Estatística Descritiva

Prof. Lorí Viali, Dr.
viali@mat.ufrgs.br
http://www.mat.ufrgs.br/~viali/

Análise Exploratória de Dados

As técnicas de análise exploratória de dados consistem em gráficos simples de desenhar que podem ser utilizados para resumir rapidamente um conjunto de dados. Uma destas técnicas é uma forma de apresentação de dados conhecida como **Caule e Folha**.

Apresentação Caule e Folha

Para ilustrar esta forma de apresentação vamos supor que o conjunto a seguir é o resultado de um teste do tipo Psicotécnico de 100 questões aplicados a 40 candidatos a um emprego em uma grande organização industrial.

Exemplo

Resultado de um teste do tipo Psicotécnico de 100 questões aplicados a 40 candidatos.

44	53	67	89	98	37	60	55
48	88	47	65	82	85	90	74
41	61	72	73	77	81	60	89
52	90	62	64	66	59	50	65
50	40	93	79	55	49	56	73

Ramo e Folha

3	7								
4	0	1	4	7	8	9	9		
5	0	0	2	3	5	5	6	9	
6	0	0	1	2	4	5	5	6	7
7	2	3	3	3	4	7	9		
8	1	2	5	5	8	8	9		
9	0	0	3	8					

Girando a representação 90 graus tem-se um diagrama semelhante a um histograma. Esta representação possui duas vantagens sobre o histograma:

É mais fácil de construir;

Apresenta os dados reais.

Exercício

Faça um representação utilizando a dezena como unidade de folha.

1565	1790	1644	1679	2008
1675	1900	1832	1756	1766
1580	1945	1733	1922	1854
1975	1870	1812	1954	1888
1634	1785	1855	2044	1965

BoxPlot – Caixa e Bigode

Outra forma de ter uma idéia do conjunto de dados é utilizar a regra dos cinco itens. Nem sempre a média e o desvio padrão são as melhores alternativas para resumir um conjunto de dados.

A média e o desvio padrão podem sofrer forte influência de valores extremos e além disso não fornecem uma idéia da assimetria do conjunto de dados. Como alternativa as seguintes cinco medidas são sugeridas (Tukey, 1977):

(i) A mediana;

(ii) Os extremos (máximo e mínimo);

(iii) Os quartis.

Estas cinco medidas são denominadas de estatísticas de ordem e são resistentes de posição de uma distribuição.

Representação

A informação fornecida por estes cinco números pode ser representada em um diagrama denominado de "Diagrama Caixa e Bigode" (*BoxPlot*). O desenho fornece uma idéia da posição, dispersão, assimetria e dados discrepantes do conjunto (*outliers*).

Traçar um retângulo tendo como extremos os quartis e englobando a mediana. Calcular a distância interquartil, isto é:

$$D_Q = Q_3 - Q_1$$

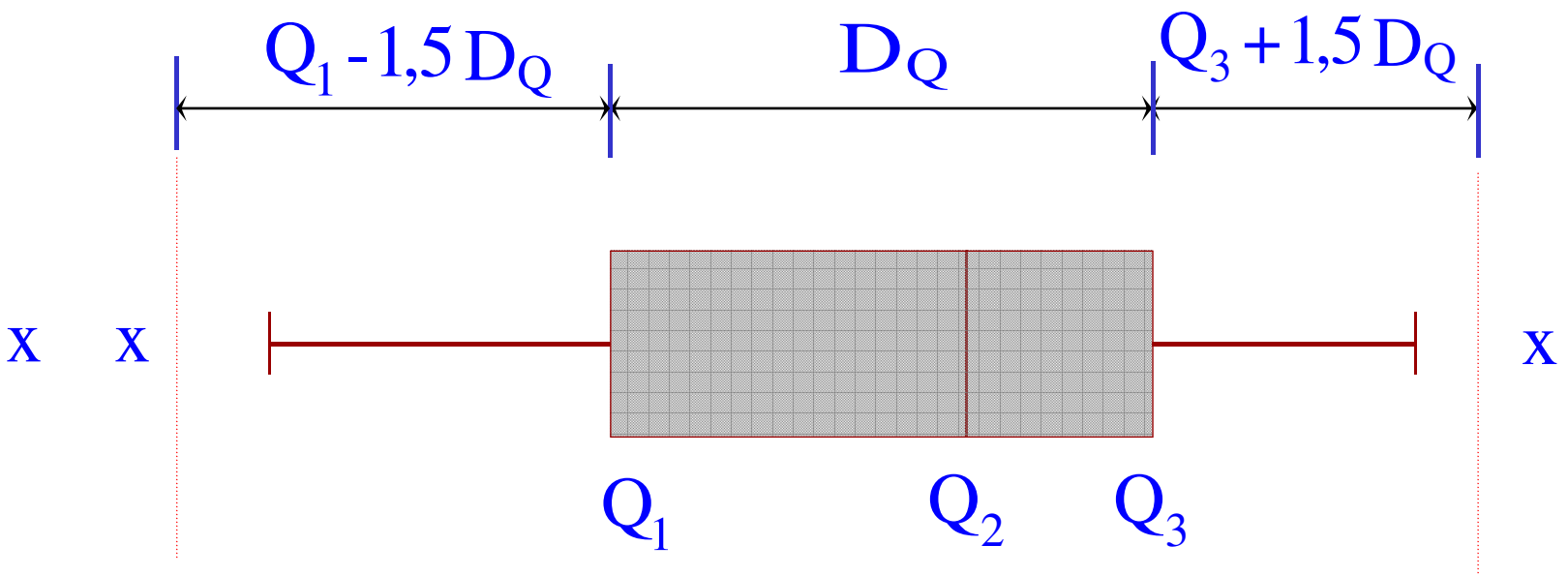
Determinar os limites dos pontos discrepantes:

$$Q_1 - 1{,}5\ D_Q$$

$$Q_3 + 1{,}5\ D_Q$$

Qualquer valor abaixo de Q1 – 1,5 DQ ou acima de Q3 + 1,5 DQ será considerado um valor discrepante (outlier). Para obter o diagrama caixa e bigode (*boxplot*) traçar duas linhas a partir do centro do retângulo e em lados opostos até o último ponto do conjunto que não seja um ponto discrepante.

BoxPlot



Exemplo

Obtenha o diagrama Caixa e Bigode para o número de paradas semanais para manutenção de uma máquina.

3	5	7	5	3	6	8	5	2
4	5	5	6	9	8	6	8	1
7	12	4	8	7	4	6		

Exemplo

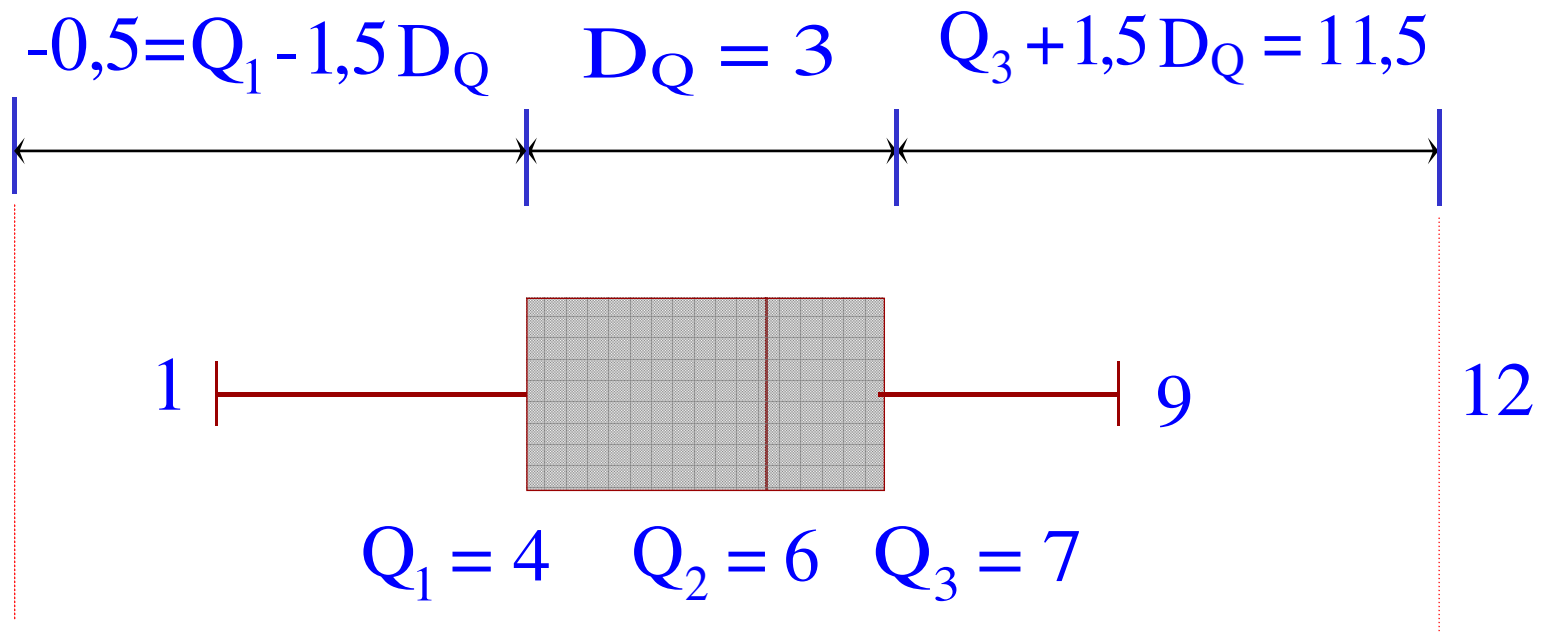
Os cinco valores são:

Mínimo	1
Quartil um	4
Mediana	6
Quartil três	7
Máximo	12

Os demais são:

D	7 – 4 = 3
Q_1 - 1,5D	-0,5
Q_3 + 1,5D	11,5
Outlier	12

BoxPlot



Wilfredo Pareto

O Diagrama de Pareto é uma homenagem ao engenheiro, filósofo, sociólogo e economista italiano Vilfredo Frederico Samaso Pareto (1848 - 1923). Pareto foi um dos pioneiros na aplicação de análises matemáticas ao estudo dos fenômenos sócio-econômicos.

Wilfredo enunciou, em 1897, o que passou a ser conhecido como "**Principio de Pareto**" que afirma: "80% das dificuldades tem origem em 20% dos problemas". Este principio poderia ser colocado como existem muitos itens triviais mas poucos vitais.

Diagrama

O Diagrama de Pareto é um gráfico de colunas simples, onde a variável está em ordem de importância freqüência de ocorrência ou custo) dos problemas ou defeitos.

Diagrama

Normalmente o diagrama envolve a freqüência simples combinada com a freqüência acumulada em um único gráfico. É, também, comum a colocação de um sistemas de eixos X'Y' auxiliares.

Exemplo

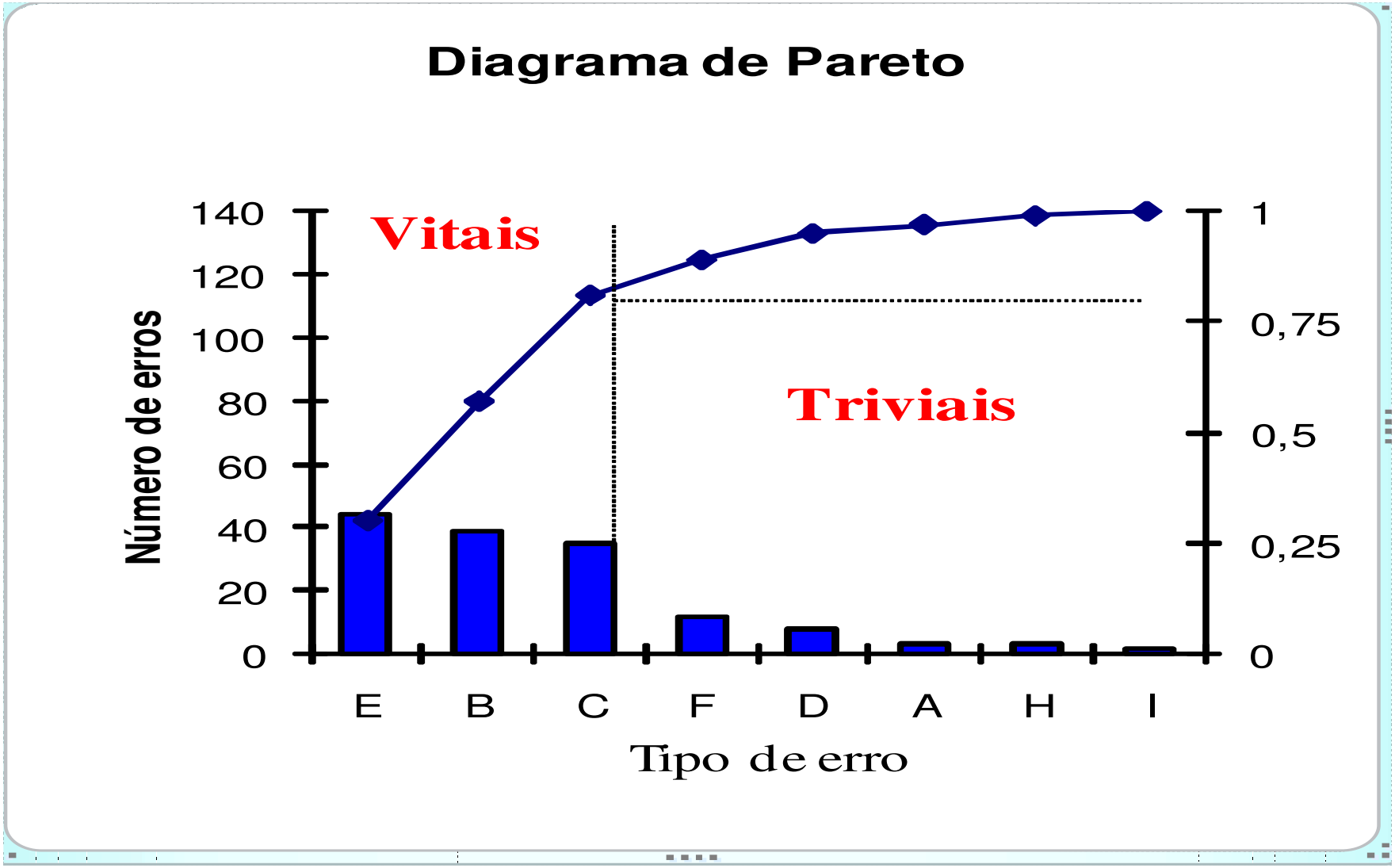


Diagrama de Pareto
Vitais
Triviais
Número de erros
140
120
100
80
60
40
20
0
1
0,75
0,5
0,25
0
E
B
C
F
D
A
H
I
Tipo de erro

Exercício

Considerando os dados sobre o "Número de defeitos" numa linha de produção de azulejos, construa o Diagrama de Pareto para a distribuição dada.

Defeitos	Número de Azulejos
Desenho	71
Esmalte	95
Lascado	97
Maior	70
Menor	83
Torto	57
Trincado	27
Total	**500**

Solução

Ordenando as freqüências dadas e calculando as freqüências relativas e relativas acumuladas, tem-se:

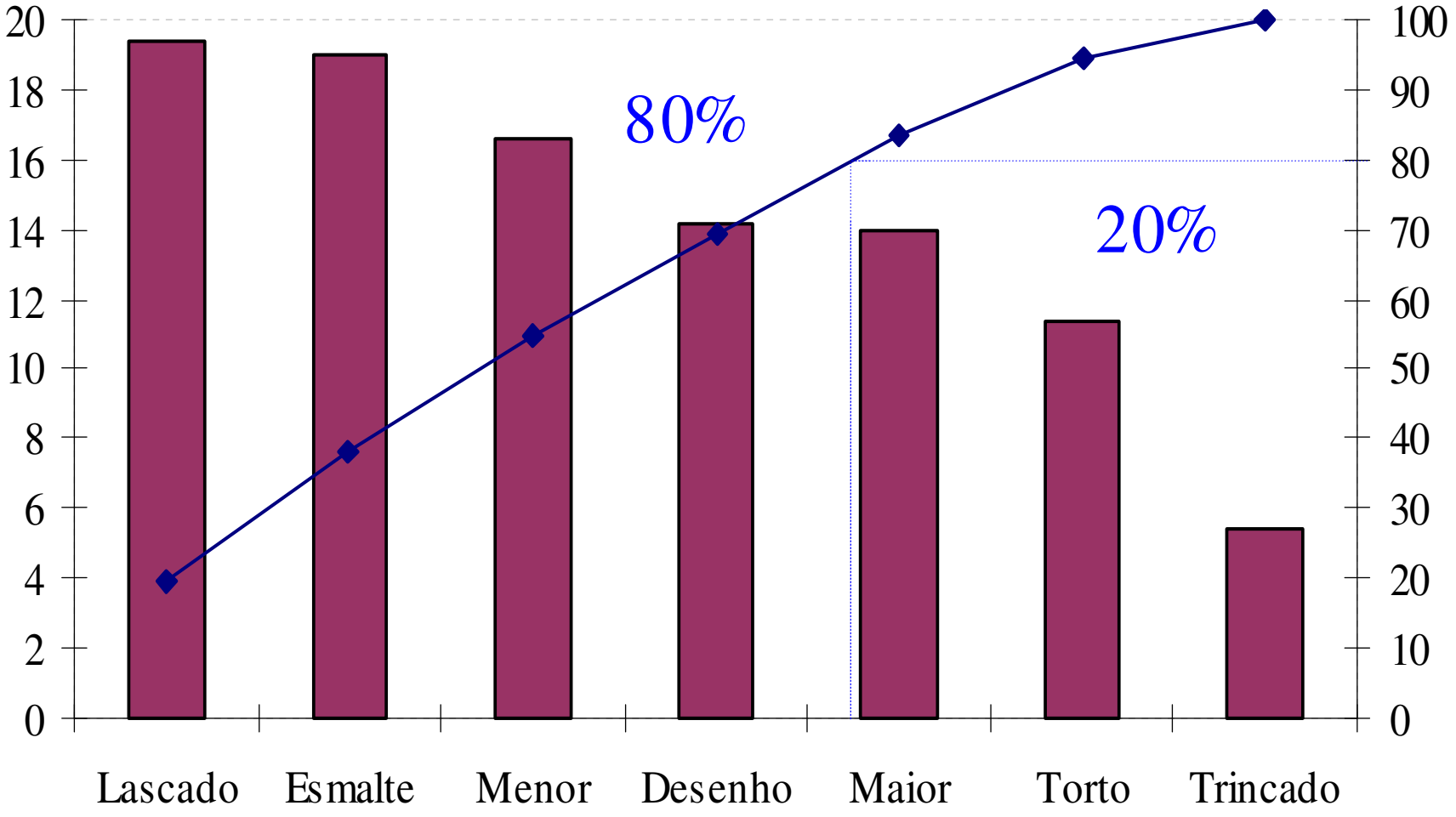
Ordenando as freqüências, tem-se:

Defeitos	Número de Azulejos
Lascado	97
Esmalte	95
Menor	83
Desenho	71
Maior	70
Torto	57
Trincado	27
Total	500

Calculando as demais freqüências:

Defeitos	**% de azulejos**	**Freq. acumulada**
Lascado	19,4	19,4
Esmalte	19,0	38,4
Menor	16,6	55,0
Desenho	14,2	69,2
Maior	14,0	83,2
Torto	11,4	94,6
Trincado	5,4	100,0
Total	**100**	**----**

Diagrama de Pareto



Posições Relativas

A média e o desvio padrão são as duas principais medidas utilizadas para descrever um conjunto de dados. Elas, também, podem ser utilizadas para comparações, isto é, para fornecer a posição relativa de um valor em relação ao conjunto como um todo.

O escore "z"

Seja $(x_1, x_2, ..., x_n)$ uma amostra de "n" observações. Sejam $\bar{x}$ e "s" a média e o desvio padrão da amostra. Então o escore z_i é o valor que fornece a posição relativa de cada x_i da amostra, tendo como ponto de referência a média e como medida de afastamento o desvio padrão.

O escore "z"

$$z_i = \frac{x_i - \bar{x}}{s}$$

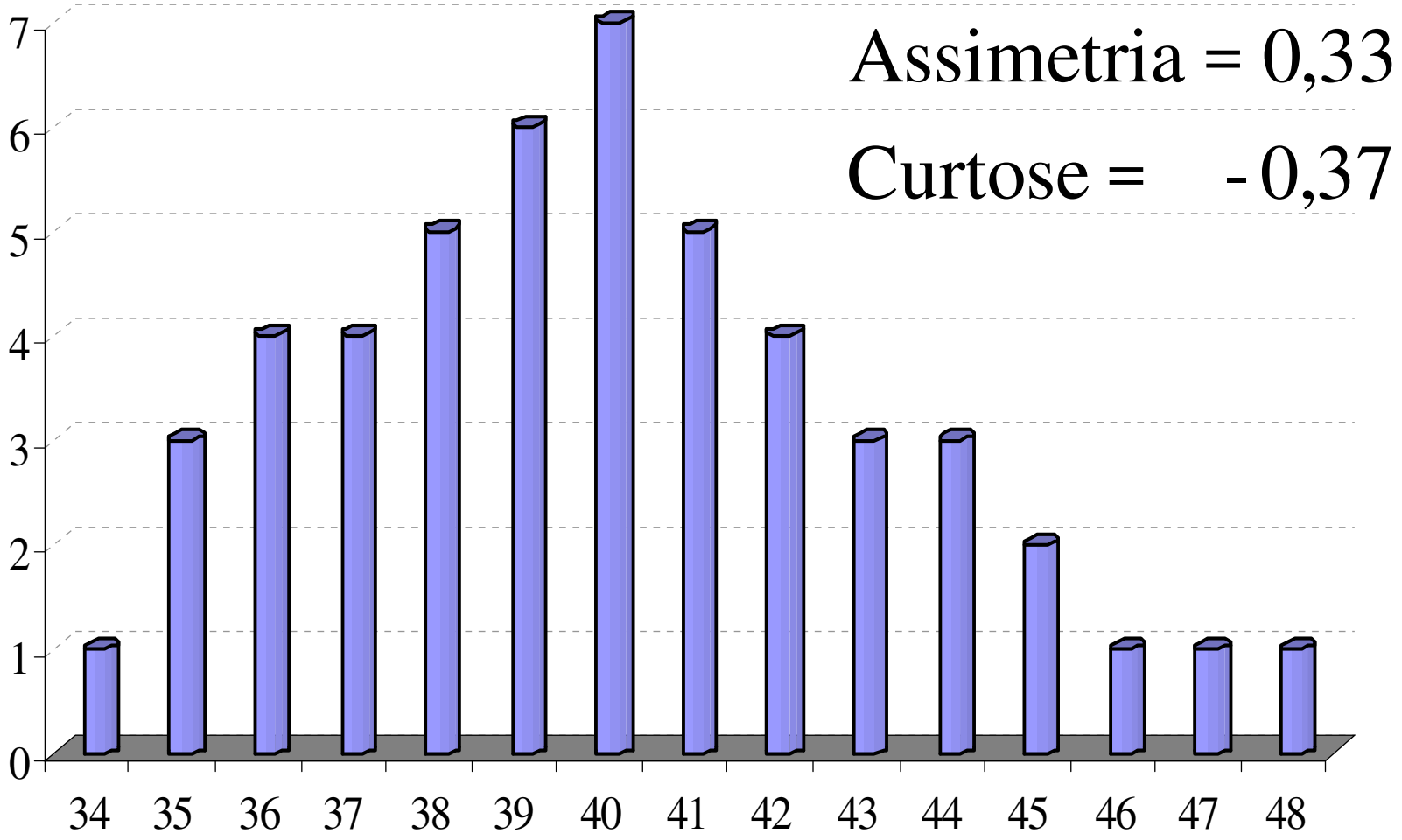
O escore z fornece o número de desvios padrão que cada valor está acima ou abaixo da média. O escore –1,5, significa que este valor está um desvio e meio abaixo da média.

O escore Z é também uma variável, que é obtida pela transformação da amostra original. Ela apresenta média igual a zero e desvio padrão igual a um.

Exemplo

Considere o seguinte amostra:

36	39	38	41	45	44	35	48	35	40
40	40	36	41	37	38	37	39	39	44
42	42	39	43	42	41	39	41	35	40
44	36	40	37	40	36	39	47	40	43
34	45	38	42	46	41	43	37	38	38

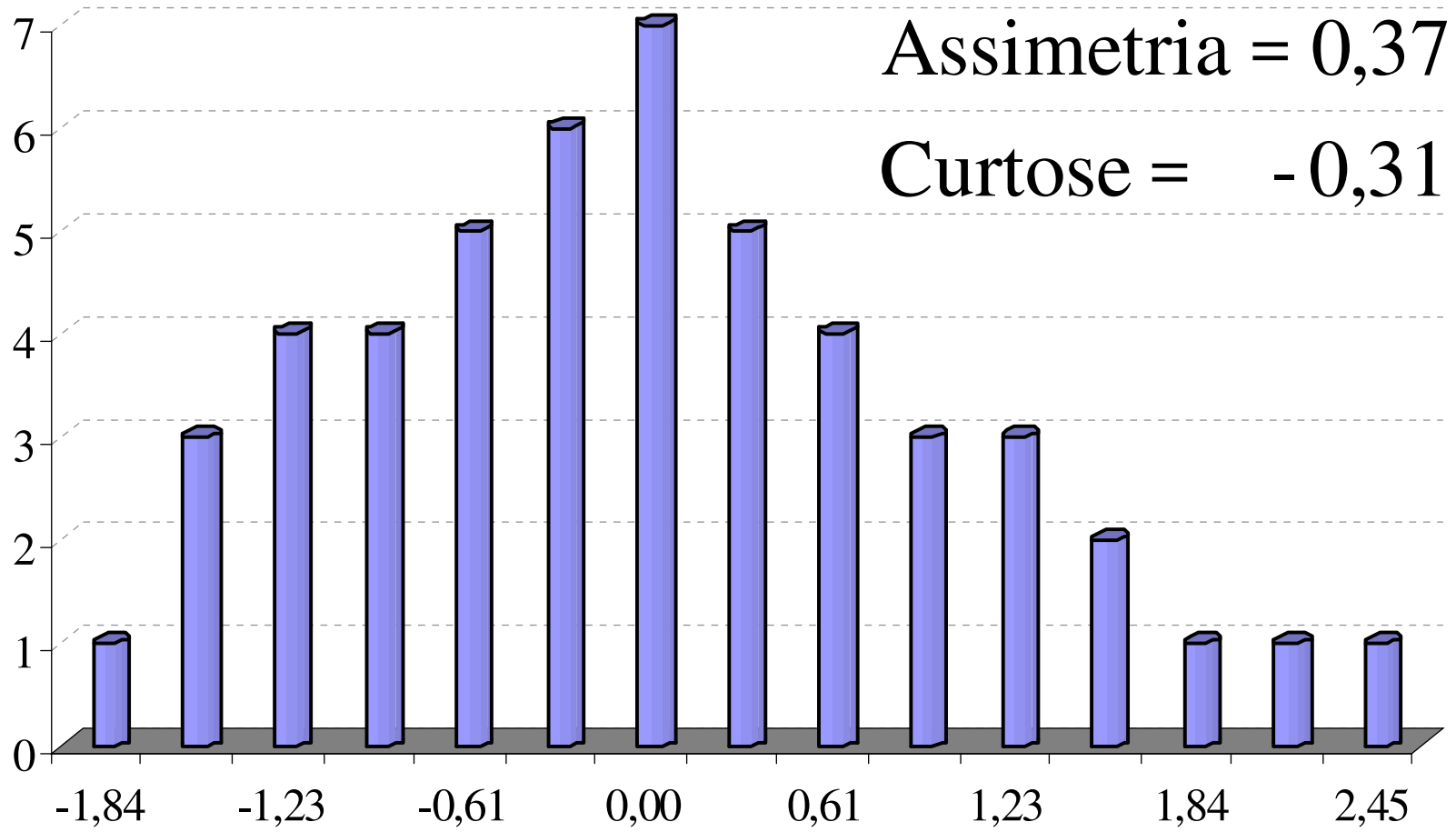

Assimetria = 0,33
Curtose = -0,37
0
1
2
3
4
5
6
7
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

Calcular os escores "**z**" para cada valor da amostra. Representar os valores da amostras e os escores em diagramas para verificar se houve alteração no formato da distribuição dos dados.

Solução:

A média e o desvio padrão da amostra são: 40 e 3,2619. Então os escores padronizados serão:

0,3066	0,9197	-0,9197	-0,6131	-0,6131
-1,2263	-0,3066	-0,6131	0,3066	1,5328
1,2263	-1,5328	2,4526	-1,5328	0,0000
0,0000	0,0000	-1,2263	0,3066	-0,9197
-0,6131	-0,9197	-0,3066	-0,3066	1,2263
0,6131	0,6131	-0,3066	0,9197	0,6131
0,3066	-0,3066	0,3066	-1,5328	0,0000
1,2263	-1,2263	0,0000	-0,9197	0,0000
-1,2263	-0,3066	2,1460	0,0000	0,9197
-1,8394	1,5328	-0,6131	0,6131	1,8394


Assimetria = 0,37
Curtose = -0,31
7
6
5
4
3
2
1
0
-1,84
-1,23
-0,61
0,00
0,61
1,23
1,84
2,45

Propriedades

- A média do escore padronizado é zero;
- O desvio padrão do escore padronizado é um.
- A forma da distribuição do escore padronizado é a mesma dos dados originais.

Escalas

O escore Z não é utilizado normalmente da forma como é calculado. É comum a utilização de uma escala linear de transformação. As duas mais utilizadas são:

Escalas

A escala T que é obtida através da seguinte transformação

$$\mathbf{T = 10.Z + 50}$$

A escala "A" que é utilizada nos vestibulares é obtida por:

$$\mathbf{A = 100.Z + 500}$$

Teorema de Chebyshev

O teorema de Chebyshev permite verificar qual é o percentual mínimo de valores de um conjunto de dados que deve estar um "certo número" de desvios em torno da média.

Em qualquer conjunto de dados com desvio padrão "s", pelo menos $(1 - 1/z^2)$ dos valores do conjunto devem estar entre "z" desvios em torno da média, onde "z" é um valor tal que $z > 1$.

Exemplos:

Assim pelo menos:

75% dos valores estão dentro de **z = 2** desvios a partir da média;

89% dos valores estão dentro de **z = 3** desvios a contar da média;

94% dos valores estão dentro de **z = 4** desvios a contar da média.

Graficamente