Estatística Básica

Prof. Fernando de Souza Bastos
fernando.bastos@ufv.br

Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa
Campus UFV - Florestal

Sumário

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a variabilidade do conjunto de observações. Por exemplo, suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

- Grupo A (Variável X): 3,4,5,6,7
- Grupo B (Variável Y): 1,3,5,7,9
- Grupo C (Variável Z): 5,5,5,5,5
- Grupo D (Variável W): 3,5,5,7
- Grupo E (Variável V): 3,5,5,6,6

Vemos que $\bar{x} = \bar{y} = \bar{z} = \bar{w} = \bar{v} = 5$. A identificação de cada uma destas séries por sua média (5, em todos os casos) nada informa sobre suas diferentes variabilidades. Notamos, então, a conveniência de serem criadas medidas que sumarizem a variabilidade de um conjunto de observações e que nos permita, por exemplo, comparar conjuntos diferentes de valores, como os dados acima, segundo algum critério estabelecido.

Um critério frequentemente usado para tal fim é aquele que mede a dispersão dos dados em torno de sua média. Duas medidas são as mais usadas: desvio médio e variância.

$$Dm(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad Var(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

O princípio básico é analisar os desvios das observações em relação à média dessas observações. Desvio é interpretado como o afastamento de uma observação em relação a uma determinada medida de posição.

Sendo a variância uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em cm, a variância será expressa em cm^2), ela pode causar problemas de interpretação. Costuma-se usar, então, o desvio padrão, que é definido como a raiz quadrada positiva da variância.

$$dp(X) = \sqrt{Var(X)} \tag{1}$$

Ambas as medidas de dispersão (Dm e dp) indicam em média qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (no caso, a média).

Coeficiente de Variação

$$CV(X) = \frac{dp(X)}{\bar{X}} \tag{2}$$

Mesmo o DP pode induzir à conclusões errôneas com relação à variabilidade. Suponha dois conjuntos de dados $D_1 = \{10, 20, 30\}$ e $D_2 = \{10000, 10010, 10020\}$. Note que nestes casos $\bar{x}_1 = 20, dp(x) = 10, \bar{x}_2 = 10010$ e $dp(x_2) = 10$. Porém, em termos percentuais, o primeiro conjunto de dados é mais heterogênio.

Medidas Complementares para Análise de Dados

- Extremos: O menor e o maior valor do conjunto de dados;
- Quartis (Q)
  - 1° Quartil: deixa um quarto dos valores abaixo, e três quartos acima dele;
  - 2° Quartil = Mediana: deixa metade dos valores abaixo, e metade acima dele;
  - 3° Quartil: deixa três quartos dos valores abaixo, e um quarto acima dele;
- Intervalo Interquartil (pode ser considerada uma medida robusta de dispersão).

Distância Interquartil

Uma medida de dispersão alternativa ao desvio padrão é a distância interquartil, definida como a diferença entre o terceiro e primeiro quartis, ou seja:

$$d_q = q_3 - q_1$$

Os cinco valores, $x_{(1)}, q_1, q_2, q_3$ *e* $x_{(n)}$ são importantes para se ter uma boa ideia da assimetria da distribuição dos dados. Para uma distribuição simétrica ou aproximadamente simétrica, deveríamos ter:

(a) $q_2 - x_{(1)} \approx x_{(n)} - q_2$;

(b) $q_2 - q_{(1)} \approx q_{(3)} - q_2$;

(c) $q_1 - x_{(1)} \approx x_{(n)} - q_3$;

(d) distâncias entre mediana e q_1, q_3 menores do que distâncias entre os extremos e q_1, q_3.

A diferença $q_2 - x_{(1)}$ é chamada dispersão inferior e $x_{(n)} - q_2$ é a dispersão superior. A condição (a) nos diz que estas duas dispersões devem ser aproximadamente iguais, para uma distribuição aproximadamente simétrica. A próxima Figura ilustra estes fatos para a chamada distribuição normal ou gaussiana.

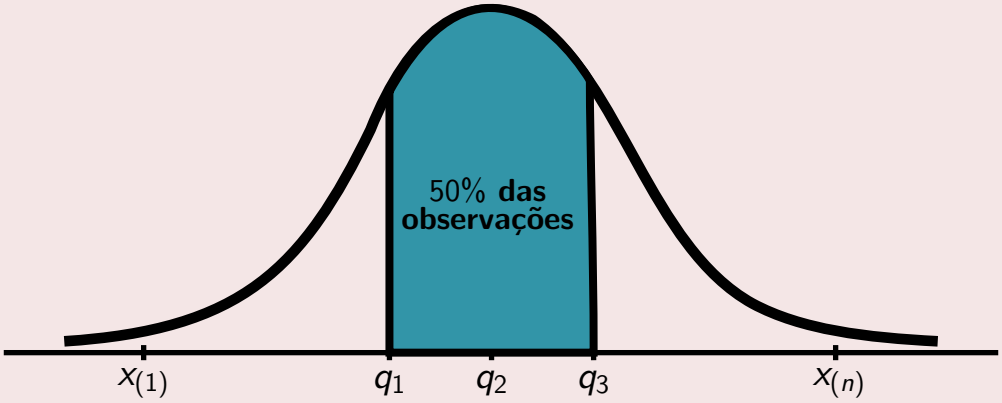


Figura: Uma distribuição simétrica: normal ou gaussiana Morettin and Bussab (2009).

As cinco estatísticas de ordem consideradas acima podem ser representadas esquematicamente como na próxima Figura, onde também incorporamos o número de observações, n. Representamos a mediana por md, os quartis por q e os extremos por E.

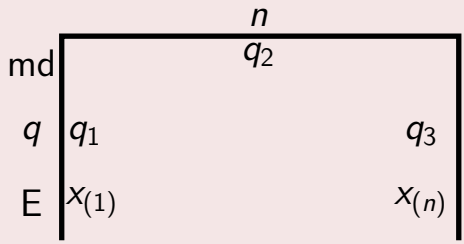


Figura: Esquema dos cinco números (Morettin and Bussab (2009)).

Box Plot

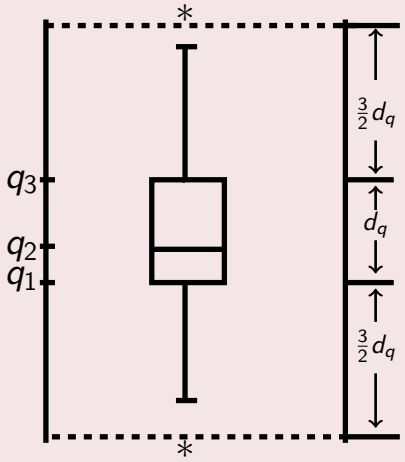
O boxplot (gráfico de caixa) é um gráfico utilizado para avaliar a distribuição empírica do dados. O boxplot é formado pelo primeiro e terceiro quartil e pela mediana. Para construir este diagrama, consideremos um retângulo onde estão representados a mediana e os quartis. A partir do retângulo, para cima, segue uma linha até o ponto mais remoto que não exceda $LS = q_3 + (1,5)d_q$, chamado limite superior. De modo similar, da parte inferior do retângulo, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que $LI = q_1 - (1,5)d_q$, chamado limite inferior.

Box-Plot

Os valores compreendidos entre esses dois limites são chamados valores adjacentes. As observações que estiverem acima do limite superior ou abaixo do limite inferior estabelecidos serão chamadas pontos exteriores e representadas por asteriscos. Essas são observações destoantes das demais e podem ou não ser o que chamamos de outliers ou valores atípicos.

Box-Plot

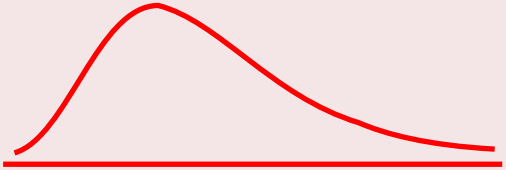
Figura: Esquema de um BoxPlot Morettin and Bussab (2009).



O box plot dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão por d_q. As posições relativas de q_1, q_2, q_3 dão uma noção da assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores remotos e pelos valores atípicos. Veja esse exemplo.

Assimetrias

Figura: Distribuições assimétricas Morettin and Bussab (2009).



Assimétrica à direita

Figura: Distribuições assimétricas Morettin and Bussab (2009).

Assimétrica à esquerda

Transformações

Vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma distribuição normal (em forma de sino) ou então mais ou menos simétrica. Mas, em muitas situações de interesse prático, a distribuição dos dados da amostra é assimétrica e pode conter valores atípicos, como vimos em exemplos anteriores.

Se quisermos utilizar tais procedimentos, o que se propõe é efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. Uma família de transformações frequentemente utilizada é

$$X^* = \begin{cases} X^p, & \text{se} \quad \text{p}>0 \\ \ln(X), & \text{se} \quad \text{p}=0 \\ -X^p, & \text{se} \quad \text{p}<0 \end{cases}$$

Normalmente, o que se faz é experimentar valores de p na sequência $\cdots, -3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, \cdots$

Para cada valor de p obtemos gráficos apropriados (histogramas, desenhos esquemáticos etc.) para os dados originais e transformados, de modo a escolhermos o valor mais adequado de p. Para dados positivos, a distribuição dos dados é usualmente assimétrica à direita. Para essas distribuições, a transformação acima com $0 < p < 1$ é apropriada, pois valores grandes de x decrescem mais, relativamente a valores pequenos. Para distribuições assimétricas à esquerda, tome $p > 1$.

Referências

P. Morettin and W. Bussab. *Estatística básica*. Editora Saraiva, São Paulo, 6 edition, 2009.