2.1.3. Diagrama de Tukey

2.1.3.1. Introdução

Os gráficos de *pontos* e o de *ramo-e-folhas* são úteis para analisar as distribuições de uma ou duas amostras pequenas. Contudo, por eles incluirem *todos* os dados da distribuição, não servem quando as amostras são grandes, pois os gráficos ficam sobrecarregados. Também não servem quando desejamos comparar simultaneamente mais de duas distribuições.

Para representar uma amostra grande, o melhor pode ser usar um *histograma*; para comparar duas amostras grandes, podemos usar o *polígono de frequências* (seção **2.1.5**). Para representar e comparar diversas amostras grandes, contudo, o que é necessário é um gráfico simplificado que mostre apenas as características principais de cada distribuição, omitindo os detalhes. Uma possibilidade é usar o diagrama de Tukey (em inglês: *boxplot* ou *box-and-whiskers*), que tem a aparência mostrada na Figura 2. John Tukey (o mesmo estatístico que criou o gráfico de *ramo-e-folhas*) não inventou realmente este gráfico, mas aperfeiçou alguns de seus detalhes e contribuiu para sua divulgação.

2.1.3.2. Como fazer o diagrama

Estes diagramas são muito simplificados, pois cada um deles é desenhado a partir de apenas cinco valores de uma distribuição: os dois extremos (o valor máximo e o mínimo) e três valores de referência, a *mediana* e os *quartis*. Estes valores são *separatrizes*, medidas que separam a distribuição em partes que contém quantidades iguais de dados. A *mediana* divide a distribuição em duas metades: a primeira contém as observações com valores iguais ou superiores ao valor da mediana, a segunda contém os valores iguais ou inferiores. Por exemplo, suponhamos a distribuição de idades de pacientes na Amostra A (já usada na Fig. **1**, seção **2.1.2**), cujos valores estão listados abaixo em ordem crescente:

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

A mediana será o valor que está no centro desta sequência. Como são 17 valores, a mediana será o 9º valor, que equivale a 32 anos. Metade destes pacientes tem idades iguais ou superiores a 32 anos; a outra metade tem idades iguais ou inferiores. A mediana é uma das *medidas de posição*, pois indica onde se encontra o centro da distribuição (estas medidas serão vistas na seção **2.2.1**).

Se tomarmos a metade à esquerda desta sequência, incluindo a mediana (dados em negrito), e localizarmos seu valor central, encontraremos o *primeiro quartil*, igual a 24 :

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

Um quarto das observações têm valor igual ou inferior a este quartil. Por outro lado, se localizarmos o valor central da metade à direita da sequência (em negrito), encontraremos o *terceiro quartil*:

13 17 21 22 24 25 26 29 **32 32 34 37 40 40 46 52 73**

No caso, será o 14º valor, igual a 40 anos; três quartos dos pacientes têm idades iguais ou inferiores a 40 anos. Representaremos o primeiro quartil por Q_1, o terceiro por Q_3. A mediana, que representaremos por $\tilde{X}$, é idêntica ao segundo quartil Q_2. Para fazer o diagrama de Tukey, marcamos num eixo a localização dos dois valores extremos e a dos três quartis, usando traços verticais, como na Fig. **1**:

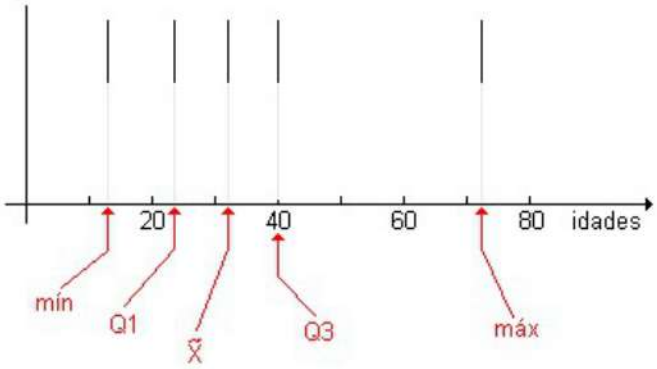


Figura 1

Em seguida, ligamos estes traços verticais por traços horizontais de uma forma convencional para completar o gráfico (Fig. **2**):

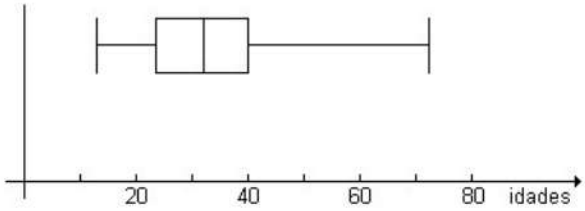


Figura 2

A parte central do gráfico, entre os quartis, é chamada de "caixa" (*box*); as partes entre os quartis e os valores extremos são às vezes chamadas de "bigodes" (*whiskers*). Alguns livros propõem que a espessura da caixa (i.e., o comprimento dos traços verticais) seja usado para representar o tamanho da amostra, mas esta convenção não tem sido seguida pela maioria dos programas de estatística. O eixo com a escala também pode ser colocado na vertical, ser for mais conveniente (como nas Figs. **9** e **10**).

Se a distribuição, ou cada metade dela, tem quantidade par de dados, o quartil será a média entre os dois valores mais próximos do centro. Como exemplo, tomemos a distribuição de idades de pacientes da Amostra B (já usada no gráfico de ramo-e-folhas da Fig. **3** da seção **2.1.2**), cujos valores estão repetidos abaixo:

Amostra B: 20 26 32 32 35 36 41 **42 43** 45 48 48 52 57 59 61

A mediana estará entre 42 e 43 anos; usaremos a média destes valores, e diremos que a mediana é igual a 42,5 anos. O primeiro quartil estará entre 32 e 35 anos:

Amostra B: 20 26 32 32 35 36 41 42 43 45 48 48 52 57 59 61

Usaremos a média, $Q_1 = 33{,}5$ anos. De forma similar, o terceiro quartil estará entre 48 e 52 anos, e receberá o valor de $Q_3 = 50$ anos.

2.1.3.3. Interpretação do diagrama de Tukey

Como exemplo do uso dos diagramas de Tukey, comparamos na Fig. **3** os dois gráficos que representam as amostras A e B:

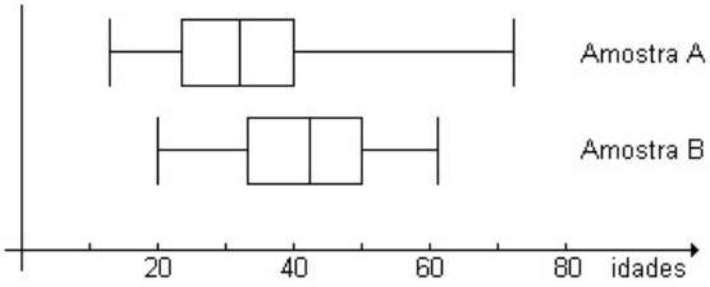


Figura 3

As *posições* das distribuições podem ser comparadas pelas medianas: vemos que os pacientes da amostra B são, em geral, um pouco mais idosos que os da amostra A. As *dispersões* podem ser comparadas pelas distâncias entre os extremos de cada distribuição, ou pelas distâncias entre os quartis. A distância entre os extremos é a "amplitude total" da distribuição; no exemplo, a amplitude total da amostra A é claramente maior que a da B. A distância entre os quartis, que corresponde à largura da caixa do gráfico, é chamada de "intervalo quartílico" (no exemplo, ambas as amostras tem aproximadamente o mesmo intervalo quartílico). Desta forma, os gráficos mostram que, neste exemplo, a dispersão das idades da Amostra A é levemente superior à da Amostra B. (Estas medidas serão vistas novamente na Seção **2.2.2**).

O diagrama de Tukey também pode indicar a assimetria de uma distribuição; isto fica claro quando comparamos os diagramas de Tukey de algumas distribuições com seus respectivos gráficos de pontos. Se uma distribuição é simétrica, os gráficos serão simétricos em relação à mediana, como no exemplo da Fig. **4.** (mediana = 3, $Q_1 = 2$, $Q_3 = 4$).

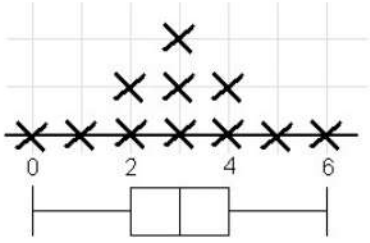
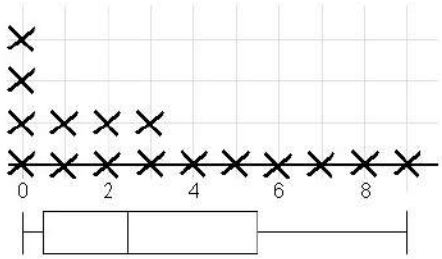
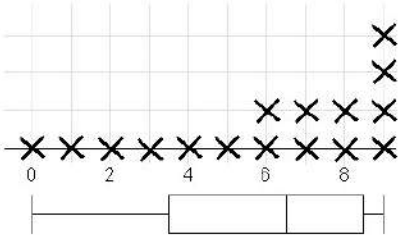


Figura 4

Se a distribuição é assimétrica positiva (o gráfico de pontos mostra uma cauda mais longa para o lado direito), o diagrama de Tukey terá a parte da caixa à direita da mediana maior que a da esquerda, e o bigode da direita mais longo que o da esquerda (Fig. **5A**). Se a distribuição for assimétrica negativa, ocorrerá o contrário (Fig. **5B**).



mediana = 2,5 $Q_1 = 0,5$ $Q_3 = 5,5$
A. Assimetria positiva



mediana = 6,5 $Q_1 = 3,5$ $Q_3 = 8,5$
B. Assimetria negativa

Figura 5. Exemplos de distribuições assimétricas

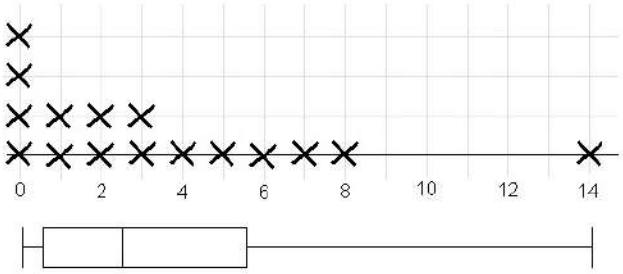
Se a distância entre um quartil e o valor extremo daquele lado for muito grande (isto é, se o bigode do diagrama de Tukey for muito longo), podemos suspeitar de que existam um ou mais valores discrepantes daquele lado. Nos gráficos de pontos e de ramo-e-folhas, a identificação dos discrepantes é feita por inspecção do gráfico; um valor que estiver isolado, afastado dos outros, pode ser considerado discrepante. No diagrama de Tukey, não é possível ver a localização de cada ponto individualmente, portanto a identificação dos discrepantes tem que ser feita de outra forma. O próprio John Tukey sugeriu a seguinte regra: será considerado discrepante o valor que estiver

acima de $Q_3 + 1,5(Q_3-Q_1)$ ou
abaixo de $Q_1 - 1,5(Q_3-Q_1)$

Esta regra tem a vantagem de permitir que os discrepantes sejam idenficados de forma automática, pelos computadores, o que facilita o trabalho de análise de dados. Suponha por exemplo uma amostra composta pelos seguintes valores:

0 0 0 0 1 1 2 2 3 3 4 5 6 7 8 14

O valor extremo 14 pode ser considerado discrepante, como fica evidente no gráfico de pontos (Fig. **6**):



mediana = 2,5 $Q_1 = 0,5$ $Q_3 = 5,5$

Figura 6

Pela regra de Tukey serão considerados discrepantes nesta amostra os valores

acima de $Q_3 + 1,5(Q_3-Q_1) = 5,5 + 1,5\times(5,5-0,5) = 13$
ou abaixo de $Q_1 - 1,5(Q_3-Q_1) = 0,5 - 1,5\times(5,5-0,5) = -7$

Isto confirma que o valor 14 deve ser considerado discrepante. Para destacar estes valores, fazemos uma alteração no diagrama de Tukey: representamos os valores discrepantes por asteriscos, e terminamos o bigode do diagrama no último valor observado não-discrepante. O gráfico do exemplo toma então a forma da Fig. 7:

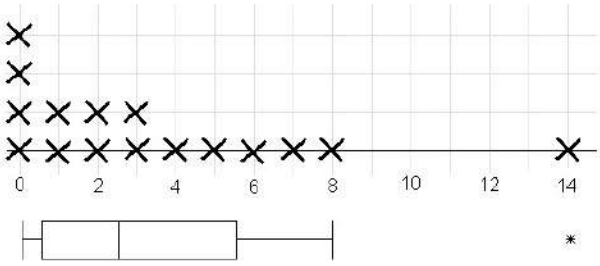


Figura 7

A maior parte dos pacotes estatísticos apresentam o gráfico desta forma, com os valores discrepantes destacados.

2.1.3.4. Quando usar o diagrama de Tukey

A principal vantagem do diagrama de Tukey é a de facilitar a comparação rápida entre mais de duas distribuições. O diagrama não mostra claramente a forma de uma distribuição, aglomerados, modas ou falhas; contudo, indica a dispersão e a assimetria, localiza claramente o centro (mediana), e assinala os pontos discrepantes existentes. Além disso, como este diagrama é feito a partir de apenas cinco valores, ele pode ser usado para qualquer quantidade de dados sem se tornar sobrecarregado, e é por isso ideal para comparar as distribuições de várias amostras grandes.

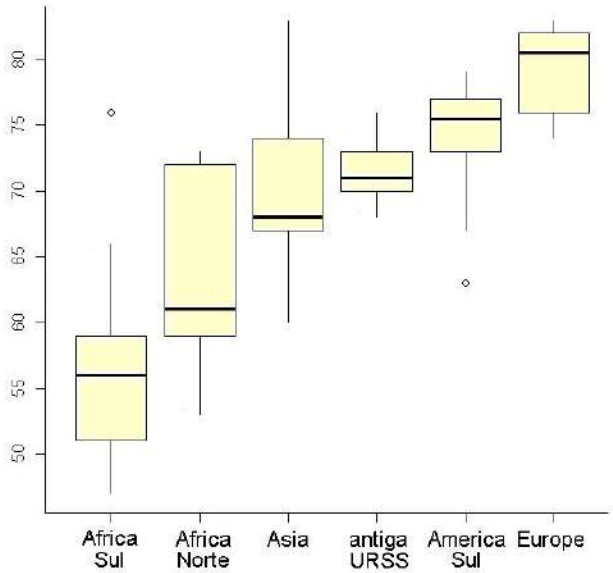


Figura 8 – Expectativas de vida de homens nos países de quatro continentes (2011)
(dados: Wikipedia)

Para exemplificar, a Figura **8** mostra as expectativas de vida de homens dos países de quatro continentes, em 2011 (a expectativa de vida é um indicador simples das condições de saúde de um país). Os diagramas mostram de forma bastante clara que os países da Europa são em geral os que tem as maiores expectativas de vida, enquanto os da África são os que tem as menores. Além disso, mostra também que os dados da África tem mais dispersão do que os outros, o que indica uma maior heterogeneidade nas condições de saúde. Numa situação intermediária está os países da Ásia, e os que faziam parte da antiga União Soviética.

Outro exemplo, na Fig. **9**, mostra o poder dos diagramas de Tukey para sintetizar uma grande quantidade de dados. Cada diagrama mostra as temperaturas a cada hora do dia, nos 365 dias do ano de 1997, na cidade do Rio de Janeiro. A figura portanto representa de forma bastante compacta um total de 24 x 365 = 8760 números, e deixa evidentes os padrões de variação da temperatura na cidade ao longo de um ano.

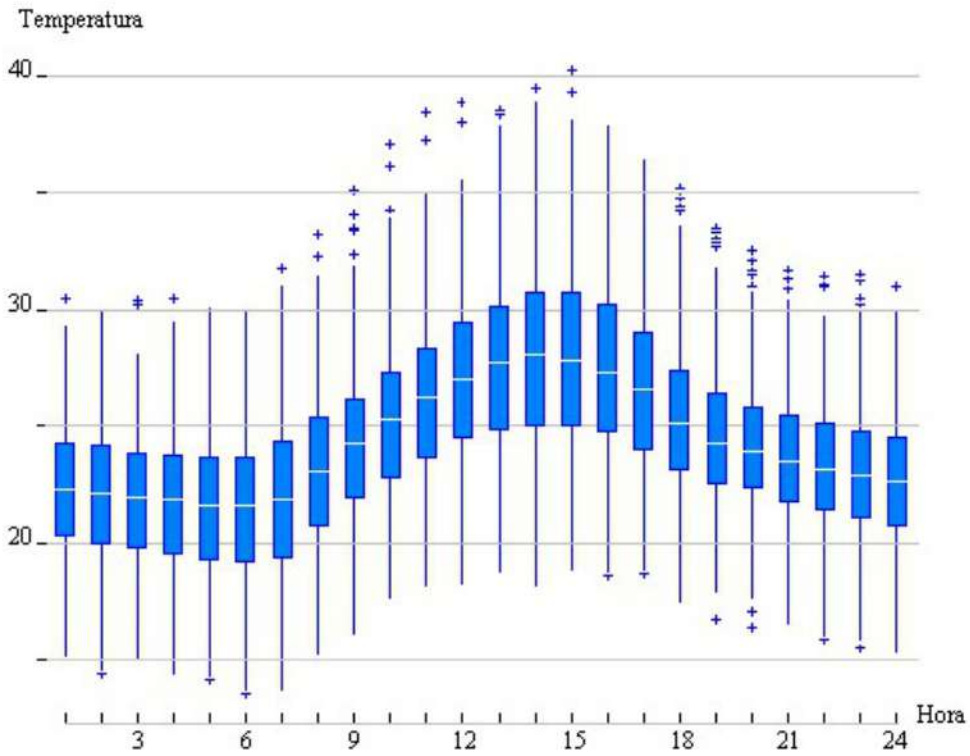


Figura 9 – Temperaturas horários na cidade do Rio de Janeiro, 1997.

2.1.3.5. Algumas variações do diagrama de Tukey

Algumas variações podem ser encontradas na forma do diagrama de Tukey:

a) os diagramas podem ser desenhados na horizontal ou na vertical;
b) os extremos dos "bigodes" podem ser destacados por um traço perpendicular ao diagrama (Figs.**1** a **7**), ou não (Figs. **8** e **9**)
c) os pontos discrepantes podem ser marcados por símbolos diferentes (compare, p. ex., a Fig. **8** com a Fig. **9**); alguns programas distinguem os valores discrepantes que estão fora da faixa de ±1,5 IQ, dos valores ainda mais extremos, que estão fora da faixa ± 3 IQ, e os representam por símbolos diferentes;

d) em alguns programas, a mediana é representada por um asterisco, em vez de um traço, para que não possa ser confundida com um dos quartis.
e) alguns autores sugerem que a espessura das caixas seja usada para representar o tamanho da amostra, quando se comparam amostras de tamanhos diferentes; outros sugerem que a espessura seja proporcional à raiz quadrada do tamanho da amostra. Alguns programas permitem que isto seja feito, mas esta prática ainda não é muito comum.

Resumo

- O diagrama de Tukey é um gráfico muito simplificado, que mostra apenas os valores extremos e os quartis de uma distribuição. Tem menos informação do que o *diagrama de ramo-e-folhas* ou o *histograma*, pois não mostra a forma da distribuição, não indica a posição de aglomerados e modas, etc.

Vantagens:

- O diagrama pode usar a *regra de Tukey* para identificar automaticamente os pontos discrepantes existentes.
- Por ser muito simplificado, o diagrama de Tukey é a melhor ferramenta para comparar graficamente várias distribuições.

Desvantagens:

- O diagrama de Tukey ainda não é bem conhecido pelo público em geral; a maior parte dos leitores não sabe como interpretá-lo.