2.1.6. Gráficos de barras, de setores e de linha

2.1.6.1. Gráfico de barras (*barchart* ou *barplot*)
(i) Para que serve; diferenças em relação ao histograma
(ii) Escala horizontal
(iii) Escala vertical
(iv) Uso de figuras em lugar dos retângulos
(v) Variações do gráfico de barras.
2.1.6.2. Gráfico de setores ou de pizza (*piechart*)
2.1.6.3. Gráfico de linha (*lineplot*)

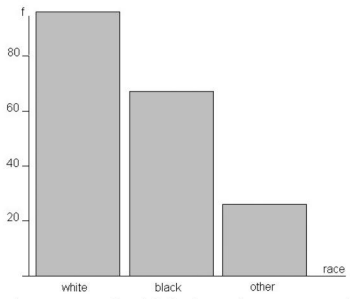
Os gráficos vistos anteriormente são os mais úteis na análise estatística, pois representam a distribuição de uma variável na amostra ou população, exibindo suas principais características - localização, dispersão, assimetria, etc. Os dois primeiros gráficos que veremos a seguir (de *barras* e de *setores*) são menos interessantes para os estatísticos, porque em geral contém pouca informação. O terceiro, o *gráfico de* linha, é usado para representar *séries temporais* (variáveis que evoluem ao longo do tempo).

Apesar de serem em geral tratados de forma sumária nos livros-texto, estes três gráficos são os mais encontrados na mídia não-especializada, como jornais e revistas – nas quais você provavelmente nunca irá encontrar um *diagrama de Tukey* ou um *diagrama de ramo-e-folhas*. Como são destinados a um público leigo, costumam ser feitos de forma chamativa, nem sempre com muita atenção aos detalhes técnicos como escalas ou unidades de medida; isto atrai a atenção dos leitores, mas às vezes pode induzi-los a tirar conclusões erradas. Veremos a seguir as características principais destes gráficos, e os cuidados que devem ser tomados para evitar erros.

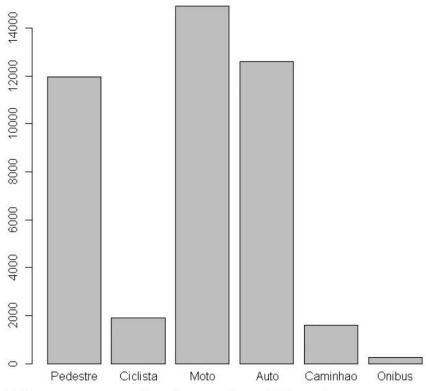
2.1.6.1. Gráfico de barras (*barchart* ou *barplot*)

(i) Para que serve; diferenças em relação ao histograma

O *gráfico de barras* (*barplot* ou *barchart*) é normalmente usado para representar a distribuição de freqüências de uma variável *qualitativa* (não-numérica); às vezes, também para variáveis *quantitativas discretas* que podem assumir poucos valores diferentes. A cada valor da variável X corresponde uma barra (retângulo), e a altura desta barra representa a freqüência (*absoluta*, *relativa* ou *percentual*) do valor.



A. Amostra de 189 de crianças recém-nascidas, classificadas pela raça da mãe



B. Mortes no trânsito, classificadas por meio de transporte, Brasil 2011. (Dados: Waiselfisz, 2013)

Figura 1. Gráficos de barras com variáveis nominais no eixo horizontal

Os gráficos (Fig. **1**) se parecem um pouco com o histograma, mas há algumas diferenças importantes. Primeiro, o histograma é usado para representar variáveis *quantitativas contínuas*, não variáveis qualitativas ou quantitativas discretas. Não existem, por isso, intervalos separando um retângulo do outro. Segundo, no histograma a freqüência é representada pela *área* do retângulo, não pela *altura*; no gráfico de colunas, a freqüência é sempre representada pela *altura*. Note que há uma certa confusão na terminologia sobre este gráfico - alguns programas estatísticos o chamam de *histogram* (por exemplo, o SPSS); contudo, preferimos manter o termo "histograma" para os gráficos que representam distribuições de frequências de dados agrupados, e chamar este de "gráfico de barras".

(ii) Escala horizontal

No gráfico de barras, se a variável for *qualitativa nominal*, os retângulos podem ser ordenados por qualquer critério. É mais comum ordená-los pelas alturas, mais isto é feito apenas para melhorar a aparência do gráfico. No gráfico da Fig. **1A**, por exemplo, a variável é *race* (a raça da mãe), numa amostra de 189 crianças nascidas num hospital americano; na Fig. **1B**, a variável é o tipo de meio de transporte. Se a variável for *ordinal* ou *quantitativa discreta*, é claro que a ordenação da variável tem que ser respeitada; um exemplo é o gráfico da Fig. **2A**, que mostra no eixo horizontal o número de gestações anteriores de cada uma de 499 parturientes em um hospital brasileiro, e no eixo vertical a freqüência relativa de cada número.

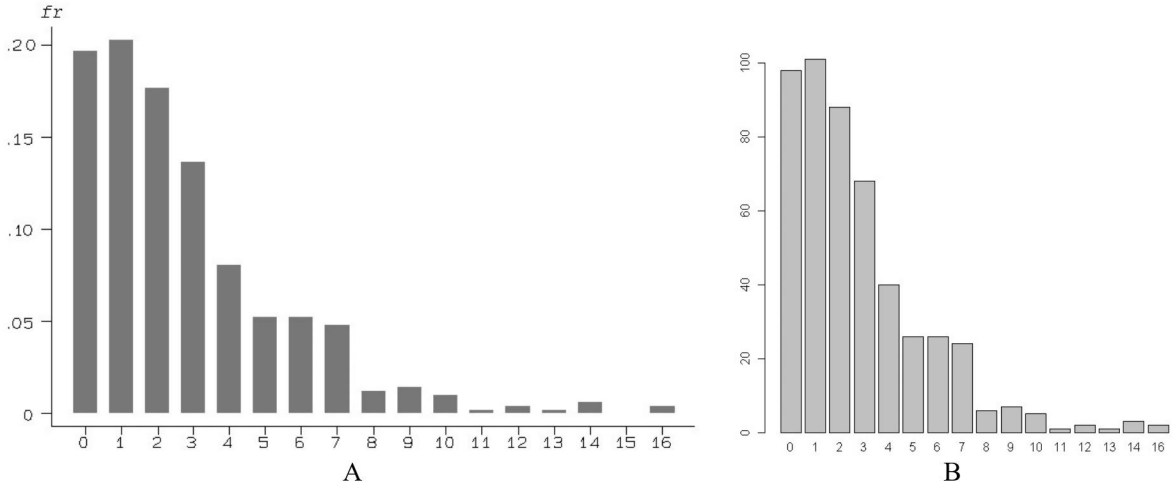


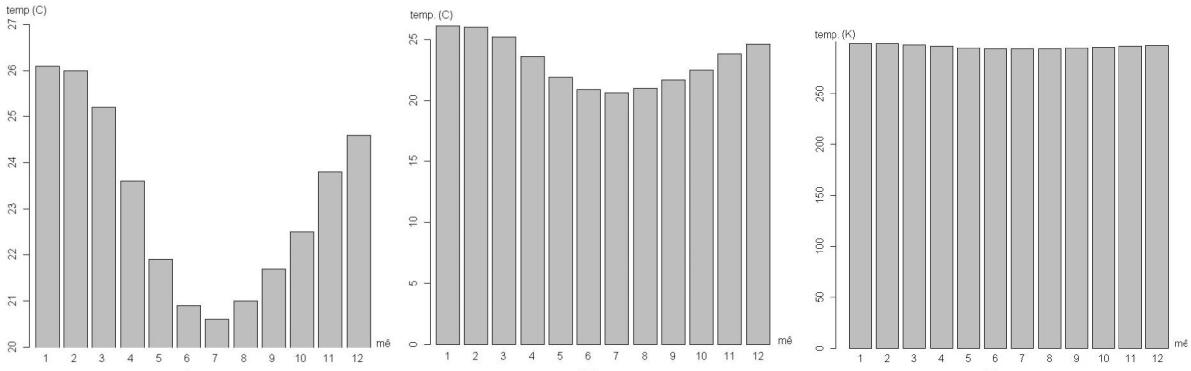
Fig. 2. Número de gestação anteriores de 499 parturientes

Para fazer este gráfico no R, é preciso verificar se algum valor da variável tem freqüência nula: o R não considera a variável na horizontal como *numérica*, mas como *nominal*, e por isso não considera as colunas vazias. Nos mesmos dados da Fig. **2A**, por exemplo, o R não acrescenta o valor "15" no eixo horizontal, e o resultado é o gráfico com escala incorreta na Fig. **2B**.

(iii) Escala vertical

Na maioria das vezes, a escolha da escala do eixo vertical não apresenta nenhuma dificuldade. Se a variável neste eixo tiver um zero natural (como nas figuras acima), a escala provavelmente irá abranger de zero até o valor máximo que deve ser representado.

Há porém situações em que isto não pode, ou não deve ser feito. A Fig. **3** mostra por exemplo as temperaturas médias mensais na cidade do Rio de Janeiro. No gráfico A, a escala vertical ficou restrita ao intervalo de 20°C a 27°C. No gráfico B, a escala incluiu o zero, e abrange de 0°C a 27°C. No gráfico C, o eixo vertical foi graduada em graus *Kelvin* (°K), que têm um zero natural (não há valores °K negativos). Qual destes é o "correto"? Com certeza não é o gráfico C; primeiro, porque os leitores não estão interessados em graus Kelvin; segundo, porque o objetivo do gráfico é mostrar o padrão de variação da temperatura ao longo dos meses, e este gráfico não mostra quase variação nenhuma. O gráfico A mostra claramente que a temperatura da cidade segue um padrão sazonal claro, com mínima em julho e máxima em janeiro. O gráfico B mostra o mesmo padrão, mas indica também que esta variação é na verdade muito pequena – a cidade tem um clima muito estável o ano todo (se comparada, por exemplo, a uma cidade européia como Budapeste, que tem média de 21,2°C no verão e -0,9 °C no inverno). Não há um gráfico que seja sempre o "correto"; o gráfico deve ser escolhido de acordo com a intenção da publicação: se pretende enfatizar a variação mensal, ou enfatizar a estabilidade do clima ao longo do ano.



A B C

Figura 3. Temperatura média mensal na cidade do Rio de Janeiro

(iv) Uso de figuras em lugar dos retângulos

Com este tipo de gráfico pode ocorrer um problema, que já foi apontado por vários estatísticos como Tufte (1998) e Huff (1982): no intuito de chamar a atenção dos leitores, os ilustradores às vezes fazem gráficos usando figuras que representam algum objeto relacionado ao tema, ao invés de retângulos; por exemplo, quando o assunto é a Economia, usando notas de dólares. Um gráfico destes está na Fig. **4**, que mostra os cinco maiores PIBs do mundo em 2020, usando figuras que representam notas de dólar. Os PIBs de cada país são proporcionais às alturas das notas, de acordo com a escala que está no eixo vertical; no entanto, a maioria dos leitores provavelmente terá a impressão de que os PIBs são proporcionais às *áreas* das notas, não às *alturas*. O ilustrador pode argumentar que os valores estão claramente indicados na escala vertical, que o leitor pode consultar; a maioria dos leitores, porém, não presta atenção nas escalas, quando folheia um jornal ou revista. Se algum leitor examinar a escala, haverá na sua mente um conflito entre a informação lógica dada na escala e a informação mais imediata dada pelos seus olhos. O PIB da China ($11,2 trilhões) é igual a 60% do PIB dos EUA ($18,6 trilhões), isto é, um pouco mais de metade; a área da nota correspondente à China, porém, é 36% da área da nota dos EUA, isto é, um pouco mais de um terço. Isto pode levar leitores a acreditar que a diferença entre as duas economias é muito maior do que é na realidade.

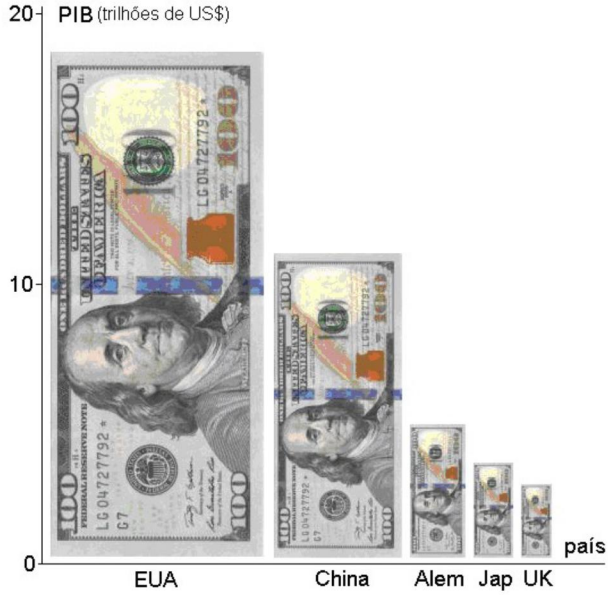


Figura 4. Os cinco maiores PIBs do mundo, 2016
(dados: *The Economist, 2019*)

A situação fica ainda pior se forem usadas figuras representado objetos sólidos, tridimensionais, como na Fig. **5**, que compara novamente os PIBs dos mesmo países, mas usando agora figuras de objetos tridimensionais ao invés de notas. O PIB da China é igual a 60% do PIB dos EUA, mas o volume do saco de dinheiro que representa a China é apenas 21% daquele dos EUA; isto dá a impressão de que a economia americana é quase cinco vezes maior do que a chinesa, e aumenta ainda mais a discrepância entre a informação fornecida pela escala e a informação apreendida intuitivamente pelos olhos.

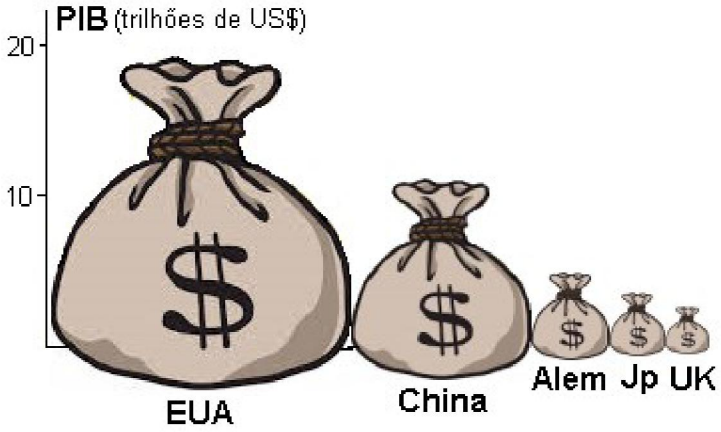


Figura 5. Cinco maiores PIBs do mundo, 2016
(dados: *The Economist, 2019*)

A correto, portanto, é fazer que o valor da variável representada seja proporcional a apenas *uma* das dimensões da figura: no gráfico de barras, proporcional à *altura* das barras, enquanto que a *largura* (e a *profundidade*, se forem usadas três dimensões) sejam constantes. No caso do gráfico dos PIBs, se o ilustrador insiste em usar figuras de objetos ao invés de simples retângulos, uma solução possível seria usar uma pilha de moedas para representar cada país; a altura das pilhas seria proporcional ao PIB, e as moedas teriam sempre os mesmos diâmetros.

(v) Variações do gráfico de barras.

Há duas variações que são encontradas com freqüência. A primeira delas é uma simples rotação do gráfico, de forma que barras fiquem na horizontal, como na Fig. **6**. Esta forma do gráfico tem exatamente a mesma informação que a forma usual, mas pode às vezes ser vantajosa em termos do *layout* da página (permite, por exemplo, que haja mais espaço para os nomes das colunas). Alguns autores preferem chamar esta forma de *gráfico de barras* e a outra de *gráfico de colunas*, mas a maioria não faz distinção entre elas.

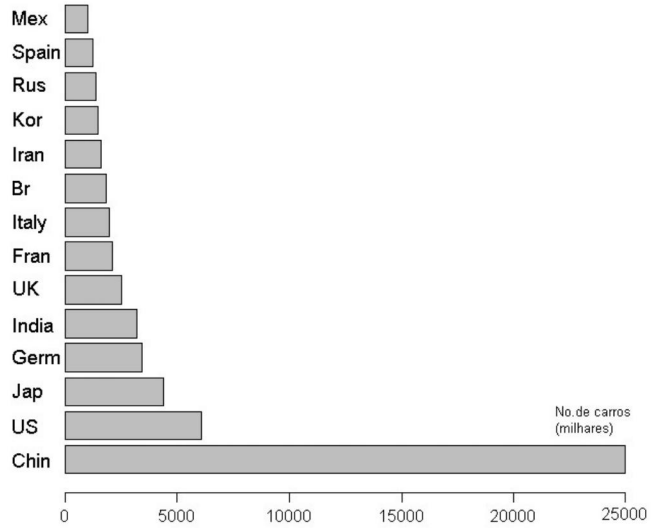


Figura 6. Número de carros novos registrados em cada país, em 2016
(dados: *The Economist, 2019*)

Outra variação é o *gráfico de barras compostas*. Neste gráfico, cada barra é subdividida de forma a mostrar as proporções que compõem o total. O gráfico da Fig. **7**, por exemplo, mostra as proporções de mortes no trânsito para cada categoria de vítimas, em 1996 e 2011, deixando evidente o crescimento do número de mortes de motociclistas. Note que este gráfico não compara os *totais* de morte nestes dois anos, mas sim as *proporções*.

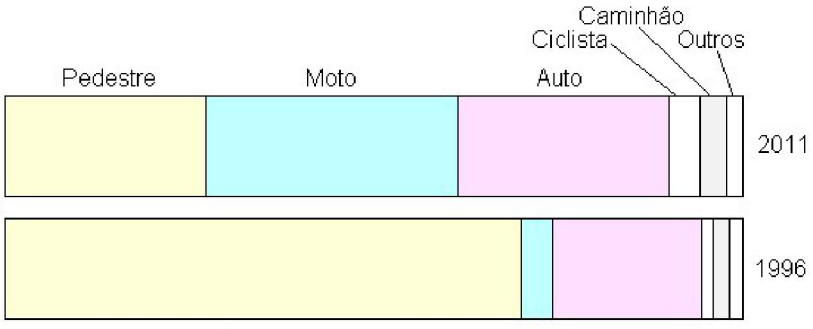


Figura 7. Mortes no trânsito, no Brasil, por categoria (dados: Waiselfisz, 2013)

A Fig. **8** compara os *totais* de capacidade instalada de geração elétrica no Brasil, e mostra como estes totais são repartidos entre as diversas formas de usinas geradoras.

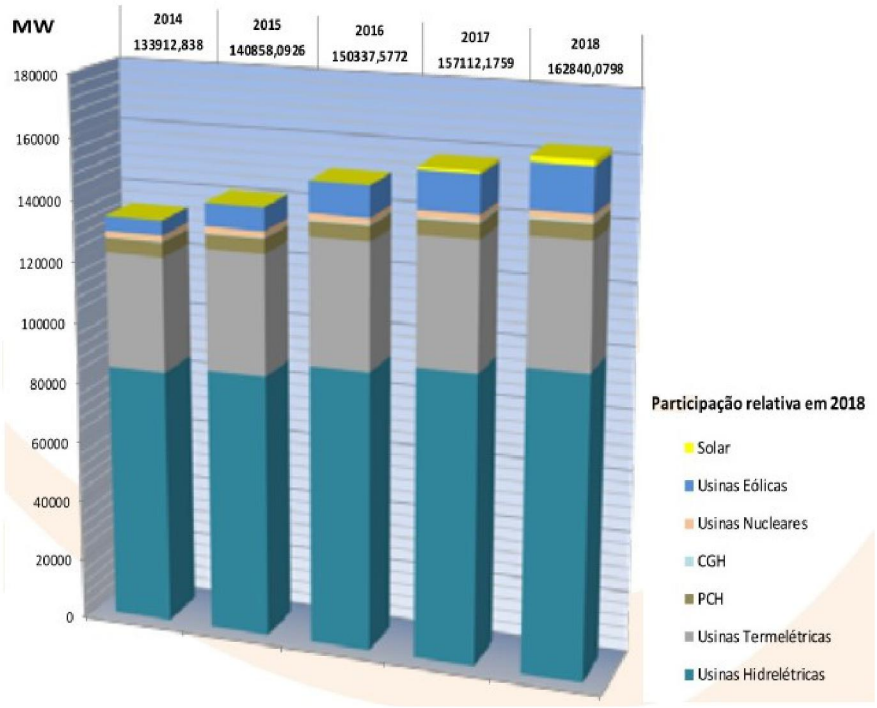


Fig. 8. Capacidade instalada de geração elétrica no Brasil (em MW)
(fonte: Empresa de Pesquisa Energética, 2019)

O gráfico da Fig. **9** mostra o número de prêmios Nobel recebidos por cada país até 2019. Neste gráfico as colunas são ordenadas por ordem alfabética do valor da variável no eixo horizontal (isto é, pelo nome do país), e não pela altura das colunas, o que é o mais usual. Com esta ordenação, fica mais fácil para os leitores encontrar o retângulo correspondente a cada país. Por outro lado, fica mais difícil comparar o número de prêmios de cada país: é bem evidente que os EUA estão em primeiro lugar, mas para descobrir quem está em segundo, terceiro, quarto, etc., os leitores terão que comparar visualmente os vários retângulos. Num gráfico destes, o objetivo principal é comparar a situação de cada país com a dos outros, e fazer algum tipo de classificação; a forma que ele foi arranjado, contudo, dificulta esta comparação, ao invés de facilitá-la.

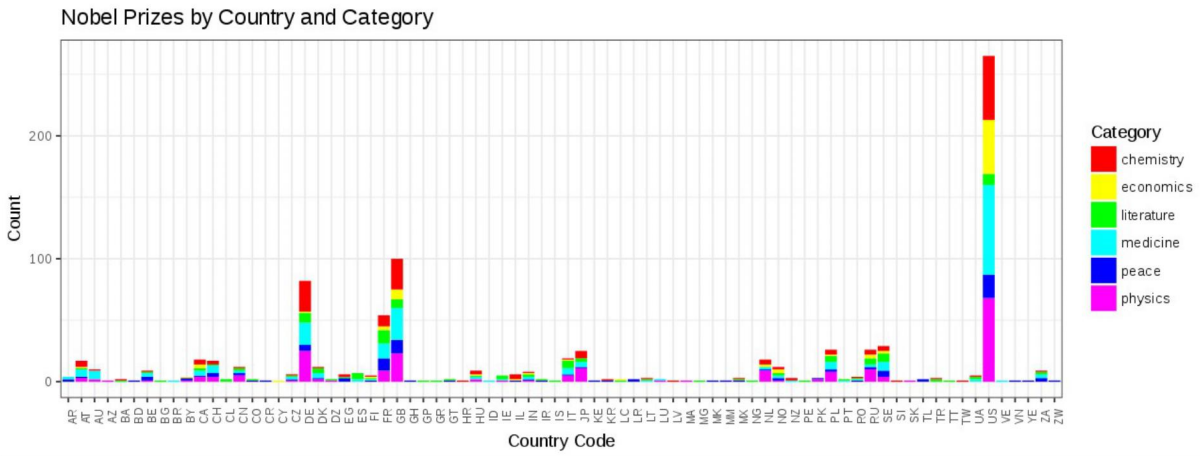
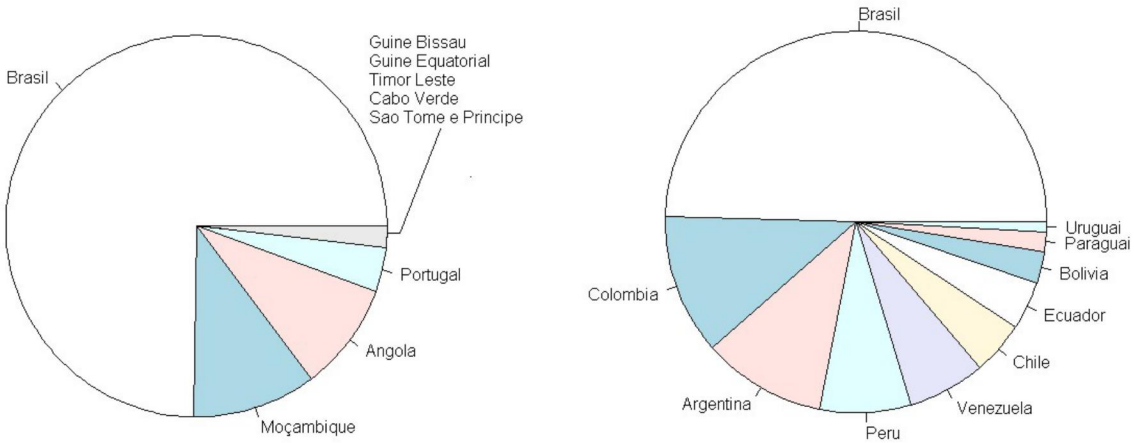


Figura 9. Número de Prêmios Nobel recebidos por cada país, até 2019 (fonte: *Wikipedia*)

2.1.6.2. Gráfico de setores ou de pizza (*piechart*)

O *gráfico de setores* tem a mesma função do *gráfico de barras compostas* visto na seção anterior - serve também para mostrar as proporções em que está dividido um total, mas usando setores de círculo ao invés de retângulos. É um dos gráficos mais encontrados na mídia, e pode ser feito através de qualquer programa de planilha eletrônica, como o *OpenOffice Calc* ou o *Microsoft Excel*.

Como exemplo, o gráfico na Fig. **10A** mostra com a população dos países pertencentes à *Comunidade dos Países de Língua Portuguesa* (países que têm o português como uma de suas línguas oficiais); o da Fig. **10B** compara as populações dos 10 maiores países da América do Sul (foram excluídos aqueles com menos de um milhão de habitantes).



A. População dos países da Comunidade dos Países de Língua Portuguesa

B. População dos maiores países da América do Sul, 2019

Figura 10. Exemplos de gráficos de setores

O gráfico de setores, no entanto, não é a melhor opção quando queremos comparar as proporções de dois totais; é muito difícil para uma pessoa comparar ângulos visualmente. Como exemplo, o gráfico na Fig. **11** mostra as proporções de mortes no trânsito brasileiro, classificadas pela categoria de vítima; os dados são os mesmos usados no gráfico de barras compostas visto antes na Fig. **7**. A comparação entre estas duas figuras mostra que o gráfico de barras compostas, apesar de menos popular, permite uma comparação mais fácil entre as proporções de 1996 e 2011 do que os gráficos de setores.

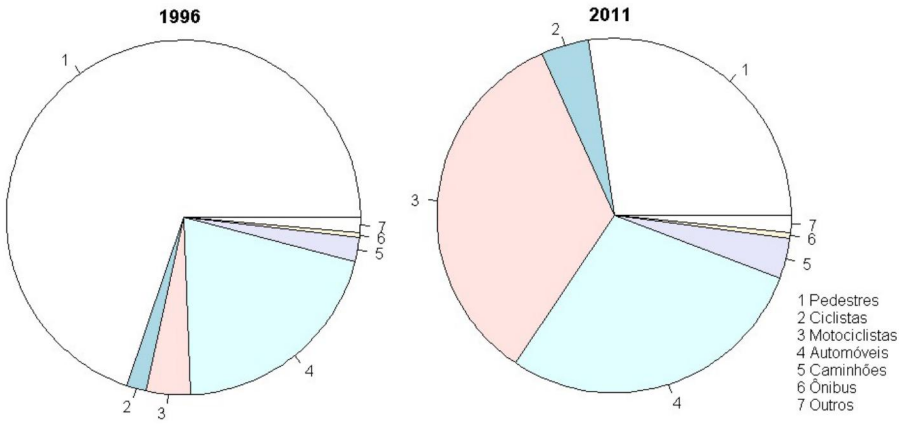
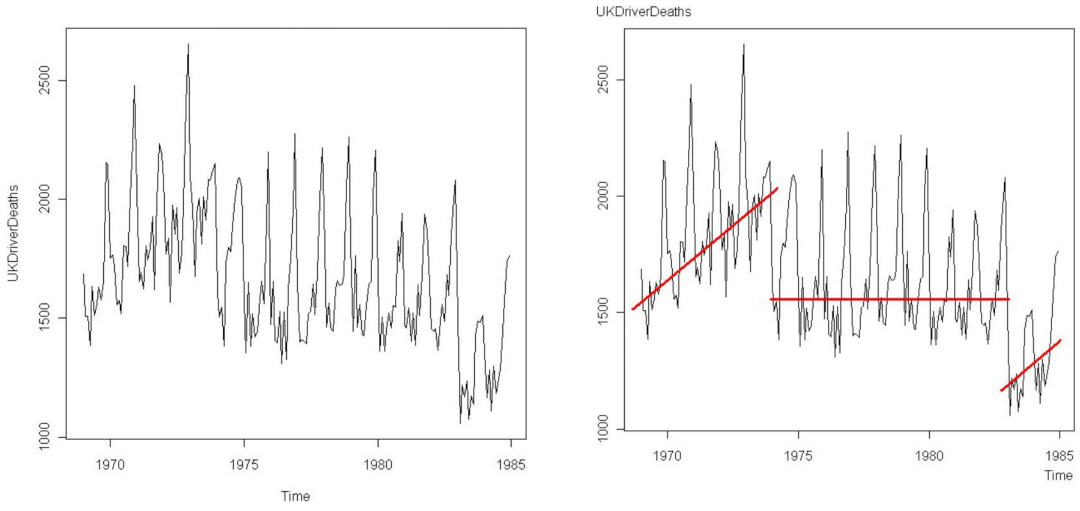


Figura 11. Mortes no trânsito, no Brasil – gráfico de setores

2.1.6.3. Gráfico de linha (*lineplot*)

O *gráfico de* linha é usado para representar variáveis que evoluem ao longo do tempo, formando as chamadas *séries temporais*. O estudo destas séries é uma sub-área da Estatística, normalmente não abordada nos cursos básicos (veja no site a seção *Séries Temporais*). Pode haver dois objetivos para este estudo: a *análise* da série, ou sua *previsão*.

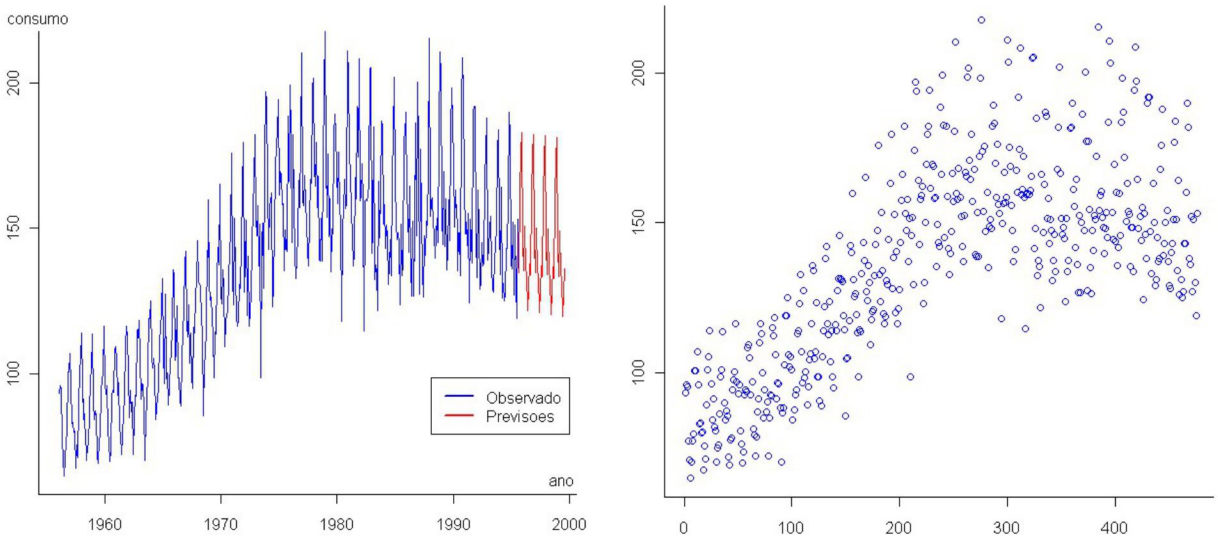
A análise procura identificar padrões no comportamento da variável, como sua *tendência* (se ela tende a subir ou a descer) e sua *sazonalidade* (variações que acontecem regularmente a cada ano, devidas ao efeito das estações). Além disso, busca os pontos onde estes padrões são quebrados, e tenta justificar estas quebras. Um exemplo desta análise, realizada de forma qualitativa (sem usar números), pode ser feito sobre o gráfico da Fig. **12A**, que mostra o número de mortos e feridos em acidentes de trânsito no Reino Unido, num período de 16 anos. Fig. **12B** mostra a variação do nível médio da série (estimado visualmente, de forma aproximada). É possível notar que havia inicialmente uma tendência de aumento linear do nível, além de um padrão que se repetia a cada ano: o número aumenta muito no início e no final de cada ano, em comparação com o observado no meio do ano (isto ocorre porque estes meses correspondem ao inverno, e o gelo nas estradas torna o tráfico mais perigoso). Houve dois momentos, porém, em que estes padrões se alteraram bruscamente. Em 1974, a tendência de crescimento desapareceu, e o nível caiu de forma repentina; a causa foi a primeira crise dos combustíveis, ocorrida quando os árabes elevaram subitamente o preço do barril, e todos os países foram obrigados a reduzir o consumo. Em 1983, ocorreu outra queda repentina no nível; desta vez, a causa foi a introdução da lei obrigando o uso de cintos de segurança nos carros, o que diminui o número de vítimas.



(A) (B)

Figura 12. Número de motoristas mortos e feridos gravemente, Reino Unido, 1969-1984.

Além da análise do *passado* de uma série temporal, podemos estar interessados em fazer previsões de valores *futuros* da série. Estas previsões são feitas por meio do ajuste de modelos probabilísticos aos dados de uma amostra da série, e da extrapolação deste modelo. A Fig. **13A** mostra a série de consumo mensal de cerveja na Austrália, entre 1956-1995 (em azul) e as previsões para quatro anos seguintes (em vermelho), feitas usando o método de *amortecimento exponencial de Holt-Winters*.



(A)

Figura. 13. Consumo mensal de cerveja na Austrália

Há duas observações importantes que devem ser feitas sobre o gráfico de linha. Primeiro, este gráfico deveria ser usado, em princípio, para representar variáveis contínuas. Freqüentemente, porém, é usado também para representar variáveis discretas, como nos exemplos acima − tanto as mortes no trânsito quanto o consumo de cerveja são totais anuais. Nos gráficos, as séries deveriam ser representadas por pontos separados, já que temos os totais para 1980 e para 1981 (por exemplo), mas não os valores intermediários; isto resultaria porém num gráfico muito confuso, como o da Fig. **13B**. Na prática, os valores de cada ano são unidos por linhas contínuas, para deixar mais evidente o percurso feito pela série.

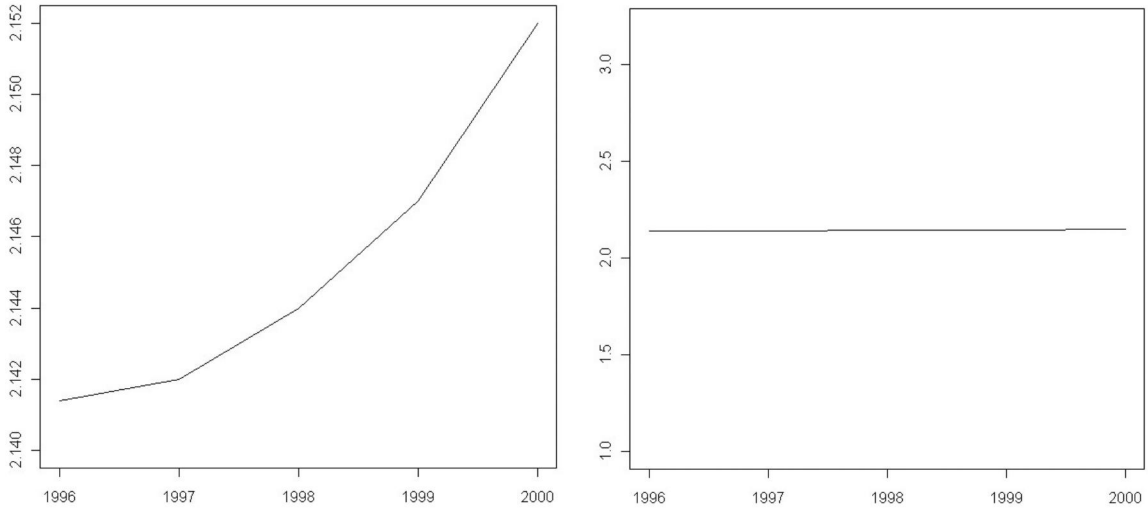


Figura 14. Taxa de juros em cinco anos (dados fictícios)

Segundo, a escala vertical do gráfico deve ser escolhida com cuidado. Assim como acontece nos gráficos de barra (seção **2.1.6.1**), a escala pode afetar a interpretação que os leitores fazem dos dados. A Fig. **14** mostra o aumento das taxas de juros num período de

cinco anos (dados fictícios), usando duas escalas diferentes, e os mesmos dados. Qual destes dois gráficos é o "correto"? Isto depende da impressão que o desenhista quer passar aos leitores: o gráfico A dá a impressão de que a taxa de juros está "explodindo"; o gráfico B, de que a taxa está praticamente constante. Provavelmente o melhor seria o gráfico B, pois o aumento na taxa está na casa do centésimos de 1% (0,01%), o que é insignificante para qualquer aplicação prática em economia ou finanças.

Um variante do gráfico de linha pode ser usada para representar as proporções em que um total se divide em instantes consecutivos do tempo; esta variante é chamada por alguns autores de *gráfico de faixas*. O gráfico da Fig. **15**, por exemplo, mostra os percentuais de mortes no trânsito, por categoria de meio de transporte, a cada ano entre 1996 e 2011, e destaca o grande aumento que houve na proporção de motociclistas entre estas mortes. Poderíamos representar esta mesma informação por meio de um conjunto de 16 diagramas de barras compostas em paralelo (um para cada ano), mas o desenho ficaria muito sobrecarregado, com um excesso de linhas desnecessárias; o gráfico de faixas é uma solução mais simples e mais elegantes.

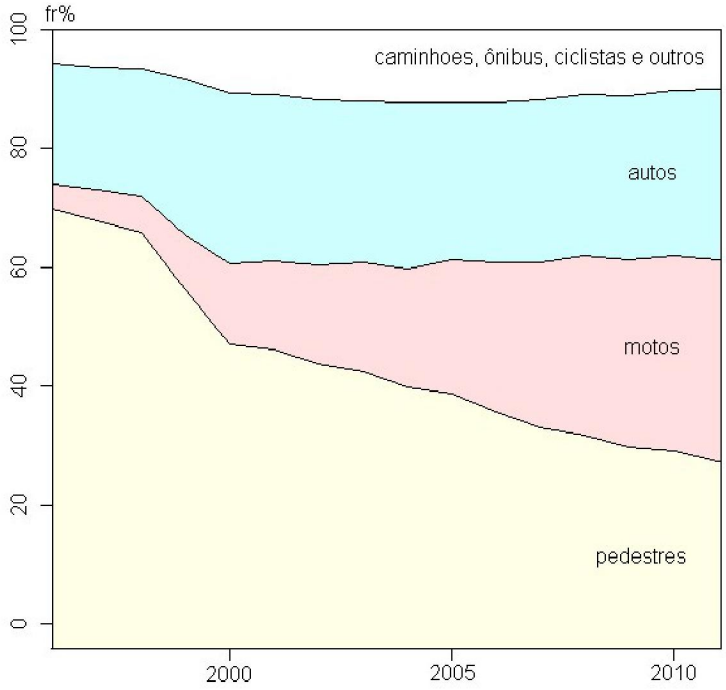


Figura 15. Mortes no trânsito, por categoria
Brasil, 1996-2011 (dados: Waiselfisz, 2013)

Referências

Waiselfisz, Júlio J. (2013). *Mapa da Violência 2013 - Acidentes de Trânsito e Motocicletas.* Rio de Janeiro: CEBELA / FLACSO.

Huff, Darrell (1982). *How to lie with statistics*. W.W.Norton, NY. 1982

Tufte, E. R. (1998). *The Visual Display of Quantitative Information.* Cheshire, Connecticut: Graphics Press.

Empresa de Pesquisa Energética (2019). *Anuário Estatístico de Energia Elétrica.* Ministério de Minas e Energia.

The Economist (2019). *Pocket World in Figures.*