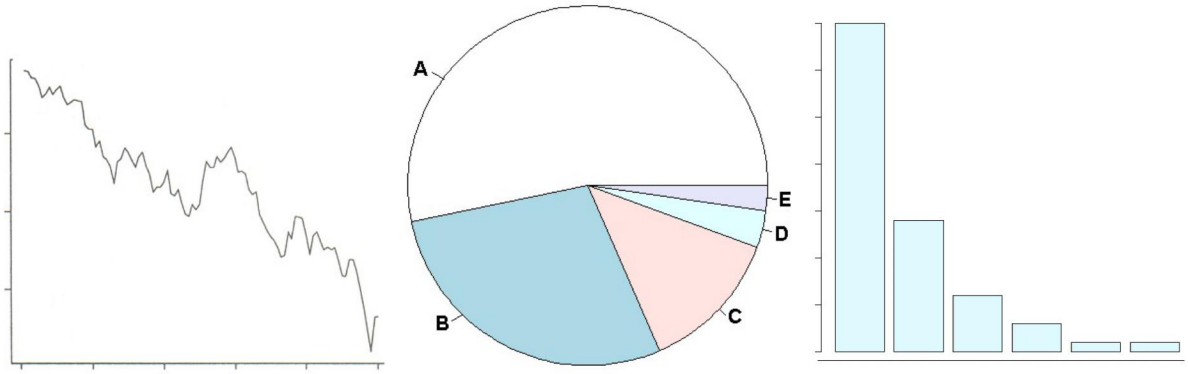
2.1.0. Gráficos

2.1.0.1. Introdução

Gráficos atualmente estão em toda parte. Em qualquer noticiário de TV, ou jornal da internet, encontramos *gráficos de linha* (Fig. **1A**) mostrando, por exemplo, o aumento do preço da gasolina, ou a queda das cotações na Bolsa de Valores no último ano; *gráficos de setores* (Fig. **1B**), mostrando a fração do mercado que é controlada por cada companhia; ou *gráficos de barras* (Fig. **1C**) comparando a economia de alguns países (veremos detalhes destes três gráficos na seção **2.1.6**).



(A) Gráfico de linha (B) Gráfico de setores (C) Gráfico de barras

Figura 1. Tipos de gráficos mais comuns

Até um passado relativamente recente, porém gráficos não eram muito comuns na mídia não-especializada, e nem mesmo nos livros de Estatística. Num livro-texto publicado em 1969, por exemplo [1], nos três primeiros capítulos há apenas 12 figuras com gráficos, das quais oito eram *histogramas*, e quatro *ogivas* de Galton (seção **2.1.5**).

Por que isto mudou? Há provavelmente duas razões. A primeira é que aumentou enormemente a quantidade de informação numérica publicada; somos hoje em dia constantemente bombardeados com "estatísticas" sobre todas as áreas – desde o desempenho da economia do país, até o número de chutes ao gol que um time fez durante um jogo de futebol –, e gráficos são a forma mais eficiente de transmitir este tipo de informação.

A segunda razão é que fazer gráficos se tornou muito fácil, depois da introdução dos computadores pessoais nos anos 1980s. Antes, tudo tinha que ser feito à mão, e qualquer gráfico poderia exigir horas de trabalho; hoje, eles podem ser feitos em segundos, com um clique do *mouse*. (Esta facilidade, é claro, também tem seu lado ruim: aumentou muito o número de gráficos publicados que são mal feitos, enganosos, ou simplesmente desnecessários; este problema é discutido em [2] e [3]).

Atualmente, os gráficos são uma parte essencial da Estatística. Existe uma expressão que diz que "uma imagem vale mais do que mil palavras", criada aparentemente por publicitários americanos no início do século XX [4]. Se isto é ou não verdade na publicidade, pode ser discutível; na Estatística, porém, é certo que um gráfico consegue passar informação de modo mais eficiente do que 1000 números. Gráficos permitem a comparação visual das magnitudes dos números que representam diferentes amostras ou populações, e mostram relações e proporções com mais impacto do que os números originais.

Suponha que, numa notícia sobre a queda da Bolsa de Valores no último vez, um jornal publique em vez do gráfico uma tabela com as cotações a cada dia do mês passado. A informação está toda nesta tabela; contudo, a maioria dos leitores não teria paciência para examinar longas séries de números, e mesmo aqueles que tivessem dificilmente conseguiriam ter uma visão global do que está acontecendo sem olhar para um gráfico.

Outro exemplo: todos sabemos que a China e a Índia são países enormes, em termos de população; contudo, a maioria das pessoas não têm uma noção muito clara da proporções entre números – especialmente entre números muito grandes, como os das populações destes dois países, e do resto do mundo. O gráfico da Fig. **2** representa as proporções das populações de cada país (classificados de acordo com os continentes em que o planeta é convencionalmente dividido), e mostra que aos países da Ásia têm mais de metade da população mundial (na verdade, 57%); além disso, mostra que a China e a Índia, juntas, têm mais de metade da população asiática.

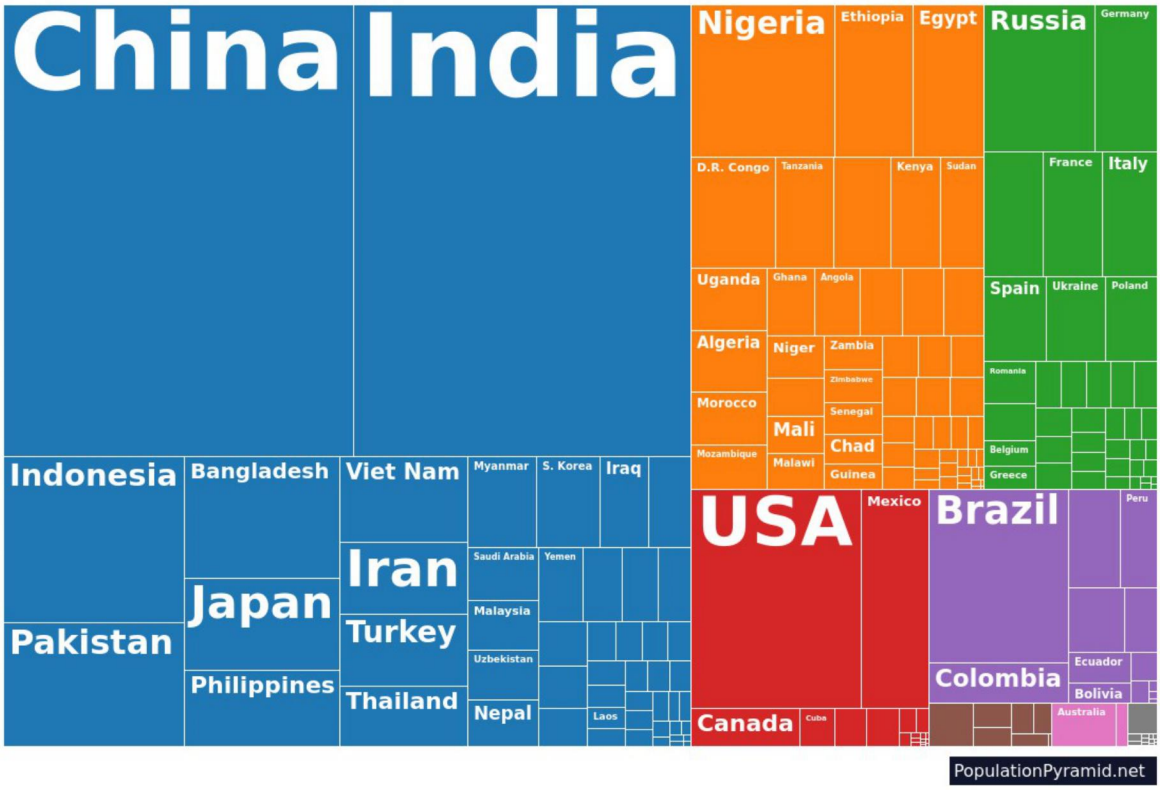
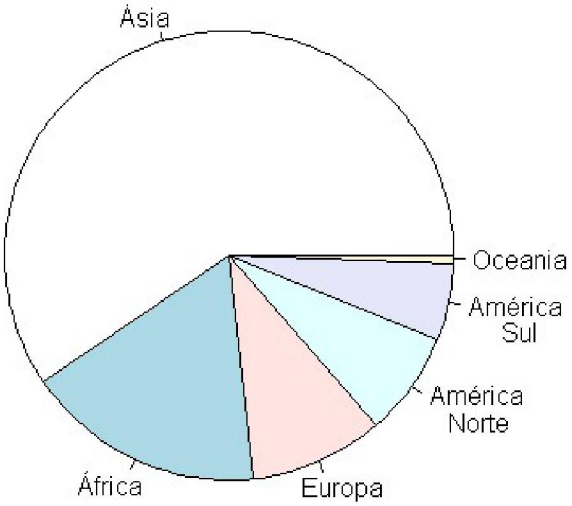


Figura 2. População dos países do mundo (fonte: *PopulationPyramid.net* [*])

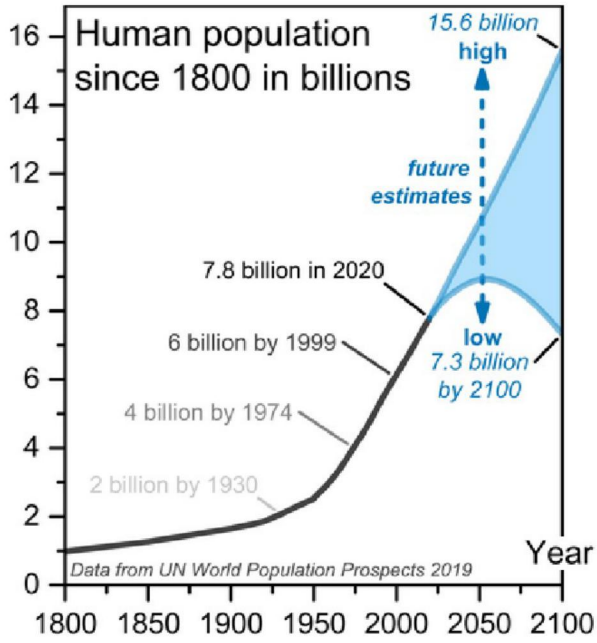
Mais de três quartos da população mundial vive na Ásia ou na África; para muita gente, provavelmente será uma surpresa ver que entre as maiores populações do mundo estão países que raramente são mencionados nas notícias, como a Indonésia, o Paquistão e a Nigéria (todos com mais de 200 milhões de habitantes). Os países normalmente considerados "desenvolvidos" estão quase todos na Europa e na América do Norte (a principal exceção é o Japão), mas estes continentes representam menos de 17% da população mundial. O gráfico mostra de forma bem clara quão pequenas são, relativamente, as populações dos principais países europeus, como Alemanha, França e a Itália.

Um outro tipo de gráfico sobre o mesmo tema é o da Fig. **3A**, que compara as populações dos continentes. Fica bem evidente a predominância das populações da Ásia e da África, e também o fato de que a população da África é maior do que as das duas Amé-

ricas juntas. Este tipo de gráfico, chamado de *gráfico de setores* ou *gráfico de pizza* (seção **2.1.6.2**), é bem conhecido, e encontrado com freqüência na mídia não-especializada (revistas, *internet*, etc.); contudo, tem menos informação do que o da Fig. **2**, pois representa apenas os continentes, não os países individualmente.



(A) Populações dos seis continentes, 2020.
(dados: www.worldmeter.info)



(B) Crescimento da população mundial, 1800-2020
(fonte: *Wikipedia*)

Figura 3. Gráfico de setores e gráfico de linha

Outro tipo de gráfico muito encontrado na mídia é o *gráfico de linha*, como o da Fig. **3B**, que mostra estimativas da população do mundo, desde 1800. É fácil ver que a população tem crescido de maneira quase linear (seguindo uma linha reta), e não parece haver ainda sinal de que esta tendência ao crescimento esteja sendo freada.

Darrell Huff, um estatístico americano que fez muito para chamar atenção do público sobre os gráficos, disse: "Quando números na forma de tabela são tabu, e as palavras não conseguem fazer bem o trabalho, como freqüentemente acontece, sobra uma resposta: faça um desenho" [2].

2.1.0.2. O conceito de "distribuição"

Gráficos como os das três figuras anteriores são muito úteis, e são encontrados com freqüência nas publicações. Muitas vezes, porém, comparar apenas totais ou médias não é suficiente. Por exemplo: ouvimos freqüentemente alguém dizer que o Brasil já não é mais um país jovem, e que sua população está envelhecendo, se comparada com a de países africanos como Moçambique e Angola; mas que ainda é mais jovem do que a da maioria dos países europeus. Como verificar se isto é verdade? Examinar as idades de todos os habitantes de cada país, um a um, não é possível. Primeiro, porque não existem listas publicadas com estas idades. Segundo, porque não conseguiríamos concluir nada depois de examinarmos estas listas, se elas existissem, já que conteriam apenas os dados "brutos" (isto é, os dados que ainda não foram depurados e organizados de alguma forma). O cérebro humano

não consegue concluir nada simplesmente examinando listas com milhões de números. Poderíamos resumir estes dados, calculando as médias das idades de cada país; isto ajuda a comparação, mas tem o defeito de perder muita informação; calculando a média do Brasil, por exemplo, reduzimos todas as idades dos 212 milhões de habitantes a um único número. (Seria melhor, aliás, usar a *mediana*, ao invés da média; seção **2.2.1**).

O que queremos na verdade é comparar as *distribuições* da variável "idade" nas populações dos diferentes países; isto é, como os valores da variável se distribuem na população. Uma variável pode assumir diversos valores diferentes; se fizermos um gráfico, marcando estes valores em um eixo, veremos que eles se distribuem ao longo do domínio da variável de acordo com um padrão, com maior freqüência em alguns lugares (alguns intervalos) do que em outros. O conceito de "distribuição" é fundamental na Estatística, e grande parte do trabalho consiste em estudar e comparar distribuições de variáveis.

Usando dados dos recenseamentos, poderíamos representar a distribuição das idades por meio de *tabelas de distribuição de frequências* (seção **2.1.4**); no entanto, um gráfico pode transmitir um resumo da informação de forma muito mais imediata do que tabelas ou médias. Os gráficos da Fig. **4** mostram a distribuição das idades (*distribuições etárias*) de três países; fica bem evidente que o Brasil está numa situação intermediária entre os países de população muito jovem (no caso, Moçambique), e o de população mais idosa do mundo, o Japão. Este tipo de gráfico é um caso particular do histograma, chamado de "pirâmide etária" (seção **2.1.5.7**). (Nota: os gráficos das Figs. **2** e **4** foram extraídos do site *PopulationPyramid.net*, que mostra as pirâmides etárias de todos os países do mundo, além de muitos outros gráficos interessantes sobre populações.)

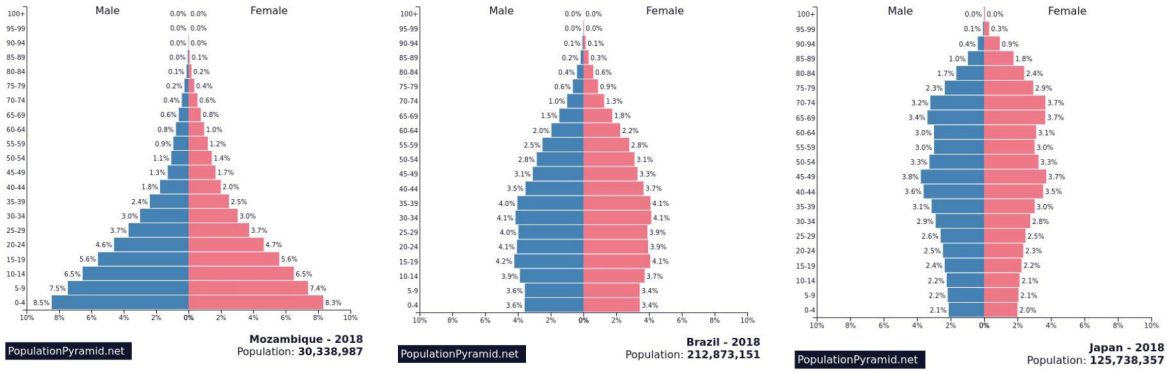


Figura 4. Pirâmides etárias de Moçambique, Brasil e Japão (fonte: *PopulationPyramid.net* [*])

Neste exemplo, temos dados sobre a população toda, obtidos por meio de recenseamentos. Em grande parte da pesquisa científica, porém, temos que tentar descobrir a forma da distribuição na população através do que observamos em amostras. Suponha por exemplo que estejamos pesquisando sobre uma variável biológica qualquer, como o "nível de bilirrubina no sangue" (a bilirrubina é uma substância encontrada no plasma sanguíneo). A maioria das pessoas (a não ser quem trabalha na área de Saúde) não tem nenhuma idéia *a priori* de como será a distribuição desta variável na população. Se um paciente faz um exame de sangue e descobre que seu nível de bilirrubina é igual a 1,5 mg/dL, que podemos concluir - este valor é alto ou baixo? Para responder a esta pergunta, precisamos conhecer algo sobre a *distribuição* do nível na população, quais faixas de valores ocorrem com mais freqüência, quais faixas com menos freqüência. Conhecer a média não é suficiente; suponha que descobrimos que o nível médio na população é 0,6 mg/dL. O valor 1,5 mg/dL está portanto acima da média. Mas está *muito* acima da média? Este é um valor preocupante,

que deve levar o médico a fazer algo, ou não tem importância? Para decidir, precisamos de saber como a variável se *distribui* em torno da média, ou pelo menos quão longe da média ela pode ir, para cima ou para baixo, num paciente saudável.

Como descobrir isto? Teremos que fazer pesquisas com amostras (não podemos examinar o sangue de todas as pessoas do mundo!). Depois, a primeira tarefa dos analistas será sempre a de fazer gráficos das variáveis que interessem. A partir da *distribuição empírica* encontrada na amostra, podemos ter uma idéia de como é a distribuição da variável na população, e de quais intervalos contém os valores mais prováveis do nível, nos pacientes saudáveis. (É importante porém lembrar que os gráficos obtidos em uma amostra não *provam* nada, apenas sugerem hipóteses, que devem ser depois confirmadas por meio de técnicas de *Inferência Estatística*.)

2.1.0.3. Papel dos gráficos na Análise Exploratória dos Dados

Dissemos acima que duas das causas do aumento da importância dos gráficos foram os crescimentos da quantidade de informação disponível e da facilidade de fazer gráficos, conseqüências da disseminação dos computares pessoais. Uma terceira razão, porém, mais técnica, foi o surgimento das idéias sobre *Análise Exploratória de Dados* (mencionadas na seção **2.0**). As ferramentas usadas nesta análise se baseiam em gráficos e ferramentas numéricas simples, e evitam em princípio o uso de modelos matemáticos complicados.

A análise por meio de gráficos se tornou por isto uma parte importante do trabalho estatístico. Primeiro, porque gráficos mostram, de forma rápida, a informação principal contida num conjunto de dados (por exemplo, um gráfico de linhas mostra imediatamente se as cotações na Bolsa de Valores no último ano subiram ou desceram; se tivéssemos apenas uma tabela com os dados originais, teríamos que gastar vários minutos comparando estes números, antes de tirar alguma conclusão). Segundo, porque gráficos nos ajudam a descobrir padrões e relações que não são evidentes nos dados originais, e esta descoberta pode gerar idéias para novos caminhos de pesquisa (por exemplo, se descobrimos que duas variáveis parecem estarem relacionadas, ou que há na amostra valores muito diferentes do que esperaríamos, etc.), Terceiro, porque gráficos podem ser usados para avaliar a qualidade dos dados. Sendo os seres humanos imperfeitos como são, praticamente todo conjunto de dados inclue erros, que têm que ser detectados. A existência de erros pode ser suspeitada a partir de indícios como valores estranhos e inesperados, a forma irregular da distribuição, ou medidas não coerentes entre si. Por fim, porque os gráficos nos ajudam a escolher quais técnicas devem ser usadas para a análise estatística dos dados, de acordo com as características da distribuição encontrada na amostra.

Estudaremos nas próximas seções os principais tipos de gráficos, e voltaremos depois a falar sobre seu uso na análise exploratória (seção **2.3**).

---

[*] Reproduzido com permissão do autor, Martin De Wulf.

referências

[1] Walker, H. M.; Lev, Joseph. *Elementary Statistical Methods*. Holt, Rinehart and Winston, 3rd ed, 1969.
[2] Huff, Darrell. *How to lie with statistics*. New York : W.W.Norton. 1982.
[3] Tufte, Edward R. *The Visual Display of Quantitative Information.* Cheshire, Conn.: Graphics Press, 1998
[4] *https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words*.