Seguro & Ensino

VOLUME I

# Estatística Básica

## PARA TOMADA DE DECISÃO

ESCOLA NACIONAL de SEGUROS
FUNENSEG

# Estatística Básica para Tomada de Decisão

## Volume 1

Coleção
Seguro & Ensino

# Estatística Básica para Tomada de Decisão

## Volume 1

**Autores**
Helio Morrone Cosentino
Álvaro Alves de Moura Jr.
André Castilho Ferreira da Costa

ESCOLA NACIONAL de SEGUROS
FUNENSEG

**Rio de Janeiro**
**2013**

1ª edição: outubro - 2013
Fundação Escola Nacional de Seguros – Funenseg
Rua Senador Dantas, 74 – Térreo, 2º, 3º e 4º andares
Rio de Janeiro – RJ – Brasil – CEP 20031-205
Tels.: (21) 3380-1000
Fax: (21) 3380-1546
Internet: www.funenseg.org.br
e-mail: faleconosco@funenseg.org.br

Impresso no Brasil/Printed in Brazil



Coordenação Editorial
*Assessoria da Diretoria Executiva/Núcleo de Publicações*

Edição
*Vera de Souza*
*Mariana Santiago*

Produção Gráfica
*Hercules Rabello*

Capa
*Grifo Design*

Diagramação
*Info Action Editoração Eletrônica Ltda. – Me*

Revisão
*Monica Teixeira Dantas Savini*

Virginia Thomé – CRB-7/3242
Responsável pela elaboração da ficha catalográfica

C767e Cosentino, Helio Morrone
Estatística básica para tomada de decisão / Hélio Morrone Cosentino; Álvaro Alves de Moura Jr.; André Castilho Ferreira da Costa. – Rio de Janeiro: Funenseg, 2013.
108 p. ; 26 cm – (Coleção Seguro & Ensino, v. 1)

ISBN nº 978-85-7052-556-7.

1. Estatística. I. Moura Jr, Álvaro Alves. II. Costa, André Castilho Ferreira da. III. Título.

0013-1239 CDU 519

# Sumário

# Apresentação

É com grande satisfação que apresento este livro desenvolvido pelos professores Helio Morrone Cosentino, Álvaro Alves de Moura Júnior e André Castilho Ferreira da Costa, experientes acadêmicos e especialistas em métodos quantitativos.

Eles têm atuado com sucesso em várias instituições de ensino superior de São Paulo, notadamente na ESNS-SP, onde lecionam para o curso de graduação em Administração.

A redação do texto, simples e objetiva, bem como a exemplificação de aplicações na área de seguros, através da planilha Excel, certamente facilitarão o aprendizado daqueles que necessitam utilizar a Estatística para o processo de tomada de decisão na indústria de seguros.

O livro certamente servirá como um adequado material de apoio nas disciplinas específicas ministradas no curso de graduação em Administração com linha de formação em Seguros e Previdência, assim como nos de pós-graduação em Seguros e Resseguro ministrados pela ESNS-SP.

Espero para breve a continuidade deste projeto, com o desenvolvimento de novos volumes pelos autores.

Uma boa leitura para todos!

**Domingos Alves Corrêa Neto**

# Características do Método Quantitativo e sua Relação com a Estatística e Nosso Dia a Dia

1

Ao se deparar com a necessidade de aprofundamento do conhecimento o pesquisador é obrigado a compreender conjuntos de dados relevantes ao seu objeto de pesquisa. Para tanto, é necessário trabalhar dados relativos a situações problemas, que posteriormente se refletirão em informações seguras para comparações e julgamentos.

O Método Quantitativo é parte da metodologia científica que tem por objetivo coletar, simplificar, analisar e modelar dados. Talvez seu aspecto mais relevante seja realizar previsões e projeções que podem auxiliar no processo de tomada de decisões, a partir de dados relativos ao fenômeno estudado. Sua principal característica é que costuma trabalhar com uma elevada quantidade de informações.

A fase inicial deste processo está associada à chamada Estatística Descritiva, objeto de estudo deste livro. A ela compete coletar os dados, organizá-los e resumi-los, apresentando-os (descrevendo-os) sob uma forma de fácil compreensão com tabelas, gráficos e valores típicos que representem o comportamento de todo o conjunto dos dados.

A quantidade de informações que temos que lidar em nosso dia a dia exige obrigatoriamente uma ferramenta capaz de controlar as vendas de uma empresa, a quantidade de leitos livres em um hospital, a programação de entregas de uma transportadora, as notas de um aluno etc.

### • • • Explorando melhor a ideia de Estatística Descritiva

A **Estatística Descritiva** é utilizada na ordenação, resumo e apresentação de dados relativos a um determinado fenômeno, de forma a tornar suas informações amplamente acessíveis.

# RESUMINDO INFORMAÇÕES

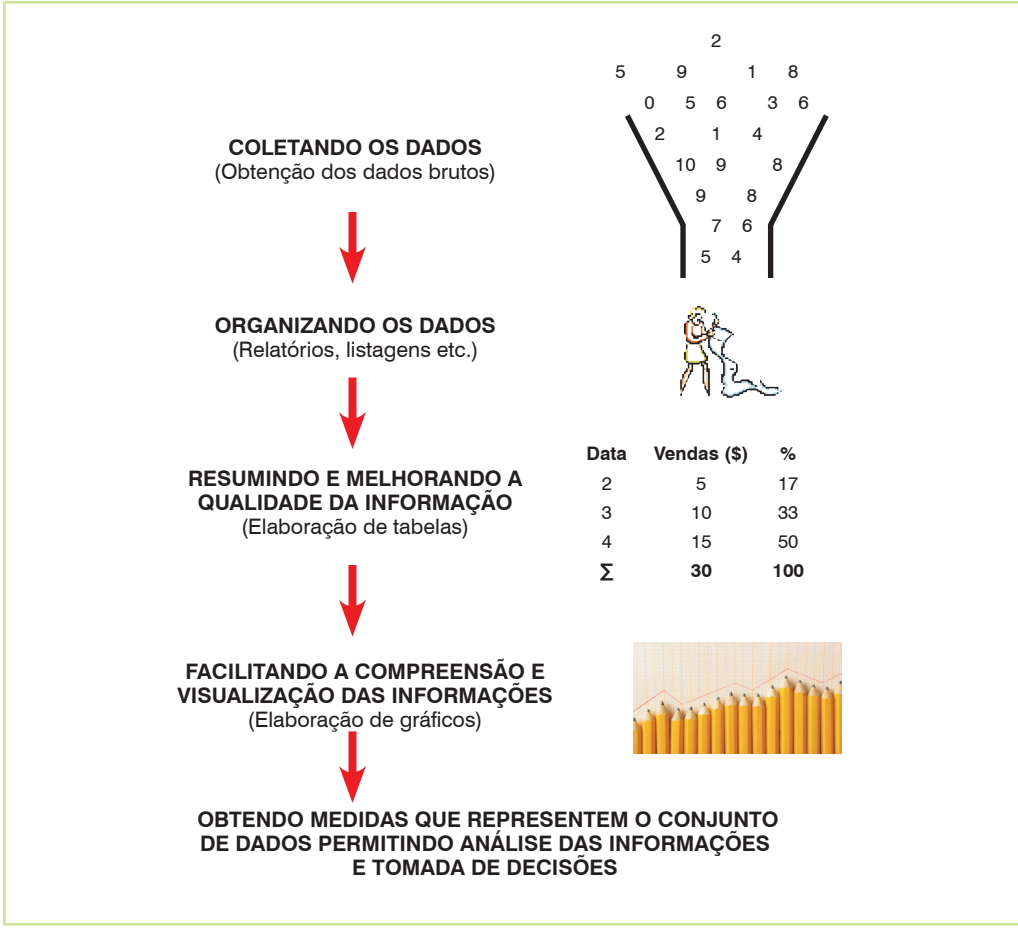
Observe que em um jornal ou revista, após lermos os títulos comumente conduzimos nossa visão diretamente para os gráficos contidos nos artigos, depois para as tabelas e, finalmente (já tendo compreendido a mensagem principal transmitida pelos gráficos e tabelas) passamos aos detalhes do texto. Muitas vezes já consideramos suficiente "ler" os gráficos e tabelas e damos a compreensão do assunto por completa, sem ao menos ler o texto.

Em qualquer empresa, que diariamente conte com diversos relatórios de controle (estoques, vendas, faturamento) seus dirigentes precisam manter um panorama atualizado da situação do negócio. Para tanto, devem consultar incontáveis listagens de dados para conhecer o desempenho de cada setor. Tarefa pouco produtiva ou mesmo impossível!

Uma forma de simplificar a ação dos executivos dessa empresa seria resumir as informações coletadas com o uso de recursos estatísticos, gerando uma forma de consulta fácil e rápida, tornando o processo de controle eficaz.

De forma mais completa, o fluxo de informações envolvido nessa empresa poderia seguir as seguintes etapas da Fig. 1:

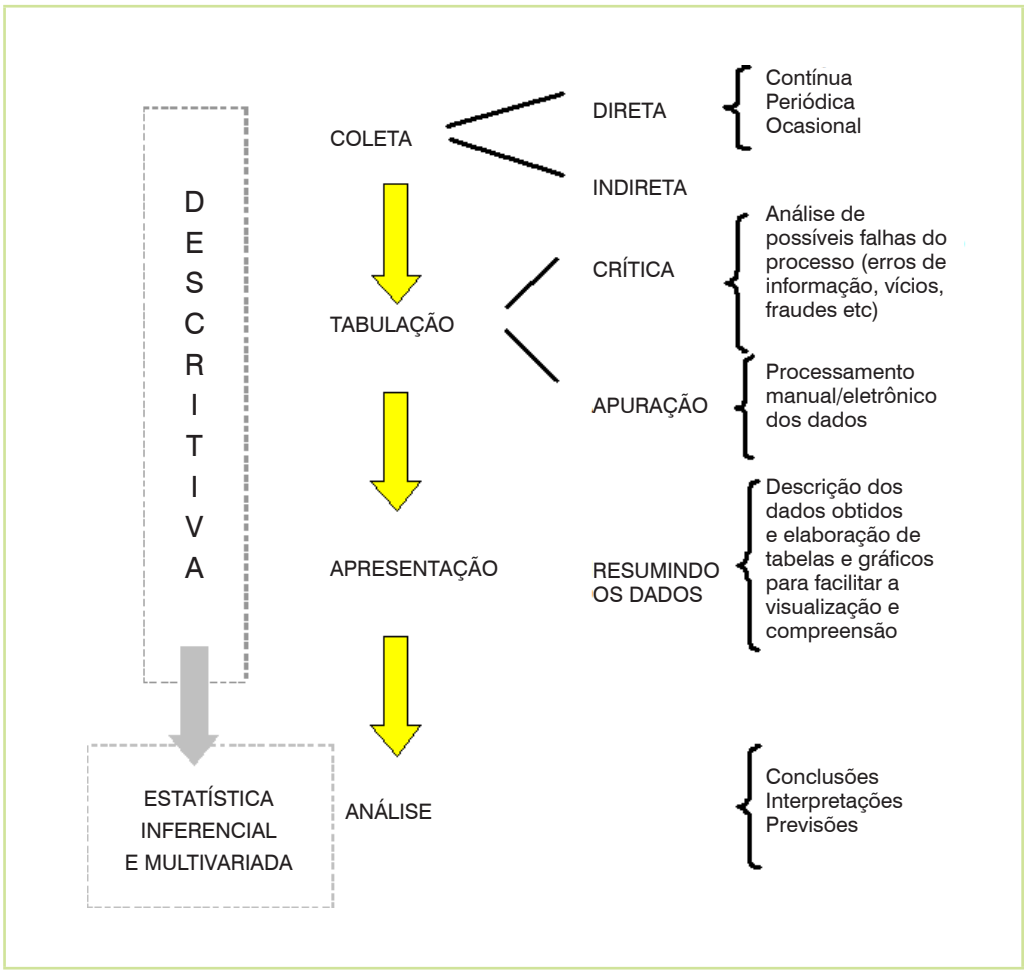
**Figura 1**
**Fluxo de Informações Típico de um Negócio**



Os processos estatísticos passam por etapas semelhantes, independentemente da situação estudada, seja na administração de empresas, medicina, engenharia ou qualquer outra atividade.

De forma mais sistemática e detalhada podemos examinar na Fig. 2 os passos de um processo estatístico básico:

**Figura 2**
**Processo Estatístico Padrão para Tratamento de Dados**




Embora a **estatística descritiva** também possibilite conclusões iniciais e interpretações sobre os fatos estudados, normalmente cabe à **estatística inferencial e multivariada** este estudo mais completo, ou seja, a estatística descritiva tem por foco relatar e organizar dados passados ou presentes que servirão de base para análises, previsões e projeções futuras sobre um fenômeno.

Em uma empresa, por exemplo, os dados relativos às vendas realizadas nos últimos três anos servem como base para gerarmos modelos de previsão das vendas do próxi-

mo ano, ou seja, com base no desempenho das vendas de um determinado produto em períodos anteriores podemos tentar estabelecer o comportamento das vendas em um dado momento futuro.

Normalmente a estatística faz uso de parcelas de uma população (amostras) para um estudo, em vez de examinar todos os elementos da população. A forma correta e adequada de fazê-lo também é objeto de estudo específico da estatística inferencial (e não da descritiva), mas será comentado em linhas gerais adiante.

## A COLETA DE DADOS

A coleta de dados **DIRETA** é realizada a partir dos elementos estudados e pode ser classificada sob três formas diferentes:

- **CONTÍNUA**: controle diário de frequência de alunos, registros de nascimentos, cotação do valor diário de uma ação.
- **PERIÓDICA**: médias semestrais dos alunos, censo demográfico (a cada 10 anos).
- **NECESSIDADES MOMENTÂNEAS**: pesquisas de opinião para lançamento de um novo produto, cadastramento de vítimas ou sobreviventes de uma epidemia.

Já a coleta de dados **INDIRETA** é elaborada a partir de dados previamente reunidos pela coleta direta (alguém já fez a coleta direta antes). Por exemplo, a quantidade de nascimentos mensais em um estado pode ser obtida (indiretamente) a partir do número de nascimentos diários registrados em cada uma das diversas cidades daquele estado. O pesquisador precisa apenas "juntar" os dados previamente obtidos.

## FORMAS DE COLETAR DADOS


A coleta de dados deve ser realizada de forma bastante criteriosa, uma vez que servirá como base para todo o estudo estatístico sobre o fenômeno.

Ela pode ser realizada através de um **CENSO**, que envolve todos os elementos de uma população, ou por **AMOSTRAGEM** quando usa apenas um subconjunto significativo dessa população.

A utilização do censo é bastante restritiva, uma vez que costuma envolver custos elevados e dificuldade da obtenção de dados.

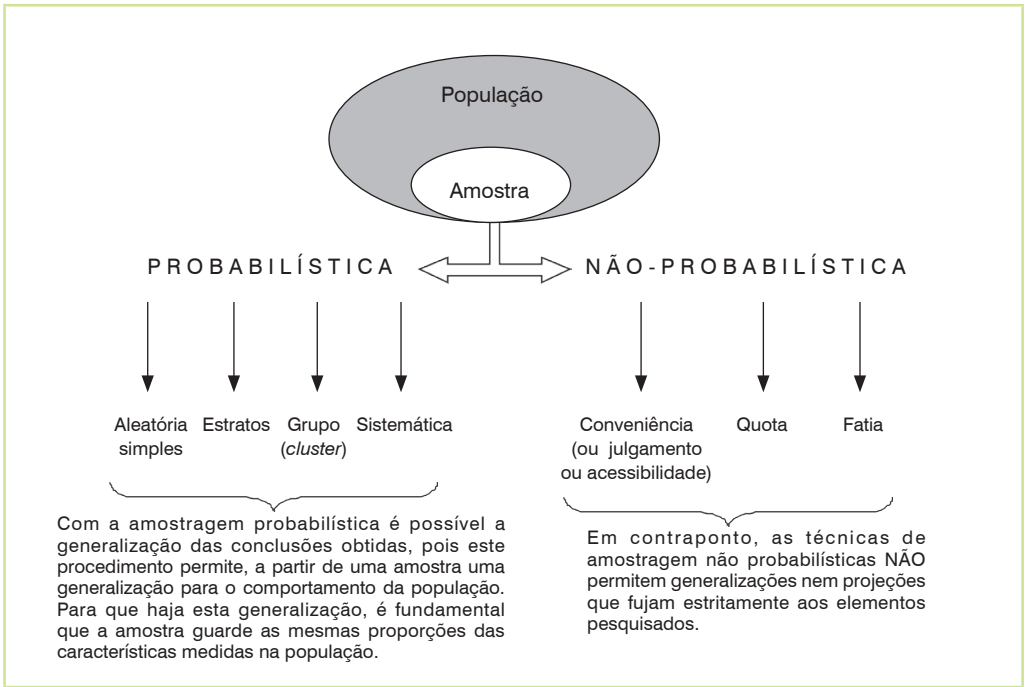
Já a utilização de **AMOSTRAS,** obtidas através de **técnicas amostrais** busca a redução de custos e uma maior facilidade na obtenção e processamento de dados. A ideia é estudar apenas parte da população e transferir as informações obtidas a partir desta amostra para toda a população.

Desta forma evita-se analisar um a um os elementos da população, minorando custos e trabalho, porém aumentando-se o risco de erro.

Matematicamente, a **POPULAÇÃO** (alguns chamam de UNIVERSO) é definida como o conjunto constituído por todos os elementos que poderiam ser investigados no estudo, já a **AMOSTRA** é um subconjunto desta população. É muito importante que a amostra seja realmente representativa da população, ou seja, ela deve guardar as mesmas características da população de onde se originou.

Costumeiramente a discussão sobre técnicas amostrais é realizada em um capítulo à parte, dada sua vital importância em estudos estatísticos. Desta forma, abordaremos aqui apenas suas principais ideias, sem esgotá-las, mas em profundidade suficiente para o entendimento dos conceitos ora abordados. A Fig. 3 demonstra as possíveis espécies de amostras.

**Figura 3**
**Tipos de Amostragem Possíveis**



Na amostra **PROBABILÍSTICA** ocorre a escolha de elementos com base em probabilidades determinadas (na verdade há uma chance igual de qualquer elemento ser selecionado), já na amostra **NÃO PROBABILÍSTICA** não são aferidas a probabilidade de suas ocorrências.

Por esses motivos, o uso de amostras não probabilísticas deve ser restrito a estudos iniciais de um fenômeno, ou quando se deseja conhecê-lo de forma inicial, devendo ser seguido depois de uma exploração mais rigorosa, caso sejam desejados resultados consistentes que permitam generalizações, estabelecimentos de leis ou obtenção de padrões de comportamento.

A utilização de procedimentos estatísticos que envolvam amostras não probabilísticas dificilmente pode ser empregada na modificação de processos de quaisquer espécies (administrativos, industriais ou mesmo teóricos), uma vez que pode ser afetado pelo **viés** da escolha da amostra. Amostras com vieses são ditas **tendenciosas,** impedem generalizações e podem implicar em erros.

### • • • Relatório Hyte, um caso típico de viés ou não?

Apesar dos estudos de Hyte terem revolucionado a cultura norte-americana e terem sido bem conduzidos do ponto de vista estatístico na ocasião, hoje discute-se sua validade. Somente as pessoas que "eram livres para discutir sexo" teriam respondido às perguntas voluntariamente, ou seja: o resultado (incorreto???) mostrou o quadro de uma população americana extremamente liberal frente a hábitos sexuais, o que de fato não ocorreria, pois uma parcela da população (os conservadores) não teria participado.

Neste capítulo discutiremos apenas as amostras probabilísticas, ficando o mérito do uso das amostras não probabilísticas para discussão em momento posterior.

### • • • Explorando um pouco mais as amostras probabilísticas

Qual o tipo de amostra probabilística é mais eficiente?
O tipo de amostragem utilizado deve ser escolhido conforme as necessidades do estudo a ser feito, mas de forma geral a amostragem aleatória simples costuma ser a mais eficiente, seguida pela amostragem sistemática, estratificada e, por último, a por grupos. Já em termos de custo-benefício, a ordem talvez se inverta.

Se houver dúvida sobre a escolha do tipo de amostra, para maior segurança recomenda-se a amostragem aleatória simples. A Fig. 4 procura estabelecer com mais detalhes:

**Figura 4**
**Tipos de Amostragens Probabilísticas, Características e Usos**

| TIPO | CARACTERÍSTICAS | ALGUNS USOS |
|---|---|---|
| ALEATÓRIA SIMPLES | Uso bastante simples em amostras pequenas e médias, sendo trabalhosa em amostras grandes. Os elementos a serem estudados são escolhidos aleatoriamente (um sorteio, por exemplo) | Controle de qualidade |
| ESTRATIFICADA | Uso quando existirem estratos, sendo necessário definir a proporcionalidade adequada para os constituintes da amostra | Quando a população tiver parte de seus componentes com características específicas que podem ser decisivas nos resultados finais do estudo |
| SISTEMÁTICA | Uso periódico, para acompanhamento de processos, sistemas e linhas de produção | Típicos de indústrias onde é necessário um controle de qualidade dos lotes. Pode também ser utilizada em serviços |
| GRUPO (*cluster*) | Usada quando a população estiver espalhada por uma grande área. Os elementos de uma população são divididos em vários grupos representativos. Um ou mais grupos são escolhidos e seus elementos são investigados em detalhes | Cidades, bairros, clientes de filiais ou departamentos de uma empresa são grupos formados naturalmente |

Alguns exemplos costumam ilustrar situações típicas.

Na amostragem ALEATÓRIA SIMPLES em uma empresa sorteiam-se alguns funcionários para responderem a uma pesquisa de opinião sobre a qualidade dos serviços do restaurante da empresa. Esse sorteio pode ser feito com um globo giratório contendo bolinhas numeradas, tabelas randômicas do Excel ou mesmo tabelas prontas pré-sorteadas.

Na amostragem ESTRATIFICADA, em uma população com 40 pessoas, contendo 30 homens e 10 mulheres, escolhem-se três homens e uma mulher para responderem a uma pesquisa de opinião sobre características de um novo modelo de veículo. Respeita-se a proporção de 3:1 (estratos de homens/mulheres).

Na SISTEMÁTICA, em uma linha de produção de lâmpadas, a cada lote contendo 1.000 unidades produzidas uma lâmpada, sistematicamente, é retirada ao acaso para testes e avaliação pelo controle de qualidade.

Já na amostragem por GRUPOS (*clusters)*, em uma cidade deseja-se estudar as características de saúde pública e poluição do ar dos bairros vizinhos a uma fábrica poluidora.

## CONSIDERAÇÕES ÉTICAS SOBRE OS PROCESSOS DE AMOSTRAGEM

Discuta com o seu professor as implicações sobre as manipulações que podem envolver um processo de amostragem.

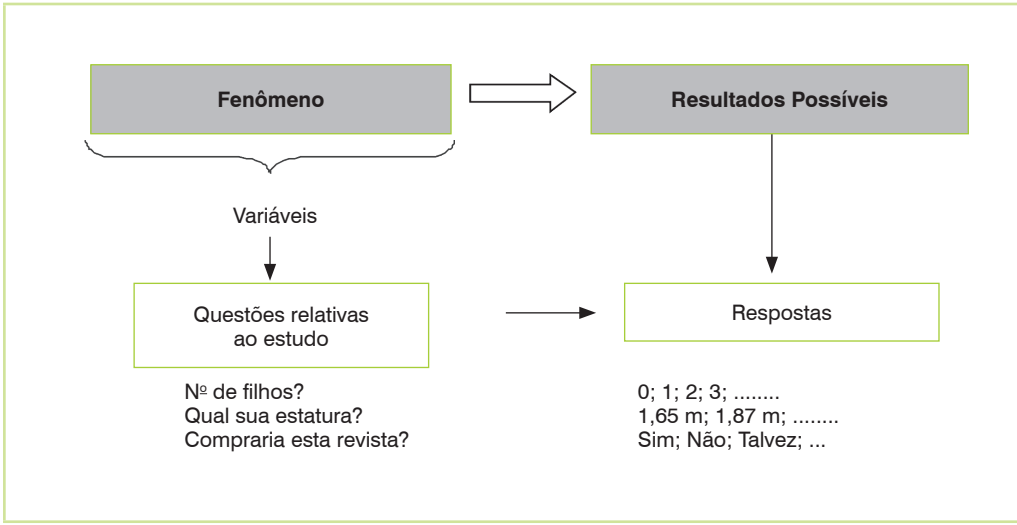
Após o entendimento de como podemos retirar uma amostra de uma população, devemos pensar no que gostaríamos de saber sobre esta população. Quais as perguntas que faríamos. Talvez essa seja a ideia básica sobre o conceito estatístico de VARIÁVEL.

## VARIÁVEIS

Embora na metodologia científica o conceito de variável seja mais amplo, sob o ponto de vista da estatística a denominação de **variável** é dada às possíveis formas de representação de um fenômeno em estudo, como representa a Fig. 5.

Em uma pesquisa de opinião, por exemplo, as diversas questões contidas no questionário buscam avaliar todas as variáveis envolvidas no problema. O termo variável é bem apropriado, visto que se espera que as respostas possam variar para cada um dos entrevistados.

**Figura 5**
**Ideia de Variável**

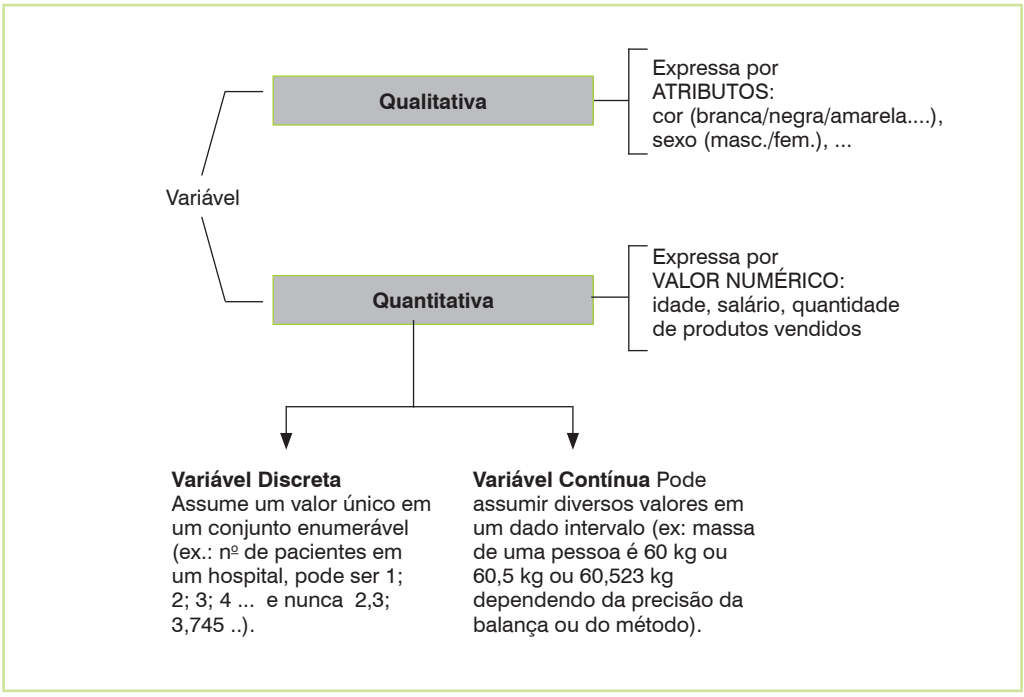



Observe que as respostas necessárias a qualquer estudo estatístico podem ser diversas, ou seja, comumente são abordadas inúmeras variáveis que envolvem dados **mensuráveis** (medidos), tais como estatura, renda, quantidade de vendas e produtos armazenados, ou dados que podem ser apenas **enumerados (categorizados)**, como raça, preferência por uma cor, religião, preferência musical.

Esquematicamente, os tipos de variáveis envolvidos na estatística descritiva podem ser divididos em duas grandes classes, conforme segue na Fig. 6:

## TIPOS DE VARIÁVEIS

**Figura 6**
**Tipos de Variáveis**




Dependendo do tipo de variável utilizado em um estudo será necessário adotar um tratamento estatístico específico. As **variáveis qualitativas não permitem cálculos ou representações matemáticas diretas** (média, por exemplo), porém aceitam abordagens estatísticas apropriadas (determinação da moda, por exemplo). Não é possível "fazer conta" com a preferência das pessoas. Se três pessoas preferem a cor verde, duas a vermelha e quatro a roxa qual seria a cor "resultante" preferida pela população? As variáveis qualitativas subdividem-se em duas categorias: ***nominais*** (apenas uma qualidade, como cor dos olhos) e ***ordinais*** (existe hierarquia, como grau de instrução). Note que, mesmo que seja associado um código a essa variável qualitativa, como a escala Likert (1 = discordo totalmente, 2 = discordo parcialmente, 3 = neutro, 4 = concordo parcialmente e 5 = concordo totalmente), ela NÃO se torna uma variável quantitativa.

As variáveis quantitativas dividem-se em duas subclasses (discretas e contínuas) e sempre aceitam ser representadas através de valores numéricos (média, mediana, moda, índices, coeficientes e taxas).

O conceito de variável discreta e contínua é relativamente simples, porém pode causar alguma confusão em tratamentos estatísticos mais elaborados.

Embora as variáveis discretas assumam somente valores inteiros, podem ser representadas durante os cálculos sob outras formas. No exemplo sobre pacientes atendidos em um hospital, sempre encontraremos como resultado valores inteiros (22; 23; 24; 25;...), porém a média de pacientes atendidos em um dia pode ser um número não inteiro (25,68 pacientes/dia). A variável permanece sendo do tipo discreta, apenas sua representação estatística modificou-se, apresentando-se como um número decimal.

## Exercícios

1) Classifique as variáveis a seguir e sugira um tipo de escala para medi-las:

**Tabela 1**
**Estudo de Enfermidades por Sexo, Gênero, Cor, Gravidade e Epsódios na Empresa T&T**

| Nº | Sexo | Idade | Cor | Gravidade | Episódios |
|---|---|---|---|---|---|
| 1 | M | 39 | branca | Leve | 2 |
| 2 | F | 8 | amarela | Leve | 1 |
| 3 | F | 24 | preta | Moderada | 3 |
| 4 | M | 13 | branca | Grave | 3 |
| 5 | M | 20 | parda | Leve | 1 |
| 6 | F | 4 | preta | Moderada | 2 |
| 7 | F | 12 | amarela | Moderada | 2 |
| 8 | F | 43 | branca | Leve | 2 |
| 9 | M | 9 | preta | Grave | 4 |
| 10 | M | 30 | branca | Grave | 5 |
| 11 | F | 12 | branca | Moderada | 3 |
| 12 | M | 11 | parda | Leve | 2 |

Fonte: Departamento Médico T&T – 2004.

# Apresentação dos Dados (Tabelas e Gráficos)

2

Os dados coletados durante um processo estatístico devem ser inicialmente tabulados, para somente depois serem apresentados e comunicados de forma conveniente.

Na fase de tabulação devem sofrer uma análise crítica, quando o pesquisador verificará se as informações possuem qualidade suficiente para utilização segura. Nesse momento são verificadas fraudes, tendenciosidades e erros que possam ter ocorrido durante o processo de coleta de dados.

Em uma pesquisa de opinião, por exemplo, é comum que o supervisor de uma equipe de entrevistadores telefone para alguns dos respondentes para verificar se, de fato, ele foi entrevistado. Assim, evita-se que algum entrevistador desonesto preencha por si próprio vários questionários, esgotando sua cota de trabalho rapidamente.

Após a eliminação de quaisquer itens suspeitos ou incorretos, o que acarretará na diminuição de tamanho da amostra poderá ocorrer a apuração manual ou eletrônica dos dados (normalmente *softwares* como Excel, SPSS, Statistica, Minitab etc) e só então a apresentação e comunicação dos resultados.

Uma pesquisa necessita de um planejamento bastante cuidadoso, abordando as variáveis estritamente necessárias. Caso contrário, corre-se o risco de obtenção de dados supérfluos que só dificultam a observação do problema focado. Deve-se evitar a tentação de aplicar inúmeros instrumentos de pesquisa ao mesmo tempo (testes, questionários, escalas), que somente gerarão dados inúteis, sendo comum o pesquisador acreditar que o uso de *softwares* estatísticos vai tratá-los com facilidade. De fato, recursos computacionais podem tratar elevada porção de informações, porém a interpretação e relação entre os resultados e as variáveis só podem ser aferidas pelo pesquisador.

## SÉRIES ESTATÍSTICAS

Uma forma bastante útil na apresentação dos dados é através das séries estatísticas, as quais os dados são dispostos em diferentes tipos de tabelas, que podem ser classificadas em quatro tipos:

i) **Série Temporal**: é um tipo de série na qual o elemento variável representa um determinado período (dia, mês, trimestre, ano etc).
**Exemplo – Série Temporal**

**Exportações Brasileiras (FOB) – em US$ milhões**

| Ano | Exportações FOB |
|---|---|
| 2000 | 55.085,60 |
| 2001 | 58.222,64 |
| 2002 | 60.361,79 |
| 2003 | 73.084,14 |
| 2004 | 96.475,24 |
| 2005 | 118.308,39 |
| 2006 | 137.807,47 |
| 2007 | 160.649,07 |
| 2008 | 197.942,44 |
| 2009 | 152.994,74 |
| 2010 | 201.915,29 |
| 2011 | 256.039,58 |

Fonte: BCB

ii) **Série Geográfica**: expressa um tipo de série em que o elemento variável é expresso por uma determinada localidade (bairro, cidade, estado, região, país etc).
**Exemplo – Série Geográfica**

**DPVAT – Indenizações Pagas por Região – 2010**

| REGIÃO | Qtde Total | Indenizações Pagas |
|---|---|---|
| NORTE | 21.254 | R$ 196.615.694,48 |
| NORDESTE | 61.316 | R$ 485.943.320,08 |
| CENTRO OESTE | 21.350 | R$ 221.743.217,62 |
| SUDESTE | 70.501 | R$ 535.016.405,38 |
| SUL | 77.930 | R$ 589.468.442,53 |
| TOTAL | 252.351 | R$ 2.028.787.080,09 |

Fonte: Seguradora Líder

iii) **Série Específica**: expressa um tipo de série em que o chamado elemento variável é um determinado fenômeno (mortes por acidente de carros, exportações por setor de atividade etc).

**Exemplo – Série Específica**

**Ocorrências Policiais Registradas, por Natureza – 2º Trimestre de 2011**

| Discriminação | Capital | Estado |
|---|---|---|
| Contra a pessoa | 34.071 | 164.887 |
| Contra o patrimônio | 112.846 | 286.517 |
| Contra os costumes | 834 | 3.519 |
| Entorpecentes | 1.520 | 11.721 |
| Contravencionais | 3.835 | 19.545 |
| Outros criminais (não inclui contravenções) | 6.490 | 32.746 |
| Total de Crimes Violentos (Hom. Doloso, Roubo, Latrocínio, Estupro e EMS) | 38.198 | 81.588 |
| Total de delitos | 159.596 | 518.935 |
| Não Criminais | 92.784 | 262.082 |

Fonte: Secretaria de Segurança Pública do Estado de São Paulo

iv) Distribuição de Frequências demonstram com que frequência (quantas vezes) o dado observado se repete. A Distribuição de Frequências pode ser estruturada de duas formas: simples (não agrupadas em classes) ou agrupada em classes, conforme mostram os exemplos abaixo.

**Exemplo – Distribuição de Frequências Simples (não agrupada em classes/ intervalos)**

**Nota de Estatística da Turma A**

| Notas ($x_i$) | N. de Alunos ($f_i$) |
|---|---|
| 1,0 | 1 |
| 2,0 | 3 |
| 3,0 | 4 |
| 4,0 | 6 |
| 5,0 | 12 |
| 6,0 | 7 |
| 7,0 | 6 |
| 8,0 | 5 |
| 9,0 | 3 |
| 10,0 | 3 |
| **Total** | **50** |

Fonte: Diário de Classe

**Exemplo – Distribuição de Frequências agrupada em classes/intervalos**

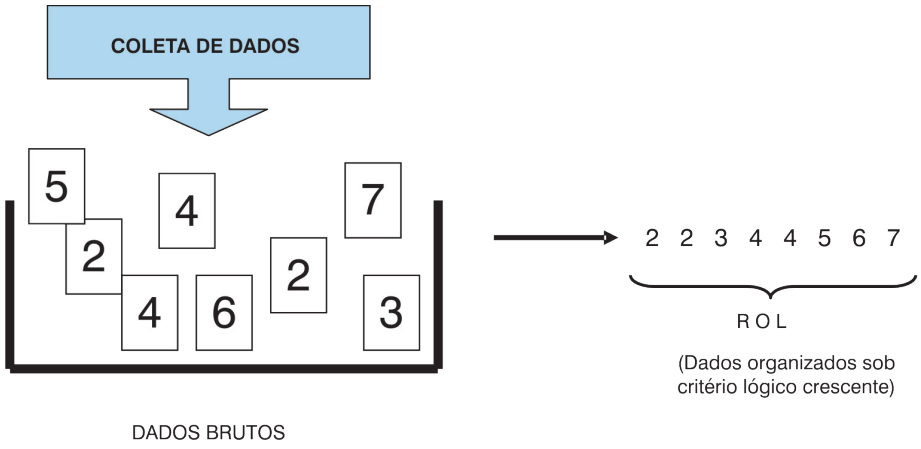
**Óbitos p/Ocorrência por Faixa Etária no Município de São Paulo – CID10: V01-V99 Acidentes de transporte Período: 2010**

| Faixa Etária ($X_i$) | Frequência ($f_i$) |
|---|---|
| Menor 1 ano | 1 |
| 1 a 4 anos | 3 |
| 5 a 9 anos | 11 |
| 10 a 14 anos | 23 |
| 15 a 19 anos | 101 |
| 20 a 29 anos | 340 |
| 30 a 39 anos | 218 |
| 40 a 49 anos | 148 |
| 50 a 59 anos | 137 |
| 60 a 69 anos | 91 |
| 70 a 79 anos | 96 |
| 80 anos e mais | 66 |
| Idade ignorada | 28 |
| Total | 1.263 |

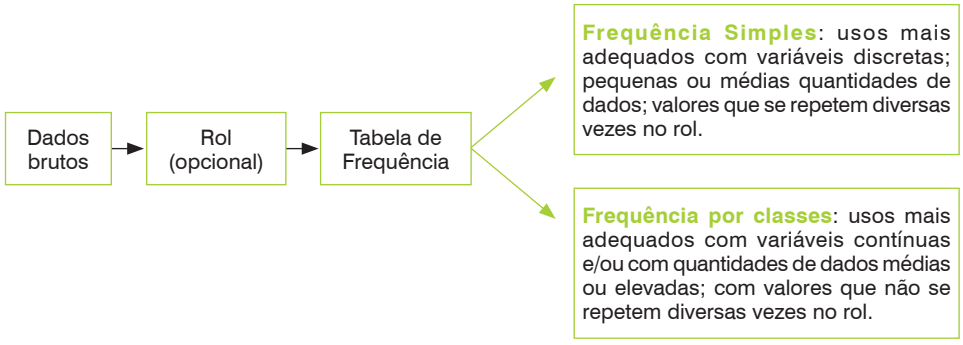
Fonte: DATASUS

## ELABORAÇÃO DE UMA DISTRIBUIÇÃO DE FREQUÊNCIAS AGRUPADA EM CLASSES

Os *softwares* realizam, como será demonstrado ao longo deste capítulo, esta tarefa com facilidade. Porém, com o objetivo de compreender como se elabora uma distribuição de frequências faremos, inicialmente, de forma manual, o que costuma exigir uma primeira ordenação lógica (**Rol**) dos dados para facilitar a contagem. A figura a seguir ilustra o processo.




A elaboração do Rol é opcional e muitas vezes não tem grande utilidade prática, permitindo-se montar diretamente dois diferentes tipos **de tabelas de distribuição de frequências**:



Observe a seguinte sequência de dados e sua apresentação em uma tabela. Por serem dados com características discretas, em pequena quantidade e se repetirem no rol, poderemos usar uma **tabela de distribuição de frequência simples**:

**Rol: 25, 25, 30, 45, 45, 50, 62, 77**

| $x_i$ | $f_i$ | $F_i$ | $f_{i\%}$ | $Fi_{\%}$ |
|---|---|---|---|---|
| 25 | 2 | 2 | 2/8 = 25% | 25% |
| 30 | 1 | 2+1 = 3 | 1/8 = 12,5 % | 37,5 % |
| 45 | 2 | 2+1+2 = 5 | 25% | 62,5% |
| 50 | 1 | 2+1+2+1 = 6 | 12,5 % | 75% |
| 62 | 1 | 2+1+2+1+1 = 7 | 12,5 % | 87,5% |
| 77 | 1 | 2+1+2+1+1 +1 = **8** | 12,5 % | **100%** |
| | 8 | | **100%** | |
| | ↓ | | ↓ | |
| ↓ Coluna das variáveis ($x_i$) indica quais as variáveis presentes no rol | ↓ Coluna das frequências simples ($f_i$) indica quantas vezes cada variável aparece no rol $\Sigma f_i$ = n = nº de elementos do rol | | ↓ Coluna das frequências (relativas) porcentuais ($f_i$ %) indica a participação % de cada uma das variáveis no todo | |

A tabela poderá conter apenas as colunas que o pesquisador julgar necessárias.

As colunas $F_i$ e $F_i$ **%** apresentam os resultados das **frequências acumuladas simples e porcentuais, respectivamente, a cada linha**.

Os resultados numéricos iguais, hachurados em cinza nas duas últimas linhas da tabela, permitem verificar se os cálculos como um todo estão corretos.

Como dissemos anteriormente, o **Rol: 25, 25, 30, 45, 45, 50, 62, 77** é constituído por pequena quantidade de variáveis discretas que se repetem no rol, motivo pelo qual utilizamos a forma mais adequada de representação que é a tabela de distribuição de frequências simples.

A título de comparação, montaremos com os mesmos dados uma **tabela de distribuição de frequências por classe**. Lembramos, no entanto, que este tipo de tabela é mais recomendado para quantidades mais elevadas de dados, mas para efeito didático vamos montar uma distribuição agrupada em classes.

Neste caso escolhemos três classes, sendo que os critérios mais comumente adotados para definir o nº de classes podem ser:

- a experiência do pesquisador;
- o interesse específico que possuímos no estudo;

- critérios matemáticos podem ser escolhidos, sendo o mais comum o **CRITÉRIO DA RAIZ** em que o nº de classes é determinado pela raiz quadrada do número de dados observados, ou seja, **K** = $\sqrt{n}$ ; em que K é o número de intervalos e *n* o número de dados observados; e

- pode-se usar ainda a regra de Sturges, em que K = 1 + 3,3 log n.

No exemplo, usando o critério da raiz, verificamos que nosso rol possui n = 8 elementos, então K = $\sqrt{8}$ = 2,8. Comumente aproximamos o resultado para o maior inteiro próximo ao valor, no caso K = 3. A título de comparação, podemos imaginar K como sendo a quantidade de gavetas que será necessária em um armário para conter todas as variáveis do rol.

No passo seguinte devemos calcular o tamanho destas gavetas, de forma que sejam suficientemente grandes para receber todos os valores.

Assim, calculamos a **Amplitude de classe (A):**

**A = (maior valor do rol – menor valor do rol) / K**

No nosso caso, A = (77 – 25) / 3 = 17,3 que pode ser utilizado desta forma, ou arredondado para 18 ou mesmo 20. Desta forma, a classe nº 1 inicia-se em 25 (menor valor do rol) e é adicionado da amplitude A = 20 e assim por diante, ou seja:

25 + 20 = 45 ou 25 ⊢ 45 e depois
45 + 20 = 65 ou 45 ⊢ 65 e finalmente
65 + 20 = 85 ou 65 ⊢ 85

A aproximação (arredondamento) da amplitude de classe para valores inteiros normalmente facilita a compreensão dos resultados, mas deve ser utilizado com parcimônia, para não distorcer muito os resultados. O resultado final seria:

| Classe | $x_i$ | $f_i$ | $F_i$ | $f_i$% | $F_i$% |
|---|---|---|---|---|---|
| 1 | 25 a 45 | 3 | 3 | 37,50% | 37,50% |
| 2 | 45 a 65 | 4 | 7 | 50,00% | 87,50% |
| 3 | 65 a 85 | 1 | 8 | 12,50% | 100,00% |
| **Total** | | **8** | **–** | **100,00%** | **–** |

Perceba no exemplo acima que os valores estão sendo repetidos, ou seja, no primeiro intervalo o limite superior da primeira classe, o valor 45, é igual ao limite inferior da

segunda classe. No entanto, para efeito de cálculo e classificação das frequências de classes devemos adotar alguns critérios para a definição desses limites de classes.

Uma maneira de se fazer isso é utilizando os seguintes critérios:

1) ⊢ (inclui o limite inferior da classe, e exclui o limite superior da classe);
2) ⊣ (exclui o limite inferior da classe, e inclui o limite inferior da classe);
3) ⊢⊣ (incluem os limites inferior e superior da classe); e
4) – (excluem os limites inferior e superior da classe).

Vale ressaltar que a escolha desses intervalos depende da disposição dos dados ao longo da distribuição de frequências, o que significa que cada caso deverá ser avaliado individualmente, sempre seguindo a máxima de tornar a interpretação o mais fácil possível para o leitor.

Dois pesquisadores diferentes podem escolher valores de **K** e de **A** diversos, pois a regra não dá uma solução exata para a disposição da distribuição de frequências, o que implicará em duas, ou mais, tabelas diferentes, que representarão o mesmo fenômeno. Se no exemplo acima tivéssemos escolhido A = 18, teríamos:

| Classe | $x_i$ | $f_i$ |
|---|---|---|
| 1 | 25 ⊢ 43 | 3 |
| 2 | 43 ⊢ 61 | 4 |
| 3 | 61 ⊢ 79 | 1 |
| **Total** | | **8** |

Observe as diferenças de valores na coluna das frequências. Note também que a 1ª tabela facilita mais a compreensão. Em uma revista ou relatório, por exemplo, a 1ª tabela provavelmente seria mais adequada à leitura e entendimento do que a outra, porém ambas estão corretas e representam adequadamente o fenômeno. Uma dica para elaborar uma distribuição de frequência é sempre buscar facilitar ao máximo o leitor, mas cuidado, não elabore distribuição com classes muito grande, sendo que parte delas tenha uma frequência muito pequena, tal procedimento pode distorcer as estatísticas (média, moda, mediana etc) calculadas a partir dessa série.

## Comparando Séries de Dados

Suponha a necessidade de comparar diversas séries de valores, como nos **róis** abaixo.

Isto poderia acontecer, por exemplo, em um estabelecimento comercial que realiza suas vendas sob três formas de pagamento diferentes (dinheiro, cheque e cartão de crédito).

Se desejarmos observar a relação entre as três modalidades de pagamentos deveremos obter tabelas com o mesmo nº de classes (K) e a mesma amplitude (A) em cada uma das classes.

O critério da raiz pode ser aplicado nos róis apenas para dar uma ideia da quantidade de classes que poderemos usar, mas nada impede ao pesquisador criar quantas classes (K) achar necessárias para estudo do problema, lembrando a utilização de intervalos de classe (A) de mesmo tamanho para facilitar o entendimento.

| Dinheiro | Cheque | Cartão |
|---|---|---|
| 2,00 | 12,00 | 26,00 |
| 2,38 | 19,00 | 29,50 |
| 4,59 | 45,00 | 32,00 |
| 12,20 | 88,33 | 44,00 |
| 12,30 | 95,40 | 48,00 |
| 15,00 | 125,12 | 56,30 |
| 15,00 | | 59,50 |
| 16,35 | | 60,00 |
| 19,27 | | 60,00 |
| 22,10 | | 132,25 |
| 23,20 | | |
| 37,50 | | |
| 48,00 | | |
| 53,75 | | |
| n = 14 | n = 6 | n = 10 |

Para criarmos um critério de classes que atenda às três séries de dados, poderíamos imaginá-las constituindo uma série única.

Para um total de (14 + 6 + 10) = 30 valores

$K = \sqrt{n}$ ou $K = \sqrt{30} = 5,48 = 6$ classes

A = (maior valor – menor valor) / K

A = (132,25 – 2,00) / 6 = 21,71 = 22,00 ou 25,00

Optamos por uma tabela de **distribuição de frequência por classes**, uma vez que temos uma quantidade média de dados, trabalhamos com uma variável contínua e existem poucos valores que se repetem nos róis.

Para melhorar ainda mais o aspecto final da tabela, além de "forçarmos" a aproximação da amplitude da classe (A) para $ 25,00 poderíamos iniciar também a classe nº 1 em zero, uma vez que o menor valor dos róis ($2,00) é próximo a este valor. Teríamos então:

| Classe | $X_i$ (R$) | Dinheiro – $f_i$ | Cheque – $f_i$ | Cartão – $f_i$ |
|---|---|---|---|---|
| 1 | 0 ⊢ 25 | 11 | 2 | 0 |
| 2 | 25 ⊢ 50 | 2 | 1 | 5 |
| 3 | 50 ⊢ 75 | 1 | 0 | 4 |
| 4 | 75 ⊢ 100 | 0 | 2 | 0 |
| 5 | 100 ⊢ 125 | 0 | 0 | 0 |
| 6 | 125 ⊢ 150 | 0 | 1 | 1 |
| Total | | 14 | 6 | 10 |

O "jogo" de aproximações usado no cálculo de **A**, **K** e do início do primeiro intervalo de classe é aprimorado com a prática do pesquisador.

Os *softwares* estatísticos montam tabelas semelhantes a estas, porém não têm a sensibilidade de fazer sozinhos estes ajustes e corremos o risco de obter algo assim:

| Classe | Intervalo ($x_i$) ($) |
|---|---|
| 1 | 2,00 ⊢ 23,71 |
| 2 | 23,71 ⊢ 45,42 |
| 3 | 45,42 ⊢ 67,13 |
| 4 | 67,13 ⊢ 88,84 |
| 5 | 88,84 ⊢ 110,55 |
| 6 | 110,55 ⊢ 132,26 |

Para K = 6, A = 21,71 e iniciando a primeira classe em $2,00 repare nos valores de difícil observação, embora não haja qualquer erro do ponto de vista estatístico.

No entanto, é possível criar alguns critérios para os softwares que nos permita elaborar tabelas de distribuição de frequências condizentes com as operações que realizamos acima.

Vamos nos basear em um novo exemplo para elaborar uma Distribuição de Frequências pelo Excel. A tabela abaixo apresenta os valores (em R$ mil) dos totais dos contratos de seguros de automóveis fechados pela Corretora Mega Seguro durante o período de 60 dias.

| Dia | Valor em R$ | Dia | Valor em R$ | Dia | Valor em R$ | Dia | Valor em R$ | Dia | Valor em R$ | Dia | Valor em R$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 15 | 11 | 16 | 21 | 22 | 31 | 24 | 41 | 31 | 51 | 28 |
| 2 | 16 | 12 | 17 | 22 | 23 | 32 | 27 | 42 | 28 | 52 | 35 |
| 3 | 17 | 13 | 18 | 23 | 15 | 33 | 29 | 43 | 25 | 53 | 35 |
| 4 | 18 | 14 | 18 | 24 | 16 | 34 | 29 | 44 | 35 | 54 | 31 |
| 5 | 19 | 15 | 19 | 25 | 17 | 35 | 34 | 45 | 25 | 55 | 32 |
| 6 | 20 | 16 | 20 | 26 | 18 | 36 | 26 | 46 | 36 | 56 | 35 |
| 7 | 20 | 17 | 20 | 27 | 19 | 37 | 34 | 47 | 30 | 57 | 35 |
| 8 | 22 | 18 | 20 | 28 | 20 | 38 | 31 | 48 | 37 | 58 | 32 |
| 9 | 23 | 19 | 21 | 29 | 21 | 39 | 25 | 49 | 33 | 59 | 34 |
| 10 | 16 | 20 | 21 | 30 | 22 | 40 | 34 | 50 | 28 | 60 | 24 |

O primeiro passo consiste em elaborar as classes, o que pode ser feito seguindo as regras já citadas, ou até mesmo a experiência do pesquisado em dispor os valores analisados. Para o exemplo acima, a distribuição terá cinco classes, sendo que a amplitude de cada classe será de R$ 5 mil, como está descrito abaixo:

| Classes | $x_i$ (R$ mil) |
|---|---|
| 1ª | 15 ⊢ 20 |
| 2ª | 20 ⊢ 25 |
| 3ª | 25 ⊢ 30 |
| 4ª | 30 ⊢ 35 |
| 5ª | 35 ⊢ 40 |
| Total | |

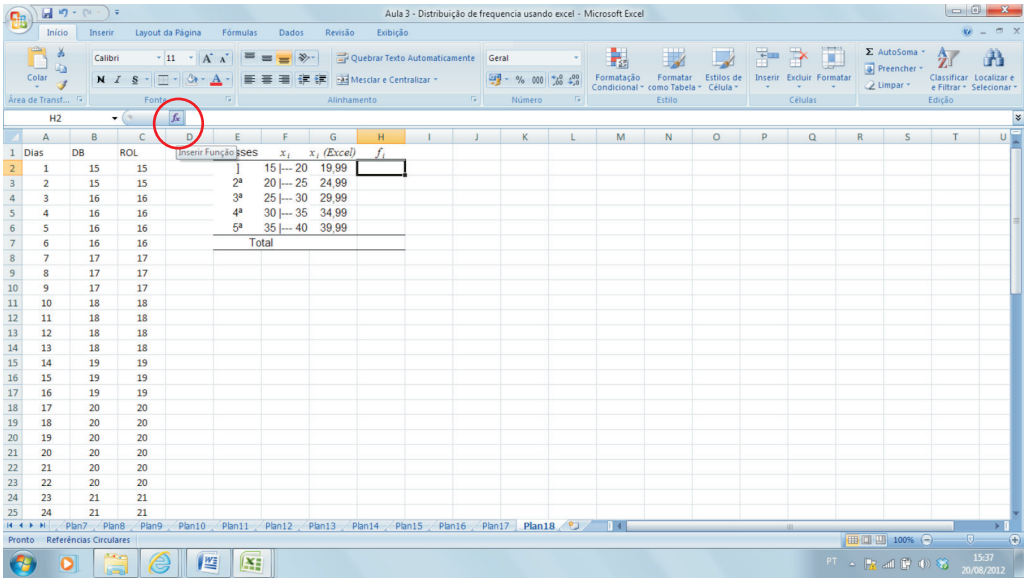
Uma vez definidas, manualmente, as classes, já é possível informar quais os parâmetros para o Excel, para que o software possa fazer os cálculos que nos permitirão elaborar uma distribuição de frequências.

Cabe lembrar que tais recursos servem para facilitar o nosso trabalho, sobretudo quando estamos trabalhando com um grande número de dados observados (500, 1.000, 10.000), o que tornaria a elaboração manual de uma tabela de distribuição bastante trabalhosa e passível de vários erros. Com o Excel, uma vez definidas as classes e a amplitude de cada classe, essa busca será rápida e precisa.

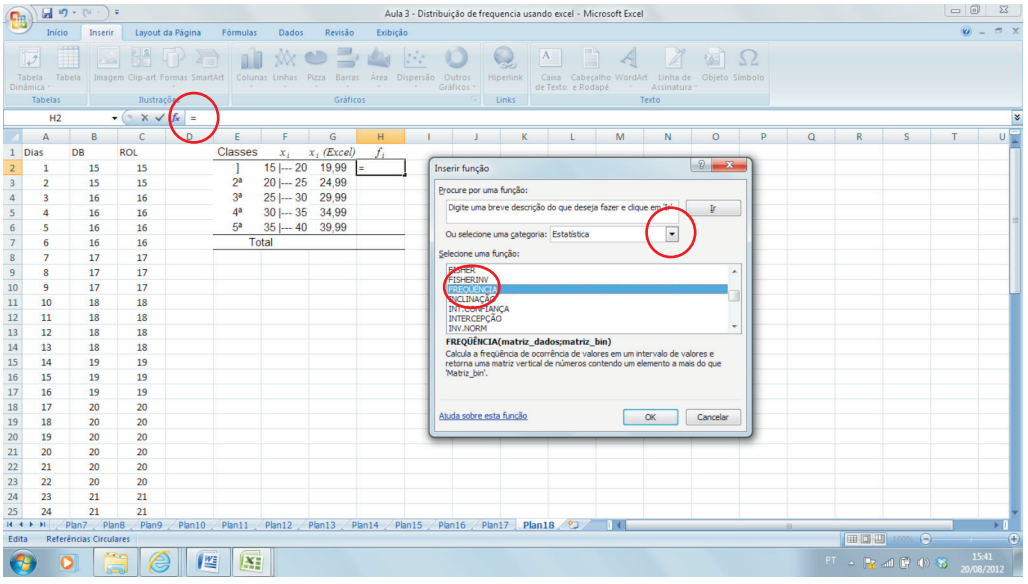
Definidos esses parâmetros, precisamos informá-los levando em consideração apenas um valor numa nova coluna, uma vez que o Excel não lê os intervalos como estão descritos acima. O novo valor indicado representa sempre o intervalo superior da classe, sendo que o valor informado também será considerado para aquela classe. Dessa forma, como a nossa distribuição acima exclui o valor superior, devemos informar que o limite superior da primeira classe será 19,99, e assim consecutivamente para as demais classes, conforme está descrito na tabela abaixo:

| Classes | $x_i$ (teórico) | $x_i$ *(Excel)* |
|---|---|---|
| 1ª | 15 ⊢ 20 | 19,99 |
| 2ª | 20 ⊢ 25 | 24,99 |
| 3ª | 25 ⊢ 30 | 29,99 |
| 4ª | 30 ⊢ 35 | 34,99 |
| 5ª | 35 ⊢ 40 | 39,99 |
| **Total** | | |

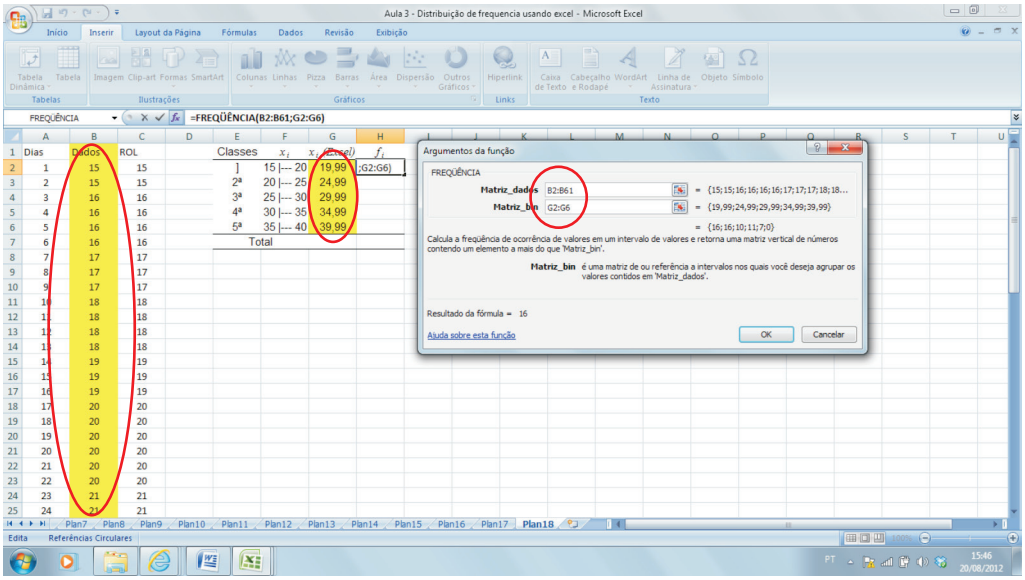
Para que o Excel faça o cálculo das frequências devemos numa nova coluna, da frequência simples, inserir a função estatística Frequência, conforme está ilustrado abaixo:



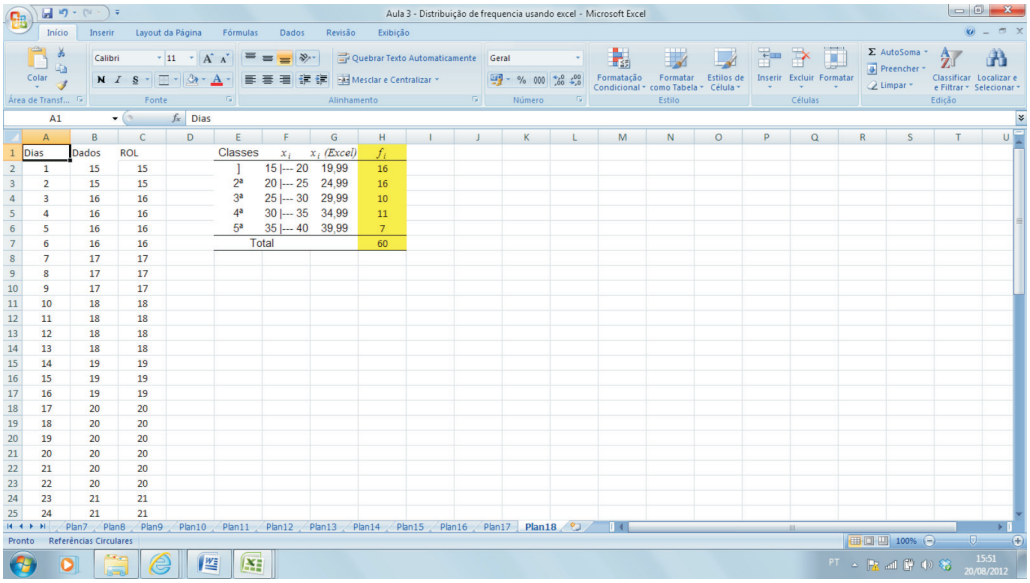

Ao clicarmos na função, abrirá uma caixa, para a qual deveremos informar a Categoria (Estatística) e a função (Frequência), conforme segue:



Em seguida, abrirá uma nova caixa solicitando a matriz dos dados (dados brutos) e a Matriz-bin (as classes):




Clique Ok. Mas ainda não terminou. Agora selecione a coluna da frequência, clique F2 (vai aparecer a fórmula da frequência na célula), e em seguida clique ***Ctrl Shift* e *Enter***** ao mesmo tempo. Atenção, não clique apenas *Enter*, é preciso utilizar o comando Ctrl Shift e *Enter* para replicar a fórmula para as demais células da frequência. Realizada essa etapa, o Excel distribuirá os dados nas classes previamente definidas, como segue:



As demais frequências poderão ser inseridas como fórmulas e replicadas para todas as classes, conforme segue:

| Classes | $x_i$ (Teórico) | $x_i$ (Excel) | $f_i$ | $F_i$ | $f_{i\%}$ | $Fi_{\%}$ |
|---|---|---|---|---|---|---|
| 1ª | 15 ⊢ 20 | 19,99 | 16 | 16 | 26,67% | 26,67% |
| 2ª | 20 ⊢ 25 | 24,99 | 16 | 32 | 26,67% | 53,33% |
| 3ª | 25 ⊢ 30 | 29,99 | 10 | 42 | 16,67% | 70,00% |
| 4ª | 30 ⊢ 35 | 34,99 | 11 | 53 | 18,33% | 88,33% |
| 5ª | 35 ⊢ 40 | 39,99 | 7 | 60 | 11,67% | 100,00% |
| Total | | | 60 | – | 100,00% | – |

Lembre-se, qualquer dúvida tecle F1 e abrirá uma janela de Ajuda do Excel!

# TÉCNICAS GRÁFICAS PARA REPRESENTAR DADOS

Gráficos têm capital função na representação dos dados e variáveis em um estudo estatístico. Alguns modelos comuns costumam atender à ampla maioria das necessidades usuais de qualquer estudo.

Aliás, uma indicação fundamental, que ressaltaremos permanentemente neste capítulo, é optar pela utilização de modelos básicos e tradicionais, objetivando a simplicidade da comunicação da informação. A seguir, vamos apresentar alguns dos tipos de gráficos utilizados com maior frequência.

## Gráfico de Colunas

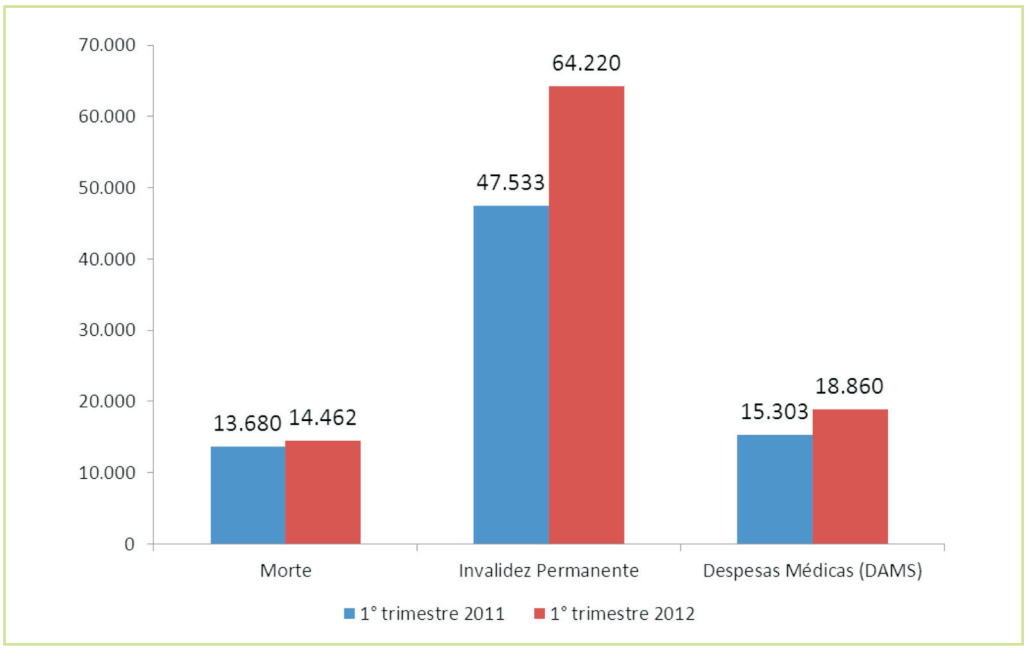
É o tipo de gráfico utilizado para a apresentação de séries cronológicas, categóricas (atributos) e de localização (departamentos, regiões).

Algumas dicas para a utilização desses dos gráficos de coluna:

- Sugere-se não utilizar esse tipo de gráfico para um número muito grande de observações e/ou variáveis.

- Tal qual o histograma, sugere-se a utilização de dois eixos (principal e secundário), quando as unidades das variáveis analisadas são diferentes. **Exemplo**: números absolutos e porcentagem.

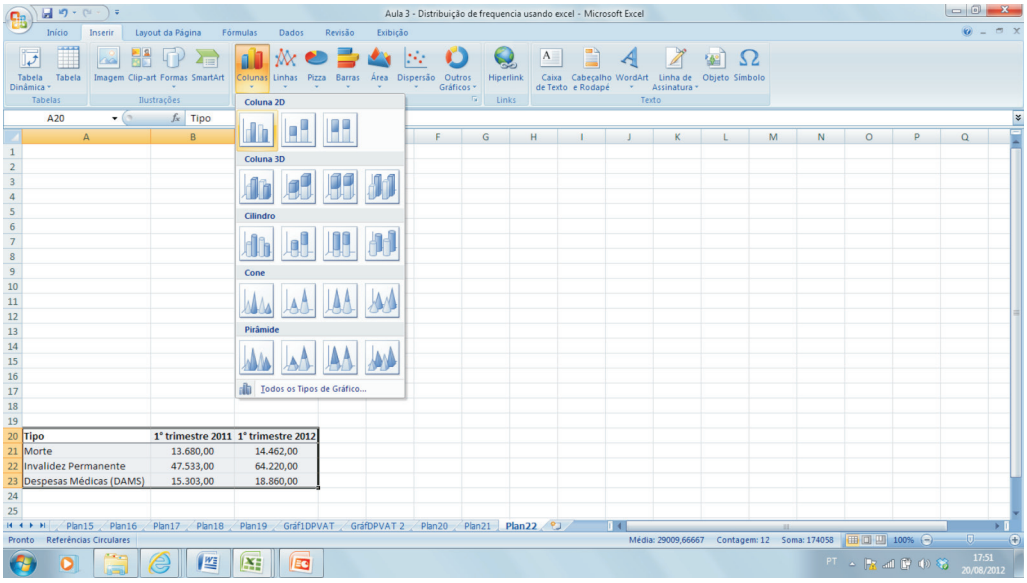
**Evolução das Indenizações do DPVAT Pagas por Natureza – Quantidade**



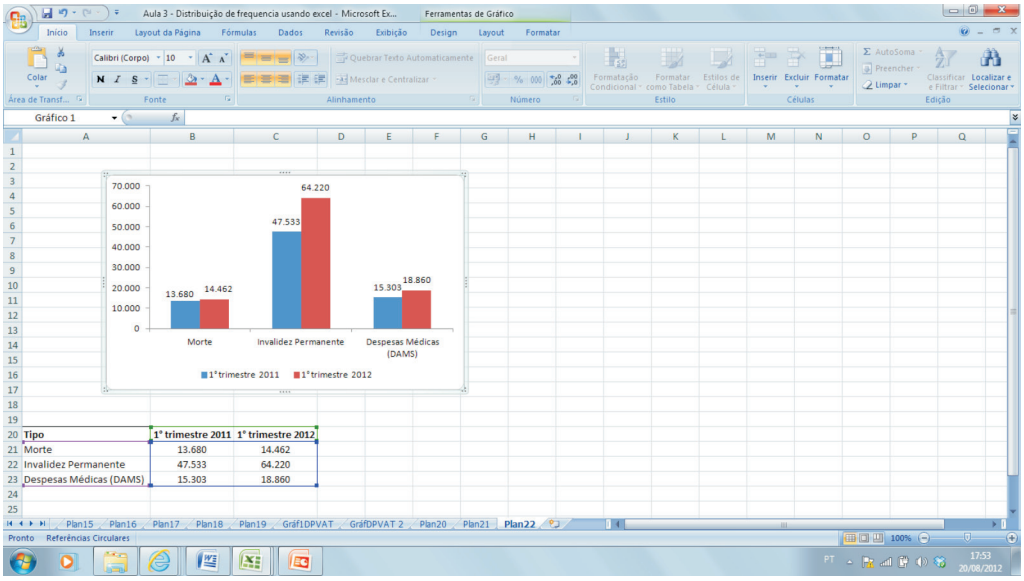

Fonte: Líder Seguradora – DPVAT

Agora vamos elaborar um gráfico de colunas no Excel. Para tanto, vamos utilizar o exemplo acima.

Selecione as células que contêm os dados do exemplo. Clique em Inserir, em seguida em Gráfico. Aparecerão várias opções de gráficos, clicar no Gráfico de Colunas, em seguida mais uma série de opções, agora só de gráficos de coluna será ofertada. Escolha aquele que desejar e for adequado ao trabalho que está realizando.

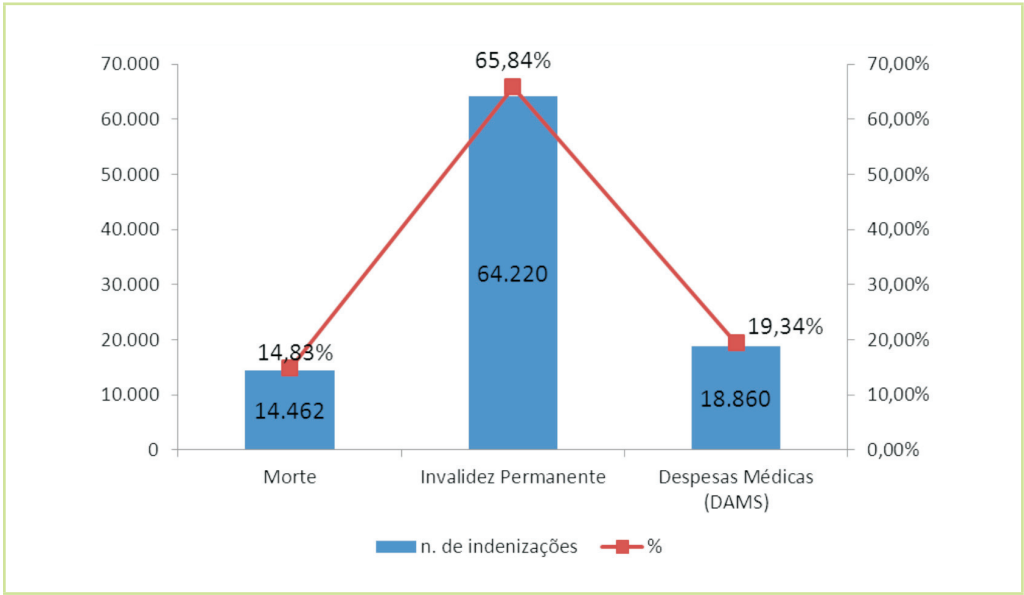


Basta escolher o tipo de gráfico de coluna desejado e clicar *enter* que o gráfico já estará pronto. Caso queira, poderá fazer uma série de ajustes para melhorar o gráfico elaborado.




No caso de duas unidades diferentes, o ideal é que cada unidade esteja associada a um eixo vertical, ou seja, deve trabalhar com um eixo vertical principal e um eixo vertical secundário, conforme o exemplo abaixo.

**Evolução das Indenizações Pagas por Natureza no 1º Trimestre de 2012 – Quantidade e %**

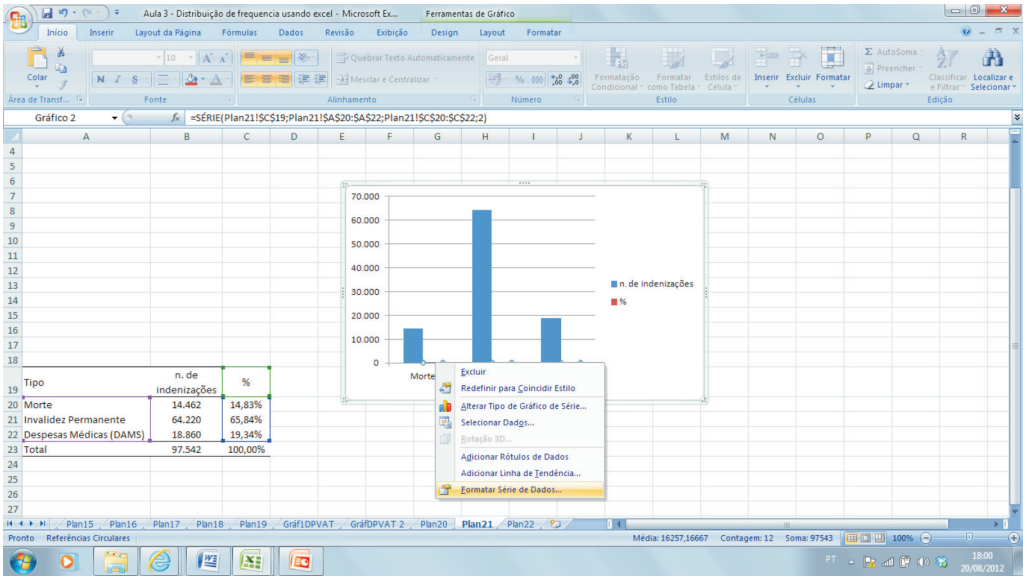



Fonte: Líder Seguradora – DPVAT

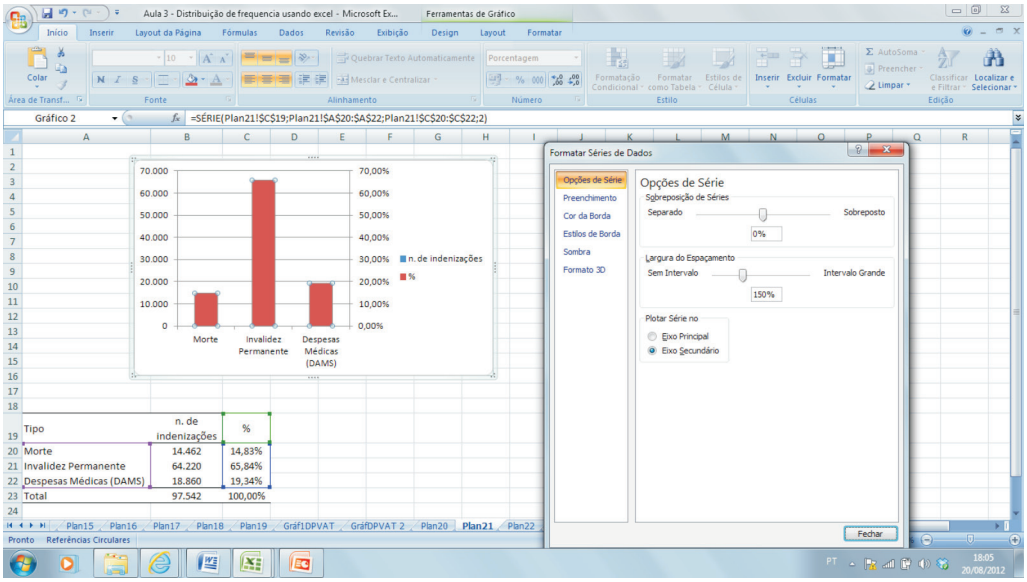
Para elaborar um gráfico de coluna e linha no Excel são necessários alguns "macetes". Tais como, criar uma coluna vertical secundária e mudar o tipo de gráfico para uma das variáveis analisadas.

Perceba no exemplo acima que o eixo vertical esquerdo está em quantidade de indenizações pagas, enquanto que o eixo vertical direito em porcentagem. Para elaborar esse gráfico no Excel serão necessários os seguintes procedimentos:

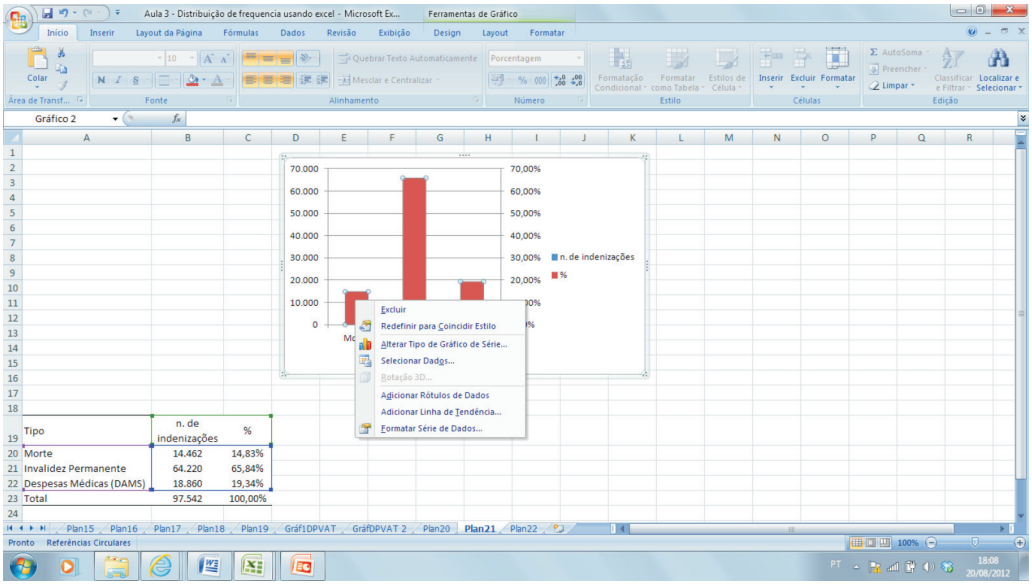
**1º)** clique com o botão direito do mouse sobre a coluna da porcentagem, aberta a Janela, clique em Formatar Série de Dados.




A janela Formatar Série de Dados se abrirá, sendo que a primeira opção para formatação é a Opções de Série, escolher Plotar série no Eixo Secundário.

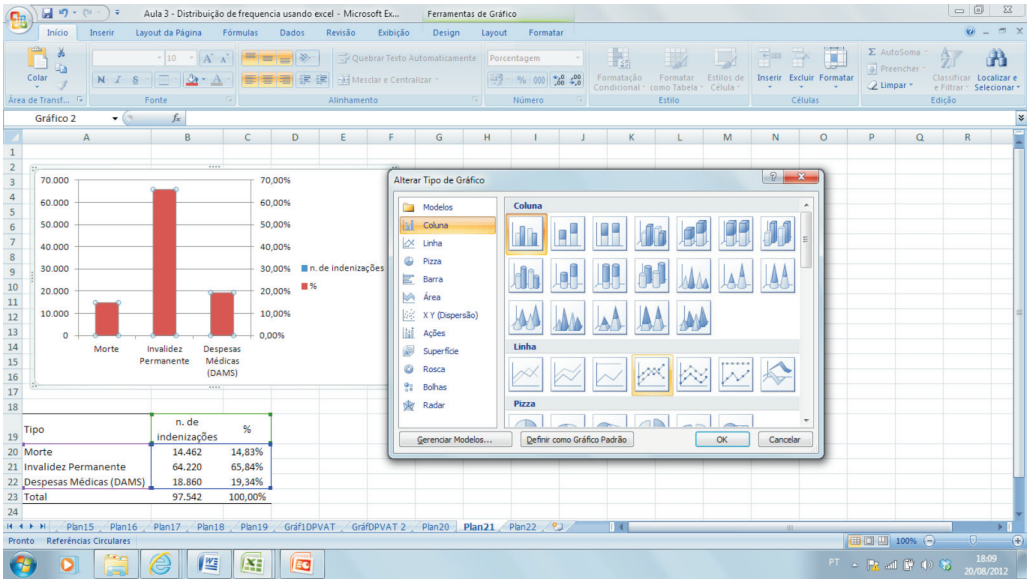


Perceba que a coluna da quantidade de indenizações sumiu. Para resolver isso, vamos mudar o tipo de gráfico da coluna %. Para tanto, clique com o botão direito do mouse sobre a coluna, e escolha Alterar tipo de gráfico na janela que se abrirá.




| Tipo | n. de indenizações | % |
|---|---|---|
| Morte | 14.462 | 14,83% |
| Invalidez Permanente | 64.220 | 65,84% |
| Despesas Médicas (DAMS) | 18.860 | 19,34% |
| Total | 97.542 | 100,00% |

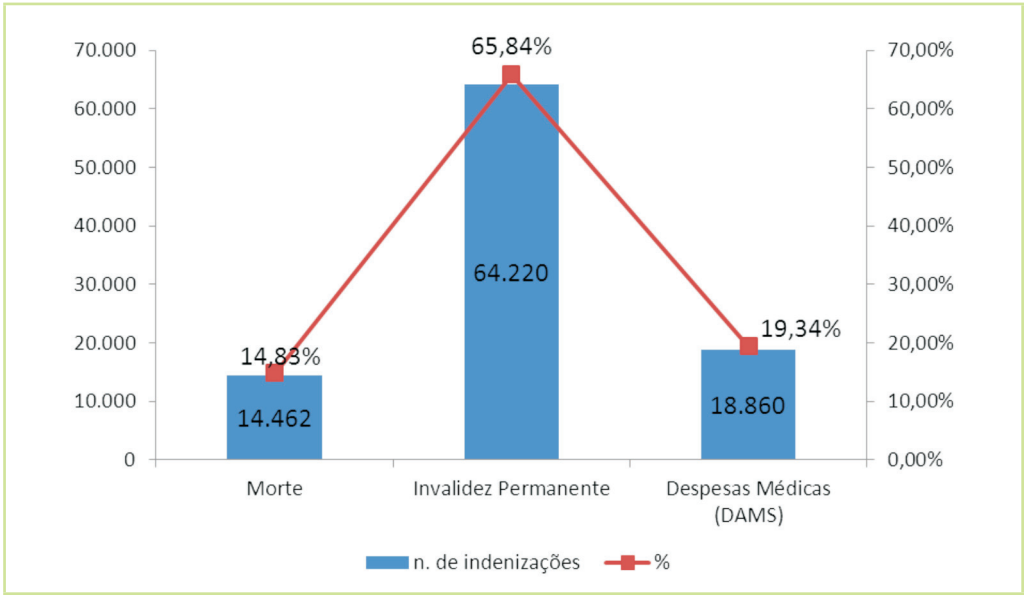
Escolha um gráfico de linha. Automaticamente as colunas e as linhas aparecerão novamente, sendo que poderemos ainda fazer mais alguns ajustes para deixar o gráfico ainda mais informativo.




| Tipo | n. de indenizações | % |
|---|---|---|
| Morte | 14.462 | 14,83% |
| Invalidez Permanente | 64.220 | 65,84% |
| Despesas Médicas (DAMS) | 18.860 | 19,34% |
| Total | 97.542 | 100,00% |

Uma dica importante, desde que a série estatística não seja muito longa é adicionar os rótulos (valores das variáveis) no gráfico. Para tanto, basta clicar com o botão direito do mouse sobre a coluna e sobre a linha, quando houver, e solicitar Adicionar rótulo de dados, e o gráfico ficará com a seguinte aparência:

**Evolução das Indenizações Pagas por Natureza no 1º trimestre de 2012 – Quantidade e %**



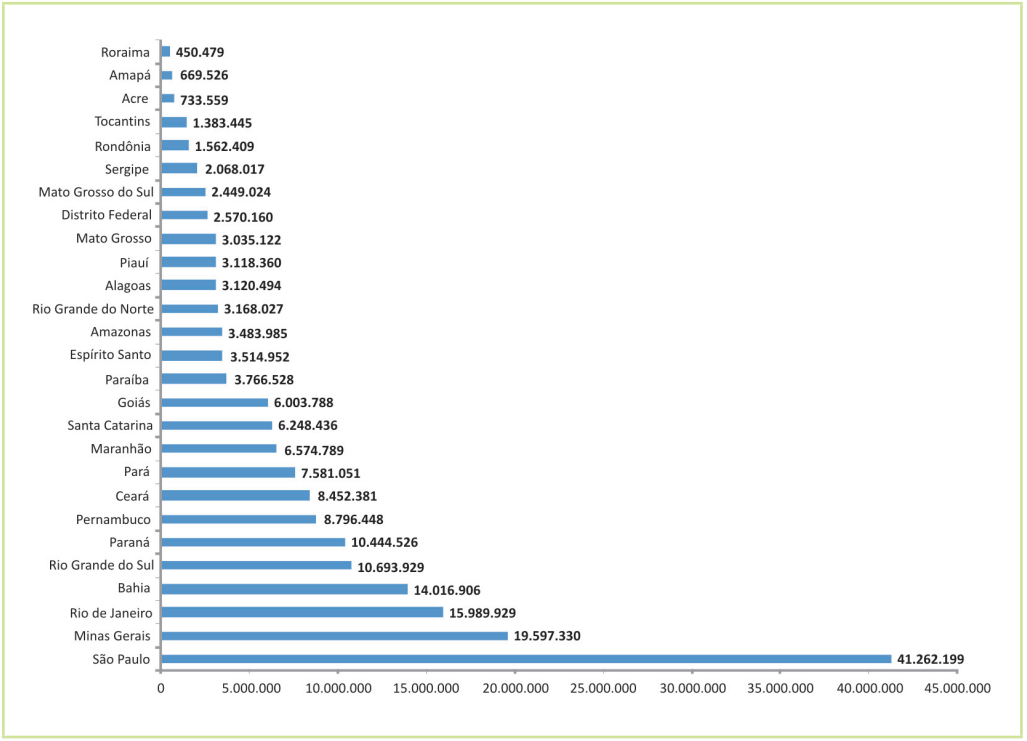

Fonte: Líder Seguradora – DPVAT

## Gráfico de Barras Horizontais

Assim como o gráfico de colunas, o gráfico de barras geralmente é utilizado para descrever um conjunto de dados que precisam evidenciar as magnitudes das variáveis analisadas. A diferença é a disposição da identificação das variáveis, que por serem extensas, ficam melhor dispostas num gráfico de barras horizontais.

Sugere-se não utilizar esse tipo de gráfico para um número muito grande observações e/ou variáveis. E, sempre que possível, informe os rótulos com os dados das variáveis para tornar a informação ainda mais visível. Veja o exemplo a seguir:

**População Residente no Brasil – por Unidade da Federação – Censo 2010**




Fonte: Censo 2010 – IBGE

Para elaborar o gráfico acima no Excel siga os mesmos passos do exemplo de Gráfico em Colunas, exceto que deverá escolher a opção de gráfico em barras.

## Histograma

É utilizado para apresentação de dados de uma tabela de distribuição de frequências. Não deixa de ser um gráfico de colunas, porém com algumas características peculiares. São representadas as variáveis ou intervalos de classes no eixo horizontal e as frequências no eixo vertical.

Além da visualização do fenômeno, histogramas indicam também características da distribuição que serão úteis na modelagem matemática para inferências, previsões e projeções sobre os fatos envolvidos no estudo. Uma vez construído o gráfico (histograma), adiciona-se a ele curvas que indicam comportamento das variáveis envolvidas, ou seja, mostram uma imagem tendencial do fenômeno.

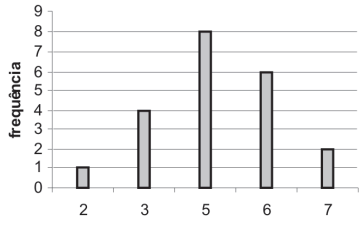
**Para uma variável discreta**:

- eixo horizontal (x) representa os valores da série
- eixo vertical (y) representa os valores das frequências
- a representação usa segmentos de reta verticais ou simplesmente pontos

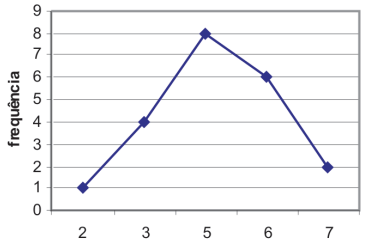
O rol **2 3 3 3 3 5 5 5 5 5 5 5 5 6 6 6 6 6 6 7 7** resulta em uma tabela de distribuição de frequência com o seguinte aspecto:

| $x_i$ | $f_i$ |
|---|---|
| 2 | 1 |
| 3 | 4 |
| 5 | 8 |
| 6 | 6 |
| 7 | 2 |

As possíveis representações gráficas abaixo podem ser utilizadas separadamente ou associadas (sobrepostas) em um mesmo gráfico:



Histograma simples



Polígono das frequências

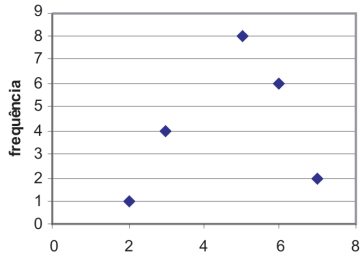
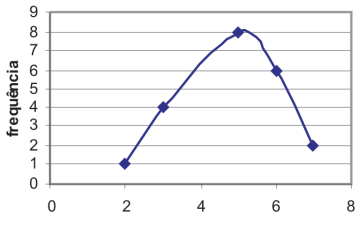


Gráfico de dispersão

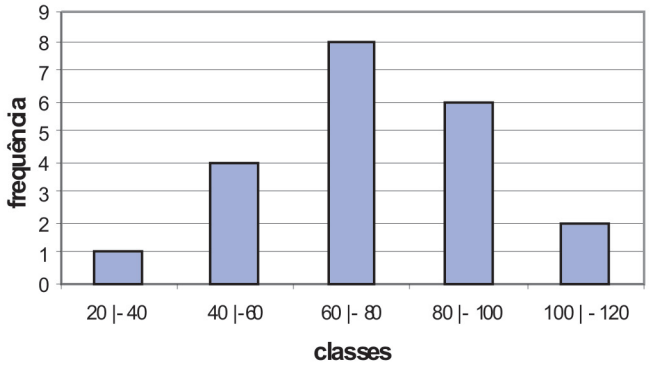


Curva polida das frequências ou ogiva

**Para uma variável contínua:**

- eixo horizontal (x) representa os intervalos (das classes) da série;
- eixo vertical (y) representa os valores das frequências; e
- a representação usa colunas, com base inferior de largura correspondente ao intervalo da classe (as colunas podem ser justapostas).

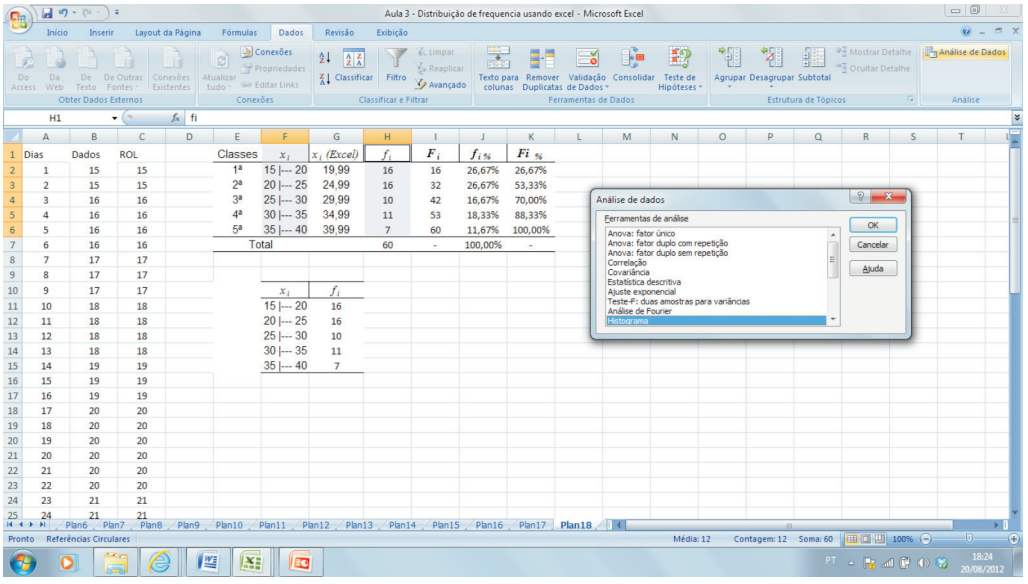
| Classes | $f_i$ |
|---|---|
| 20 \|- 40 | 1 |
| 40 \|-60 | 4 |
| 60 \|- 80 | 8 |
| 80 \|- 100 | 6 |
| 100 \| - 120 | 2 |



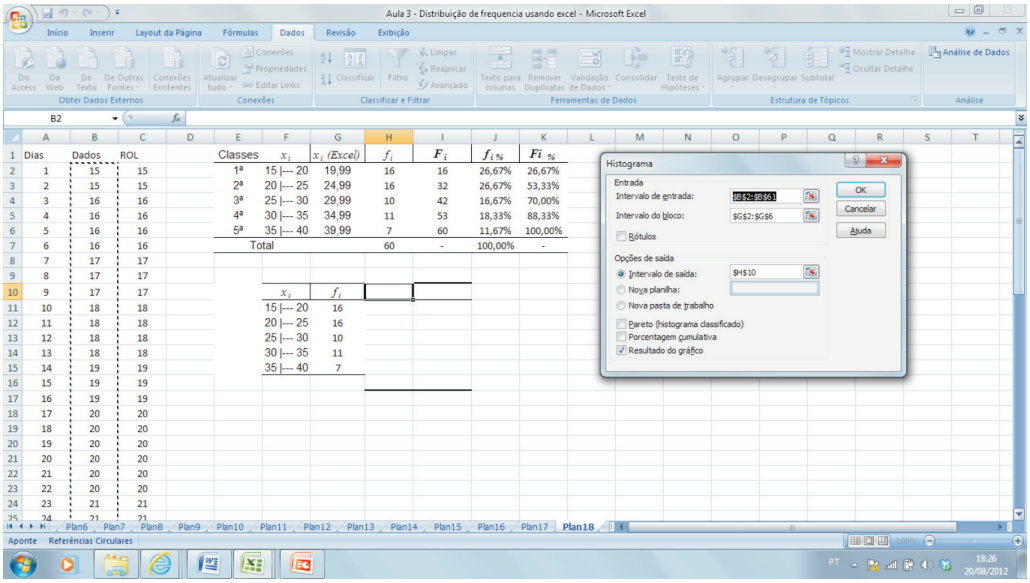

Num histograma para variáveis discretas, o polígono das frequências é uma linha que une os pontos médios das bases superiores das colunas. A área do polígono é igual à área do histograma (colunas).

Vamos nos basear no exemplo acima para elaborar um Histograma no Excel. No caso do Histograma o Excel possui uma Ferramenta de Análise que nos permite não apenas elaborar uma tabela de Distribuição de Frequências, igual àquela que já elaboramos, mas também preparar o gráfico correspondente, ou seja, o histograma da distribuição.

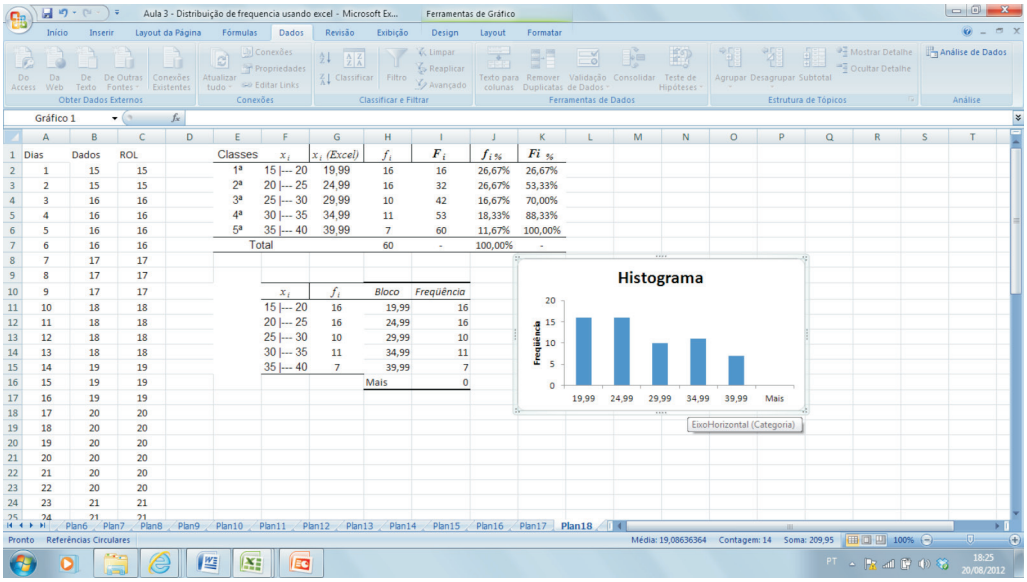
Para tanto, clique em Dados, e em seguida em Análise de Dados. Abrirá uma janela com diversas ferramentas de análise, escolha a opção Histograma.



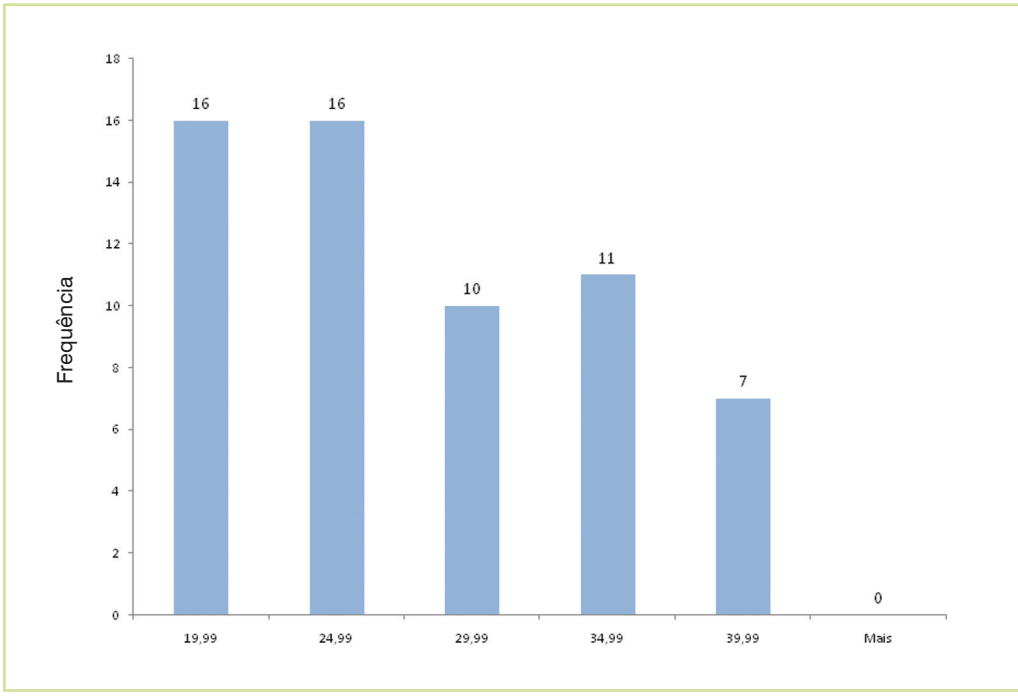
Em seguida, uma nova janela se abrirá solicitando algumas informações acerca da distribuição. A primeira delas é o Intervalo de Entrada, que corresponde aos dados (brutos), informe as células em que esses valores se encontram; em seguida, informe as células que compõem o Intervalo do Bloco (classes); o Intervalo de saída, que corresponde à célula em que a distribuição será apresentada; e, por fim, o resultado gráfico, que corresponde ao Histograma.




Ao finalizar essas operações clique *enter*, e o Histograma já está pronto.



Assim como demonstrado anteriormente, a apresentação do gráfico pode ser aprimorada para melhorar a compreensão do mesmo, basta explorar os diversos recursos do Excel.



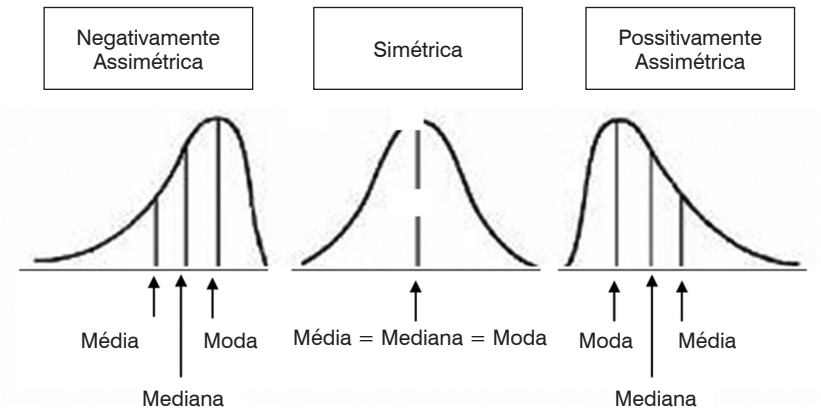

Assim como a tabela de distribuição de frequências, o histograma pode ser complementado a partir do cálculo das demais frequências (absolutas e relativas) já discutidas, lembrando que quando houver diferentes unidades é aconselhável a utilização de dois eixos verticais (o principal e o secundário).

## Características das Curvas de Frequências – Assimetria e Curtose

Com relação às curvas de frequências, merecem ser analisadas, duas medidas que também contribuem para descrever o comportamento de um conjunto de dados observados. São: o grau de assimetria e de curtose.

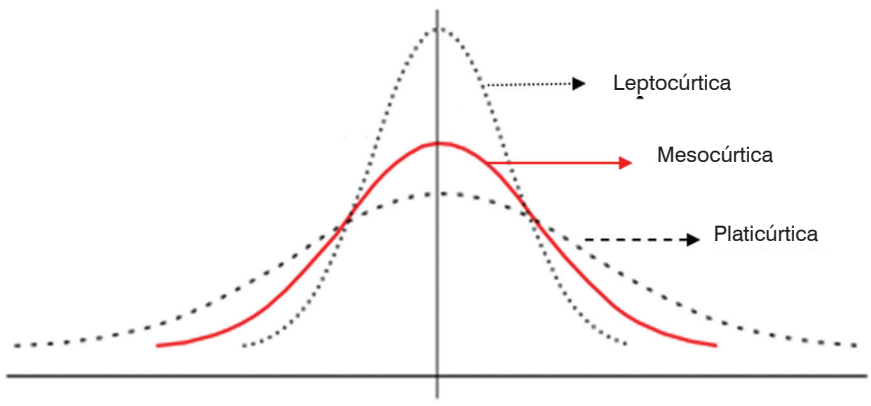
O grau de assimetria nos permite avaliar a dispersão dos dados observados em relação às medidas de tendência central. Quando não se registra tal dispersão, dizemos que a distribuição é simétrica, o que significa que a Média é igual à Mediana e à Moda, de tal modo que 50% dos dados observados estão exatamente abaixo dessas estatísticas e 50% acima, conforme mostra a figura a seguir.

**Assimetria**




Dando prosseguimento à avaliação do comportamento de uma distribuição, outra importante característica pode ser observada quanto à curva de frequência, que está associada ao seu "achartamento ou afilamento" , ou seja, seu grau de curtose. Por exemplo, uma distribuição pode apresentar elevado achatamento, o que mostra que os dados observados estão fortemente distribuídos ao longo da curva (distribuição platicúrtica).

**Curtose**



Existem técnicas estatísticas que nos permitem calcular tanto o grau de assimetria de uma curva de frequência quanto de Curtose. No entanto, nos restringiremos apenas às avaliações conceitual e gráfica dessas medidas.

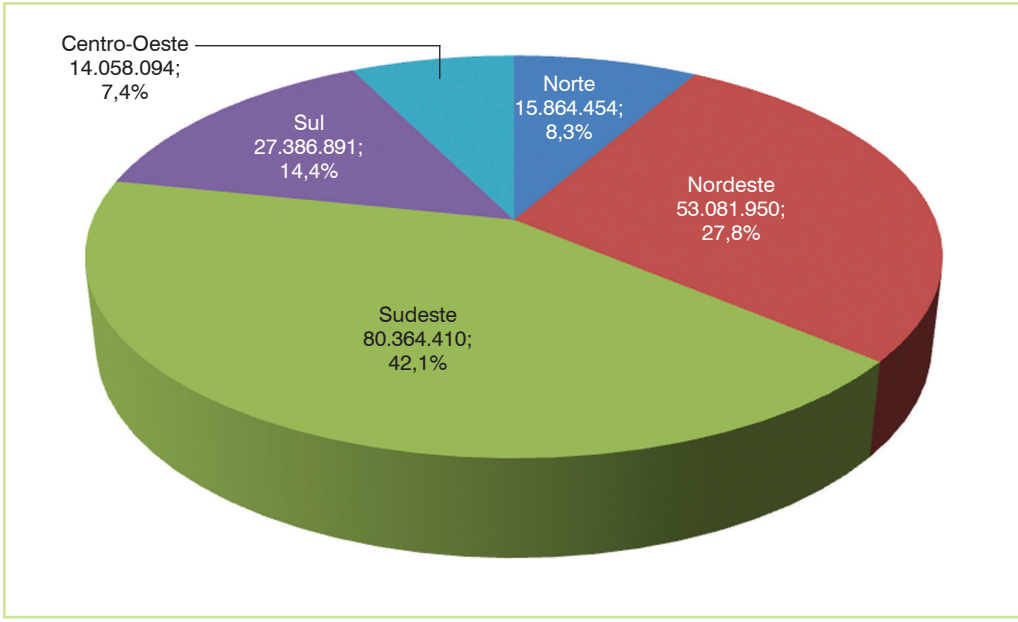
## Gráfico de Setores ou Pizza

É utilizado para mostrar e comparar a importância entre as várias proporções envolvidas em um estudo. Ao contrário dos demais gráficos, não permitem análises mais profundas do fenômeno, não serve para elaboração de projeções matemáticas, mas possui um bom apelo visual para apresentações visuais e gráficas.

Sugere-se para esse tipo de gráfico um número reduzido de observações, haja vista que seu objetivo é de propiciar uma imediata noção dos valores ou porcentuais.

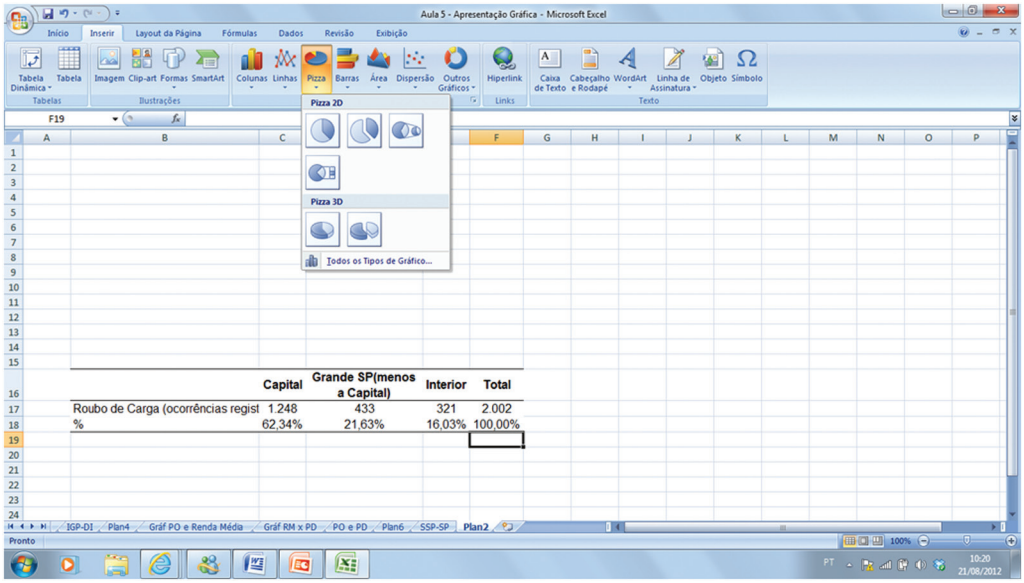
Também é possível apresentar, em um gráfico de setores, diferentes unidades de medida, como mostra o exemplo a seguir.

**População Residente no Brasil em 2010 – por Região – em N. de Habitantes e Distribuição %**

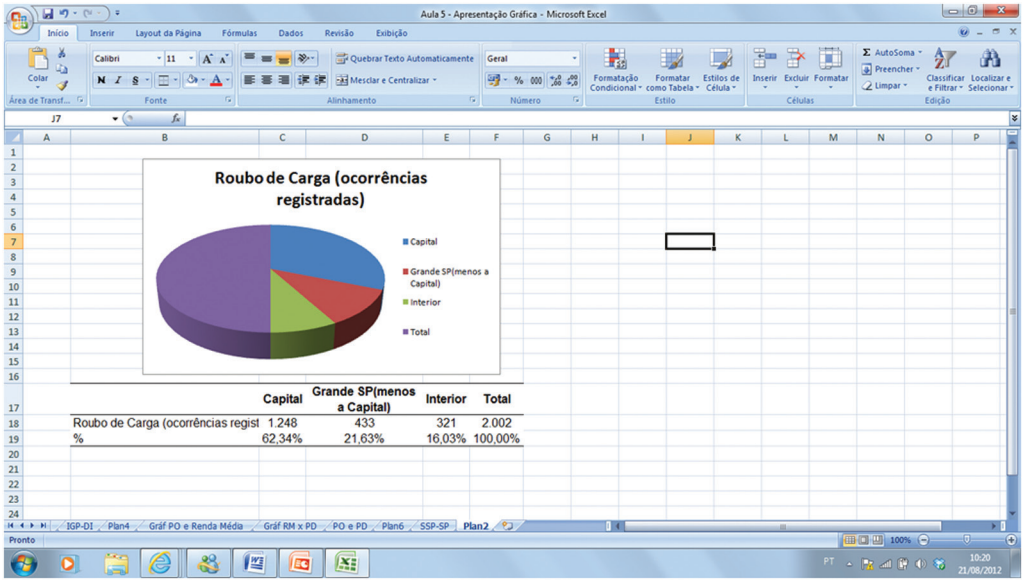



Fonte Censo – IBGE

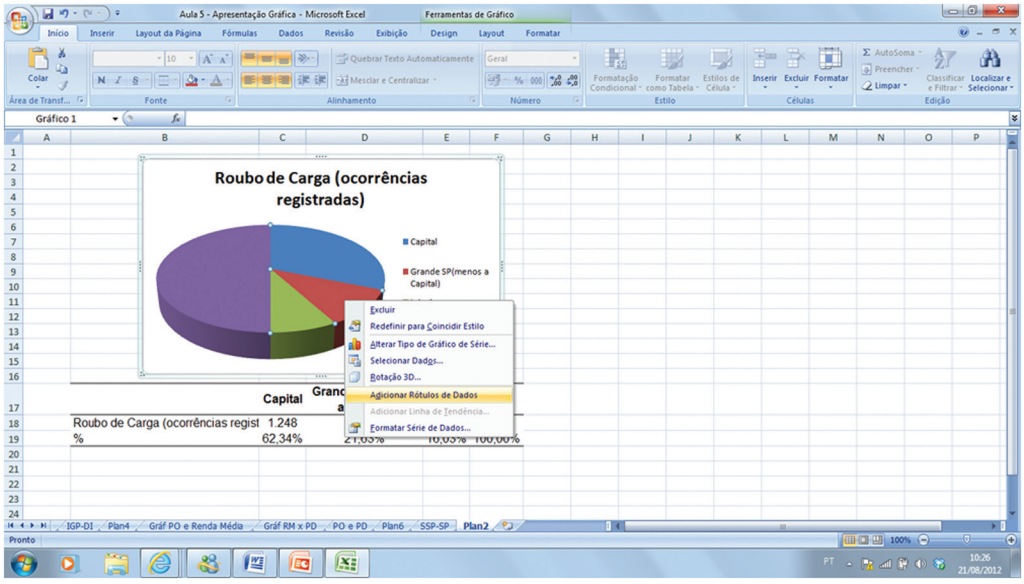
A elaboração de um gráfico de setores no Excel requer um procedimento semelhante aos demais gráficos, basta selecionar as células em que as informações estão dispostas e pedir para inserir gráfico de Pizza.

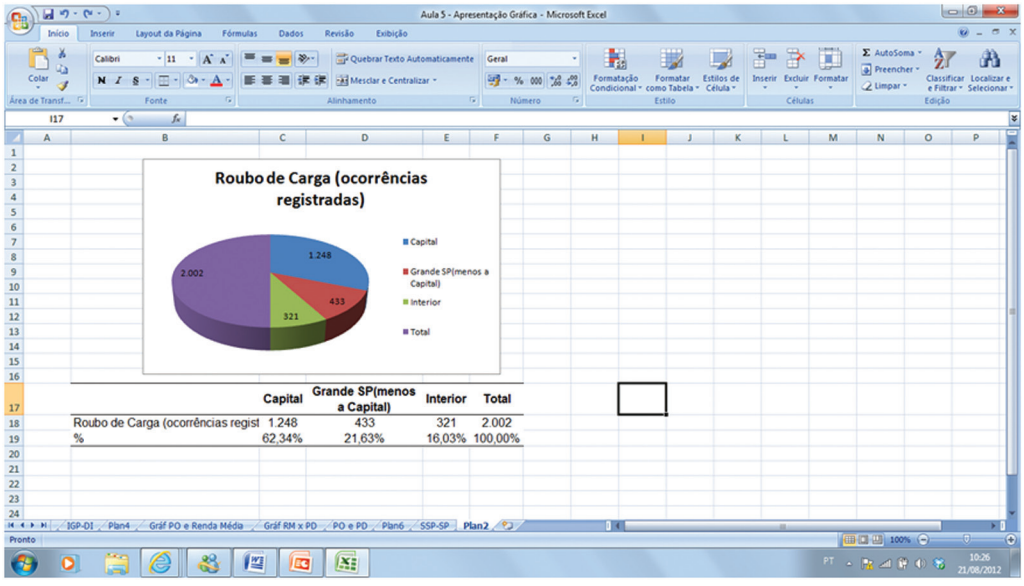



Basta clicar no gráfico selecionado e a operação será realizada. Uma série de melhorias no aspecto visual do gráfico poderá ser feita, apenas clique com o botão direito do mouse sobre o gráfico para efetuá-las.

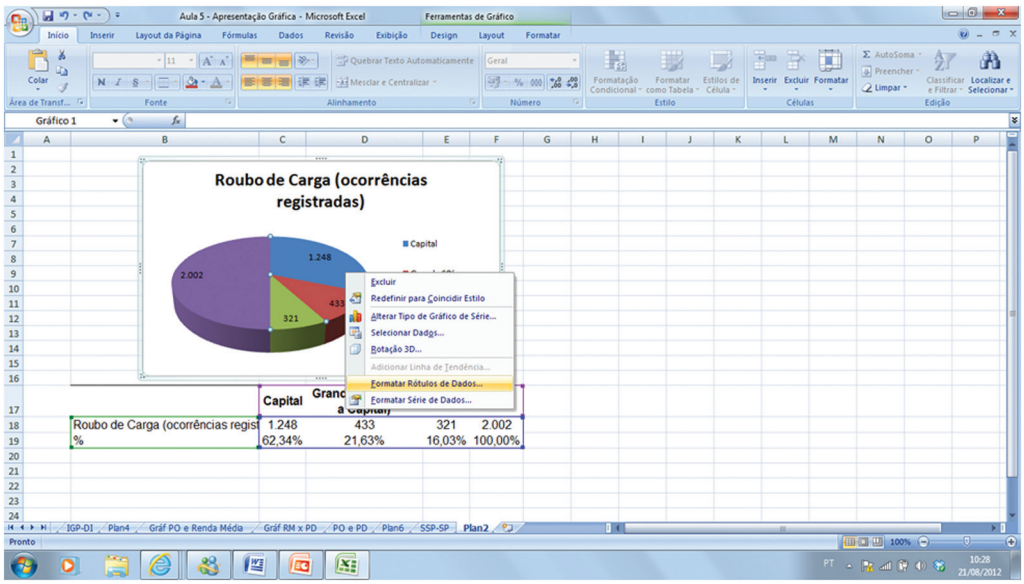


Uma das melhorias sugeridas é a inclusão dos valores da série no próprio gráfico. Para tanto, clique com o botão direito exatamente sobre o gráfico, ao abrir a janela escolha Adicionar Rótulo de Dados.

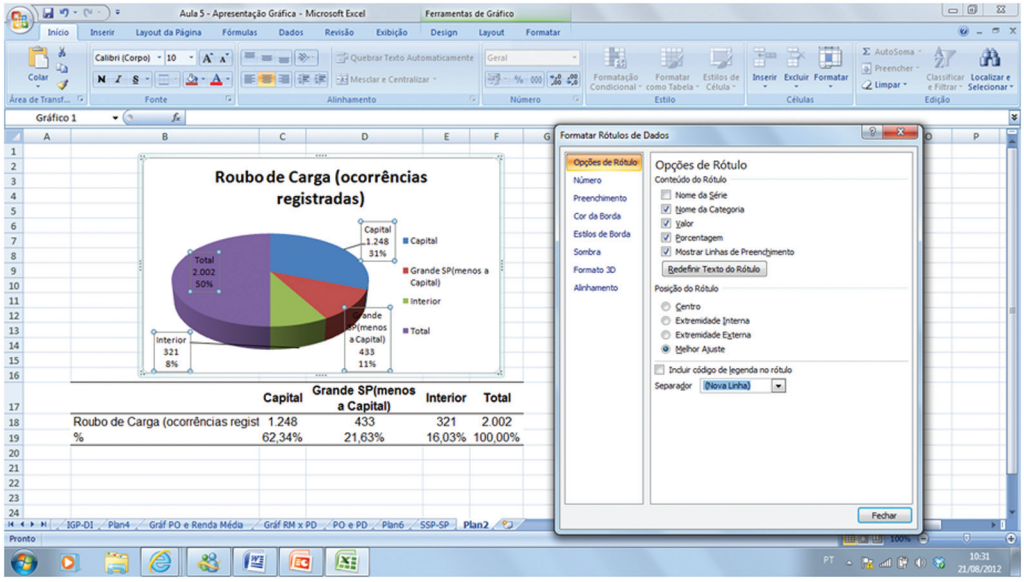


Mas ainda é possível incluir os percentuais. Para tanto, clique novamente sobre o gráfico e, em seguida, em Formatar Rótulos de Dados na janela aberta.




Ao abrir a Janela Formatar Rótulo de Dados, clique nas opções desejadas, sempre objetivando tornar ainda mais clara a interpretação dos resultados.




Busque sempre explorar todas as opções que o *software* nos fornece, e se não gostar do resultado é só desfazer a opção escolhida.

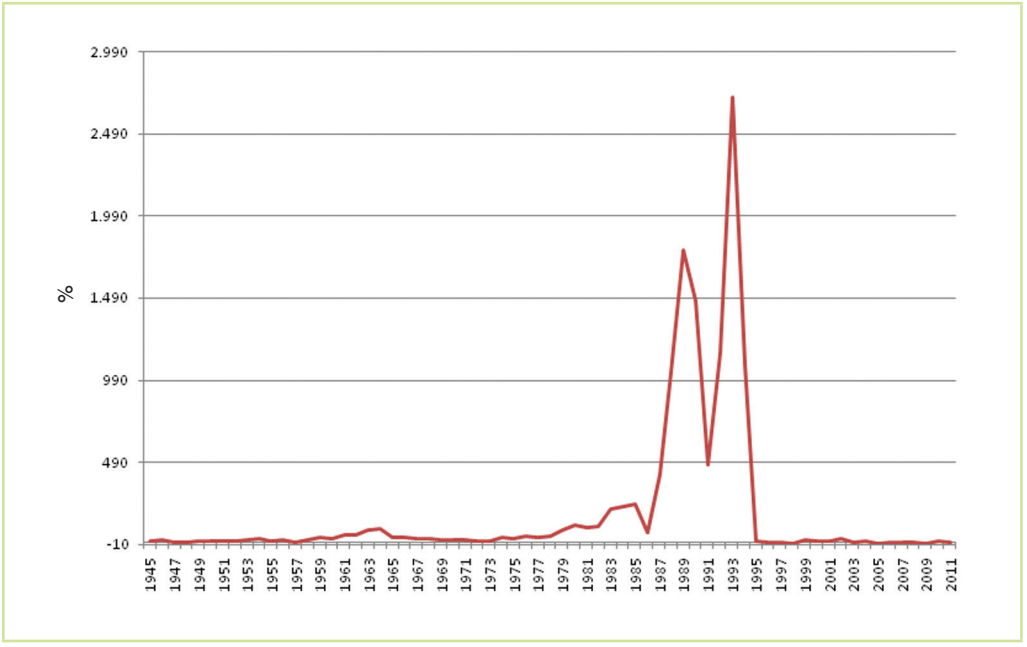
## Gráfico de Linhas

Geralmente utilizado para descrever séries temporais, principalmente quando são bastante longas. O gráfico de linhas, por descrever uma série temporal, permite que o leitor identifique a tendência do fenômeno analisado ao longo do tempo.

Sugere-se que se a variável apresentar uma amplitude muito grande de um momento para o outro, o gráfico seja separado de tal modo que a tendência possa ser mais visível.

No exemplo abaixo é possível ver uma situação em que uma grande amplitude não permite ver, em detalhes, o comportamento do índice de inflação a partir de 1994.

**Índice Geral de Preços – IGP-DI – 1945/2011 – em %**

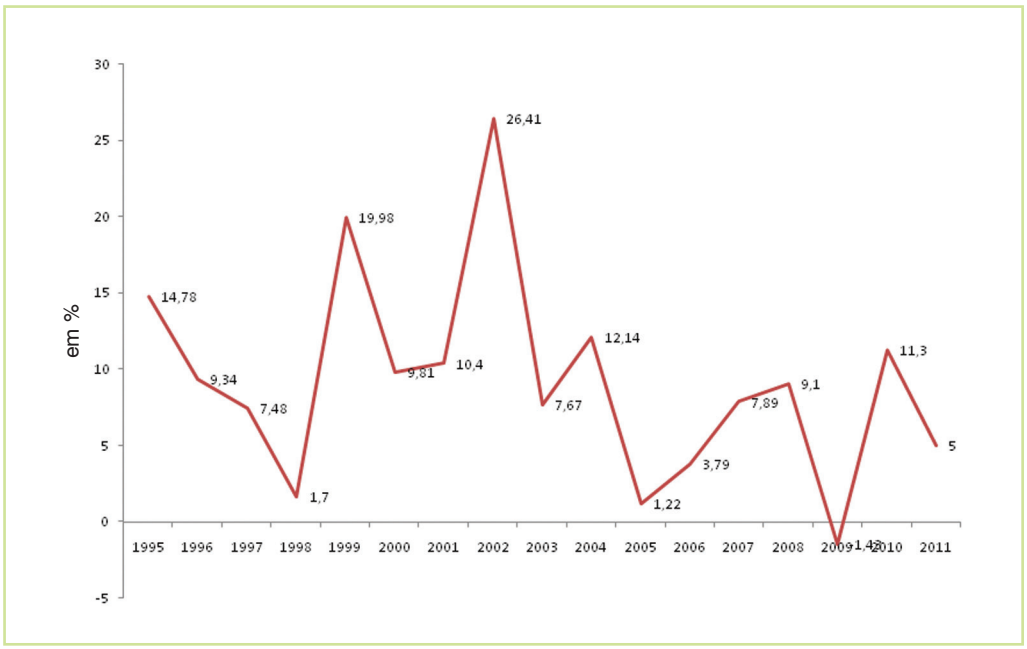



Fonte: FGV

Para esse caso reitera-se a sugestão de separar o gráfico, de tal modo que se possa detalhar o comportamento do aumento dos preços. Perceba, no exemplo abaixo, que o índide de inflação tem oscilado bastante após 1994. No gráfico anterior essas variações não eram visíveis, em função dos elevados índices registrados entre os anos de 1986 e 1995.

Assim como os demais gráficos, o gráfico de linha pode ser complementado com rótulos, bem como uma série de ajustes podem ser feitos para melhorar a apresentação e facilitar sua compreensão.

**Índice Geral de Preços – IGP-DI e Linha de Tendência (em %)**




Fonte: FGV

Recomenda-se, ainda, que para séries temporais curtas não seja utilizado o gráfico de linhas. Nesse caso, o gráfico de colunas será mais apropriado.

## Gráfico de Pontos ou Dispersão

É um tipo de gráfico geralmente utilizado para relacionar duas variáveis. Ademais, é considerada uma técnica relativamente simples de projeções e previsões de cenários futuros, uma vez que pode, complementarmente, se apoiar no uso de modelos matemáticos (regressão linear, múltipla etc).

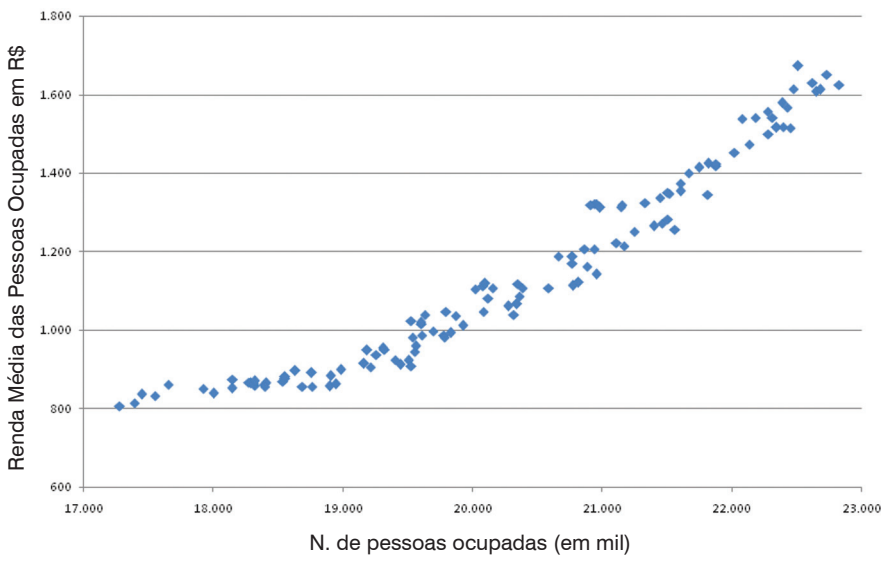
Apesar da existência de técnicas estatísticas mais acuradas para estudar a relação entre variáveis, o gráfico de dispersão é recorrentemente utilizado para mostrar o grau de associação entre duas variáveis.

No exemplo a seguir, busca-se relacionar a renda média dos trabalhadores com carteira assinada ao número de pessoas ocupadas. Ambas as variáveis são calculadas pela Pesquisa Mensal de Emprego do IBGE.

Observe no gráfico a seguir que é possível identificar um comportamento que expressa a relação entre as duas variáveis analisadas. Esse comportamento aponta para uma tendência de crescimento da renda dos trabalhadores associado ao crescimento do número de pessoas empregadas.

Perceba que os pontos estão bastante próximos uns dos outros, e seguem uma tendência crescente, corroborando a hipótese de associação entre renda e aumento do número de pessoas empregadas, que tem como contrapartida a redução do número de pessoas desempregadas.

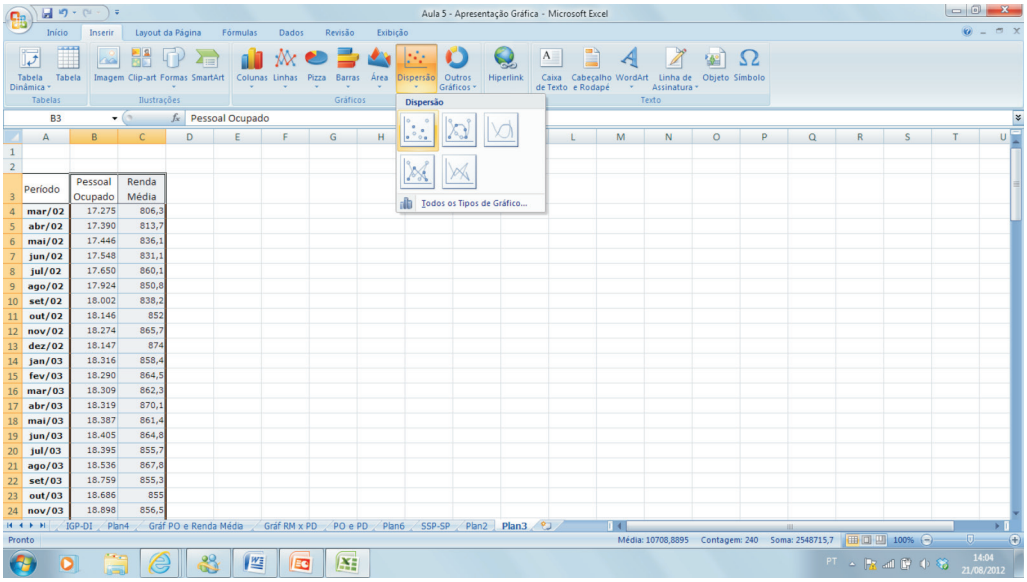
**Pessoas Ocupadas × Renda Média (R$) (PME/IBGE)**



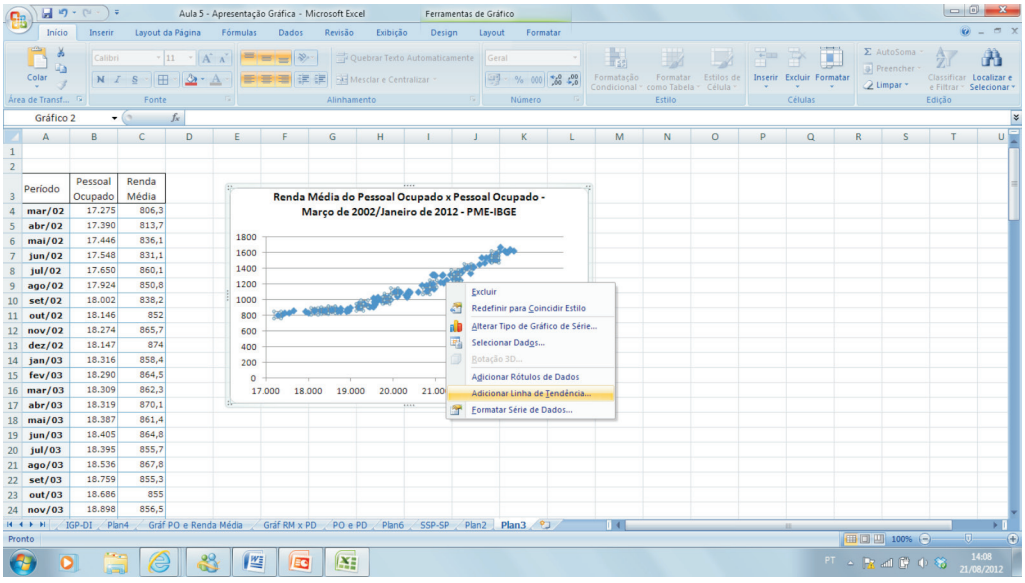

Para reforçar essa constatação, que é apenas visual, o Excel dispõe de um recurso adicional, que permite a inclusão de uma linha de tendência, que mede o grau de associação entre as duas variáveis. Para tanto, o software estima uma função matemática *Renda Média = f (Número de pessoas ocupadas)*, e mede o Coeficiente de Determinação – $R^2$ (que é o quadrado do Coeficiente de Correlação entre variáveis).

Vamos elaborar o gráfico do exemplo anterior, e incluir a linha de tendência para o gráfico acima.

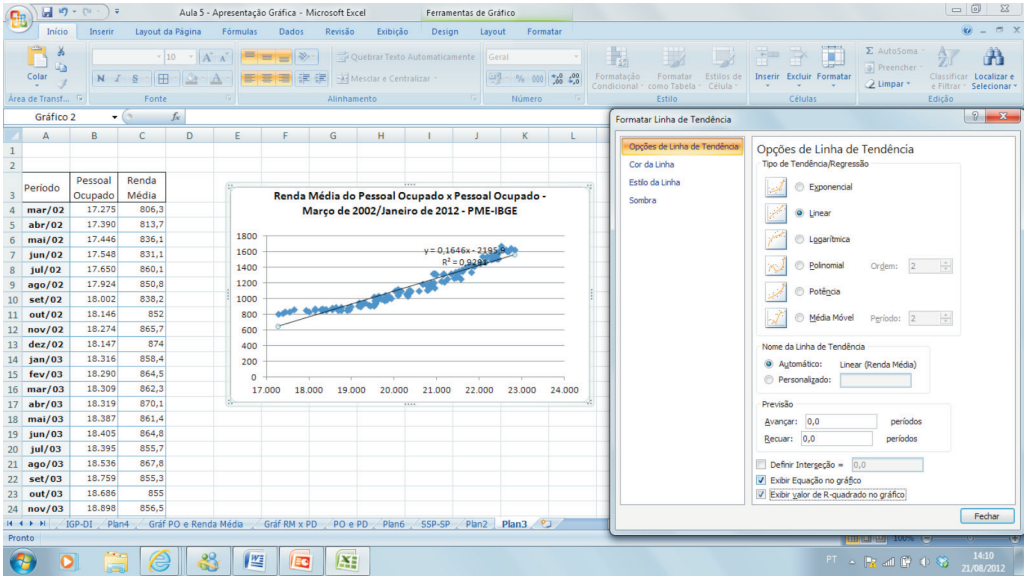
Para tanto, selecione as células em que se encontram os dados observados para as duas variáveis (apenas as colunas cujos dados serão utilizados para efeito de construção do gráfico. Observe que a coluna do período não fará parte do gráfico, logo não está sendo selecionada. Em seguida, Clique em inserir, Gráficos e escolha um dos Gráficos de Dipersão apresentados.




Uma vez elaborado o gráfico, clique com o botão direito do mouse sobre os dados e peça para Adicionar Linha de Tendência.



Em seguida uma janela, chamada Formatar Linha de Tendência se abrirá. Nessa janela será solicitada a escolha de tipo de Tendência (função matemática) que será utilizada, o que requer uma certa experiência por parte do pesquisador, ou que a mesma já seja definida previamente pelo estudo em questão. No nosso exemplo escolhemos uma tendência Linear. Várias outras informações são requeridas, vamos escolher apenas duas delas: que seja informada a Equação da linha de tendência e o Coeficiente de Determinação que mede o grau de associação entre as variáveis.

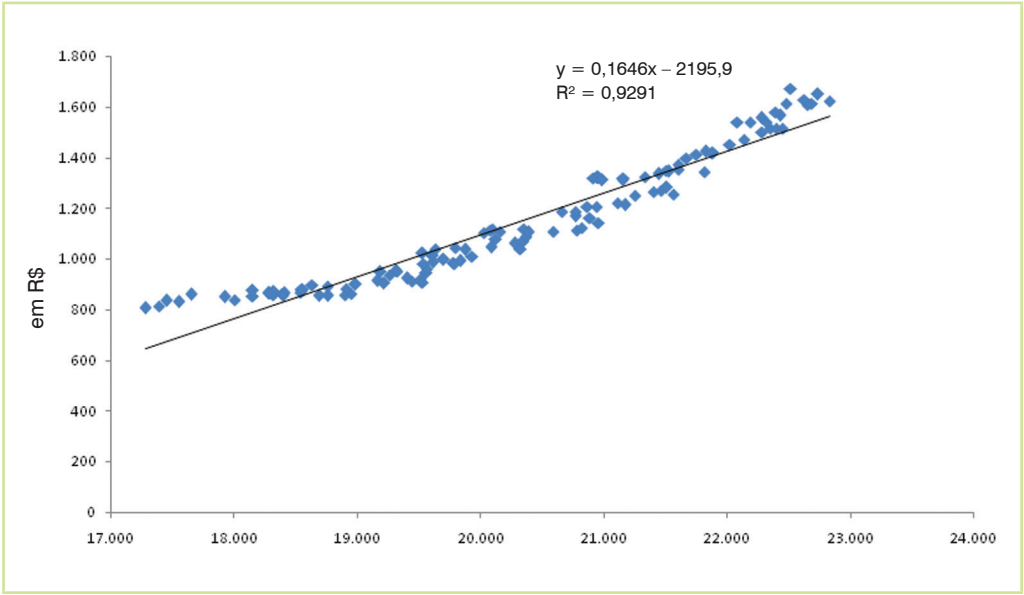



| Período | Pessoal Ocupado | Renda Média |
|---|---|---|
| mar/02 | 17.275 | 806,3 |
| abr/02 | 17.390 | 813,7 |
| mai/02 | 17.446 | 836,1 |
| jun/02 | 17.548 | 831,1 |
| jul/02 | 17.650 | 860,1 |
| ago/02 | 17.924 | 850,8 |
| set/02 | 18.002 | 838,2 |
| out/02 | 18.146 | 852 |
| nov/02 | 18.274 | 865,7 |
| dez/02 | 18.147 | 874 |
| jan/03 | 18.316 | 858,4 |
| fev/03 | 18.290 | 864,5 |
| mar/03 | 18.309 | 862,3 |
| abr/03 | 18.319 | 870,1 |
| mai/03 | 18.387 | 861,4 |
| jun/03 | 18.405 | 864,8 |
| jul/03 | 18.395 | 855,7 |
| ago/03 | 18.536 | 867,8 |
| set/03 | 18.759 | 855,3 |
| out/03 | 18.686 | 855 |
| nov/03 | 18.898 | 856,5 |

No gráfico a seguir já é possível identificar os resultados da inclusão da linha de tendência, que inclusive nos permite, a partir da Equação estimada, fazer projeções para o futuro. Por exemplo, ao supor-se que o número de pessoas empregadas seja de 25 milhões – substituir na variável × da equação – pode-se inferir que a renda média estimada será de R$1.919,1.

Ademais, deve-se ressaltar que o grau de associação entre as duas variáveis é bastante elevado, uma vez que o $R^2$ é 0,9291. Se $R^2$ for igual a 1, o grau de associação é considerado perfeito.

**Renda Média do Pessoal Ocupado × Pessoal Ocupado**
**Março de 2002/Janeiro de 2012 – PME-IBGE**



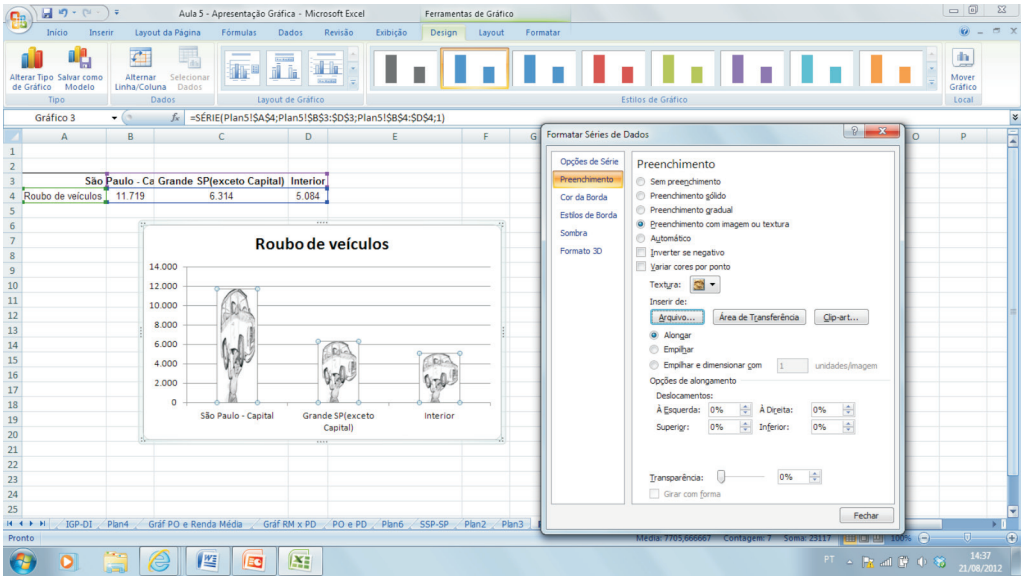

Apesar de bastante útil, essa ferramenta de análise estatística é limitada. Técnicas bem mais sofisticadas, e confiáveis, são utilizadas para a elaboração de estudos que precisam ter uma confiabilidade maior.

## Pictrogramas

Geralmente são gráficos de barras horizontais ou de colunas, mas com a diferença que as barras e colunas são substituídas por figuras, cujo tamanho procura reproduzir as proporções das variáveis analisadas.

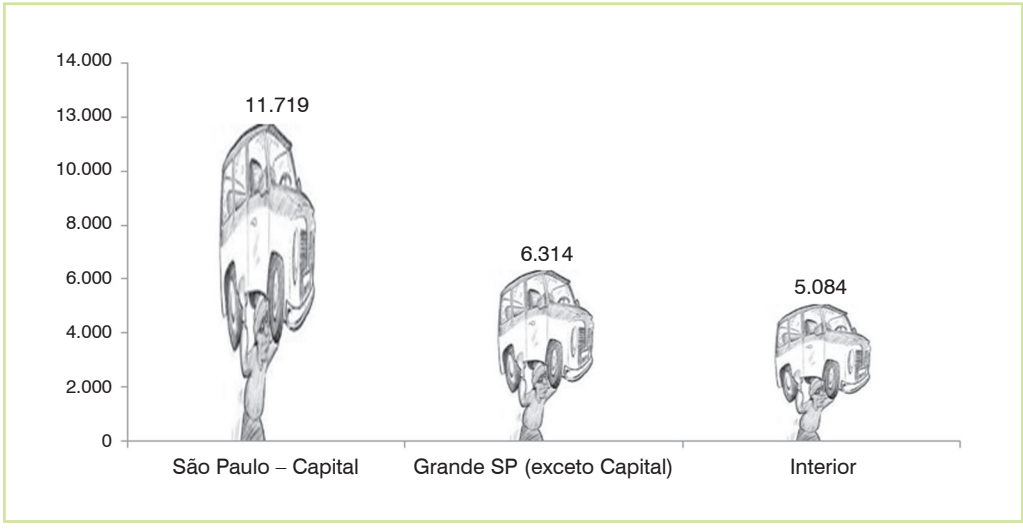
Esse tipo tem um objetivo mais visual do que técnico, e são mais apropriados para apresentações ou textos menos formais.

Para elaborar um Pictograma no Excel é necessário, inicialmente, criar um gráfico de colunas (ou de barras horizontais) primeiro. Uma vez criado o gráfico, clique com o botão direito do mouse sobre as colunas, e em seguida sobre Formatar Série de Dados. Escolha Preenchimento, e em seguida Preenchimento da Imagem ou Textura, em Arquivos escolha uma imagem previamente gravada e pronto, ela será anexada às colunas do gráfico.




No gráfico a seguir é possível ter uma melhor visualização do resultado final de um Pictograma.

**Roubo de Veículos – 1º Trimestre de 2012 – SSP/SP**




> Lembre-se, as figuras utilizadas num pictograma precisam ser previamente gravadas numa pasta, para que possam ser reproduzidas pelo Excel.

Bom, esses são os principais gráficos utilizados para complementar a tarefa de apresentar um conjunto de dados. Porém, outros tipos de gráficos podem ser encontrados tanto nos livros de estatística quanto no próprio Excel.

**Em caso de dúvidas, sobre o tipo de gráfico a utilizar, opte pelo gráfico de colunas**. Pode-se dizer que ele reúne todas as qualidades necessárias para o entendimento dos fenômenos estudados e possui atributos técnicos e visuais suficientes para artigos científicos, relatórios, apresentações, revistas etc.

Evite usar colunas em três dimensões (3D), que apesar de melhorarem o aspecto geral da figura diminuem sua compreensão, especialmente quando existirem várias colunas.

## Resumindo a Utilização dos Recursos Computacionais para Elaboração de Gráficos

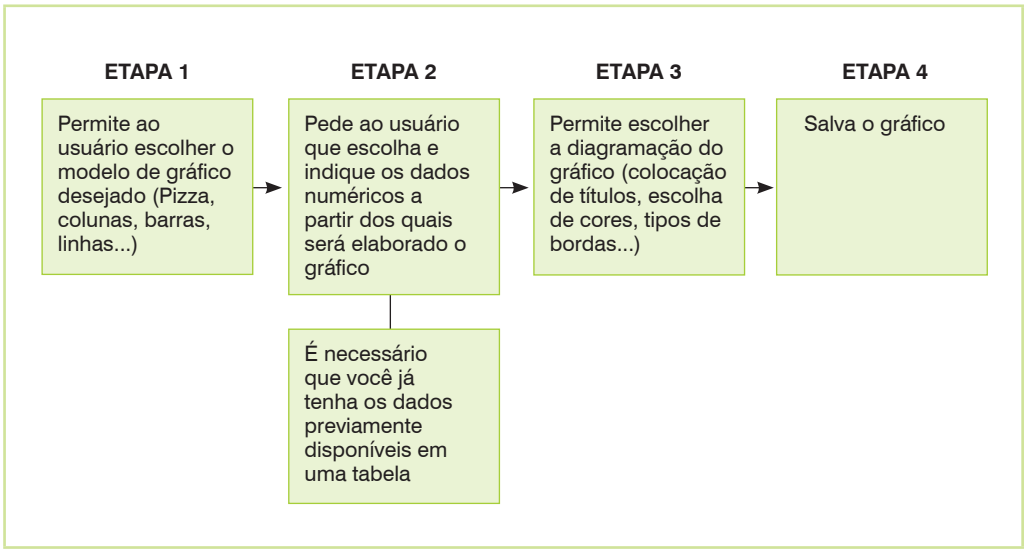
Na prática diária do mundo dos negócios é costume a manipulação de grandes volumes de dados. Temos como exemplo relatórios de vendas e extratos de movimentação de contas bancárias.

Comumente estes dados são expressos por valores numéricos, e a eles são associadas funções matemáticas, o que permite sua apresentação através de gráficos.

A elaboração de gráficos é um poderoso recurso para resumir e facilitar a visualização destes dados. Uma maneira fácil para realizar tal tarefa é através de recursos computacionais, com o uso do aplicativo Excel.

Os gráficos podem ser construídos passo a passo ao acessar o assistente de gráficos selecionando no menu *Inserir* e a seguir *Gráficos*.

Os passos que o assistente percorrerá com você serão:




Para saber mais sobre montagem de gráficos no Excel, veja uma aula virtual na página **www.mat 1 xyz.....com.br**. Lá você poderá ver com detalhes a confecção de gráficos.

# CARACTERÍSTICAS DE TABELAS E GRÁFICOS DE QUALIDADE

**A regra básica para elaboração de tabelas e gráficos é optar pela simplicidade**.

Cada gráfico/tabela deverá preferencialmente conter uma única informação, evitando-se agrupar grandes conjuntos de dados, o que causa dificuldade de compreensão ao leitor.

Durante a diagramação, evite bordas, cores, linhas de grade e todos os elementos visuais que possam "carregar" visualmente seu aspecto. Veja nas figuras a seguir alguns exemplos do que fazer ou não.

Uma **tabela inadequada**:

| REGIÃO | Participação no mercado | % de participação | Previsão | | CONCORRÊNCIA | Crescimento |
|---|---|---|---|---|---|---|
| N | R$ 12.000 | 12% | +12 % | A | Região Norte | 25 % |
| S | R$ 135.000 | 5% | +4 % | B | Região Sul | 12 % |
| L | R$ 120.000 | 22% | + 76 % | C | Região Leste | 15 % |
| O | R$ 2.000 | 13% | – 5% | D | Região Oeste | 15% |
| CO | R$ 20.000 | 18% | +23% | E | Região Oeste | 5 % |
| SU | R$ 132.000 | Não determ. | ------ | Tot | ---------- | 72% |

**Comentários sobre a tabela inadequada**:

1) A tabela traz duas informações diferentes ao mesmo tempo. Na verdade são duas tabelas agrupadas, misturando diversas informações de uma empresa com a de seu concorrente. Separe-as e identifique-as.

2) Falta título na tabela. A forma correta é constar um título, que forneça informações detalhadas ao leitor, sem que ele tenha que consultar quaisquer elementos no texto do relatório (a tabela em si é autoexplicativa). Deverá constar sempre o período de tempo a que ela se refere, bem como a fonte de dados.

| REGIÃO | | | | | CONCORRÊNCIA | |
|---|---|---|---|---|---|---|

Poderia ser assim:

TABELA Nº 01 – Participação no mercado nacional da ACMEX Ltda. – situação atual e previsões para o biênio 2004/2005.

3) Evitar misturar "siglas" com palavras ou repetir informações que sobrecarreguem o visual da tabela.

TABELA Nº 01 – Participação..... TABELA Nº 02 – .....

| REGIÃO | | | | | CONCORRÊNCIA | |
|---|---|---|---|---|---|---|
| N | | | | | Região Norte | |
| S | | | | | Região Sul | |
| L | | | | | Região Leste | |
| O | | | | | Região Oeste | |

PQ não escrever Norte, Sul, Leste... (ou N, S, L ... para o concorrente). Padronize!

Repetição da palavra "Região". Seria melhor o título CONCORRÊNCIA POR REGIÃO

4) Alinhe os valores numéricos à sua direita. Isto permite melhor visualização das grandezas envolvidas. Caso haja a presença de valores negativos eles podem ser coloridos ou negritados para chamar a atenção. É comum em atos contábeis colocá-los entre parênteses. Neste caso use uma legenda logo abaixo da tabela indicando algo como ***( ) ...valores negativos***.

| Participação no mercado | % de participação | Previsão |
|---|---|---|
| R$ 12.000 | 12% | +12 % |
| R$ 135.000 | 5% | +4 % |
| R$ 120.000 | 22% | + 76 % |
| R$ 2.000 | 13% | – 5% |
| R$ 20.000 | 18% | +23% |
| R$ 132.000 | Não determ. | ----- |

MELHOR

| Participação no mercado (em R$) | Participação (em %) | Previsão (em %) |
|---|---|---|
| 12.000 | 12 | 12 |
| 135.000 | 5 | 4 |
| 120.000 | 22 | 76 |
| 2.000 | 13 | **– 5** |
| 20.000 | 18 | 23 |
| 132.000 | ---------- | ------- |

Evite repetir unidades como R$ ou %. **O correto é que elas constem exclusivamente no título**. No entanto, na prática cotidiana é comum (e aceitável) a repetição do símbolo de %.

Valores indeterminados devem ser representados por traços. Nunca misture notações diferentes, ou use "zero" ou ϕ (vazio). Use "zero" exclusivamente para "zero".

Evite usar sinais de +. Identifique e destaque somente valores negativos.

5) Na diagramação da tabela é recomendável:
   a) evitar linhas verticais;
   b) evitar linhas horizontais (são aceitáveis para separar títulos de colunas e linhas de totalizadores, ou seja, limitar a tabela em cima e embaixo);
   c) evitar cores e sombreamentos desnecessários; use-as somente para destacar algo imprescindível;
   d) evitar usar negritos e/ou letras maiúsculas nas células internas da tabela (são aceitáveis somente nos títulos);
   e) evitar abreviações nos títulos das colunas e linhas (se necessário aumente sua tabela);
   f) evitar linhas e colunas de tamanhos diferentes; é melhor procurar uniformizar suas larguras;
   g) quando houver várias tabelas em um texto, na medida do possível, todas deverão ser de mesmas dimensões e alinhadas igualmente;
   h) Quando houver mais de uma tabela no texto, cada uma delas deverá ser numerada e identificada com seu título;
   i) evitar colocar duas ou mais tabelas lado a lado, sendo melhor posicioná-las uma seguida da outra, verticalmente;
   j) centralizar cada uma das tabelas na página; e
   k) o conteúdo das células deverão ser centralizados.

**Como ficaria a tabela melhor apresentada**[1]:

**Tabela 1**
**Participação no Mercado Nacional da ACMEX Ltda**
**Situação Atual e Previsões para o Biênio 2006/2008**

| REGIÃO | PARTICIPAÇÃO NO MERCADO (em R$) | PARTICIPAÇÃO (em %) | PREVISÃO |
|---|---|---|---|
| NORTE | 12.000 | 12 | 12 |
| SUL | 135.000 | 5 | 4 |
| LESTE | 120.000 | 22 | 76 |
| OESTE | 2.000 | 13 | **(5)** |
| CENTRO-OESTE | 20.000 | 18 | 23 |
| SUDESTE | 132.000 | ------- | ------ |

( ) ...valores negativos
Fonte: Diretoria de Planejamento e Marketing (agosto – 2009)

[1] As características apresentadas neste capítulo podem não possuir rigor adequado para trabalhos acadêmicos e científicos em níveis mais elaborados. Nosso objetivo foi proporcionar ao aluno alguns conceitos básicos de comunicação visual, uma vez que ele está tomando os primeiros contatos com o tema de maneira mais aplicada à sua futura prática de trabalho. Ao longo de seu curso de graduação, espera-se que o aluno aprimore tais conhecimentos.

## Exercícios

1) Os dados a seguir representam as notas dadas em uma pesquisa de satisfação geral de um restaurante:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7,0 | 8,0 | 8,0 | 5,0 | 6,0 | 8,0 | 7,0 | 6,0 | 7,0 | 7,0 |
| 3,0 | 9,0 | 8,0 | 8,0 | 6,0 | 4,0 | 6,0 | 5,0 | 2,0 | 10,0 |
| 8,0 | 8,0 | 9,0 | 6,0 | 7,0 | 10,0 | 10,0 | 5,0 | 8,0 | 8,0 |

a) Organize o rol.
b) Monte uma tabela de frequências, indicando a frequência simples, a frequência relativa simples, a frequência acumulada e a frequência relativa acumulada.
c) Com a ajuda do Excel, monte gráficos que mostrem o desempenho do restaurante.

2) A tabela abaixo indica o tamanho do estabelecimento por número de funcionários, em 31 de dezembro de 2011, no setor de atividades auxiliares dos serviços financeiros, seguros, previdência complementar, planos de saúde e de resseguros no município de São Paulo, de acordo com o tamanho do estabelecimento:

| Tamanho do Estabelecimento – por número de funcionários ($x_i$) | Número de Estabelecimentos ($f_i$) |
|---|---|
| De 1 a 4 | 3.011 |
| De 5 a 9 | 2.480 |
| De 10 a 19 | 3.228 |
| De 20 a 49 | 5.906 |
| De 50 a 99 | 6.217 |
| De 100 a 249 | 9.271 |
| De 250 a 499 | 11.070 |
| De 500 a 999 | 7.102 |
| **Total** | **48.285** |

a) Complete a tabela calculando a frequência simples, a frequência relativa simples, a frequência acumulada e a frequência relativa acumulada.
b) Com a ajuda do Excel, monte gráficos que facilitem a compreensão dos dados para uma pessoa que não conheça estatística.

# 3 Medidas Resumo

Temos insistido ao longo dos capítulos anteriores sobre a principal função da Estatística Descritiva, que seria a capacidade de resumir dados e apresentá-los de forma acessível e de fácil compreensão. Para tanto, exploramos nos capítulos anteriores a forma de fazê-lo, através de tabelas e gráficos.

Uma outra forma de obter elementos que permitam caracterizar séries de valores é estabelecer algumas **medidas resumo**, ou seja, encontrar um ou mais números que representem todos os demais valores que compõem o estudo. São as chamadas **medidas de posição (medidas de tendência central** e **separatrizes**).

Tais medidas representam um valor ao redor do qual os elementos da série estão distribuídos; em uma representação gráfica, posicionaria a série em um eixo horizontal, localizando um determinado valor em torno do qual a série se concentra.

Num passo seguinte, no próximo capítulo, será discutida a qualidade destas medidas resumo como representantes do todo, ou seja, estudaremos se, de fato, este valor é competente para representar toda uma população ou amostra.

O quadro a seguir, mostra algumas características destas medidas de tendência central:

| Tipo | Representação | Usos e Características |
|---|---|---|
| Média | $\bar{x}$ (amostra) ou $\mu$ (população) | • deve ser utilizada quando houver forte concentração de valores na "área central" da série organizada (rol);<br>• não representa adequadamente séries que possuam valores extremados. |
| Mediana | $m_d$ | • pode ser utilizada quando houver forte concentração de dados no início ou no final da série;<br>• é mais importante a quantidade de elementos presentes na série, do que propriamente seus valores;<br>• não é afetada por valores extremados presentes no conjunto de medidas, como acontece para a média. |
| Moda | $m_o$ | • pode ser utilizada quando houver forte concentração de dados no início ou no final da série;<br>• é utilizada em séries que apresentam um elemento típico, que se repete com maior frequência do que os demais. |

# MÉDIAS

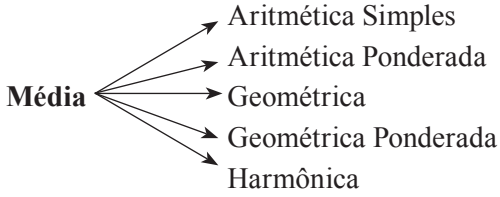
Talvez sejam as medidas de posição mais comumente utilizadas, o que acaba muitas vezes representando um problema, uma vez que são usadas indiscriminadamente (quando seria mais adequado o uso da mediana ou da moda), causando percepções distorcidas do fenômeno estudado.

Costuma ser usada de forma "incompleta", pois a utilização da média implica na necessidade do cálculo de outra grandeza chamada **desvio (dispersão)**, que será estudada no capítulo seguinte.

Por exemplo, se um aluno tirou, em três avaliações, as notas zero, cinco e dez, supondo que o peso de cada prova é o mesmo, a média desse aluno será cinco. Perceba que a média desse aluno não reflete com exatidão o seu desempenho, uma vez que ele tirou uma nota muito ruim (zero) e uma nota máxima (dez). Para avaliar o comportamento desse fenômeno, e de qualquer outro, o ideal é utilizar outras medidas estatísticas, sobretudo aquelas que vão medir a dispersão dos dados observados – no exemplo as três notas – em relação à média.

Portanto, expressar uma média sem seu correspondente valor de desvio pode implicar em grave erro de avaliação de um fenômeno.

Os diferentes tipos de médias são representados a seguir:




## Média Aritmética Simples e Ponderada ($\bar{x}$)

Vamos retomar o exemplo descrito acima, do aluno que tirou nota 0, 5 e 10 nas três avaliações que fez. Para cálculo da média aritmética dessas notas é bastante simples, basta somar as três notas e dividir por três.

$$Média = \frac{0+5+10}{3} = \frac{15}{3} = 5$$

Mas agora vamos formalizar esse cálculo, utilizando as fórmulas apropriadas para tanto. Neste caso, como representa uma média aritmética simples, a fórmula utilizada para tanto é a seguinte:

$$\bar{X} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3}{3} = \frac{0+5+10}{3}$$

Nesse exemplo, a MÉDIA ARITMÉTICA SIMPLES é assim denominada, porque os valores da frequência simples dos três dados observados é igual a 1. Porém, quando temos dados observados com frequência superior a 1, o ideal é que passemos a utilizar uma ponderação. Por exemplo, a nota média dos 10 alunos que realizaram o exame final de estatística pode ser calculada somando-se as dez notas e dividindo por 10. Suponha que essas são as notas desse grupo de alunos: 4, 4, 5, 5, 5, 6, 6, 7, 7 e 8.

A média do grupo será:

$$\bar{X} = \frac{\sum x_i}{n} = \frac{4+4+5+5+5+6+6+7+7+8}{10} = \frac{57}{10} = 5{,}7$$

Podemos refazer esse cálculo levando em consideração as frequências das notas, para tanto vamos dispor essas notas numa tabela de distribuição de frequência comum:

| $x_i$ | $f_i$ |
|---|---|
| 4 | 2 |
| 5 | 3 |
| 6 | 2 |
| 7 | 2 |
| 8 | 1 |
| Total | 10 |

Neste caso, vamos utilizar uma nova fórmula que utilize as frequências como uma forma de ponderar as notas.

$$\bar{X} = \frac{\sum x_i f_i}{\sum f_i}$$

Perceba que o elemento no numerador da fórmula passou a ser multiplicado (ponderado) pela frequência, e o denominador a somatória da frequência, que corresponde exatamente ao número de dados observados. Precisamos, primeiramente, multiplicar os dados observados (notas) pelas respectivas frequências:

| $x_i$ | $f_i$ | $x_i f_i$ |
|---|---|---|
| 4 | 2 | 8 |
| 5 | 3 | 15 |
| 6 | 2 | 12 |
| 7 | 2 | 14 |
| 8 | 1 | 8 |
| Total | 10 | 57 |

Agora basta utilizar esses resultados na fórmula:

$$\bar{X} = \frac{\sum x_i f_i}{\sum f_i} = \frac{8+15+12+14+8}{10} = \frac{57}{10} = 5{,}7$$

É bastante simples, e a interpretação é exatamente a mesma, ou seja, a nota média dos dez alunos que realizaram o exame de estatística é 5,7. Note que esse mesmo cálculo pode ser considerado para o primeiro exemplo, as três notas de um aluno (0, 5 e 10). A única diferença é que a ponderação, no caso a frequência, de cada nota é igual a 1. Neste caso a distribuição de frequência ficará assim:

| $x_i$ | $f_i$ | $x_i f_i$ |
|---|---|---|
| 0 | 1 | 0 |
| 5 | 1 | 5 |
| 10 | 1 | 10 |
| Total | 3 | 15 |

E o cálculo da média será:

$$\bar{X} = \frac{\sum x_i f_i}{\sum f_i} = \frac{0+5+10}{3} = \frac{15}{3} = 5$$

Você deve estar se perguntando porque complicar tanto o cálculo de uma simples média. O fato é que situações mais complexas, tais como a utilização de uma grande quantidade de dados observados ou séries estatísticas já dispostas em distribuições de frequências, passam a requerer essa forma de calcular a média. Os exemplos utilizados realmente não demandam a utilização da fórmula para o cálculo da média, mas com esses exemplos procuramos demonstrar como a interpretação dessa medida estatística é fácil.

O exemplo seguinte demonstra o cálculo da média para uma variável contínua ou para uma distribuição de frequências agrupada em classes. Assim como no exemplo dos dez alunos, neste caso calculamos **a média aritmética ponderada** que pode ser obtida a partir de uma tabela de distribuição de frequências.

Vale lembrar que nesses casos os dados observados são apresentados em intervalos, como no exemplo abaixo, em que na primeira classe os valores observados vão de 2 (inclusive) a 6 (exclusive), e assim por diante.

| $x_i$ | $f_i$ |
|---|---|
| 2 ⊢ 6 | 4 |
| 6 ⊢ 10 | 2 |
| 10 ⊢ 14 | 1 |
| 14 ⊢ 18 | 3 |
| Total | 10 |

Para calcular a média, entre outras estatísticas, não podemos utilizar os intervalos. Nesse caso, utilizamos o valor médio de cada intervalo de classe, que representa a média simples do limite inferior e superior de cada intervalo de classe. Para esse mesmo exemplo, observe que no primeiro intervalo o limite inferior da primeira classe é igual a 2 e o limite superior 6, então a média desse intervalo será (2+6)/2 = 4. Esse valor corresponderá ao dado observado ($x_i$) para efeito de cálculo da média, conforme veremos abaixo:

| $x_i$ | $x_i$ | $f_i$ | $x_i f_i$ |
|---|---|---|---|
| 2 ⊢ 6<br>(2+6) / 2 | 4 | 4 | 4 . 4 = 16 |
| 6 ⊢ 10<br>(6 + 10 ) / 2 | 8 | 2 | 8 . 2 = 16 |
| 10 ⊢ 14<br>(10+14)/2 | 12 | 1 | 12 . 1 =12 |
| 14 ⊢ 18<br>(14+18)/2 | 16 | 3 | 16 . 3 = 48 |
| | | 10 | 94 |

Uma vez identificado o ponto médio de cada intervalo, basta aplicar a fórmula da média tal qual fizemos no exemplo anterior:

$$\overline{X} = \frac{\sum X_i f}{\sum f_i} = \frac{16+16\_12+48}{10} = \frac{94}{10} = 9{,}4$$

## Média Ponderada

Outra medida de posição comumente utilizada é a média ponderada. Essa medida é semelhante à obtida a partir da distribuição de frequências. Como observamos, em uma distribuição de frequências, elas assumem a função de ponderação na composição da somatória dos dados observados.

No caso a seguir essa mesma perspectiva existe. A diferença é que passamos a chamar tal componente de peso ou ponderação ($p_i$). No exemplo a seguir temos uma empresa que está buscando apurar a sua margem de lucro, que é obtida a partir da venda de seis produtos com participações e taxas de retorno diferentes.

Para que possamos identificar o lucro auferido com os seis produtos devemos, inicialmente, calcular a participação de cada produto no volume total de vendas e, assim, compor os pesos ou ponderações de cada um deles.

Uma vez que calculamos o peso de cada produto devemos multiplicá-los pela respectiva margem de lucro. A somatória desses resultados compõe a média ponderada da margem de lucro da empresa.

| Produtos | Vendas | Margem de Lucro ($x_i$) | Participação nas Vendas ($p_i$) | $x_i p_i$ |
|---|---|---|---|---|
| Produto A | R$ 60.000 | 4,00% | 0,3158 | 1,2632% |
| Produto B | R$ 40.000 | 6,00% | 0,2105 | 1,2632% |
| Produto C | R$ 30.000 | 8,00% | 0,1579 | 1,2632% |
| Produto D | R$ 25.000 | 3,00% | 0,1316 | 0,3947% |
| Produto E | R$ 20.000 | 5,00% | 0,1053 | 0,5263% |
| Produto F | R$ 15.000 | 7,00% | 0,0789 | 0,5526% |
| Total | R$ 190.000 | – | 1 | 5,26% |

Veja o cálculo da média ponderada abaixo.

$$\bar{x}_p = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p} = \frac{5,26}{1} = 5,26\%$$

Observe que a somatória do peso tem que ser igual a 1.

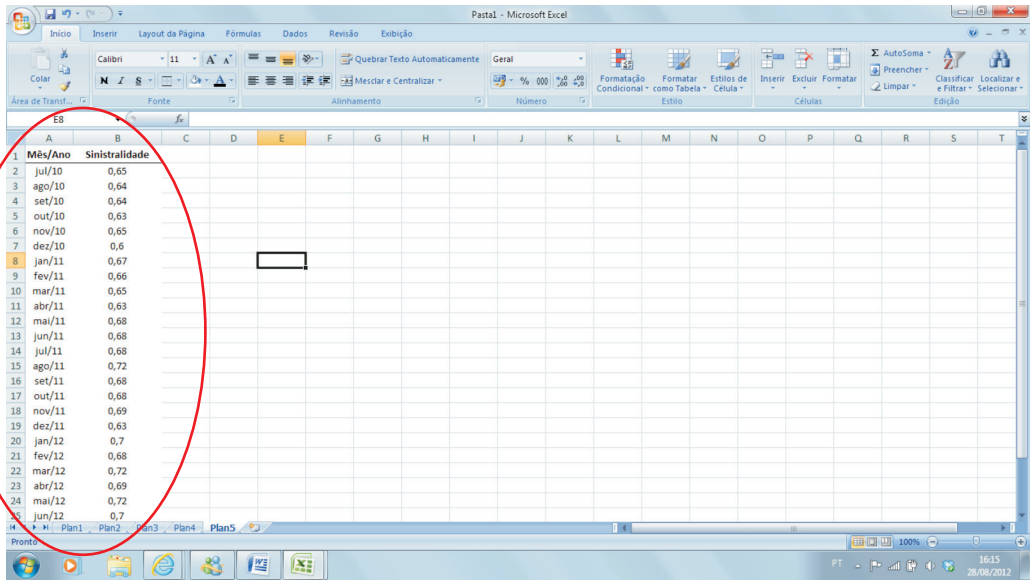
## Usando o Excel para Calcular a Média

O cálculo da média utilizando o Excel é bastante simples, e pode ser feito de duas maneiras: a partir da inclusão de uma função estatística; e, por meio da ferramenta de análise de dados. Como a ferramenta de análise de dados nos fornece um conjunto de resultados denominado estatística descritiva, vamos trabalhar, por enquanto, apenas com a primeira opção. Para tanto, vamos utilizar os dados apresentados na tabela abaixo, que representam o índice de sinistralidade do seguro de automóvel, entre julho de 2010 e junho de 2011.

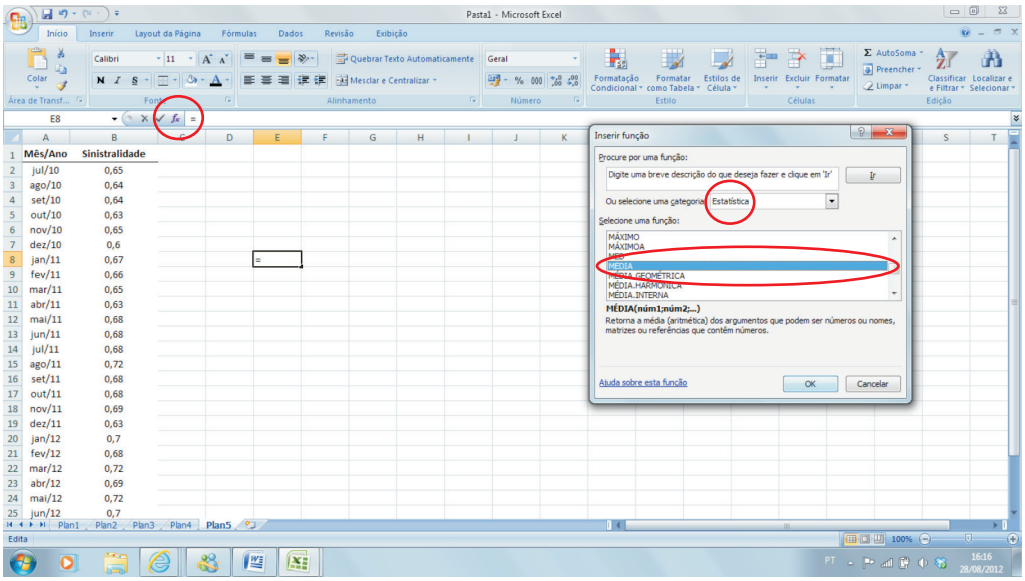
| Mês/Ano | Sinistralidade | Mês/Ano | Sinistralidade |
|---|---|---|---|
| jul/10 | 0,65 | jul/11 | 0,68 |
| ago/10 | 0,64 | ago/11 | 0,72 |
| set/10 | 0,64 | set/11 | 0,68 |
| out/10 | 0,63 | out/11 | 0,68 |
| nov/10 | 0,65 | nov/11 | 0,69 |
| dez/10 | 0,6 | dez/11 | 0,63 |
| jan/11 | 0,67 | jan/12 | 0,7 |
| fev/11 | 0,66 | fev/12 | 0,68 |
| mar/11 | 0,65 | mar/12 | 0,72 |
| abr/11 | 0,63 | abr/12 | 0,69 |
| mai/11 | 0,68 | mai/12 | 0,72 |
| jun/11 | 0,68 | jun/12 | 0,7 |

Fonte: SUSEP

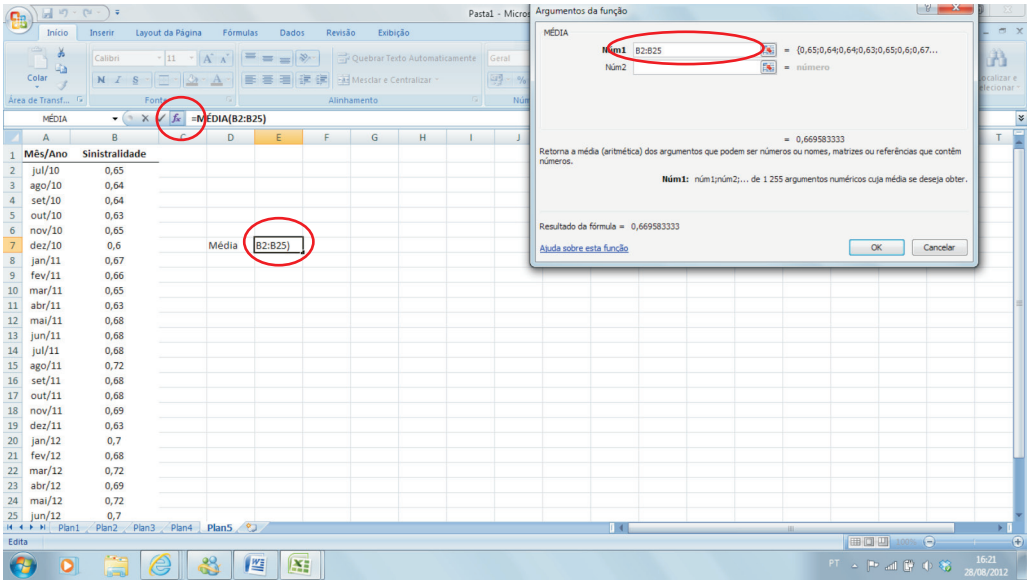
Para que o Excel identifique todos os dados da tabela, lembre-se que os mesmos deverão estar dispostos numa única coluna.




Em seguida clique em $f_x$. A janela Inserir função se abrirá, selecione a categoria estatística e, sem seguida MÉDIA e clique OK, conforme está demonstrado abaixo.



Em seguida uma nova janela se abrirá, solicitando as células onde se encontram os dados observados, que neste exemplo são de B2 até B25.




| | Mês/Ano | Sinistralidade |
|---|---|---|
| 2 | jul/10 | 0,65 |
| 3 | ago/10 | 0,64 |
| 4 | set/10 | 0,64 |
| 5 | out/10 | 0,63 |
| 6 | nov/10 | 0,65 |
| 7 | dez/10 | 0,6 |
| 8 | jan/11 | 0,67 |
| 9 | fev/11 | 0,66 |
| 10 | mar/11 | 0,65 |
| 11 | abr/11 | 0,63 |
| 12 | mai/11 | 0,68 |
| 13 | jun/11 | 0,68 |
| 14 | jul/11 | 0,68 |
| 15 | ago/11 | 0,72 |
| 16 | set/11 | 0,68 |
| 17 | out/11 | 0,68 |
| 18 | nov/11 | 0,69 |
| 19 | dez/11 | 0,63 |
| 20 | jan/12 | 0,7 |
| 21 | fev/12 | 0,68 |
| 22 | mar/12 | 0,72 |
| 23 | abr/12 | 0,69 |
| 24 | mai/12 | 0,72 |
| 25 | jun/12 | 0,7 |

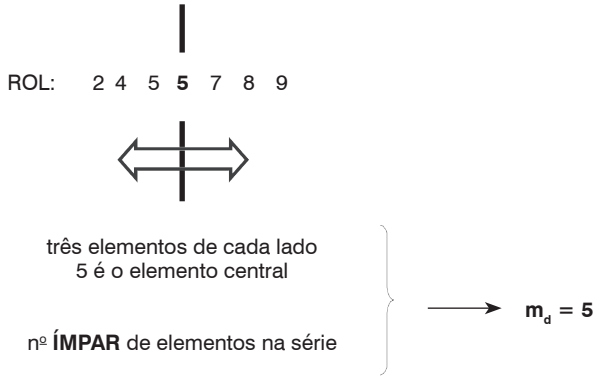
Clique Ok e o resultado, a média do conjunto de dados observados, será apresentado na célula escolhida. Nesse exemplo, o resultado é 0,67, o que significa dizer que a sinistralidade média para o seguro de automóveis no período analisado é 0,67.

## Mediana

A mediana ($m_d$) é um valor que separa o rol em duas partes com o mesmo número de elementos, ou seja, o valor que ocupa a posição central do conjunto de dados observados.

Cálculo da **mediana para variáveis discretas**:

Na série 8, 4, 5, 9, 5, 2, 7 teríamos:



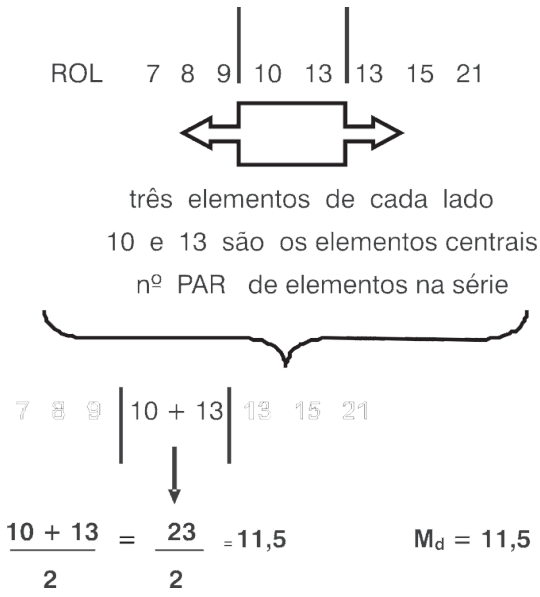
Portanto, a Mediana do conjunto de dados acima é Md=5, ou seja, esse valor divide o rol exatamente ao meio. Perceba que 50% dos dados estão abaixo da mediana e 50% acima. Para calcular a Mediana é necessário identificar o Elemento Mediano (EMd), que corresponde à posição, no rol, em que se encontra a Mediana.

Quando o número de dados observados é ímpar, neste caso são sete dados, o elemento mediano é calculado da seguinte maneira:

$EMd = \frac{n+1}{2} = \frac{7+1}{2} = 4$, ou seja, a Mediana está na quarta posição, ou é o quarto elemento da série de dados, que para esse caso é igual a 5.

| Rol | | | | | | | |
|---|---|---|---|---|---|---|---|
| Posição | 1ª | 2ª | 3ª | 4ª | 5ª | 6ª | 7ª |
| Dados | 2 | 4 | 5 | 5 | 7 | 8 | 9 |

Quando a série possui um número par de elementos, temos dois termos centrais: (n /2) e (n/2 + 1). **Exemplo**: na série 7, 21, 13, 15, 10, 8, 9, 13



A regra de bolso para o cálculo da Mediana quando o número de elementos é par é a seguinte:

**1º passo)** Calcule o elemento mediano:

$$Emd = \frac{n}{2} = \frac{8}{2} = 4$$

De acordo com a fórmula acima, o elemento mediano está na quarta posição. Porém, como já foi demonstrado, esse valor não está dividindo a série exatamente ao meio, por isso se faz necessário o segundo passo:

**2º passo)** Calcule a média simples do valor que está na quarta posição e na posição imediatamente posterior, ou seja, na quinta posição, que para esse exemplo são os valores 10 e 13.

| Rol | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Posição | 1ª | 2ª | 3ª | 4ª | 5ª | 6ª | 7ª | 8ª |
| Dados | 7 | 8 | 9 | 10 | 13 | 13 | 15 | 21 |

Portanto, o resultado da Mediana será: $Md = \frac{10+13}{2} = 11{,}5$

## Cálculo da Mediana para Variáveis Contínuas

Da mesma forma que fizemos algumas adaptações para o cálculo da média de uma série disposta numa distribuição de frequências agrupada em classes, o cálculo da mediana também requer que adotemos algumas regras de bolso para encontrar o valor mediano. No exemplo abaixo, temos a distribuição de idades dos funcionários de uma empresa:

| Classe | Idades ($x_i$) | $f_i$ |
|---|---|---|
| 1 | 18 ⊢ 22 | 2 |
| 2 | 22 ⊢ 26 | 5 |
| 3 | 26 ⊢ 30 | 8 |
| 4 | 30 ⊢ 34 | 4 |
| 5 | 34 ⊢ 38 | 2 |
| Total | | 21 |

O primeiro passo para calcular a mediana dessa distribuição é identificar o elemento Mediano:

$$Emd = \frac{n}{2} = \frac{21}{2} = 10{,}5$$

Portanto, o valor mediano na posição 10,5. Para identificar a classe em que se encontra esse valor mediano precisamos complementar a distribuição com a Frequência Acumulada.

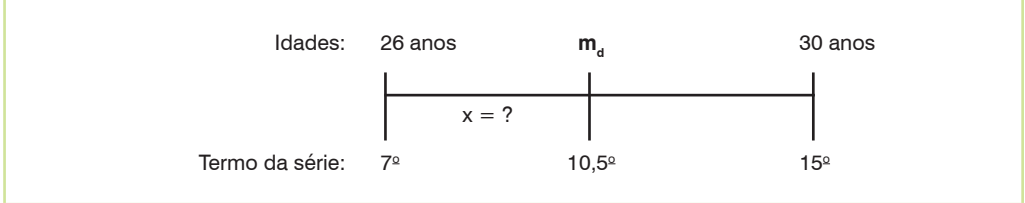
| Classe | Idades ($x_i$) | $f_i$ | $F_i$ |
|---|---|---|---|
| 1 | 18 ⊢ 22 | 2 | 2 |
| 2 | 22 ⊢ 26 | 5 | 7 |
| 3 | 26 ⊢ 30 | 8 | 15 |
| 4 | 30 ⊢ 34 | 4 | 19 |
| 5 | 34 ⊢ 38 | 2 | 21 |
| Total | | 21 | – |

Agora que temos a coluna da frequência acumulada, é possível identificar a classe em que se encontra o valor mediano. Como o elemento mediano é igual a 10,5, a mediana se encontra na terceira classe (de 26 a 30 exclusive), que indica os valores que estão entre 8ª e a 15ª posições. Lembre-se que os dados estão agrupados em ordem crescente, por isso podemos identificar as posições de cada classe a partir da frequência acumulada.

Ademais, como o Elemento Mediano (10,5) é um valor decimal, o mesmo indica que a Mediana está localizada entre o 10º e 11º elemento da série. Observando a coluna das frequências acumuladas ($F_i$), note que o 10º e 11º elementos da série estão localizados na terceira classe, portanto esta classe será considerada como a classe mediana.

| | Classe | Intervalo (idades) | $f_i$ | $F_i$ |
|---|---|---|---|---|
| | | | | 7 |
| **Classe mediana →** | **3** | **26 ⊢ 30** | **8** | **15** |

Este intervalo de quatro elementos (26 a 30 anos) possui oito funcionários, então podemos dividi-lo de forma adequada:



Ou seja: $\frac{15-7}{4} = \frac{10{,}5-7}{x}$ ou seja $\frac{8}{4} = \frac{10{,}5-7}{x} \rightarrow x = \frac{10{,}5-7}{8} \cdot 4$

Desta forma, $M_d = 26 + x \rightarrow m_d = 26 + \underbrace{\frac{10{,}5-7}{8} \cdot 4}_{x} \rightarrow m_d = 27{,}8$

Generalizando o cálculo, temos a seguinte fórmula para o cálculo da mediana para variáveis contínuas:

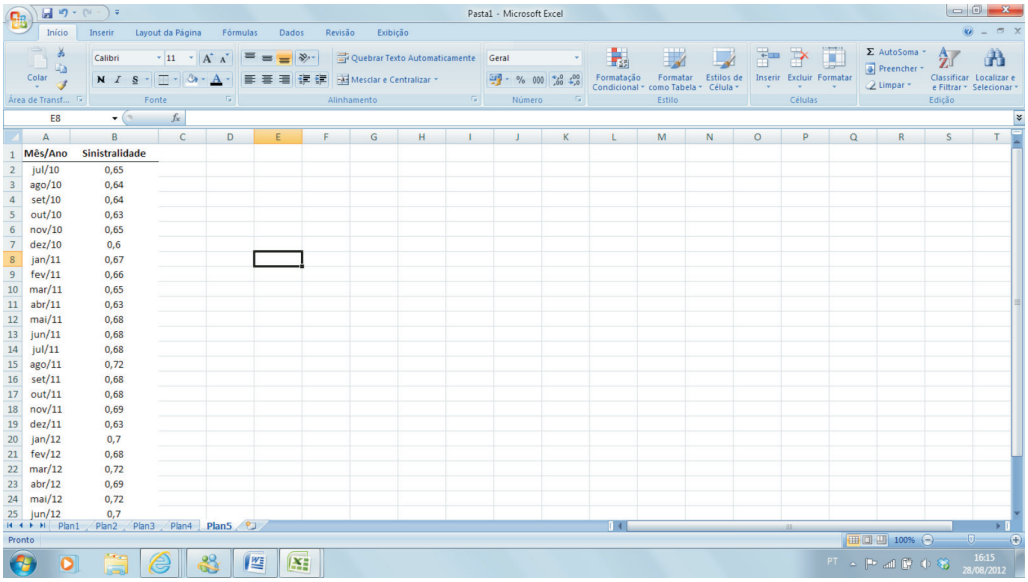
$$Md = l_{Md} + \frac{E_{Md} - F_{Ant}}{f_{Md}} h = 26 + \frac{10{,}5-7}{8} \times 4 = 27{,}8$$

onde:
$l_{md}$ – Limite inferior da classe onde se encontra o valor mediano;
$EMd$ – Elemento mediano;
$F_{ant}$ – Frequência acumulada da classe anterior à classe onde se encontra o valor mediano;
$f_{Md}$ – Frequência simples da classe onde se encontra o valor mediano; e
$h$ – Amplitude do intervalo onde se encontra o valor mediano.

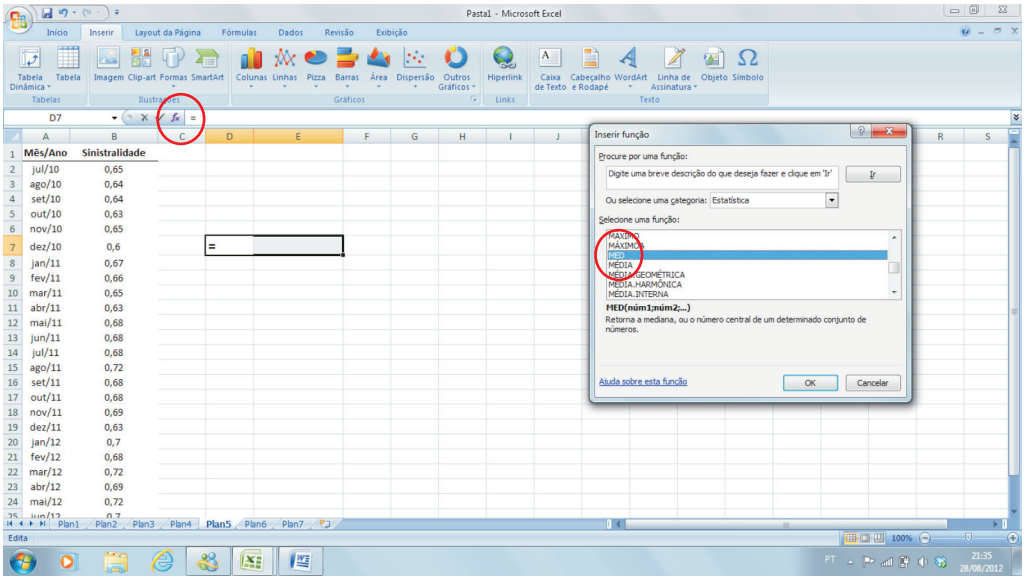
## Usando o Excel para Cálculo da Mediana

Assim como a média, o cálculo da mediana no Excel pode ser feito utilizando a inclusão de uma função estatística, ou por meio da ferramenta de análise Estatística Descritiva. Por enquanto vamos nos ater apenas à primeira forma, ou seja, vamos incluir uma função estatística.

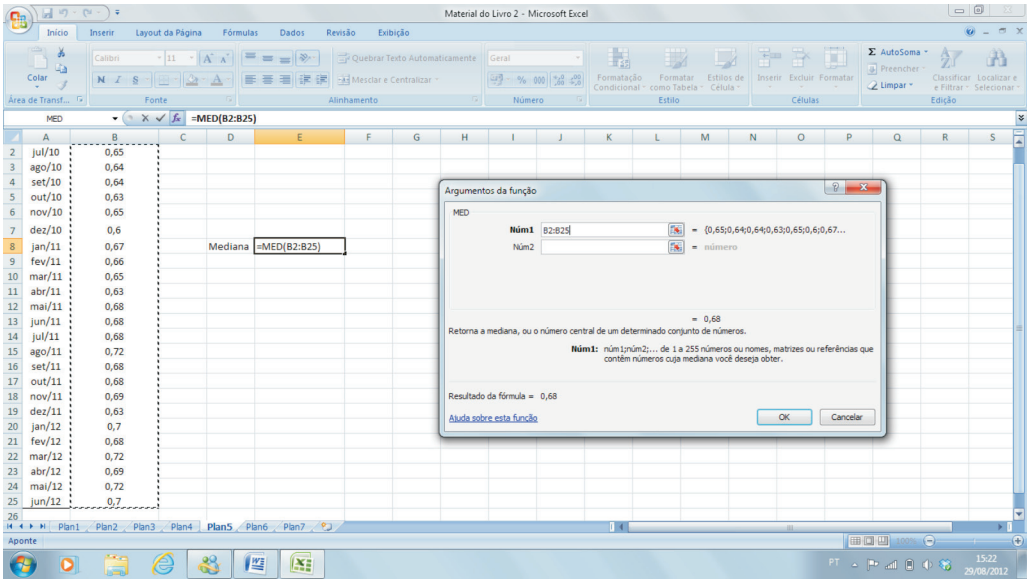
Para tanto, vamos nos basear no exemplo da Sinistralidade das seguradoras no Brasil entre julho de 2010 e junho de 2012, lembrando que os dados devem estar dispostos numa única coluna.




Escolha uma célula e clique em *fx*, e uma nova janela – Inserir função – se abrirá. Uma vez aberta a janela Inserir função escolha a função Med (de Mediana), e em seguida clique OK.



Ao clicar OK abrirá a janela Argumentos da função, na qual serão informadas as células onde se encontram a série de dados, que para esse exemplo são de B2 até B25.




Dê OK e o valor da mediana aparecerá imediatamente na célula escolhida, que para o exemplo em questão é igual a 0,68.

## Moda

A moda ($m_o$) é o valor que possui maior frequência em uma série, ou seja, é o valor que mais se repete na série.

No caso de tratarmos com **variáveis discretas**, basta identificar no rol ou na tabela de distribuição a variável que possui maior frequência.

### • • • Exemplo 1

No rol 5 5 6 6 6 9 10 12 12

$M_o = 6$

Caso tivéssemos este mesmo rol expresso na forma de uma tabela de distribuição:

| $x_i$ | $f_i$ | |
|---|---|---|
| 5 | 2 | |
| **6** | **3** | → $M_o = 6$ |
| 9 | 1 | |
| 10 | 1 | |
| 12 | 2 | |

**• • • Exemplo 2**

No Rol 2 3 **4 4 5 5** 6 7 8 9.
Existem dois valores modais: Mo=4 e Mo=5. Uma série pode ter mais do que dois valores modais.

**• • • Exemplo 3:**

No Rol 1 2 3 4 5 6 7 8 9 10.
Neste caso, não existe nenhum dado observado que se repete, logo, dizemos que a série é **AMODAL**.

No caso de duas variáveis diferentes possuírem as mesmas maiores frequências, diríamos que a série é bimodal, existem casos em que mais do que duas variáveis possuem as mesmas maiores frequências, assim como uma série pode não ter nenhum valor que se repete mais do que as demais de uma determinada série, o que caracteriza uma série amodal.

Como os dados estão agrupados em classes (intervalos de valores), o cálculo da Moda será feito por aproximação. Para tanto, vamos utilizar a seguinte fórmula de King:

$$Mo = l_{Mo} + \frac{f_{post}}{f_{ant} + f_{post}} h$$

Onde,
$M_o$ é o valor modal;
$l_{mo}$ é o limite inferior da classe em que se encontra o valor modal;
$f_{post}$ frequência simples da classe posterior àquela em que se encontra o valor modal;
$f_{ant}$ é o limite inferior da classe anterior àquela em que se encontra o valor modal; e
$h$ é a amplitude da classe em que se encontra o valor modal.

Calculemos a Moda para o exemplo abaixo, que apresenta a distribuição de idades dos 21 funcionários de uma empresa:

| Classe | Idades ($x_i$) | $f_i$ |
|---|---|---|
| 1 | 18 ⊢ 22 | 2 |
| 2 | 22 ⊢ 26 | 5 |
| 3 | 26 ⊢ 30 | 8 |
| 4 | 30 ⊢ 34 | 4 |
| 5 | 34 ⊢ 38 | 2 |
| Total | | 21 |

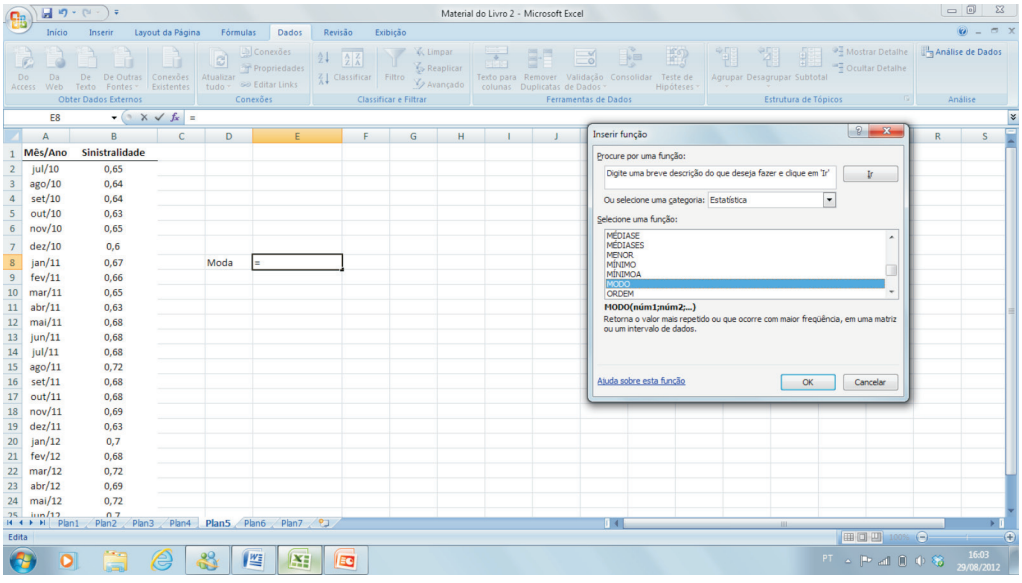
O primeiro passo consiste em identificar a classe em que se encontra o valor modal, que para o exemplo em questão é a terceira, uma vez que este intervalo apresenta a maior frequência. Em seguida, apliquemos a fórmula identificando cada um dos seus elementos:

$$Mo = l_{Mo} + \frac{f_{post}}{f_{ant} + f_{post}} h = 26 + \frac{4}{5+4} 4 = 27{,}8$$

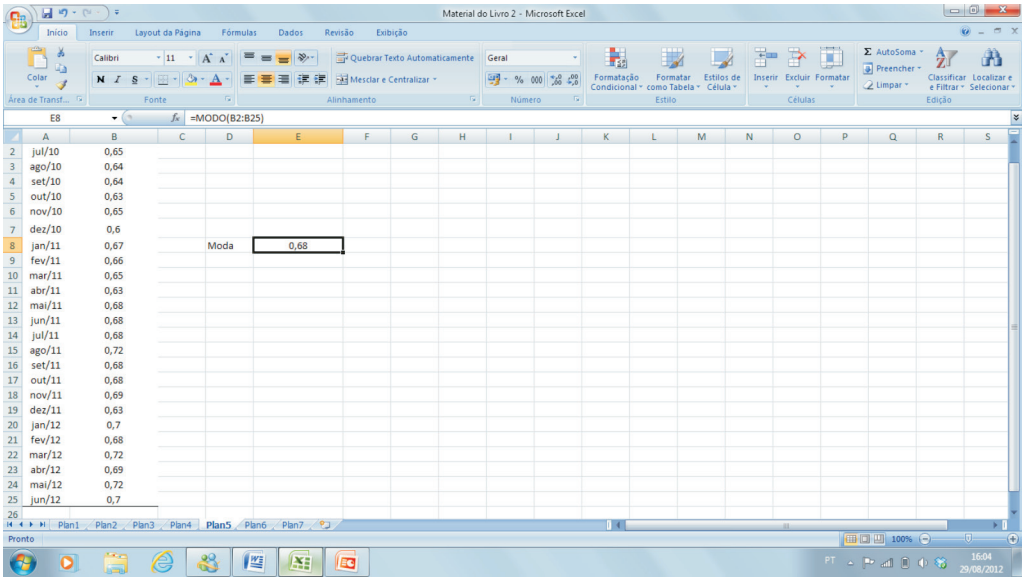
Portanto, a idade modal da série em questão é 27,8 anos.

## Usando o Excel para Cálculo da Moda

Voltemos ao exemplo da Sinistralidade das seguradoras no Brasil entre julho de 2010 e junho de 2012, reiterando que a série deve estar disposta numa única coluna. Escolha uma célula e, em seguida, clique no ícone fx. Em seguida, na janela aberta selecione a categoria estatística, que apresentará uma série de funções. Escolha a função MODO, ou seja, o Moda, conforme está demonstrado a seguir:



Dê OK, e o valor da Moda será apresentado na célula escolhida, que para o presente exemplo é igual a 0,68.




Portanto, o resultado indica que para o período analisado o índice de sinistralidade modal, o que mais se repetiu ao longo da série, é 0,68.

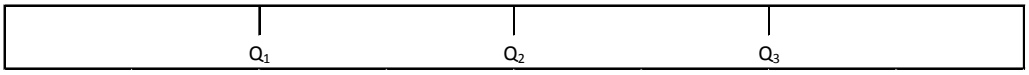
## Medidas Separatrizes (ou Medidas de Posição Relativa)

Além da média, mediana e moda, outras medidas de posição podem ser utilizadas para caracterizar uma série. Permitem comparar valores de conjuntos de dados diferentes, ou mesmo dentro de um mesmo conjunto de dados, sendo denominadas genericamente por QUANTIS, mais especificamente: Decil, Quartil, e PERCENTIL.

Assim como a mediana, o quartil, o decil e o centil divide a série de dados de tal modo que se torna possível avaliá-la de forma segmentada. São chamadas medidas separatrizes, pois assim como a mediana que divide a série em duas partes iguais, o quartil dividirá a série em quatro partes iguais, o decil em dez partes iguais e o centil em cem partes iguais, conforme descreve a figura a seguir:

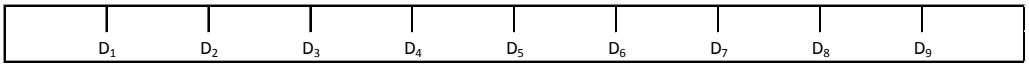
O quartil divide a série em quatro partes iguais, de tal modo que podemos calcular o 1º, o 2º e o 3º quartil. Para tanto, precisamos identificar o elemento quartil $EQ_i = \frac{n \times i}{4}$, assim como fizemos no cálculo da Mediana, lembrando que tal elemento identifica a posição no rol do valor quartil.

Observe que o primeiro quartil ($Q_1$) separa a sequência ordenada deixando 25% (¼) de seus valores à sua esquerda e o restante (75%) à direita. Já o 2º quartil, que coincide com a mediana, separa a série em duas partes simétricas.



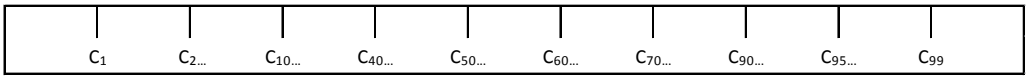

**Mediana**

Da mesma forma, o Elemento Decil: $ED_i = \frac{n \times i}{10}$ nos dá a posição do i-ésimo decil que temos o interesse em calcular.




**Mediana**

ER, por fim, o Elemento Centil: $EC_i = \frac{n \times i}{100}$ indica a posição do i-ésimo decil da série de dados.




**Mediana**

Observe nas figuras acima que para as três medidas separatrizes é possível identificar um valor mediano correspondente, pois essa estatística além de ser uma medida de tendência central, também tem a função de uma separatriz ao dividir exatamente ao meio uma série de dados. Portanto, o valor da mediana é o mesmo do segundo quartil, do quinto decil e do quinquagésimo centil.

Vamos trabalhar sobre um exemplo, para que possamos calcular as três medidas separatrizes, bem como interpretar os seus resultados. A série abaixo, apresentada numa distribuição de frequências comum, representa a idade de 110 clientes da Corretora Alpha com idade entre 18 e 39 anos, que contrataram seguro de automóvel num determinado período.

| $x_i$ | $f_i$ |
|---|---|
| 18 | 4 |
| 19 | 7 |
| 20 | 10 |
| 21 | 12 |
| 22 | 15 |
| 23 | 18 |
| 24 | 21 |
| 25 | 23 |
| **Total** | **110** |

Vamos calcular o primeiro quartil ($Q_1$). Para tanto, precisamos identificar a posição desse valor, ou seja, precisamos o elemento do primeiro quartil $EQ_1 = \frac{110 \times 1}{4} = 27{,}5$.

O resultado acima indica que o $Q_1$ está na 27,5ª posição no rol. Para identificarmos essa posição na série precisamos de mais uma informação, a frequência acumulada:

| $x_i$ | $f_i$ | $F$ |
|---|---|---|
| 18 | 4 | 4 |
| 19 | 7 | 11 |
| 20 | 10 | 21 |
| 21 | 12 | 33 |
| 22 | 15 | 48 |
| 23 | 18 | 66 |
| 24 | 21 | 87 |
| 25 | 23 | 110 |
| **Total** | **110** | – |

Agora conseguimos identificar a classe, bem como o valor do primeiro quartil ($Q_1$). Observe que a terceira classe, que corresponde aos clientes de 21 anos, inclui do 22º ao 33º dados da série (ou do rol), como o primeiro elemento quartil corresponde à 25ª posição, podemos concluir que o $Q_1$=21. Esse resultado indica que 25% dos clientes da amostra têm entre 19 e 21 anos de idade.

Vamos utilizar o mesmo exemplo para calcular o terceiro quartil ($Q_3$). Neste caso o elemento decil será: $EQ_3 = \frac{110 \times 3}{4} = 82{,}5$. O terceiro quartil se encontra na 82,5ª posição, ou seja, está na sétima classe, o que significa que 75% dos clientes dessa amostra têm até 24 anos de idade.

Baseando-nos no mesmo exemplo, vamos calcular agora o quarto decil ($D_4$), para o qual o elemento decil é $ED_4 = \frac{110 \times 4}{10} = 44$. De acordo com o resultado, o quarto decil está na 44ª posição da série, ou na quinta classe, logo $D_4$=22, o que significa que 40% dos clientes dessa amostra têm entre 18 e 22 anos.

Ainda em relação ao exemplo acima, vamos calcular o 60º centil. Neste caso, o elemento centil será $Ec_{60} = \frac{110 \times 60}{100} = 66$. Considerando que o 60º centil se encontra na 66ª posição, temos que o $D_{60}$=23. Portanto, 60% dos clientes da amostra têm entre 18 e 23 anos de idade.

Como os dados do exemplo anterior estão dispostos numa distribuição de frequência simples ou comum, foi possível realizar o cálculo das medidas separatrizes baseando-se

apenas nos correspondentes elementos, que ao indicarem a posição das medidas nos forneceram a classe (ou a idade dos clientes). Porém, quando trabalhamos com variáveis contínuas ou com uma distribuição de frequências agrupadas em classes, precisamos calcular um valor aproximado que corresponderá aos quartis, decis ou centis desejados.

Suponha que a Corretora Alpha tenha 140 clientes que contrataram seguro residencial, e que estamos interessados em avaliá-los de acordo com a faixa salarial (em salários mínimos) de cada um deles. Para tanto, foi elaborada a seguinte distribuição de frequências:

| $x_i$ | $f_i$ | $F$ |
|---|---|---|
| 5 ⊢ 10 | 10 | 10 |
| 10 ⊢ 15 | 17 | 27 |
| 15 ⊢ 20 | 25 | 52 |
| 20 ⊢ 25 | 35 | 87 |
| 25 ⊢ 30 | 21 | 108 |
| 30 ⊢ 35 | 18 | 126 |
| 35 ⊢ 40 | 14 | 140 |
| **Total** | **140** | **–** |

Baseado no exemplo acima, vamos calcular o primeiro quartil (Q1). Assim como na distribuição anterior, precisamos identificar o elemento quartil, que neste caso será $Q_1 = \frac{140 \times 1}{4} = 35$. De acordo com o resultado, o Q1 está na 35ª posição, ou seja, na terceira classe. Lembre-se, nessa classe se encontram os valores da série (ou do rol) que estão dispostos entre 28ª e a 52ª posição. Porém, como a distribuição está agrupada em classes, precisaremos recorrer a uma fórmula, semelhante a da mediana, para poder encontrar o valor do primeiro quartil, que será a seguinte:

$$Q_i = l_{Qi} + \frac{EQ_i - F_{Ant}}{f_{Qi}} h$$

Sendo que:
$l_{Qi}$ é o limite inferior da classe em que se encontra o i-ésimo quartil;
$EQ_i$ é o i-ésimo quartil;
$F_{Ant}$ é a frequência acumulada da classe anterior àquela em que se encontra o i-ésimo quartil;
$f_{Qi}$ a frequência simples da classe em que se encontra o i-ésimo quartil; e
$h$ é a amplitude da classe em que se encontra o i-ésimo quartil.

Voltemos ao exemplo, como o $EQ_i = 35$, temos que:

$Q_i = l_{Qi} + \frac{EQ_i - F_{Ant}}{f_{Qi}} h = 15 + \frac{35-27}{25} \times 5 = 16{,}6$. Portanto, 25% dos clientes dessa carteira ganham **até** 16,6 salários mínimos.

Baseando-nos no mesmo exemplo, calculemos agora o quinto decil ($D_5$), para o qual precisamos utilizar a seguinte fórmula: $D_i = l_{Di} + \frac{ED_i - F_{Ant}}{f_{Di}} h$.

Muito bem, agora precisamos encontrar o Elemento Decil, que neste caso será $D_5 = \frac{140 \times 5}{10} = 70$. Como o quinto decil está na 70ª posição da série temos que o mesmo se encontra na quarta classe (20 ⊢ 25 salários mínimos). Neste sentido, temos: $D_i = l_{Di} + \frac{ED_i - F_{Ant}}{f_{Di}} h = 20 + \frac{70-52}{35} \times 5 = 22{,}6$. Portanto, 50% dos clientes dessa carteira ganham **até** 22,6 salários mínimos.

Lembre-se que 5º decil, assim como o 2º quartil e o 50º centil correspondem à mediana da série, assim sendo, o valor calculado acima também representa o valor mediano da série. Aplique a fórmula da mediana para tirar a prova do que está sendo afirmado.

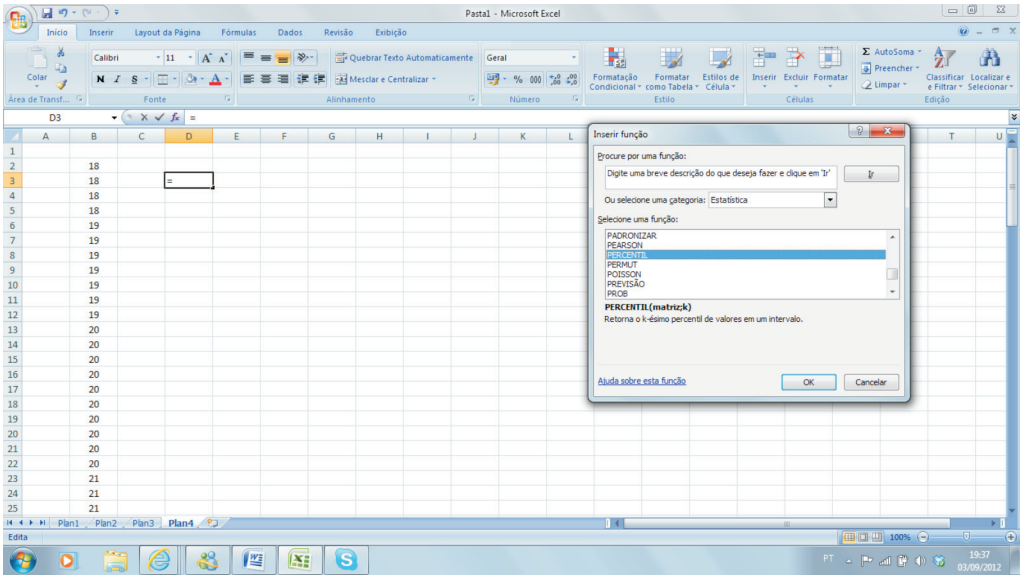
Para finalizar as medidas separatrizes, vamos calcular o 65º centil, sendo que a fórmula para o cálculo dos decis é: $C_i = l_{Ci} + \frac{EC_i - F_{Ant}}{f_{Ci}} h$. Para o exemplo o elemento centil é $C_{50} = \frac{140 \times 65}{100} = 91$. Aplicando a fórmula, temos que:

$C_i = l_{Ci} + \frac{EC_i - F_{Ant}}{f_{Ci}} h = 25 + \frac{91-87}{21} \times 5 = 26$.

O resultado indica que 65% dos clientes da carteira de seguro residencial da Corretora Alpha tem uma renda de até 26 salários mínimos.

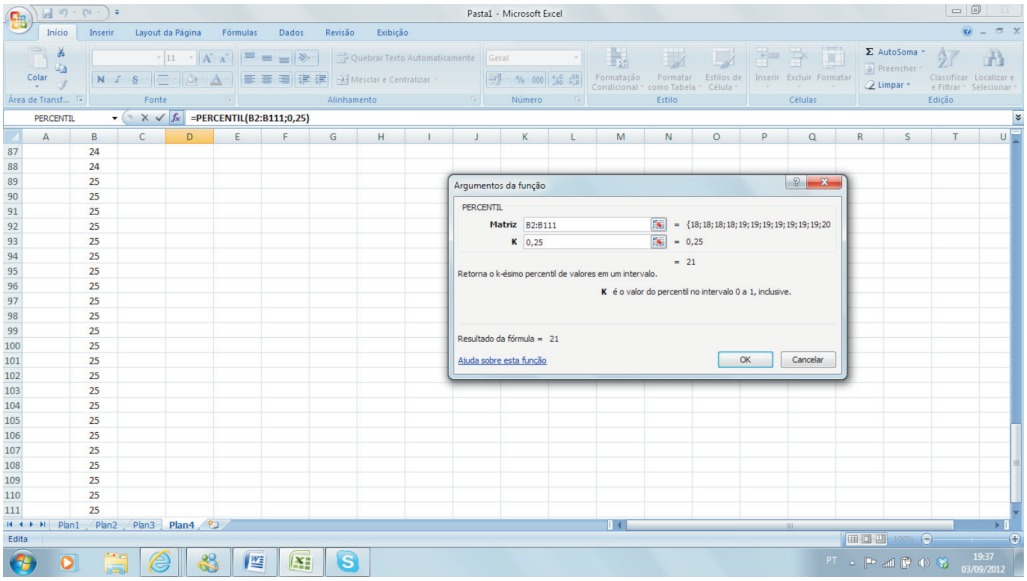
## Usando o Excel para o Cálculo dos Quartis, Decis e Centis

Para realizar esse o cálculo das separatrizes (Quartis, Decis e Centis) no Excel disponha os dados da série numa única coluna e escolha uma célula para incluir uma função. Ao clicar em $f_x$ a janela Inserir Função será aberta, em seguida escolha categoria Estatística, e selecione a função Percentil.



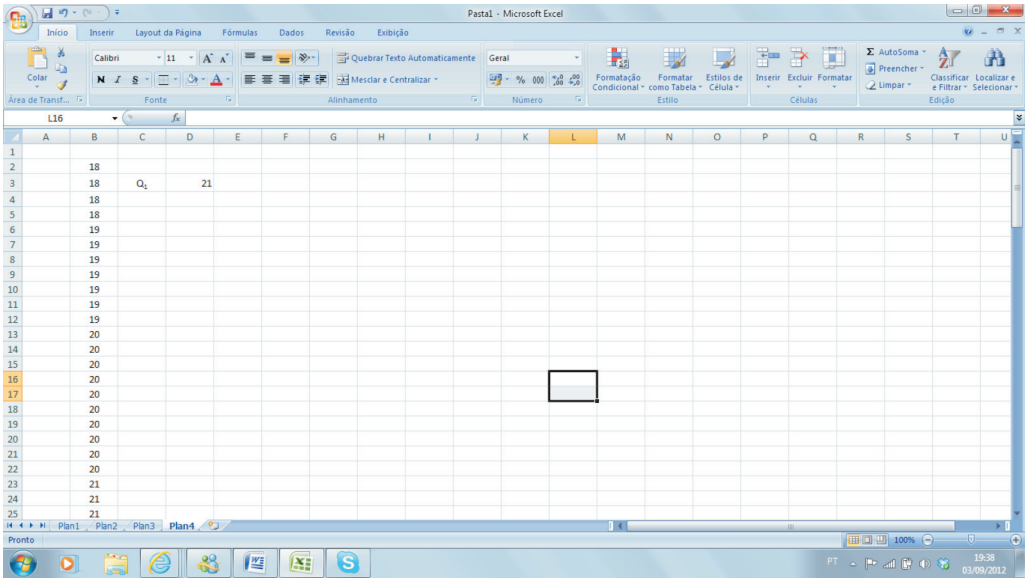

Em seguida uma nova janela, denominada Argumentos da Função se abrirá, solicitando que seja informada a Matriz, que corresponde à série de dados, e o K, que se refere ao i-ésimo percentil. Note que tanto o cálculo dos quartis quanto dos decis e dos centis será realizado a partir dessa função Percentis. Basta indicar o valor percentil proporcional à medida que busca calcular, sendo que o intervalo desse valor vai de 0 a 1.

No exemplo a seguir vamos calcular o 1º quartil, que também corresponde ao 25º centil. Para tanto, indique que o valor percentil é igual a 0,25.

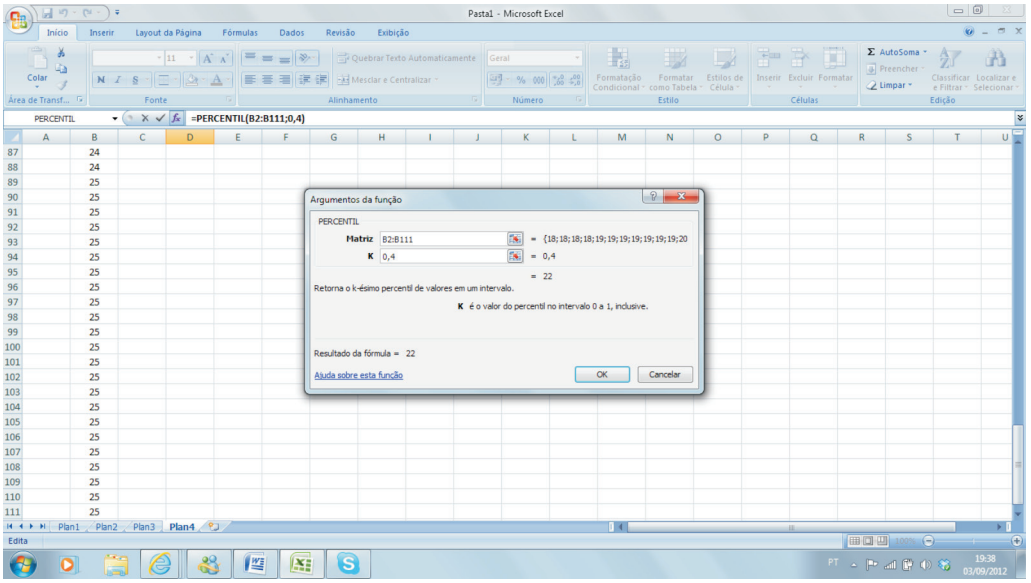



Basta clicar OK e o valor do 1º quartil aparecerá na célula escolhida para a inserção da função. Seguindo a mesma lógica, caso queira calcular o 2º quartil (que corresponde ao 5º decil, 50º centil e à mediana), o valor Percentil a ser informado deverá ser 0,5.

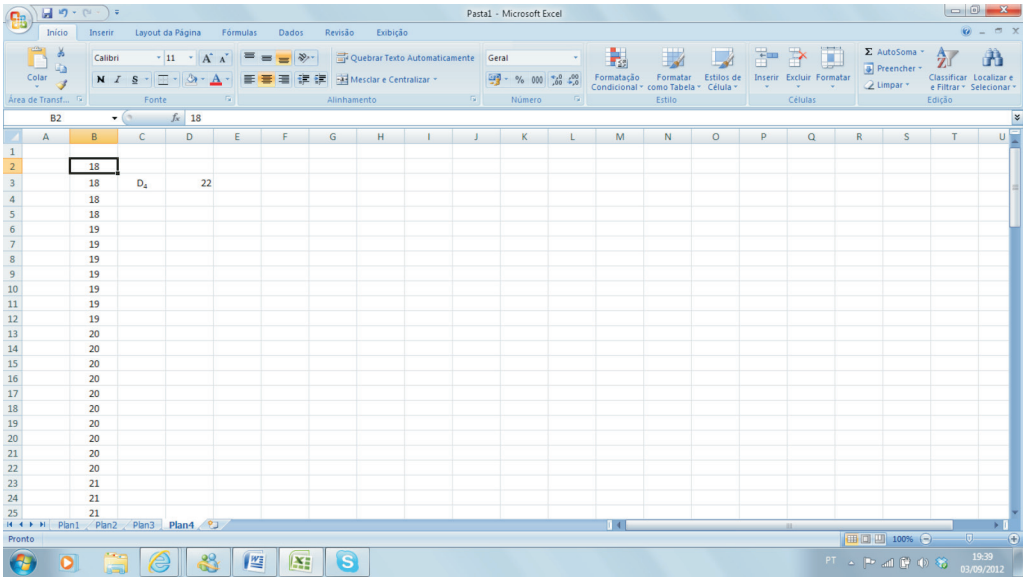
Por fim, para o cálculo do 3º quartil o valor do Percentil é igual a 0,75, sendo que o mesmo, obviamente, corresponde ao 75º centil.



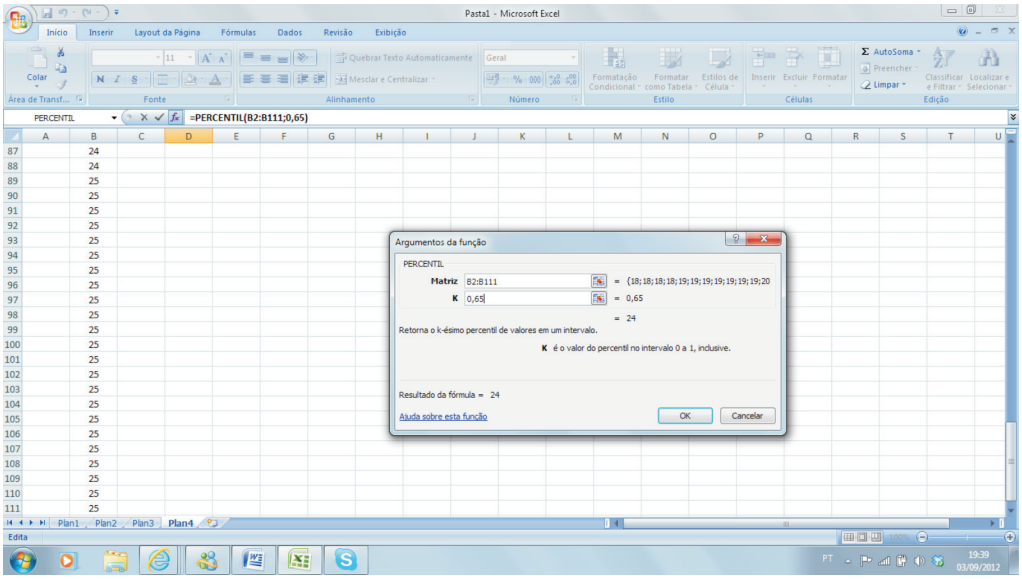
Para o cálculo dos decis a lógica é a mesma. Por exemplo, para calcular o 4º decil (que corresponde ao 40º centil), indique para o valor percentil 0,4.



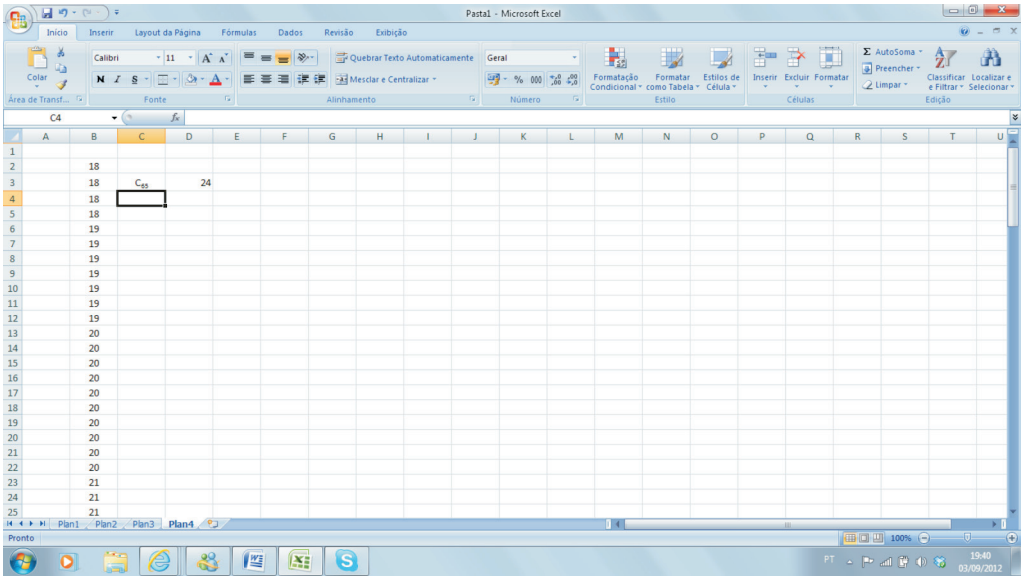

Ao clicar OK aparecerá o valor desejado na célula escolhida.



Por fim, o cálculo dos centis corresponde ao próprio Percentil, sendo que qualquer valor indicado entre 0 e 1 resultará no i-ésimo centil. Por exemplo, o cálculo do 65º centil.




O centil calculado será igual a 24, lembrando que de acordo com esse resultado 65% dos clientes analisados pela amostra têm até 24 anos de idade.




Uma vez abordadas as medidas estatísticas denominadas de Tendência Central e Separatrizes, vamos abordar, agora, as medidas que permitem complementar a análise descritiva de uma série de dados a partir de medidas que nos permitem analisar a dispersão (ou variabilidade) dos dados observados da série em relação às medidas de Tendência Central.

# Exercícios

1) Os dados a seguir representam as notas dadas em uma pesquisa de satisfação geral de um restaurante:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7,0 | 8,0 | 8,0 | 5,0 | 6,0 | 8,0 | 7,0 | 6,0 | 7,0 | 7,0 |
| 3,0 | 9,0 | 8,0 | 8,0 | 6,0 | 4,0 | 6,0 | 5,0 | 2,0 | 10,0 |
| 8,0 | 8,0 | 9,0 | 6,0 | 7,0 | 10,0 | 10,0 | 5,0 | 8,0 | 8,0 |

a) Calcule a moda, média e mediana do conjunto de dados.
b) Mostre os quartis desse conjunto de dados.

2) A tabela abaixo indica o Tamanho do Estabelecimento, a parir do número de funcionários com registro em carteira em 31 de dezembro de 2011, no setor de atividades auxiliares dos serviços financeiros, seguros, previdência complementar, planos de saúde e de resseguros no município de São Paulo, de acordo com o tamanho do estabelecimento:

| Tamanho do Estabelecimento – por número de funcionários registrados ($X_i$) | Número de Estabelecimentos ($f_i$) |
|---|---|
| De 1 a 4 | 3.011 |
| De 5 a 9 | 2.480 |
| De 10 a 19 | 3.228 |
| De 20 a 49 | 5.906 |
| De 50 a 99 | 6.217 |
| De 100 a 249 | 9.271 |
| De 250 a 499 | 11.070 |
| De 500 a 999 | 7.102 |
| Total | 48.285 |

a) Calcule a média e analise o resultado.
b) Calcule a moda e analise o resultado.
c) Calcule a mediana e analise o resultado.
d) Calcule o 1º e o 3º quartil e analise os resultados.
e) Calcule o 3º e o 7º decil e analise os resultados.
f) Calcule o 33º e o 67º centil e analise os resultados.

# Medidas de Dispersão

# 4

As medidas de dispersão são úteis para avaliar a representatividade da média, ou seja, o quanto a média representa bem um conjunto de dados e, consequentemente, na possibilidade de ser utilizada para a tomada de decisões.

As principais medidas de dispersão são a **amplitude (A)**, o **desvio-médio simples (DMS)** que também é chamado de **desvio-médio absoluto**, a **variância ($\sigma^2$)** e o **desvio-padrão ($\sigma$)**.

Estatisticamente, o conceito de dispersão é sinônimo de: variabilidade, diferença, desvio, distância.

Como dissemos no capítulo anterior, a média é uma das medidas mais utilizadas na estatística, porém sua utilização sem critérios pode causar erros de julgamento. Para que tenha real utilidade, seu cálculo deverá estar obrigatoriamente associado ao cálculo de seu desvio.

Um exemplo ilustrativo: suponha a necessidade de escolher um piloto de aviões recém-formado. Caso você examine somente as médias finais que alcançaram durante seus cursos, não observará qualquer diferença (nota 5 para ambos).

| | Piloto A | Piloto B |
|---|---|---|
| **Decolar** | 10 | 5 |
| **Pousar** | zero | 5 |
| **Nota média:** | **5** | **5** |

Tomando sua decisão somente pelo valor da média você incorrerá em um erro de julgamento grave, pois o piloto A não sabe sequer pousar o avião. Já analisando a distribuição de notas dos candidatos, podemos observar um maior equilíbrio das habilidades do piloto B, que seria uma melhor opção.

O fato de preferirmos o piloto B deve-se à comparação da sua nota média com cada uma das suas habilidades, ou seja:

| | Piloto A | Piloto B |
|---|---|---|
| Decolar | 10 | 5 |
| Pousar | zero | 5 |
| Nota média: | **5** | **5** |
| | | |
| Σ **desvios** | **10** | **0** |

**Desvio: 5 – 5 = 0 para as duas habilidades do piloto B**

Para o piloto A teríamos as seguintes diferenças (desvios) entre a média e suas notas, que devem ser calculadas em valores absolutos (módulo):

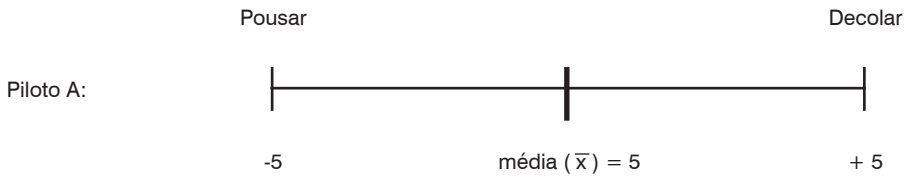
$\Sigma d_A = d_1 + d_2 = |10-5| + |0-5| = 5 + 5 = 10$ (maior desvio total)

Já para o piloto B:

$\Sigma d_B = d_1 + d_2 = |5-5| + |5-5| = 0 + 0 = 0$ (menor desvio total)

### • • • Cuidado

Se não utilizássemos o valor absoluto, teríamos para o piloto A: (5-10) + (5-0) = -5 + 5 = 0 o que daria a falsa impressão que não existem diferenças entre os dois pilotos, pois ambos teriam desvios iguais a zero! De qualquer forma, devemos lembrar que "não existem notas negativas"; o fato de ocorrer um desvio -5 indica apenas que ele está abaixo da média, e +5 está acima da média, como em uma escala.



## CÁLCULO DO DESVIO-MÉDIO SIMPLES (OU DESVIO ABSOLUTO) PARA UMA POPULAÇÃO

Costumeiramente não trabalhamos com o desvio total, mas sim com o chamado **desvio-médio simples (DMS)**, ou seja:

$$DMS = \frac{\sum d_i}{n}$$ ou ainda, de forma expandida $$DMS = \frac{\sum | x_i - \bar{X} |}{n}$$

Onde:

$\Sigma d_i$ .... somatório dos valores absolutos dos desvios

$n$ ..... número de elementos da série envolvidos

$x_i$ ..... variáveis

Ademais, é comum associarmos ao cálculo dos desvios um **coeficiente de variação percentual** (também chamado de erro percentual), expresso por uma razão entre o desvio e a média da série, ou seja:

$$CV(\%) = \frac{DMS}{\bar{X}} \times 100$$

No exemplo dos pilotos teríamos o seguinte tratamento completo:

- **Para o primeiro piloto**:

| | Piloto A | Desvios ($d_i$) |
|---|---|---|
| Decolar | 10 | $d_1 = \|10 - \mathbf{5}\| = 5$ |
| Pousar | Zero | $d_2 = \|\ 0 - \mathbf{5}\ \| = 5$ |
| $\bar{x}$ | **5** | |

$$\Sigma d_i = 5+5 = 10$$

$$\begin{cases} DMS_{PilotoB} = \dfrac{\sum d_i}{n} = \dfrac{\sum | x_i - \bar{X} |}{n} = \dfrac{10}{2} = 5 \\ CV(\%)_{PilotoB} = \dfrac{DMS}{\bar{X}} \times 100 = \dfrac{5}{5} \times 100 = 100\% \end{cases}$$

- **Para o outro piloto, temos**:

| | Piloto B | Desvios ($d_i$) |
|---|---|---|
| Decolar | 5 | $d_1 = \|5 - \mathbf{5}\| = 0$ |
| Pousar | 5 | $d_2 = \|5 - \mathbf{5}\| = 0$ |
| $\bar{x}$ | **5** | |

$$\Sigma d_i = 0 + 0 = 0$$

$$\begin{cases} DMS_{PilotoB} = \dfrac{\sum d_i}{n} = \dfrac{\sum |x_i - \bar{X}|}{n} = \dfrac{0}{2} = 0 \\ CV(\%)_{PilotoB} = \dfrac{DMS}{\bar{X}} \times 100 = \dfrac{0}{5} \times 100 = 0 \end{cases}$$

Comparando ambos os pilotos, e acrescentando um novo conceito de dispersão, denominado **Amplitude,** definido como **a diferença entre o maior e o menor valor da série (A = maior valor – menor valor)**, teríamos:

| | $\bar{X}$ | Amplitude | DMS | $CV_{\%}$ |
|---|---|---|---|---|
| **Piloto A** | 5 | 10 – 0 = 10 | 5 | 100 % |
| **Piloto B** | 5 | 5 – 5=0 | 0 | 0% |

A média 5 para ambos os pilotos não possibilitaria uma tomada de decisão segura, porém a maior amplitude para o piloto A, apoiada pelo maior desvio, indica que ele tem uma maior variação de suas notas (no caso 100%, um valor muito grande), o que permitiria escolhermos o piloto B, que possui menor desvio.

Estatisticamente, o piloto que oferece menor risco para seus passageiros é o piloto B.

## CÁLCULO DA VARIÂNCIA E DO DESVIO-PADRÃO PARA UMA POPULAÇÃO

Outra maneira de calcularmos o desvio de uma série de valores é utilizarmos o chamado desvio-padrão, que se assemelha conceitualmente ao anterior. Em vez de utilizar-se do valor absoluto (módulo) para evitar valores negativos de desvios (valores da série inferiores à média), ele faz uso do quadrado, ou seja, calcula as diferenças e eleva-as ao quadrado, garantindo sempre resultados positivos.

Esta fase do cálculo recebe o nome de cálculo da **variância**, que é representada pela **letra grega sigma elevada ao quadrado ($\sigma^2$)**.

Elevar um valor de desvio ao quadrado evita que tenhamos valores negativos, porém aumenta seu valor, que fica bastante ampliado em relação à realidade. Uma forma de "compensarmos" este aumento indevido é determinarmos a raiz quadrada da variância, que é o chamado **desvio-padrão**, representado pela **letra grega sigma (σ)**

Temos então:

$$\sigma^2 = \frac{\sum (x_i - \bar{X})^2}{n} \text{ e } \sigma = \sqrt{\sigma^2}$$

É comum associarmos ao cálculo dos desvios-padrão um **Coeficiente de variação percentual**, expresso por uma razão entre o desvio-padrão e a média da série, ou seja:

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

**Por que utilizar este artifício, elevar ao quadrado e depois extrair a raiz quadrada? Não seria mais simples trabalharmos com o valor absoluto (módulo), ou seja, usar o desvio-médio simples (DMS) em vez do desvio-padrão (σ)?**

Na realidade o cálculo do valor absoluto não é uma operação algébrica (adição, multiplicação, extração de raízes ou potenciação com valores inteiros ou fracionados), o que causará dificuldades em cálculos posteriores de inferências estatísticas.

Tal fato não acontece com a variância, que possui maior versatilidade algébrica. Por exemplo, a variância possui a propriedade aditiva: duas populações independentes, com variâncias distintas, ao terem um elemento de cada população escolhidos aleatoriamente e somados, apresentarão uma variância igual à soma da variância das populações de onde foram extraídos.

Já o uso de valores absolutos criaria dificuldades neste tipo de cálculo.

Desta forma, **é mais comum o uso do desvio-padrão (σ) do que o desvio-médio simples (DMS), no dia a dia da estatística**. Compare as duas formas de cálculo:

Na série: $\underbrace{2\ 4\ 4\ 4\ 5\ 6\ 6\ 6\ 9\ 9}$ ⟶ x = (2 + 4.3 + 5 + 6.3 + 9.2) / 10

n = 10 x = (2 + 12 + 5 + 18 + 18) / 10 = 5,5

Repare que nos cálculos dos desvios abaixo, poderíamos calcular o desvio de cada uma das medidas individualmente, um a um, ou multiplicá-los pela sua frequência ($f_i$). Por exemplo, o valor 4 aparece três vezes no rol, o que permite calcular uma vez seu desvio e multiplicá-lo por três agilizando o trabalho.

## Cálculo do DMS para Populações

Calculamos inicialmente o desvio de cada uma das medidas:

| $x_i$ | $f_i$ | $\lvert x_i - \bar{X} \rvert f_i$ |
|---|---|---|
| 2 | 1 | $\lvert 2 - 5{,}5 \rvert \cdot 1 = 3{,}5$ |
| 4 | 3 | $\lvert 4 - 5{,}5 \rvert \cdot 3 = 4{,}5$ |
| 5 | 1 | $\lvert 5 - 5{,}5 \rvert \cdot 1 = 0{,}5$ |
| 6 | 3 | $\lvert 6 - 5{,}5 \rvert \cdot 3 = 4{,}5$ |
| 9 | 2 | $\lvert 9 - 5{,}5 \rvert \cdot 2 = 9{,}0$ |
| – | 10 | $\Sigma = 22{,}0$ |

Podemos calcular agora o desvio-médio simples (DMS):

$$DMS = \frac{\sum |x_i - \bar{X}| f_i}{n} = \frac{22}{10} = 2{,}2$$

$$CV = \frac{DMS}{\bar{X}} \times 100 = \frac{2{,}2}{5{,}5} \times 100 = 40\%$$

## Cálculo do Desvio-Padrão para Populações

Calculamos inicialmente a variância de cada uma das medidas:

$$\sigma^2 = \frac{\sum (x_i - \bar{X})^2 f_i}{n}$$

| $x_i$ | $f_i$ | $(x_i - \bar{X})^2 f_i$ |
|---|---|---|
| 2 | 1 | $(2 - 5{,}5)^2 \cdot 1 = 3{,}5^2 \cdot 1 = 12{,}3$ |
| 4 | 3 | $(4 - 5{,}5)^2 \cdot 3 = 1{,}5^2 \cdot 3 = 6{,}8$ |
| 5 | 1 | $(5 - 5{,}5)^2 \cdot 1 = 0{,}5^2 \cdot 1 = 0{,}3$ |
| 6 | 3 | $(6 - 5{,}5)^2 \cdot 3 = 1{,}5^2 \cdot 3 = 6{,}8$ |
| 9 | 2 | $(9 - 5{,}5)^2 \cdot 2 = 3{,}5^2 \cdot 2 = 24{,}5$ |
| – | 10 | $\Sigma = 50{,}7$ |

Antes de calcularmos o desvio-padrão, devemos calcular a variância ($\sigma^2$) da série:

$$\sigma^2 = \frac{\sum (x_i - \bar{X})^2 f_i}{n} = \frac{50,7}{10} = 5,1$$

Agora, o cálculo do desvio-padrão ($\sigma$):

$$\sigma = \sqrt{\sigma^2} = \sqrt{5,1} = 2,3$$

$$CV = \frac{\sigma}{\bar{X}} \times 100 = \frac{2,3}{5,5} = 42\%$$

- observe que os cálculos dos dois tipos de desvios resultam em valores diferentes, porém ambos podem ser usados para medir a dispersão de uma série em torno de sua média;
- o cálculo do desvio-padrão ($\sigma$) necessita do cálculo prévio da variância ($\sigma^2$);
- o cálculo da variância ($\sigma^2$), neste primeiro momento, só tem utilidade como passo intermediário para determinação do desvio-padrão ($\sigma$), mas em tópicos avançados de estatística terá papel fundamental na análise de dados;
- as unidades dimensionais da média e dos desvios correspondem às dimensões originais dos dados, ou seja, se estivermos tratando com idades, o valor médio será expresso, por exemplo, em anos, assim como qualquer um dos desvios. A dimensão da variância ($\sigma^2$), no entanto, não terá significado prático (anos$^2$);
- qualquer um dos tipos de desvios adotados acima (DMS ou $\sigma$), apoiados pelo cálculo da amplitude (A), permitem uma avaliação segura de dados;
- se estivermos trabalhando com uma única população o cálculo do coeficiente de variação ($CV_{\%}$) é opcional;
- a maior utilidade do coeficiente de variação ($CV_{\%}$) é o de permitir uma comparação entre duas populações diferentes; e
- as ideias desenvolvidas até o momento neste capítulo são aplicáveis a populações e precisam de algumas adaptações para o uso em amostras, conforme veremos a seguir.

# TRABALHANDO COM DESVIOS EM AMOSTRAS

O cálculo dos desvios para amostras é ligeiramente diferente do cálculo utilizado para populações. Basicamente, ocorre uma alteração nos denominadores, que envolvem o tamanho da população (n) e da amostra (n-1). Algumas alterações na nomenclatura também são usuais, observe no quadro a seguir:

| Cálculo dos desvios em POPULAÇÕES | Cálculo dos desvios em AMOSTRAS |
|---|---|
| **Desvio-médio simples**<br>$DMS = \frac{\sum d_i}{n}$<br>$DMS = \frac{\sum \lvert x_i - \bar{X} \rvert}{n}$<br>$CV(\%) = \frac{DMS}{\bar{X}} \times 100$ | **Desvio-médio simples**<br>$DMS = \frac{\sum d_i}{n-1}$<br>$DMS = \frac{\sum \lvert x_i - X \rvert}{n-1}$<br>$CV(\%) = \frac{DMS}{\bar{X}} \times 100$ |
| **Variância e desvio-padrão**<br>$\sigma^2 = \frac{\sum (x_i - \bar{X})^2 f_i}{n}$<br>$\sigma = \sqrt{\sigma^2}$<br>$CV = \frac{\sigma}{\bar{X}} \times 100$ | **Variância e desvio-padrão**<br>$s^2 = \frac{\sum (x_i - \bar{X})^2 f_i}{n-1}$<br>$s = \sqrt{s^2}$<br>$CV = \frac{s}{\bar{X}} \times 100$ |

Note que ao trabalharmos com amostras, é comum denominarmos a variância por **$s^2$** e o desvio-padrão por **s,** ambos derivados do termo inglês *sample* (amostra).

O fato de dividirmos por **(n-1)** na amostra, e não por **n** como na população, deve-se ao motivo de apenas **(n-1)** valores poderem ser associados a qualquer número, antes que o último valor possa ser determinado no cálculo individual das variâncias.

A prática demonstra que a divisão por **(n-1)** faz o valor da **variância amostral ($s^2$)** tender ao valor da variância populacional **($\sigma^2$)**; caso a variância amostral fosse dividida por **n**, o valor da variância populacional ficaria subestimado.

Suponha que o exemplo da Corretora Alpha de 140 clientes que contrataram seguro residencial seja uma amostra extraída do total de clientes de todos os tipos de seguro comercializado pela corretora. Vamos avaliar esses clientes, de acordo com a faixa salarial (em salários mínimos) de cada um deles, calculando a renda média (Média) e as dispersões em relação à média (Desvio-Médio, Desvio-Padrão e Coeficiente de Variação).

| $x_i$ | $f_i$ |
|---|---|
| 5 ⊢ 10 | 10 |
| 10 ⊢ 15 | 17 |
| 15 ⊢ 20 | 25 |
| 20 ⊢ 25 | 35 |
| 25 ⊢ 30 | 21 |
| 30 ⊢ 35 | 18 |
| 35 ⊢ 40 | 14 |
| **Total** | **140** |

## Calculando a Média, o Desvio-Médio e o Coeficiente de Variação

- **Média**: $\bar{X} = \frac{\sum x_i f_i}{\sum f_i} = \frac{3200}{140} = 22{,}8571$
- **Desvio-Médio**: $DMS = \frac{\sum |x_i - \bar{X}| f_i}{n-1} = \frac{952{,}1414}{140-1} = 6{,}85$
- **Coeficiente de Variação**: $CV = \frac{6{,}85}{22{,}8571} \times 100 = 29{,}97\%$

| $x_i$ | $f_i$ | $x_i$ | $x_i f_i$ | $\lvert x_i - \bar{X} \rvert$ | $\lvert x_i - \bar{X} \rvert f_i$ |
|---|---|---|---|---|---|
| 5 ⊢ 10 | 10 | 7,5 | 75 | 15,3571 | 153,571 |
| 10 ⊢ 15 | 17 | 12,5 | 212,5 | 10,3571 | 176,0707 |
| 15 ⊢ 20 | 25 | 17,5 | 437,5 | 5,3571 | 133,9275 |
| 20 ⊢ 25 | 35 | 22,5 | 787,5 | 0,3571 | 12,4985 |
| 25 ⊢ 30 | 21 | 27,5 | 577,5 | 4,6429 | 97,5009 |
| 30 ⊢ 35 | 18 | 32,5 | 585 | 9,6429 | 173,5722 |
| 35 ⊢ 40 | 14 | 37,5 | 525 | 14,6429 | 205,0006 |
| **Total** | **140** | **–** | **3.200** | **–** | **952,1414** |

De acordo com os resultados, a média salarial da amostra de clientes da corretora é de 22,86 salários mínimos, e o desvio-médio em relação a esse valor é 6,85 salários mínimos, o que representa uma dispersão dos dados observados em relação à média de 29,97%, de acordo com o coeficiente de variação.

## Calculando a Variância, o Desvio-Padrão e o Coeficiente de Variação

| $x_i$ | $f_i$ | $x_i$ | $x_i f_i$ | $(x_i - \bar{X})^2$ | $(x_i - \bar{X})^2 f_i$ |
|---|---|---|---|---|---|
| 5 ⊢ 10 | 10 | 7,5 | 75 | 235,8405 | 2.358,4052 |
| 10 ⊢ 15 | 17 | 12,5 | 212,5 | 107,2695 | 1.823,5818 |
| 15 ⊢ 20 | 25 | 17,5 | 437,5 | 28,69852 | 717,4630 |
| 20 ⊢ 25 | 35 | 22,5 | 787,5 | 0,12752 | 4,4632 |
| 25 ⊢ 30 | 21 | 27,5 | 577,5 | 21,55652 | 452,6869 |
| 30 ⊢ 35 | 18 | 32,5 | 585 | 92,98552 | 1.673,7394 |
| 35 ⊢ 40 | 14 | 37,5 | 525 | 214,4145 | 3.001,8033 |
| **Total** | **140** | **–** | **3.200** | **–** | **10.032,1429** |

- **Variância**: $S^2 = \dfrac{\sum (x_i - \bar{X})^2 f_i}{n-1} = \dfrac{10.032,1429}{139} = 72,1737$

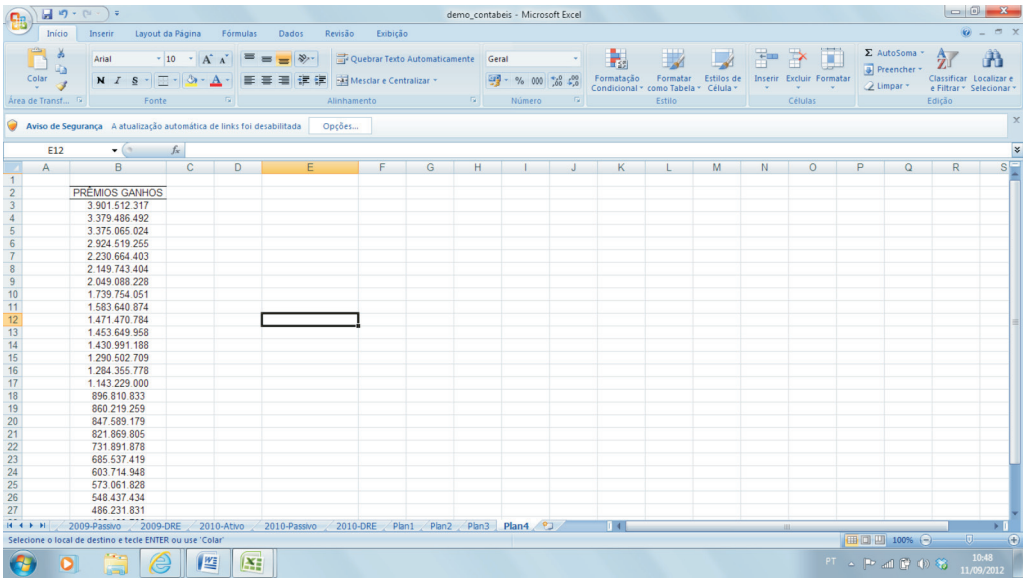
- **Desvio-Padrão**: $S = \sqrt{S^2} = \sqrt{72,1737} = 8,5$

$$CV = \frac{8,5}{22,8571} \times 100 = 37,19\%$$

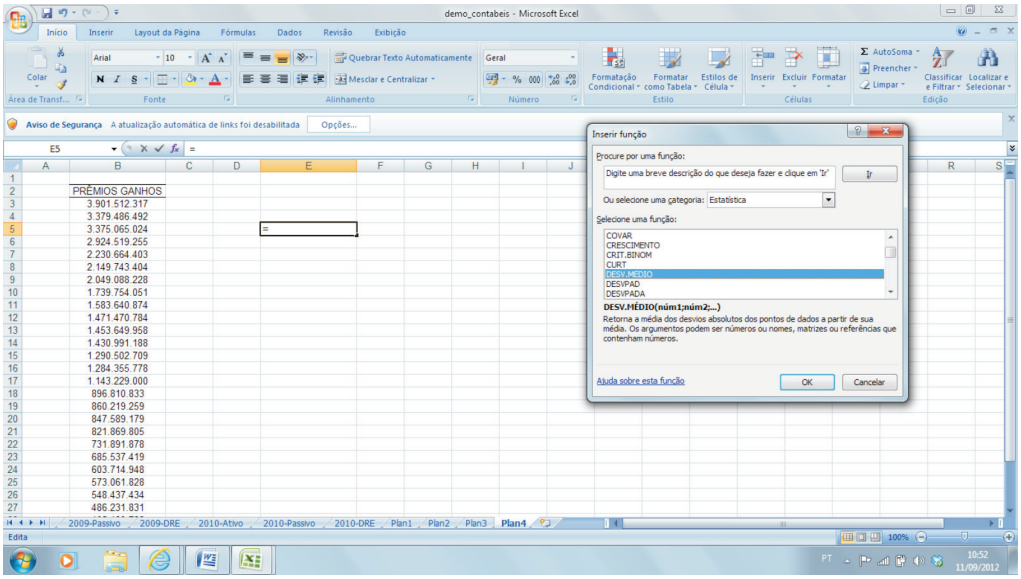
De acordo com os resultados, a média salarial da amostra de clientes da corretora é de 22,86 salários mínimos, o Desvio-Padrão é de 6,85 salários mínimos, o que representa uma dispersão dos dados observados em relação à média de 37,19%, de acordo com o coeficiente de variação.

## Usando o Excel para o Cálculo do Desvio-Médio e do Desvio-Padrão

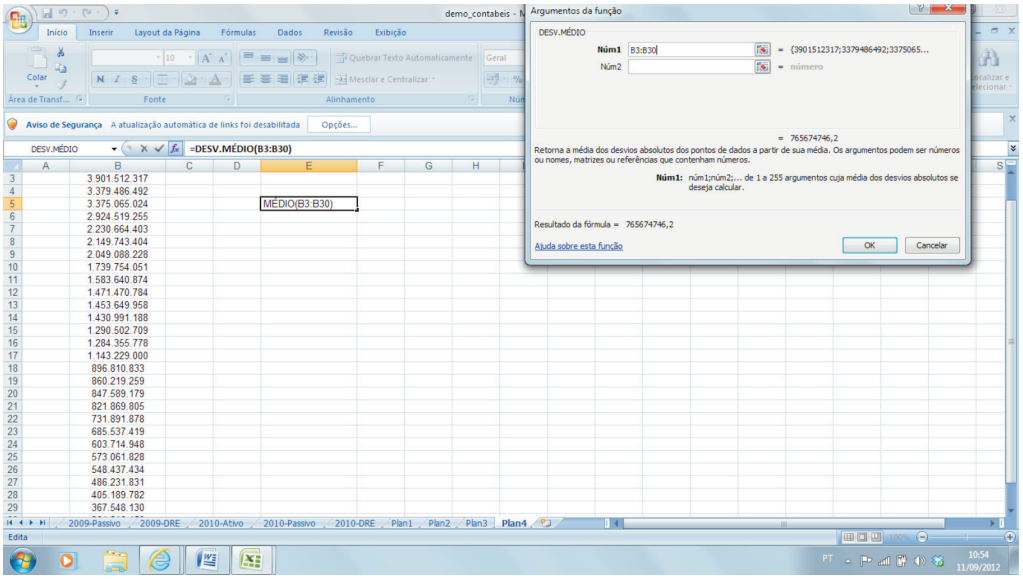
No exemplo abaixo estão dispostos os Prêmios Ganhos por 30 seguradoras no ano de 2010 (dados da SUSEP), cuja média para a amostra é de R$1.414.178.294,00.



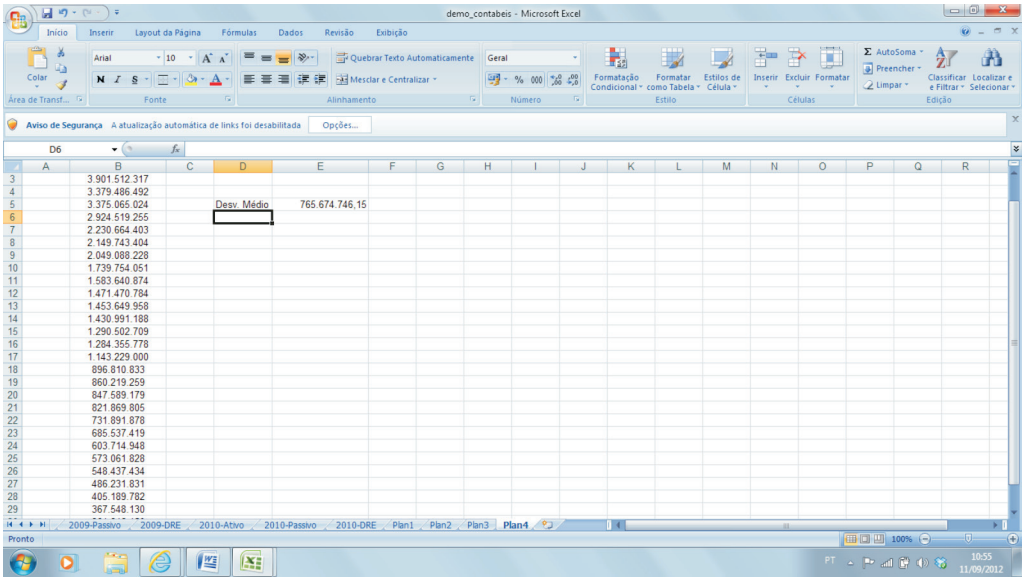

Para calcular o Desvio-Médio lembre-se que os dados devem estar dispostos numa única coluna. Escolha uma célula em branco e clique em inserir função, ou no ícone *fx*, para abrir a janela função. Selecione a Categoria Estatística, e a função Desv. Médio.



Selecionar e informar as células em que se encontram a série de dados e clicar Ok. O valor do Desvio-Médio aparecerá na célula inicialmente escolhida.

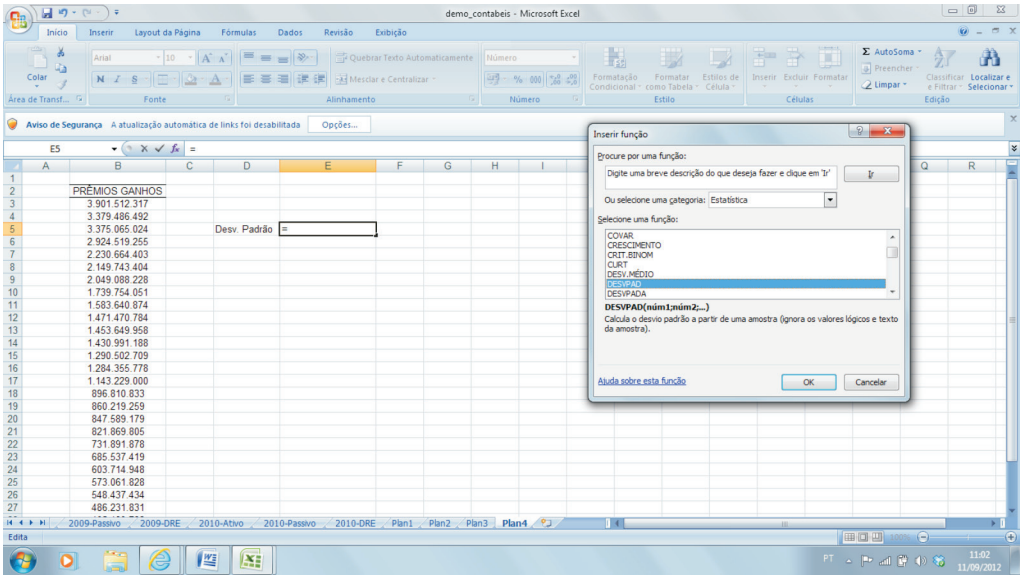



Neste exemplo, o Desvio-Médio será igual a R$765.674.746,15.

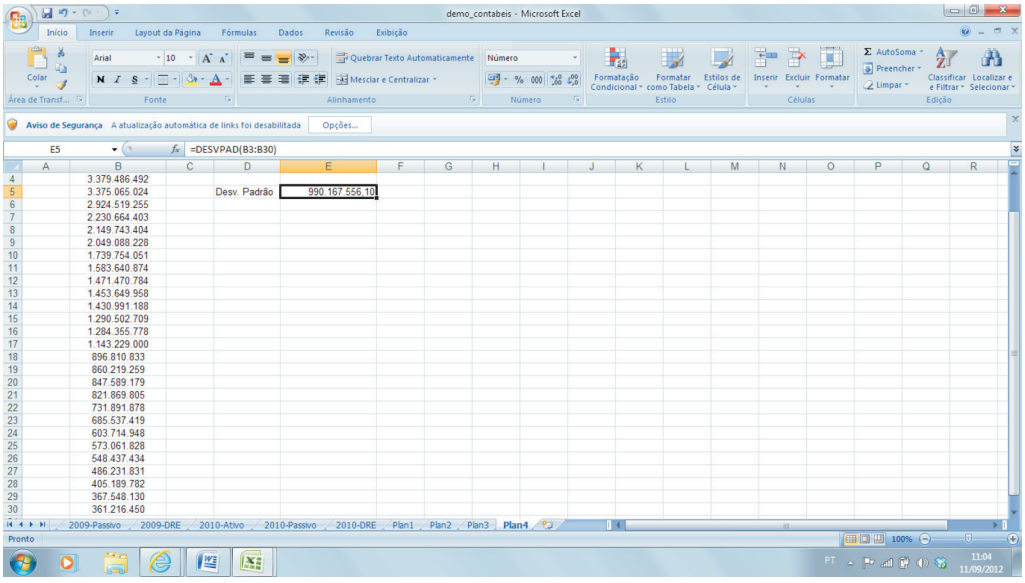


Como a Média é R$1.414.178.294,00, temos que o Coeficiente de Variação é $CV = \frac{765.674.746}{1.414.178.294} \times 100 = 54{,}14\%$. Cabe ressaltar que o Excel não tem a função Coeficiente de Variação.

Para o cálculo da Variância e do Desvio-Padrão o procedimento é o mesmo, apenas escolha a função DESVPAD:




Portanto, o valor do Desvio-Padrão da Amostra é R$990.167.556.



De acordo com esses resultados, o Coeficiente de Variação será:

$$CV = \frac{990.167.556}{1.414.178.294} \times 100 = 70{,}02\%,$$

ou seja, os prêmios recebidos pelas 30 seguradoras apresentam uma dispersão de cerca de 70% em relação ao Prêmio Médio.

## Usando a Ferramenta de Análise de Dados Estatística Descritiva do Excel

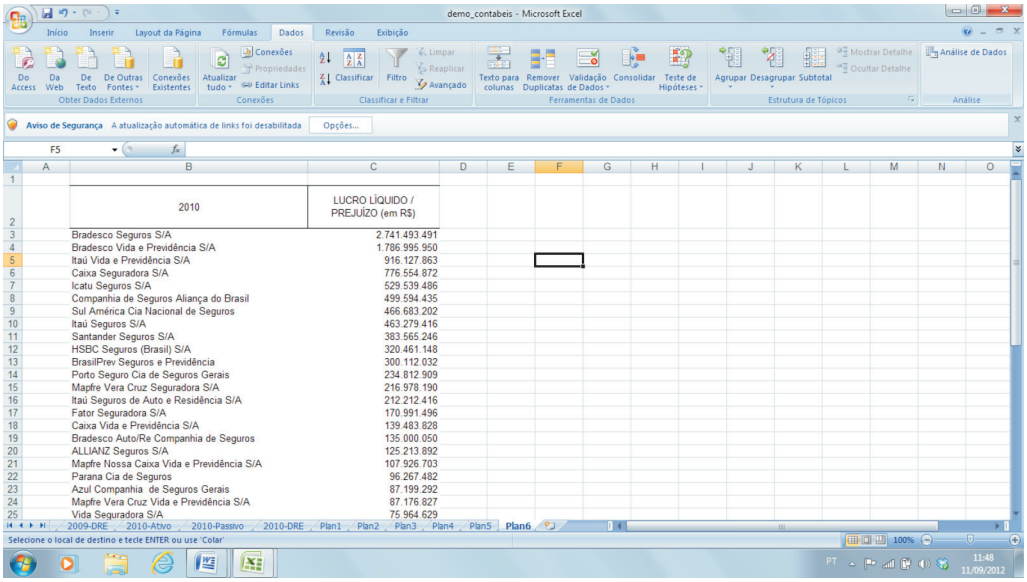
Vários dos cálculos que realizamos pelo Excel foram baseados na inclusão de uma função estatística. No entanto, o Excel tem uma Ferramenta de Análise, chamada Estatística Descritiva, que nos permite obter essas estatísticas, bem como outras que não foram contempladas pelo livro.

- **Exemplo**: Na tabela abaixo estão dispostas as 50 maiores seguradoras do Brasil, classificadas pelo Lucro Líquido (FUNENSEG):

| 2010 | LUCRO LÍQUIDO / PREJUÍZO (em R$) | 2010 | LUCRO LÍQUIDO / PREJUÍZO (em R$) |
|---|---|---|---|
| Bradesco Seguros S/A | 2.741.493.491 | Safra Vida e Previdência S/A | 58.006.644 |
| Bradesco Vida e Previdência S/A | 1.786.995.950 | Itaú Unibanco Seguros Corporativos S.A. | 47.771.541 |
| Itaú Vida e Previdência S/A | 916.127.863 | Capemisa Seguradora de Vida e Previdência | 40.402.057 |
| Caixa Seguradora S/A | 776.554.872 | J.Malucelli Seguradora S/A | 37.798.968 |
| Icatu Seguros S/A | 529.539.486 | Metropolitan Life Seguros e Previdência S/A | 34.109.573 |
| Companhia de Seguros Aliança do Brasil | 499.594.435 | HSBC Vida e Previdência (Brasil) S/A | 32.492.328 |
| Sul América Cia Nacional de Seguros | 466.683.202 | ACE Seguradora S/A | 31.028.692 |
| Itaú Seguros S/A | 463.279.416 | CHUBB do Brasil Cia de Seguros | 30.491.997 |
| Santander Seguros S/A | 383.565.246 | Cardif do Brasil Vida e PrevidênciaS/A | 28.102.140 |
| HSBC Seguros (Brasil) S/A | 320.461.148 | Companhia de Seguros Gralha Azul | 27.873.123 |
| BrasilPrev Seguros e Previdência | 300.112.032 | Panamericana de Seguros S/A | 26.447.459 |
| Porto Seguro Cia de Seguros Gerais | 234.812.909 | Mares Mapfre Riscos Especiais Seguradora S/A | 25.171.874 |
| Mapfre Vera Cruz Seguradora S/A | 216.978.190 | Santander Brasil Seguros S/A | 24.147.781 |
| Itaú Seguros de Auto e Residência S/A | 212.212.416 | Sul América Seguros de Pessoas e Previdência S/A | 24.053.828 |
| Fator Seguradora S/A | 170.991.496 | Companhia de Seguros Aliança da Bahia | 23.456.702 |
| Caixa Vida e Previdência S/A | 139.483.828 | Virgina Surety Companhia de Seguros do Brasil | 19.911.154 |
| Bradesco Auto/Re Companhia de Seguros | 135.000.050 | Liberty Seguros S/A | 16.260.060 |
| ALLIANZ Seguros S/A | 125.213.892 | Porto Seguro Vida e Previdência S/A | 15.585.224 |
| Mapfre Nossa Caixa Vida e Previdência S/A | 107.926.703 | Marítima Seguros S/A | 15.012.155 |
| Parana Cia de Seguros | 96.267.482 | Royal & Sunalliance Seguros (Brasil) S/A | 14.163.101 |
| Azul Companhia de Seguros Gerais | 87.199.292 | COFACE do Brasil Seguros de Crédito Interno | 14.146.448 |
| Mapfre Vera Cruz Vida e Previdência S/A | 87.176.827 | Indiana Seguros S/A | 13.530.295 |
| Vida Seguradora S/A | 75.964.629 | Mongeral Aegon Seguros e Previdência S/A | 13.256.485 |
| Unimed Seguradora S/A | 67.657.097 | Aliança do Brasil Seguros S/A | 12.808.821 |
| HDI Seguros S/A | 62.431.075 | Luizaseg Seguros S/A | 12.390.914 |

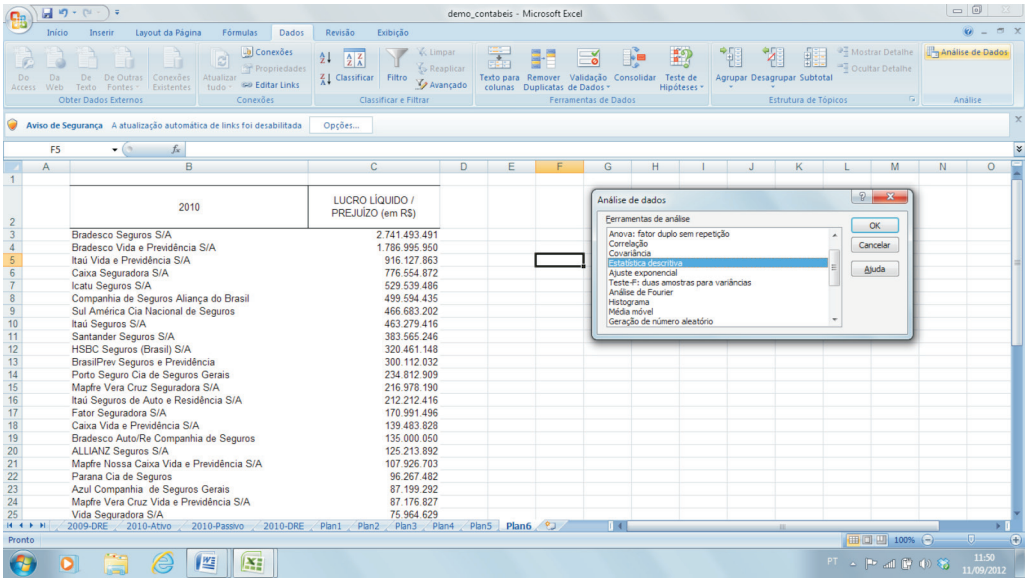
Fonte: FUNENSEG

Para calcular a Estatística Descritiva dessa série clique em Dados, e em seguida em Análise de Dados, lembrando que a série deve estar disposta em uma única coluna:

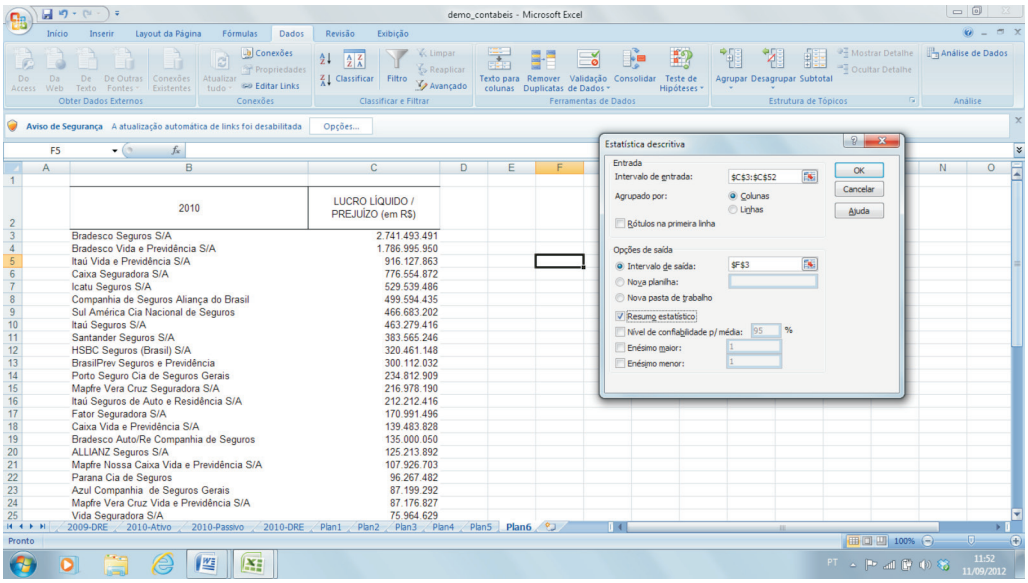



| 2010 | LUCRO LÍQUIDO / PREJUÍZO (em R$) |
|---|---|
| Bradesco Seguros S/A | 2.741.493.491 |
| Bradesco Vida e Previdência S/A | 1.786.995.950 |
| Itaú Vida e Previdência S/A | 916.127.863 |
| Caixa Seguradora S/A | 776.554.872 |
| Icatu Seguros S/A | 529.539.486 |
| Companhia de Seguros Aliança do Brasil | 499.594.435 |
| Sul América Cia Nacional de Seguros | 466.683.202 |
| Itaú Seguros S/A | 463.279.416 |
| Santander Seguros S/A | 383.565.246 |
| HSBC Seguros (Brasil) S/A | 320.461.148 |
| BrasilPrev Seguros e Previdência | 300.112.032 |
| Porto Seguro Cia de Seguros Gerais | 234.812.909 |
| Mapfre Vera Cruz Seguradora S/A | 216.978.190 |
| Itaú Seguros de Auto e Residência S/A | 212.212.416 |
| Fator Seguradora S/A | 170.991.496 |
| Caixa Vida e Previdência S/A | 139.483.828 |
| Bradesco Auto/Re Companhia de Seguros | 135.000.050 |
| ALLIANZ Seguros S/A | 125.213.892 |
| Mapfre Nossa Caixa Vida e Previdência S/A | 107.926.703 |
| Parana Cia de Seguros | 96.267.482 |
| Azul Companhia de Seguros Gerais | 87.199.292 |
| Mapfre Vera Cruz Vida e Previdência S/A | 87.176.827 |
| Vida Seguradora S/A | 75.964.629 |

Ao abrir a janela Análise de Dados, escolha a Ferramenta de Análise Estatística Descritiva:

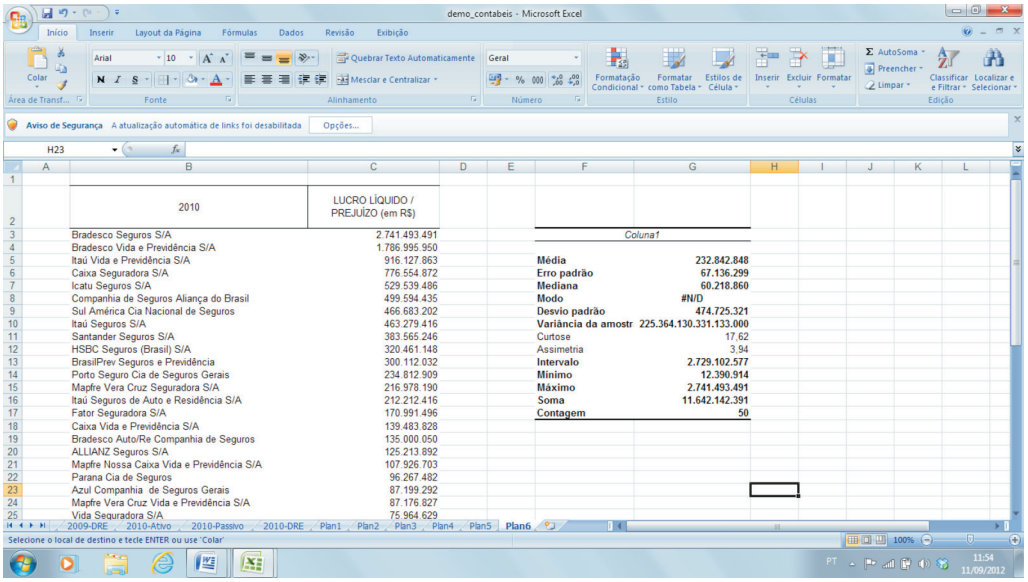


Em seguida informe: o Intervalo de Entrada, que corresponde às células em que se encontram os valores da série, o intervalo de saída, que representa a célula a partir da qual os resultados serão dispostos, clique em resumo estatístico, e OK.




| | 2010 | LUCRO LÍQUIDO / PREJUÍZO (em R$) |
|---|---|---|
| 3 | Bradesco Seguros S/A | 2.741.493.491 |
| 4 | Bradesco Vida e Previdência S/A | 1.786.995.950 |
| 5 | Itaú Vida e Previdência S/A | 916.127.863 |
| 6 | Caixa Seguradora S/A | 776.554.872 |
| 7 | Icatu Seguros S/A | 529.539.486 |
| 8 | Companhia de Seguros Aliança do Brasil | 499.594.435 |
| 9 | Sul América Cia Nacional de Seguros | 466.683.202 |
| 10 | Itaú Seguros S/A | 463.279.416 |
| 11 | Santander Seguros S/A | 383.565.246 |
| 12 | HSBC Seguros (Brasil) S/A | 320.461.148 |
| 13 | BrasilPrev Seguros e Previdência | 300.112.032 |
| 14 | Porto Seguro Cia de Seguros Gerais | 234.812.909 |
| 15 | Mapfre Vera Cruz Seguradora S/A | 216.978.190 |
| 16 | Itaú Seguros de Auto e Residência S/A | 212.212.416 |
| 17 | Fator Seguradora S/A | 170.991.496 |
| 18 | Caixa Vida e Previdência S/A | 139.483.828 |
| 19 | Bradesco Auto/Re Companhia de Seguros | 135.000.050 |
| 20 | ALLIANZ Seguros S/A | 125.213.892 |
| 21 | Mapfre Nossa Caixa Vida e Previdência S/A | 107.926.703 |
| 22 | Parana Cia de Seguros | 96.267.482 |
| 23 | Azul Companhia de Seguros Gerais | 87.199.292 |
| 24 | Mapfre Vera Cruz Vida e Previdência S/A | 87.176.827 |
| 25 | Vida Seguradora S/A | 75.964.629 |

Os resultados serão os seguintes:




| | 2010 | LUCRO LÍQUIDO / PREJUÍZO (em R$) |
|---|---|---|
| 3 | Bradesco Seguros S/A | 2.741.493.491 |
| 4 | Bradesco Vida e Previdência S/A | 1.786.995.950 |
| 5 | Itaú Vida e Previdência S/A | 916.127.863 |
| 6 | Caixa Seguradora S/A | 776.554.872 |
| 7 | Icatu Seguros S/A | 529.539.486 |
| 8 | Companhia de Seguros Aliança do Brasil | 499.594.435 |
| 9 | Sul América Cia Nacional de Seguros | 466.683.202 |
| 10 | Itaú Seguros S/A | 463.279.416 |
| 11 | Santander Seguros S/A | 383.565.246 |
| 12 | HSBC Seguros (Brasil) S/A | 320.461.148 |
| 13 | BrasilPrev Seguros e Previdência | 300.112.032 |
| 14 | Porto Seguro Cia de Seguros Gerais | 234.812.909 |
| 15 | Mapfre Vera Cruz Seguradora S/A | 216.978.190 |
| 16 | Itaú Seguros de Auto e Residência S/A | 212.212.416 |
| 17 | Fator Seguradora S/A | 170.991.496 |
| 18 | Caixa Vida e Previdência S/A | 139.483.828 |
| 19 | Bradesco Auto/Re Companhia de Seguros | 135.000.050 |
| 20 | ALLIANZ Seguros S/A | 125.213.892 |
| 21 | Mapfre Nossa Caixa Vida e Previdência S/A | 107.926.703 |
| 22 | Parana Cia de Seguros | 96.267.482 |
| 23 | Azul Companhia de Seguros Gerais | 87.199.292 |
| 24 | Mapfre Vera Cruz Vida e Previdência S/A | 87.176.827 |
| 25 | Vida Seguradora S/A | 75.964.629 |

| *Coluna1* | |
|---|---|
| **Média** | **232.842.848** |
| **Erro padrão** | **67.136.299** |
| **Mediana** | **60.218.860** |
| **Modo** | **#N/D** |
| **Desvio padrão** | **474.725.321** |
| **Variância da amostr** | **225.364.130.331.133.000** |
| Curtose | 17,62 |
| Assimetria | 3,94 |
| **Intervalo** | **2.729.102.577** |
| **Mínimo** | **12.390.914** |
| **Máximo** | **2.741.493.491** |
| **Soma** | **11.642.142.391** |
| **Contagem** | **50** |

Perceba que quase todos os resultados constantes da Ferramenta de Análise Estatística Descritiva foram abordados ao longo do livro, basta agora fazer as devidas análises.

Exercícios

1) Os dados a seguir representam as notas dadas em uma pesquisa de satisfação geral de um restaurante:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7,0 | 8,0 | 8,0 | 5,0 | 6,0 | 8,0 | 7,0 | 6,0 | 7,0 | 7,0 |
| 3,0 | 9,0 | 8,0 | 8,0 | 6,0 | 4,0 | 6,0 | 5,0 | 2,0 | 10,0 |
| 8,0 | 8,0 | 9,0 | 6,0 | 7,0 | 10,0 | 10,0 | 5,0 | 8,0 | 8,0 |

a) Calcule o desvio-médio.
b) Calcule o coeficiente de variação.
c) Calcule o desvio-padrão.
d) Calcule o coeficiente de variação.

2) A tabela abaixo indica o tamanho do estabelecimento por número de funcionários, em 31 de dezembro de 2011, no setor de atividades auxiliares dos serviços financeiros, seguros, previdência complementar, planos de saúde e de resseguros no município de São Paulo, de acordo com o tamanho do estabelecimento:

| Tamanho do Estabelecimento – por número de funcionários registrados ($X_i$) | Número de Estabelecimentos ($f_i$) |
|---|---|
| De 1 a 4 | 3.011 |
| De 5 a 9 | 2.480 |
| De 10 a 19 | 3.228 |
| De 20 a 49 | 5.906 |
| De 50 a 99 | 6.217 |
| De 100 a 249 | 9.271 |
| De 250 a 499 | 11.070 |
| De 500 a 999 | 7.102 |
| Total | 48.285 |

a) Calcule o desvio-médio.
b) Calcule o coeficiente de variação.
c) Calcule o desvio-padrão.
d) Calcule o coeficiente de variação.

Exercícios

## Exercícios Complementares

As tabelas a seguir indicam os vínculos empregatícios ativos em 31 de dezembro de 2011, no setor de atividades auxiliares dos serviços financeiros, seguros, previdência complementar, planos de saúde e de resseguros no município de São Paulo.

1) De acordo com a remuneração em salários mínimos:

| Remuneração em Salário Mínimo ($x_i$) | Número de Funcionários Registrados ($f_i$) |
|---|---|
| Até 0,50 | 11 |
| 0,51 a 1,00 | 274 |
| 1,01 a 1,50 | 3.964 |
| 1,51 a 2,00 | 5.265 |
| 2,01 a 3,00 | 8.617 |
| 3,01 a 4,00 | 6.009 |
| 4,01 a 5,00 | 3.983 |
| 5,01 a 7,00 | 5.372 |
| 7,01 a 10,00 | 4.800 |
| 10,01 a 15,00 | 4.076 |
| 15,01 a 20,00 | 2.052 |
| Total | 44.423 |

Fonte: RAIS – Relação Anual de Informações Sociais

Pede-se:

a) Complete a tabela de distribuição de frequência com a fr (%), F e Fr (%).
b) Calcule a média e analise o resultado.
c) Calcule a moda e analise o resultado.
d) Calcule a Mediana e analise o resultado.
e) Calcule o 1º e o 3º quartil e analise os resultados.
f) Calcule o 3º e o 7º Decil e analise os resultados.
g) Calcule o Coeficiente de Variação a partir do Desvio-Médio.
h) Calcule o Coeficiente de Variação a partir do Desvio-Padrão.
i) Calcule o 33º e o 67º centil e analise os resultados.
j) Calcule o Coeficiente de Variação a partir do Desvio-Médio.
h) Calcule o Coeficiente de Variação a partir do Desvio-Padrão.
i) Analise os resultados.

Exercícios

2) De acordo com a faixa etária:

| Faixa Etária ($X_i$) | Número de Funcionários Registrados ($f_i$) |
|---|---|
| 10 a 14 | 1 |
| 15 a 17 | 293 |
| 18 a 24 | 8.871 |
| 25 a 29 | 11.212 |
| 30 a 39 | 16.966 |
| 40 a 49 | 7.710 |
| 50 a 64 | 3.050 |
| Total | 48.103 |

Fonte: RAIS – Relação Anual de Informações Sociais

Pede-se:

a) Complete a tabela de distribuição de frequência com a fr (%), F e Fr (%).
b) Calcule a média e analise o resultado.
c) Calcule a moda e analise o resultado.
d) Calcule a Mediana e analise o resultado;
e) Calcule o 1º e o 3º quartil e analise os resultados.
f) Calcule o 3º e o 7º Decil e analise os resultados.
g) Calcule o Coeficiente de Variação a partir do Desvio-Médio.
h) Calcule o Coeficiente de Variação a partir do Desvio-Padrão.
i) Calcule o 33º e o 67º centil e analise os resultados.
j) Calcule o Coeficiente de Variação a partir do Desvio-Médio.
h) Calcule o Coeficiente de Variação a partir do Desvio-Padrão.
i) Analise os resultados.

3) De acordo com o grau de escolaridade

| Grau de Instrução ($X_i$) | Número de Funcionários ($f_i$) |
|---|---|
| Analfabeto | 1 |
| Até 5ª incompleto | 75 |
| 5ª completo | 132 |
| 6ª a 9ª fundamental | 241 |
| Fundamental completo | 655 |
| Médio incompleto | 1.173 |
| Médio completo | 14.668 |
| Superior incompleto | 8.871 |
| Superior completo | 22.119 |
| Mestrado | 289 |
| Doutorado | 61 |
| Total | 48.285 |

Fonte: RAIS – Relação Anual de Informações Sociais

a) Complete a tabela de distribuição de frequências.
b) Analise os resultados.

# FORMULÁRIO RESUMO

1) **Média Aritmética**

a) **Dados Brutos (Não Agrupados)**

$$\bar{x} = \frac{\sum x_i}{n}$$

b) **Dados Agrupados em Tabelas de Frequência – Variáveis Discretas**

$$\bar{x} = \frac{\sum x_i \, f_i}{\sum f_i}$$

c) **Dados Agrupados em Tabelas de Frequência – Variáveis Contínuas**

$$\bar{x} = \frac{\sum x_i \, f_i}{\sum f_i}$$ neste caso $x_i$ = ponto médio do intervalo da classe *i*.

2) **Média aritmética**

3) **Mediana**

a) **Dados Agrupados em Tabelas de Frequência – Variáveis Contínuas**

$$Md = l_{Md} + \frac{E_{Md} - F_{Ant}}{f_{Md}} h$$

Onde:
$l_{md}$ – Limite inferior da classe onde se encontra o valor mediano;
*EMd* – Elemento mediano;
$F_{ant}$ – Frequência acumulada da classe anterior à classe onde se encontra o valor mediano;
$f_{Md}$ – Frequência simples da classe onde se encontra o valor mediano; e
*h* – Amplitude do intervalo onde se encontra o valor mediano.

4) **Moda**

a) **Dados Agrupados em Tabelas de Frequência – Variáveis Contínuas**
**Moda de King:**

$$Mo = l_{Mo} + \frac{f_{post}}{f_{ant} + f_{post}}$$

Onde:
$M_o$ é o valor modal;
$l_{mo}$ é o limite inferior da classe em que se encontra o valor modal;
$f_{post}$ frequência simples da classe posterior àquela em que se encontra o valor modal;
$f_{ant}$ é o limite inferior da classe anterior àquela em que se encontra o valor modal; e
*h* é a amplitude da classe em que se encontra o valor modal.

5) **Amplitude total**

$A_t = x_{máx} - x_{mín}$, onde: $x_{máx}$ – maior valor de $x_i$ e $x_{mín}$ – menor valor de $x_i$

6) **Desvio-médio simples (DMS)**

a) **Dados Brutos (Não Agrupados)**

$$DMS = \frac{\sum |x_i - \bar{x}|}{n}$$

b) **Dados Agrupados em Tabelas de Frequência – Variáveis Discretas**

$$DMS = \frac{\sum |x_i - \bar{x}| \ f_1}{\sum f_1}$$

c) **Dados Agrupados em Tabelas de Frequência – Variáveis Contínuas**

$DMS = \frac{\sum |x_i - \bar{x}| \ f_1}{\sum f_1}$ neste caso $x_i$ = ponto médio do intervalo da classe *i*.

7) **Variância – cálculo para população**

a) **Dados Brutos (Não Agrupados)**

$$\sigma^2(x) = \frac{\sum (x_i - \bar{x})^2}{n}$$

b) **Dados Agrupados em Tabelas de Frequência – Variáveis Discretas**

$$\sigma^2(x) = \frac{\sum (x_i - \bar{x})^2 \ f_i}{\sum f_i}$$

c) **Dados Agrupados em Tabelas de Frequência – Variáveis Contínuas**

$\sigma^2(x) = \frac{\sum (x_i - \bar{x})^2 \ f_i}{\sum f_i}$ neste caso $x_i$ = ponto médio do intervalo da classe *i*.

8) **Variância – cálculo para amostra**

a) **Dados Brutos (Não Agrupados)**

$$s^2(x) = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

b) **Dados Agrupados em Tabelas de Frequência – Variáveis Discretas**

$$s^2(x) = \frac{\sum (x_i - \bar{x})^2 \ f_i}{\sum f_i - 1}$$

c) **Dados Agrupados em Tabelas de Frequência – Variáveis Contínuas**

$$s^2(x) = \frac{\sum (x_i - \bar{x})^2 \ f_i}{\sum f_i - 1}$$ neste caso $x_i$ = ponto médio do intervalo da classe *i*.

9) **Desvio-padrão – cálculo da população**

$\sigma(x) = \sqrt{\sigma^2(x)}$ onde: $\sigma^2(x)$ = variância da população.

10) **Desvio-padrão – cálculo para amostra**

$s(x) = \sqrt{s^2(x)}$ onde: $s^2(x)$ = variância da amostra.

11) **Coeficiente de variação – cálculo para população**

$$CV = \frac{\sigma(x)}{\bar{x}}$$

12) **Coeficiente de variação – cálculo para amostra**

$$CV = \frac{s(x)}{\bar{x}}$$

13) **Variância relativa – cálculo para população**

$$V(x) = \frac{\sigma^2(x)}{(\bar{x})^2}$$