no ícone fx, em Estatística, em DIST.NORM e em OK;
(d) entra-se com o valor (ou com as respectivas células) de X (referência), com a média, com o desvio-padrão e com a palavra Verdadeiro;
(e) caso haja necessidade (no caso de ser solicitada uma porcentagem acima de um determinado valor), faz-se uma última operação complementar: 1 - o percentual obtido após a realização do passo (d).