



Engenharia de Dados Experimentais

Tabelas de Frequência e Medidas

Raquel Guiné



Conteúdo

1. Tipos de medidas
2. Tabelas de frequências em séries simples
3. Organização dos dados em classes
4. Tabela de frequências em séries classificadas
5. Medidas de tendência central
6. Medidas de dispersão
7. Medidas de tendência não central (Quantis)
8. Distribuição normal
9. Medidas de forma
10. Erro padrão

1. TIPOS DE MEDIDAS

Medidas de tendência central ou de localização:

- Média (\bar{X})
- Moda (Mo)
- Mediana (Md ou \tilde{X})

Medidas de tendência não central ou separatrizes:

- Percentil,
- Quartil,
- Decil...

Medidas de dispersão ou variabilidade:

- Desvio padrão (s)
- Variância (s^2)
- Amplitude de variação (AV)
- Coeficiente de Variação ($CV = \frac{s}{\bar{X}} \times 100$)

Medidas de forma:

- Achatamento ou curtose (K)
- Enviezamento (SK)

2. TABELAS DE FREQUÊNCIAS EM SÉRIES SIMPLES

Foi realizada uma colheita de dados (idade de um conjunto de pessoas) em que se obteve uma série:

19, 20, 21, 20, 19, 20, 22, 23, 21, 20

19, 18, 20, 25, 23, 22, 21, 19, 20, 22

22, 20, 18, 25, 24, 20, 22, 20, 18, 19

caos estatístico.

1º passo: Ordenar os dados

2.º passo: Condensar os dados numa tabela de frequências

Elementos obrigatórios a apresentar numa tabela de frequências

Série não
classificada

X (idade)	F	f	%	Fac	fac	$fac\%$	F.X
18	3	0,10	10,0	3	0,10	10	54
19	5	0,16	16,0	8	0,26	26	95
20	9	0,30	30,0	17	0,56	56	180
21	3	0,10	10,0	20	0,66	66	63
22	5	0,16	16,0	25	0,82	82	110
23	2	0,06	6,0	27	0,88	88	46
24	1	0,03	3,0	28	0,91	91	24
25	2	0,06	6,0	30	1,00	100	50
N	30	1,00	100,0				622

A soma pode não dar exatamente 1,00 por causa dos arredondamentos

X – Variável idade (quantitativa, intervalar)

F – número de vezes que o valor da variável se repete (Frequência absoluta)

N – Efetivo da amostra: $N = \sum F$

f – frequência relativa: $f = \frac{F}{N}$

O somatório das frequências relativas é igual à unidade

% – frequência percentual: $\% = \frac{F}{N} \times 100$

O somatório das frequências percentuais é igual a cem

Elementos obrigatórios a apresentar numa tabela de frequências

Série não
classificada

X (idade)	F	f	%	Fac	fac	$fac\%$	F.X
18	3	0,10	10,0	3	0,10	10	54
19	5	0,16	16,0	8	0,26	26	95
20	9	0,30	30,0	17	0,56	56	180
21	3	0,10	10,0	20	0,66	66	63
22	5	0,16	16,0	25	0,82	82	110
23	2	0,06	6,0	27	0,88	88	46
24	1	0,03	3,0	28	0,91	91	24
25	2	0,06	6,0	30	1,00	100	50
N	30	1,00	100,0				622

A soma pode não dar exatamente 1,00 por causa dos arredondamentos

Fac – frequência absoluta acumulada

$$Fac_i = Fac_{i-1} + F_i$$

$$Fac \text{ (última)} = N$$

fac – frequência relativa acumulada

$$fac_i = fac_{i-1} + f_i$$

$$fac \text{ (última)} = 1$$

Limites da série

As séries têm limites

li (limite inferior) – corresponde ao menor valor da série

Li (limite superior) – corresponde ao maior valor da série

No nosso exemplo:

$li = 18 \text{ anos}$ \Rightarrow limite inferior da série e limite inferior da 1ª classe

$Li = 25 \text{ anos}$ \Rightarrow limite superior da série e limite superior da última classe

A diferença entre o limite superior da série menos o limite inferior da série é a amplitude de variação, ou amplitude total (uma medida de dispersão)

Amplitude de variação

AV (Amplitude de variação) = Limite superior da série – Limite inferior da série

$$AV = 25 - 18 = 7$$

3. ORGANIZAÇÃO DOS DADOS EM CLASSES

As classes são agrupamentos de dados que permitem sintetizar os dados.
São intervalos de variação de uma variável.

Limites das classes

Uma classe i tem sempre um limite inferior designado por li e um limite superior designado por Li

Amplitude de classe

h (Amplitude de classe) = $h = Li - li$ (é sempre igual em todas as classes)

Existem formas automáticas de cálculo do nº de classes e da amplitudes das classes, mas por vezes essa divisão obedece a critérios definidos pelo investigador, e a amplitude de classe pode até ser igual.

Exemplos de classes não uniformes:

Classification	BMI(kg/m ²)	
	Principal cut-off points	Additional cut-off points
Underweight	<18.50	<18.50
Severe thinness	<16.00	<16.00
Moderate thinness	16.00 - 16.99	16.00 - 16.99
Mild thinness	17.00 - 18.49	17.00 - 18.49
Normal range	18.50 - 24.99	18.50 - 22.99
		23.00 - 24.99
Overweight	≥25.00	≥25.00
Pre-obese	25.00 - 29.99	25.00 - 27.49
		27.50 - 29.99
Obese	≥30.00	≥30.00
Obese class I	30.00 - 34.99	30.00 - 32.49
		32.50 - 34.99
Obese class II	35.00 - 39.99	35.00 - 37.49
		37.50 - 39.99
Obese class III	≥40.00	≥40.00

Source: Adapted from WHO, 1995, WHO, 2000 and WHO 2004.

4. TABELA DE FREQUÊNCIAS EM SÉRIES CLASSIFICADAS

Agrupamos os nossos dados do exercício anterior em classes de limites reais.

Série classificada

X (idade)	F	f	%	Fac	X'	F.X'
[18-20[8	0,27	27,00	8	19 *	152 \$
[20-22[12	0,40	40,00	20	21 #	252
[22-24[7	0,23	23,00	27	23	161
[24-26[3	0,10	10,00	30	25	75
N	30	1,00	100,00			640

Fechado à esquerda e aberto à direita

Os significados de **F**, **f**, **%** e **Fac** são iguais ao caso anterior, i.e., na série não classificada

X' – ponto médio ou marca da classe

$$X'(\text{pontomediodaclasse}) = \frac{\text{Limite superior classe} + \text{Limite inferior classe}}{2}$$

$$* (18+20)/2 = 19$$

Série
classificada

X (idade)	F	f	%	Fac	X'	F.X'
[18-20[8	0,27	27,00	8	19 *	152 \$
[20-22[12	0,40	40,00	20	21 #	252
[22-24[7	0,23	23,00	27	23	161
[24-26[3	0,10	10,00	30	25	75
N	30	1,00	100,00			640

Fechado à esquerda e aberto à direita

Depois de calcular o primeiro ponto médio ou marca, podemos obter os outros pontos somando o h (amplitude da classe)

$$\# 19 + 2 = 21$$

Cálculo de **F.X'**

$$\$ 19 * 8 = 152$$

5. MEDIDAS DE TENDÊNCIA CENTRAL

As medidas de tendência central ou medidas de localização são **Média / Moda / Mediana**, e permitem ver como os valores se aproximam de um valor médio.

\bar{X} – Média

Mo – Moda

\check{X} – Mediana (também pode aparecer como - Md)

Média

Há várias médias (aritmética ponderada, geométrica,...), mas usa-se com mais frequência a média aritmética simples.

$$\bar{X}(\text{idades da turma}) = \frac{\text{soma de todas as idades}}{30} \text{ ou seja } \bar{X} = \frac{\sum Xi}{N}$$

Mas se os dados já estiverem agrupados por frequências:

$$\bar{X} = \frac{\sum (Xi.Fi)}{N}$$

No nosso exemplo:

X (idade)	F	f	%	Fac	fac	fac %	F.X
18	3	0,10	10,0%	3	0,10	10%	54*
19	5	0,16	16,0%	8	0,26	26%	95
20	9	0,30	30,0%	17	0,56	56%	180
21	3	0,10	10,0%	20	0,66	66%	63
22	5	0,16	16,0%	25	0,82	82%	110
23	2	0,06	6,0%	27	0,88	88%	46
24	1	0,03	3,0%	28	0,91	91%	24
25	2	0,06	6,0%	30	1,00	100%	50
N	30	1	100,0%				622

*18x3=54

$$\bar{X} \text{ (Série não classificada)} = \frac{622}{30} = 20,73 \text{ anos} \cong 21 \text{ anos}$$

Ponto médio da classe

X (idade)	F	f	%	Fac	X'	F.X'
18-20	8	0,27	27,00	8	19	152*
20-22	12	0,40	40,00	20	21	252
22-24	7	0,23	23,00	27	23	161
24-26	3	0,10	10,00	30	25	75
N	30	1,00	100,00			640

*8x19=152

$$\bar{X} \text{ (série classificada)} = \frac{640}{30} = 21,33 \approx 21 \text{ anos}$$

Mediana

- \tilde{X} é o valor central, ou seja, divide a amostra ao meio.
- A mediana é a medida de tendência central que divide a série em duas partes iguais (50% dos casos estão à esquerda e 50% à direita)

Exemplo (N ímpar):

$$2, 5, \boxed{8}, 9, 10 \Rightarrow \tilde{X} = 8$$

Exemplo 2 (N par):

$$2, 5, \boxed{8, 9}, 10, 11 \Rightarrow \tilde{X} = \frac{8+9}{2} = 8,5$$

Esta operação tem de ser feita com a série ordenada

- Seja N par ou ímpar utiliza-se a seguinte fórmula para determinar o termo

$$\tilde{X} = \frac{N + 1}{2}$$

No nosso exemplo:

$$\tilde{X} = \frac{30+1}{2} = \frac{31}{2} = 15,5 \text{ (esta é a localização da Mediana)}$$

A **mediana** fica localizada no décimo quinto termos e meio.

Vamos às frequências acumuladas e vemos onde fica o décimo quinto termo e meio:
vai cair nas pessoas que têm **20 anos**

18, 18, 18, 19, 19, 19, 19, 19, 20, 20, 20, 20, 20, 20, **20, 20**, 20,



Também podemos consultar as tabelas: nas Fac (frequências acumuladas) vemos que FAC = 17 (i.e., já passou os 15,5) nos 20 anos

Moda

- É o valor da variável que mais vezes se repete; é o valor mais frequente.
- Pode haver mais do que uma moda.
- É pouco usada nas variáveis de natureza quantitativa, é mais usada nos dados qualitativos.

No nosso exemplo:

Verifica-se na tabela na coluna das frequências absolutas qual é o valor mais elevado

$$M_o = 20 , \text{ repete-se } 9 \text{ vezes } (F = 9)$$

Nota final

- A **moda** pode ser utilizada em todos os tipos de variáveis: nominal, ordinal, intervalar ou de razão.
- A **mediana** só pode ser utilizada nos dados de natureza intervalar ou superior: intervalar e de razão
- A **média** também só pode ser utilizada nos dados de natureza intervalar ou de razão.

6. MEDIDAS DE DISPERSÃO

As medidas de dispersão ou de variabilidade são medidas que tendem a ver como os valores se afastam do valor médio central.

AV – Amplitude de variação

s^2 – Variância

s – Desvio padrão

cv – Coeficiente de variação

Amplitude de Variação

- Diferença entre o limite superior e o limite inferior da série.

$$AV = Li - li$$

- Nos dados em classes, é o limite superior da última classe menos o limite inferior da 1ª classe.

No nosso exemplo: $AV = 25 - 18 = 7$

Variância

- A variância é dada pela seguinte fórmula:

$$s^2 = \frac{\sum(F \cdot X^2)}{N - 1} - \bar{X}^2$$

- Para calcular a variância tem de na tabela se criar uma coluna para F. X²

$$s^2 = \frac{13\ 004}{30 - 1} - (20,73)^2 = \frac{13\ 004}{29} - 429,73 = 18,68$$

X (idade)	F	f	%	Fac	fac	fac %	F.X ²
18	3	0,10	10,0%	3	0,10	10%	972 *
19	5	0,16	16,0%	8	0,26	26%	1805
20	9	0,30	30,0%	17	0,56	56%	3600
21	3	0,10	10,0%	20	0,66	66%	1323
22	5	0,16	16,0%	25	0,82	82%	2420
23	2	0,06	6,0%	27	0,88	88%	1058
24	1	0,03	3,0%	28	0,91	91%	576
25	2	0,06	6,0%	30	1,00	100%	1250
N	30	1	100,0%				13 004

*3x18²=972

Desvio padrão

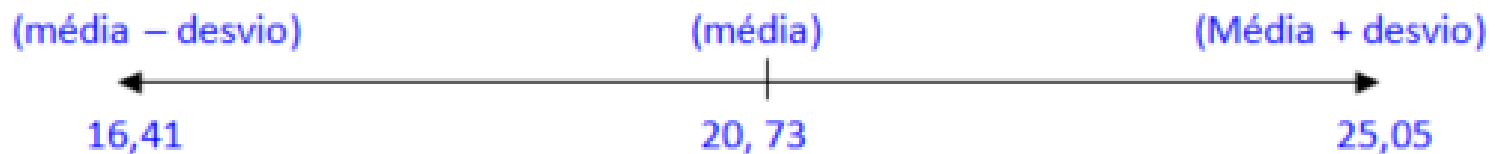
- O desvio padrão é dado por:

$$s = \sqrt{s^2}$$

- No nosso problema dá:

$$s = \sqrt{18,68} = 4,32$$

- Os resultados significam que a média pode oscilar entre 16,41 (média menos o desvio padrão) e 25,09 (média mais o desvio padrão).
- Numa **distribuição normal** a média varia dentro destes intervalos (média menos o desvio padrão) e (média mais o desvio padrão).



- A variância e o desvio padrão são sempre positivos.
- A variância é menos utilizada, sendo mais usado o desvio padrão.

Coeficiente de variação

O coeficiente de variação é dado por:

$$CV = \frac{s}{\bar{X}} \times 100 \text{ (em percentagem)}$$

- No nosso problema dá:

$$CV = \frac{4,32}{20,73} \times 100 = 20,84\%$$

Classificação dos CV:

➤ <i>Dispersão baixa</i>	➡	$CV < 15\%$
➤ <i>Dispersão moderada</i>	➡	$15\% \leq CV \leq 30\%$
➤ <i>Dispersão elevada</i>	➡	$CV > 30\%$

- *O CV pode ultrapassar os 100% quando o desvio padrão é muito alto e a média é baixa.*

7. MEDIDAS DE TENDÊNCIA NÃO CENTRAL (QUANTIS)

- Separatrizes ou medidas de tendência não central ou quantis, são medidas que dividem a série em várias partes iguais

Quantis

- ✓ Quartis – divide a série em 4 partes iguais, 1 quartil = 25%
- ✓ Quintis – divide a série em 5 partes iguais, 1 quintil = 20%
- ✓ Decis – divide a série em 10 partes iguais, 1 decil = 10%
- ✓ Centis/percentis – divide a série em 100 partes iguais, 1 centil = 100%



$$\text{Intervalo interquartílico (IQ)} = Q_3 - Q_1 = 75\% - 25\% = 50\%$$

Para o nosso exemplo:

X (idade)	F	f	%	Fac	fac	fac %	F.X ²
18	3	0,10	10,0%	3	0,10	10%	972 *
19	5	0,16	16,0%	8	0,26	26%	1805
20	9	0,30	30,0%	17	0,56	56%	3600
21	3	0,10	10,0%	20	0,66	66%	1323
22	5	0,16	16,0%	25	0,82	82%	2420
23	2	0,06	6,0%	27	0,88	88%	1058
24	1	0,03	3,0%	28	0,91	91%	576
25	2	0,06	6,0%	30	1,00	100%	1250
N	30	1	100,0%				13 004

*3x18²=972

a) Calcular o terceiro quartil

Localização do termo: $lQ3 = \frac{i \cdot \sum F}{4} = \frac{3 \cdot 30}{4} = 22,5$

Quando a frequência acumulada é 25 (Fac = 25) esta contém o termo com a posição 22,5, o que significa que o valor do 3º quartil corresponde ao valor da variável (idade) para essa frequência:

$$Q3 = 22 \text{ anos}$$

Assim, 75% da amostra tem idade inferior a 22 anos (< 22)

25% tem idade superior ou igual a 22 anos (≥ 22)

b) Calcular o decil 6

Localização do termo: $lD6 = \frac{i \cdot \sum F}{10} = \frac{6 \cdot 30}{10} = 18$

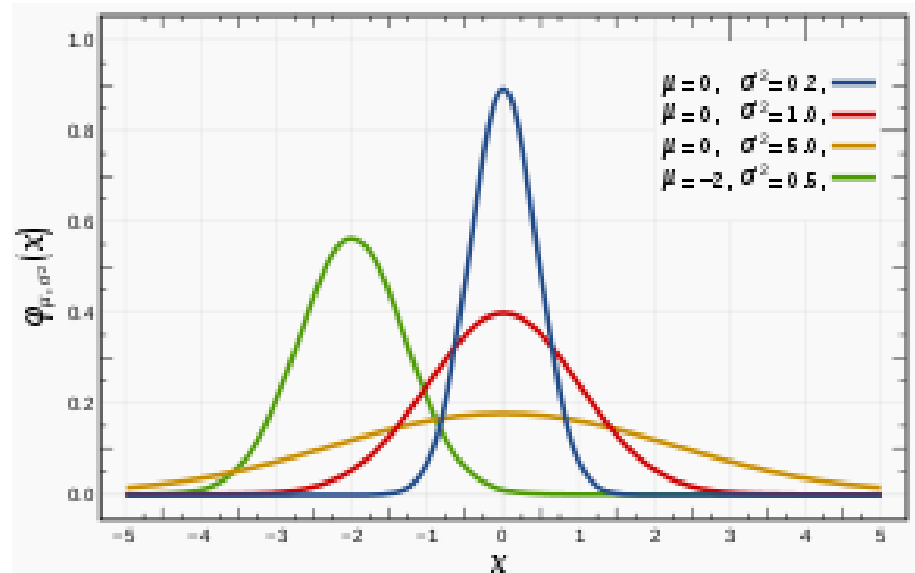
O termo 18 situa-se na Fac = 20, ou seja D6 = 21 anos.

Assim, 60% < 21 anos e

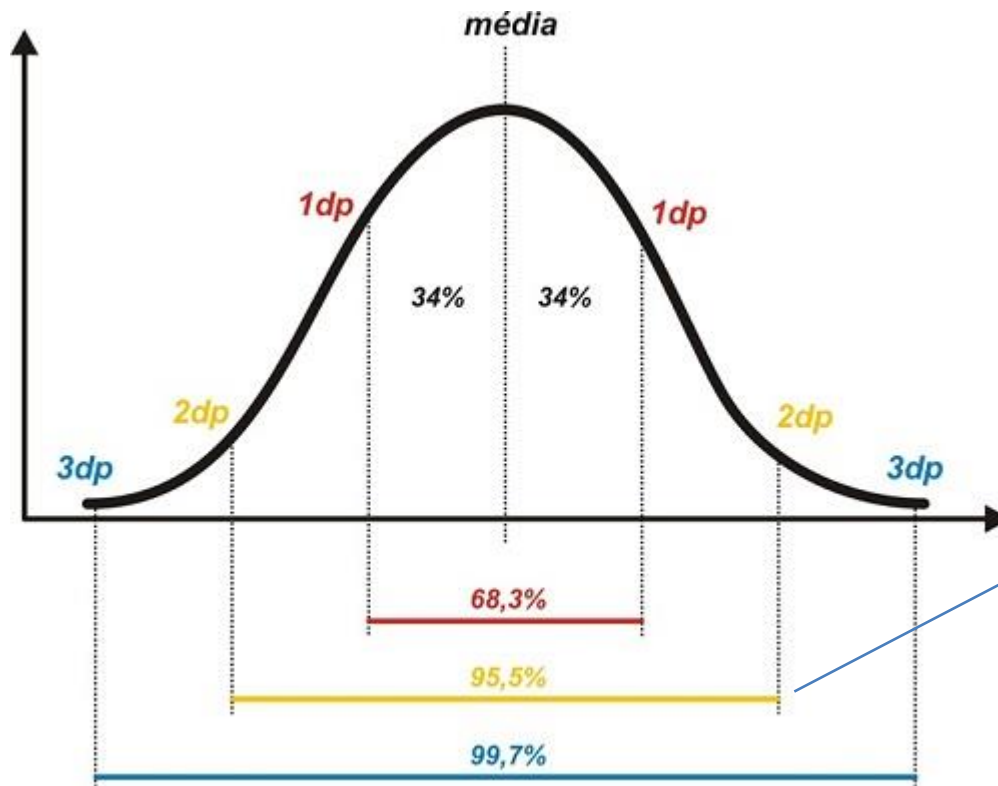
40% tem idade \geq 21 anos

8. DISTRIBUIÇÃO NORMAL

- A distribuição normal conhecida também como distribuição gaussiana é a mais importante distribuição contínua.
- Diversos fenômenos físicos e sociais tem como resultado uma distribuição normal.
- Mesmo que os dados não sejam distribuídos segundo uma normal a média dos dados converge para uma distribuição normal conforme o número de dados aumenta.
- A distribuição é normal quando tem a forma de "sino"
- A distribuição normal com média nula e desvio padrão unitário é chamada de distribuição normal centrada e reduzida ou **distribuição normal padrão.**

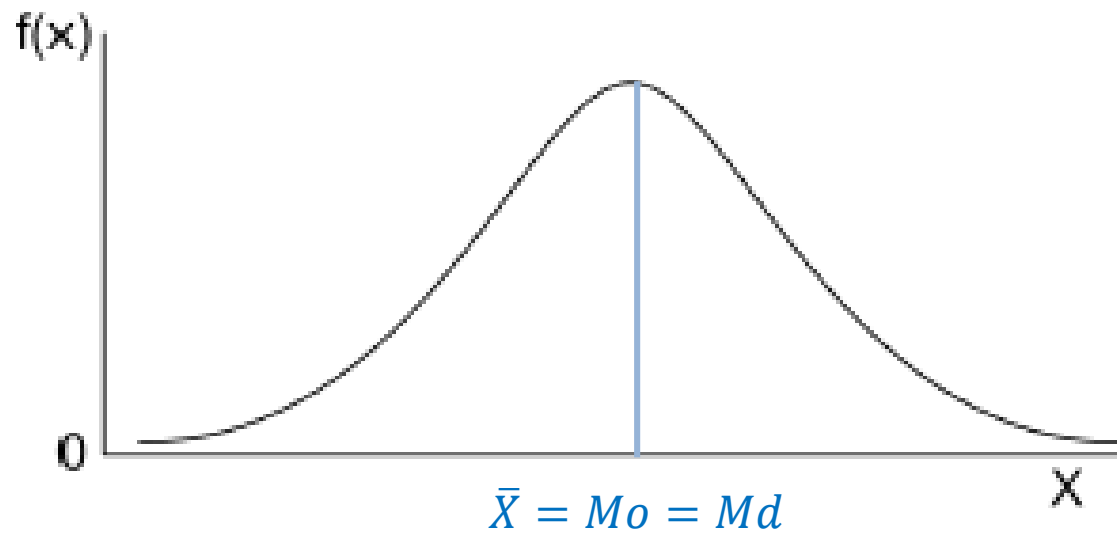


- Os dois parâmetros importantes caracterizadores da DN são a média μ e o desvio padrão σ .
- Quando μ e σ são desconhecidos (caso mais comum), estes valores são estimados por \bar{X} (média) e s (desvio padrão), respectivamente, a partir da amostra.



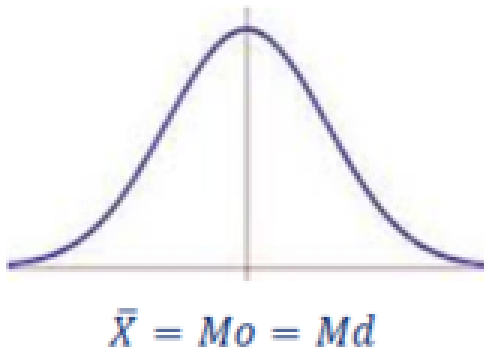
*95% de confiança
 (significância de 5%) é
 normalmente usado em
 testes estatísticos*

Para a distribuição normal: Média = moda = mediana



9. MEDIDAS DE FORMA

- As medidas de forma são duas: a curtose (K) e o enviezamento (SK), e medem o desvio em relação à distribuição normal.
- Para a distribuição normal = curva de Gauss = curva Gaussiana os valores de K e SK encontram-se no intervalo $]-2 ; +2[$, e esta denomina-se por curva mesocurtica ou curva normocurtica.



Média = moda = mediana

Neste caso, considerando 5% de significância

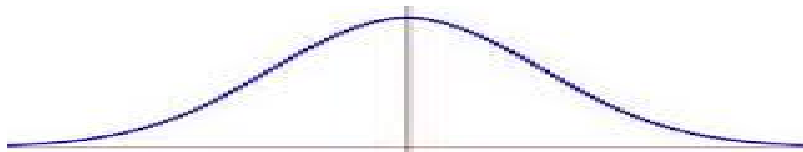
(SPSS): $-2 < K < +2$

$-2 < SK < +2$

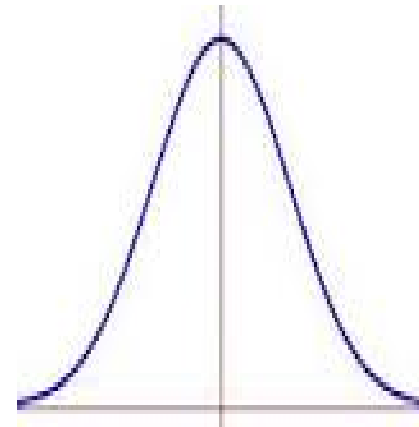
- As medidas de forma (K e SK) são significativas nos seguintes termos:
 - ✓ valor absoluto $> 1,96$ é significativo com $p < 0,05$
 - ✓ valor absoluto $> 2,58$ é significativo com $p < 0,01$
 - ✓ valor absoluto $> 3,29$ é significativo com $p < 0,001$

Curtose (K)

- Mede o achatamento
- O valor absoluto da curtose não tem limite superior, podendo ser tão elevado quanto possível.
- O limite inferior depende do nível de significância considerado, mas a 5% é 2.
- Se só houver curtose, verifica-se na mesma a condição: **Média = moda = mediana**



Curva platycurtica
Dados muito afastados
Neste caso: $K \leq -2$ (ex: $K = -5$)



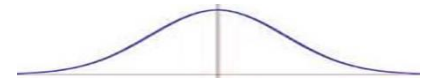
Curva leptocurtica
Dados muito concentrados
Neste caso: $K \geq +2$ (ex: $K = 4$)

Fórmula de cálculo manual da curtose:

$$Curtose = \frac{Q_3 - Q_1}{2(D_9 - D_1)}$$

Quando calculada por esta fórmula manual, a interpretação dos resultados é diferente do SPSS, e por isso:

❑ Se curtose > 0,263 ➡ a curva é platicurtica



❑ Se curtose = 0,263 ➡ a curva é normocurtica

❑ Se curtose < 0,263 ➡ a curva é leptocurtica

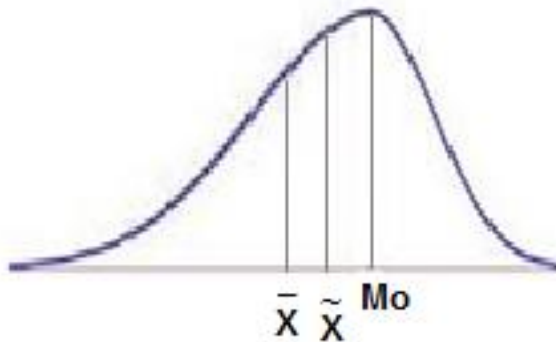


Enviesamento ou assimetria (SK)

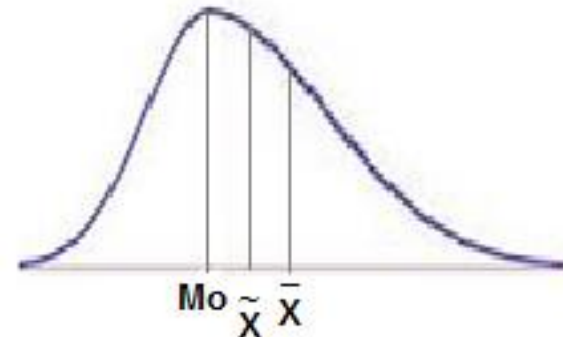
- Mede a assimetria em relação à média.
- Se houver assimetria significativa, **não** se verifica a condição:

$$\text{Média} = \text{moda} = \text{mediana}$$

- Os limites para o enviesamento são os mesmos considerados para a curtose.



Curva enviesada à direita
Média deslocada à esquerda
Enviesamento negativo ($SK \leq -2$)



Curva enviesada à esquerda
Média deslocada à direita
Enviesamento positivo ($SK \geq +2$)

NOTA: Pode haver cruzamento entre os dois fatores.
Ex: Curva platycurtica e enviesada à esquerda.

No nosso exemplo:

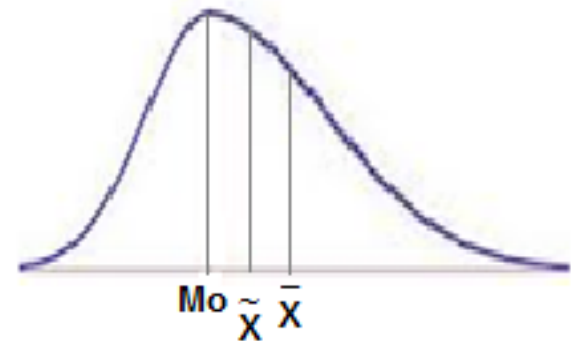
X (idade)	F	f	%	Fac	fac	fac %	F.X ²
18	3	0,10	10,0%	3	0,10	10%	972 *
19	5	0,16	16,0%	8	0,26	26%	1805
20	9	0,30	30,0%	17	0,56	56%	3600
21	3	0,10	10,0%	20	0,66	66%	1323
22	5	0,16	16,0%	25	0,82	82%	2420
23	2	0,06	6,0%	27	0,88	88%	1058
24	1	0,03	3,0%	28	0,91	91%	576
25	2	0,06	6,0%	30	1,00	100%	1250
N	30	1	100,0%				13 004

* $3 \times 18^2 = 972$

$$\bar{X} \text{ (média)} = 21$$

$$\tilde{X} \text{ (mediana)} = 20$$

$$Mo \text{ (moda)} = 20$$



Logo a curva é enviesada à esquerda

No nosso exemplo:

$$Curtose = \frac{22-19}{2(23-18)} = 0,3$$

A curva é platicurtica.

X (idade)	F	f	%	Fac	fac	fac %	F.X ²
18	3	0,10	10,0%	3	0,10	10%	972 *
19	5	0,16	16,0%	8	0,26	26%	1805
20	9	0,30	30,0%	17	0,56	56%	3600
21	3	0,10	10,0%	20	0,66	66%	1323
22	5	0,16	16,0%	25	0,82	82%	2420
23	2	0,06	6,0%	27	0,88	88%	1058
24	1	0,03	3,0%	28	0,91	91%	576
25	2	0,06	6,0%	30	1,00	100%	1250
N	30	1	100,0%				13 004

* 3x18²=972

Localização dos termos

Termos

$$lQ3 = \frac{i \cdot \sum F}{4} = \frac{3 \cdot 30}{4} = 22,5$$

Q3 = 22 anos (Fac = 25 > 22,5)

$$lQ1 = \frac{i \cdot \sum F}{4} = \frac{1 \cdot 30}{4} = 7,5$$

Q1 = 19 anos (Fac = 8 > 7,5)

$$lD9 = \frac{i \cdot \sum F}{10} = \frac{9 \cdot 30}{10} = 27$$

D9 = 23 anos (Fac = 27)

$$lD1 = \frac{i \cdot \sum F}{10} = \frac{1 \cdot 30}{10} = 3$$

D1 = 18 anos (Fac = 3)

10. ERRO PADRÃO

- O erro padrão ou erro amostral mede o erro que se comete ao trabalhar com a média, isto é, mede o quanto a amostra é representativa da população

$$E = \frac{s}{\sqrt{n}}$$

Para o nosso exemplo:

$$E = \frac{4,32}{\sqrt{30}} = 0,79$$

- *É um erro elevado, pois a amostra é pequena (Ter um erro de 0,79 numa amostra de 30 pessoas é grande!)*
- *O erro diminui à medida que aumenta o tamanho da amostra (Quanto mais próxima estiver a amostra da população, menor é o erro amostral)*

FIM