Técnicas Computacionais em Probabilidade e Estatística I

Aula III

Chang Chiann
MAE 5704- IME/USP
1º Sem/2008

Arquivo *PULSE* do Minitab

Refere-se a um experimento feito por alunos. Cada um deles registrou sua altura, peso, sexo, hábito de fumar e nível de atividade física. Depois, todos eles jogaram moedas e aqueles que tiraram "CARA" fizeram **corrida estacionária por 1 minuto**, registrando a pulsação antes de correr e a pulsação depois de correr. Os demais registraram a pulsação após 1 minuto, mesmo sem ter corrido.

Informações do arquivo *PULSE*

MTB > INFO

Information of the worksheet

Column	Count	Name	
C1	92	Pulse1	
C2	92	Pulse2	
C3	92	Ran	1- fez corrida 2- não fez corrida
C4	92	Smokes	1- fuma 2- não fuma
C5	92	Sex	1- masculino 2- feminino
C6	92	Height	
C7	92	Weight	
C8	92	Activity	0- não tem 1- leve 2- moderada 3- intensa

Variáveis qualitativas →

- Ran
- Smokes
- Sex — Nominal
- Activity — Ordinal

Variáveis quantitativas →

- Pulse 1
- Pulse 2 — Discreta
- Height
- Weight — Contínua

Planilha (parcial)

Row	Pulse1	Pulse2	Ran	Smokes	Sex	Height	Weight	Activity
1	64	88	1	2	1	66,00	140	2
2	58	70	1	2	1	72,00	145	2
3	62	76	1	1	1	73,50	160	3
4	66	78	1	1	1	73,00	190	1
5	64	80	1	2	1	69,00	155	2
6	74	84	1	2	1	73,00	165	1
7	84	84	1	2	1	72,00	150	3
8	68	72	1	2	1	74,00	190	2

•
•
•

Variáveis Quantitativas

Medidas de posição

Média ($\bar{x}$)

Mediana (md)

Quartis ($Q1$, $Q3$)

Máximo (máx)

Mínimo (min)

Medidas de dispersão

Variância (s^2)

Desvio padrão (s)

Intervalo-interquartil ($Q3 - Q1$)

Coeficiente de variação (CV)

MTB > describe c1 c6 c7

Descriptive Statistics

Variable	*N*	Mean	Median	Tr Mean	StDev	SE Mean
Pulse1	92	72,87	71	72,61	11,01	1,15
Height	92	68,72	69	68,784	3,659	0,382
Weight	92	145,15	145	144,52	23,74	2,48

Variable	Min	Max	*Q*1	*Q*3	*CV*
Pulse1	48	100	64	80	**11,01/72,87=0,15**
Height	61	75	66	72	**3,659/68,717=0,05**
Weight	95	215	125	156,5	**23,74/145,15=0,16**

Os dados também podem ser resumidos construindo-se uma tabela de distribuição de freqüências .

Distribuição de freqüências de uma variável é uma lista dos valores individuais ou dos intervalos de valores que a variável pode assumir, com as respectivas freqüências de ocorrência.

Resumo de um conjunto de dados

$x_1, x_2, \ldots, x_n$

Podemos construir uma tabela contendo k classes:

a) As freqüências absolutas n_i, i = 1, ..., k

b) As freqúências relativas $f_i = n_i/n$, i=1, ..., k

c) As densidades de freqüência $d_i = f_i/h_i$, i = 1, ..., k

No arquivo ***PULSE***

MTB > tally c1

Não há perda de informação

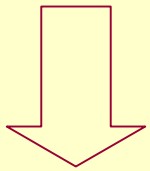
Summary Statistics for Discrete Variables

Pulse1	Count	Percent
48	1	1,09
54	2	2,17
58	3	3,26
60	4	4,35
61	1	1,09
62	9	9,78
64	4	4,35
66	5	5,43
68	11	11,96
70	6	6,52
72	6	6,52
74	5	5,43
76	5	5,43
78	5	5,43
80	3	3,26
82	3	3,26
84	4	4,35
86	1	1,09
87	1	1,09
88	3	3,26
90	4	4,35
92	2	2,17
94	1	1,09
96	2	2,17
100	1	1,09
N=	92	

Alternativa: construir intervalos de classe

Classe de pulsação	frequência
48 |- 54	1
54 |- 60	5
60 |- 66	18
66 |- 72	22
72 |- 78	16
78 |- 84	11
84 |- 90	9
90 |- 96	7
96 |- 102	3

Informações mais resumidas



Perda de informação

Exemplo 2:

Variável: altura (*height*) ⇨ contínua ⇨ Construir intervalos de classe

Distribuição de freqüência para altura

(arquivo *PULSE*)

Classes de altura	f	fr
60,25 |- 61,75	1	0,011
61,75 |- 63,25	10	0,109
63,25 |- 64,75	2	0,022
64,75 |- 66,25	13	0,141
66,25 |- 67,75	7	0,076
67,75 |- 69,25	20	0,217
69,25 |- 70,75	7	0,076
70,75 |- 72,25	15	0,163
72,25 |- 73,75	9	0,098
73,75 |- 75,25	8	0,087
Total	92	1

Variáveis Quantitativas

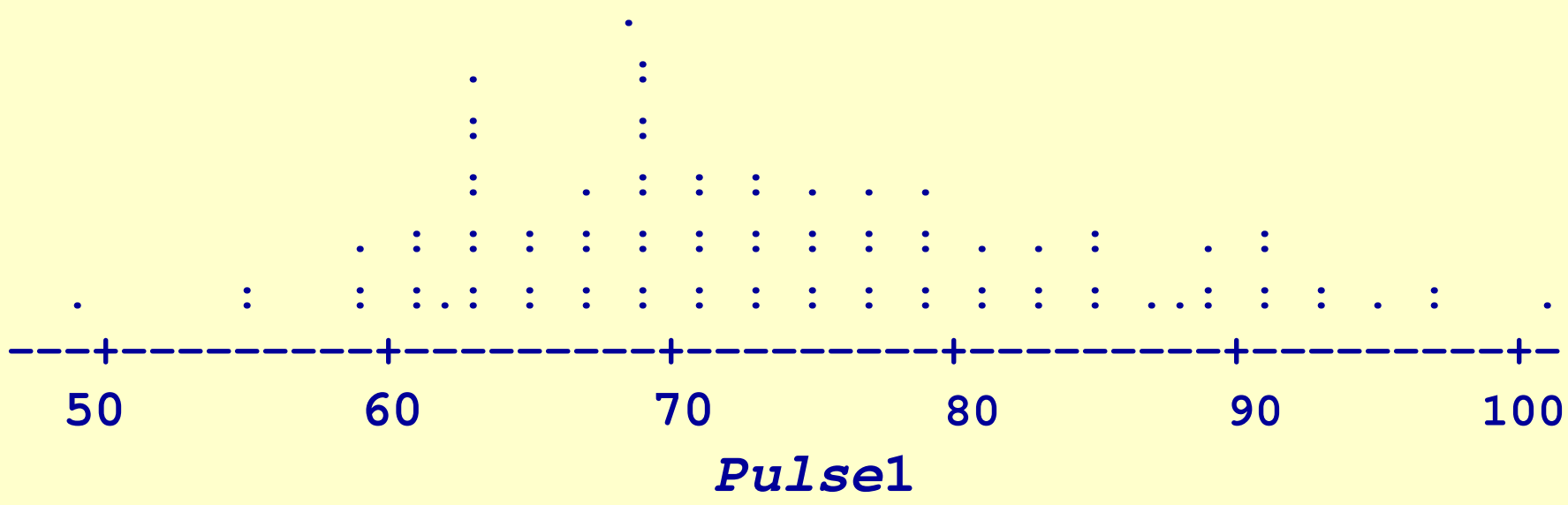
Gráficos

- "Dotplot "
- Histograma
- Ramos e folhas
- Função distribuição empírica

DOTPLOT

Arquivo ***PULSE*** – *Dotplot* da pulsação em repouso (*PULSE1*)

```
MTB > DOTPLOT C1
```



Arquivo *PULSE* – *Dotplot* da pulsação em repouso (*PULSE1*) segundo Sexo (*SEX*)

```
MTB > DotPlot 'Pulse1' ;
SUBC> Same;
SUBC> By 'Sex'.
```

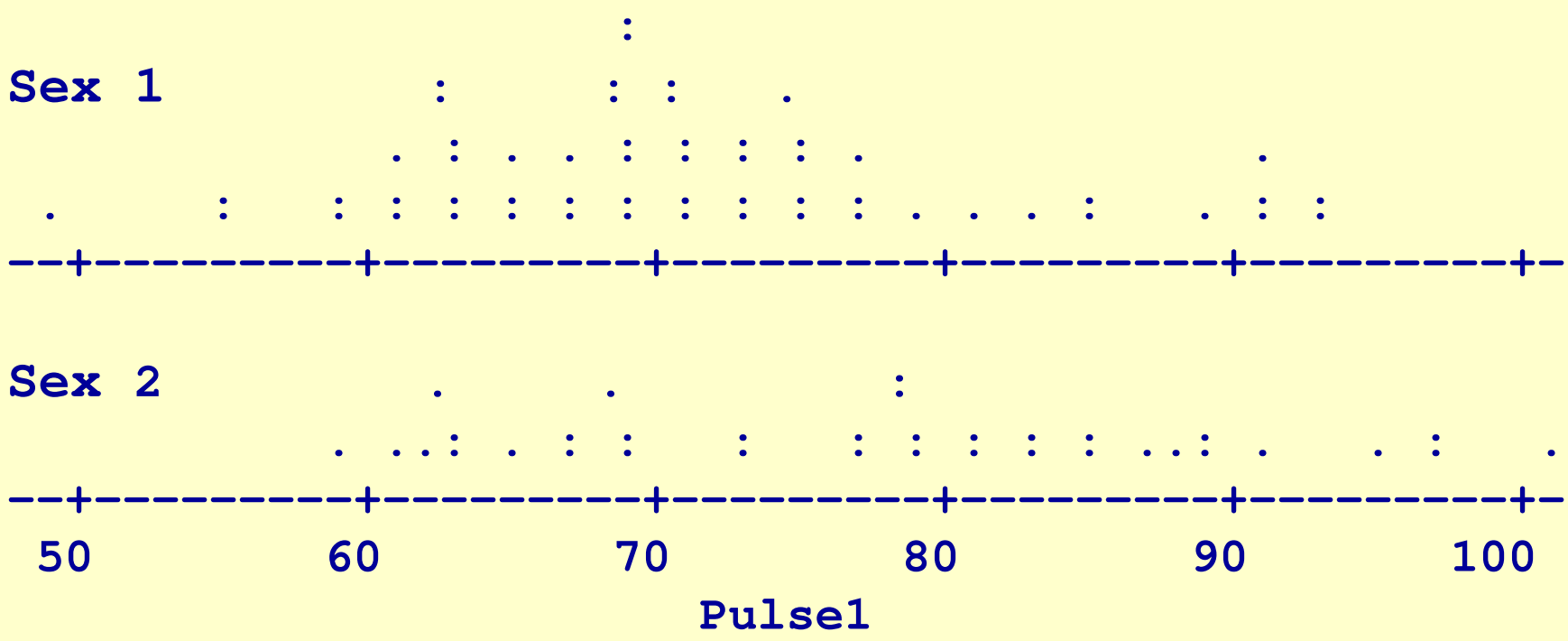
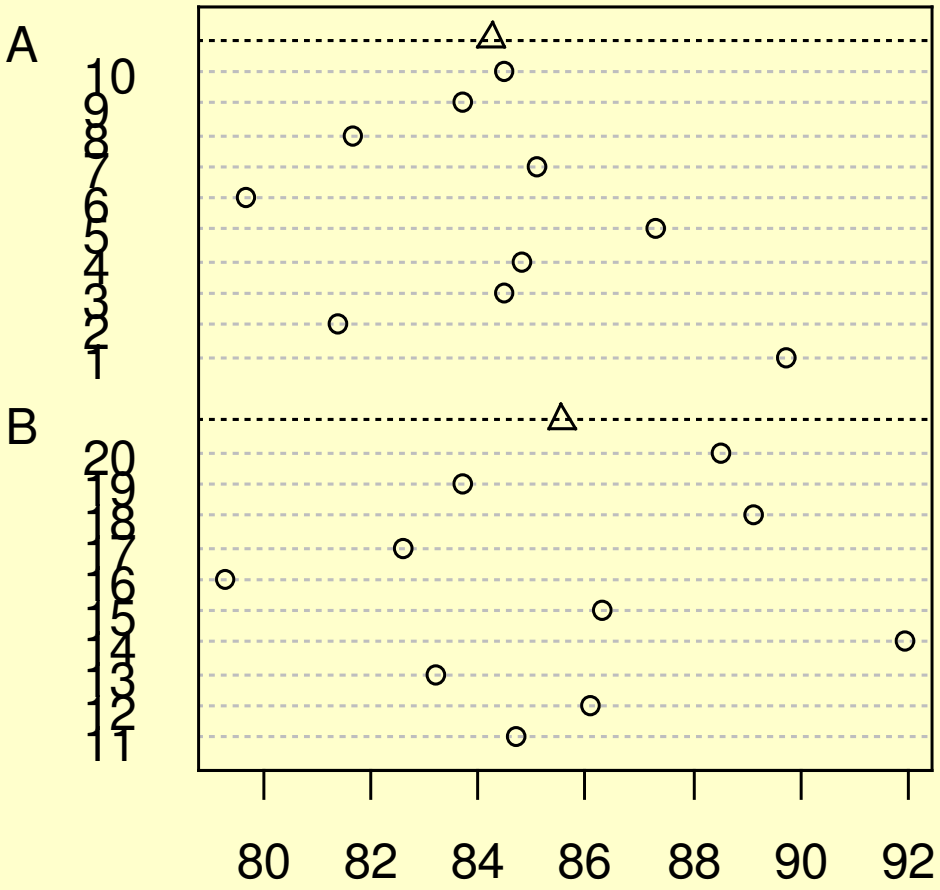


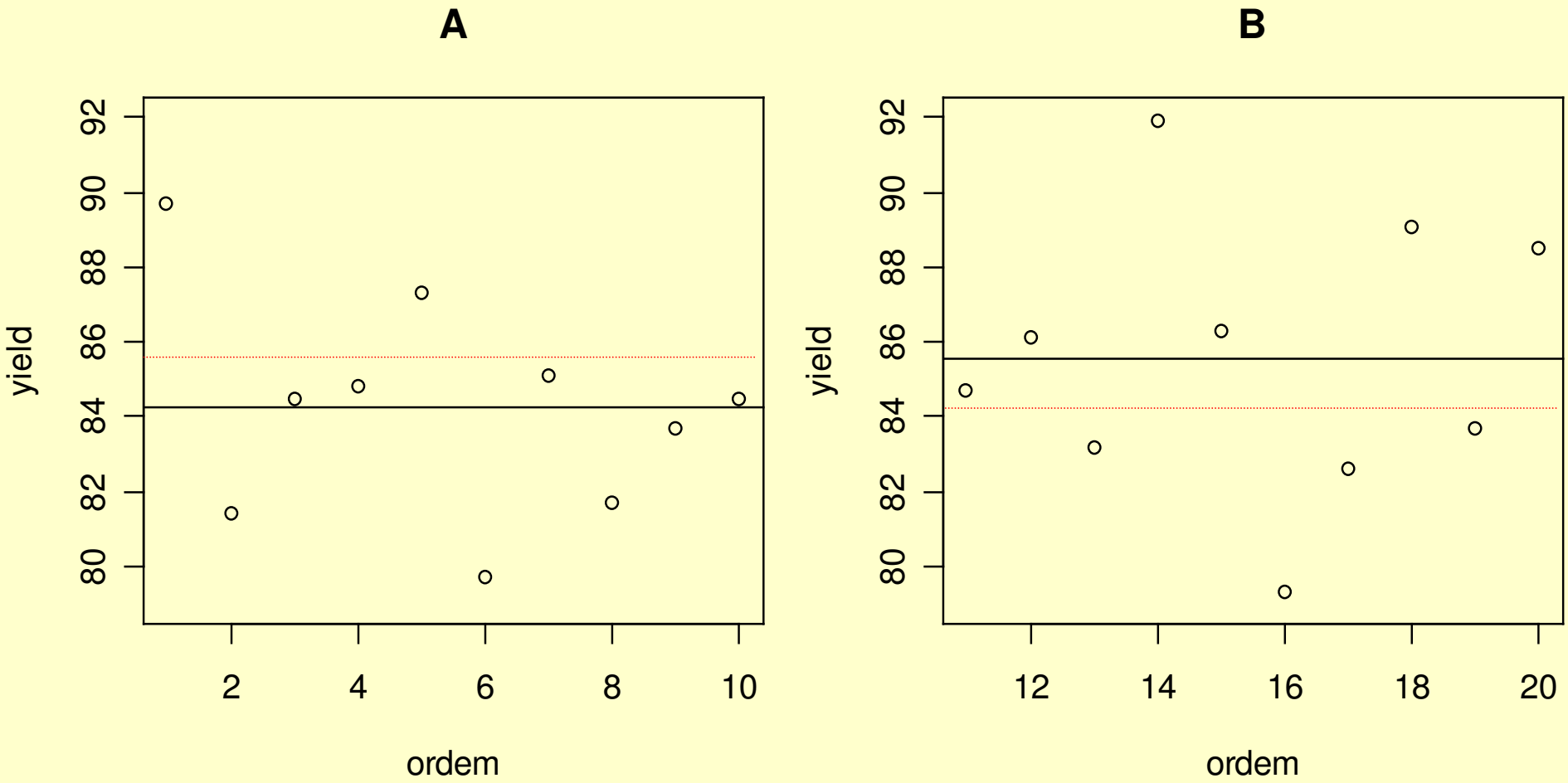
Gráfico de Dispersão de Pontos



- Representação útil para caracterizar padrões de dependência serial

```
> dotchart(yield,labels=ordem,groups=metodo,gdata=c(84.24,85.54),gpch=2)
```

Gráfico de Dispersão de Pontos




```
> plot(yield ~ ordem,subset = metodo == "A",ylim=c(79,92),main='A')
> abline(h=mean(yield[metodo == "A"]))
```

Histograma

Agrupar os dados em intervalos de classes (distribuição de freqüências)

Bases iguais

Construir um retângulo para cada classe, com base igual ao tamanho da classe e *altura proporcional à freqüência da classe (f).*

Bases diferentes

Construir um retângulo para cada classe, com base igual ao tamanho da classe e *área do retângulo igual a freqüência relativa da classe(fr).*

A altura será dada por

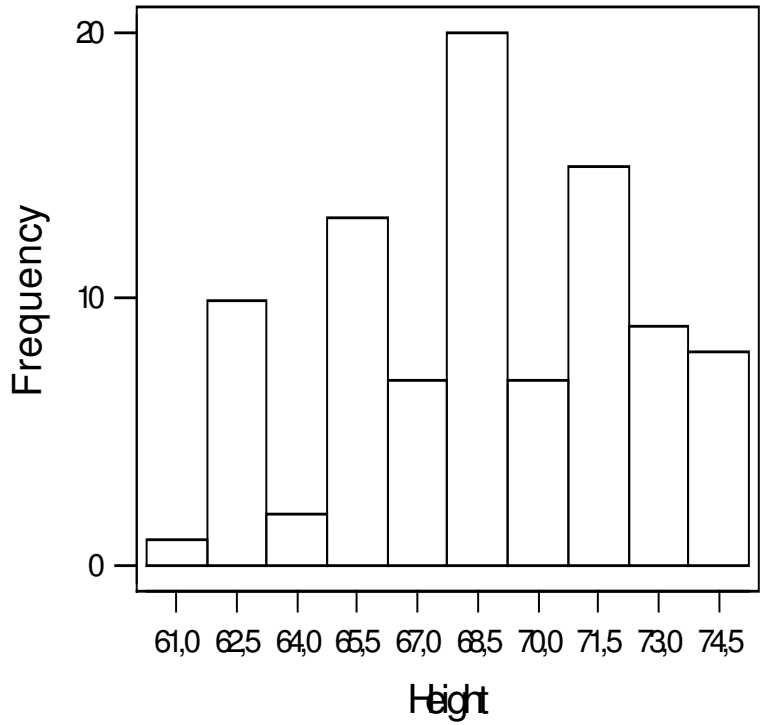
h = fr/base (densidade de freqüência).

Arquivo *PULSE* – Histograma da altura (*Height*)

Distribuição de freqüência para altura (arquivo *PULSE*)

Classe de altura	f	fr
60,25 ⊢ 61,75	1	0,011
61,75 ⊢ 63,25	10	0,109
63,25 ⊢ 64,75	2	0,022
64,75 ⊢ 66,25	13	0,141
66,25 ⊢ 67,75	7	0,076
67,75 ⊢ 69,25	20	0,217
69,25 ⊢ 70,75	7	0,076
70,75 ⊢ 72,25	15	0,163
72,25 ⊢ 73,75	9	0,098
73,75 ⊢ 75,25	8	0,087
Total	92	1

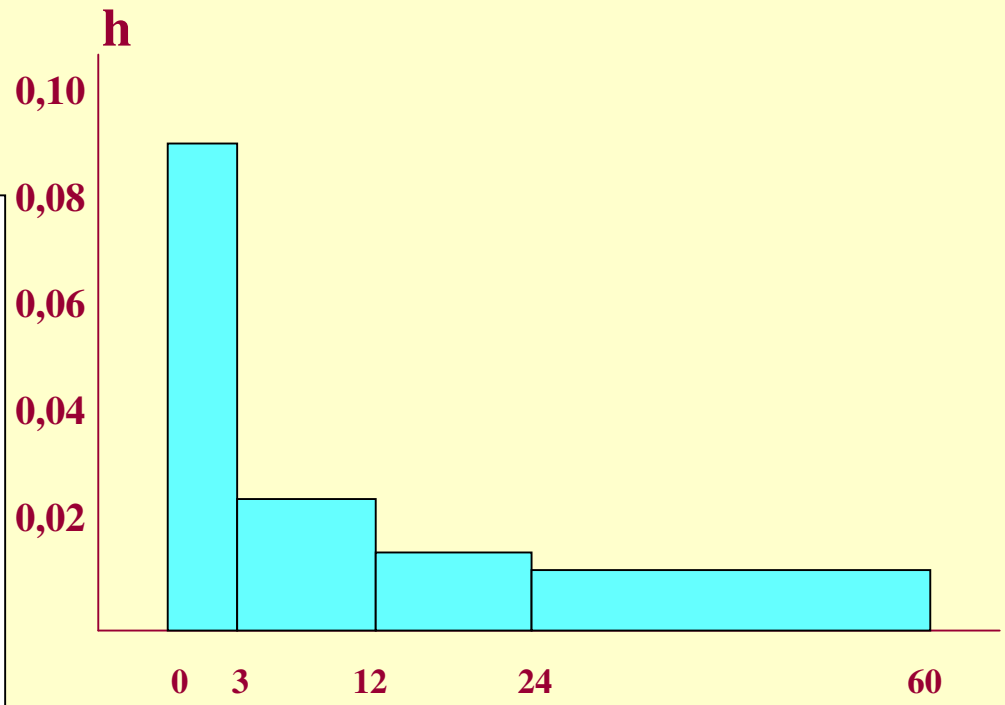
```
MTB > HIST C6
```

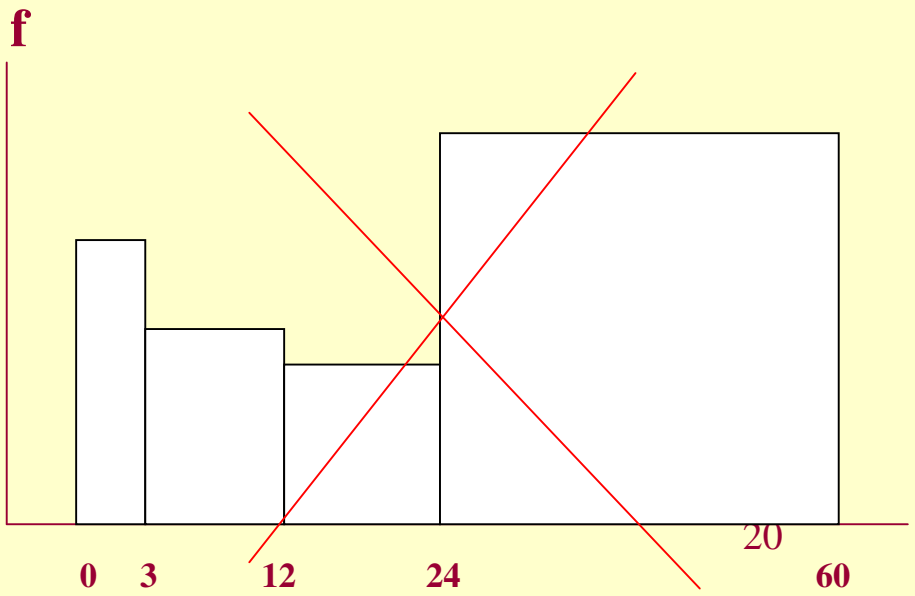


Exemplo: Classes desiguais

Vacinação Infantil

Classes (meses)	f	fr	h
0 |- 3	140	0,28	0,093
3 |- 12	100	0,20	0,022
12 |-24	80	0,16	0,013
24 |-60	180	0,36	0,010
Total	500	1,00	





Observações:

a) O número de classes utilizadas é obtido aproximadamente por

$$c \approx [x_{(n)} - x_{(1)}]/h$$

b) h grade: poucas classes e o histograma pode não revelae dados importantes;

c) H pequeno: muitas classes e algumas poderão ser vazias.

Freedman e Diaconis (1981):

$$H = 1{,}349\ s\ (\log n/n)^{1/3}$$

Onde s: estimador robusto de σ

Gráfico Ramos-e-Folhas

A

1	79	7
1	80	
3	81	47
3	82	
4	83	7
(3)	84	558
3	85	1
2	86	
2	87	3
1	88	
1	89	7

B

1	79	3
1	80	
1	81	
2	82	6
4	83	27
5	84	7
5	85	
5	86	13
3	87	
3	88	5
2	89	1
1	90	
1	91	9

- Representação gráfica das observações sem qualquer perda de informação sobre os dados originais
- Os valores da profundidade de cada linha auxilia no cálculo
- Quando há muitas folhas num ramo, podemos considerar ramos subdivididas.

ROL

Ind	Ciclo (dias)
1	22.9
2	26.3
3	26.6
4	26.8
5	26.9
6	26.9
7	27.5
8	27.6
9	27.6
10	28.0
11	28.4
12	28.4
13	28.5
14	28.8
15	28.8
16	29.4
17	29.9
18	30.0
19	30.3
20	31.2
21	31.8

Observações Ordenadas

uma maneira simples de garantir resistência

Gráfico Ramo-e-Folhas

	1	22	9
		23	
		24	
		25	
	6	26	3 6 8 9 9
	9	27	5 6 6
n=21	(6)	28	0 4 4 5 8 8
	6	29	4 9
	4	30	0 3
	2	31	2 8

$\Rightarrow$ *A profundidade do valor 30.3 é 3*

Gráfico Ramo-e-Folhas

- **escolha de uma Escala para dispor os dados**

$$L = 10 \ \log_{10} n : \ \#\text{de ramos}$$

$$\lambda = AV / L \ : \text{comprimento dos intervalos}$$

1	22	9					
	23						
	24						
	25						
6	26	3	6	8	9	9	
9	27	5	6	6			
(6)	28	0	4	4	5	8	8
6	29	4	9				
4	30	0	3				
2	31	2	8				

- stem(ciclo)
- stem(ciclo,scale=2)

1	22	9			
2	26 *	3			
6	26 ·	6	8	9	9
	27 *				
9	27 ·	5	6	6	
(3)	28 *	0	4	4	
9	28 ·	5	8	8	
6	29 *	4			
5	29 ·	9			
4	30 *	0	3		
	30 ·				
2	31 *	2			
1	31 ·	8			

Função distribuição empírica

$F_e(x) = n(x)/n$, qq x real

. n(x): no. de observações $\leq$ x;

. $F_e(x)$: estimador de F(x)

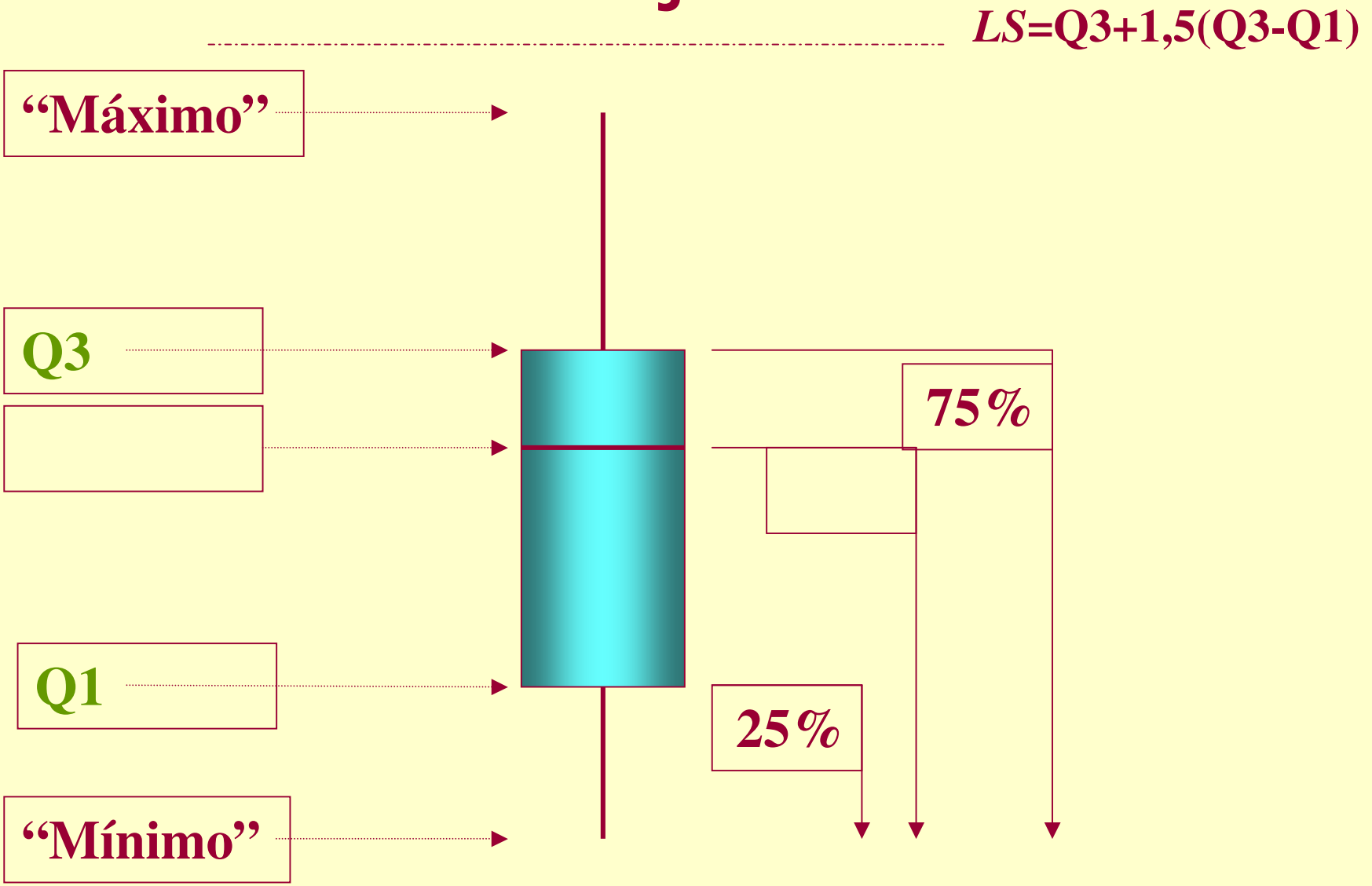
Exemplo: 1, 2, 3, 4, 8

$F_e(x_{(i)}) = i/n$, i=1, ..., n

Boxplot

Representa os dados através de um retângulo construído com os quartis e fornece informações sobre os valores extremos.

Construção



"Máximo" é o maior valor menor que *LS*;

"Mínimo" é o menor valor maior que *LI*.

Exemplo

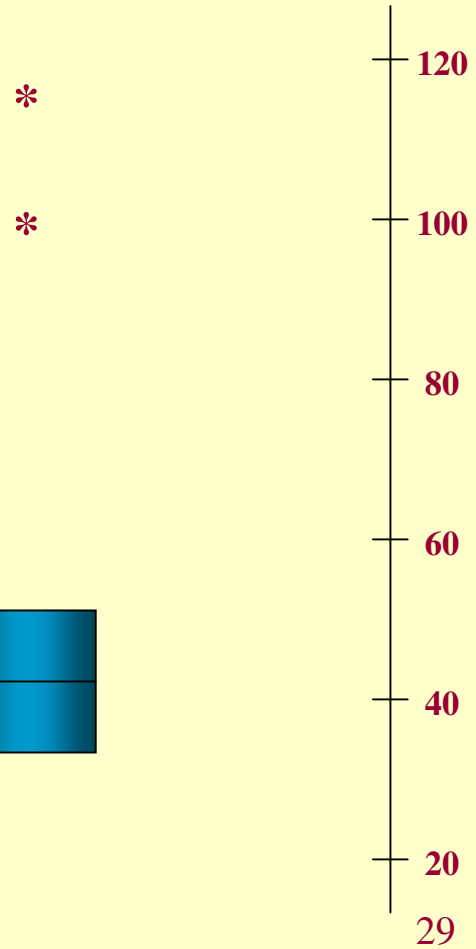
Resistência do Box Plot

Tempo de Sobrevivência (dias)

Dados Ordenados (n=36)

18	21	21	23	23	25
27	29	30	31	32	32
32	34	35	36	38	41
42	42	43	44	45	46
46	47	48	50	54	56
57	58	60	61	98	116

Me = 41,5 Q1 = 30,25 Q3 = 49,5

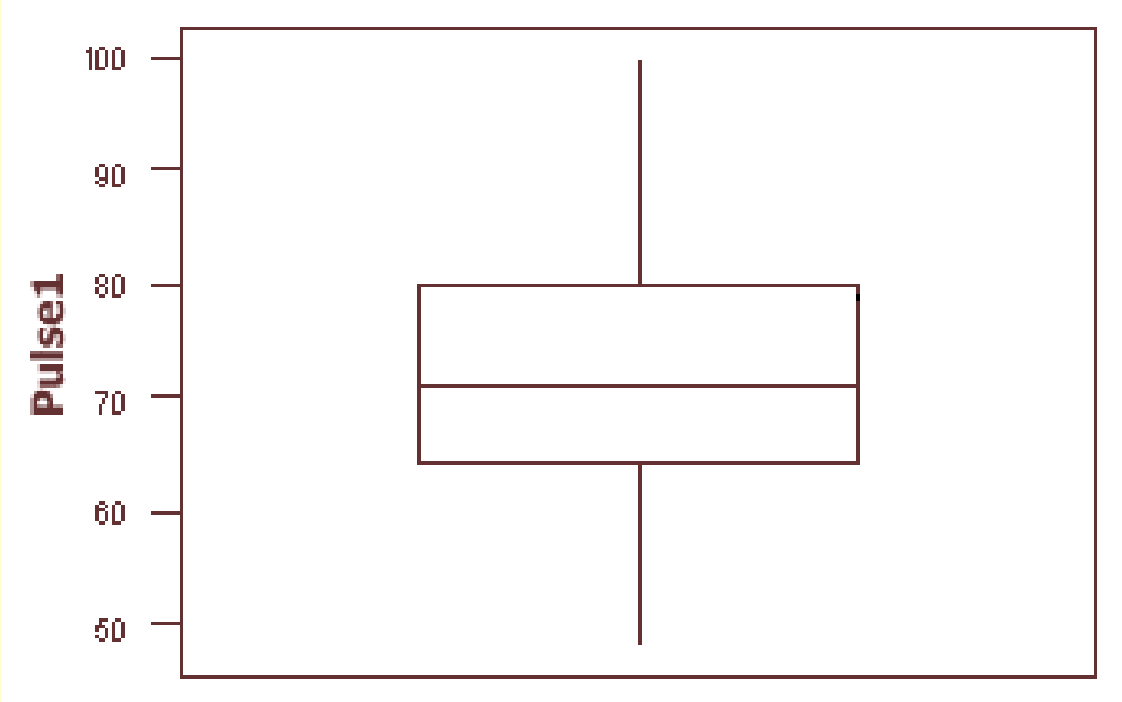


Observações aberrantes ?

Q1 - 1,5(Q3 - Q1) **Q3 + 1,5(Q3 -Q1)**

Arquivo *PULSE* – *Boxplot* da pulsação em repouso (*PULSE1*)

```
MTB > BOXPLOT C1
```

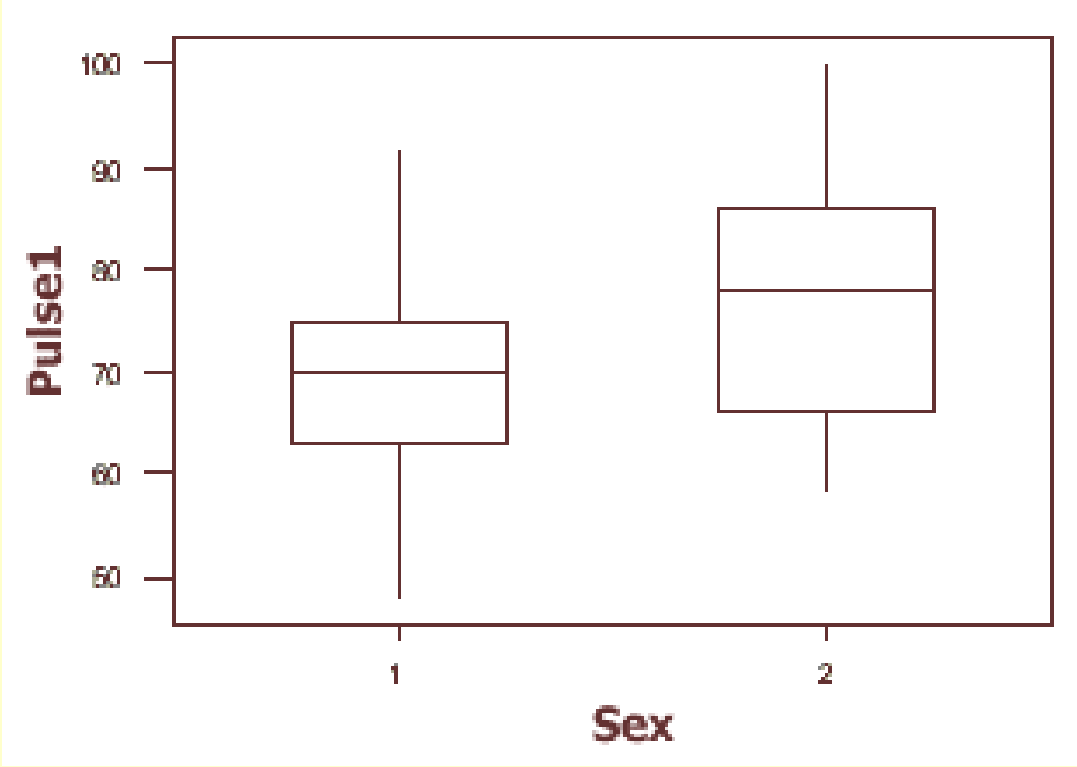


Alguns Comentários:

- **não há observações discrepantes;**
- **a distribuição dos valores é aproximadamente simétrica.**

Arquivo *PULSE* – *Boxplots* da pulsação em repouso (*PULSE1*) por sexo (*SEX*)

```
MTB > BOXPLOT C1*C5
```

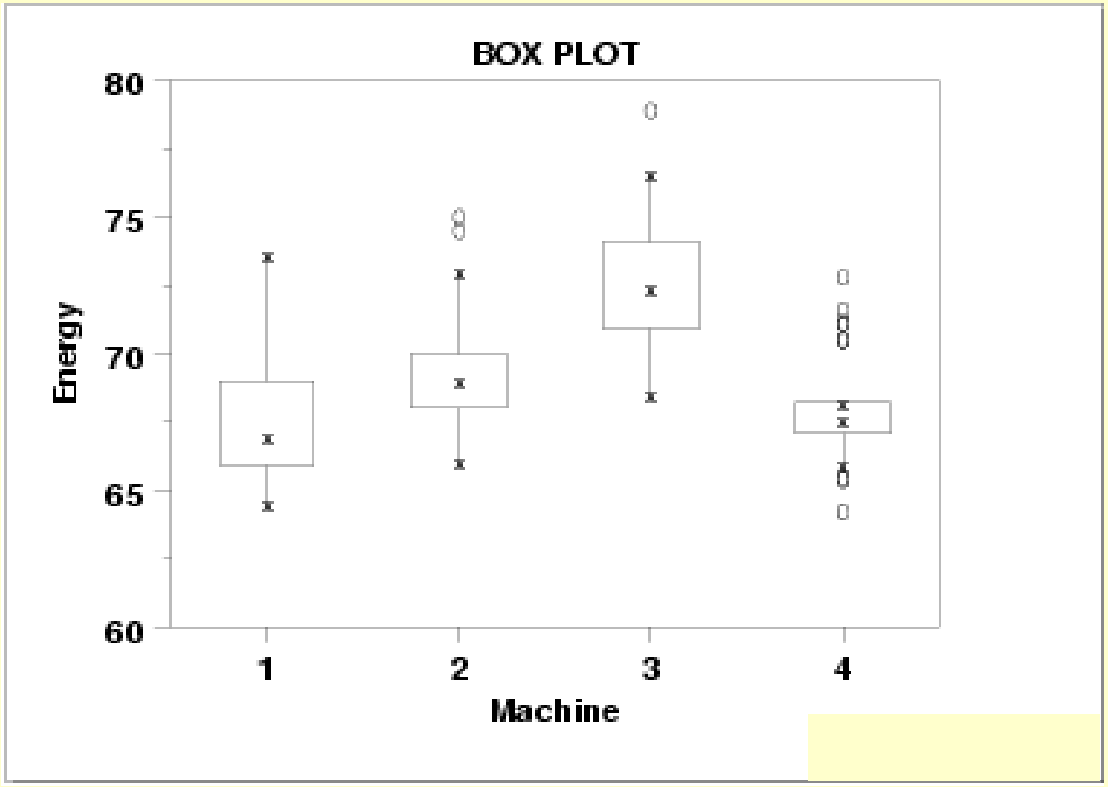


Alguns Comentários:

- **não há observações discrepantes;**
- **as medidas de posição são maiores para o sexo feminino;**
- **não há fortes evidências de assimetria nos dois grupos.**

Valores Amostrais Típicos e Outliers

Atípicos



Sob a Normal, estes pontos outliers são esperados ocorrer a quantos desvios padrão distante da média ?

L1 = Q1 - 1.5*IQ L2 = Q1 - 3.0*IQ

U1 = Q3 + 1.5*IQ U2 = Q3 + 3.0*IQ

Parâmetros de Posição e Escala

Y; $f_Y(y \mid \theta, \lambda)$

$f_{Y-\theta}(y-\theta \mid \lambda)$ $\Rightarrow \theta$ é parâmetro de posição (a distrib. da variável (Y-θ) não depende de θ)

$f_{\frac{Y}{\lambda}}\left(\frac{y}{\lambda} \mid \theta\right)$ $\Rightarrow$ λ é parâmetro de escala

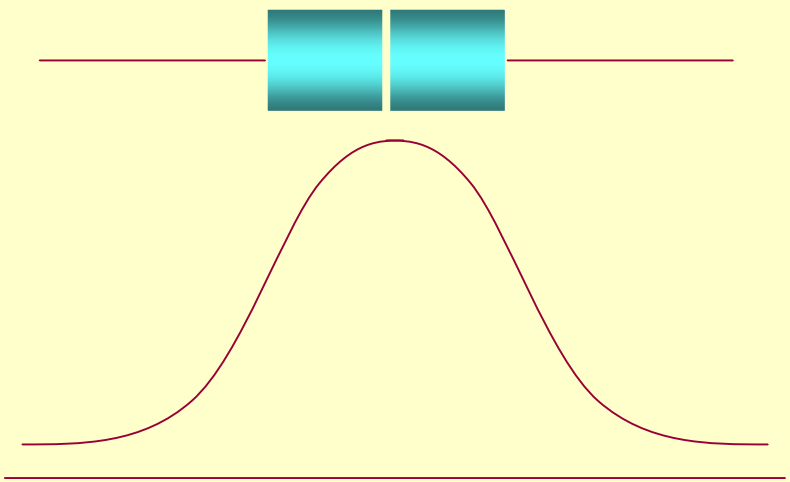
$Y=(Y_1,\ldots,Y_n)$; $T(a+y_1,\ldots,a+y_n)=a+T(y_1,\ldots,y_n)$ $\Rightarrow$ T é estimador do parâmetro de posição

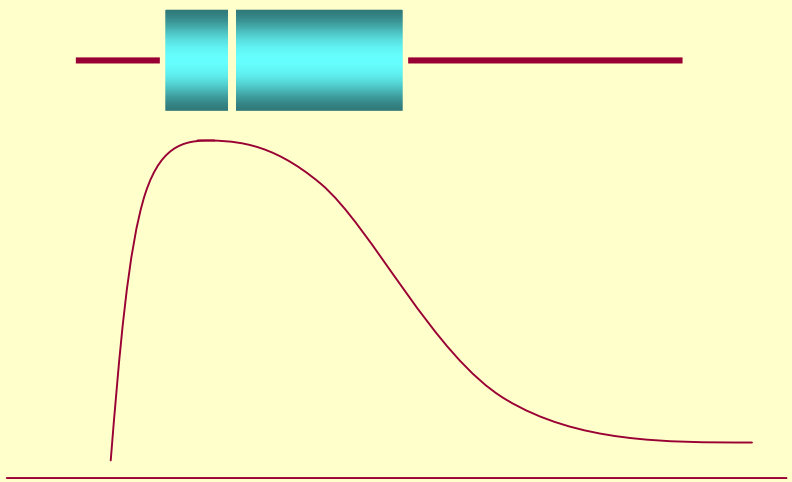
$Y=(Y_1,\ldots,Y_n)$; $T(a+b\ y_1,\ldots,a+b\ y_n)=|b|\ T(y_1,\ldots,y_n)$

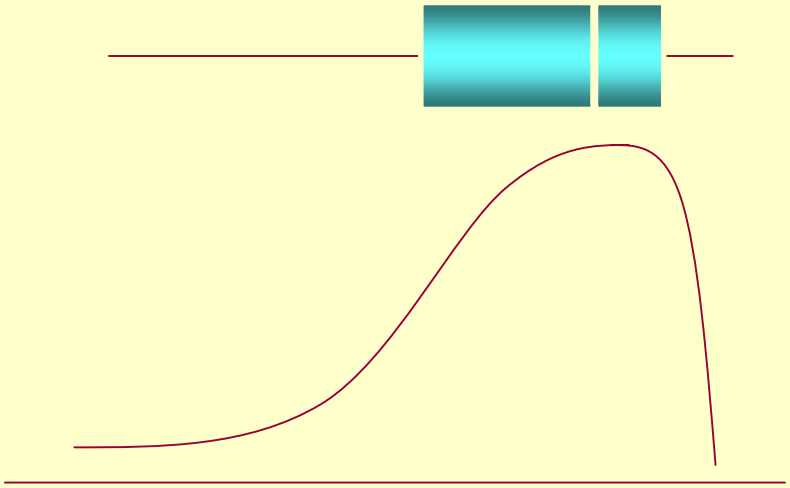
$\Rightarrow$ T é estimador do parâmetro de escala

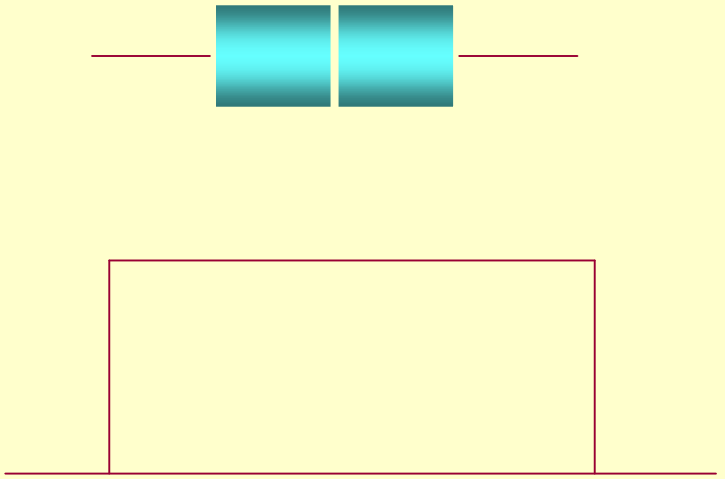
- Obtenção de estimadores robustos para parâmetros de posição e escala
- Dificuldade: conhecer o comportamento dos estimadores pode variar $\Rightarrow$ devido a fugas da Normalidade, devido ao efeito das caudas das distribuições

Forma da Distribuição



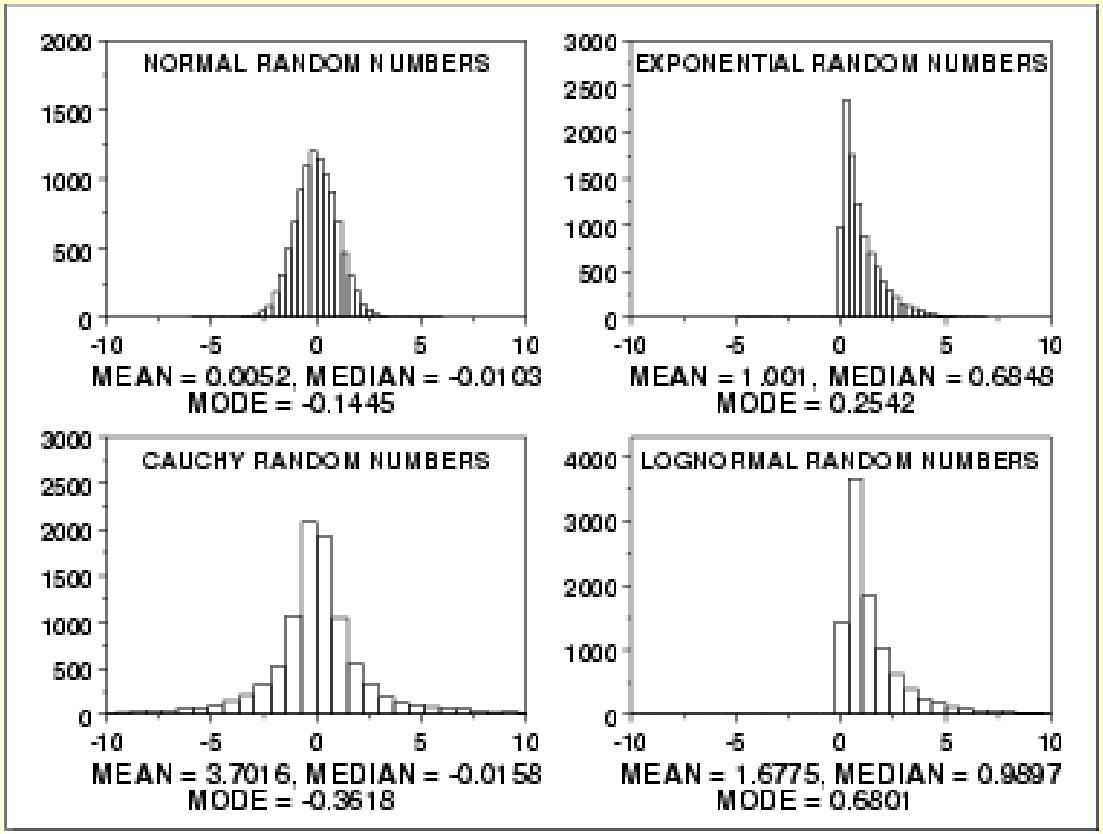






Forma da Distribuição
Estimadores de Posição

Geração de Números Aleatórios



Garantir robustez e resistência

Casos de caudas pesadas: uma alternativa é adotar medidas que penalizam os dados

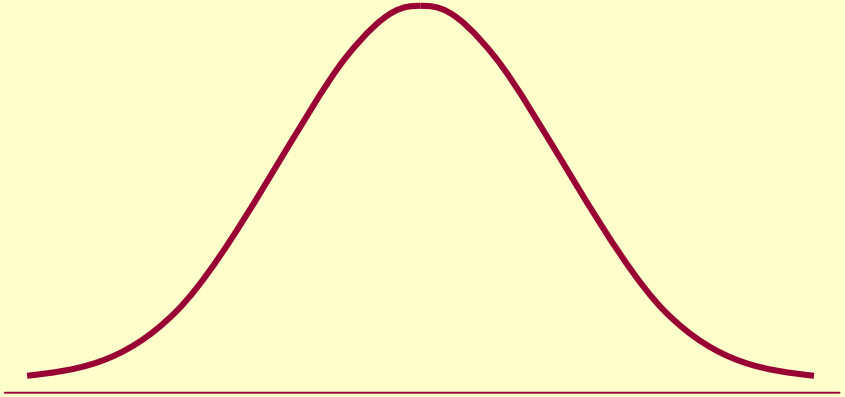
Estimador do parâmetro de posição:

Normal: média

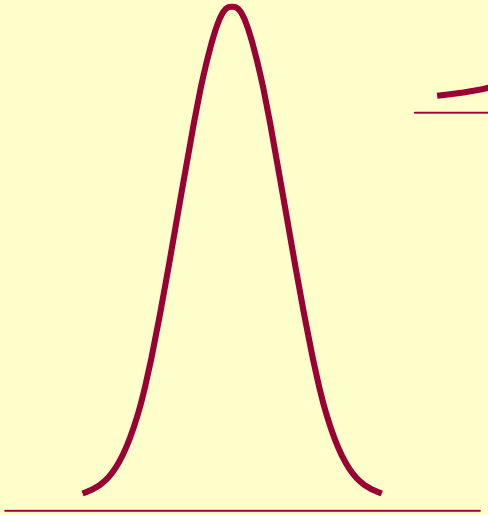
Exponencial: distr assimétricas não há um consenso sobre qual estimador adotar (média, mediana, moda, média tri,…)

Cauchy: propriedade interessante $\Rightarrow$ aumentando *n* não aumenta a precisão da média (distr amostral da média é igual à distr original dos dados). Neste caso a mediana é adotada, sendo um valor amostral típico

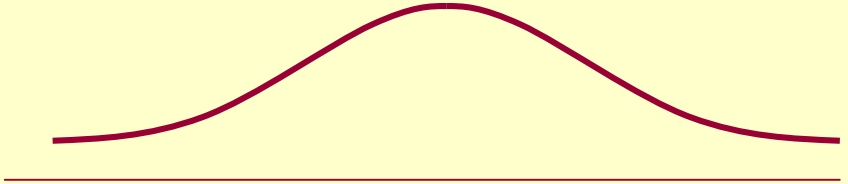
Forma da Distribuição
Classe das Distribuições Simétricas



Distribuição Mesocúrtica
Cauda Neutra



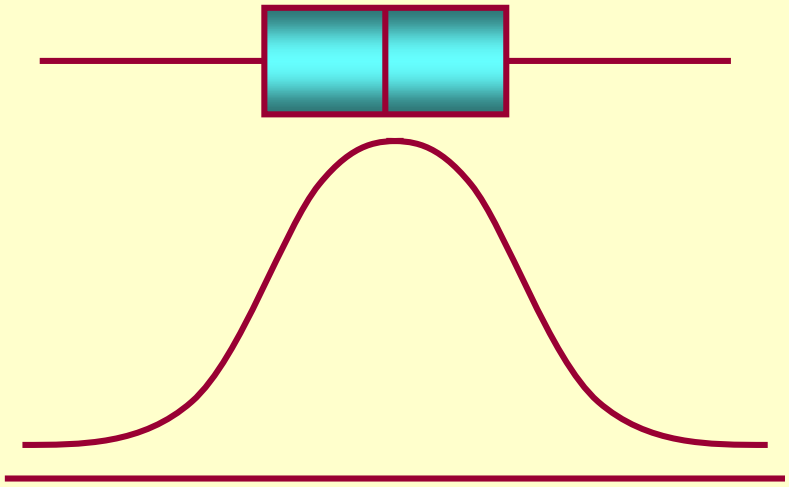
Distribuição Leptocúrtica
Cauda “Pesada” (curta)



Distribuição Platicúrtica
Cauda “Leve” (longa)

$\Rightarrow$ Parâmetros de Locação e Escala das distribuições

Forma da Distribuição
Obtenção de Estimadores na "Classe Simétrica"



Simetria da Distribuição Normal:

$$Q_2 - Y_{(1)} \approx Y_{(n)} - Q_2$$
$$Q_2 - Q_1 \approx Q_3 - Q_2$$
$$Q_1 - Y_{(1)} \approx Y_{(n)} - Q_3$$

Se uma distribuição é aproximadamente Simétrica:

$$Q_2 - Y_{(j)} = Y_{(n-j+1)} - Q_2; \qquad j = 1,2,...,[(n+1)/2]$$

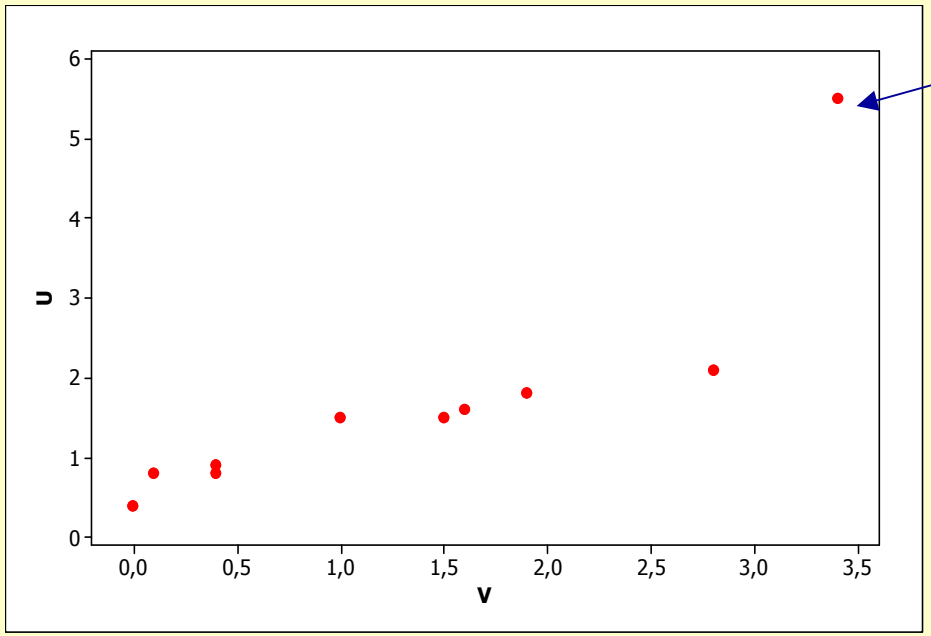
$\Rightarrow$ **Gráfico de Simetria:** $u_j = Q_2 - Y_{(j)} \quad \times \quad v_j = Y_{(n-j+1)} - Q_2$

Forma da Distribuição "Classe Simétrica"

⇒ **Gráfico de Simetria:** $u_j = Q_2 - Y_{(j)} \quad \times \quad v_j = Y_{(n-j+1)} - Q_2$

22,9 26,3 26,6 26,8 26,9 26,9 27,5 27,6 27,6 28,0 28,4 28,4 28,5 28,8 28,8 29,4 29,9 30,0 30,3 31,2 31,8

Q2=Md=28,4



Ponto atípico