Quantis

Como vimos nos estudos de probabilidade, uma forma de caracterizar uma variável é a utilização dos quantis, que são valores que dividem os dados ordenados em q subconjuntos de dados de dimensões essencialmente iguais. Os quantis são estabelecidos a partir de pontos de corte que determinam as fronteiras entre os subconjuntos consecutivos.

Alguns dos quantis mais comumente utilizados são os quartis, que dividem os dados ordenados em quatro partes de mesma dimensão. O primeiro quartil Q_1 é o ponto de corte tal que 25% das observações são menores ou iguais a ele e consequentemente 75% das observações são maiores ou iguais a Q_1, o segundo quartil Q_2 é igual à mediana e divide os dados ordenados em duas partes iguais de forma que metade das observações são menores ou iguais a Q_2 e a outra metade das observações são maiores ou iguais a Q_2.

Considere o seguinte conjunto de dados ordenados:

$$5\ 7\ 12\ 14\ 15 \quad 22 \quad 25\ 30\ 36\ 42\ 53$$

Note que à esquerda do número central 22 temos um conjunto com cinco valores menores que ele, e a direita temos outro conjunto com a mesma dimensão de cinco elementos maiores que ele. Dessa forma podemos dizer que metade dos dados são menores ou iguais a 22 e metade dos dados são maiores ou iguais a 22 e assim esse valor é definido como a mediana ou segundo quartil desse conjunto de dados.

Considerando um caso com um conjunto de dados ordenados com um número par de observações:

$$5\ 7\ 12\ 14\ 15\ 22 \quad 25\ 30\ 36\ 42\ 53\ 65$$

Nesse caso não dispomos de um valor central que divida os dados em subconjuntos de mesma dimensão. Nesses casos a estratégia adotada é considerar como segundo quartil ou mediana, um valor médio entre os dois valores centrais do conjunto de dados. Assim definimos $\frac{22+25}{2} = 23,5$ como mediana desse conjunto de dados, pois podemos dizer que metade das observações são menores ou iguais a $23,5$ e a outra metade das observações são maiores ou iguais a $23,5$.

Nos dois exemplos citados, o que fizemos para encontrar as medianas foi estabelecer em qual posição encontrava-se a medida central e usar seu valor como ponto de corte entre os grupos, no caso em que não existia uma única medida central, determinou-se quais eram as duas medidas centrais e calculou-se uma média entre elas. As medidas do tipo quantil são chamadas medidas de posição.

Lembrando que podemos dividir os dados em quantas partes quisermos, consideremos a divisão em 100 partes iguais e teremos os chamados percentis como pontos de corte entre os grupos. Seguindo a ideia utilizada no caso da mediana ou segundo quartil, precisamos encontrar em qual posição dos dados encontra-se o ponto de corte que dividirá os grupos da forma desejada.

Considerando um conjunto de n observações ordenadas utilizaremos a notação $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, em que o índice representa a posição dentre as observações ordenadas, sendo $x_{(1)}$ a menor observação e $x_{(n)}$ a maior. Um determinado percentil p de um conjunto de dados com n observações ordenadas é dado por $x_{(L)}$, sendo L obtido por:

$$L = n\frac{p}{100} \begin{cases} \text{se,} \quad L \quad \text{é um número inteiro tomamos} \quad \frac{x_{(L)}+x_{(L+1)}}{2} \\ \text{se,} \quad L \quad \text{é decimal, arredonda-se para o maior inteiro e toma-se} \quad x_{(L)}. \end{cases}$$

Exemplos:

No conjunto de dados:

5 7 12 14 15 22 25 30 36 42 53

para encontrar o primeiro quartil temos $L = 11\frac{25}{100} = 2,75$ e assim definimos o primeiro quartil $Q_1 = x_{(3)}$ e como 12 é a terceira observação ordenada temos $Q_1 = 12$ e podemos dizer que 25% das observações são menores ou iguais a 12 e 75% das observações são maiores ou iguais a 12.

Para encontrar o segundo quartil temos $L = 11\frac{50}{100} = 5,5$ e assim definimos a mediana ou segundo quartil $Q_2 = x_{(6)}$ e portanto a mediana é igual a 22.

Para encontrar o terceiro quartil temos $L = 11\frac{75}{100} = 8,25$ e assim definimos a mediana ou segundo quartil $Q_3 = x_{(9)}$ e portanto o terceiro quartil é igual a 36.

No conjunto de dados:

5 7 12 14 15 22 25 30 36 42 53 65

para encontrar o primeiro quartil temos $L = 12\frac{25}{100} = 3$ e assim definimos o primeiro quartil como $Q_1 = \frac{x_{(3)}+x_{(4)}}{2}$ e portanto o primeiro quartil é igual a $\frac{12+14}{2} = 13$.

Para encontrar o segundo quartil temos $L = 12\frac{50}{100} = 6$ e assim definimos o primeiro quartil como $Q_2 = \frac{x_{(6)}+x_{(7)}}{2}$ e portanto o primeiro quartil é igual a $\frac{22+25}{2} = 23,5$.

Para encontrar o terceiro quartil temos $L = 12\frac{75}{100} = 9$ e assim definimos o primeiro quartil como $Q_3 = \frac{x_{(9)}+x_{(10)}}{2}$ e portanto o primeiro quartil é igual a $\frac{36+42}{2} = 39$.

Uma medida muito usada, que descreve a variabilidade dos dados, é a amplitude inter quartílica: $q_3 - q1 = A_q$.

Box Plot ou Diagrama em Caixas

Esse tipo de gráfico descreve simultaneamente diversas informações de um conjunto de dados, tais como, centro, dispersão, simetria e identificação de valores extremos.

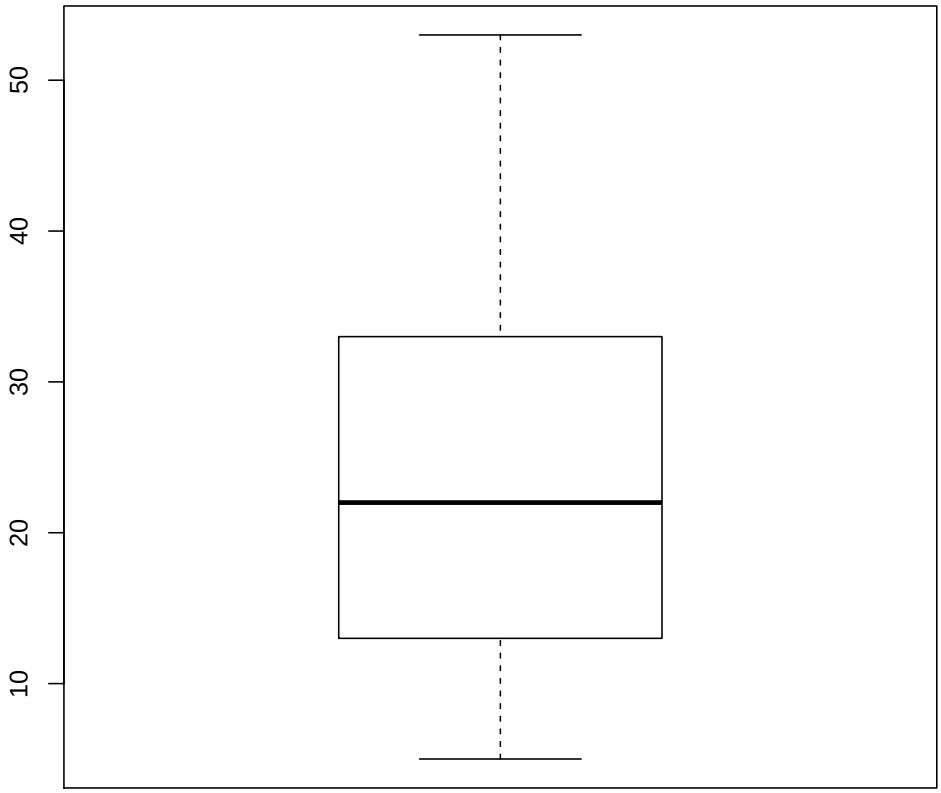
Para construir esse diagrama, traça-se dois retângulos: um representando o espaço entre o quartil inferior e a mediana e outro entre a mediana e o quartil superior, representando a faixa dos 50% dos valores mais típicos da distribuição. A partir do retângulo, para cima, segue uma linha até o ponto mais remoto que não exceda $LS = q_3 + 1,5A_q$, chamado limite superior. De

forma análoga, da parte inferior do retângulo, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que $LI = q_1 - 1,5A_q$, chamado limite inferior. As observações que estiverem acima do limite superior ou abaixo do limite inferior serão chamadas de pontos extremos e representados por pontos ou asteriscos.

Exemplo: Vejamos o boxplot do exemplo com conjunto de dados com as 11 observações

5 7 12 14 15 22 25 30 36 42 53

vimos que $Q_1 = 12$, $Q_2 = 22$ e $Q_3 = 36$



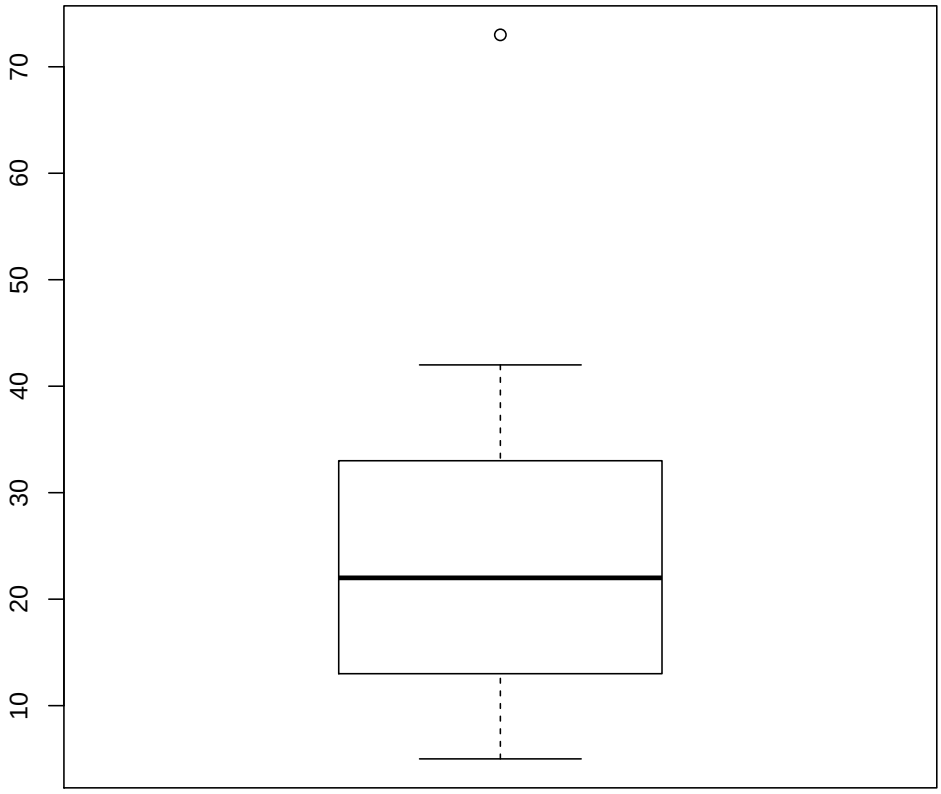
Note que os maiores e menores valores observados encontram-se dentro dos limites estabelecidos para valores extremos. Temos $1,5(Q_3 - Q_1) = 1.5(36 - 12) = 36$ e assim qualquer observação menor que $Q_1 - 36 = -24$ ou maior que $Q_3 + 36 = 72$ são considerados valores extremos e devem ser identificados no boxplot. Como todas as observações estão contidas nesse intervalo as pontas das retas que se estendem do retângulo vão até os valores mínimo $x_{(1)} = 5$ e máximo $x_{(11)} = 53$.

Considerando o conjunto de dados alternativo:

$$5\ 7\ 12\ 14\ 15\ 22\ 25\ 30\ 36\ 42\ 73$$

Os valores dos quartis se mantem inalterados $Q_1 = 12$, $Q_2 = 22$ e $Q_3 = 36$, mas note que $1,5(Q_3 - Q_1) = 1.5(36 - 12) = 36$ e assim qualquer observação menor que $Q_1 - 36 = -24$ ou maior que $Q_3 + 36 = 72$ são considerados valores extremos e devem ser identificados no boxplot.

A maior observação $x_{(11)} = 73$ se encontra fora desse intervalo e recebe a identificação de valore extremo no boxplot com a marcação do ponto por um símbolo do tipo $*$. As pontas das retas que se estendem do retângulo vão até os valores mais extremos contidos dentro dos intervlos $x_{(1)} = 5$ e $x_{(10)} = 42$.



Este tipo de gráfico é muito útil para comparar de forma simples e visual diferentes grupos:

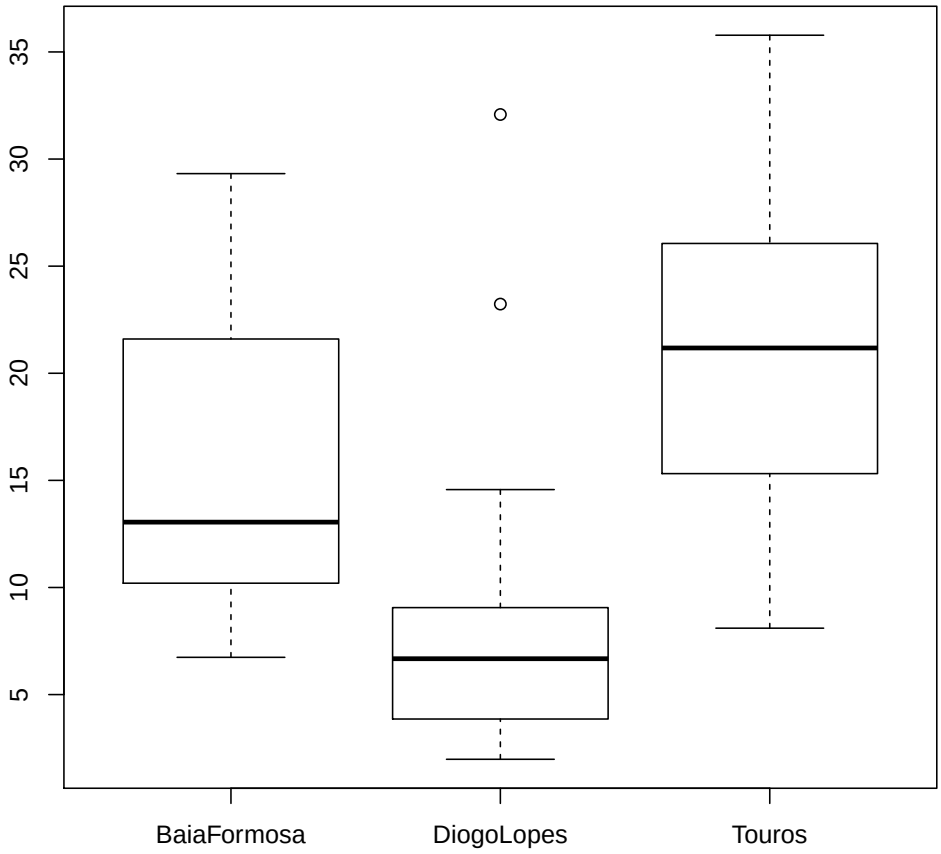


Figura 1: Distribuições dos pesos de camarões coletados em diferentes localidades.

Nesta figura podemos visualizar as diferenças entre os pesos dos camarões coletados em 3 diferentes localidades. Pode-se perceber que grande parte dos camarões mais pesados foram pescados na localidade de Touros.