ESTATÍSTICAS DESCRITIVAS PARA DADOS SIMBÓLICOS

Renata Maria C. R. de Souza

Centro de Informática U·F·P·E

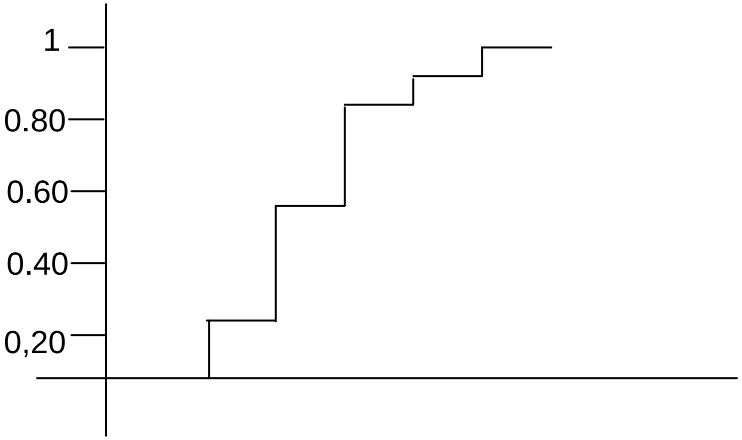
Introdução

Média

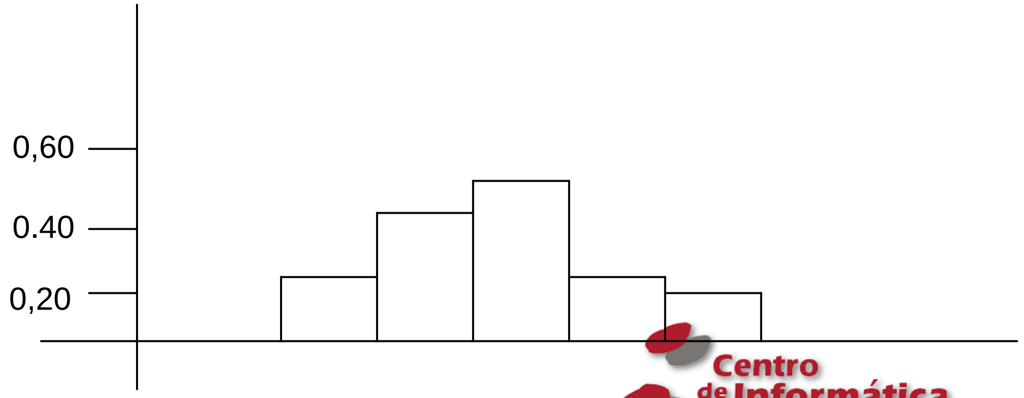
Variância

Distribuição de Freqüências

Função de Distribuição Acumulada



Histograma



Estatísticas para dados clássicos

E - {1,2,...,n} conjunto de n elementos

Y - uma variável descritora

O - conjunto dos possíveis valores de Y

y = (Y(1),...,Y(n))' vetor de valores de Y para n elementos

$\{V_1, V_2, ..., V_t\}$ conjunto desses valores distintos de Y onde

$$V_1 \leq V_2 \leq _{...} \leq V_t$$

$freq_Y(V_i)$ = número de exemplos do valor V_i

$i \in \{1,....,t\}$

Distribuição de freqüências - todos os pares $(V_j, freq(V_j))$

Função de distribuição - acumulada $F_Y(V) = \frac{1}{n} \sum_{V_j \le V} freq(V_j)$

Média - $\bar{y} = \frac{1}{n} \sum_{j=1}^{t} \mathrm{Freq}(V_j) \times V_j$

Variância - $s_Y^2 = \frac{1}{n} \sum_{j=1}^{t} freq(V_j)(V_j - \bar{y})^2$

Histograma - Considere m subintervalos $I_1,...,I_m$ de $[V_1,V_t]$ e

$$m_h = |k \in E / Y(k) \in I_h| \quad \text{ou} \quad m_h = \frac{|k \in E / Y(k) \in I_h|}{n}$$

Representação gráfica dos pares (I_h, m_h)

Estatísticas para dados simbólicos

E - {1,2,...,n} conjunto de n unidades

$Y=\{Y_1,\ldots, Y_p\}$ – conjunto de variáveis simbólicas (multivalorada, modal ou intervalo) sendo O_j- conjunto dos possíveis valores de Y_j (domínio de Y_j)

$O =\{O \times \ldots \times O_p\}$

$\mathbf{d}_i =(\xi_{i1},\ldots, \xi_{ip})$ – descrição simbólica

Tabela de dados simbólicos

E	Y_1	...	Y_j	...	Y_p
1					
i	ξ_{i1}		ξ_{ij}		ξ_{ip}
n					

Centro de Informática U·F·P·E

Estatísticas para Dados Simbólicos

- Envolve
  - Descrições individuais
  - Dependências lógicas nos dados
  - Extensão virtual

Vetor de Descrição Individual

Seja

d = $(D_1,\ldots, D_p)$ um vetor de descrição e D=($D_1 \times \ldots \times D_p$) o conjunto de descrição cartesiano associado a **d**

x é chamado de **vetor de descrição individual** se $D_j \subset O_j$ é um único valor ou seja

x $= (\{x_1\},\ldots,\{x_p\}) = (x_1,\ldots,x_p)$

Exemplos:
Multivalorada:Y=cor de pássaro
ξ ={vermelho, azul, verde} (descrição simbólica associada a Y)
Descrições individuais associadas a ξ : x=vermelho, x=azul, x=verde

Intervalo: Y=altura ξ=[65,70] x=todos os pontos no intervalo [65,70]

Dependências Lógicas nos Dados

A dependência lógica pode ser representada pela regra

$$r: \text{ Se } \boldsymbol{x} \in A \text{ então } \mathbf{x} \in B$$

onde $A \subseteq D$ e $B \subseteq D$ são dois conjuntos de descrições e r uma função de O em [0,1].

Um vetor de descrição individual **x** satisfaz a regra r se e somente se $\mathbf{x} \in A \cap B$ ou $\mathbf{x} \notin A$. A notação r(**x**) =1 indicará que **x** satisfaz a regra e r(**x**) =0 caso contrário

Exemplo. Considere $\mathbf{d}_1$=({10,26},{1})) um descrição simbólica de Y_1=idade e Y_2= número de filhos. Suponha que existe a regra r: If A={$Y_1 \leq 12$} então B={Y_2=0}.

Temos dois vetores de descrições individuais:
$\mathbf{x}_1$=(10,1) e $\mathbf{x}_2$=(26,1) . Então $r(\mathbf{x}_1)$=0 e $r(\mathbf{x}_2)$=1

Uso de Dependências Lógicas nos Dados

- Tratamento com grandes conjuntos de dados agregados pode produzir unidades que tem valores de descrições individuais não praticáveis.

	Indivíduo	Idade	No de Filhos
Clássico	1	10	0
	2	26	1
	3	10	0
	4	10	0
	5	26	1
Simbólico	w	[10,26]	{0,1}

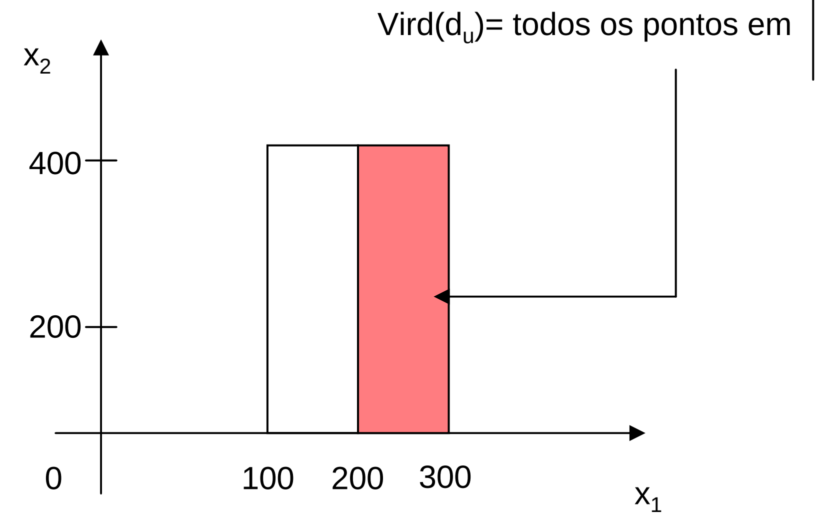
Centro de Informática U·F·P·E

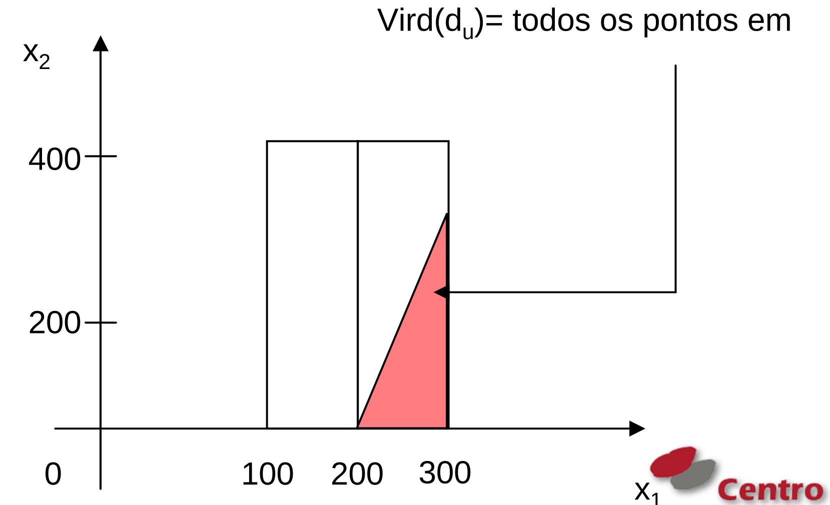
Exemplos

- Considere um par de variáveis simbólicas de tipo intervalo (Y_1,Y_2) com Y_1= custo de medicação e Y_2= cobertura do seguro e a regra
- Regra r_1:
- $A_1=\{Y_1=\text{Custo} < \$200\}$ e $B_1=\{Y_2=\$0\}$.
- Suponha
- $d_u=Y(u)=\xi_u=([100,300],[0,400])$

- Regra r_2:
- $A_2=\{Y_1=\text{Custo}\}$ e $B_2=\{Y_2 \le Y_1\}$.

Considerando as duas regras





Extensão Virtual de um vetor de descrição

Considere **d=**$(D_1,...,D_p)$, $D=(D_1\times...\times D_p)$ e R um conjunto de regras de dependências lógicas.

A extensão virtual do vetor de descrição d vir(d) é o conjunto de vetores de descrição individual que satisfaz todas as regras em R ou seja

$$vir(d) := \{ \mathbf{x} \in D / \; r(\mathbf{x}) = 1 \} \;\; p/ \text{ toda } r \in R$$

Exemplo:

E	Y_1	Y_2
1	{a,b}	{3}
2	{b,c}	{2,3}
3	{b}	{1,2}
4	{c}	{1,3}
5	{a}	{1,3}

O_1 ={a,b,c}

O_2 ={1,2,3}

r: Se A={$y_1 \in$ {b,c}} então B= {$y_2 \in$ {2}}

A={(b,1),(b,2),(b,3), (c,1),(c,2),(c,3)}
B={(a,2),(b,2),(c,2)}

r(**x**)=1 se x $\in A \cap B$ ou x $\notin$ A

d_1	d_2	d_3	d_4	d_5
{a},{3} {b},{3}	{b},{2} {b},{3} {c},{2} {c},{3}	{b},{1} {b},{2}	{c},{1} {c},{3}	{a},{1} {a},{3}

Extensão Virtual

$Vir(d_1) = \{\mathbf{x} \in \{a,b\} \times \{3\} \,/\, r(x) = 1\} = \{(a,3)'\}$

$Vir(d_2) = \{(b,2)', (c,2)'\}$

$Vir(d_3) = \{(b,2)'\}$

$Vir(d_4) = \varnothing$

$Vir(d_5) = \{(a,1)', (a,3)'\}$

Centro de Informática U·F·P·E

Distribuição de Freqüências para Variáveis Multivaloradas

Seja Y_j = Z uma variável simbólica multivalorada que toma valores $\xi \in O_z$. Esses valores podem ser categóricos ou discretos.

Para um conjunto de observações multivaloradas $\{\xi_1, \ldots, \xi_n\}$, a **freqüência observada** de Z é dada por

$$Freq_z(\xi) = \sum_{k \in E} w_z(\xi : k)$$

onde

$$w_z(\xi ; k) = \frac{\left| \boldsymbol{x} \in vir(d_k) / x_{[z]} = \xi \right|}{\left| vird(d_k) \right|}$$

Freq(ξ) pode não ser um valor inteiro como no caso clássico.

$x[_z]$ valor individual da variável simbólica Z

Distribuição de Freqüências para Variáveis Multivaloradas

- A **distribuição de frequencia empírica** é o conjunto de pares $\{\xi, Freq_z(\xi)\}$ para todo $\xi \in O_z$.
- A **distribuição de frequencia relativa** ou o histograma é o conjunto para todo $\xi \in O_z$

$$\{\xi, Freq_z(\xi) = \frac{1}{m}\sum_{k \in E} w_z(\xi : k)\}$$

- $m=n-n_0$ onde n_0 é a quantidade de k cujo $|vird(d_k)=0|$

Distribuição de Freqüências para Variáveis Multivaloradas

- A **Função de distribuição empírica** é dada por

$$F_z(\xi) = \frac{1}{m} \sum_{\xi_k \leq \xi} Freq_z(\xi_k)$$

- A **Média e Variância amostral simbólica** são

$$\bar{z} = \frac{1}{m} \sum_{\xi_k \in O_z} \xi_k Freq(\xi_k) \qquad S_z^2 = \frac{1}{m} \sum_{\xi_k \in O_z} (\xi_k - \bar{z})^2 Freq(\xi_k)$$

- A **Mediana amostral simbólica** é dada por

$$\tilde{z} = \xi \,/\, F_z(\xi) = 1/2$$

Centro de Informática U·F·P·E

Exemplo 1: Com dependência lógica

$$Freq_z(\xi) = \begin{cases} 0+0+0+1/2=0.5 & \text{se } \xi = 1 \\ 2 & \text{se } \xi = 2 \\ 1.5 & \text{se } \xi = 3 \\ 0 & \text{outro caso} \end{cases}$$

$$\bar{z} = \frac{1}{4}\left(1\times 0.5 + 2\times 2 + 3\times 1.5\right) = 2{,}125$$

$$S_z^2 = \frac{1}{4}\ (0.5\ (1 - 2.125)^2 + 2(2-2.125)^2 + 1.5(3 - 2.125)^2) = 1{,}81$$

$$\tilde{z} = \xi / F_z(\xi) = 0{,}50 = 2$$

Centro de Informática U·F·P·E

Exemplo 1: Com dependência lógica

Função de distribuição
Acumulada

$$F_Z(\xi) = \begin{cases} 0 & \text{se } \xi < 1 \\ 0{,}125 & \text{se } 1 \le \xi < 2 \\ 0{,}625 & \text{se } 2 \le \xi < 3 \\ 1 & \text{outro caso} \end{cases}$$

Exemplo 2: Sem dependência lógica

E	Y_1	Y_2
1	{a,b}	{3}
2	{b,c}	{2,3}
3	{b}	{1,2}
4	{c}	{1,3}
5	{a}	{1,3}

d_1	d_2	d_3	d_4	d_5
{a},{3} {b},{3}	{b},{2} {b},{3} {c},{2} {c},{3}	{b},{1} {b},{2}	{c},{1} {c},{3}	{a},{1} {a},{3}

Exemplo 1: Sem dependência lógica

$$\text{Freq}_Z(\xi) = \begin{cases} 0+0+0.5+0.5+0.5=1.5 & \text{se } \xi = 1 \\ 1 & \text{se } \xi = 2 \\ 2.5 & \text{se } \xi = 3 \\ 0 & \text{outro caso} \end{cases}$$

$$\bar{z} = \frac{1}{4}(1 \times 1.5 + 2 \times 1 + 3 \times 2{,}5) = 3{,}25$$

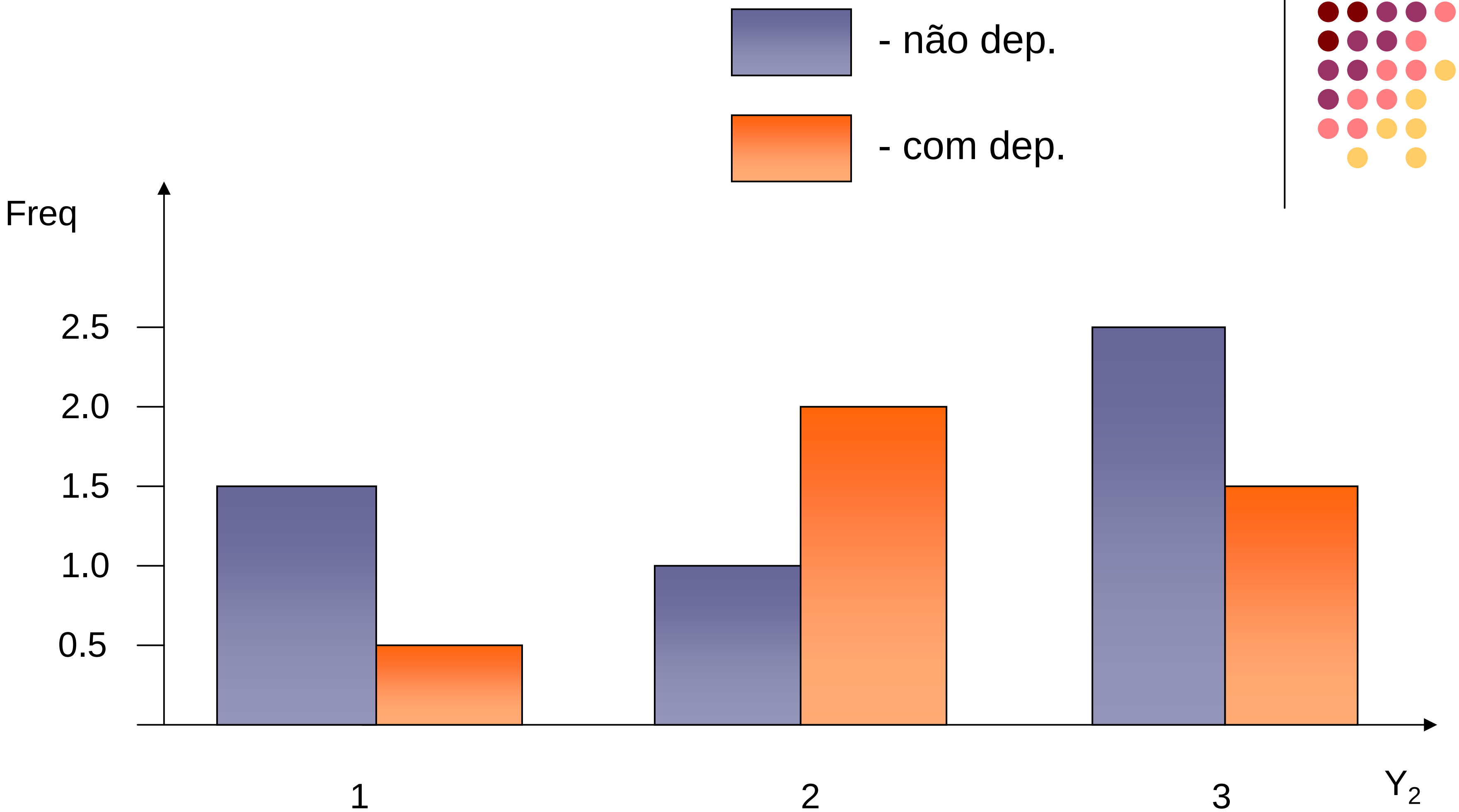
$$S_z^2 = \ \ 2{,}71$$

$$\tilde{z} = \xi \,/\, F_z(\xi) = 0{,}5 = 2$$

Exemplo 1: Sem dependência lógica

Função de distribuição Acumulada

$$F_Z(\xi) = \begin{cases} 0 & \text{se } \xi < 1 \\ 0.375 & \text{se } 1 \leq \xi < 2 \\ 0.875 & \text{se } 2 \leq \xi < 3 \\ 1 & \text{outro caso} \end{cases}$$


- não dep.
- com dep.
Freq
2.5
2.0
1.5
1.0
0.5
1
2
3
Y_2

Distribuição de Freqüência para Variáveis Intervalo

Seja Z uma variável intervalo. Para cada $u \in E$, $Z(u) = [a_u, b_u]$.

Hipóteses

1 - Cada $u \in E$ é selecionado com a mesma probabilidade 1/n

2 - Os valores $x_{[z]}$ para $\underline{\mathbf{x}} \in vir(d_u)$ são uniformemente distribuído no intervalo $Z(u) = [a_u, b_u]$.

Z é considerada uma variável aleatória cuja distribuição é uma mistura de n distribuições uniformes

$$\Pr(x_{[z]} \le \xi / \boldsymbol{x} \in vir(d_u)) = \begin{cases} 0 & se \;\; \xi < a_u \\ \dfrac{\xi - a_u}{b_u - a_u} & se \;\; a_u \le \xi < b_u \\ 1 & se \;\; b_u \le \xi \end{cases}$$

A Função de Distribuição Empírica é dada por

$$F_z(\xi) = \frac{1}{n} \sum_{u \in E} \Pr(x_{[z]} \le \xi / \boldsymbol{x} \in vir(d_u))$$

$$F_z(\xi) = \frac{1}{n} \sum_{u:\xi \in Z(u)} \frac{\xi - a_u}{b_u - a_u} + \frac{\left|\{u / \xi \ge b_u\}\right|}{n}$$

A Função de Densidade Empírica é dada por

$$f_z(\xi) = \frac{1}{n} \sum_{u:\xi \in Z(u)} \frac{1}{b_u - a_u} = \frac{1}{n} \sum_{u \in E} \frac{1l_{Z(u)}(\xi)}{L(Z(u))} \quad \xi \in lR$$

onde $1l_{Z(k)}(\xi)$ é uma função indicadora e L a largura do intervalo Z(u)

A média empírica é dada por

$$\bar{z} = \int_{-\infty}^{+\infty} \xi \; f_z(\xi) \; d_\xi = \frac{1}{n} \sum_{u \in E} \int_{\underline{Z}_u}^{\overline{Z}_u} \frac{1l_{Z(u)}(\xi)}{L(Z(u))} \; \xi \, d_\xi = \frac{1}{n} \sum_{u \in E} \frac{1}{b_u - a_u} \int_{a_u}^{b_u} \xi \, d_\xi$$

$$\bar{z} = \frac{1}{2n} \sum_{u \in E} \frac{b_u^{\,2} - a_u^{\,2}}{b_u - a_u} = \frac{1}{n} (b_u + a_u)/2$$

O desvio padrão empírico é dada por

$$s_z = \sqrt{\int_{-\infty}^{+\infty} \xi^2 \ f_Z(\xi) \ d_\xi - \bar{z}^2}$$

Segundo momento

$$M_2 = \int_{-\infty}^{+\infty} f_Z(\xi) \ d_\xi = \frac{1}{n} \sum_{u \in E} \int_{a_u}^{b_u} \frac{\xi^2}{L(Z(u))} d_\xi$$

$$M_2 = \frac{1}{n} \sum_{u \in E} \frac{b_u^{\ 3} - a_u^{\ 3}}{3 \times L(Z(u))} = \frac{1}{3n} \sum_{u \in E} (b_u^{\ 2} + b_u a_u + a_u^{\ 2})$$

$$s_z = \sqrt{\frac{1}{3n} \sum_{u \in E} (b_u^{\ 2} + b_u a_u + a_u^{\ 2}) - \frac{1}{4n^2} \left[\sum_{u \in E} (a_u + b_u) \right]^2}$$

Exemplo

E = {1,...,8}

Z (E) = {[0,2];[1,3];[1.5,2.5];[2,4];[3.5,5];[4.5,5.5];[5,7];[6.5,7.5]}

I=[0,7.5]

u	1	2	3	4	5	6	7	8
$\frac{a_u + b_u}{2}$	2	2	3	4.25	4.75	0.50	6	0.50

Média = 3.781

Desvio Padrão = 2.045

Histogramas sem Dependências Lógicas

Considere uma partição de I = [min $\underline{z}_u$ / k $\in$ E}, max {z_u k $\in$ E}] em r subintervalos $I_j = [u_{g-1}, u_g]$ para g $\in$ {1,...,r} que estão em ordem crescente.

A freqüência observada é dada por

$$f_g = \sum_{u \in E} \frac{L(Z(u) \cap I_g)}{L(Z(u))}$$

A freqüência relativa é dada por $p_g = \frac{1}{n} f_g$

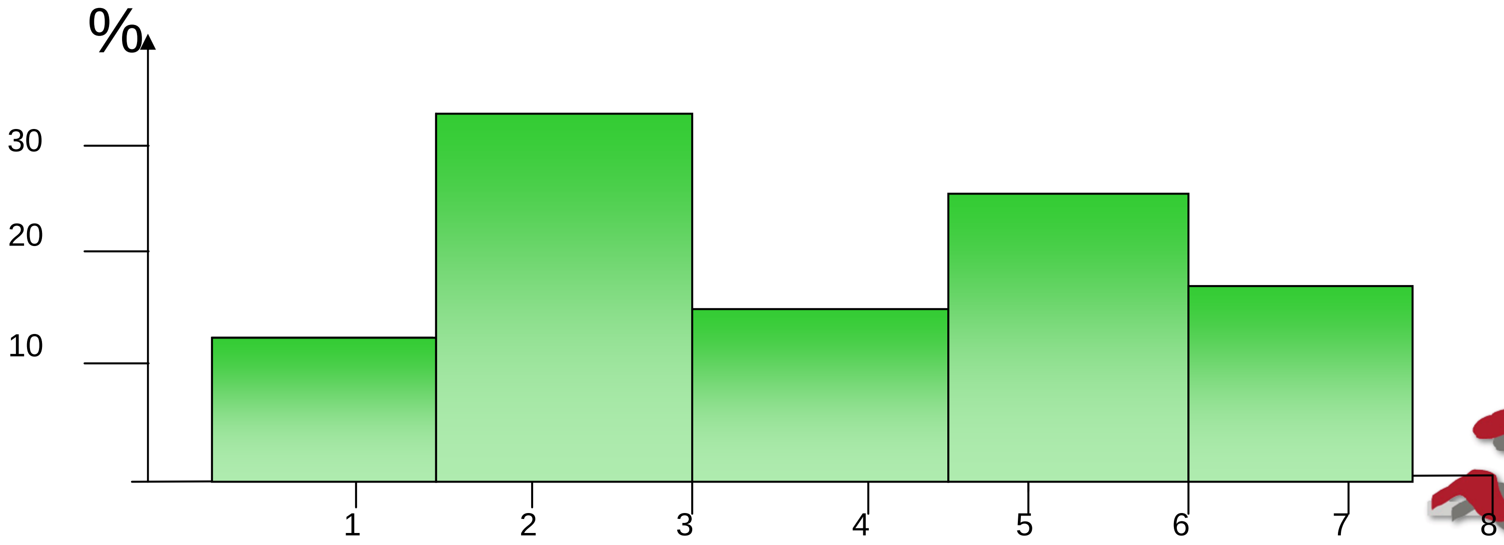
O histograma é a representação gráfica {I_g,f_g } para g=1,...,r

Centro de Informática U·F·P·E

$$p_j = \frac{1}{n}\sum_{u\in E}\frac{L(Z(u)\cap I_j)}{L(Z(u))}$$

Distribuição de Frequências

j	1	2	3	4	5
I	[0,1.5]	[1.5,3.0]	[3.0,4.5]	[4.5,6.0]	[6.0,7.5]
p	0.125	0.312	0.145	0.229	0.187



Histogramas com Dependências Lógicas

u	pulso	Pressão sistólica	Pressão diastólica
u_1	[44,68]	[90,110]	[50,70]
u_2	[60,72]	[90,130]	[70,90]
u3	[56,90]	[140,180]	[90,100]
u_4	[70,112]	[110,142]	[80,108]
u_5	[54,72]	[90,100]	[50,70]
u_6	[70,100]	[134,142]	[80,110]
u_7	[72,100]	[130,160]	[76,90]
u_8	[76,98]	[110,190]	[70,110]
u_9	[86,96]	[138,180]	[90,110]
u_{10}	[86,100]	[110,150]	[78,100]
u_{11}	[53,55]	[160,190]	[205,219]
u_{12}	[50,55]	[180,200]	[110,125]
u_{13}	[73,81]	[125,138]	[78,99]
u_{14}	[60,75]	[175,194]	[90,100]
u_{15}	[42,52]	[105,115]	[70,82]

Centro de Informática U·F·P·E

Histogramas com Dependências Lógicas

- r_1 Se A = {$Y_1 \leq 55$} então B= {$Y_2 \leq 170$, $Y_3 \leq 100$}
- O conjunto A inclue os elementos u=1,5,11,12,15.
- Os elementos u=11,12 não são válidos na região de B
- Então os elementos u=11,12 são omitidos do conjunto.
- Considerando os intervalos I_{jj} (j=1,5) I_1=[80,100), I_2=[100,120)], I_3=[120,140)], I_4=[140,160)], I_5=[160,180)], I_6=[180,200)] para variável presssão sistólica temos que

$$f_5 = 0 + 0 + \left(\frac{180-160}{180-140}\right)_{u=3} + 0 + \ldots + 0 + \left(\frac{180-160}{180-138}\right)_{u=9} + 0_{u=10} + 0_{u=13} + \left(\frac{180-175}{194-175}\right) = 1{,}489$$

$$p_5 = 1{,}489/13 = 0{,}114$$

$\bar{z} = 133{,}769$ Média

$S_Z^2 = 730$ Desvio Padrão

Centro de Informática U·F·P·E

Histogramas com Dependências Lógicas

- r_2 Se A $= \{Y_1 < 55\}$ então B$= \{Y_2 \leq 170\}$
- O conjunto A inclue os elementos u=1,5,11,12,15.
- O elemento u=12 não é válido na região de B e é omitido do conjunto.
- A extensão virtual da descrição de u=11 são todos os pontos em $[53,55] \times [160,170]$
- Considerando os intervalos I_{jj} (j=1,5) $I_1=[80,100)$, $I_2=[100,120)]$, $I_3=[120,140)]$, $I_4=[140,160)]$, $I_5=[160,180)]$, $I_6=[180,200)]$ para variável presssão sistólica temos que

$\bar{z} = 136,$ Média

$S_Z^2 = 743{,}88$ Desvio Padrão

$$f_5 = 0 + 0 + \left(\frac{180-160}{180-140}\right)_{u=3} + 0 + \ldots + 0 + \left(\frac{180-160}{190-110}\right)_{u=8} +$$

$$\left(\frac{180-160}{180-138}\right)_{u=9} + 0_{u=10} + \left(\frac{170-160}{170-160}\right)_{u=11} + 0 + \left(\frac{180-160}{194-175}\right) + 0_{u=14}$$

$$= 2{,}489$$

Histogramas com Dependências Lógicas

- r_3 Se A $= \{Y_1 \le 70\}$ então B$= \{Y_2 \le 170\}$
- O elemento u=12 não é logicamente verdadeiro e é omitido do conjunto.
- Partes da descrição de u=3,11,14 não são válidas.
- A descrição virtual de u=3 é:
- R=[56,90]×[140,170] ∪ [70,90]×[170,180]

$$\frac{(90-56)\times(170-140)}{(90-56)\times(170-140)+(90-70)\times(180-70)}$$

$$f_5 = 0+0+\left\{\left(\frac{170-160}{170-140}\right)(0.836)_{R_1} + \left(\frac{180-170}{180-170}\right)(0.164)_{R_2}\right\}_{u=3} + 0 + \ldots + \left(\frac{180-160}{190-110}\right)_{u=8} +$$

$$\left(\frac{180-160}{180-138}\right)_{u=9} + 0_{u=10} + \left(\frac{180-160}{190-160}\right)_{u=11} + 0 + 0 + \left(\frac{180-160}{194-175}\right)_{u=14} + 0$$

$$= 2{,}432$$

$$\bar{z} = 135{,}87$$ Média

$$S_Z^2 = 737{,}05$$ Desvio Padrão

Distribuição de Frequencias para Variáveis Modais

- Y_j uma variável modal (multivalorada modal)
- ξ_{jk} - categorias ou valores de Y_j com frequencias relativas p_{jk} (k=1,..,s) tal que

$$\sum p_{jk} = 1$$

- Assumindo Z= Y_j a frequencia observada de Z é dada por:

$$O_Z(\xi_k) = \sum_{u \in E} \pi_z(\xi_k : u)$$

$$\pi_z(\xi_k : u) = P(Z = \xi_k / x \in virt(d_u), u)$$

$$P(Z = \xi_k / x \in virt(d_u), u) = \frac{\sum_x P(x = \xi_k / x \in virt(d_u), u)}{\sum_x \sum_{k=1}^{s} P(x = \xi_k / x \in virt(d_u), u)}$$

Distribuição de Frequencias para Variáveis Modais

u	Tipo de Combustível					Indicadora	
	{Nada	Gasolina	Óleo	Eletricidade	Carvão}	{Não	Sim}
u_1	{0,0	0,87	0,07	0,05	0,01}	{0,09	0,91}
u_2	{0,0	0,71	0,11	0,10	0,08}	{0,12	0,88}
u_3	{0,0	0,83	0,08	0,09	0,00}	{0,23	0,77}
u_4	{0,0	0,76	0,06	0,11	0,07}	{0,19	0,81}
u_5	{0,0	0,78	0,06	0,09	0,07}	{0,12	0,88}
u_6	{0,0	0,90	0,01	0,08	0,01}	{0,22	0,78}
u_7	{0,0	0,87	0,01	0,10	0,02}	{0,22	0,78}
u_8	{0,0	0,78	0,02	0,13	0,07}	{0,13	0,87}
u_9	{0,0	0,91	0,00	0,09	0,00}	{0,24	0,76}
u_{10}	{0,25	0,00	0,41	0,14	0,20}	{0,09	0,91}

Centro de Informática U·F·P·E

Distribuição de Frequencias para Variáveis Modais

Frequencias para variável Y_1

Frequencias para variável Y_2

$$O_{Y_1}(Nada) = 0{,}25$$

$$O_{Y_2}(Não) = 1{,}65$$

$$O_{Y_1}(Gas.) = (0{,}87 + ... + 0{,}91 + 0{,}00) = 7{,}41$$

$$O_{Y_2}(Sim) = 8{,}35$$

$$O_{Y_1}(Óleo) = 0{,}83$$

$$\sum_{k=1}^{10} O_Y(\xi_k) = n$$

$$O_{Y_1}(Eletric.) = 0{,}98$$

$$O_{Y_1}(Carvão) = 0{,}53$$

Distribuição de Frequencias para Variáveis Modais

Frequencias Relativas para variável Y_1

Frequencias Relativas para variável Y_2

$$P_{Y_1}(Nada) = 0{,}25/10 = 0{,}025$$

$$P_{Y_1}(Gas.) = 0{,}741$$

$$P_{Y_1}(Óleo) = 0{,}083$$

$$P_{Y_1}(Eletric.) = 0{,}098$$

$$P_{Y_1}(Carvão) = 0{,}053$$

$$P_{Y_2}(Não) = 0{,}165$$

$$P_{Y_2}(Sim) = 0{,}835$$

$$\sum_{k=1}^{10} P_Y(\xi_k) = 1$$

Distribuição de Frequencias para Variáveis Modais

- Considere a regra

$$r = \begin{cases} r_1 : Y_2 \in \{Não\} \Rightarrow Y_1 \in \{Nada\} \\ r_2 : Y_2 \in \{Sim\} \Rightarrow Y_1 \in \{Gas, Óleo, Eletric.Carvão\} \end{cases}$$

- Descrições individuais válidas

$$C = \{(\xi_0, Não); (\xi_1, Sim); (\xi_2, Sim); (\xi_3, Sim); (\xi_3, Sim)\}$$

- Assumindo que Y_1 e Y_2 são independents, será obtido a probabilidade conjunta de (Y_1, Y_2)

Distribuição de Frequencias para Variáveis Modais

- Considere u=10 e a tabela de probabilidades conjunta

Y_1, Y_2	(Nada, Não)	(Gas., Sim)	(Óleo, Sim)	(Eletric ,Sim)	(Carvão, Sim)	TOTAL
u_{10}	0,0225	0,00	0,3731	0,1274	0,1820	0,7050

$0{,}25 \times 0{,}09$

Probabilidades Padronizadas

Y_1, Y_2	(Nada, Não)	(Gas., Sim)	(Óleo, Sim)	(Eletric ,Sim)	(Carvão, Sim)
u_{10}	0,0319	0,00	0,5292	0,1807	0,2582

Centro de Informática U·F·P·E

Distribuição de Frequencias para Variáveis Modais

Y_1 , Y_2	(Nada, Não)	(Gas., Sim)	(Óleo, Sim)	(Eletric ,Sim)	(Carvão, Sim)
u_1	0,0900	0,7917	0,0637	0,0455	0,0091
u_2	0,1200	0,6248	0,0968	0,0880	0,0704
u_3	0,2300	0,6391	0,0616	0,0693	0,0000
u_4	0,1900	0,6156	0,0486	0,0891	0,0567
u_5	0,1200	0,6864	0,0528	0,0792	0,0616
u_6	0,2200	0,7020	0,0078	0,0624	0,0078
u_7	0,2200	0,6786	0,0078	0,0708	0,0156
u_8	0.1300	0,6786	0,0174	0,1131	0,0609
u_9	0,2400	0,6916	0,0000	0,0684	0,0000
u_{10}	0,0319	0,00	0,5292	0,1807	0,2582

Centro de Informática U·F·P·E

Distribuição de Frequencias para Variáveis Modais

Frequencias para variável Y_1

Frequencias para variável Y_2

$$O_{Y_1}(Nada) = 1{,}5919$$

$$O_{Y_1}(Gas.) = 6{,}1084$$

$$O_{Y_1}(Óleo) = 0{,}0637 + \ldots + 0{,}000 + 0{,}5292 = 0{,}5403$$

$$O_{Y_1}(Eletric.) = 0{,}8737$$

$$O_{Y_1}(Carvão) = 0{,}403$$

$$O_{Y_2}(Não) = 1{,}6$$

$$O_{Y_2}(Sim) = 8{,}4$$