CC-226
Introdução à Análise de Padrões

Prof. Carlos Henrique Q. Forster

Variáveis, Estatísticas e Distribuições de Probabilidades

Tópicos de hoje

- Definições
- Alguns estimadores estatísticos
- Distribuições de probabilidades
- Histogramas

Fenômeno

- **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 + ∞ e -∞).
- Variável **discreta** é aquela cujo espaço amostral é finito.

Estatística

- Inferência Estatística consiste na generalização das informações a respeito de uma amostra, para a sua população.
- A Probabilidade considera modelos para estimar informações sobre instâncias. É um processo de dedução lógica.
- A Estatística considera informações sobre instâncias pra gerar um modelo para toda a população. É um processo de raciocínio indutivo.

Exercício

- Considere um dado de seis lados. Qual a média esperada para jogadas desse dado?

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

- Suponha que joguei o dado 5 vezes e obtive: 2, 3, 3, 6, 1, o que é plenamente possível. Qual foi a média amostral obtida?

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

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 amostras de tamanho n $\{x_1; x_2; \ldots ; x_n\}$ a média amostral é definida como

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

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 e médias aparadas

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

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

- 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.

Revisão de 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.

Distâncias 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.

Descritores de mínima distância

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 e variância amostral

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.

Probabilidade

- Probabilidade caracteriza um fenômeno aleatório e é um modelo para a freqüência que ocorre um evento quando se tende a um número infinito de experimentos, jogadas, amostras.
- Seja A um evento, então:

1. $P(A) \geq 0$
2. Se $A \cap B = \emptyset$, então $P(A \cup B) = P(A) + P(B)$.
3. Seja S o espaço amostral, então $P(S) = 1$.

Outras propriedades

- $P(\emptyset) = 0$.
- $\bar{A}$ é complemento de A, então $P(A) = 1 - P(\bar{A})$.
- E é claro que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Quando $A \cap B = \emptyset$ dizemos que os eventos A e B são mutuamente exclusivos.

Densidade de uma Variável Aleatória

- Definimos a distribuição de probabilidades ou função de densidade de probabilidade (pdf - probability density function) sobre pontos da reta de Borel.
- No caso de variáveis discretas, o valor da função de densidade de probabilidade corresponde à freqüência relativa de que o resultado de um experimento seja igual ao argumento da função.
- $$P(X = 5) = f(5)$$
- No caso de variáveis contínuas, o valor da densidade de probabilidade é tal que a integral da função sobre um intervalo corresponda à freqüência relativa do resultado de um experimento caia dentro do intervalo.
- $$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Distribuição uniforme

- Na distribuição uniforme discreta, cada elemento do espaço amostral é igualmente provável. No caso contínuo, a probabilidade é proporcional ao tamanho do intervalo (desde que dentro do intervalo em que a distribuição é definida). Para um intervalo [A;B] utilizamos a definição:

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B \\ 0, & \text{caso contrário} \end{cases}$$

Distribuição de Bernoulli

Variável aleatória de Bernoulli apresenta como possíveis valores 0 ou 1. Isto é, o espaço amostral é binário $= \{0, 1\}$.

Distribuição de Bernoulli

$$Bern(x; \alpha) = \begin{cases} 1 - \alpha & \text{se } x = 0 \\ \alpha & \text{se } x = 1 \\ 0 & \text{caso contrário} \end{cases}$$

Em geral utilizamos $p = \alpha$ e $q = 1 - \alpha$.

Exemplo: Quantos compradores levam monitores de CRT?

$$P(1) = 0,2$$

$$P(0) = 0,8$$

(soma deve ser 1)

$$P(x) = Bern(x; 0, 2)$$

α é uma parâmetro, isto é, uma quantidade que define a distribuição dentre uma família de distribuições.

Exemplo

Exemplo: (Devore) Quantos nascimentos até nascer um menino?

$$P(B) = p$$
$$P(G) = 1 - p$$

$$p(1) = P(B) = p$$
$$p(2) = P(G) \cdot P(B) = (1-p)p$$
$$p(3) = P(G)P(G)P(B) = (1-p)^2 p$$
$$p(x) = \begin{cases} (1-p)^{x-1}p & x = 1, 2, 3, \ldots \\ 0 & \text{caso contrário} \end{cases}$$

Rodadas de Bernoulli

- Experimentos de Bernoulli (jogadas, rodadas, tentativas)
  - n experimentos chamados tentativas;
  - resultado de cada experimento é sucesso S ou falha F;
  - tentativas são independentes;
  - probabilidade de sucesso (p) é constante de uma tentativa para outra.

Exemplo

Repetimos um experimento binomial de Bernoulli n vezes.
Quantas vezes foi obtido "sucesso", isto é, resposta 1?
Resultados possíveis e igualmente prováveis de 3 tentativas:
SSS SSF SFS SFF FSS FSF FFS FFF

Agrupar por número de sucessos

3	SSS
2	SSF SFS FSS
1	SFF FSF FFS
0	FFF

Distribuição Binomial

- Distribuição binomial é definida por:

$$P(x) = Bin(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n \\ 0, & \text{caso contrário} \end{cases}$$

- Lembrando números binomiais

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}$$

Função de Densidade Acumulada

- A função de densidade acumulada (cdf - cumulative density function) é definida para variáveis discretas como
- $$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$$
- No caso contínuo, a definição é a seguinte:
- $$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(y)dy$$
- Assim, a probabilidade de um intervalo pode ser obtida por:
- $$P(a < X < b) = F(b) - F(a)$$

CDF – Propriedades

- Os casos contínuos e discretos podem ser unificados utilizando ou funções impulso de Dirac ou definição da integral de Lebesgue sobre espaços mensuráveis (incluindo sigma-álgebras).
- $F(-\infty) = 0$
- A $F(+\infty) = 1$ crescente.
- A cdf é diferenciável à direita
- A pdf é a derivada: $f(x) = F'(x)$

Amostras aleatórias sintéticas

- Para fins de simulação, se possuímos um gerador de números pseudo-aleatórios entre 0 e 1 (exclusive) e com distribuição uniforme, podemos utilizar a cdf para obter números aleatórios sorteados de acordo com uma determinada distribuição de probabilidades.
- Se F é a cdf da distribuição de que queremos obter amostras, então a probabilidade de obtermos um valor no intervalo (a; b) é F(b)-F(a). Como F varia de 0 a 1, assim como o nosso gerador de números aleatórios, e é crescente, então se obtivermos um valor sorteado uniformemente entre F(a) e F(b) podemos considerar como um valor no intervalo (a; b) sorteado de acordo com a distribuição almejada.
- Assim, basta sortear uniformemente um valor x entre 0 e 1 e aplicar a inversa da cdf:

$$y = F^{-1}(x)$$

Mediana e quantis populacionais pela CDF

- A mediana de uma distribuição corresponde ao valor que separa 50% da probabilidade, assim:

$$\tilde{x} = F^{-1}(50\%)$$

- Da mesma forma qualquer quantil (quartis ou percentis) podem ser obtidos.

Gráficos

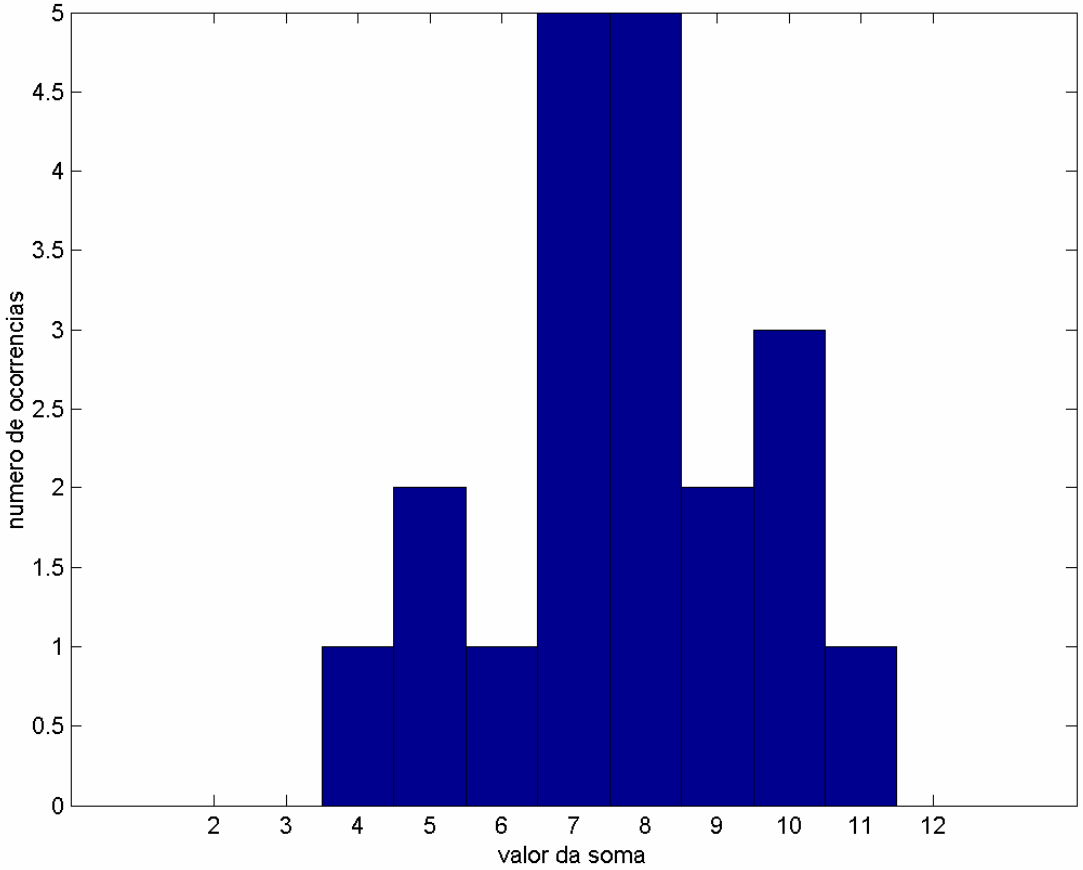
- Há três tipos de gráficos mais importantes para descrever uma ou mais variáveis aleatórias.
  - Histogramas
  - Diagramas de dispersão (scatterplots ou scattergrams)
  - Boxplots

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

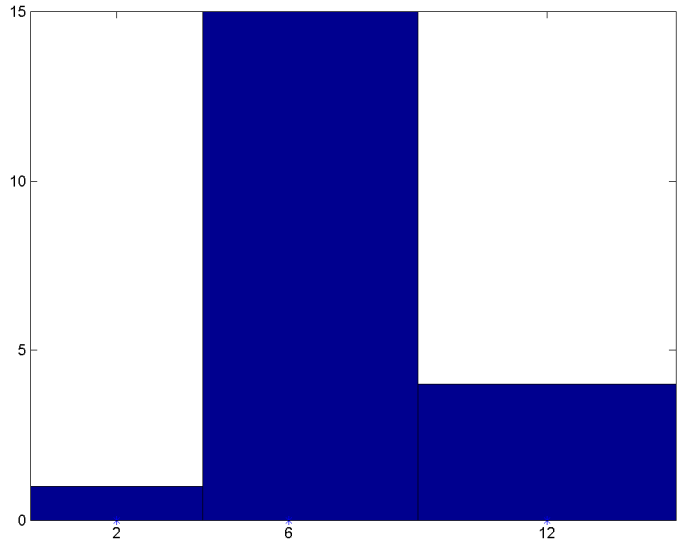
Histograma

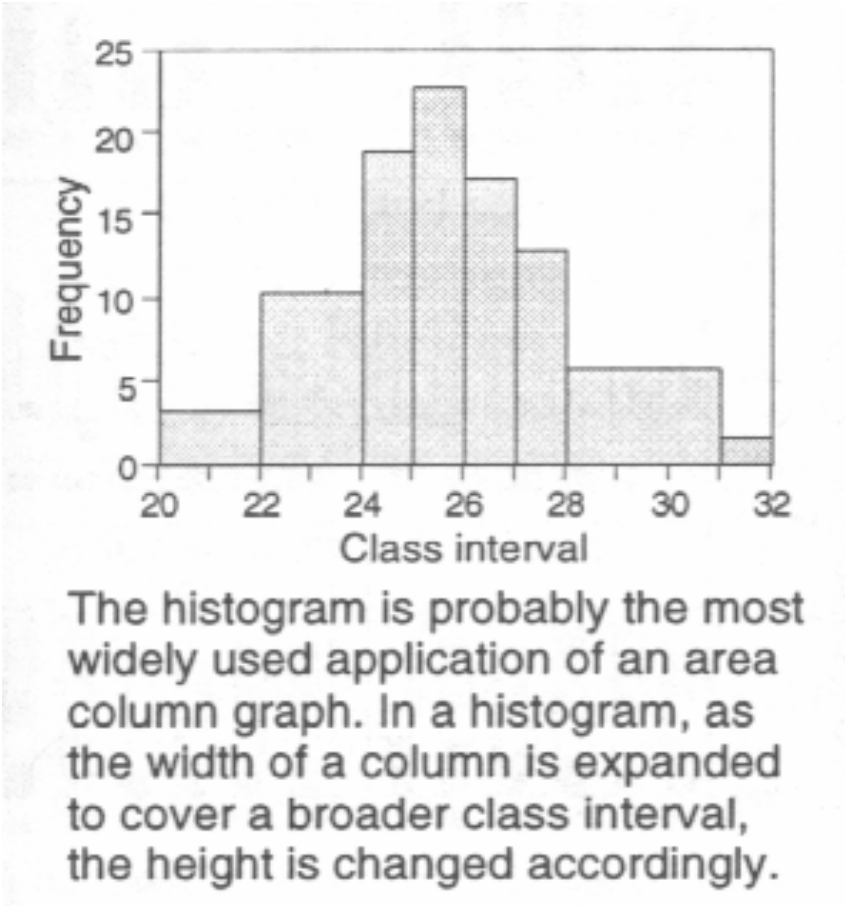
- Para partições de mesmo tamanho, a altura do retângulo é proporcional à freqüência relativa.



Variações do histograma

- Partições não uniformes





Escala de densidade

- 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:

altura do retângulo =

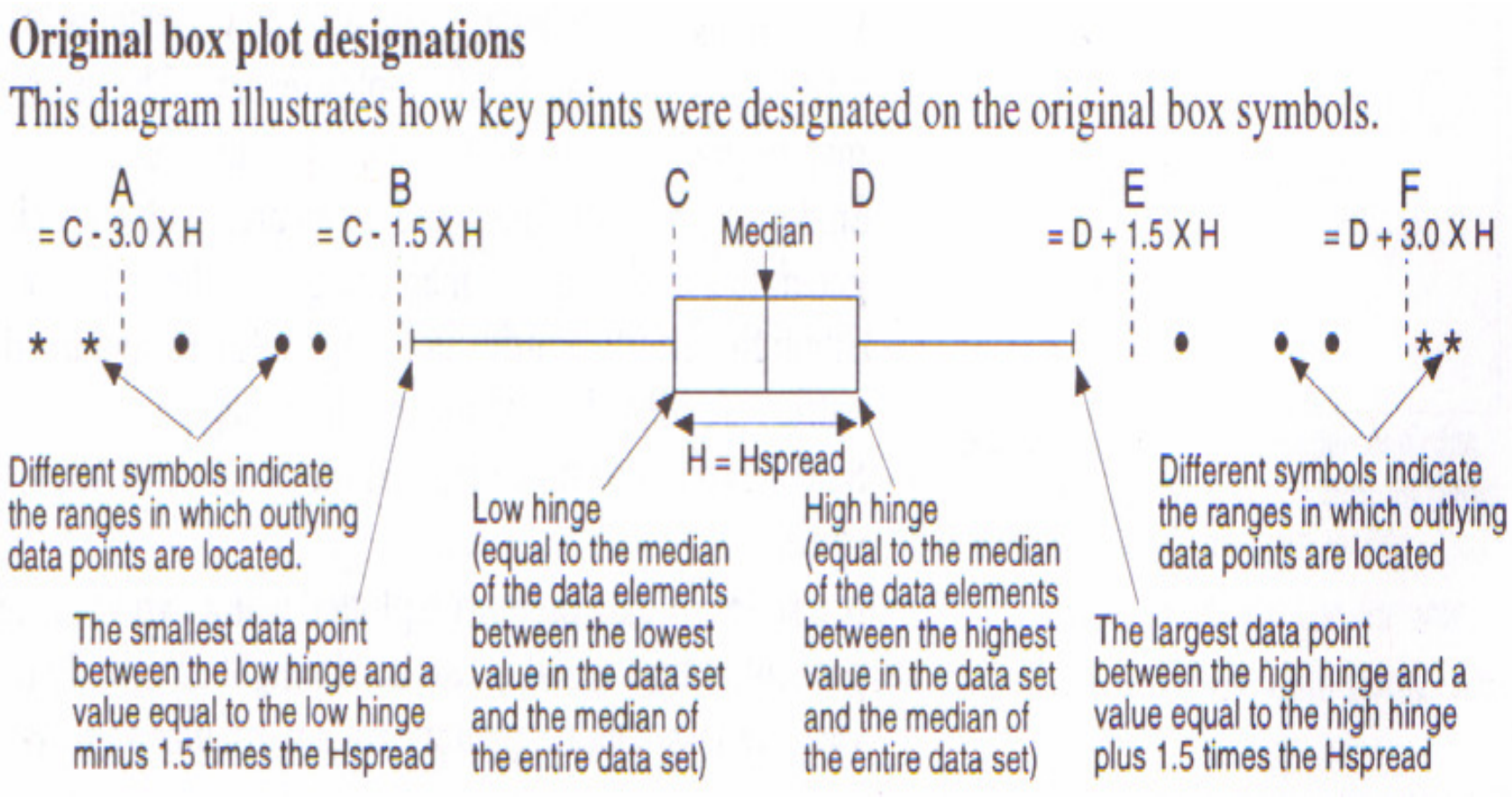
freq relativa da classe / largura da classe

- onde

freq relativa de um valor =

ocorrências do valor / número de observ

Box-plot


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

Esperança

- Esperança é o valor médio esperado de uma variável aleatória.

$$E(x) = \sum_{x \in D} x \cdot p(x)$$

- No caso contínuo:

$$E(x) = \int_{-\infty}^{+\infty} x \cdot p(x) dx$$

Exemplos

Trata-se do somatório da série harmônica que não converge. Dessa forma, a média
não é uma boa medida para caracterizar esse tipo de distribuição.

- Crianças são distribuídas na escala Apgar de 0 a 10.

Apgar	0	1	2	3	4	5	6	7	8	9	10
%	0,002	0,001	0,002	0,005	0,02	0,04	0,18	0,37	0,25	0,12	0,01

$$E(x) = \mu = 0 \cdot 0,002 + 1 \cdot 0,001 + \ldots + 10 \cdot 0,01 = 7,15$$

- X é o número de entrevistas pelas quais um estudante passa antes de conseguir um emprego.

$$p(x) = \begin{cases} \frac{k}{x^2} & x = 1, 2, 3, \ldots \\ 0 & \text{caso contrário} \end{cases}$$

k é tal que $\sum_{x=1}^{\infty} \frac{k}{x^2} = 1$ e não precisa ser calculado (basta ver que é finito).

$$E(x) = \frac{1 \cdot k}{1} + \ldots + \frac{x \cdot k}{x^2} + \ldots = \sum_{x=1}^{\infty} \frac{k}{x} \to \infty$$

Esperança de uma função

- Definição:

$$E[f(x)] = \int_{x \in D} f(x)p(x)dx$$

- Propriedade de operador linear

$$E[aX + b] = aE[X] + b$$

Variância da distribuição

- Seja μ o valor médio esperado dado por μ=E(x)
- A variância é o valor esperado pelo quadrado dos desvios:

$$Var(x) = E[(x-\mu)^2]$$

- Outras fórmulas que podem ser utilizadas para obter a variância:

$$Var(x) = \int_{x \in D} (x-\mu)^2 p(x) dx = E[x^2] - E[x]^2$$

Momentos Estatísticos

- Além da média e variância, é possível definir descritores de ordem mais alta da distribuição.
- O momento de ordem n é definido como a esperança de x^n.
- $m^0 = E(x^0) = E(1) = \int p(x)dx = 1$
- $m^1 = E(x) = \mu$
- $m^2 = E(x^2)$
- $m^3 = E(x^3)$

Momentos Centrais

- A partir dos momentos de ordem 2, podem-se utilizar momentos baseados nos desvios em relação à média. Esses são momentos centrais.

$$\mu_2 = E[(x-\mu)^2] = \sigma^2$$
$$\mu_3 = E[(x-\mu)^3]$$

Obliquidade e Curtose

- Duas medidas importantes para caracterizar uma distribuição não-normal são os coeficientes de skewness e de kurtosis. No caso do skewness, coeficiente próximo de zero significa simetria, caso contrário, uma tendência à esquerda para números negativos e, à direita para números positivos.

$$\text{skewness} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$$

- A kurtosis mede a concentração próxima a média (ou pico). No caso da normalidade, o valor é 3. Menos que 3, a distribuição é mais achatada chamada platykurtic. Maior que 3, o pico é mais acentuado e a distribuição é chamada leptokurtic.

$$\text{kurtosis} = \frac{\mu_4}{\mu_2^2}$$

-

$$\text{kurtosis} = \frac{\mu_4}{\mu_2^2} - 3$$

Desigualdades interessantes sobre momentos

Desigualdade de Chebyshev se aplica a qualquer distribuição.

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

Uma interpretação pode ser obtida para $a = k\sigma$

$$P(|X - \mu| \geq \sigma) \leq \frac{1}{k^2}$$

A probabilidade do valor de X cair numa distância maior ou igual a k desvios-padrão da média é de no máximo $\frac{1}{k^2}$. Isso para qualquer tipo de distribuição. Para 3 sigmas, a probabilidade é menor ou igual a 1/9. Para 6 sigmas, a probabilidade é no máximo 1/36 ou 2,7%.

Desigualdade de Markov se aplica a variáveis não-negativas.

$$P(X \geq a) \leq \frac{\mu}{a}$$

Em ambos os casos, $a > 0$.

Entropia

A entropia é uma medida da aleatoriedade de uma distribuição, definida como

$$H(X) = E\left[\ln \frac{1}{P(X)}\right] = -\int_{x \in D} p(x) \ln p(x) dx$$

Se o logaritmo for na base 2, a unidade de medida é o bit. (Para ln, diz-se que é o nit).

Verificar que $\lim_{x \to 0} xlnx = 0$.

Considere uma variável aleatória de Bernoulli com probabilidade p de sucesso. Pela definição de entropia (vamos utilizar log na base 2),

$$H(p) = -p \lg p - (1-p) \lg(1-p)$$

Pelos limites, temos que

$$H(0) = H(1) = 0$$

Interpretação: total determinismo se 100% de chance de ser 1 ou de ser 0.

Exemplo

Exemplo (Mitzenmacher e Upfal): Entropia de duas moedas viciadas, uma com 3/4 de probabilidade de ser coroa e outra com 7/8.

$$H(\frac{3}{4}) = -\frac{3}{4}\lg\frac{3}{4} - \frac{1}{4}\lg\frac{1}{4} \approx 0,8113$$

$$H(\frac{7}{8}) = -\frac{7}{8}\lg\frac{7}{8} - \frac{1}{8}\lg\frac{1}{8} \approx 0,5436$$

A primeira moeda é aquela que apresenta distribuição com maior entropia. Logo, menos se pode dizer sobre o resultado obtido antes de observá-lo.

Agora, queremos determinar p para que a entropia seja máxima.

$$\frac{\partial H(p)}{\partial p} = -\lg p + \lg(1-p) = \lg\frac{1-p}{p}$$

Assim, $\lg p = \lg(1-p)$ que acontece quando p vale 1/2 e $H(1/2) = 1$ bit.

Exemplo

O lançamento de uma moeda não-viciada tem a aleatoriedade de um bit.
Suponha uma roleta de 8 posições de probabilidade uniforme. Calcular a entropia:

$$H = 8 \times \left(-\frac{1}{8} \lg \frac{1}{8} \right) = 3$$

São necessários 3 bits para codificar o resultado da roleta.

Distribuição de máxima entropia

- Encontrar a distribuição de máxima entropia consiste em determinar a pdf p(x) que maximiza H sob as restrições que regram as pdfs. Assim, procura-se maximizar:

$$H = -\int_D p(x) \ln p(x) dx$$

- Sujeito a:

$$\int_D p(x) dx = 1$$

- Vamos procurar a pdf de máxima entropia, dado que conhecemos a média e a variância. As restrições são:

$$\mu = \int_{-\infty}^{+\infty} x p(x) dx, \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx$$

Continuação

- Formulando com multiplicadores de Lagrange, o novo funcional a minimizar é

$$F = \begin{aligned} & -\int_{-\infty}^{+\infty} p(x) \ln p(x) dx + \\ & \lambda_1 \left(\int_{-\infty}^{+\infty} p(x) dx - 1 \right) + \\ & \lambda_2 \left(\int_{-\infty}^{+\infty} x p(x) dx - \mu \right) + \\ & \lambda_3 \left(\int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx - \sigma^2 \right) \end{aligned}$$

Continuação

- Derivando em função de p e igualando a zero, obtemos que

$$p(x) = e^{-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2}$$

- Substituindo p(x) nas restrições, determinamos os multiplicadores.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A Distribuição Gaussiana (ou Normal)

Para média μ e variância σ^2, a distribuição normal é definida pela expressão:

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Para média zero e variância unitária (e desvio-padrão), definimos a distribuição normal padrão:

$$p(z) = N(z;0,1) = \frac{1}{\sqrt{2\pi}} \exp -\frac{z^2}{2}$$

Padronização da normal

- A função cumulativa de densidade da normal padrão é baseada na função de erro:

$$\Phi(z) = \int_{-\infty}^{z} N(y; 0, 1)dy$$

- Qualquer distribuição normal pode ser padronizada utilizando a transformação linear:

$$Z = \frac{X - \mu}{\sigma}$$

Propriedade dos desvio-padrão da distribuição normal

- A probabilidade de uma amostra ser obtida dentro de 1 desvio-padrão da média é dada por:

$$\Phi(1) - \Phi(-1)$$

- Vamos tabelar para alguns desvios-padrão de distância

k	dentro de $k\sigma$	fora	Chebyshev $1/k^2$
1	0,6826	0,3173	1
2	0,9545	0,0455	0,25
3	0,9973	0,0027	0,1111
6	0,9999	0,2e-8	0,0278

Esperança e variância da binomial

- A esperança e a variância de uma distribuição binomial são dadas por:
- $E(x) = n\ p$
- $Var(x) = n\ p\ (1-p)$
- A distribuição binomial pode aproximar uma normal com média np e variância np(1-p)