CC-226 Introdução à Análise de Padrões

Probabilidades e Estatísticas Descritivas

Carlos Henrique Q. Forster[1]

[1]Divisão de Ciência da Computação
Instituto Tecnológico de Aeronáutica

3 de março de 2008

Outline

Definições

- **Fenômeno aleatório** é um fenômeno empírico caracterizado pela propriedade que sua observação sob um dado conjunto de circunstâncias não leva sempre ao mesmo resultado observado, mas a outros resultados mantendo uma regularidade estatística. (Parzen)
- **Evento aleatório** é aquela condição cuja freqüência de ocorrência aproxima-se de um valor limite estável quando o número de observações tende ao infinito. (Parzen)
- **Espaço de descrição amostral** de um fenômeno é o espaço das descrições de todos os possíveis resultados do fenômeno. (Parzen)

Eventos

- Formalmente, eventos são representados por conjuntos e podem ser definidos através das operações de complemento e uniões contáveis de conjuntos. O conjunto de todos os eventos (um conjunto de conjuntos) mais as operações de complemento e uniões contáveis formam uma σ-álgebra.
- Definidos complemento e união, a intersecção é conseqüência do Teorema de DeMorgan. Assim, operações booleanas podem ser aplicadas a eventos.

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

Mais definições

- **Amostra**, observação ou instância é a descrição do resultado observado de um fenômeno aleatório.
- **População** é o conjunto de objetos de interesse. O conjunto de amostra é um subconjunto da população. (Devore)
- Uma **variável** é qualquer característica (associada a um valor) que pode mudar de um objeto a outro da população. (Devore)
- Dados univariados, bivariados e multivariados contém respectivamente uma, duas ou múltiplas variáveis.
- Uma **variável aleatória** é um mapa do espaço amostral sobre a reta de Borel (reta real mais os símbolos $+\infty$ e $-\infty$).
- Variável **discreta** é aquela cujo espaço amostral é finito.

Gráficos

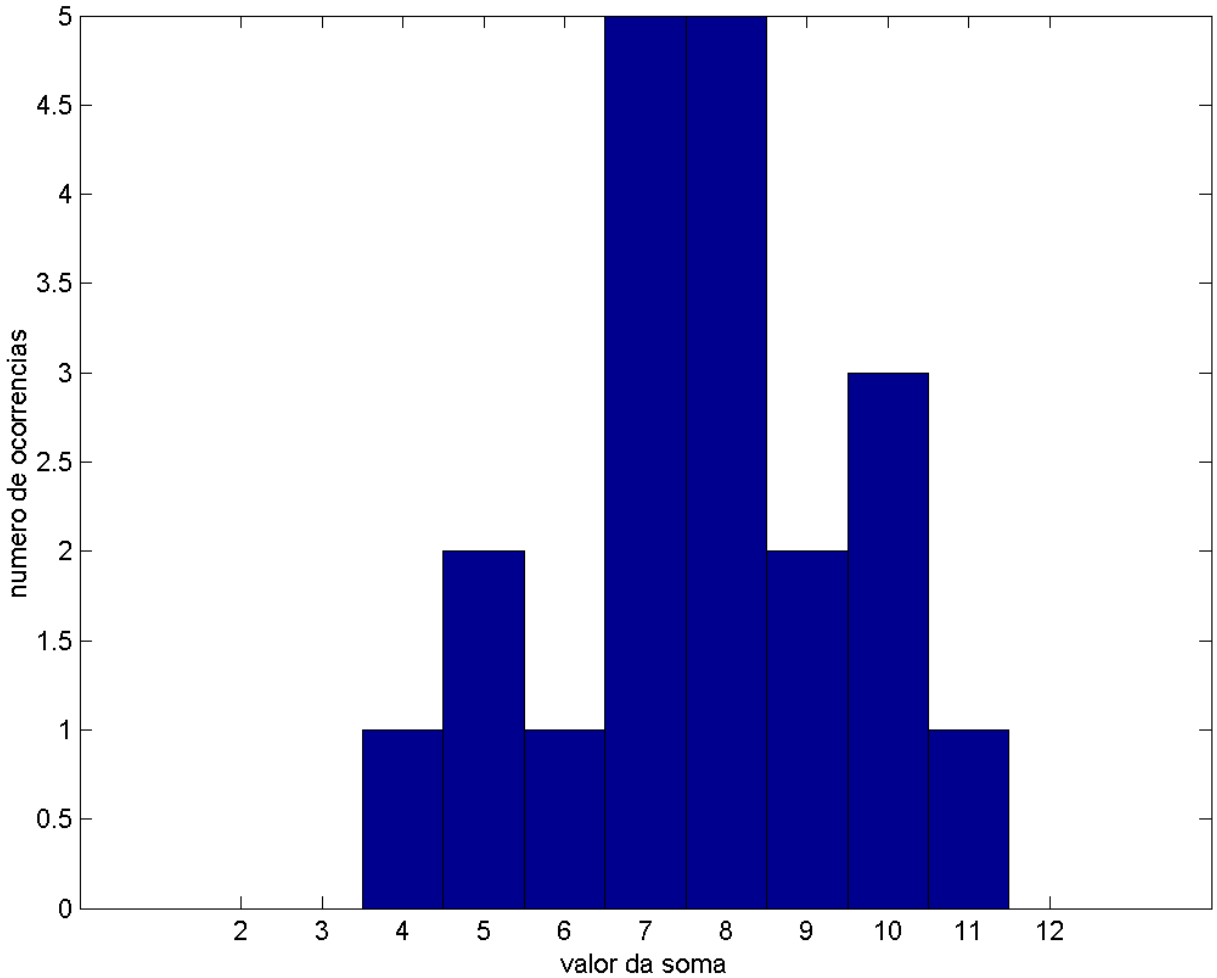
Há três tipos de gráficos mais importantes para descrever uma ou mais variáveis aleatórias.

- Histogramas
- Diagramas de dispersão (scatterplots ou scattergrams)
- Boxplots

Construção do histograma

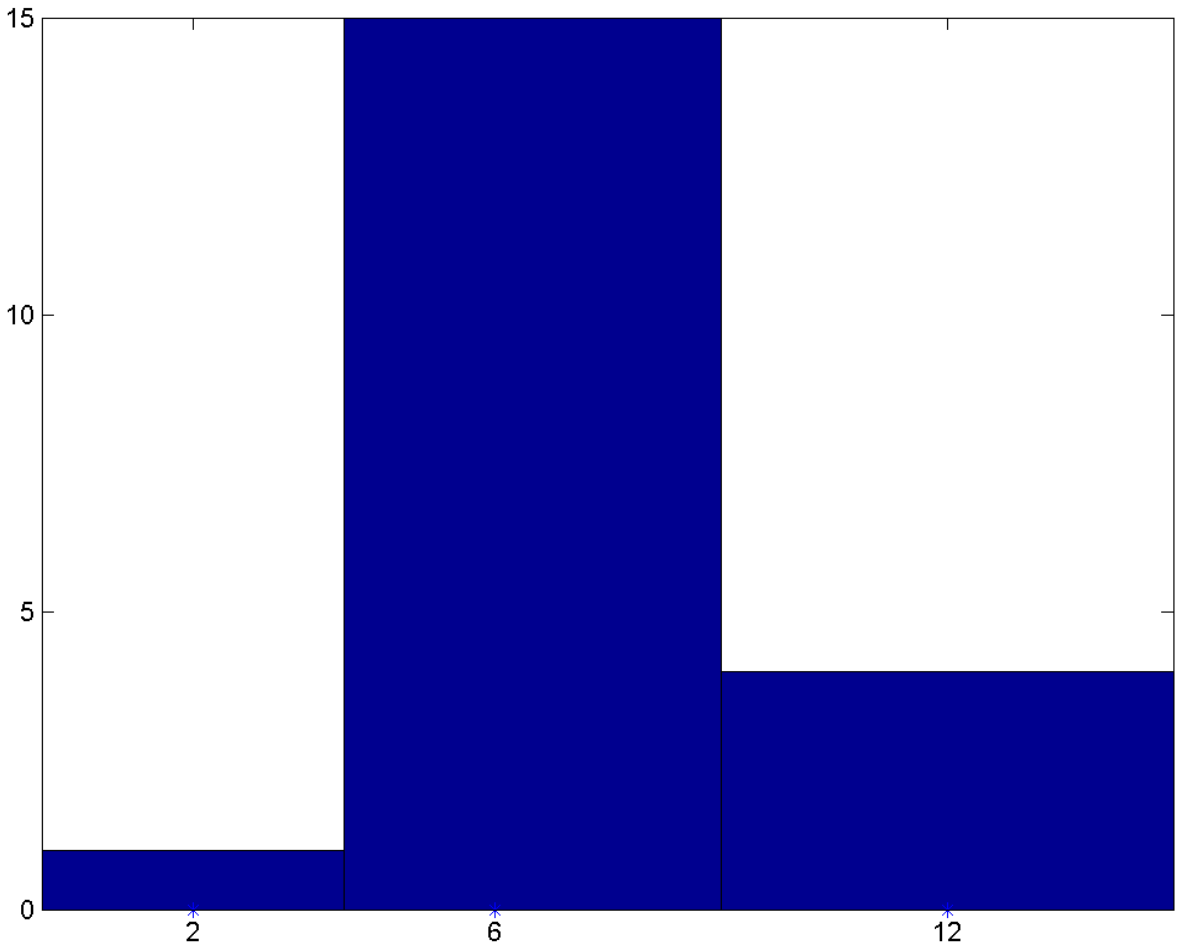
- A construção do histograma de uma variável (dadas várias observações) compreende a partição do espaço em um conjunto de classes e plotar o número de ocorrências ou a freqüência relativa de um valor dentro de cada partição.
- Fenômeno: jogar pares de dados e obter a soma.
- Valores obtidos: 9 7 10 7 10 8 8 5 5 6 7 7 8 8 4 10 7 9 11 8

Para partições de mesmo tamanho, a altura do retângulo é proporcional à freqüência relativa.



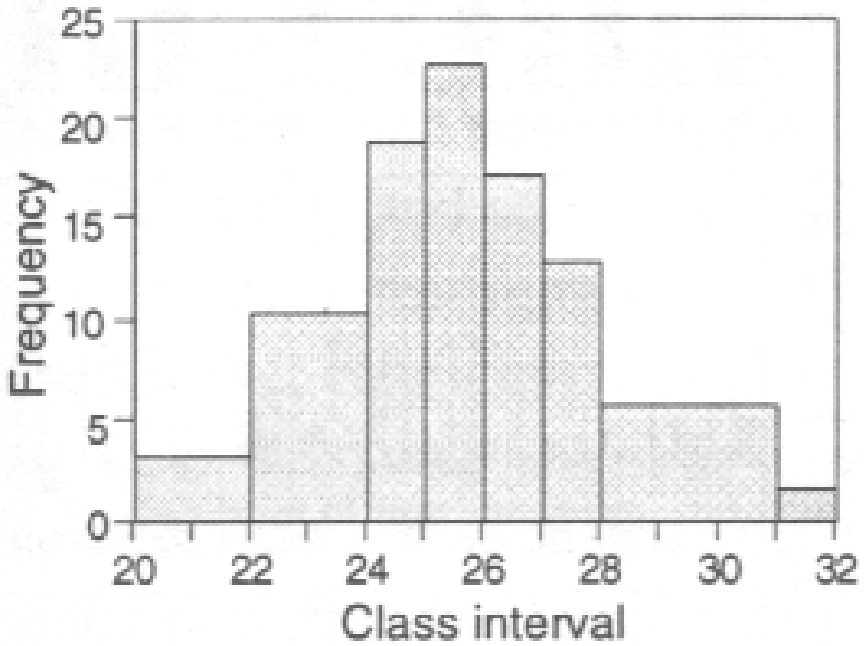
Histogramas com partições não uniformes

A partição não precisa ser uniforme...



Histogramas com partições não uniformes

A partição não precisa ser uniforme...



The histogram is probably the most widely used application of an area column graph. In a histogram, as the width of a column is expanded to cover a broader class interval, the height is changed accordingly.

- O certo seria utilizar uma escala baseada na área do retângulo, de forma que esta represente a densidade dos pontos.
- Para partições uniformes, isso já é verdade.
- Para partições não-uniformes, devemos calcular a altura do retângulo para que a sua área seja proporcional à freqüência relativa correspondente.
- Essa á chamada de escala de densidade.

No caso em que as partições não são uniformes, a altura do retângulo deve representar a densidade e pode ser calculada da seguinte maneira:

$$\texttt{altura do retângulo} = \frac{\texttt{freq relativa da classe}}{\texttt{largura da classe}}$$

onde

$$\text{freq relativa de um valor} = \frac{\text{ocorrências do valor}}{\text{número de observações total}}$$

Código MATLAB para histograma dos dados

```
dados=floor(rand(20,2)*6+1);
somas=sum(dados')';
disp(somas');
hist(somas,[2:12]);
xlabel('valor da soma');
ylabel('numero de ocorrencias');
```

Primeiro Exercício de Implementação

- Escrever uma função em MATLAB para desenhar um histograma de forma a particionar o espaço amostral em classes com aproximadamente a mesma freqüência de ocorrência e, portanto, os retângulos devem ter áreas semelhantes.
- A entrada deve ser um vetor de números reais correspondendo a amostras de uma distribuição ou experimento e um número inteiro correspondendo ao número de classes.
- Apresentar gráficos para o experimento anterior variando o número de dados e o número de jogadas.
- Apresentar gráficos para o experimento: jogar 5 dados, somar e elevar ao quadrado, 1000 jogadas, aproximadamente 20 retângulos.

Código para desenhar retângulos

```
figure; hold on;
for i=1:3
    x0=5+8*i; x1=10+8*i; y0=3+i*i; y1=6+i*i;
    plot([x0,x0,x1,x1,x0],[y0,y1,y1,y0,y0],'r-');
end;
```

Descritores de tendência central

- Descritores de tendência central buscam representar uma variável aleatória por um único valor representativo.

  média relacionada ao centro de massa. Valores muito discrepantes têm grande influência sobre a medida.

  mediana freqüência de valores acima é igual à freqüência de valores abaixo. Não importa a posição desses valores, só se são maiores ou menores que a mediana. Valores distantes não afetam a mediana.

  moda representa o valor mais freqüente. Pode-se falar em mais de uma moda quando há tendência de freqüência alta em valores díspares ou há mistura de modelos.

  mediatriz representa o ponto central do intervalo que contém as amostras. Depende apenas da amostra de valor mínimo e da de valor máximo.

Média amostral

Para n amostras $x_1, x_2, \ldots, x_n$ a média amostral $\bar{x}$ é definida como

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \sum_{i=1}^{n} x_i$$

Seu significado é diferente da média da população μ.

Diferença médias amostral e populacional

Exemplo da diferença da média da população para a média amostral. Considere um dado de seis lados. Qual a média esperada para jogadas desse dado?

$$\frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3,5$$

Suponha que joguei o dado 5 vezes e obtive: 2, 3, 3, 6, 1. O que é plenamente possível.
A média amostral é dada por

$$\frac{2+3+3+6+1}{5} = 3,0$$

Assim, $\mu = 3,5$ e $\bar{x} = 3,0$.

Média de uma variável binária

No caso de uma variável binária, faz-se seus valores 0 e 1.

$$x_i = \begin{cases} 1 & \text{se pertence à categoria} \\ 0 & \text{caso contrário} \end{cases}$$

Neste caso a média fornece a proporção amostral.

Mediana amostral

Considere os dados ordenados. A mediana amostral $\tilde{x}$ é definida por

$$\tilde{x} = \begin{cases} \frac{n+1}{2}\text{-ésimo valor}, & \text{para } n \text{ ímpar} \\ \text{média dos valores de índices } \frac{n}{2} \text{ e } \frac{n}{2}+1, & \text{para } n \text{ par} \end{cases}$$

Médias aparadas

Considere os dados ordenados. A média aparada consiste na média dos elementos centrais, descartando, por exemplo, os valores 10% maiores e os 10% menores. Quando a porcentagem descartada se aproxima de zero, a média aparada equivale à média, quando se aproxima de 100%, equivale à mediana.

Revendo Distâncias

Definimos uma **métrica** (ou distância) como uma função real positiva sobre um par de elementos do mesmo espaço $\rho : X \times X \rightarrow \Re^+$ que possui as seguintes propriedades:

$$\rho(A, A) = 0$$

$$\rho(A, B) = \rho(B, A)$$

$$\rho(A, C) \leq \rho(A, B) + \rho(B, C)$$

Quando queremos apenas uma medida de dissimilaridade para comparar objetos, a propriedade da desigualdade triangular não é necessária.

Distâncias de Minkowski

Consideramos como exemplo dois pontos no espaço 3D:

$$P_1 = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \text{e} P_2 = \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix}$$

A distância de Minkowski é dada por

$$d_{L_p}(P_1, P_2) = \sqrt[p]{|x_1 - x_2|^p + |y_1 - y_2|^p + |z_1 - z_2|^p}$$

Trata-se de uma forma geral para definir distâncias particularmente importantes.

A distância de Manhattan ou distância L_1 conta o número de quarteirões que separa dois pontos, andando sempre nas direções dos eixos ordenados.

$$d_{L_1}(P_1, P_2) = |x_1 - x_2| + |y_1 - y_2| + |z_1 - z_2|$$

A distância de Chebyshev ou distância L_∞ corresponde à maior dimensão do retângulo de arestas paralelas aos eixos e que contém os dois pontos como vértices opostos.

$$d_{L_\infty}(P_1, P_2) = \max\{|x_1 - x_2|, |y_1 - y_2|, |z_1 - z_2|\}$$

A distância Euclidiana ou distância L_2 corresponde ao comprimento da reta que une os dois pontos.

$$d_{L_2}(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

Algumas outras distâncias

- A distância de Hamming de duas cadeias de bits de mesmo comprimento corresponde ao número de bits invertidos de uma cadeia para outra. No caso de conjuntos corresponde ao número de elementos presente em A e não-presente em B, mais o número de elementos presente em B, mas não-presente em A.
- Distâncias de edição entre dois objetos corresponde ao número de operações de edição que devem ser efetuadas para transformar um objeto no outro. Exemplo: distância de Levenshtein.
- A distância de Haussdorff para conjuntos de pontos A e B corresponde à maior distância mínima entre um ponto de A e um ponto de B.

O valor de mínima distância Euclidiana aos dados

Procuramos $\bar{x}$ que é um valor cuja distância euclidiana aos dados $x_i, i = 1 \ldots N$ é mínima.
Definimos uma função energia a minimizar.

$$E = \sum_{i=1}^{N} (\bar{x} - x_i)^2$$

A raiz quadrada foi omitida por ser uma função crescente em $(0, +\infty)$.
Encontramos o menor E igualando-se seu gradiente em função de $\bar{x}$ a zero. No caso,

$$\nabla E = \frac{\partial E}{\partial \bar{x}}$$

Igualando a zero,

$$\frac{\partial E}{\partial \bar{x}} = 2\sum_{i=1}^{N}(\bar{x} - x_i) = 0$$

Observe que o valor que procuramos anula a soma dos desvios:

$$\sum_{i=1}^{N}(\bar{x} - x_i) = 0$$

Reescrevendo, obtemos

$$\sum_{i>1}^{N}\bar{x} - \sum_{i=1}^{N}x_i = N\bar{x} - \sum_{i=1}^{n}x_i = 0$$

Logo, obtemos a média amostral.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i$$

A mediana como mínima distância L_1

Mostramos de forma análoga que a mediana satisfaz a minimização da distância L_1. Desconsideramos as descontinuidades, por simplicidade.

$$E = \sum_{i=1}^{N} |\tilde{x} - x_i|$$

Derivando e igualando a zero,

$$\frac{\partial E}{\partial \tilde{x}} = \sum_{i=1}^{N} sgn(\tilde{x} - x_i) = 0$$

Separando o somatório,

$$\sum_{i:x_i>\tilde{x}} 1 + \sum_{i:x_i<\tilde{x}} (-1) = 0$$

Assim, o número de instâncias maior que $\tilde{x}$ deve ser igual ao número de instâncias maiores que $\tilde{x}$.

$$\sum_{i:x_i>\tilde{x}} 1 = \sum_{i:x_i<\tilde{x}} 1$$

Medidas de Variabilidade ou Dispersão

- A amplitude é a diferença entre o maior e o menor valor.
- A diferença inter-quartil é a diferença entre o quartil superior e o quartil inferior. Os quartis são valores que separam 25% dos dados.
- A variância é uma medida de dispersão relacionada a um modelo de inércia da amostra. Considere os desvios em relação à média amostral $x_i - \bar{x}$. O somatório dos desvios é nulo. Para número de elementos da amostra n grande, a variância é dada pela média dos quadrados dos desvios chamada σ^2.
- O desvio-padrão σ é a raiz-quadrada da variância. No caso de uma distribuição normal, referimo-nos ao número de sigmas que uma amostra está distante da média.

Variância

Para amostras grandes a variância é definida como:

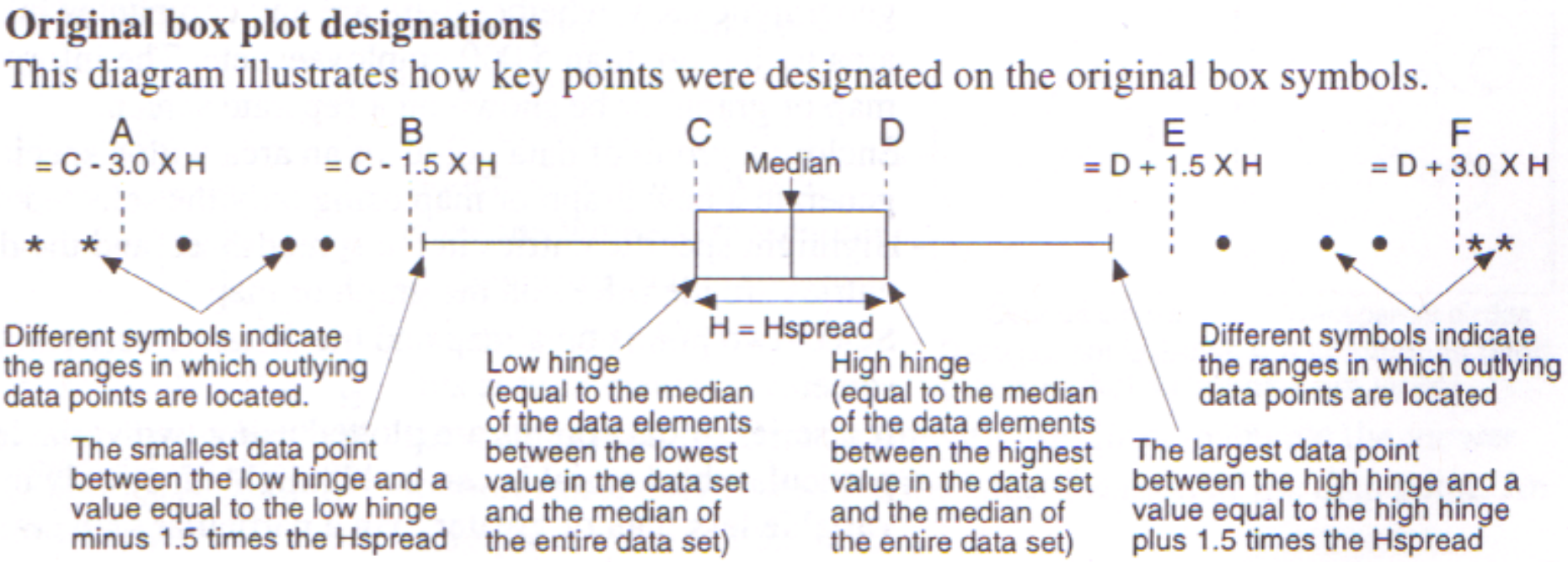
$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

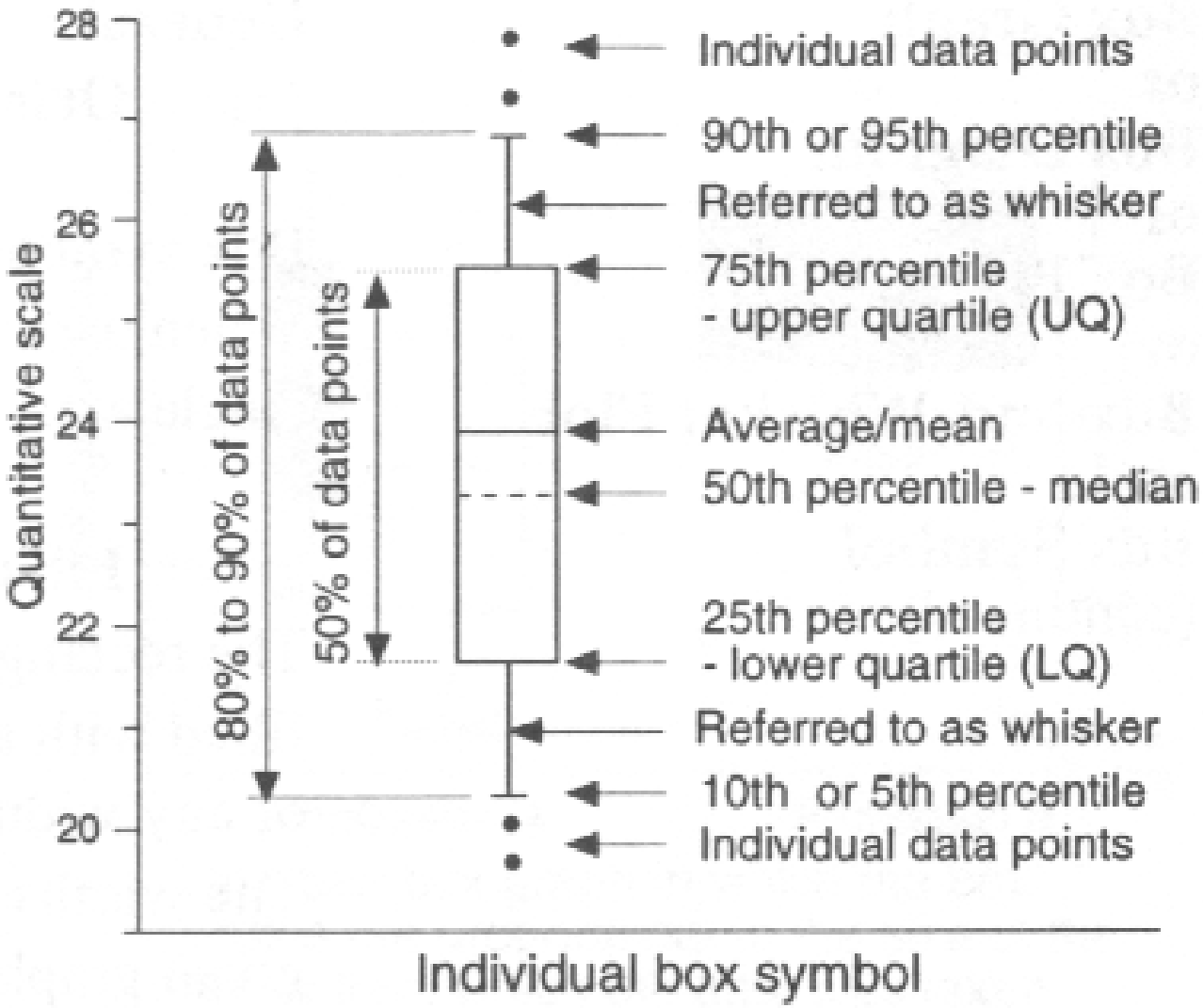
Para número pequeno de elementos da amostra, a variância é dada por:

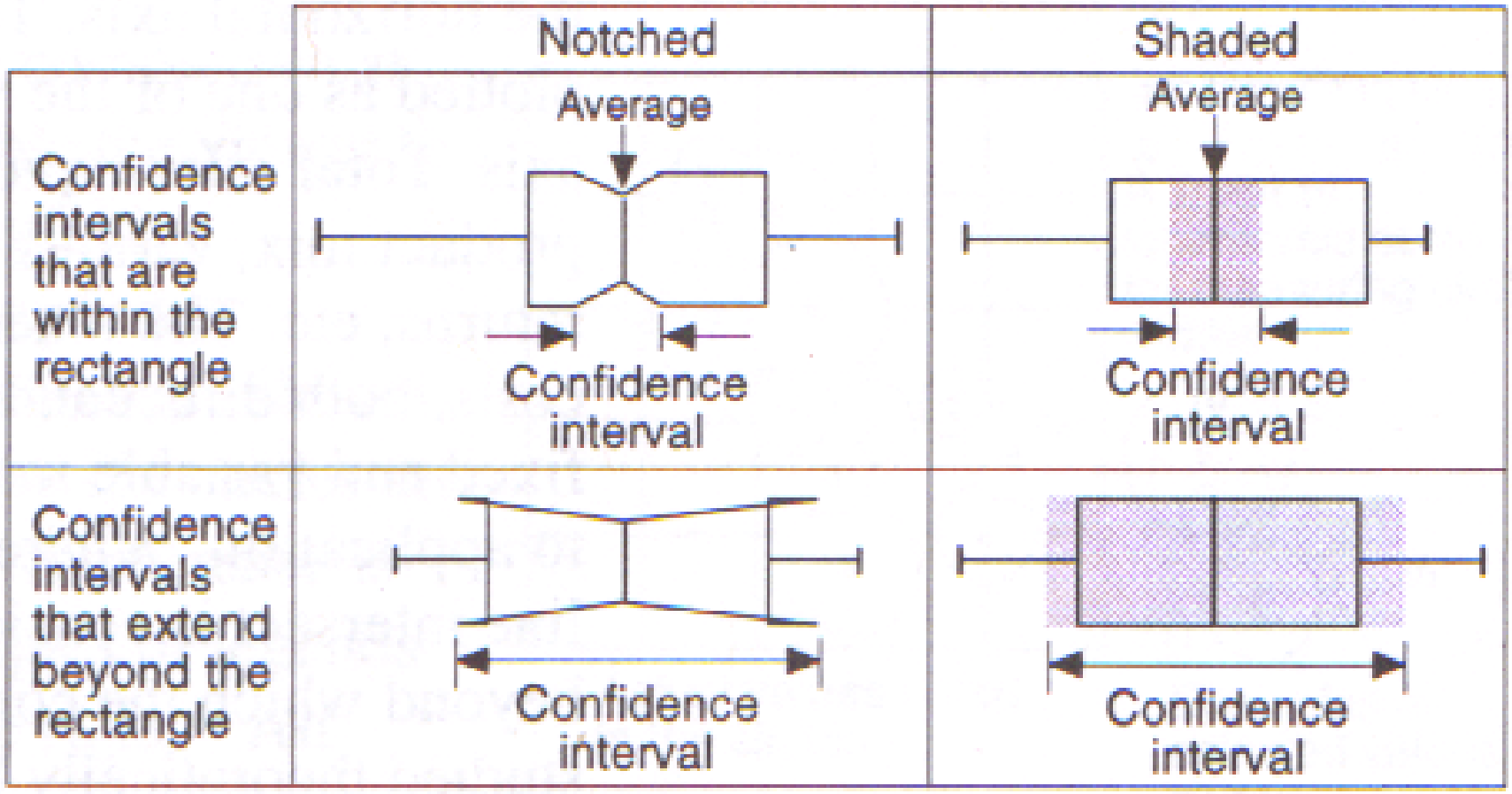
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

A diferença acontece porque no segundo caso a discrepância entre μ e $\bar{x}$ passa a ser relevante e σ^2 subestimaria o valor real da variância.

Boxplots


Original box plot designations
This diagram illustrates how key points were designated on the original box symbols.
A
= C - 3.0 X H
B
= C - 1.5 X H
C
Median
D
E
= D + 1.5 X H
F
= D + 3.0 X H
H = Hspread
Different symbols indicate the ranges in which outlying data points are located.
The smallest data point between the low hinge and a value equal to the low hinge minus 1.5 times the Hspread
Low hinge (equal to the median of the data elements between the lowest value in the data set and the median of the entire data set)
High hinge (equal to the median of the data elements between the highest value in the data set and the median of the entire data set)
The largest data point between the high hinge and a value equal to the high hinge plus 1.5 times the Hspread
Different symbols indicate the ranges in which outlying data points are located


28
26
24
22
20
Quantitative scale
80% to 90% of data points
50% of data points
Individual data points
90th or 95th percentile
Referred to as whisker
75th percentile
- upper quartile (UQ)
Average/mean
50th percentile - median
25th percentile
- lower quartile (LQ)
Referred to as whisker
10th or 5th percentile
Individual data points
Individual box symbol



Examples of notches and shading used to designate confidence intervals in box plots

Função `boxplot` do Statistics Toolbox do MATLAB.