2.1.1. Gráfico de pontos (*dotplot*)

2.1.1.1. Introdução

O *gráfico de pontos* é uma maneira simples e rápida de organizar pequenas quantidades de dados. Por meio dele podemos ver a forma geral da distribuição, identificar os padrões e regularidades existentes, localizar os aglomerados (*clusters*), os pontos discrepantes (*outliers*), se existirem, e os valores máximos e mínimos. Este gráfico é chamado em inglês de *dotplot* ; em português, não existe uma tradução uniformemente aceita, e o usaremos a expressão *gráfico de pontos*.

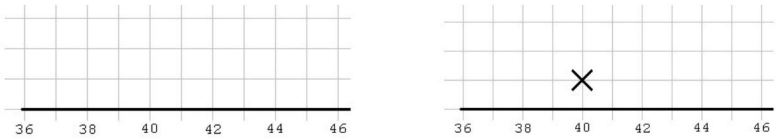
Este gráfico pode ser feito rapidamente, à mão, sobre um papel milimetrado ou quadriculado. Poucos programas de computador, porém, têm rotinas para fazê-lo; em geral, os programas preferem o gráfico de *ramo-e-folhas*, que será visto depois.

2.1.1.2. Como fazer o gráfico

Os dados que usaremos nesta seção vêm de uma amostra de estudantes de Estatística do curso de Medicina. Suponha que estejamos interessados na variável "número dos sapatos" destes estudantes, e desejamos representá-la por um gráfico de pontos. Para os 43 estudantes do sexo masculino, os valores são:

```
40 41 43 40 40 41 42 38 39 40 43 40 41 42 41 40 40 40 42 39
41 40 39 41 39 30 40 40 41 42 38 43 43 43 46 40 42 41 40 40
40 40 42
```

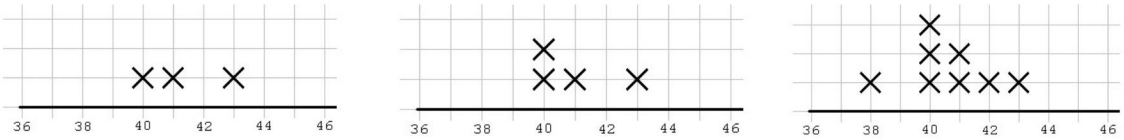
Para começar, fazemos um eixo horizontal com uma escala apropriada e uma grade (Fig. **1A**). O primeiro valor, 40, é marcado com um X ou outro símbolo qualquer (Fig. **1B**).



A B

Figura 1

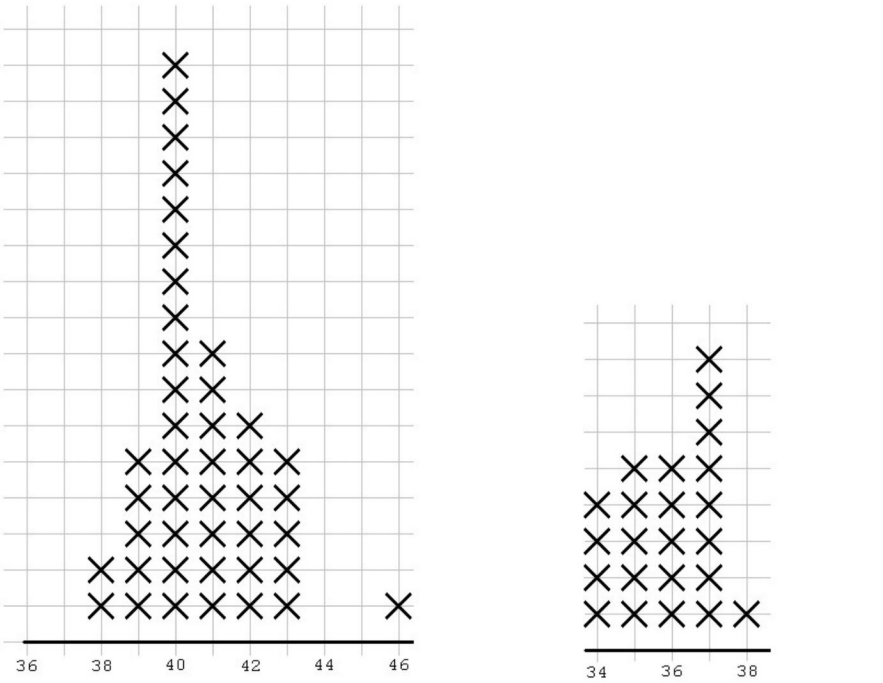
Os valores seguintes são 41 e 43, também marcados com X (Fig. **2A**). O próximo valor é novamente 40; para marcá-lo, fazemos um X acima do 40 anterior (Fig. **2B**). Os valores repetidos que forem surgindo deverão ser marcados acima dos já existentes, de forma que tenhamos uma pilha de Xs. Marcando os valores seguintes, 40, 41, 42 e 38, obtemos o gráfico da Fig. **2C**.



A B C

Figura 2

O resultado final, depois de marcados todos os pontos, será parecido com o gráfico da Fig. **3A**.



A. Números dos sapatos dos homens B. Números dos sapatos das mulheres

Figura 3

Para as 23 estudantes do sexo feminino, os valores são:

```
37 35 37 36 36 34 37 37 37 36 34 37 35 38 36 35 37 36 35 35 34 37 34
```

Marcando estes valores num outro eixo, obtemos o gráfico da Fig. **3B**.

2.1.1.3. Alguns conceitos importantes: posição, moda, assimetria, valor discrepante.

Usaremos os gráficos das Fig. **3** para ilustrar alguns conceitos importantes na análise de dados. Primeiro, estes gráficos representam a *distribuição* da variável "número do sapato" nas duas amostras; isto é, mostram como os valores se distribuem ao longo da escala, com maior frequência em alguns lugares (intervalos) do que em outros. Este conceito é fundamental em Estatística Descritiva, pois uma boa parte do trabalho consiste em analisar a distribuição de uma variável, ou em comparar as distribuições de duas ou mais variáveis. Em ambas as figuras, as distribuições seguem o padrão encontrado na maioria das variáveis naturais: os pontos formam um *aglomerado* (*cluster*) que tem forma similar

ao perfil de uma montanha. As *posições* destas distribuições são porém diferentes: a da Fig. **3A** tem um pico no valor 40, que foi o tamanho de sapato mais comum entre os homens; a da Fig. **3B** tem um pico no valor 36. O pico do gráfico, que indica o valor que foi encontrado com maior freqüência, é chamado de *moda* (*mode*); distribuições que têm apenas uma moda são chamadas de *unimodais*. A moda é uma das *medidas de posição*; veremos outras na seção **2.2.1**. Em ambas as figuras, as distribuições são levemente assimétricas; a maior parte das variáveis naturais têm distribuições *simétricas*, ou levemente *assimétricas*. A distribuição das Figs. **6** e **9**, por outro lado, são fortemente assimétricas; este tipo de assimetria, no qual a maior parte dos valores estão aglomerados no lado esquerdo do gráfico, e há alguns valores isolados formando um longa cauda à direita, é chamado de *assimetria positiva*. O gráfico da Fig. **3A** tem um ponto isolado no valor 46, distante do aglomerado; este valor é chamado de *ponto discrepante* (*outlier*). Localizar e tentar justificar a existência destes valores é uma tarefa importante da análise de uma distribuição. Neste exemplo, provavelmente o valor 46 não tem nenhum significado especial; valores discrepantes, porém, às vezes indicam que houve erros de medição, de registro, ou algum outro problema que tem que ser investigado.

2.1.1.4. Algumas variações do gráfico de pontos

Há algumas variações possíveis deste gráfico. Uma delas é o gráfico *duplo*, que permite que duas distribuições sejam comparadas num mesmo eixo. Outras são os gráficos em que os dados, ao invés de serem marcados com X, são marcados com siglas, letras, ou qualquer outro sinal que permita que o leitor identifique cada dado.

(i). Gráficos duplos

Em geral, não conseguimos comparar duas distribuições simplesmente representando todos os pontos num mesmo eixo. Por exemplo, se representarmos os números dos sapatos de homens e de mulheres num mesmo eixo, obtemos o gráfico da Fig. **4**.

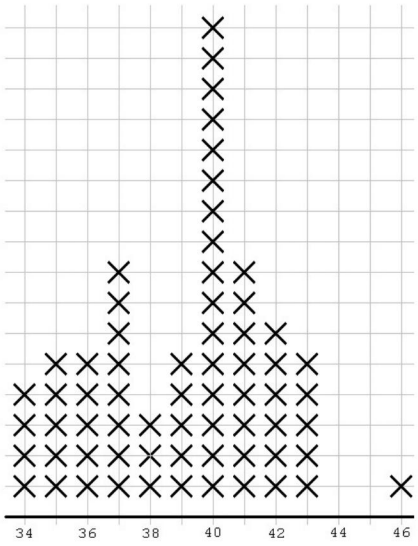


Figura 4

Este tipo de gráfico não tem muita utilidade; não podemos reconhecer quais dados se referem aos homens, e quais às mulheres, e portanto não podemos fazer nenhuma comparação útil. O fato de o gráfico ser *bimodal* (ter duas modas) já indica que algo está errado – variáveis raramente tem distribuições com esta forma. No caso, a existência de duas modas indica que estamos misturando duas distribuições diferentes, cada uma com sua própria moda. O melhor seria representar os dados em lados opostos do eixo horizontal, em cima e embaixo, como na Fig. **5**. Este gráfico mostra claramente, como esperávamos, que as distribuições são bastante distintas, e quase não se sobrepõem (há apenas um valor, 38, que ocorre nas duas distribuições).

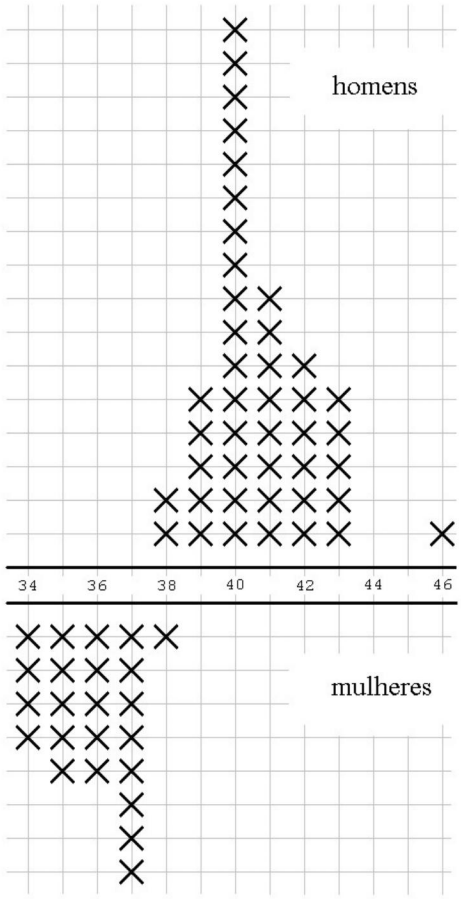


Figura 5. Números dos sapatos, homens e mulheres

(ii). Usando outros símbolos em lugar do X

Podemos fazer os gráficos usando outros símbolos, em lugar dos X; por exemplo, cruzes (+), círculos (o), asteriscos (*), ou letras, siglas, etc. Isto pode servir para facilitar a localização dos valores no gráfico.

Por exemplo, o gráfico da Fig. **6** mostra um gráfico de pontos duplo que compara a freqüência (em porcentagem) com que cada letra aparece em textos escritos, em português e em inglês. Este tipo de informação tem algumas aplicações práticas. No código Morse, por exemplo, usado para comunicação por telégrafo, todas as letras são codificadas por seqüências de pontos (·) e traços (−). As duas letras usadas com mais freqüência em inglês

receberam os símbolos mais simples: *E* (·) e *T* (−). As letras menos freqüentes receberam em geral símbolos mais complicados, como J (· − − −), Q (− − · −), X (− · · −) e Z (− − · ·).

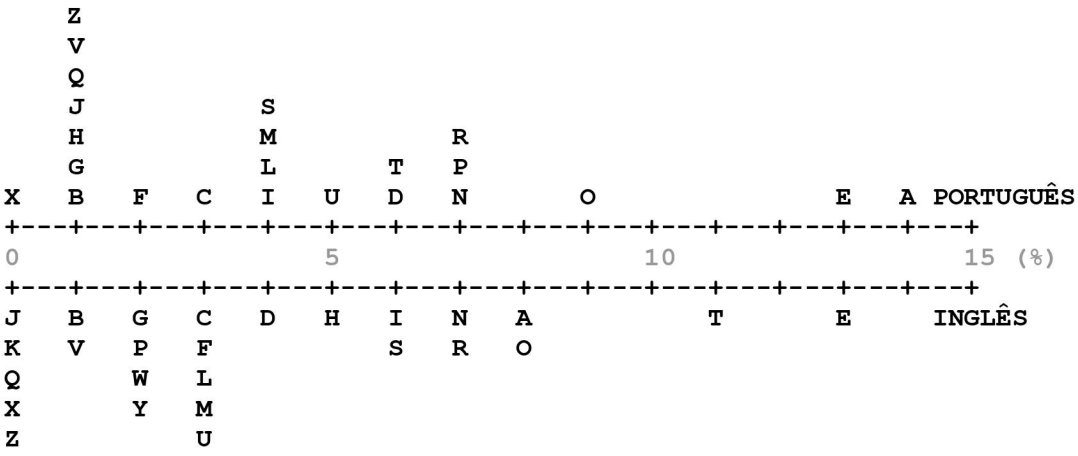


Fig. 6 - Freqüência das letras do alfabeto em textos escritos (em %)
dados: *National Council of Teachers of Mathematics* (inglês); o autor (português)

2.1.1.5. Quando usar e quando não usar o gráfico de pontos

O gráfico de pontos é mais útil quando não há muitos dados, e a variabilidade é pequena (as observações podem assumir poucos valores diferentes, de preferência inteiros); é o que acontece, por exemplo, com os *números dos sapatos*, nos exemplos acima. Se a variabilidade é grande, o gráfico pode não ser adequado. Por exemplo, suponha que os dados abaixo representem a distribuição de idades dos pacientes em uma amostra A:

Amostra A: 13 17 21 22 24 25 26 29 32 32 34 37 40 40 46 52 73

As idades variam muito, e quase nunca se repetem. Se tentarmos fazer um gráfico de pontos com estes dados, obtemos o que está na Fig. 7. Este gráfico não diz muita coisa; como os pontos estão excessivamente espalhados, não tem sentido falar em moda, simetria, etc. (Estes mesmos dados estão melhor representados no *gráfico de ramo-e-folhas*, Fig. 1 da seção 2.1.2.2).

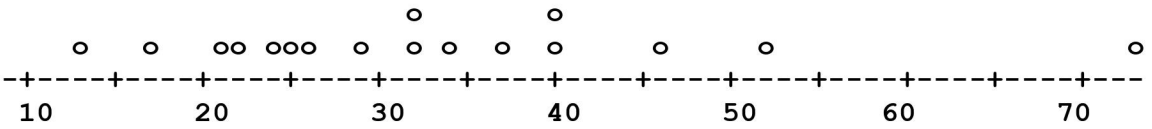


Figura 7. Idade de pacientes

Outro exemplo: obteremos um resultado ainda pior se tentarmos fazer o gráfico do nível de colesterol numa amostra de 100 homens, dos quais reproduzimos algumas linhas abaixo.

```
134   147   157   161   162   164   165   166   171   173
176   176   178   179   179   180   181   181   183   184 (...)
268   272   279   286   287   289   290   296   298   382
```

Como pode ser visto, o nível varia de 134 a 382, e os valores quase não se repetem. O gráfico mostrará apenas pontos espalhados ao longo do eixo, sem chegar a uma forma definida. Uma solução poderia ser a de arredondar os dados – por exemplo, desprezar as unidades, e marcar os X nas dezenas mais próximas (130, 140, ...); isto resulta no gráfico da Fig. **8**. Para estes dados, porém, também seja mais útil usar o *gráfico de ramo-e-folhas* (seção **2.1.2**), ou o *histograma* (seção **2.1.5**).

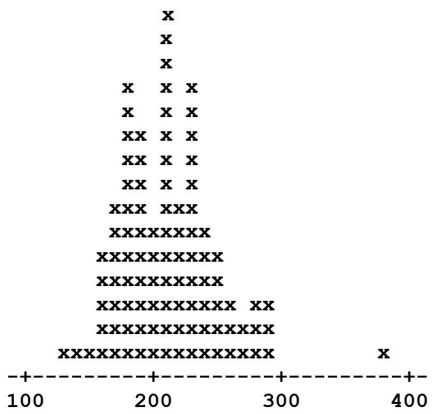


Figura 8 - Nível de colesterol sérico em miligramas percentuais, em 100 homens com idades entre 45 e 67 anos

Uma vantagem do gráfico de pontos é que ele é mais fácil de entender, para um leitor leigo, do que o de ramo-e-folhas. Além disso, vale a pena insistir no gráfico de pontos quando queremos representar os valores por siglas, letras, etc. Por exemplo, o gráfico da Fig. **9** mostra a densidade demográfica dos estados brasileiros (habitantes/km^2), usando a sigla dos estados para representar cada valor.

A quantidade de dados não é grande (são 27 estados), mas as densidades variam muito, de 2 hab/km^2 (AM e RR) até 480 hab/km^2 (DF). Arredondamos os valores para o múltiplo de 20 mais próximo (a escala progride de 20 em 20: 0, 20, 40, 60, etc.). A vantagem desta forma de gráfico (em relação ao gráfico de ramo-e-folhas, por exemplo), é que cada estado é representado por sua sigla, o que permite que o leitor entenda imediatamente o que está sendo mostrado, sem ter que consultar uma legenda. No gráfico da Fig. **6** também foi preciso fazer arredondamentos, mas ali o problema era mais simples: bastou aproximar o valor de cada freqüência até o inteiro mais próximo.

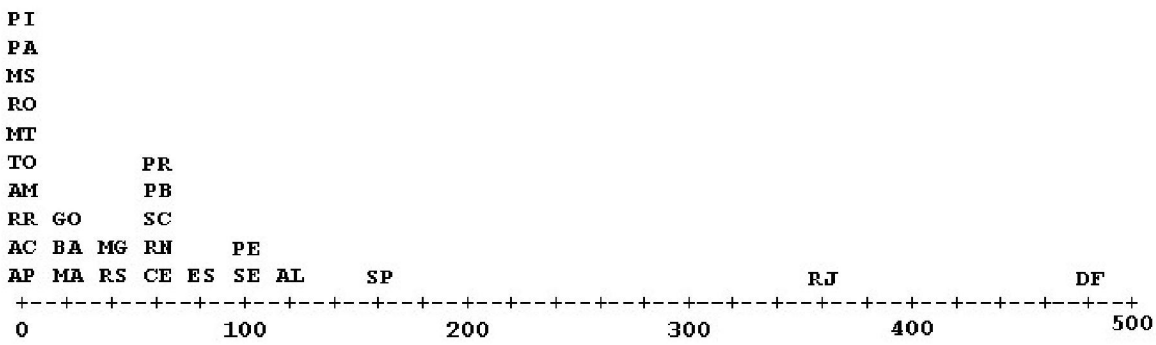


Figura 9 – Densidade demográfica dos estados brasileiros, em 2013 (hab / km2)

Note que a distribuição é extremamente assimétrica; a maioria dos estados têm densidades abaixo de 100 hab/km^2, mas há valores discrepantes com 360 (RJ) e 480 (DF). Esta assimetria é bastante comum em dados de populações, ou de variáveis econômicas. Para comparação: entre os países grandes, o de maior densidade é Bangladesh (1.271 hab/km^2). Na Europa, o de maior densidade é a Holanda (421); Portugal e França têm densidades em torno de 110-120. Países africanos têm geralmente baixas densidades: Moçambique e Angola, por exemplo, têm em torno de 25-30 hab/km^2.

Resumo

- O gráfico de pontos serve para representar a distribuição de uma amostra, se são poucos os dados e a variabilidade é pequena.
- Se a variabilidade é grande, e for necessário agrupar os dados ou arredondá-los, talvez seja melhor usar outro gráfico (como o de *ramo-e-folhas*, ou o *histograma*).
- Vale a pena usar o gráfico de pontos quando os dados podem ser representados por siglas, letras, etc., o que torna o gráfico mais atraente para os leitores, e mais fácil de entender.
- Conceitos importantes apresentados:
    - *distribuição* de uma variável
    - *posição* de uma distribuição; *moda*, distribuição *unimodal* / *bimodal*;
    - *simetria* / *assimetria* de uma distribuição
    - *pontos discrepantes*