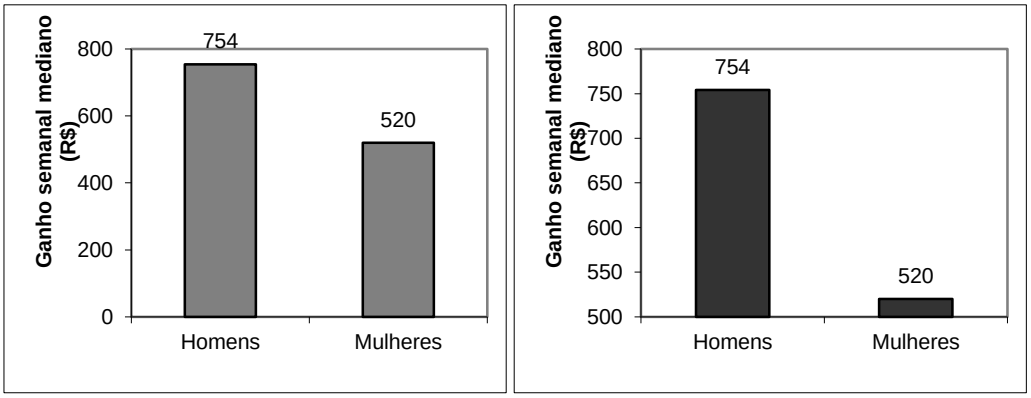
Gráficos

A organização de dados em tabelas de freqüências proporciona um meio eficaz de estudo do comportamento de características de interesse. Muitas vezes, a informação contida nas tabelas pode ser mais facilmente visualizada através de gráficos.

Deve ser notado, entretanto, que a utilização de recursos visuais na criação de gráficos deve ser feita cuidadosamente; um gráfico desproporcional em suas medidas pode dar falsa impressão de desempenho e conduzir a conclusões equivocadas. Obviamente, questões de manipulação incorreta da informação podem ocorrer em qualquer área e não cabe culpar a Estatística. O uso e a divulgação ética e criteriosa de dados devem ser pré-requisitos indispensáveis e inegociáveis.

Regras gerais para construção de gráficos:

- Os gráficos devem ser construídos de tal maneira que a frase “basta olhar para entender” seja válida.
- Um gráfico deve conter um título e os seus eixos (X e Y) devem estar rotulados incluindo as unidades (gr, R$, Kg, etc.).
- Um gráfico deve ser completo de tal maneira que todas as informações necessárias para entendê-lo devem estar sobre ele, o que o torna auto-explicativo.
- Gráficos utilizados para comparação entre si devem ter a mesma escala, caso contrário a comparação fica comprometida.
- Quando for construir gráficos de barras o valor zero deve ser o ponto de início dos eixos, caso isto não seja feito, ocorre uma distorção no gráfico.
    - <u>Gráficos enganosos:</u> muitos dispositivos visuais – como gráficos em barras ou setores – podem ser utilizados para exagerar ou diminuir a verdadeira natureza de um conjunto de dados. Veja o exemplo a seguir:



Os principais tipos de gráficos usados na representação estatística são:

- **Gráfico de setores:**

É um gráfico muito comum para representar distribuições de freqüências de variáveis qualitativas. É construído repartindo um disco em setores correspondentes às freqüências relativas de cada valor ou categoria. É particularmente útil quando o número de categorias não é grande e as categorias não obedecem a alguma ordem específica. **Vantagem:** todas as informações contidas na tabela de freqüências podem ser transportadas para o gráfico.

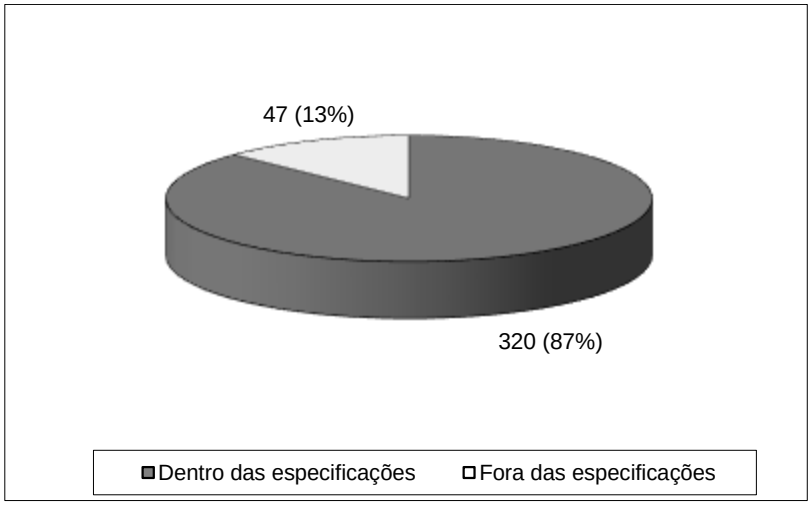


Figura 1 – Distribuição de peças produzidas por determinado setor de uma empresa de acordo com a qualidade

- **Gráfico de colunas ou barras:**

Utiliza o plano cartesiano com os valores ou categorias da variável no eixo das abcissas e as frequências ou porcentagens no eixo das ordenadas. Para cada valor da variável desenha-se uma barra com altura correspondendo à sua freqüência ou porcentagem. Esse tipo de gráfico se adapta bem à variáveis qualitativas (ou categóricas).

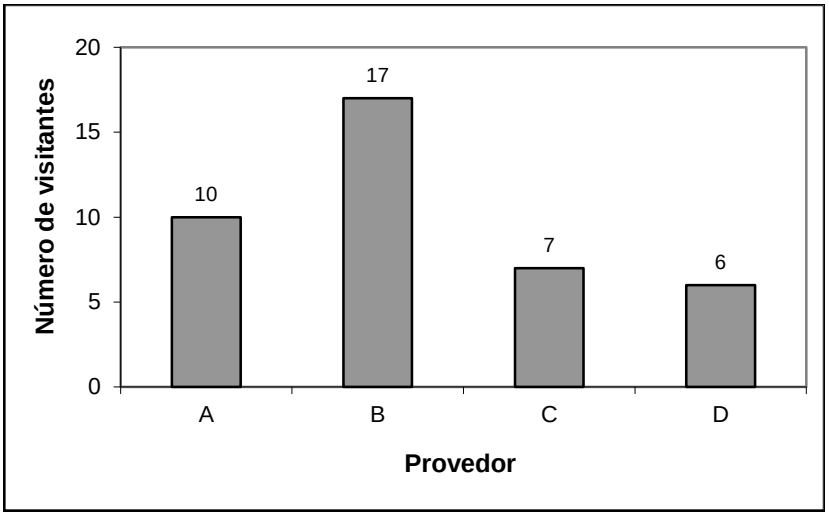


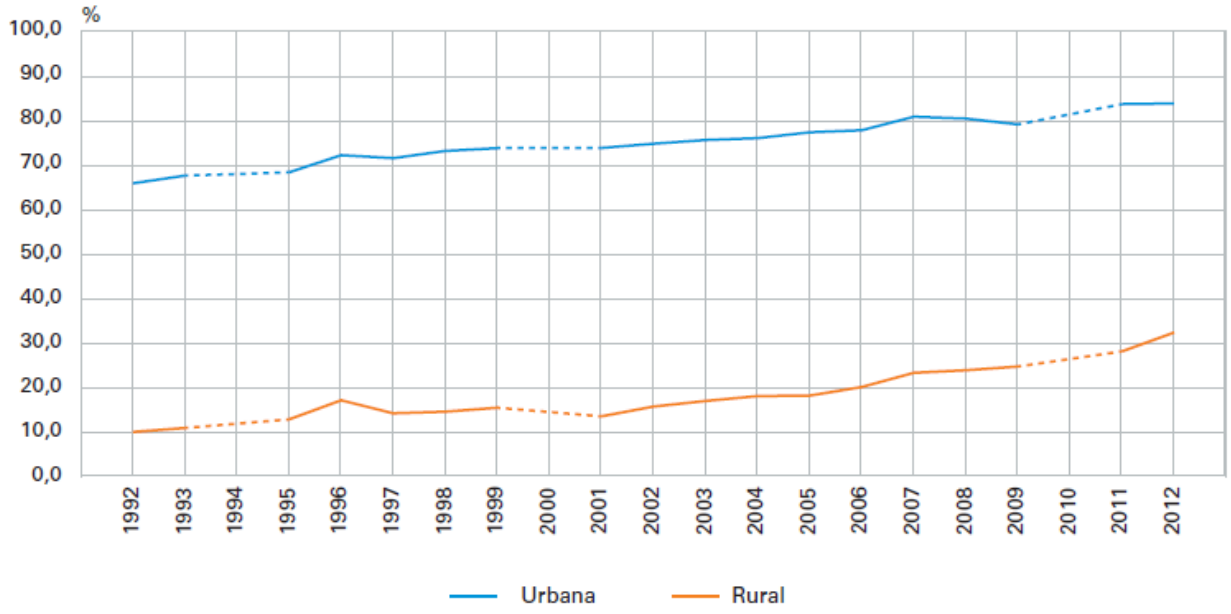
Figura 2 – Distribuição do provedor utilizado pelo visitante de um determinado *site*.

- **Gráficos de Linhas**

Sua construção é muito semelhante à do gráfico de barras porém utiliza-se linhas interligadas ao invés de barras para representar a freqüência ou a porcentagem de cada valor ou categoria da variável estudada.

É particularmente indicado para retratar séries temporais, ou seja, dados relativos a variáveis observadas anualmente, mensalmente, trimestralmente, de hora em hora, diariamente, etc.

Gráfico 62 - Proporção de moradores em domicílios particulares permanentes com esgotamento sanitário adequado, por situação do domicílio - Brasil 1992/2012



Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios 1992/2012.

Notas: 1. Exclusive população rural de Rondônia, Acre, Amazonas, Roraima, Pará e Amapá entre os anos de 1992 e 2003, a partir de 2004 a amostra inclui todo o território nacional, constituindo-se numa nova série.
2. Não houve pesquisa nos anos 1994, 2000 e 2010.

- **Histograma:**

É a forma mais usual de apresentação de distribuições de freqüências de variáveis contínuas ou de variáveis discretas cuja tabela de freqüências contém intervalos de valores (classes). Também é construído utilizando o plano cartesiano em que o eixo X é representado pelas faixas de valores da variável e o eixo Y representa as freqüências ou as porcentagens associadas a cada uma das faixas de valores. Observe a tabela a seguir:

Tabela 1 - Distribuição de freqüência do diâmetro (em centímetros) de peças fabricadas por uma indústria:

Diâmetro (em cm)	Freqüência	Porcentagem	Freqüência acumulada	Porcentagem acumulada
1,810 ⊢ 1,822	7	14,0	7	14,0
1,822 ⊢ 1,834	14	28,0	21	42,0
1,834 ⊢ 1,846	18	36,0	39	78,0
1,846 ⊢ 1,858	7	14,0	46	92,0
1,858 ⊢ 1,870	4	8,0	50	100,0
Total	50	100,0	-	-

A partir desses dados podemos construir o histograma utilizando, na escala do eixo X, os intervalos mostrados na tabela:

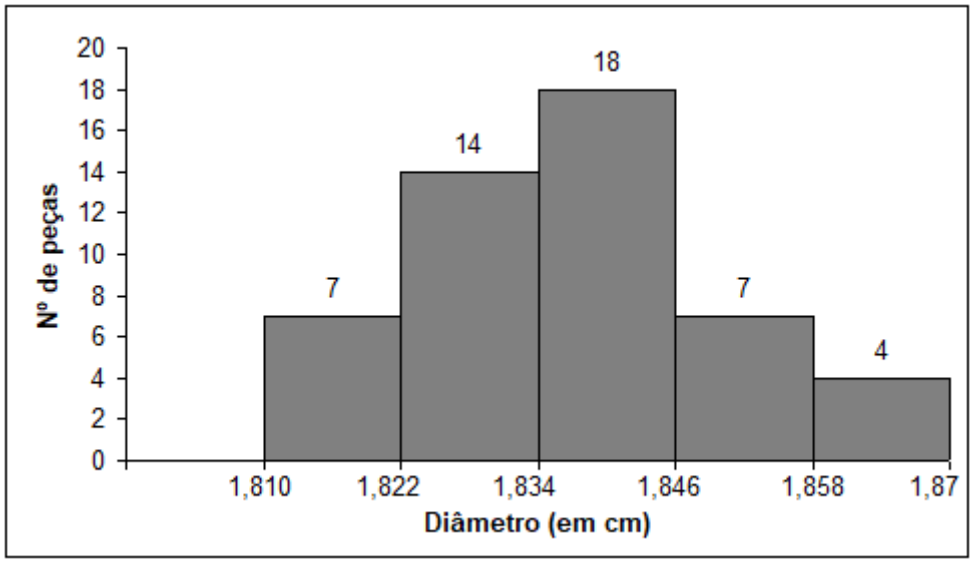


Figura 3 – Histograma para o diâmetro (em centímetros) das peças produzidas por uma indústria.

Vemos que 64% das peças foram produzidas com um diâmetro variando de 1,822 a 1,845 cm e que somente 8% das peças tinham diâmetro variando de 1,858 a 1,869cm.

O Histograma permite ver como os dados de uma variável numérica se espalham no intervalo formado pelo menor e maior valor da amostra. Simplesmente olhando o gráfico podemos perceber se os dados tendem a estar acumulados numa pequena região dentro do intervalo delimitado pelos extremos (máximo e mínimo). Ao invés disso, os dados podem estar igualmente bem espalhados dentro daquele intervalo ou pode ter duas pequenas regiões de grande concentração.

- **Gráfico de pontos:**

Quando os dados consistem em um pequeno conjunto de números, estes podem ser representados traçando-se uma reta com uma escala que abranja todas as mensurações observadas e grafando-se as respectivas freqüências como pontos acima da reta.

Observe o exemplo a seguir:

Os tempos, em segundos, entre carros que passam por um cruzamento, viajando na mesma direção foram quantificados:

6	3	5	6	4	3	5
4	6	3	4	5	2	18

O gráfico de pontos correspondente a estes dados é:

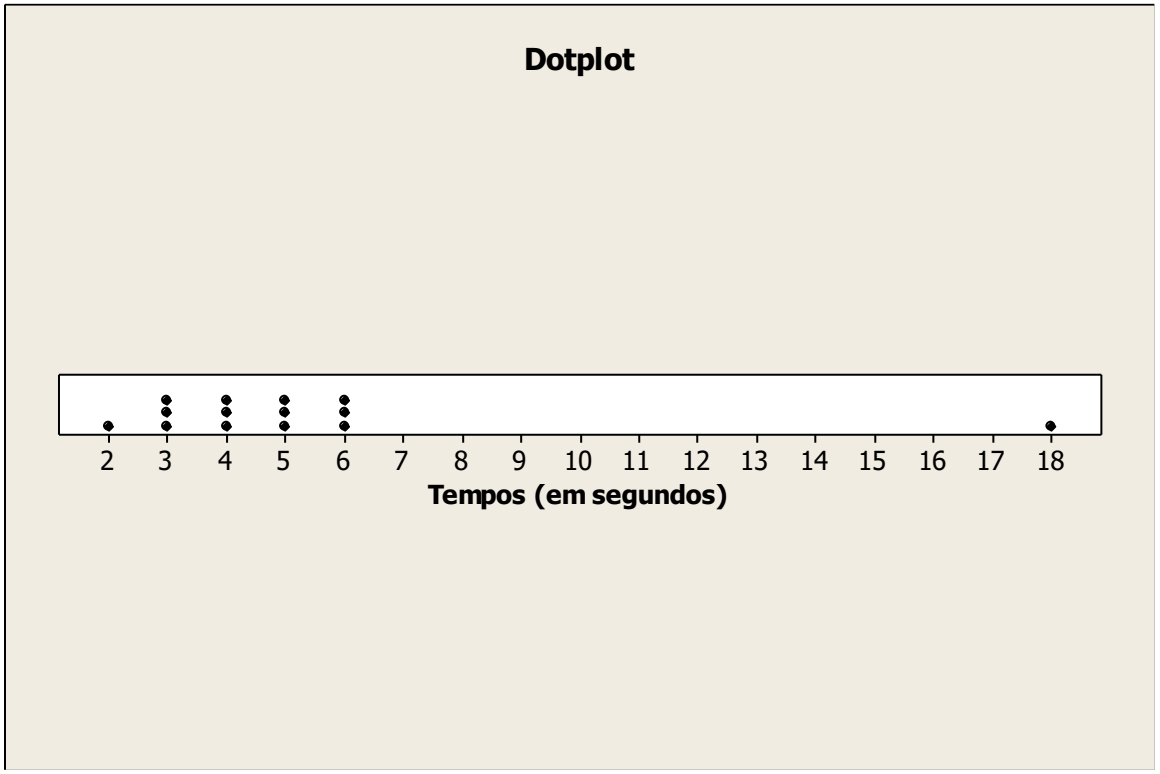


Figura 4 – Gráfico de pontos para o tempo (em segundos) entre carros que passam por um cruzamento.

Note que os valores tendem a se agrupar em torno de 4 ou 5, com exceção do último valor, 18, que se afasta grandemente do conjunto. Casos como esse, com valores extremos, devem sempre ser investigados, com vistas a uma explicação, e para que se determine se os extremos devem ou não ser excluídos do conjunto.

- **Gráfico de colunas agrupadas e de colunas empilhadas:**

Quando os dados encontram-se representados em tabelas de contingência os gráficos de colunas agrupadas e de colunas empilhadas mostram-se como boas opções de representação dos dados:

Colunas agrupadas:

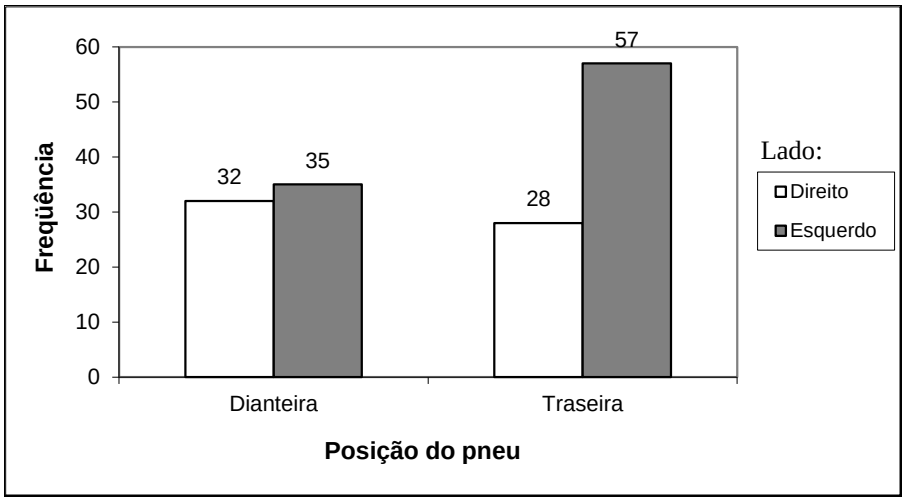
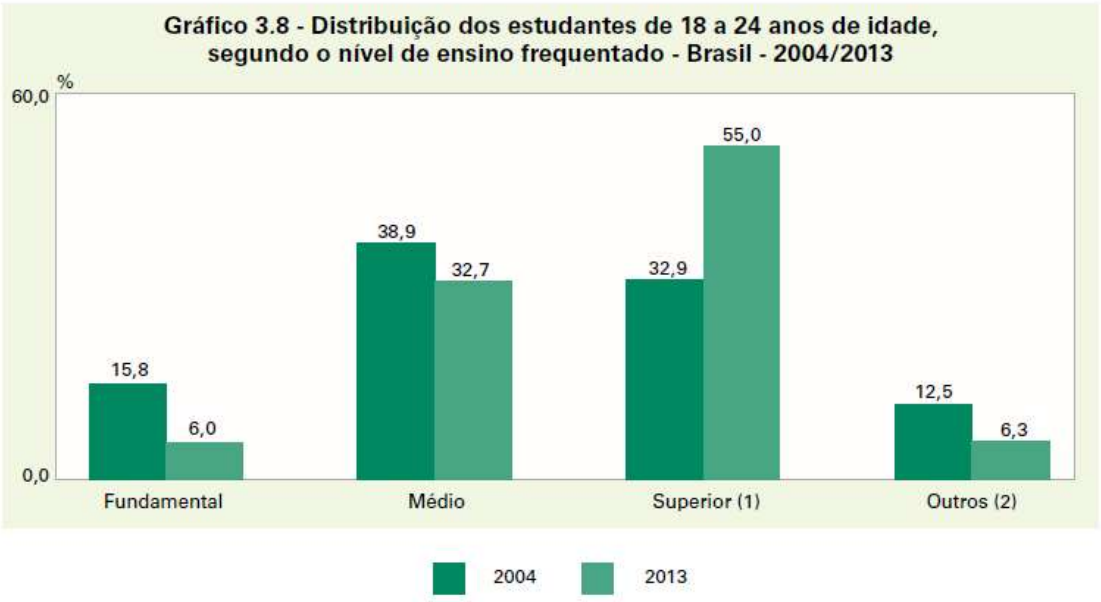


Figura 5 – Distribuição de freqüência do número de defeitos de acordo com a posição do pneu em veículos utilitários.



Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios 2004/2013.
(1) Inclusive mestrado e doutorado. (2) Pré-vestibular, supletivo e alfabetização de adultos.

Os gráficos de colunas agrupadas e de colunas empilhadas podem ser construídos utilizando-se no eixo Y, a frequência ou o percentual, porém, podemos utilizar três tipos de percentuais distintos (% da linha, % da coluna e % do total). Como identificar qual deles está sendo exibido no gráfico?

Observe que o gráfico 3.8 foi construído utilizando o percentual referente ao ano em que a pesquisa foi realizada. Como verificar isso? Somando-se os percentuais dos níveis fundamental, médio, superior e outros para o ano de 2004, por exemplo, obtemos 100%. O mesmo acontece se realizarmos as somas para o ano de 2013.

Colunas empilhadas:

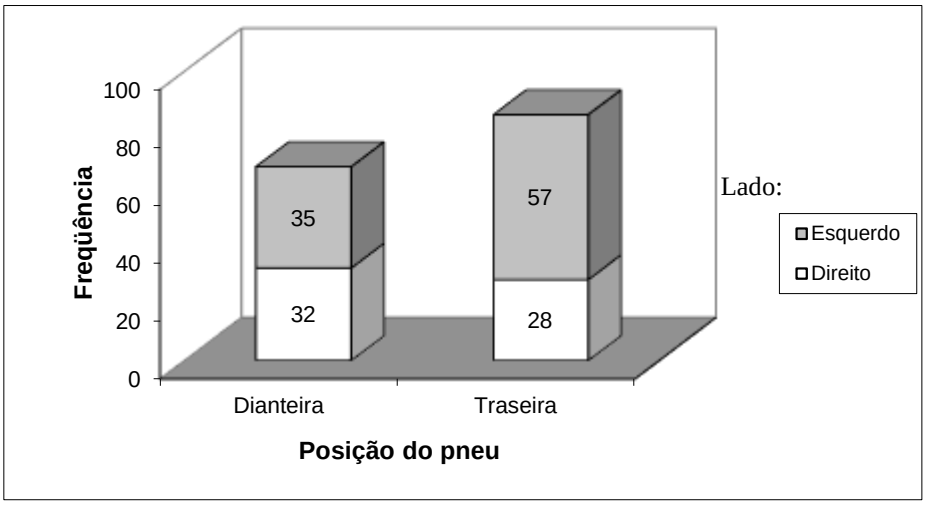


Figura 6 – Distribuição de freqüência do número de defeitos de acordo com a posição do pneu em veículos utilitários.

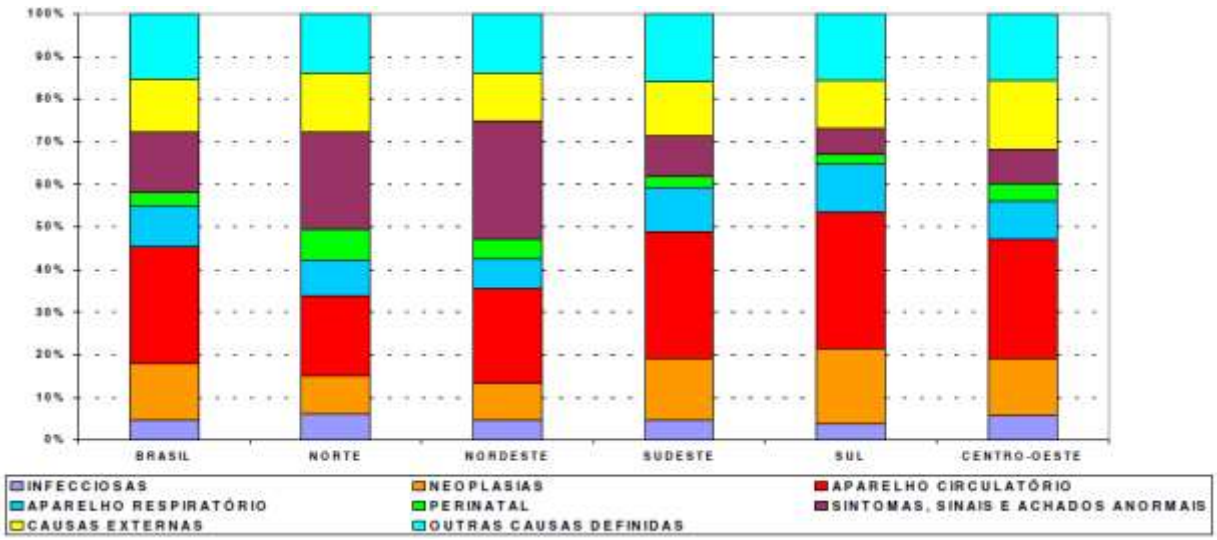


Figura 7 – Mortalidade proporcional por causas, segundo regiões, Brasil, 2001.
Fonte: DASIS/SVS – Ministério da Saúde.

A única diferença entre o gráfico de colunas agrupadas e o gráfico de colunas empilhadas é a posição das barras que podem ser dispostas lado a lado (agrupadas) ou empilhadas. Ambos são utilizados para as mesmas situações.

- **Gráfico de Ramo e Folhas (Stem-and-Leaf):**

Se a quantidade de dados é pequena, o gráfico de ramo-e-folhas (stem-and-leaf, em inglês) é bem útil. O ramo e folhas pode ser feito à mão rapidamente e permite visualizar toda a distribuição dos dados na sua faixa de variação. A ideia básica é usar os próprios dígitos dos valores que queremos visualizar para construir um histograma

No gráfico de Ramo e Folhas os dados quantitativos são separados em duas partes. Por exemplo, se tivéssemos na amostra o número 257, o primeiro ou os dois primeiros algarismos (25) formariam o ramo e o outro algarismo (7) formaria a folha do gráfico.

Para ilustrar a construção de um diagrama de ramo e folhas considere a amostra a seguir que representam as idades (em anos) de um grupo de indivíduos:

21	23	25	30	31
33	33	33	34	38
39	41	41	42	42
43	44	44	45	46
46	47	47	48	48
48	49	49	50	50
51	52	52	53	53
54	55	55	57	57
58	59	64	64	65
68	69	70	75	

Com base na amostra já ordenada, seria possível construir o gráfico de ramo e folhas mostrando na primeira coluna as dezenas e na segunda coluna as unidades de todos os dados. Observe:

```
2  135
3  01333489
4  11223445667788899
5  00122334557789
6  44589
7  05
```

Figura 8 – Gráfico de ramo e folhas para a idade em que cada ramo representa uma dezena.

A interpretação do gráfico de ramo e folhas permite visualizar onde estão concentradas as maiores e as menores freqüências. No exemplo, há uma grande concentração de freqüências nos valores centrais, de 41 a 59 anos.

Deitando a página, podemos ver a distribuição desses dados. No exemplo, vemos que a distribuição é aproximadamente simétrica (ver o material de Medidas Descritivas). Eis a grande vantagem do gráfico ramo e folhas. Podemos visualizar a

distribuição dos dados e, ainda assim, conservar toda a informação da lista original; se necessário, podemos recompor a relação original de valores.

Se o número de dados analisados for muito grande podemos, podemos dividir os ramos, duplicando-os. Dessa forma os algarismos com finais 0, 1, 2, 3 e 4 devem ser colocados na primeira parte do ramo e os algarismos com finais 5, 6, 7, 8 e 9 devem ser colocados na segunda parte do ramo. Observe:

```
2  13
2  5
3  013334
3  89
4  1122344
4  5667788899
5  00122334
5  557789
6  44
6  589
7  0
7  5
```

Figura 9 – Gráfico de ramo e folhas para a idade em que cada ramo representa meia dezena.

- ***Boxplot* ou Diagrama em Caixa:**

O *Boxplot* é um gráfico para representar dados quantitativos sendo conveniente para revelar medidas de tendência central, dispersão, distribuição dos dados e a presença de *outliers (valores discrepantes*). A construção de um diagrama em caixa exige que tenhamos o valor mínimo, o primeiro quartil (Q_1), a mediana (ou segundo quartil Q_2), o terceiro quartil (Q_3), o valor máximo e a distância interquartílica (DIQ).

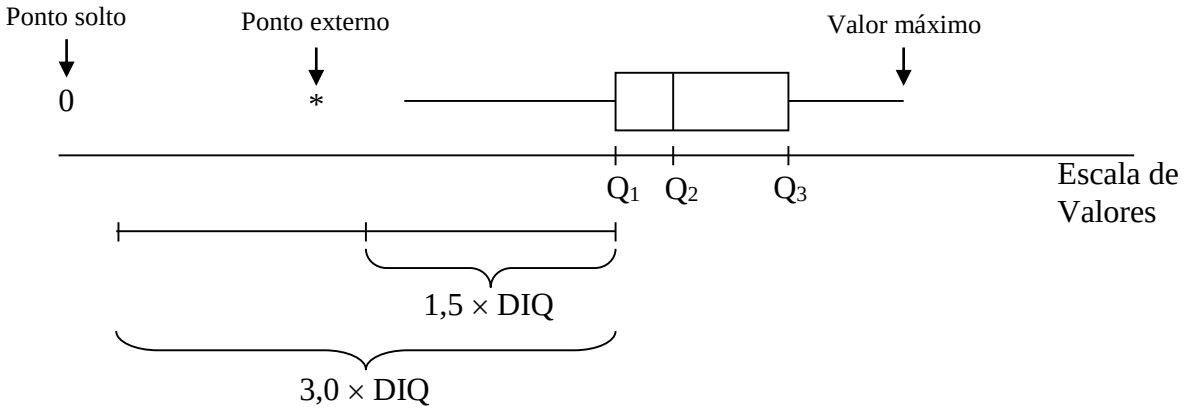
A distância interquartílica (DIQ) é obtida pela distância entre o terceiro e o primeiro quartil: $DIQ = Q_3 - Q_1$.

Para identificar a presença de valores discrepantes na amostra deve-se, primeiramente, calcular as seguintes medidas: $Q_1 – 1{,}5xDIQ$ e $Q_3 + 1{,}5xDIQ$. Elas representam os limites para detecção de outliers. Assim:

- se houver na amostra valores inferiores a ($Q_1 – 1{,}5xDIQ$), eles serão considerados valores discrepantes.

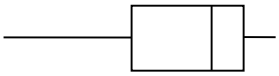
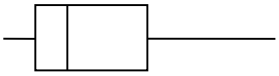
- se houver na amostra valores superiores a ($Q_3 + 1{,}5xDIQ$), eles serão considerados valores discrepantes.

Podemos esquematizar a construção de um boxplot da seguinte maneira:



Quando alguns dados apresentam-se de forma irregular em relação aos demais – com valores muito altos ou muito baixos – também denominados *outliers* ou valores discrepantes, estes pontos específicos são destacados dos demais (como mostrado no esquema acima). O destaque do *outlier* possibilita uma análise posterior mais aprofundada destes valores e a sua eventual exclusão dos estudos.

A forma da distribuição dos dados pode ser analisada por meio do diagrama em caixas da seguinte maneira:

		
Simétrica	Assimétrica à esquerda	Assimétrica à direita

Observe o exemplo abaixo que mostra a distribuição dos salários por hora trabalhada (em US$) de homens e mulheres de uma empresa:

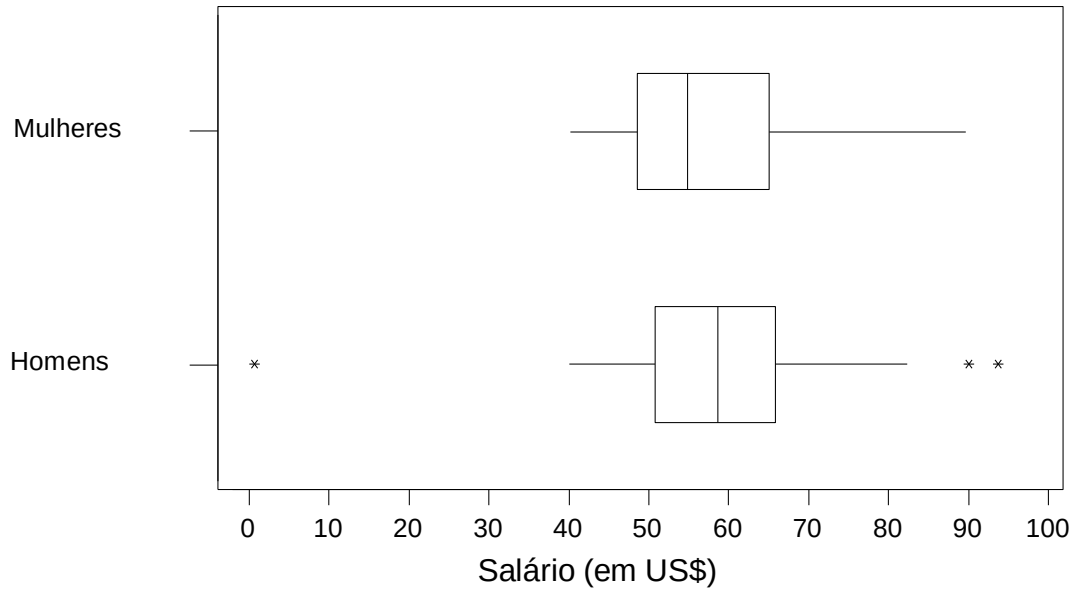


Figura 10 – Diagrama em caixa para o salário por hora trabalhada (em US$) de homens e mulheres.

Nota-se que há valores discrepantes em relação ao salário dos homens (marcados com *) o que não ocorre entre as mulheres. Percebe-se também que a distribuição do salário das mulheres é assimétrica à direita tendo uma menor mediana que a dos homens.

Um outro exemplo de aplicação do diagrama em caixa é mostrado a seguir:

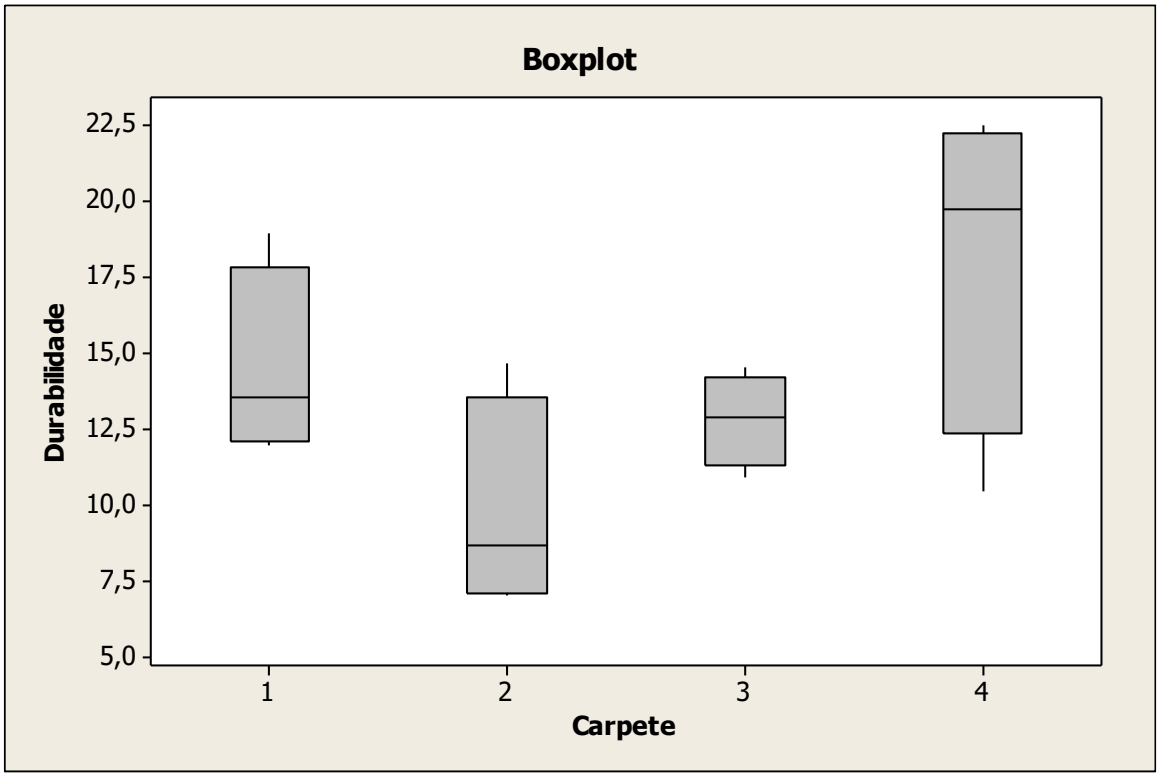


Figura 11 – Diagrama em caixa para a durabilidade de quatro marcas de carpete.

No *boxplot* mostrado na Figura 11 temos a distribuição da durabilidade de quatro marcas de carpete. Podemos observar que a duração mediana das marcas 1 e 3 são próximas, em contrapartida a marca 2 apresentou menor mediana e a marca 4 a maior mediana de durabilidade. Observa-se também que a marca 3 apresentou menor dispersão quando comparada às outras marcas analisadas.

Falhas na elaboração de gráficos

Os gráficos simplificam e agilizam o processo de análise de um conjunto de dados. Embora a visualização de informações em gráficos seja mais simples e fácil, em determinadas situações, os gráficos podem representar uma armadilha, transmitindo um conteúdo nem sempre verdadeiro.

Entre os principais erros na elaboração de gráficos, podem-se mencionar:

- Gráfico sucata: Figura demais, informação de menos. Às vezes o uso excessivo de figuras pode ocultar a informação que se deseja transmitir. Por exemplo, a evolução do salário mínimo exibida na primeira parte da Figura 2.14 com cédulas mascara a verdadeira evolução (aumento de 280%). A informação referente ao aumento é muito melhor transmitida no gráfico em forma de linha:

Figura 2.14 *Gráfico sucata* versus *representação melhorada.*

Fonte: Bruni, 2007.

- Ausência de base relativa: Os gráficos podem ocultar a verdadeira informação transmitida em função da base empregada ou sugerida na análise. Por exemplo, os gráficos da Figura 2.15 mostram a repetência em quatro classes distintas: FR, SO, JR e SR. No gráfico à esquerda, nota-se que a repetência na primeira classe foi maior. No primeiro gráfico, está exibida, apenas, a freqüência simples dos alunos que perderam o ano. Se as classes possuírem tamanhos diferentes, uma melhor transmissão de informação é feita através do uso de base relativa (%). Em relação à repetência, nota-se que, em termos percentuais e relativos, não existiram diferenças entre as classes:

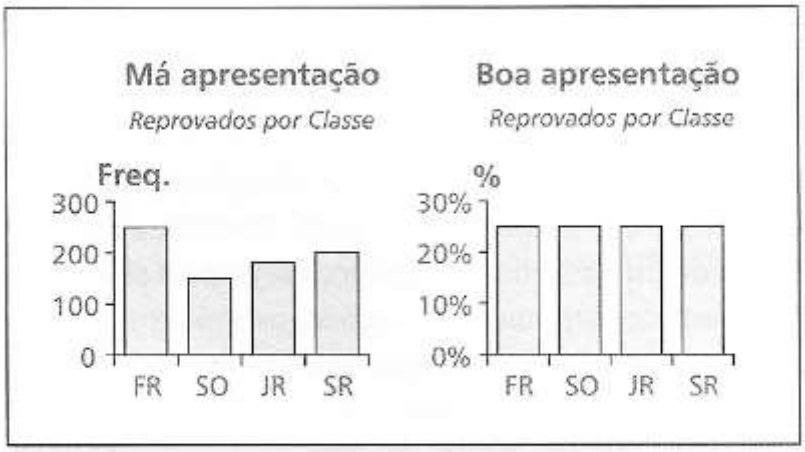


Figura 2.15 *Apresentação melhorada com base em* $F_i\%$.

Fonte: Bruni, 2007.

- Eixo vertical comprimido: Um gráfico deve ser utilizado para transmitir a informação da melhor maneira possível. As escalas utilizadas devem ser coerentes com o tamanho da figura exibida. Na Figura 2.16, na parte exibida à esquerda estão apresentadas as evoluções de vendas quadrimestrais de determinada empresa. Em função da escala utilizada, números variando de 0 a 200, as diferenças de vendas pouco podem ser percebidas. Porém, uma análise mais detalhada e melhor percepção pode ser vista na figura à direita, que adotou uma escala mais apropriada, de 0 a 50.

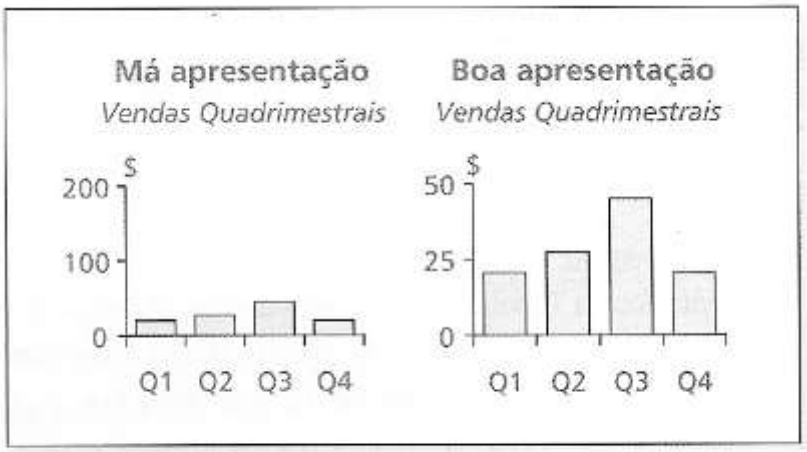


Figura 2.16 *Eixo vertical comprimido.*

Fonte: Bruni, 2007.

- Ausência do ponto zero: A ausência do ponto zero em uma figura pode disfarçar a análise, aumentando demasiadamente eventuais variações:

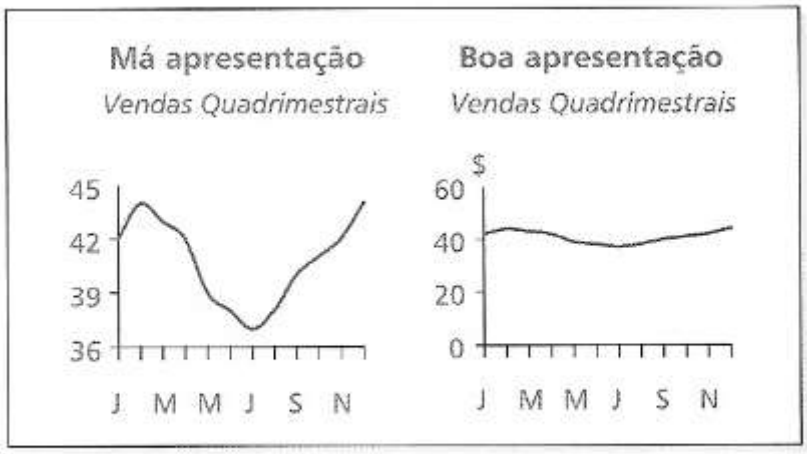


Figura 2.17 *Ausência do ponto zero.*

Fonte: Bruni, 2007.