Prof. MSc. Marcos Alexandruk

E-mail: alexandruk@uni9.pro.br

https://github.com/alexandruk/analisededados

Estatística

Estatística

A palavra estatística tem origem no latim "**status**" e relaciona-se com "**estado**".

No início, a palavra era usada para se referir ao "**cidadão político**".

Posteriormente, passou a ser utilizada em alemão com o sentido de "**conjunto de dados do Estado**", de onde decorre o seu significado desde o século XIX.

BATISTA, Carolina. Estatística. Toda Matéria, 2021. Disponível em: https://www.todamateria.com.br/estatistica-conceito-fases-metodo/. Acesso em: 23/02/2021.

Estatística

“Estatística é uma ciência exata que estuda a coleta, a organização, a análise e registro de dados por amostras.

Utilizada desde a Antiguidade, quando se registravam os nascimentos e as mortes das pessoas, é um método de pesquisa fundamental para tomar decisões. Isso porque fundamenta suas conclusões nos estudos realizados.”

BATISTA, Carolina. Estatística. Toda Matéria, 2021. Disponível em: https://www.todamateria.com.br/estatistica-conceito-fases-metodo/. Acesso em: 23/02/2021.

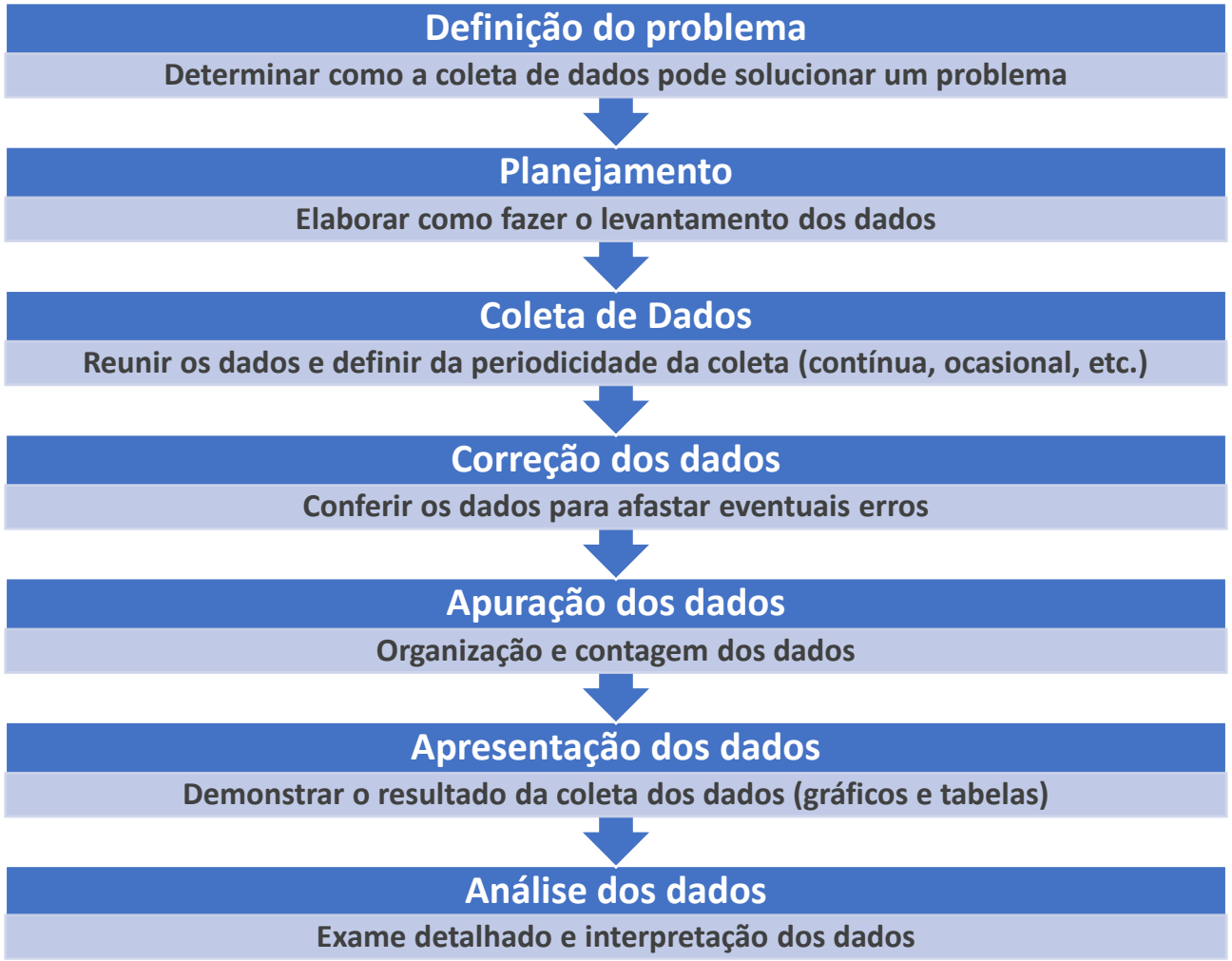
Estatística

Estatística é ciência que tem por fim a **pesquisa e a comparação dos fatos gerais e particulares** verificados no movimento das sociedades.

Objetivo geral da estatística

O objetivo da estatística é a análise e interpretação dos fenômenos sociais de qualquer natureza, para planejamento de ações.

Fases do Método Estatístico



BATISTA, Carolina. Estatística. Toda Matéria, 2021. Disponível em: https://www.todamateria.com.br/estatistica-conceito-fases-metodo/. Acesso em: 23/02/2021. (adaptado)

Importância da Estatística na Engenharia

A probabilidade e estatística pode contribuir para mitigação dos erros e favorecer a análise de um projeto em construção, considerando as mais diversas situações, de forma que dados estatísticos podem auxiliar os testes de desempenho e o controle de qualidade.

Análise Descritiva

A análise descritiva dos dados se limita a calcular algumas medidas de posição e variabilidade, como a média e variância, por exemplo.

Inferência

Inferência estatística é um ramo da Estatística cujo objetivo é fazer afirmações a partir de um conjunto de valores representativo (amostra) sobre um universo.

Em geral, a inferência estatística está associada à coleta, à redução, à análise e à modelagem dos dados.

Tipos de Variáveis

Variáveis qualitativas: apresentam algum tipo de atributo do elemento pesquisado. (educação, estado civil, sexo, etc.)

Variáveis quantitativas: apontam para um impacto no elemento pesquisado e contribuem na análise.

Variáveis quantitativas discretas: Quando podemos expressar as variáveis por um número inteiro em certa contagem, chamamos de variável quantitativa discreta. (número de filhos, quantidade de veículos, etc.)

Variáveis quantitativas contínuas: Quando destacamos uma variável por intermédio de uma medida, chamamos de variável quantitativa contínua. (tempo, temperatura, pressão, etc.)

Distribuições de frequências

No estudo de uma variável, devemos dispor um maior interesse em conhecer a distribuição dessa variável por meio das possíveis realizações dela e dispor seus valores, de modo que se tenha uma boa ideia global dessa distribuição.

Distribuições de frequências

Frequência de porcentagens de 20 empregados segundo o grau de instrução:

Grau de instrução	Contagem	Frequência	Proporção	Porcentagem
1º grau	8	8	0,4	40%
2º grau	7	7	0,35	35%
Superior	5	5	0,25	25%
Total	**20**	**20**	**1,00**	**100%**

Distribuições de frequências

Frequência absoluta acumulada e frequência relativa acumulada:

Grau de instrução	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1º grau	8	40%	8	40%
2º grau	7	35%	8 + 7 = 15	40% + 35% = 75%
Superior	5	25%	15 + 5 = 20	75% + 25% = 100%
Total	**20**	**100%**	**20**	**100%**

As frequências acumuladas são extremamente úteis quando o objetivo é saber a quantidade ou a porcentagem até determinada característica.

Amplitude total

Alturas de 32 crianças de 1 a 4 anos:

73,93	71,51	66,83	64,17	66,16	65,7	64,78	65,81
63,15	62,56	61,88	60,94	60,3	60,15	56,57	55,86
71,47	70,09	64,44	63,27	66,06	65,09	64,73	64,16
62,69	61,91	61,49	60,73	60,24	59,37	56,03	**55,77**

Amplitude total = Valor Máximo – Valor Mínimo

Amplitude total = 73,93 – 55,77

Amplitude total = 18,16

Números de classes

Alturas de 32 crianças de 1 a 4 anos:

73,93	71,51	66,83	64,17	66,16	65,7	64,78	65,81
63,15	62,56	61,88	60,94	60,3	60,15	56,57	55,86
71,47	70,09	64,44	63,27	66,06	65,09	64,73	64,16
62,69	61,91	61,49	60,73	60,24	59,37	56,03	**55,77**

Número de classes = SQRT (n)

Número de classes = SQRT (32)

Número de classes = 5,65 (aproximado para 6)

SQRT => Raiz Quadrada

Amplitude do intervalo

Amplitude do intervalo = Amplitude total / número de classes

Amplitude do intervalo = 18,16 / 6

Amplitude do intervalo = 3,02

Classes	fi	Fi	fr	Fr
55 ⊢	4	4	12,50%	12,50%
58 ⊢	6	10	18,75%	31,25%
61 ⊢	7	17	21,88%	53,13%
64 ⊢	11	28	34,38%	87,50%
67 ⊢	1	29	3,13%	90,63%
70 ⊢ [ERRO]	3	32	9,38%	100%
Total	**32**		**100%**	

Amplitude do intervalo

Amplitude do intervalo = Amplitude total / número de classes

Amplitude do intervalo = 18,16 / 6

Amplitude do intervalo = 3,02

Classes	fi	Fi	fr	Fr
55 ⊢ 58	4	4	12,50%	12,50%
58 ⊢ 61	6	10	18,75%	31,25%
61 ⊢ 64	7	17	21,88%	53,13%
64 ⊢ 67	11	28	34,38%	87,50%
67 ⊢ 70	1	29	3,13%	90,63%
70 ⊢ 73 **[ERRO]**	3	32	9,38%	100%
Total	**32**		**100%**	

[ERRO] O maior valor é 73,93 (está acima de 73)

Amplitude do intervalo

Amplitude do intervalo = Amplitude total / número de classes

Amplitude do intervalo = 18,16 / 6

Amplitude do intervalo = 3,02 (arredondar para 4)

Classes	fi	Fi	fr	Fr
55 ⊢ 59				
59 ⊢ 63				
63 ⊢ 67				
67 ⊢ 71				
71 ⊢ 75				
75 ⊢ 79				
Total	**32**		**100%**	

Regra de Sturges

$k = 1 + 3{,}3 * LOG(n)$

k = Número de classes

n = Total de dados

$k = 1+3{,}3*LOG(20)$

$k = 5{,}293399$ → $k \approx 5$

$A_{Total} = Valor_{Max} - Valor_{Min}$

$A_{Total} = 42 - 15$ → $A_T = 27$

$h = A_{Total}/k$

$h = 27/5$ → $h = 5{,}4$ → $h \approx 6$

h = Amplitude do Intervalo

(Arredondar para cima)

Pesquisa: Idade				
17	18	16	24	23
42	40	36	15	18
26	23	23	24	28
41	16	18	20	27

IDADE	fi
15 ⊢ 21	8
15 ⊢ 27	6
27 ⊢ 33	2
33 ⊢ 39	1
39 ⊢ 45	3

Exercício

SALÁRIOS					
20,50	9,50	15,30	17,20	24,10	19,90
15,40	12,70	7,40	16,50	15,30	26,20
14,90	7,80	23,30	15,90	11,80	18,40
13,40	14,30	16,20	16,70	9,20	16,80
9,80	20,10	17,80	17,10	12,60	15,90

Classes	fi
7,40 ⊢ 10,40	
10,40 ⊢ 13,60	
13,60 ⊢ 16,80	
16,80 ⊢ 20,00	
20,00 ⊢ 23,20	
23,20 ⊢ 26,40	

Amplitude Total (A_{Total}) =	

Total de dados (n) =	

Número de classes (k) =	

Amplitude do intervalo (h) =	

Exercício

SALÁRIOS					
20,50	9,50	15,30	17,20	24,10	19,90
15,40	12,70	7,40	16,50	15,30	26,20
14,90	7,80	23,30	15,90	11,80	18,40
13,40	14,30	16,20	16,70	9,20	16,80
9,80	20,10	17,80	17,10	12,60	15,90

Classes	fi
7,40 ⊢ 10,60	
10,60 ⊢ 13,80	
13,80 ⊢ 17,00	
17,00 ⊢ 20,20	
20,20 ⊢ 23,40	
23,40 ⊢ 26,60	

$k = 1 + 3{,}3 * LOG(n)$

k = Número de classes

n = Total de dados

$A_{Total} = Valor_{Max} - Valor_{Min}$

$h = A_{Total}/k$

h = Amplitude do Intervalo

$k = 1+3{,}3*LOG(30)$

$k = 6$

$A_{Total} = 26{,}20 - 7{,}40$

$h = 18{,}80/6$

$k \approx 6$

$A_T = 18{,}80$

$h = 3{,}13$

$h \approx 3{,}20$

(Arredondar para cima)

Exercício

SALÁRIOS					
20,50	9,50	15,30	17,20	24,10	19,90
15,40	12,70	7,40	16,50	15,30	26,20
14,90	7,80	23,30	15,90	11,80	18,40
13,40	14,30	16,20	16,70	9,20	16,80
9,80	20,10	17,80	17,10	12,60	15,90

Classes	fi
7,40 ⊢ 10,60	5
10,60 ⊢ 13,80	4
13,80 ⊢ 17,00	11
17,00 ⊢ 20,20	6
20,20 ⊢ 23,40	2
23,40 ⊢ 26,60	2
TOTAL VALORES	30

Amplitude Total (A_{Total}) =	18,80

Total de dados (n) =	30

Número de classes (k) =	6

Amplitude do intervalo (h) =	3,20

Exercício

Classes	fi	Fi	fr	Fr
7,40 ⊢ 10,60	5	5	16,66%	16,66%
10,60 ⊢ 13,80	4	9	13,33%	30,00%
13,80 ⊢ 17,00	11	20	36,66%	66,66%
17,00 ⊢ 20,20	6	26	20,00%	86,66%
20,20 ⊢ 23,40	2	28	6,66%	93,33%
23,40 ⊢ 26,60	2	30	6,66%	100,00%
Total	30		100%%	

fi = frequência absoluta

Fi = frequência absoluta acumulada

fr = frequência relativa

Fr = frequência relativa acumulada

Referências

DACHS J. N. W. Análise de dados e regressão. São Paulo: IME USP, 1978.

LEVIN J. Estatística aplicada a Ciências Humanas. São Paulo: Harper e Row do Brasil, 1978.

MORETTIN P. A. Introdução a estatística para ciências exatas. São Paulo: Atual Editora, 1981.

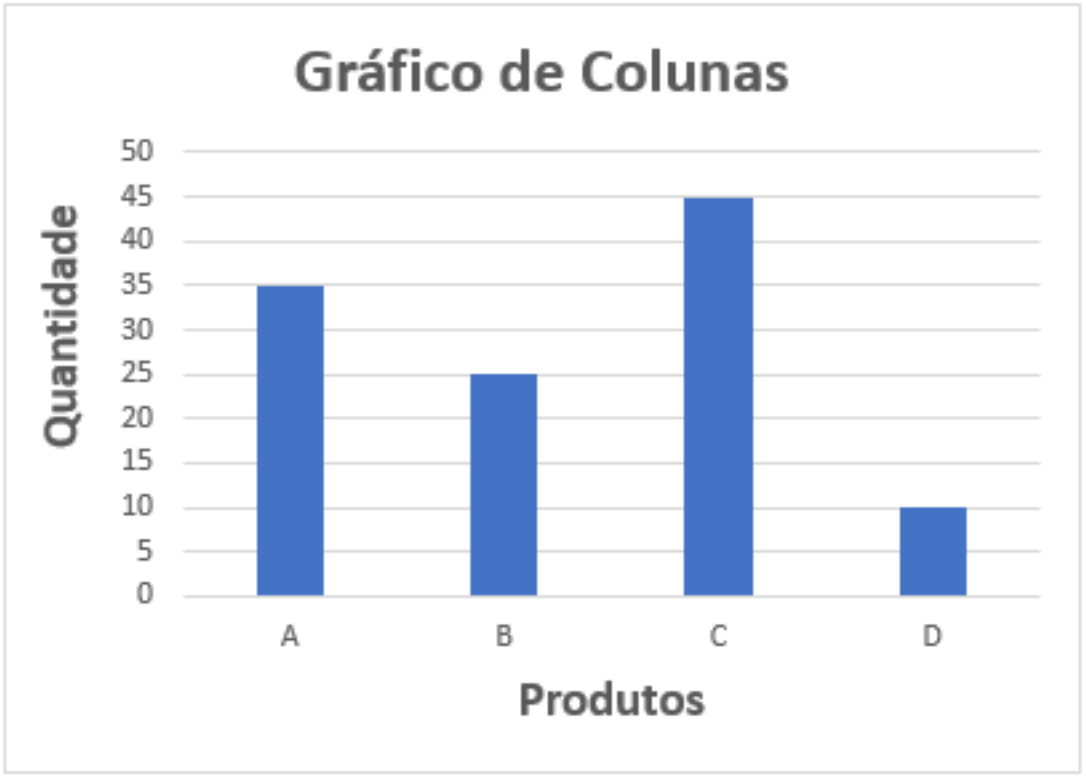
Prof. MSc. Marcos Alexandruk

E-mail: alexandruk@uni9.pro.br

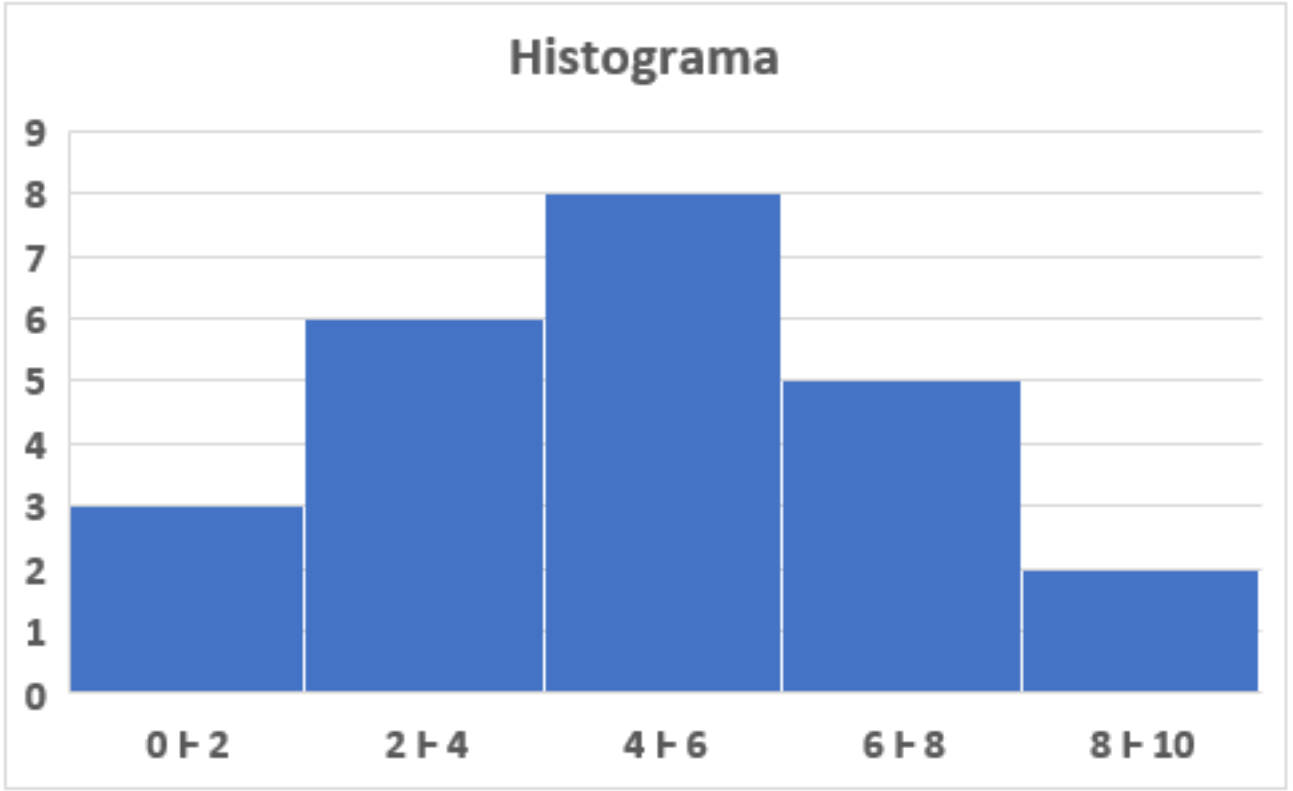
https://github.com/alexandruk/analisededados

Representações gráficas

Produtos	Quantidade
A	35
B	25
C	45
D	10



Classes	Frequências
0 ⊢ 2	3
2 ⊢ 4	6
4 ⊢ 6	8
6 ⊢ 8	5
8 ⊢ 10	2

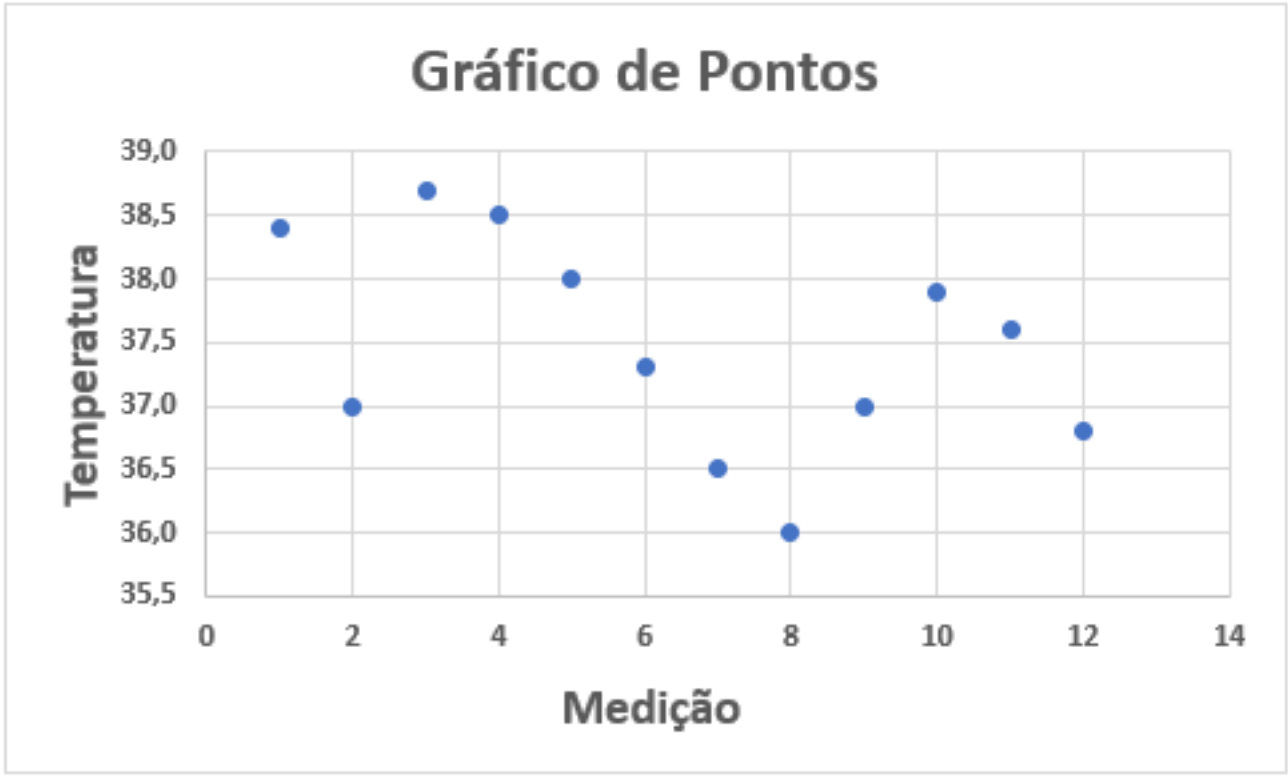


Análise de Dados

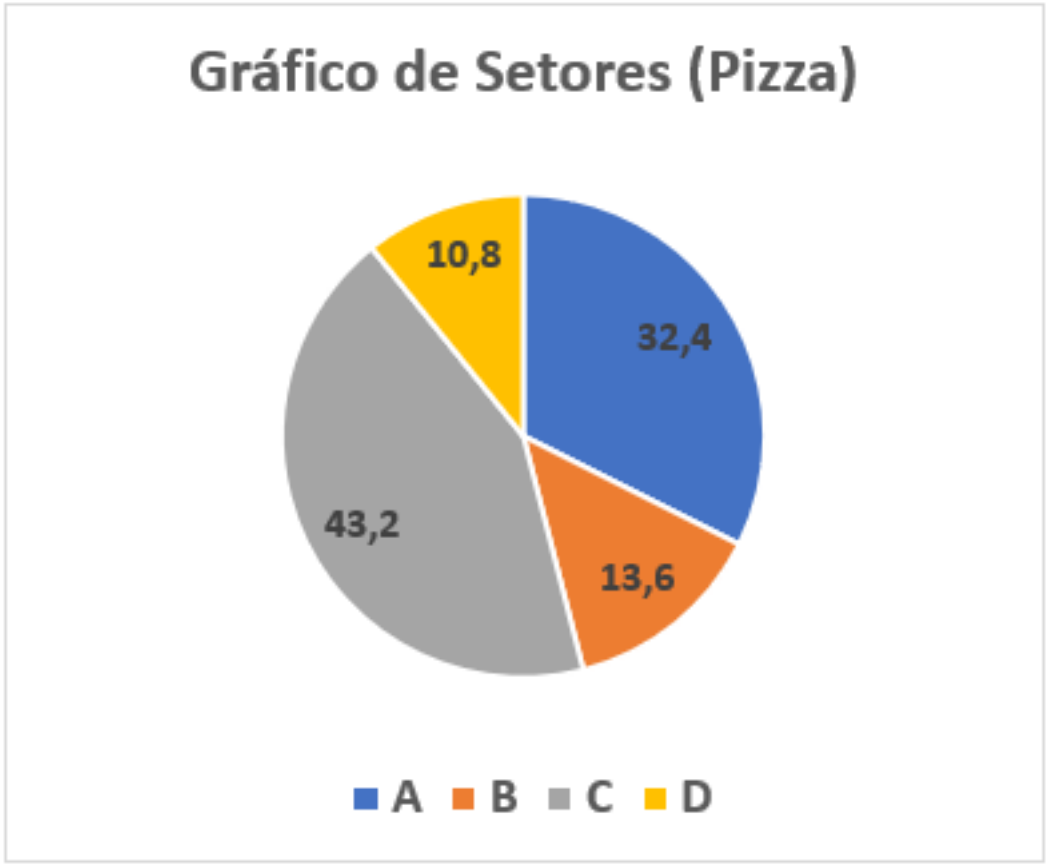
Medição	Temperatura
1	38,4
2	37,0
3	38,7
4	38,5
5	38,0
6	37,3
7	36,5
8	36,0
9	37,0
10	37,9
11	37,6
12	36,8

Análise de Dados

Medição	Temperatura
1	38,4
2	37,0
3	38,7
4	38,5
5	38,0
6	37,3
7	36,5
8	36,0
9	37,0
10	37,9
11	37,6
12	36,8

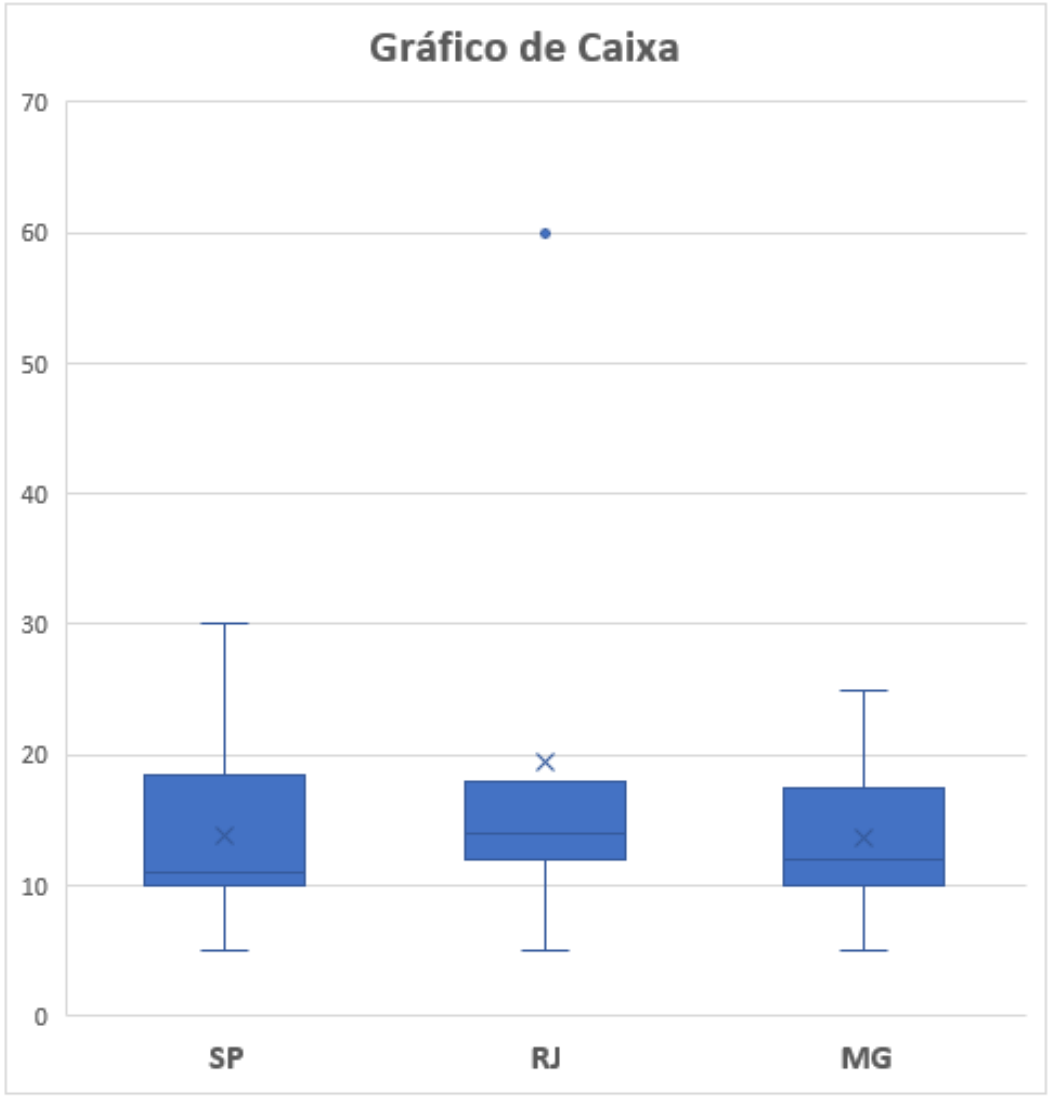


Produtos	Quantidade (%)
A	32,4
B	13,6
C	43,2
D	10,8



Análise de Dados

SP	10
SP	5
SP	30
SP	12
SP	10
SP	20
SP	14
SP	10
RJ	12
RJ	60
RJ	5
RJ	15
RJ	18
RJ	12
RJ	14
MG	10
MG	10
MG	12
MG	5
MG	14
MG	25
MG	12
MG	20
MG	15



Prof. MSc. Marcos Alexandruk

E-mail: alexandruk@uni9.pro.br

https://github.com/alexandruk/analisededados

Medidas de tendência central:
média aritmética; média geométrica; média harmônica

Média aritmética

1º caso: dados não agrupados

A média aritmética dos valores x_1 , x_2 , x_3 , ... , x_n é o quociente entre a soma desses valores e o seu número total n.

$$\bar{x} = \frac{x_1+x_2+x_3+\cdots+x_n}{n} \text{ ou } \bar{x} = \frac{\Sigma x_i}{n}$$

Exemplo: Determinar a média aritmética dos valores: 3, 7, 8, 10 e 11.

$$\bar{x} = \frac{3+7+8+10+11}{5} = 7{,}8$$

Média aritmética

2º caso: dados agrupados sem intervalos

Se os elementos x_1 , x_2 , x_3 , ... , x_n apresentam, respectivamente, frequências f_1 , f_2 ,f_3 , ... , f_n , então:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_n f_n}{n} \text{ ou } \bar{x} = \frac{\Sigma x_i f_i}{n}$$

Exemplo: dada a amostra: 2, 5, 5, 5, 5, 6, 6, 6, 8, 8, a média será:

$$\bar{x} = \frac{2 \,.\, 1 + 5 \,.\, 4 + 6 \,.\, 3 + 8 \,.\, 2}{10} = \frac{56}{10} = 5{,}6$$

x_i	f_i	$x_i f_i$
2	1	2
5	4	20
6	3	18
8	2	16
Total	10	56

Média aritmética

3º caso: dados agrupados com intervalos

Quando os dados estão agrupados, aceita-se, por convenção, que as frequências se distribuam uniformemente ao longo da classe e que, portanto, o seu ponto médio (x) é o valor representativo do conjunto. Então:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_n f_n}{n} \text{ ou } \bar{x} = \frac{\Sigma x_i f_i}{n}$$

Exemplo: dada a amostra conforme a tabela, a média será:

$$\bar{x} = \frac{3{,}5 \,.\, 1 + 6{,}5 \,.\, 10 + 9{,}5 \,.\, 8 + 12{,}5 \,.\, 1}{20} = \frac{157}{20} = 7{,}85$$

Classe	x_i	f_i	$x_i f_i$
2 ⊢ 5	3,5	1	3,5
5 ⊢ 8	6,5	10	65
8 ⊢ 11	9,5	8	76
11 ⊢ 14	12,5	1	12,5
Total		20	157

Média geométrica

A média geométrica de um conjunto de números positivos é definida como o **produto de todos os membros do conjunto elevado ao inverso do número de membros**. Indica a tendência central ou o valor típico de um conjunto de números usando o produto dos seus valores.

A média geométrica é frequentemente utilizada quando comparamos diferentes itens – encontrando uma única "figura representativa" para esses itens – quando cada um desses itens possuem múltiplas propriedades que possuem diferentes escalas numéricas. Por exemplo, a média geométrica pode nos dar uma "média" significativa para comparar duas companhias que estão sendo classificadas numa escala de 0 a 5 para suas sustentabilidades ambientais e sendo classificadas de 0 a 100 para suas viabilidades financeiras. Se a média aritmética fosse usada em vez da média geométrica, a viabilidade financeira pesaria mais pois seu alcance numérico é grande, logo uma pequena mudança percentual na classificação financeira (por exemplo: uma mudança de 80 para 90) faria uma grande diferença na média aritmética do que uma grande diferença percentual na classificação da sustentabilidade ambiental (por exemplo uma mudança de 2 para 5 na escala).

Média geométrica

Sejam x1 , x2 , x3 , ... , xn valores da variável X, associadas, respectivamente, às frequências f1 , f2 , f3 , ... , fn. Então, a média geométrica de x é definida por:

$$M_g = \sqrt[n]{x_1^{f_1} . x_2^{f_2} . x_3^{f_3} . \ldots . x_n^{f_n}}$$

Em particular, se f_1 , f_2 , f_3 , ... , f_n = 1. temos:

$$M_g = \sqrt[n]{x_1 . x_2 . x_3 . \ldots . x_n}$$

Exemplo: Dada a tabela de distribuição de frequências, temos:

$$M_g = \sqrt[22]{1^8 . 2^6 . 3^5 . 5^3} = \sqrt[22]{1944000} = 1{,}93$$

x_i	f_i
1	8
2	6
3	5
5	3
Total	22

Média harmônica

A média harmônica é definida como a quantidade de elementos no conjunto, divida pela soma do inverso dos elementos do conjunto.

A média aritmética é muitas vezes utilizada erroneamente em casos que exigem a média harmônica. Um exemplo é o cálculo da velocidade média em um percurso de ida e volta em uma mesma via, em que a ida é percorrida a 60 km/h e a volta a 40 km/h a média aritmética de 50 está incorreta. A velocidade média no percurso total é a média harmônica de 40 e 60, ou seja 48km/h.

Exemplo:

Distância de A a B = 120 Km | Velocidade média = 40 Km/h | Duração da viagem = 120 /40 = 3 horas

Distância de B a A = 120 Km | Velocidade média = 60 Km/h | Duração da viagem = 120 /60 = 2 horas

Distância total de A a B + de B a A = 120 Km + 120 Km = 240 Km

Duração total da viagem = 3 horas + 2 horas = 5 horas

Velocidade média da viagem (de A a B + de B a A) = 240 / 5 = 48 Km/h

Média harmônica

Se os elementos x_1 , x_2 , x_3 , ... , x_n apresentam, respectivamente, frequências f_1 , f_2 , f_3 , ... , f_n , então, a média harmônica é definida como o inverso da média aritmética do inverso dos valores:

$$M_h = \frac{n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} + \cdots + \frac{f_n}{x_n}} = \frac{n}{\sum \frac{f_i}{x_i}}$$

Em particular, se f_1 , f_2 , f_3 , ... , f_n = 1. temos:

$$M_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_i}}$$

Exemplo: Dada a tabela de distribuição de frequências, temos:

$$M_h = \frac{22}{\frac{8}{1} + \frac{6}{2} + \frac{5}{3} + \frac{3}{5}} = \frac{22}{\frac{398}{30}} = 22.\frac{30}{398} = \frac{330}{199} = 1{,}66$$

x_i	f_i
1	8
2	6
3	5
5	3
Total	22

Prof. MSc. Marcos Alexandruk

E-mail: alexandruk@uni9.pro.br

https://github.com/alexandruk/analisededados

Medidas de tendência central: moda e mediana

Objetivo

Calcular e interpretar as medidas de tendência central: a **moda** e a **mediana** de uma distribuição, destacando as suas diferenças e usos.

Média, moda e mediana

Apesar de ser bastante utilizada a média aritmética, nem sempre é a medida mais adequada para se analisar um agrupamento de dados.

Veja o exemplo. Numa certa empresa com 200 empregados, os salários são os seguintes:

Salários (em salários mínimos)	Número de empregados
1	100
2	30
3	30
4	5
5	25
10	5
25	3
40	2

Calculando o salário médio desses empregados, obtemos 3 salários mínimos.

Este número está correto do ponto de vista aritmético, mas não é representativo da condição salarial da maioria dos empregados. Afinal, 130 (65% do total) deles, ganham menos do que este valor. Por outro lado, de acordo com a tabela, 5 empregados (2,5%) ganham mais do que 20 salários mínimos, o que "puxa" a média para cima.

Neste caso, é mais conveniente usarmos outro tipo de medida como valor representativo do salário dos empregados, conforme veremos nesta aula.

Moda

Dada uma coleção de números, a moda é o valor que ocorre com **maior frequência**.

Assim, no exemplo citado, o salário mais frequente é o salário mínimo, que é recebido por 100 empregados, isto é, 1 salário mínimo.

Observações:

- Existem casos em que a moda não existe — os valores não se repetem ou todos os valores têm a mesma frequência (distribuição amodal).
- Em alguns casos, pode haver mais de uma moda, ou seja, a distribuição dos valores pode ser bimodal, trimodal, etc.

Moda (M_o)

1º Caso: dados não agrupados

É o valor de maior frequência ou que aparece mais vezes em um conjunto de dados.

Exemplo: **7, 8, 8, 9, 10, 10, 10, 12, 15**

O elemento de maior frequência é o 10, que aparece três vezes.

Portanto $\mathbf{M_o = 10}$ (distribuição unimodal).

Exemplo: **3, 5, 8, 10, 12 e 13**

Todos os elementos da série apresentam a mesma frequência, logo, a série é **amodal**.

Exemplo: **2, 2, 5, 5, 8, 9**

Os elementos 2 e 5 têm frequência 2. Logo, temos $\mathbf{M_o = 2}$ **e** $\mathbf{M_o = 5}$ (distribuição **bimodal**).

Moda (M_o)

2º Caso: dados agrupados sem intervalos

Basta identificar o elemento de maior frequência.

x_i	f_i
0	2
2	4
3	**5**
4	3
6	1

Portanto, $\mathbf{M_o = 3}$

Moda (M_o)

3º Caso: dados agrupados com intervalos

Neste caso, consideramos como moda o valor compreendido entre os limites da **classe modal**, ou seja, **aquela que apresenta a maior frequência**. Tal valor é dado por:

$$M_o = l_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} . h$$

Em que:

$l_i =$ limite inferior da classe modal

$\Delta_1 =$ diferença entre a frequência (f_i) da classe modal e a imediatamente anterior

$\Delta_2 =$ diferença entre a frequência (f_i) da classe modal e a imediatamente posterior

$h =$ amplitude da classe modal

Moda (M_o)

3º Caso: dados agrupados com intervalos

Exemplo: Dada a tabela:

classe	f_i
0 ⊢ 10	1
10 ⊢ 20	3
20 ⊢ 30	6
30 ⊢ 40	2

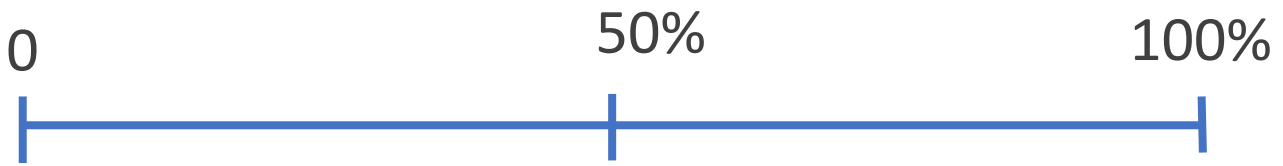
1º passo: identifica-se a classe modal (aquela que apresenta maior frequência)
No caso, trata-se da 3ª classe 20 ⊢ 30 (fi=6)

2º passo: aplica-se a fórmula. No caso temos:

$$l_i = 20$$
$$\Delta_1 = 6 - 3 = 3$$
$$\Delta_2 = 6 - 2 = 4$$
$$h = 30 - 20 = 10$$
$$M_o = 20 + \frac{3}{3+4} . 10$$
$$M_o = 20 + \frac{30}{7}$$
$$M_o = \frac{140 + 30}{7}$$
$$M_o = \frac{170}{7}$$
$$M_o = 24{,}29$$

Mediana ($\bar{x}$)

Dada uma coleção de números colocados em ordem crescente, a mediana ($\bar{x}$) é o valor que divide a amostra em duas partes iguais.



50% dos valores da série são valores menores ou iguais a $\bar{x}$ e 50% dos valores da série são maiores ou iguais a $\bar{x}$.

Mediana ($\bar{x}$)

1º Caso: dados não agrupados

Quando temos um número ímpar de elementos, dispostos em ordem crescente, a mediana é definida como sendo o elemento central, de ordem $\frac{n+1}{2}$

Exemplo: **1, 2, 3, 4, 5**

$\bar{x} = 3$

Se a coleção tiver um número par de elementos, também dispostos em ordem crescente, a mediana é definida como a média aritmética dos dois valores centrais, de ordem $\frac{n}{2}$ e $\frac{n}{2} + 1$

Exemplo: **2, 4, 6, 8, 10, 12**

$$\bar{x} = \frac{6+8}{2} = \frac{14}{2} = 7$$

Mediana ($\bar{x}$)

2º Caso: dados agrupados sem intervalos

Basta considerar a frequência acumulada e localizar a mediana, procedendo da mesma forma que no caso anterior.

Exemplo 1: Dada a distribuição:

x_i	f_i
12	1
14	2
15	1
16	2
17	1
20	2
Total	9

Como n = 9 é ímpar, logo será o elemento de ordem $\frac{n+1}{2}$ ou seja:

$$\bar{x} = \frac{n+1}{2} = \frac{9+1}{2} = \frac{10}{2} = 5 \text{ (5º elemento)}$$

Mediana ($\bar{x}$)

2º Caso: dados agrupados sem intervalos

Exemplo 2: Dada a distribuição:

x_i	f_i
12	2
14	2
15	2
16	2
17	2
20	2
Total	12

Como n = 12 é par, logo será o elemento de ordem $\frac{n+1}{2}$ e $\frac{n+1}{2} + 1$ ou seja:

$$\bar{x} = \frac{n}{2} e \frac{n}{2} + 1 = \frac{12}{2} \text{ e } \frac{12}{2} + 1 = 6 \text{ e } 7 \text{ (6º e 7º elemento)}$$

Mediana ($\bar{x}$)

3º Caso: dados agrupados com intervalos

Neste caso, devemos inicialmente localizar a classe mediana. Para isso seguimos os seguintes passos:

1º passo: calculamos a ordem $\frac{1}{2}$. Independente se n é par ou ímpar.

2º passo: pela F_i (Frequência acumulada) identificamos a classe que contém a mediana.

3º passo: utilizamos a fórmula:

$$\bar{x} = l_i + \frac{\left(\frac{n}{2} - \Sigma_f\right)}{F_{Md}} \cdot h$$

Em que:

$l_i =$ limite inferior da classe modal

$n =$ tamanho total da amostra ou número de elementos

$\sum_f =$ soma das frequências anteriores à classe mediana

$h =$ amplitude da classe mediana

$F_{Md} =$ Frequência da classe mediana

Mediana ($\bar{x}$)

3º Caso: dados agrupados com intervalos

Exemplo: Dada a tabela:

classe	f_i	F_i
3 ⊢ 6	2	2
6 ⊢ 9	5	7
9 ⊢ 12	8	15
12 ⊢ 15	3	18
15 ⊢ 18	1	19
Total	19	

1º passo: calcula-se $\frac{n}{2}$. Como n = 19, temos:

$\frac{19}{2} = 9{,}5$ (elemento)

2º passo: Identifica-se a classe mediana pela F_i.

Neste caso a classe mediana é a 3ª: 9 ⊢ 12

3º passo: Aplica-se a fórmula:

$l_i = 9$

$\Sigma_f = 7$

$h = 12 - 9 = 3$

$F_{Md} = 8$

Portanto:

$$\bar{x} = l_i + \frac{\left(\frac{n}{2} - \Sigma_f\right)}{F_{Md}} \cdot h$$

$$\bar{x} = 9 + \frac{9{,}5-7}{8} \cdot 3 = 9 + \frac{2{,}5}{8} \cdot 3 = 9 + \frac{7{,}5}{8} = 9{,}9375$$

Conclusão 50% dos valores são menores ou iguais a 9,94 e 50% são maiores ou iguais a 9,94

Prof. MSc. Marcos Alexandruk

E-mail: alexandruk@uni9.pro.br

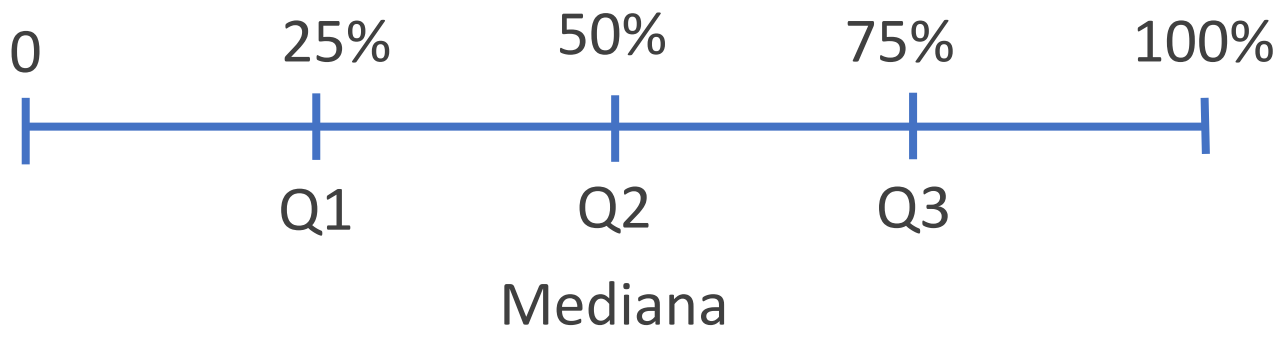
https://github.com/alexandruk/analisededados

Separatrizes

Quartil, Decil, Percentil

Quartis (Q_1, Q_2 (Md) e Q_3)

Os **quartis** dividem um conjunto ordenado de dados em quatro partes iguais, com cada parte representando 25%.



Quartis (Q_1, Q_2 (Md) e Q_3)

As notas de nove alunos em uma determinada prova estão apresentadas a seguir:

73, 74, 77, 52, 85, 59, 73, 84, 92

Determine a mediana, o 1º e o 3º quartil.

1º passo: Ordenar os elementos em ordem crescente:

52, 59, 73, 73, 74, 77, 84, 85, 92

2º passo: Determinar a mediana (o elemento central):

Q2 (Mediana) = 52, 59, 73, 73, 74, 77, 84, 85, 92

3º passo: Determinar Q1 (1º Quartil):

52, 59, 73, 73, 74, 77, 84, 85, 92

Q1 = (59+73)/2 = 122/2 = 66

4º passo: Determinar Q3 (3º Quartil):

52, 59, 73, 73, 74, 77, 84, 85, 92

Q3 = (84+85)/2 = 169/2 = 84,5

Quartis (Q_1)

Dados agrupados com intervalos

Exemplo: Dada a tabela:

classe	f_i	F_i
160 ⊢ 164	7	7
164 ⊢ 168	4	11
168 ⊢ 172	5	16
172 ⊢ 176	8	24
176 ⊢ 180	16	40
	$\sum f_i = 40$	

$$* = l_i + \frac{(k.\sum fi - F_{i\ anterior})}{f_i} \cdot h$$

Q_1: calcular $\frac{1 \cdot n}{4}$. Como n = 40, temos:

$\frac{40}{4} = 10º$ (elemento)

2º passo: Identifica-se a classe do Q_1 pela F_i.

Neste caso a classe Q_1 é a 2ª: 164 ⊢ 168

3º passo: Aplica-se a fórmula:

$l_i = 164$

$k = \frac{1}{4}$

$\Sigma_{fi} = 40$

$F_{i\ anterior} = 7$

$f_i = 4$

$h = 168 - 164 = 4$

Portanto:

$$Q_1 = 164 + \frac{\left(\frac{1}{4}.40 - 7\right)}{4} \cdot 4$$

$$Q_1 = 164 + \frac{(10 - 7)}{\cancel{4}} \cdot \cancel{4}$$

$$Q_1 = 164 + 3 = 167$$

Quartis (Q_2)

Dados agrupados com intervalos

Exemplo: Dada a tabela:

classe	f_i	F_i
160 ⊢ 164	7	7
164 ⊢ 168	4	11
168 ⊢ 172	5	16
172 ⊢ 176	8	24
176 ⊢ 180	16	40
	$\sum f_i = 40$	

$$* = l_i + \frac{(k.\sum fi - F_{i\,anterior})}{f_i} \cdot h$$

Q_2: calcular $\frac{2\,.\,n}{4}$. Como n = 40, temos:

$\frac{80}{4} = 20º$ (elemento)

2º passo: Identifica-se a classe do Q_2 pela F_i.

Neste caso a classe Q_2 é a 4ª: 172 ⊢ 176

3º passo: Aplica-se a fórmula:

$l_i = 172$

$k = \frac{1}{2}$

$\Sigma_{fi} = 40$

$F_{i\,anterior} = 16$

$f_i = 8$

$h = 176 - 172 = 4$

Portanto:

$$Q_2 = 172 + \frac{\left(\frac{1}{2}.40 - 16\right)}{8} \cdot 4$$

$$Q_2 = 172 + \frac{(20 - 16)}{8} \cdot 4$$

2

$$Q_2 = 172 + 2 = 174$$

Quartis (Q_3)

Dados agrupados com intervalos

Exemplo: Dada a tabela:

classe	f_i	F_i
160 ⊢ 164	7	7
164 ⊢ 168	4	11
168 ⊢ 172	5	16
172 ⊢ 176	8	24
176 ⊢ 180	16	40
	$\sum f_i = 40$	

$$* = l_i + \frac{(k.\sum fi - F_{i\ anterior})}{f_i} \cdot h$$

Q_3: calcular $\frac{3\ .\ n}{4}$. Como n = 40, temos:

$\frac{120}{4} = 30º$ (elemento)

2º passo: Identifica-se a classe do Q_3 pela F_i.

Neste caso a classe Q_3 é a 5ª: 176 ⊢ 180

3º passo: Aplica-se a fórmula:

$l_i = 176$

$k = \frac{3}{4}$

$\Sigma_{fi} = 40$

$F_{i\ anterior} = 24$

$f_i = 16$

$h = 180 - 176 = 4$

Portanto:

$$Q_3 = 176 + \frac{\left(\frac{3}{4}.40 - 24\right)}{16} \cdot 4$$

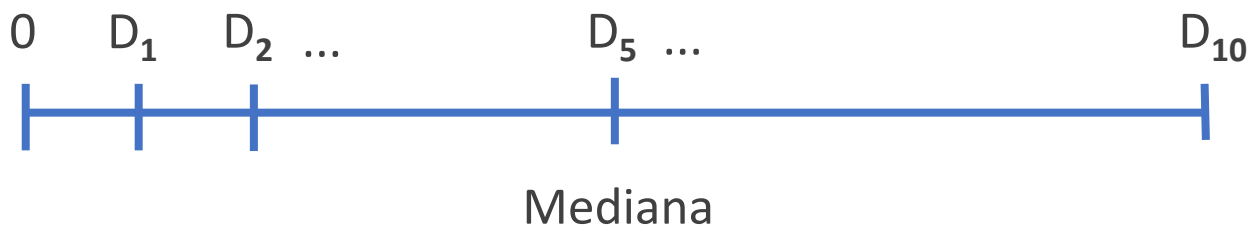
$$Q_3 = 176 + \frac{(30 - 24)}{16} \cdot \cancel{4}$$

4

$$Q_3 = 176 + \frac{6}{4} = 176 + 1{,}5 = 177{,}5$$

Decil (D_1, D_2 , D_3 ... D_{10})

Os **decis** dividem um conjunto ordenado de dados em 10 partes iguais, com cada parte representando 10%.



Decil

Dados agrupados com intervalos

Exemplo: Dada a tabela:

classe	f_i	F_i
160 ⊢ 162	7	7
162 ⊢ 164	4	11
164 ⊢ 166	8	19
166 ⊢ 168	9	28
168 ⊢ 170	12	40
	$\sum f_i = 40$	

$$* = l_i + \frac{(k.\sum fi - F_{i\ anterior})}{f_i} \cdot h$$

D_2: calcular $\frac{2\ .\ n}{10}$. Como n = 40, temos:
$\frac{80}{10} = 8º$ (elemento)

2º passo: Identifica-se a classe D_2 pela F_i.

Neste caso a classe D_2 é a 2ª: 162 ⊢ 164

3º passo: Aplica-se a fórmula:

$l_i = 162$
$k = \frac{2}{10}$
$\Sigma_{fi} = 40$
$F_{i\ anterior} = 7$
$f_i = 4$
$h = 164 - 162 = 2$

Portanto:

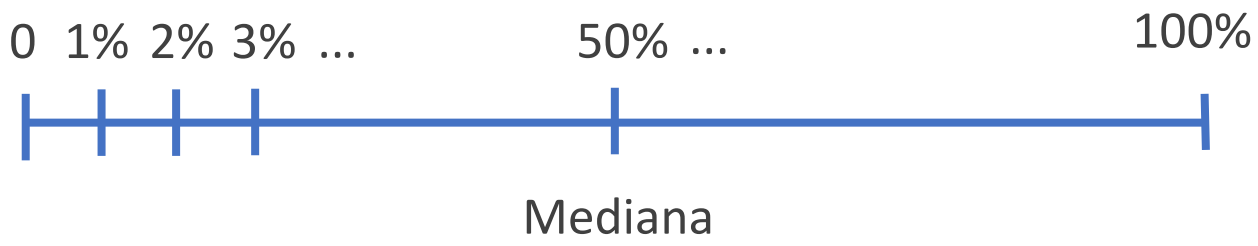
$$D_2 = 162 + \frac{\left(\frac{2}{10}.40 - 7\right)}{4} \cdot 2$$

$$D_2 = 162 + \frac{(8-7)}{\cancel{4}\ 2} \cdot \cancel{2}$$

$$D_2 = 162 + \frac{1}{2} = 162 + 0{,}5 = 162{,}5$$

Percentil (P_1, P_2 , P_3 ... P_{100})

Os **percentis** dividem um conjunto ordenado de dados em 100 partes iguais, com cada parte representando 1%.



Percentil

Dados agrupados com intervalos

Exemplo: Dada a tabela:

classe	f_i	F_i
160 ⊢ 162	7	7
162 ⊢ 164	4	11
164 ⊢ 166	8	19
166 ⊢ 168	9	28
168 ⊢ 170	12	40
	$\sum f_i = 40$	

$$* = l_i + \frac{(k.\sum fi - F_{i\ anterior})}{f_i} \cdot h$$

P_{20}: calcular $\frac{20\,.\,n}{100}$. Como n = 40, temos:
$\frac{800}{100} = 8º$ (elemento)

2º passo: Identifica-se a classe P_{20} pela F_i.

Neste caso a classe P_{20} é a 2ª: 162 ⊢ 164

3º passo: Aplica-se a fórmula:

$l_i = 162$
$k = \frac{20}{100}$
$\Sigma_{fi} = 40$
$F_{i\ anterior} = 7$
$f_i = 4$
$h = 164 - 162 = 2$

Portanto:

$$P_{20} = 162 + \frac{\left(\frac{20}{100} \cdot 40 - 7\right)}{4} \cdot 2$$

$$P_{20} = 162 + \frac{(8-7)}{\cancel{4}\,2} \cdot \cancel{2}$$

$$P_{20} = 162 + \frac{1}{2} = 162 + 0{,}5 = 162{,}5$$

R (quartil, decil e percentil)

```
x <- c(69, 70, 75, 66, 83, 88, 66, 63, 61, 68, 73, 57, 52, 58, 77)
```

quartis

```
quantile(x)
```

decis

```
quantile(x, prob = seq(0, 1, length = 11))
```

percentis

```
quantile(x, prob = seq(0, 1, length = 101))
```

resumo

```
summary(x)
```

R (entrada de dados externos – arquivo .csv)

Criar o arquivo teste.txt:

```
nome,idade
Fulano,20
Beltrano,30
Sicrano,40
```

Importar os dados:

```
teste<-read.table("c:/Aulas/teste.txt",header=T,sep=",")
```

Medidas de Dispersão

desvio médio, variância, desvio padrão e coeficiente de variação

Média, moda e mediana

A **média**, apesar de ser uma medida muito utilizada em Estatística, é muitas vezes insuficiente para caracterizar aceitavelmente uma distribuição.

A **moda** e a **mediana** também são medidas que nem sempre são suficientes para caracterizar um conjunto de dados.

Em alguns casos, temos que recorrer a outros parâmetros, chamados de medidas de dispersão.

As medidas de dispersão são medidas estatísticas utilizadas para avaliar o grau de variabilidade ou dispersão dos valores em torno da média. Servem para medir a representatividade da média.

Média, moda e mediana

Observe as séries:

a. 10, 1, 18, 20, 35, 3, 7, 15, 11, 10
b. 12, 13, 13, 14, 12, 14, 12, 14, 13, 13
c. 13, 13, 13, 13, 13, 13, 13, 13, 13, 13

Estes dados possuem a mesma média 13. No entanto, são sequências completamente distintas do ponto de vista da variabilidade de dados.

Na série "c" não há dispersão.

Comparando-se as séries "a" e "b", percebe-se que "a" apresenta maior dispersão em torno da média do que "b".

Isso indica que necessitamos de outro tipo de medida para distinguir e comparar os três conjuntos de dados.

O critério frequentemente usado para tal fim é aquele que mede a maior ou menor dispersão dos dados em torno da média, e as medidas mais usadas são:

- **desvio médio**
- **variância**
- **desvio padrão**
- **coeficiente de variação**

Desvio médio (Dm)

É a análise dos desvios em torno da média. Calculamos inicialmente a média da amostra ($\bar{x}$):

Em seguida, identificamos a distância de cada elemento da amostra para sua média:

$$|d_i| = |x_i - \bar{x}|$$

Finalmente, calculamos o desvio médio:

$$\frac{\sum |d_i| F_i}{n} \quad \text{ou} \quad \frac{\sum |x_i - \bar{x}| F_i}{n}$$

Onde x_i é a variável, $\bar{x}$ a média e n o número de dados da amostra.

Dessa forma, o desvio médio é a média aritmética dos valores absolutos dos desvios.

x_i	F_i	$x_i F_i$	$\lvert d_i \rvert = \lvert x_i - \bar{x} \rvert$	$\lvert d_i \rvert F$
2	5	10	|2 - 4,17| = 2,17	2,17 x 5 = 10,85
3	4	12	|3 - 4,17| = 1,17	1,17 x 4 = 4,68
5	4	20	|5 - 4,17| = 0,83	0,83 x 4 = 3,32
6	2	12	|6 - 4,17| = 1,83	1,83 x 2 = 3,66
7	3	21	|7 - 4,17| = 2,83	2,83 x 3 = 8,49
Total	18	75		31

$$\bar{x} = \frac{\sum x_i F_i}{n} = \frac{75}{18} = 4{,}17$$

$$Dm = \frac{\sum |\mathrm{di}| F_i}{n} = \frac{31}{18} = 1{,}72$$

Variância (Var)

É a média aritmética dos quadrados dos desvios. Logo:

$$Var = \frac{\sum d_i^2 F_i}{n}$$

x_i	F_i	$x_i F_i$	$\lvert d_i \rvert = \lvert x_i - \bar{x} \rvert$	di^2	$d_i^2 F_i$
2	5	10	|2 - 4,17| = 2,17	4,71	23,55
3	4	12	|3 - 4,17| = 1,17	1,37	5,48
5	4	20	|5 - 4,17| = 0,83	0,69	2,76
6	2	12	|6 - 4,17| = 1,83	3,35	6,7
7	3	21	|7 - 4,17| = 2,83	8,01	24,03
Total	18	75			62,52

$$Var = \frac{\sum d_i^2 F_i}{n} = \frac{62{,}52}{18} = 3{,}47$$

Desvio padrão (Dp)

Como para calcular a variância trabalhamos com os quadrados dos desvios, podemos ter uma incompatibilidade em relação às unidades dos valores da variável considerada.

Para contornar esse problema, temos o desvio padrão, que é a raiz quadrada da variância:

$$Dp = \sqrt{Var}$$

x_i	F_i	$x_i F_i$	$\lvert d_i \rvert = \lvert x_i - \bar{x} \rvert$	di^2	$d_i^2 F_i$
2	5	10	|2 - 4,17| = 2,17	4,71	23,55
3	4	12	|3 - 4,17| = 1,17	1,37	5,48
5	4	20	|5 - 4,17| = 0,83	0,69	2,76
6	2	12	|6 - 4,17| = 1,83	3,35	6,7
7	3	21	|7 - 4,17| = 2,83	8,01	24,03
Total	18	75			62,52

$$Var = \frac{\sum d_i^2 F_i}{n} = \frac{62{,}52}{18} = 3{,}47$$

$$Dp = \sqrt{Var} = \sqrt{3{,}47} = 1{,}86$$

Resumindo: a distribuição possui média **4,17**. Isto é, seus valores estão em torno de **4,17** e seu grau de concentração é de **1,72**, medido pelo desvio médio e de **1,86**, medido pelo desvio padrão.

Coeficiente de Variação (CV)

O desvio padrão por si só não nos diz muita coisa; para contornar esta dificuldade, usamos o coeficiente de variação.

Trata-se de uma medida relativa de dispersão útil para a comparação em termos relativos do grau de concentração em torno da média de séries distintas.

É expresso em porcentagens e dado por:

$$CV = \frac{Dp}{\bar{x}} \cdot 100$$

Onde Dp é o desvio padrão e $\bar{x}$, a média da distribuição.

Diz-se que a distribuição possui pequena variabilidade (dispersão) quando o CV apresentar valor até 15%; média dispersão quando estiver acima de 15% até 30% e grande dispersão quando superar 30%.

Coeficiente de Variação (CV)

Considere a tabela abaixo:

x_i	F_i	x_iF_i	$\|d_i\| = \|x_i - \bar{x}\|$	di^2	$d_i^2F_i$
2	5	10	|2 - 4,17| = 2,17	4,71	23,55
3	4	12	|3 - 4,17| = 1,17	1,37	5,48
5	4	20	|5 - 4,17| = 0,83	0,69	2,76
6	2	12	|6 - 4,17| = 1,83	3,35	6,7
7	3	21	|7 - 4,17| = 2,83	8,01	24,03
Total	18	75			62,52

Baixa dispersão CV ≤ 15%

Média dispersão: 15% < CV < 30%

Alta dispesão: CV ≥ 30%

$$Var = \frac{\sum d_i^2 F_i}{n} = \frac{62,52}{18} = 3,47$$

$$Dp = \sqrt{Var} = \sqrt{3,47} = 1,86$$

$$\bar{x} = 4,17$$

$$CV = \frac{Dp}{\bar{x}} \cdot 100$$

$$CV = \frac{1,86}{4,17} .100 = 44,60\%$$

alta dipersão

Coeficiente de Variação (CV)

Exemplo: Numa empresa, o salário médio dos homens é de R$ 4.000,00, com desvio padrão de R$ 1.500,00 e, o das mulheres, é em média de R$ 3.000,00, com desvio padrão de R$ 1.200,00. Então:

$$CV_H = \frac{1500}{4000} . 100 = 37{,}5\%$$

$$CV_M = \frac{1200}{3000} . 100 = 40\%$$

Logo, podemos concluir que os salários das mulheres apresentam maior dispersão que os dos homens.

De modo geral, quanto menor o CV, menos dispersos estão os dados em torno da média, que passa a ser mais representativa do conjunto de dados.

Medidas de Dispersão (Exemplo)

Encontre o desvio médio, o desvio padrão e o coeficiente de variação da distribuição:

Classes	x_i	F_i	$x_i F_i$	$\lvert d_i \rvert$	$\lvert d_i \rvert F_i$	di^2	$d_i^2 F_i$
2 ⊢ 4	3	2	6	4,2	8,4	17,64	35,28
4 ⊢ 6	5	4	20	2,2	8,8	4,48	19,36
6 ⊢ 8	7	7	49	0,2	1,4	0,04	0,28
8 ⊢ 10	9	4	36	1,8	7,2	3,24	12,96
10 ⊢ 12	11	3	33	3,8	11,4	14,44	43,32
		20	144		37,2		111,20

Introdução à teoria da amostragem

Introdução à teoria da amostragem

Objetivo:

Determinar o espaço amostral, os eventos desse espaço e calcular o número de elementos destes conjuntos.

Experimento aleatório (E)

Experimentos aleatórios são aqueles que, mesmo repetidos várias vezes sob condições semelhantes, **apresentam resultados imprevisíveis**.

Exemplo:

Em uma afirmação do tipo: "é provável que meu time ganhe a partida de hoje" pode resultar:

- Que o time perca.
- Que o time ganhe.
- Que o time empate.

O resultado é imprevisível e depende do acaso. Fenômenos como esses são chamados **fenômenos aleatórios** ou **experimentos aleatórios**.

Espaço amostral (S)

É o conjunto de todos os possíveis resultados de um experimento aleatório (E). Indicamos o espaço amostral por S e o número de elementos de S por n(S).

Exemplo:

E: jogar um dado cúbico e observar o número da face de cima:

S = {1, 2, 3, 4, 5, 6}

n(S) = 6

E: jogar uma moeda e observar o resultado:

S = {cara, coroa}

n(S) = 2

E: lançar duas moedas e observar o resultado na face de cada uma:

S = {(cara, cara), (cara, coroa), (coroa, cara), (coroa, coroa)}

n(S) = 4

Evento

É qualquer subconjunto do espaço amostral S de um experimento aleatório (E).

Exemplo:

E: lançar um dado cúbico e observar o número da face de cima:

S = {1, 2, 3, 4, 5, 6}

n(S) = 6

Exemplos de eventos de S:

A: sair número par: {2, 4, 6}

n(A) = 3

B: sair número primo: {2, 3, 5}

n(B) = 3

Combinações de eventos

Uma urna contém 3 bolas pretas e 3 bolas vermelhas. Dessa urna são retiradas, sucessivamente, 3 bolas.
Determine o espaço amostral.
Determine os eventos.
A: as 3 bolas têm a mesma cor
B: exatamente 2 das bolas são pretas
C: as 3 bolas são vermelhas
D: o número de bolas pretas é igual ao número de bolas vermelhas

Solução:

S = {(P, P, P), (P, P, V), (P, V, P), (P, V, V), (V, P, P), (V, P, V), (V, V, P), (V, V, V)}
n(S) = 8
A = {(P, P, P), (V, V, V)}
n(A) = 2
B = {(P, P, V), (P, V, P), (V, P, P)}
n(B) = 3
C = {(V, V, V)}
n(C) = 1
D = { }
n(D) = 0

O conjunto vazio é chamado evento impossível.
Quando o evento coincide com o espaço amostral, ele é chamado evento certo.

Combinações de eventos

União de dois eventos

Sejam **A** e **B** dois eventos, então **A U B** é um evento que ocorrerá se, e somente se, **A** ou **B** ocorrem.

Considere o exemplo:

E: lançamento um dado cúbico e observação do número voltado para cima

S = {1, 2, 3, 4, 5, 6}

A: ocorrência de um número ímpar: **{1, 3, 5}**

B: ocorrência de um número par primo: **{2}**

Logo, A U B: ocorrência de um número ímpar ou um número par primo:

A U B = {1, 2, 3, 5}

Combinações de eventos

União de dois eventos

Sejam **A** e **B** dois eventos, então **A ∩ B** é um evento que ocorrerá se, e somente se, **A** e **B** ocorrem simultaneamente.

Observação: em particular, se **A ∩ B = Ø** , **A** e **B** são chamados mutuamente exclusivos.

Considere o exemplo:

E: lançamento um dado cúbico e observação do número voltado para cima

S = {1, 2, 3, 4, 5, 6}

A: ocorrência de um número par: **{2, 4, 6}**

B: ocorrência de um número múltiplo de 4: **{4}**

Logo, A ∩ B: ocorrência de um número par e múltiplo de 4:

A ∩ B = {4}

Combinações de eventos

Complementar de um evento

Dado um evento **A**, o conjunto formado pelos elementos de S que não pertencem a **A** se chama evento complementar de **A** em relação a S e indica-se por $\overline{\mathbf{A}}$.

Considere o exemplo:

E: lançamento um dado cúbico e observação do número voltado para cima

S = {1, 2, 3, 4, 5, 6}

A: ocorrência de um número par: **{2, 4, 6}**

$\overline{\mathbf{A}}$ **= {1, 3, 5}**

A U $\overline{\mathbf{A}}$ **= {1, 2, 3, 4, 5, 6} = S**

A ∩ $\overline{\mathbf{A}}$ **= Ø**

Fatorial, permutação, arranjo e combinação

Fatorial

O fatorial de um número natural n é representado por **n!** (lê-se n fatorial), em que:

n! = n . (n – 1) . (n – 2) . ... . 3 . 2 . 1, para n ≥ 2

1! = 1

0! = 1

Exemplos:

4! = 4 . 3 . 2 . 1 = 24

7! = 7 . 6 . 5 . 4 . 3 . 2 . 1 = 5040

n! = n . (n – 1)! ou n . (n – 1) . (n – 2)! ...

Exemplo: simplificar as expressões:

$$\frac{8!}{5!.3!} = \frac{8.7.6.\cancel{5!}}{\cancel{5!}.3!} = \frac{8.7.6}{3.2.1} = 56$$

$$\frac{(n+1)!}{n!} = \frac{(n+1).\cancel{n!}}{\cancel{n!}} = n+1$$

Permutação

Dado um conjunto de n elementos, chama-se permutação simples dos n elementos, qualquer sequência – agrupamento ordenado – desses n elementos.

$P_n = n!$

Exemplos:

Dados três cartões coloridos: Amarelo (A), Laranja (L) e Vemelho (V). Quantas sequências diferentes podemos obter ao enfileirar os três cartões?

$P_n = 3! = 3 . 2 . 1 = 6$

(A,L,V) (A,V,L) (L,A,V) (L,V,A) (V,A,L) (V,L,A)

Um restaurante funciona cinco dias por semana. Elabrou, portanto, cinco menus diferentes, designados como M1, M2, M3, M4 e M5. De quantas maneiras é possível escolher esses menus de forma que não haja repetição na semana?

$P_n = 5! = 5 . 4 . 3 . 2 . 1 = 120$

Arranjo

Chama-se arranjo simples de n elementos distintos tomados p a p (p ≤ n), todo agrupamento **ordenado** formado por p elementos escolhidos entre os n elementos dados. *(muda a ordem, muda o grupo: AB ≠ BA)*

Arranjo Simples

Considere um conjunto com **n** elementos distintos. Qualquer sequência de **p** desses elementos (todos distintos) é chamada de Arranjo Simples (0 ≤ p ≤ n, com n e p naturais). Dizemos arranjo simples de n elementos tomados p a p, e simbolizamos por $A_{n,p}$

Esse arranjo simples pode ser calculado da seguinte forma:

$$A_{n,p} = \frac{n!}{(n-p)!}$$

Arranjos com a totalidade dos elementos, ou seja, $A_{n,n}$ são as permutações de n elementos.

$$A_{n,n} = P_n = n!$$

Exemplo:

Em uma urna de sorteio de prêmios existem dez bolas enumeradas de 0 a 9. Determine o número de possibilidades existentes num sorteio cujo prêmio é formado por uma sequência de 6 algarismos.

$$A_{10,6} = \frac{n!}{(n-p)!} = \frac{10!}{(10-6)!} = \frac{10!}{4!} = \frac{10 \,.\, 9 \,.\, 8 \,.\, 7 \,.\, 6 \,.\, 5 \,.\, \cancel{4!}}{\cancel{4!}} = 10 \,.\, 9 \,.\, 8 \,.\, 7 \,.\, 6 \,.\, 5 = 151200$$

Combinação

Denominam-se combinações simples de n elementos tomados p a p (p ≤ n), aos diferentes subconjuntos que contêm p elementos, **sem referência à ordem**. *(muda a ordem, o grupo permanece o mesmo: AB = BA)*

Uma combinação simples é representada da seguinte forma:

$$C_{n,p} = \frac{n!}{p!\,(n-p)!}$$

Exemplo:

Dado o conjunto de 4 elementos {a,b,c,d}, quantas combinações formadas com 2 elementos são possíveis?

$$C_{4,2} = \frac{n!}{p!\,(n-p)!} = \frac{4!}{2!\,(4-2)!} = \frac{4\,.\,3\,.\,2!}{2!\,.\,2!} = \frac{12}{2} = 6$$

{a,b} {a,c} {a,d} {b,c} {b,d} {c,d}

$C_{n,p}$ ou $\binom{n}{p}$ são formas de indicar o número de combinações de n elementos p a p, com n ≥ p.

$\binom{n}{p}$ também é chamado **coeficiente binomial** porque é utilizado em uma fórmula matemática chamada **teorema binomial**.

Exercícios

Fatorial: Calcule os seguintes fatoriais:

$4!$

$\frac{10!}{8!}$

Permutação: Dados cinco cartões coloridos, quantas sequências diferentes podemos obter ao enfileirar os cinco cartões?

Arranjo: Suponha que, em uma corrida de oito cavalos, você esteja tentando acertar a ordem de chegada dos três primeiros finalistas, sem nada saber sobre os cavalos. Quantas finais são possíveis?

Combinação: Sabe-se que um júri foi formado por 7 pessoas, selecionadas de um grupo de 21 pessoas. Neste caso, temos um agrupamento de ordem 7 (as 7 pessoas que formam o júri). A ordem de escolha dessas 7 pessoas não muda o grupo. Portanto, quantas combinações são possíveis?

Quantos jogos diferentes (com seis números em cada jogo) uma pessoa precisaria fazer para ter 100% de certeza de ganhar o prêmio em um determinado sorteio da Mega Sena? (São sorteados 6 números dentre 60 em cada sorteio.)

Probabilidade de um evento

Probabilidade de ocorrência de um evento

Dado um experimento aleatório, sendo S o seu espaço amostral, vamos admitir que todos os elementos de S tenham a mesma chance de acontecer, chamamos de probabilidade de um evento A o número real P(A), tal que:

$$P(A) = \frac{n(A)}{n(S)}$$

Em que:

n(A) = número de elementos do evento A

n(S) = número de elementos do espaço amostral S

Probabilidade de ocorrência de um evento

Considerando o lançamento de um dado, determine a probabilidade de ocorrer na face superior:

Um número par

Solução:

Temos que: S = {1, 2, 3, 4, 5, 6}, logo n(S) = 6

A = {2, 4, 6}, logo n(A) = 3

Então: P(A) = $\frac{3}{6} = \frac{1}{2}$ ou 50%

Probabilidade de ocorrência de um evento

Considerando o lançamento de um dado, determine a probabilidade de ocorrer na face superior:

O número 2

Solução:

Temos que: S = {1, 2, 3, 4, 5, 6}, logo n(S) = 6

B = {2}, logo n(B) = 1

Então: P(B) = $\frac{1}{6}$, ou 16,67%

Probabilidade de ocorrência de um evento

Considerando o lançamento de um dado, determine a probabilidade de ocorrer na face superior:

Um número menor ou igual a 6

Solução:

Temos que: S = {1, 2, 3, 4, 5, 6}, logo n(S) = 6

C = {1, 2, 3, 4, 5, 6}, logo n(C) = 6

Então: P(C) = $\frac{6}{6}$ = 1, ou 100%

Probabilidade de ocorrência de um evento

Considerando o lançamento de um dado, determine a probabilidade de ocorrer na face superior:

Um número maior que 6

Solução:

Temos que: S = {1, 2, 3, 4, 5, 6}, logo n(S) = 6

D = {...}, logo n(D) = 0

Então: P(D) = $\frac{0}{6}$ = 0, ou 0%

Eventos complementares

Sabemos que um evento pode ocorrer ou não. Sendo **p** a probabilidade de que ele ocorra (**sucesso**) e **q** a probabilidade de que ele não ocorra (**insucesso**), então: **p + q = 1** ou **q = 1 – p**:

Exemplo: vimos que, no lançamento de um dado, a probabilidade de ocorrer o número 2 na face superior é $\frac{1}{6}$ ou 16,67%.

Logo, a probabilidade de não tirar 2 no lançamento é $1-\frac{1}{6}=\frac{5}{6}$ ou 83,33%.

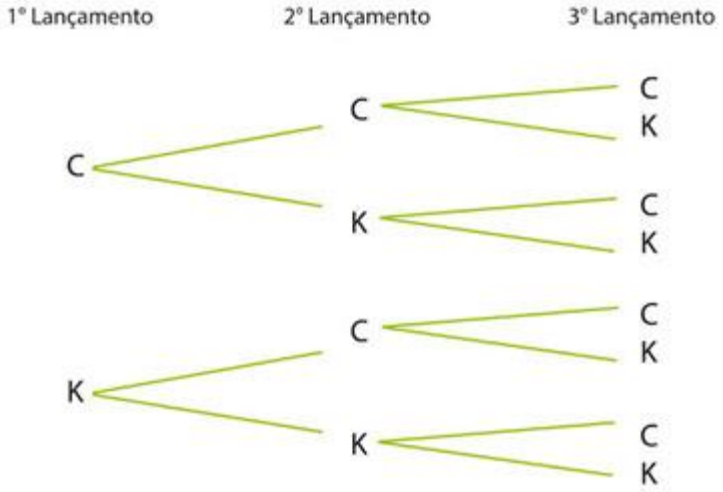
Exercício resolvido 1

Uma moeda é lançada 3 vezes sucessivamente. Qual a probabilidade de obtermos:

- Resultados iguais: 3 vezes cara ou 3 vezes coroa
- Pelo menos uma cara
- Exatamente uma cara
- Número de coroas maior que o número de caras

Exercício resolvido 1

- **Resultados iguais: 3 vezes cara (C) ou 3 vezes coroa (K)**



Solução:

$S = \{(CCC), (CCK), (CKC), (CKK), (KCC), (KCK), (KKC), (KKK)\}$

$n(S) = 8$

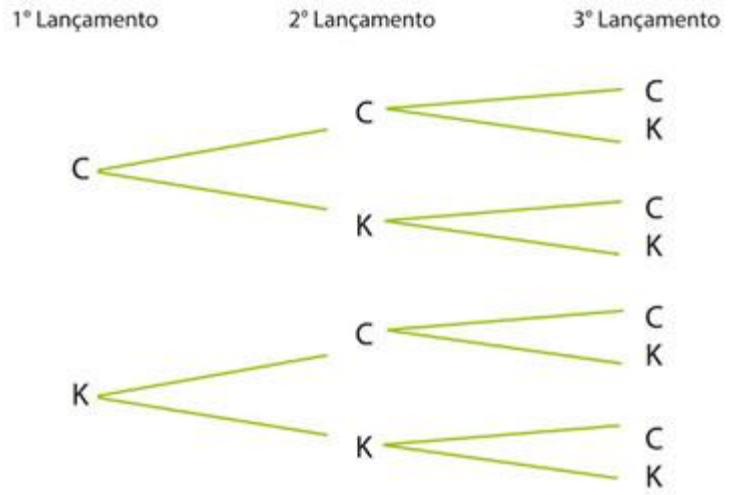
$A = \{(CCC), (KKK)\}$

$n(A) = 2$

$\text{Logo}, P(A) = \frac{2}{8} = \frac{1}{4} = 0{,}25 = 25\%$

Exercício resolvido 1

- **Pelo menos uma cara (C)**



Solução:

$S = \{(CCC), (CCK), (CKC), (CKK), (KCC), (KCK), (KKC), (KKK)\}$

$n(S) = 8$

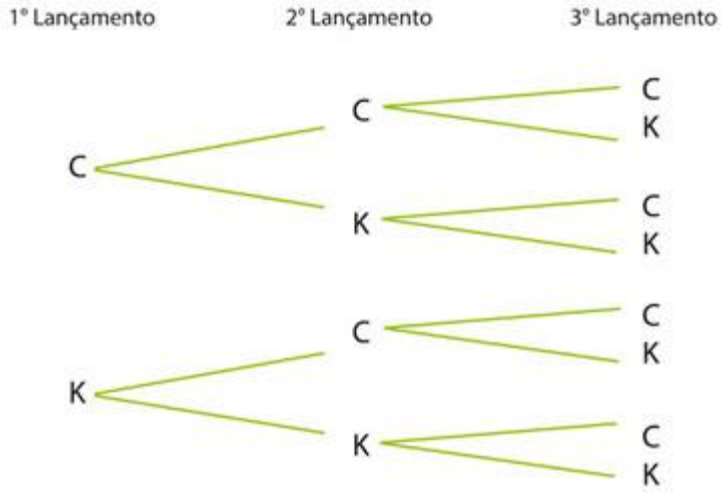
$B = \{(CCC), (CCK), (CKC), (CKK), (KCC), (KCK), (KKC)\}$

$n(B) = 7$

$\text{Logo}, P(B) = \frac{7}{8} = 0{,}875 = 87{,}5\%$

Exercício resolvido 1

- **Exatamente uma cara (C)**



Solução:

$S = \{(CCC), (CCK), (CKC), (CKK), (KCC), (KCK), (KKC), (KKK)\}$

$n(S) = 8$

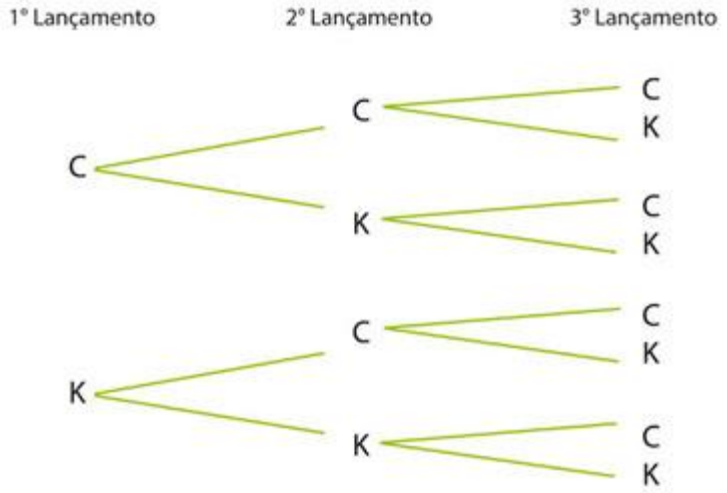
$C = \{(CKK), (KCK), (KKC)\}$

$n(C) = 3$

$\text{Logo}, P(C) = \frac{3}{8} = 0{,}375 = 37{,}5\%$

Exercício resolvido 1

- **Número de coroas (K) maior que o número de caras (C)**



Solução:

$S = \{(CCC), (CCK), (CKC), (CKK), (KCC), (KCK), (KKC), (KKK)\}$

$n(S) = 8$

$D = \{(CKK), (KCK), (KKC), (KKK)\}$

$n(D) = 4$

$\text{Logo}, P(D) = \frac{4}{8} = \frac{1}{2} = 0{,}50 = 50\%$

Exercício resolvido 2

Uma equipe de **doze pessoa**s é formada por **nove homens** e **três mulheres**. Dessas pessoas, duas serão sorteadas para compor uma comissão. Qual é a probabilidade de a comissão ser formada por:

- Duas mulheres
- Dois homens
- Um homem e uma mulher

Solução: Vamos, primeiramente, calcular o número de elementos do espaço amostral S; para isso, devemos considerar um grupo de 12 pessoas, do qual serão retirados 2 elementos, não importando a ordem, o que corresponde ao número de combinações de 12, tomados 2 a 2:

$$C_{n,p} = \frac{n!}{p!\,(n-p)!}$$

$$n(S) = C_{12,2} = \frac{12!}{(12-2)!\;\;2!} = \frac{12!}{10!\;\;2!} = \frac{10 \cdot 11 \cdot 10!}{10!\;\;2!} = \frac{12 \cdot 11}{2 \cdot 1} = \frac{132}{2} = 66$$

Exercício resolvido 2

- **Comissões formadas por 2 mulheres, de um total de 3 mulheres:**

$$n(A) = C_{3,2} = \frac{3!}{(3-2)!\,2!} = \frac{3!}{1!\,2!} = \frac{6}{2} = 3$$

$$Logo, P(A) = \frac{n(A)}{n(S)} = \frac{3}{66} = \frac{1}{22} = 0{,}045 = 4{,}5\%$$

Exercício resolvido 2

- **Comissões formadas por 2 homens, de um total de 9 homens:**

$$n(B) = C_{9,2} = \frac{9!}{(9-2)!\,2!} = \frac{9\,.\,8\,.\,7!}{7!\,2!} = \frac{72}{2} = 36$$

$$Logo\ P(B) = \frac{n(B)}{n(S)} = \frac{36}{66} = \frac{6}{11} = 0{,}545 = 54{,}5\%$$

Exercício resolvido 2

- **Comissões formadas por 1 homem (de um total de 9) e 1 mulher (de um total de 3):**

$$n(C) = C_{9,1} \,.\, C_{3,1} = \frac{9!}{(9-1)!\,1!} \cdot \frac{3!}{(3-1)!\,1!} = \frac{9!}{8!} \cdot \frac{3!}{2!} = \frac{9\,.\,8!}{8!} \cdot \frac{3\,.\,2!}{2!} = 9\,.\,3 = 27$$

$$Logo\ P(C) = \frac{n(C)}{n(S)} = \frac{27}{66} = \frac{9}{22} = 0{,}409 = 40{,}9\%$$

Teoremas de cálculo de probabilidade

Teorema da soma: probabilidade da união de dois eventos

Considere o experimento: Lançamento simultâneo de dois dados cúbicos, um preto e um vermelho, e a observação da soma dos números que aparecem nas faces superiores.

Qual é a probabilidade de obtermos soma par **ou** soma múltipla de 3?

O espaço amostral S é:

S = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

n(S) = 36

O evento A – sair soma par é:

A = {(1, 1), (1, 3), (1, 5), (2, 2), (2, 4), (2, 6), (3, 1), (3, 3), (3, 5), (4, 2), (4, 4), (4, 6), (5, 1), (5, 3), (5, 5), (6, 2),(6, 4), (6, 6)}

n(A) = 18

$$Logo, P(A) = \frac{n(A)}{n(s)} = \frac{18}{36} = \frac{1}{2}$$

O evento B – sair soma múltipla de 3 é:

B = {(1, 2), (1, 5), (2, 1), (2, 4), (3, 3), (3, 6), (4, 2), (4, 5), (5, 1), (5, 4), (6, 3), (6, 6)}

n(B) = 12

$$Logo, P(B) = \frac{n(B)}{n(S)} = \frac{12}{36} = \frac{1}{3}$$

Queremos a probabilidade de obter soma par **ou** soma múltipla de 3, ou seja, procuramos a probabilidade de ocorrer o evento AÈ B:

AUB = {(1, 1), (1, 2), (1, 3), (1, 5), (2, 1), (2, 2), (2, 4), (2, 6), (3, 1), (3, 3), (3, 5), (3, 6), (4, 2), (4, 4), (4, 5), (4, 6), (5, 1), (5, 3), (5, 4), (5, 5), (6, 2), (6, 3), (6, 4), (6, 6)}

n(AUB) = 24

$$Logo, P(A \bigcup B) = \frac{n(A \bigcup B)}{n(S)} = \frac{24}{36} = \frac{2}{3}$$

Teorema da soma: probabilidade da união de dois eventos

Suponha agora que queremos calcular a probabilidade de obter soma par e soma múltipla de 3, ou seja, procuramos a probabilidade de ocorrer o evento A∩B:

A∩B = {(1, 5), (2, 4), (3, 3), (4, 2), (5, 1), (6, 6)}

n(A∩B) = 6

O espaço amostral S é:

S = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

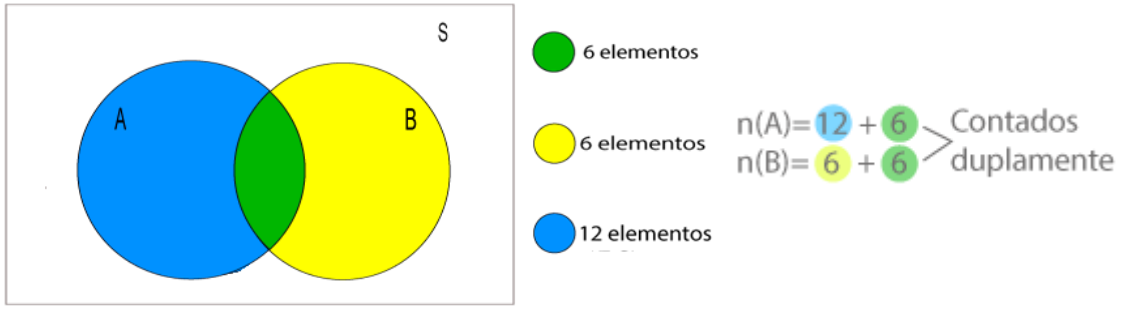
n(S) = 36

$$Logo, P(A \cap B) = \frac{n(a \cap B)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

Considerando os resultados obtidos de P(A), P(B), P(A∪B) e P(A∩B), verifica-se a igualdade:

$$\frac{2}{3} = \frac{1}{2} + \frac{1}{3} - \frac{1}{6}, \quad ou \quad seja, \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Podemos também comprovar esse resultado no diagrama seguinte, chamado Diagrama de Venn:

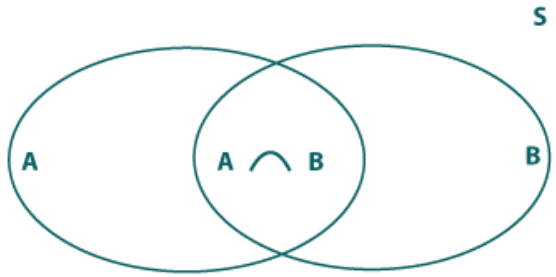


Teorema da soma: probabilidade da união de dois eventos

Se A e B são eventos do mesmo espaço amostral S, então a probabilidade de ocorrer A ou B (A∪B) é dada por:

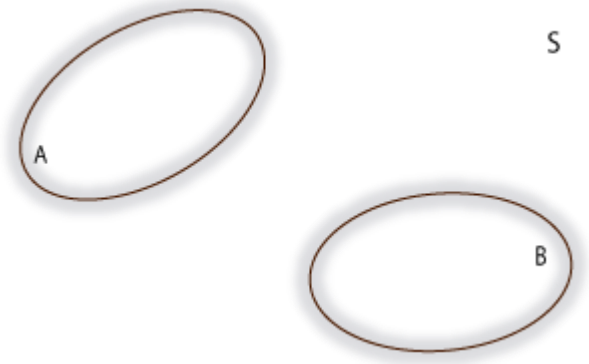
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Podemos comprovar este resultado no Diagrama de Venn a seguir:



Em particular, se A e B são eventos mutuamente exclusivos (A∩B = Ø), então:

$$P(A \cup B) = P(A) + P(B) - P(\emptyset) = P(A) + P(B)$$



Teorema da soma: probabilidade da união de dois eventos

Uma urna contém 100 bolas idênticas numeradas de 1 a 100 e uma delas é escolhida ao acaso. Qual a probabilidade de:
Obtermos um múltiplo de 6 ou de 8.
O espaço amostral S é:
S = {1, 2, 3, ... , 100}
n(S) = 100
Evento A – sair múltiplo de 6:
A = {6, 12, 18, 24, 30, 36, 42, 48, 54, 60, 66, 72, 78, 84, 90, 96}
n(A) = 16

$$Logo, P(A) = \frac{n(A)}{n(S)} = \frac{16}{100} = 16\%$$

Evento B – sair múltiplo de 8:
B = {8, 16, 24, 32, 40, 48, 56, 64, 72, 80, 88, 96}
n(B) = 12

$$Logo,\ \ P(B) = \frac{n(B)}{n(S)} = \frac{12}{100} = 12\%$$

A∩B = {24, 48, 72, 96}
n(A∩B) = 4

$$Logo, P(A \bigcap B) = \frac{n(A \bigcap B)}{n(S)} = \frac{4}{100} = 4\%$$

Portanto, P(A∪B) = P(A) + P(B) – P(A∩B)

$$P(A \cup B) = \frac{16}{100} + \frac{12}{100} - \frac{4}{100} = \frac{24}{100} = 24\%$$

Teorema da soma: probabilidade da união de dois eventos

Uma urna contém 100 bolas idênticas numeradas de 1 a 100 e uma delas é escolhida ao acaso. Qual a probabilidade de:

Observarmos um número não múltiplo de 5.

O espaço amostral S é:

S = {1, 2, 3, ... , 100}

n(S) = 100

Evento C – sair múltiplo de 5:

C = {5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100}

n(C) = 20

$$Logo, P(C) = \frac{n(C)}{n(S)} = \frac{20}{100} = 20\%$$

Evento B – sair múltiplo de 8:

Como queremos a probabilidade de não ser um múltiplo de 5, então queremos o evento complementar de C.

$$P(C) + P(\overline{C}) = 1$$

$$P(\overline{C}) = 1 - \frac{20}{100} = \frac{100-20}{100} = \frac{80}{100} = 80\%$$

Probabilidade condicional
e
eventos independentes

Probabilidade condicional

Probabilidade condicional refere-se à probabilidade de um evento A sabendo que ocorreu um outro evento B e representa-se por P(A|B). (Lê-se: "probabilidade de A dependente da condição B" ou "probabilidade condicional de A dado B" ou ainda probabilidade de A condicionada por B.)

A probabilidade de A condicionada por B é definida por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$sendo: P(B) > 0$$

Nota-se que a probabilidade de A muda após o evento B ter acontecido. Isso porque o resultado de A é uma das possibilidades de B. É preciso, portanto, calcular os eventos que são comuns a B e também a A, ou seja A ∩ B.

Exemplo

Uma urna contém 15 bolas numerada de 1 a 15. Retira-se da urna uma bola ao acaso e observa-se que o número é maior que 6. Qual é a probabilidade desse número ser múltiplo da 3?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad sendo: P(B) > 0$$

A: Múltiplos de $3 = \{3, 6, 9, 12, 15\}$

B: Números $> 6 = \{7, 8, 9, 10, 11, 12, 13, 14, 15\}$

$$P(B) = \frac{9}{15}$$

$$A \cap B = \{9, 12, 15\}$$

$$P(A \cap B) = \frac{3}{15}$$

$$P(A|B) = \frac{\frac{3}{15}}{\frac{9}{15}} = \frac{3}{9} = \frac{1}{3}$$

Solução alternativa:

Redução do Espaço Amostral

$n(S) = 9$

A: Múltiplos de $3 = \{9, 12, 15\}$

$$P(A) = \frac{3}{9} = \frac{1}{3}$$

Eventos independentes

Dizemos que dois eventos A e B são independentes quando a realização de um dos eventos não afeta a probabilidade da realização do outro.

Dessa forma, para a ocorrência simultânea dos dois eventos independentes, temos:

$$P(A \cap B) = P(A).P(B)$$

Exemplo:

Lançamento de dois dados:

Probabilidade de obtermos 1 no primeiro dado = $\frac{1}{6}$

Probabilidade de obtermos 5 no segundo dado = $\frac{1}{6}$

Logo, a probabilidade de obtermos, simultaneamente, 1 no primeiro dado e 5 no segundo dado é: $\frac{1}{6} . \frac{1}{6} = \frac{1}{36}$

Exercícios

Jogou-se um dado e a face superior apresentou valor maior que dois. Qual a probabilidade do valor obtido no lançamento ser par?

Uma urna contém 12 bolas, das quais, quatro são azuis e o restante, amarelas. Retira-se da urna uma bola ao acaso e observa-se que a cor é azul. A bola **retirada é novamente colocada na urna**. A seguir, retira-se novamente uma bola da urna. Qual é a probabilidade da bola também ser azul?

Variáveis aleatórias discretas
Distribuição equiprovável
Distribuição de Bernoulli

Variáveis aleatórias discretas X Variáveis aleatórias contínuas

Diz-se que uma **variável aleatória** é discreta se todos os seus valores podem ser listados e pertencem a um conjunto finito ou infinito enumerável *.

Exemplos: número de chegadas a uma fila, número de caras em uma jogada de duas moedas, resultado da jogada de um dado.

Uma variável **aleatória é contínua** se os seus valores não podem ser listados, mas podem assumir um número infinito de valores em um intervalo finito ou infinito.

Exemplo: intervalo de tempo da vida de uma lâmpada, altura de pessoas em uma sala.

* Um conjunto é enumerável quando existe correspondência um a um entre os seus elementos e os números naturais.

Variáveis aleatórias discretas

Assim como um evento aleatório é algo que não sabemos ao certo se ocorrerá, mas cuja probabilidade de ocorrência podemos calcular, **variável aleatória** é uma que não sabemos ao certo que valor tomará, mas para a qual **podemos calcular a probabilidade de tomar determinado valor**.

Se chamarmos de **X** o **número de caras que aparecem no lançamento de uma moeda**, então **X é uma variável aleatória**.

As **variáveis aleatórias** são geralmente representadas por **letras maiúsculas**: X, Y, W, etc. e os **valores que a variável aleatória pode assumir** são geralmente representados por letras **minúsculas**: x1, x2 ... xk, por exemplo. As respectivas **probabilidades** de ocorrência de cada valor são **representadas por: p(x1), p(x2) ... p(xk)**; dessa forma, podemos escrever uma **função (x, p(x))** que é chamada de **função de probabilidade da variável aleatória X**. Também podemos usar a notação **p(x) = P(X = x)**; em que **x é o valor de X e p(x) é a probabilidade de X tomar o valor x**.

Variáveis aleatórias discretas

Exemplo:

Lançamos uma moeda duas vezes. Vamos definir a variável aleatória **X = número de caras obtidas nos dois lançamentos**.

Solução:

Se C representa cara e K representa coroa, então o espaço amostral é:

S = {(CC), (CK), (KC), (KK)}

E cada um desses resultados tem probabilidade. Podemos montar uma tabela:

Resultados	Probabilidades	X (nº de caras)
CC	1/4	2
CK	1/4	1
KC	1/4	1
KK	1/4	0

Variáveis aleatórias discretas

Temos então:

$$p(0) = P(X = 0) = P(KK) = \frac{1}{4}$$

$$p(1) = P(X = 1) = P(CK\ ou\ KC) = \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = \frac{1}{2}$$

$$p(2) = P(X = 2) = P(CC) = \frac{1}{4}$$

Podemos, então, esquematizar a distribuição de probabilidades da variável aleatória X:

xi	0	1	2	Total
p(xi)	1/4	1/2	1/4	1

Portanto, vimos que a cada ponto do espaço amostral a variável em consideração associa um valor numérico, o que corresponde em Matemática ao conceito de função, mais precisamente, a uma função definida no espaço amostral S e assumindo valores reais.

Média

Considerando a variável X do exemplo anterior (número de caras obtidas em dois lançamentos de uma moeda). Qual é a média esperada?

Da tabela de distribuição de probabilidades que construímos, observamos 25% de chance de não termos nenhuma cara, 50% de chance de obtermos uma cara e 25% de chance de conseguirmos 2 caras.

Logo, em média, esperamos ter: (0,25 x 0) + (0,50 x 1) + (0,25 x 2) = 1 cara.

(Significa que ao lançar duas moedas, esperamos que na média cara "apareça" uma vez.)

Ou seja, escrevemos a média em termos da função de probabilidade:

Média = [p(2) x 2] + [p(1) x 1] + [p(0) x 0]

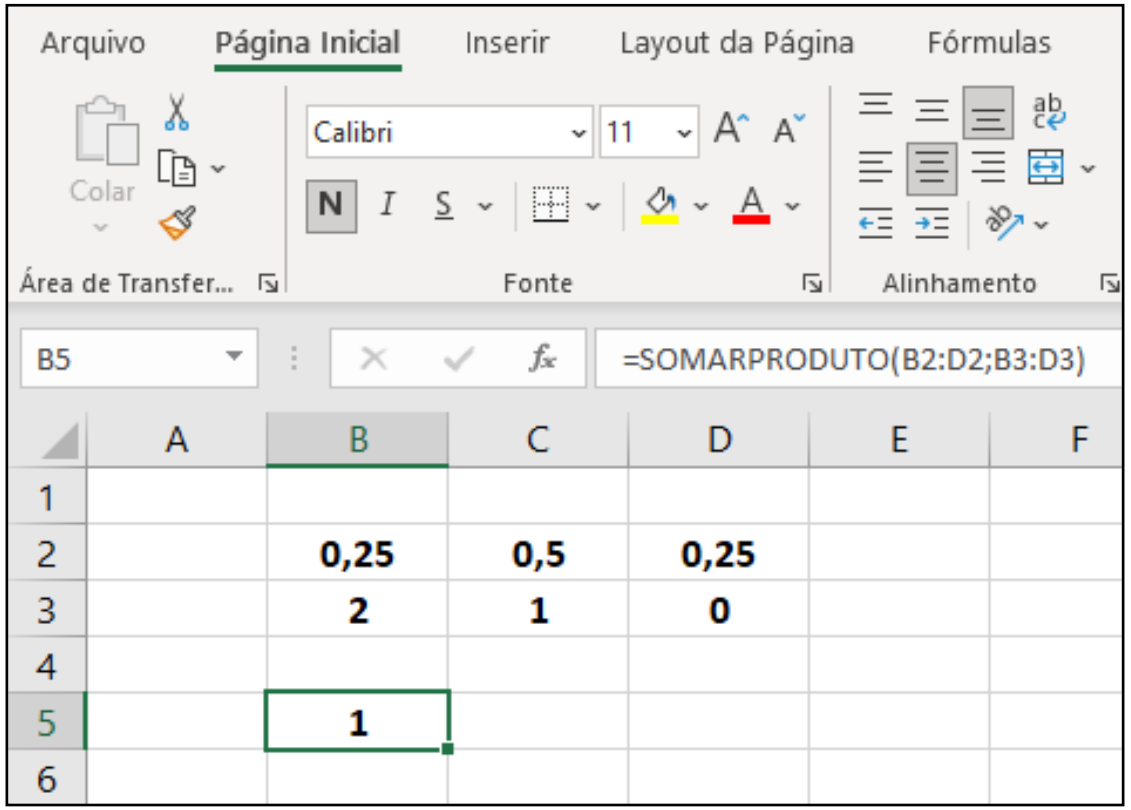
Que é o **valor esperado de X**, ou **esperança de X**, ou **média de X**.

Portanto, a média de uma variável aleatória discreta X é dada por:

$E(X) = \mu(X) = \sum x_i p_i$

Média

Calcular a média das probabilidades no Excel:



Variância e desvio padrão

Podemos também definir todas as outras medidas vistas. As mais importantes são:

Variância de X: $Var(X) = \sigma^2(X) = (x_1 - \mu(X))^2.p(x_1) + \ldots (x_k - \mu(X))^2.p(x_k) = \sum(x_i - \mu(X))^2.p(x_i)$

Desvio padrão de X: $DP(X) = \sigma(x) = \sqrt{Var(X)}$

xi	0	1	2	Total
p(xi)	1/4	1/2	1/4	1
	0,25	0,50	0,25	

$\mu(X) = (0,25 \times 0) + (0,50 \times 1) + (0,25 \times 2) = 1$

$Var(X) = (0 \times 1)^2 \times 0,25 + (1 \times 1)^2 \times 0,50 + (2 \times 1)^2 \times 0,25 = 0 + 0,50 + 1 = 1,50$

$\sigma(x) = \sqrt{Var(X)} = \sqrt{1,50} = 1,2247$

Distibuição equiprovável

Dizemos que uma distribuição de probabilidades é equiprovável se p1 = p2 = p3 = ... = pk, isso é, se todos os valores da variável aleatória tiverem a mesma probabilidade de ocorrência.

Definição: a variável aleatória discreta X, assumindo os valores x_1, x_2 ... x_k, tem distribuição equiprovável se, e somente se:

$$P(X = X_i) = p(x_i) = \frac{1}{k}, para\ todo\ i = 1, 2, \dots, k$$

Dessa forma, essa distribuição tem:

$$E(X) = \frac{1}{k} \sum x_i$$

$$Var(X) = \frac{1}{k}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{k}\right)$$

Distibuição equiprovável

Exemplo: Quando jogamos um dado, temos uma distribuição equiprovável. Assim, podemos calcular a média e a variância usando as fórmulas:

$$E(X) = \frac{1}{k}\sum x_i = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = 3{,}5$$

$$Var(X) = \frac{1}{k}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{k}\right) = \frac{1}{6}\left[(1+4+9+16+25+36) - \frac{21^2}{6}\right] = \frac{1}{6}\left[91 - \frac{441}{6}\right] =$$

$$= \frac{1}{6}\cdot\frac{105}{6} = \frac{35}{12} = 2{,}9167$$

Distribuição de Bernoulli

A distribuição de Bernoulli é aplicada a experimentos em que os resultados possíveis apresentam ou não determinada característica.

Exemplos:

- Uma moeda é lançada: o resultado é cara ou não.
- Uma peça é escolhida ao acaso de um lote de 1000 peças: ela é defeituosa ou não.
- Um dado é lançado: ocorre face 6 ou não.

Em cada experimento podem ocorrer apenas dois resultados: sucesso (ocorrência de cara, peça defeituosa, face 6) ou fracasso (ocorrência de coroa, peça boa, face diferente de 6).

Observação: os nomes sucesso e fracasso não têm, aqui, o significado que lhes damos na linguagem cotidiana. Servem apenas para designar os dois resultados de cada experimento. Dessa forma, no primeiro exemplo, poderíamos chamar de sucesso o resultado coroa e de fracasso o resultado cara.

Distribuição de Bernoulli

Para cada experimento, podemos definir uma variável aleatória X que assume apenas dois valores: o valor 1, se ocorre sucesso, e o valor 0, se ocorre fracasso.

A probabilidade de ocorrer sucesso em cada caso será indicada por p e, consequentemente, a probabilidade de ocorrer fracasso será **q = 1 – p**.

Definição: a variável aleatória discreta X, assumindo apenas os valores 1 e 0 com a função de probabilidade:

x	0	1	Total
p(x)	1 - p	p	1

É chamada variável aleatória de Bernoulli.

Dessa forma, essa distribuição tem:

E(x) = p

Var(X) = p - p2 = p(1 - p)

Observação: Experimentos que resultam numa variável aleatória de Bernoulli são chamados ensaios de Bernoulli.

Distribuição de Bernoulli

Exemplo: Consideremos o lançamento de uma dado e observação da ocorrência ou não da face 6:

x	0	1	Total
p(x)	5/6	1/6	1

$$E(X) = \frac{1}{6}$$

$$Var(X) = \frac{1}{6}\left(1 - \frac{1}{6}\right) = \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$$

Distribuição binomial

Distribuição de Poisson

Distribuição binomial

Esta distribuição é aplicada em casos de experimentos repetidos, onde existem dois possíveis resultados: sucesso ou fracasso, cara ou coroa, item defeituoso ou item não defeituoso, etc. A probabilidade de cada resultado pode ser calculada utilizando a regra da multiplicação, ou com o uso do diagrama de árvore, porém é muito mais simples e eficiente utilizar uma equação generalizada.

Assim, uma variável aleatória poderá ter sua distribuição de probabilidade modelada de forma binomial caso atenda os seguintes pressupostos:

- o resultado é completamente aleatório;
- existem somente dois possíveis resultados (experimento Bernoulli);
- todas as tentativas possuem a mesma probabilidade para um resultado em particular (as tentativas ou realizações do experimento são independentes);
- existe uma probabilidade **p** de sucesso **constante em cada tentativa**;
- o número de tentativas, **n**, é um valor fixo, um número **inteiro** e **positivo**.

Distribuição binomial

A variável aleatória discreta X corresponde ao número de sucessos em n tentativas independentes.

O modelo é dado pela seguinte função massa de probabilidade (PMF):

$$p_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

onde,

$C_{n,x}$ = número de cobinação de n elementos x a x

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

para $x = 0, 1, \ldots, n$:

Exemplos

No lançamento de uma moeda cinco vezes, qual a probabilidade de serem obtidas exatamente duas caras?

$$Solução: n = 5; x = 2; p = \frac{1}{2}$$

$$P(x = 2) = \frac{5!}{2!(5-2)!} . \frac{1}{2}^2 . \left(1 - \frac{1}{2}\right)^{5-2} = \frac{20}{2} . \frac{1}{4} . \frac{1}{8} = \frac{20}{128} = \frac{5}{16} = 0{,}3125 = 31.25\%$$

Usando o R para resolver:

n = 5 - número de tentativas
p = 0.5 - probabilidade de resultar cara
x = 3 - contagem dos sucessos

```
> dbinom(2,5,0.5)
[1] 0.3125
```

Exercício

Considere uma cartela com meia dúzia de ovos e que cada um tem 5% de probabilidade de ser quebrado durante o manuseio, no transporte ou nas gôndolas do mercado. Qual é a probabilidade de uma caixa chegar até o consumidor com dois ovos quebrados?

Distribuição de Poisson

A distribuição de Poisson (publicada por Siméon Denis Poisson, em 1838) é uma distribuição discreta de probabilidade aplicável a ocorrências de um número de eventos em um intervalo específico (tempo, comprimento, superfície, volume etc.).

Para aplicar a distribuição de Poisson, três aspectos devem ser observados:

- O experimento calcula quantas vezes um evento ocorre em um determinado intervalo (tempo, área, volume, etc.);
- A probabilidade do evento ocorrer é a mesma para cada intervalo;
- O número de ocorrências de um intervalo é independente do outro.

A distribuição Poisson tem apenas um parâmetro, **λ** (lambda) que é interpretado como uma taxa média de ocorrência do evento, e a probabilidade de ocorrerem exatamente **x** eventos é dada por:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{n!}$$

onde:

e (número de Euler) ≈ 2,7183

$\lambda > 0$

Distribuição de Poisson

Exemplo: Um usuário recebe em média 5 mensagens em seu WhatsApp a cada 15 minutos. Qual a probabilidade desse usuários receber 8 mensagens no mesmo intervalo (15 minutos)?

λ = 5

n = 8

***e* = 2.7183**

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{n!} \Rightarrow P(X = x) = \frac{5^8 e^{-5}}{8!} \Rightarrow 0.065278$$

Usando o R para resolver:

```
> dpois(8,5)
[1] 0.06527804 #6.5278%
```

Exercício

Supondo que as consultas em um banco de dados ocorrem de forma independente e aleatória, com uma taxa média de três consultas por minuto. Qual a probabilidade que no próximo minuto ocorram cinco consultas?

Distribuição Geométrica
Distribuição Hipergeométrica

Distribuição Geométrica

Consideremos que temos uma série de ensaios de Bernoulli (amostragens independentes com probabilidade constante "p" de sucesso cada uma). Agora, em vez de usarmos um número fixo de amostragens, como fizemos na distribuição binomial, elas serão realizadas até que o sucesso seja obtido.

Dizemos que a variável aleatória "X" (número de tentativas até que o sucesso seja obtido) tem distribuição geométrica com parâmetro "p", onde 0 < p < 1, se a sua função de probabilidade é dada pela fórmula:

$$P(X = k) = p(1-p)^{k-1}, \qquad k = 0, 1, 2, \ldots$$

k = nº de tentativas até o aparecimento do primeiro sucesso.

p = probabilidade de sucesso no ensaio individual.

1 – p = a probabilidade de fracasso no ensaio individual.

Essa distribuição é utilizada para modelar, por exemplo, o número de repetições do lançamento de uma moeda até a primeira ocorrência de cara; o instante de tempo medido em unidades de tempo inteira até a chegada do próximo consumidor em uma fila, ou até a próxima falha em um equipamento.

Distribuição Geométrica

A **esperança** de uma variável aleatória que tem distribuição geométrica com parâmetro "p" é dada por:

$$E(X) = \frac{1}{p}$$

p = probabilidade de sucesso no ensaio individual.

A **variância** de uma variável aleatória que tem distribuição geométrica com parâmetro "p" é dada por:

$$Var(X) = \frac{1-p}{p^2}$$

1 – p = a probabilidade de fracasso no ensaio individual.

p = probabilidade de sucesso no ensaio individual.

Distribuição Geométrica

Exemplo

Um aluno do curso de Educação Física faz vários arremessos livres na cesta de basquete da sua faculdade e só para quando consegue acertar um.

A probabilidade de ele acertar um arremesso é 0,2 (20%). Nessas condições, responda:

Qual a probabilidade de acertar o arremesso na 2ª tentativa?

p = 0,2 (probabilidade de sucesso no ensaio individual)

1 - p = 1 - 0,2 = 0,8 (probabilidade de fracasso no ensaio individual)

k = 2 (nº de tentativas até o aparecimento do primeiro sucesso)

$$P(X = k) = p(1 - p)^{k-1}$$

$$P = (X = 2) = 0{,}2\ (0{,}8)^{2-1} = 0{,}16 = 16\%$$

Distribuição Geométrica

Exercício

Um aluno do curso de Educação Física faz vários arremessos livres na cesta de basquete da sua faculdade e só para quando consegue acertar um.

A probabilidade de ele acertar um arremesso é 0,2 (20%). Nessas condições, responda:

Qual a probabilidade de acertar o arremesso na 4ª tentativa?

p = 0,2 (probabilidade de sucesso no ensaio individual)

1 - p = 1 - 0,2 = 0,8 (probabilidade de fracasso no ensaio individual)

k = 4 (nº de tentativas até o aparecimento do primeiro sucesso)

$$P(X = k) = p(1 - p)^{k-1}$$

$$P = (X = 2) = 0{,}2\ (0{,}8)^{4-1} = 02\ .\ 0{,}512 = 0{,}1024 = 10{,}24\%$$

Distribuição Geométrica

Exemplo

Um aluno do curso de Educação Física faz vários arremessos livres na cesta de basquete da sua faculdade e só para quando consegue acertar um.

A probabilidade de ele acertar um arremesso é 0,2 (20%). Nessas condições, responda:

Quantas tentativas, em média, terá que realizar até acertar o 1º arremesso?

p = 0,2 (probabilidade de sucesso no ensaio individual)

$$E(X) = \frac{1}{p} = \frac{1}{0{,}2} = 5$$

Distribuição Geométrica

Exercício

Um aluno do curso de Educação Física faz vários arremessos livres na cesta de basquete da sua faculdade e só para quando consegue acertar um.

A probabilidade de ele acertar um arremesso é 0,25 (25%). Nessas condições, responda:

Quantas tentativas, em média, terá que realizar até acertar o 1º arremesso?

p = 0,2 (probabilidade de sucesso no ensaio individual)

$$E(X) = \frac{1}{p} = \frac{1}{0{,}25} = 4$$

Distribuição Geométrica

Exemplo

Um casal com problemas para engravidar recorreu a uma técnica de inseminação artificial no intuito de conseguir o primeiro filho.

A eficiência da referida técnica é de 25% e o custo de cada inseminação é R$2000,00. Nessas condições, responda:

Qual a probabilidade de que o casal obtenha êxito na 3ª tentativa?

p = 0,25 (probabilidade de sucesso no ensaio individual)

1 - p = 1 - 0,25 = 0,75 (probabilidade de fracasso no ensaio individual)

k = 3 (nº de tentativas até o aparecimento do primeiro sucesso)

$$P(X = k) = p(1 - p)^{k-1}$$

$$P = (X = 3) = 0{,}25\ (0{,}75)^{3-1} = 0{,}14 = 14\%$$

Distribuição Geométrica

Exemplo

Um casal com problemas para engravidar recorreu a uma técnica de inseminação artificial no intuito de conseguir o primeiro filho.

A eficiência da referida técnica é de 25% e o custo de cada inseminação é R$2000,00. Nessas condições, responda:

Qual o custo esperado deste casal para obter o primeiro filho?

p = 0,25 (probabilidade de sucesso no ensaio individual)

R$ 2000,00 (custo de cada inseminação)

$$E(X) = \frac{1}{p} = \frac{1}{0{,}25} = 4$$

Como o valor esperado é 4. Portanto, significa que o casal deve esperar ter o filho somente na 4ª tentativa gastando:

4 x R$ 2000,00 = R$ 8000,00

Distribuição Hipergeométrica

A distribuição hipergeométrica descreve o número de sucessos em uma sequência de "n" amostras de uma população finita, sem reposição, dividida segundo dois atributos.

Nesses casos, não podemos usar a distribuição binomial, pois não satisfazem o critério de probabilidade constante (p) em todas as provas do experimento, uma vez que os eventos são dependentes entre si. Dessa forma, **a binomial corresponde ao esquema de extrações com reposição**, enquanto **na hipergeométrica o esquema é de extrações sem reposição**.

Exemplos

Em uma população de "N" objetos, "D" objetos têm o atributo "A" e "N" – "D" têm o atributo "B". Um grupo de "n" elementos é escolhido ao acaso sem reposição. A distribuição hipergeométrica calcula a probabilidade de que esse grupo contenha "k" elementos com o atributo "A".

Considere uma carga com "N" objetos dos quais "D" têm defeito. A distribuição hipergeométrica descreve a probabilidade de que em uma amostra de "n" objetos distintos escolhidos da carga aleatoriamente, sem reposição, exatamente "k" objetos sejam defeituosos.

Distribuição Hipergeométrica

Em geral, se uma variável aleatória "X" (número de elementos na amostra que têm o atributo "A") segue uma distribuição hipergeométrica com parâmetros "N" (total de objetos), "D", e "n" (objetos distintos escolhidos aleatoriamente), então, a probabilidade de termos exatamente "k" sucessos é dada por:

$$P(X = k) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{n}{k}}, \qquad 0 \le k \le \min(D, n)$$

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ = coeficiente binomial [95]

$\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$ = possíveis amostras sem reposição

$\binom{D}{k} = \frac{D!}{k!\,(D-k)!}$ = maneiras de escolher "k" objetos com atributo "A"

$\binom{N-D}{n-k} = \frac{N!}{n!\,(N-n)!} - \frac{D!}{k!\,(D-k)!}$ = maneiras de preencher o resto da amostra com objetos que têm atributo "B"

Distribuição Hipergeométrica

Em um lote de cem peças, dez são defeituosas. Escolhendo-se cinco peças sem reposição, qual a probabilidade de não se obter peças defeituosas?

Resolução

Temos: N = 100; D = 10 e n = 5. Queremos a probabilidade de que nenhuma peça seja defeituosa, ou seja, k = 0:

$$P(X=0) = \frac{\binom{10}{0}\binom{100-10}{5-0}}{\binom{100}{5}} = \frac{\binom{10}{0}\binom{90}{5}}{\binom{100}{5}} = \frac{\frac{10!}{0!(10-0)!}\cdot\frac{90!}{5!(90-5)!}}{\frac{100!}{5!(100-5)!}} =$$

$$1.\frac{\frac{90!}{5!\ 85!}}{\frac{100!}{5!\ 95!}} = \frac{\frac{90.89.88.87.86.85!}{5!\ 85!}}{\frac{100.99.98.97.96.95!}{5!\ 95!}} = \frac{90.98.88.87.86}{100.99.98.97.96} = 0{,}584$$

Modelos Probabilísticos Contínuos

Modelos Probabilísticos Contínuos

A distribuição contínua descreve as probabilidades dos possíveis valores de uma variável aleatória contínua.

Uma variável aleatória contínua apresenta um conjunto de valores possíveis (conhecidos como intervalos) que é infinito e incontável.

As probabilidades de variáveis aleatórias contínuas são definidas como a área sob a curva da sua distribuição.

Assim, apenas faixas de valores podem ter uma probabilidade diferente de zero. A probabilidade de que uma variável aleatória contínua seja exatamente igual a algum valor é sempre zero.

Portanto, é possível calcular a probabilidade de que um homem pesa entre 70 e 80 quilos. No entanto, a probabilidade de um homem pesar exatamente 70 quilos para a precisão infinita é zero. Porque a área sob a curva em um único ponto (que não tem nenhuma largura) é zero.

Modelos Probabilísticos Contínuos

Alguns dos principais modelos probabilísticos contínuos são:

- Uniforme Contínuo
- Normal ou Gaussiana
- Qui-quadrado
- t de Student
- F de Snedecor
- Gama
- Beta
- Exponencial
- Weibull
- Logística

Nas próximas aulas abordaremos dois modelos utilizados com muita frequência:

- Distribuição Normal ou Gaussiana
- Distribuição Exponencial

Distribuição Normal (Gaussiana)

A distribuição normal é um dos principais modelos de probabilidade para **variáveis aleatórias contínuas**.

Exemplos de variáveis aleatórias contínuas: comprimento, peso, temperatura, pressão, tempo, etc.

Em uma coleta aleatória de dados de fontes independentes, geralmente observa-se que a distribuição dos dados é normal. Ou seja, ao traçar um gráfico com o valor da variável no eixo horizontal **x** e a contagem dos valores no eixo vertical **y**, obtemos uma curva em forma de sino. O **centro da curva** representa a **média** do conjunto de dados. No gráfico, cinquenta por cento dos valores ficam à esquerda da média e os outros cinquenta por cento ficam à direita. Isso é conhecido como distribuição normal.

Variáveis aleatórias com diferentes médias μ e variâncias σ^2 podem ser modeladas pelas funções de densidade de probabilidade normal. A média μ determina o centro do gráfico em forma de sino e a variância σ^2 determina a largura da distribuição.

Distribuição Normal (Gaussiana)

Gerando valores aleatórios para uma distribuição normal:

A função **`rnorm()`** abaixo gera **dois conjuntos diferentes** com 10 valores aleatórios de uma distribuição normal:

```
> rnorm(10)
 [1]  0.3599659 -1.5875803 -0.3890161  0.1307562  0.5890599
 [6] -2.1539719  1.4859364 -0.4329219  0.6385853  1.9210244
> rnorm(10)
 [1] -1.15513692 -2.11647932 -1.13449016  1.61930642 -1.02581667
 [6] -0.45670775 -0.05491379 -0.58710610  0.34518977  1.49852742
```

A função **`set.seed()`** é utilizada para reproduzir os resultados dos geradores de números pseudo-aleatórios. Isto é importante, por exemplo, em simulações ou para ajustar um modelo de classificação com dois subconjuntos dos dados, um para treinar o modelo e outro para o testar. O valor do seed (semente) não é importante desde que a sua utilização seja consistente.

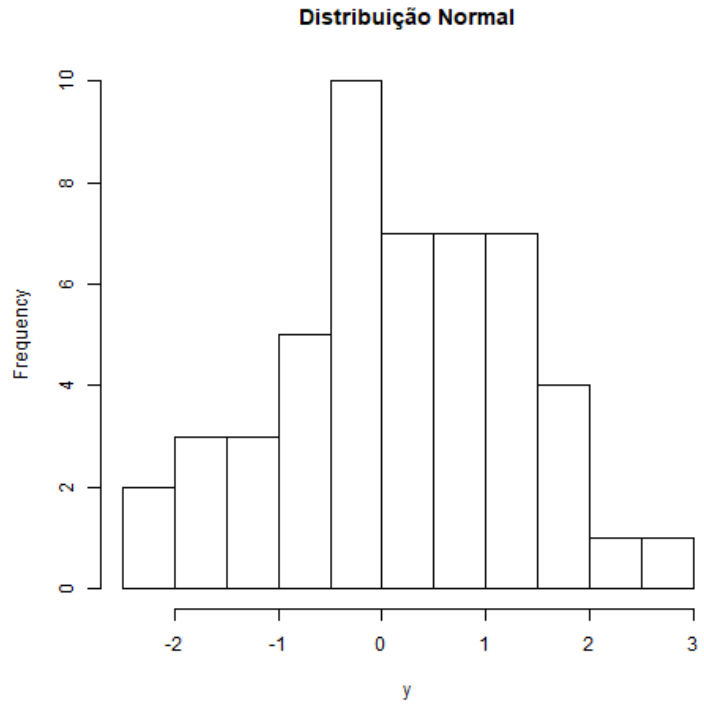
A função **`rnorm()`** abaixo gera **dois conjuntos iguais** com 10 valores aleatórios de uma distribuição normal:

```
> set.seed(128); rnorm(10)
 [1]  0.596771824  0.482611999  1.664405236 -0.025995522  1.209026965
 [6]  0.580352168 -0.458345286  0.004552681  0.209954522  1.511140204
> set.seed(128); rnorm(10)
 [1]  0.596771824  0.482611999  1.664405236 -0.025995522  1.209026965
 [6]  0.580352168 -0.458345286  0.004552681  0.209954522  1.511140204
```

Distribuição Normal (Gaussiana)

`rnorm()` Gera números aleatórios cuja distribuição é normal. Usa o tamanho da amostra como entrada e gera os números aleatórios. O histograma apresenta a distribuição dos números gerados.

```
# Cria uma amostra de com 50 números normalmente distribuídos
y <- rnorm(50)
# Gera o histograma para amostra criada
hist(y, main = "Distribuição Normal")
```

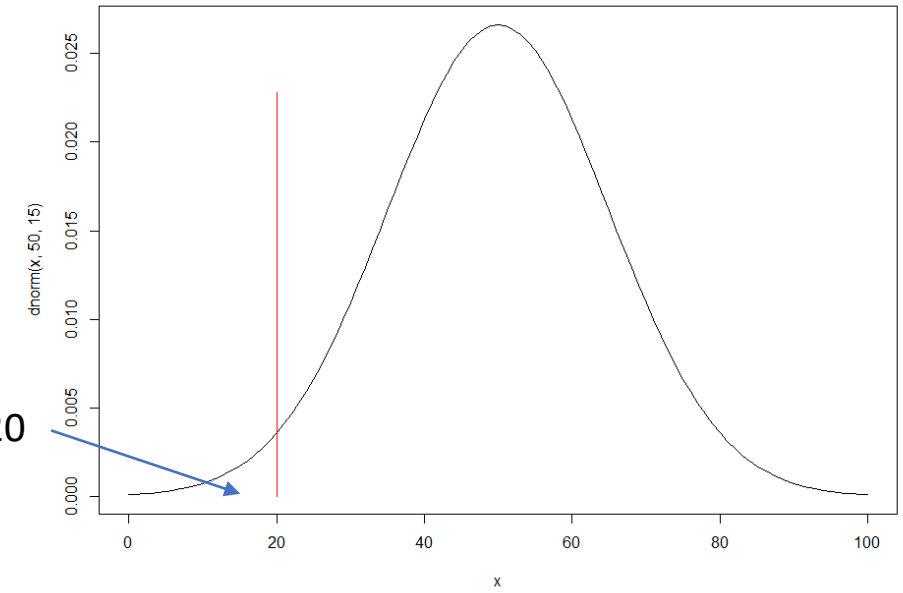


Distribuição Normal (Gaussiana)

Exemplo 1: Qual a probabilidade de ocorrência de um valor **menor** que 20 em uma distribuição normal de média igual a 50 e desvio padrão igual a 15?

```
# x=20; mean=50; sd=15

pnorm(20, 50, 15)

[1] 0.02275013  # ≈ 2.28%

curve(dnorm(x, 50, 15),0,100)

lines(c(20,20),c(0,0.0228),col="red")
```

Probabilidade de valor < 20



Distribuição Normal (Gaussiana)

Exemplo 2: Qual a probabilidade de ocorrência de um valor **maior** que 20 em uma distribuição normal de média igual a 50 e desvio padrão igual a 15?

```
1-pnorm(20, 50, 15)
```

```
[1] 0.9772499 # ≈ 97.72%
```

Distribuição Exponencial

A distribuição exponencial, muito utilizada em engenharia, descreve o tempo de chegada de uma sequência de eventos independentes recorrentes aleatoriamente.

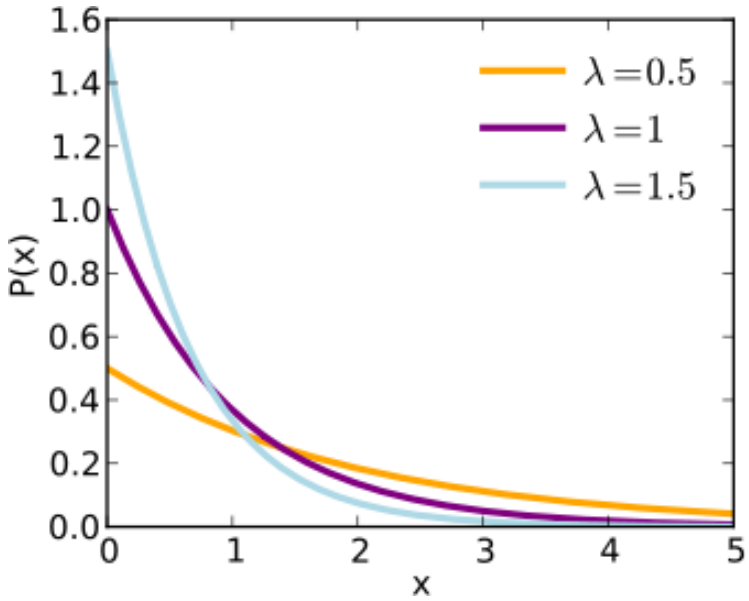
O modelo ajusta-se bem a dados de tempo para ocorrência de um evento. Exemplo: tempo para atendimento de uma chamada.

É uma distribuição também muito utilizada na prática para modelar tempo de falha de objetos. Por exemplo, pode ser usada para modelar o tempo que demora até uma lâmpada falhar.

Se μ é o tempo médio de espera para a próxima recorrência do evento, sua função de densidade de probabilidade é:

$$f(x) = \begin{cases} \lambda e^{\lambda . x} & se\ x \geq 0 \\ 0 & se\ x < 0 \end{cases}$$

em que λ é o parâmetro de taxa da distribuição e deve satisfazer λ > 0. Neste caso, λ é o tempo médio de vida e x é um tempo de falha.
O parâmetro deve ter a mesma unidade do tempo da falha x. Isto é, se x é medido em horas, λ também será medido em horas.



Distribuição Exponencial

O tempo médio de checkout de um caixa de supermercado é de três minutos. Qual a probabilidade do checkout de um cliente ser concluído pelo caixa em menos de dois minutos?

Tomando como base o enunciado, verifica-se que a taxa de processamento do checkout é igual a um dividido pelo tempo médio de conclusão do checkout. Portanto, a taxa de processamento é de 1/3 checkouts por minuto. Em seguida, aplica-se a função pexp da distribuição exponencial com taxa = 1/3.

```
> pexp(2, rate=1/3)
[1] 0.4865829
```

A probabilidade de concluir um checkout em menos de dois minutos no caixa é de 48.7%.

A distribuição exponencial é a única distribuição contínua que possui **perda de memória**.

Portanto, suponha que o cliente já esperou um minuto para conclusão do checkout. Isto significa que a probabilidade de ser atendido no próximo minuto seja maior do que no primeiro minuto? A resposta é não. Não importa o quanto tempo o cliente tenha esperado. Esta propriedade da distribuição exponencial é chamada de perda de memória.

Regressão Linear Simples
e
Correlação

Regressão Linear Simples

Regressão linear simples

A regressão linear simples constitui uma tentativa de estabelecer uma equação matemática linear (reta) com apenas uma variável dependente que descreva o relacionamento entre duas variáveis.

A reta de regressão (equação linear) apresenta como principais características:

- **Coeficiente angular** (α) da reta, dado pela tangente da reta;
- **Coeficiente linear** (β), a cota da reta em determinado ponto (o valor de y quando x for igual a zero).

$$y = \alpha + \beta x + \varepsilon$$

Onde:

x: variável independente que busca explicar y

y: variável dependente a ser prevista

ε: erro que corresponde ao desvio entre o valor real e o aproximado (pela reta) de y

Equação linear

A equação linear pode ser obtida no R através da função **`lm()`** que calcula a regressão linear simples.

A regressão linear simples é obtida por meio do seguinte comando:

```
lm(y~x,data)
```

lm: linear model

y~x: y depende de x

data: conjunto de dados (data.frame) valores das "colunas" que correspondem a **x** e **y**

Equação linear

Exemplo:

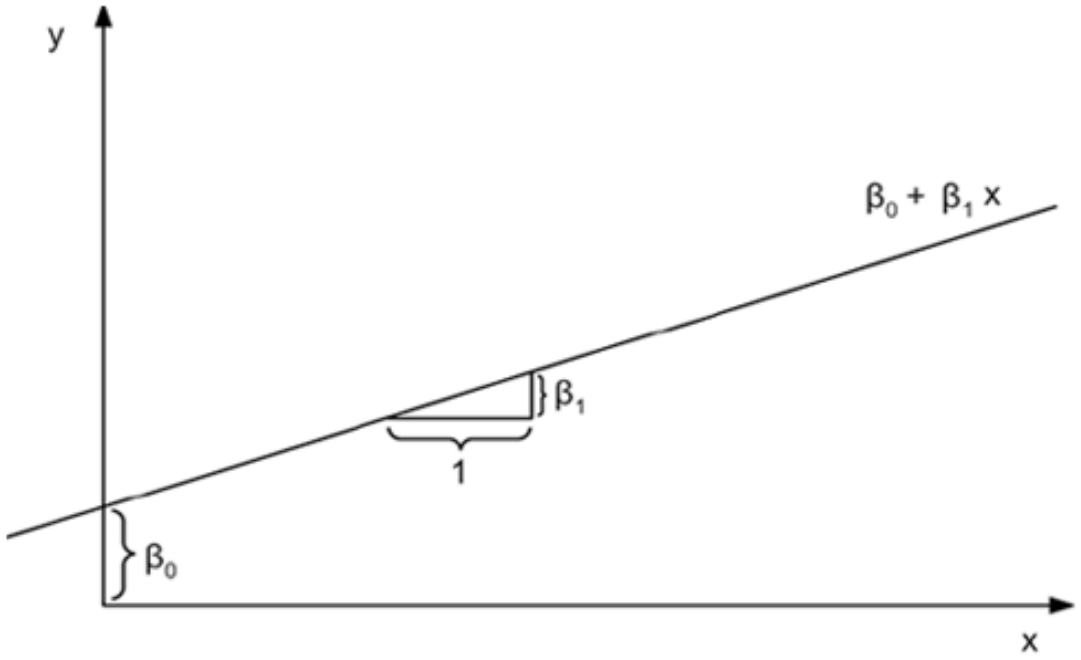
```
> x <- c(205,225,305,380,560,600,685,735)
> y <- c(15,20,21,22,22,25,28,28)
> dados <- data.frame(x,y) #cria um data.frame
> regressão <- lm(y~x,data=dados) #ou regressão <- lm(y~x)
> regressão

Call:
lm(formula = y ~ x, data = dados)

Coefficients:
(Intercept)            x
   13.85362      0.01899
```

Equação da reta: $y = 13.85362 + 0.01899x$

Equação linear

O parâmetro β_0 é chamado intercepto ou coeficiente linear e representa o ponto em que a reta corta o eixo dos y's, quando $x = 0$.

O parâmetro β_1 representa a inclinação da reta e é chamado de coeficiente de regressão ou coeficiente angular.



Equação linear

regressão é uma lista com os seguintes elementos:

```
> names(regressão)
[1] "coefficients"  "residuals"     "effects"     "rank"
[5] "fitted.values" "assign"        "qr"          "df.residual"
[9] "xlevels"       "call"          "terms"       "model"
```

Valores preditos da variável resposta para cada elemento da amostra (previsão):

```
> regressão$fitted.values
       1        2        3        4        5        6        7        8
17.74673 18.12655 19.64582 21.07013 24.48847 25.24811 26.86233 27.81187
```

Erro ou resíduos (valor observado - valor predito) para cada ponto da amostra:

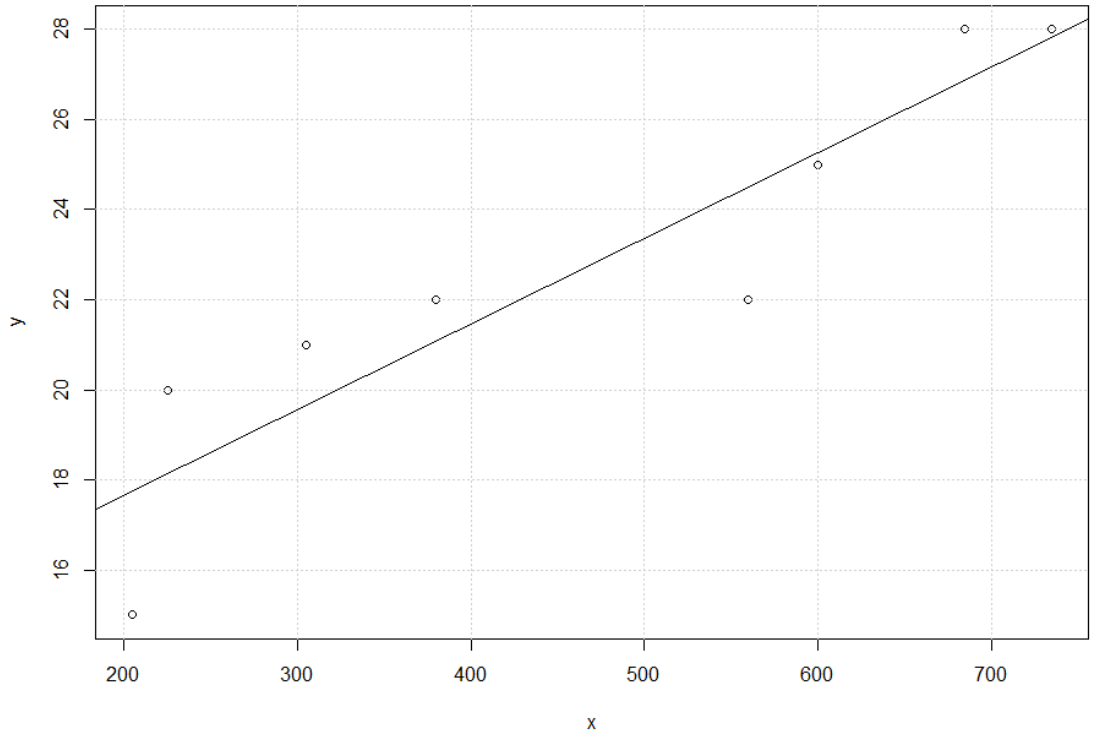
```
> regressão$residuals
 1          2          3          4          5          6          7          8
-2.7467348  1.8734490  1.3541839  0.9298729 -2.4884736 -0.2481061  1.1376747  0.1881341
```

Estimativa dos coeficientes da regressão:

```
> regressão$coefficients
(Intercept)           x
13.85361800  0.01899081
```

Equação linear

```
z = plot(x,y)
grid(z) #aplicando grid ao gráfico
abline(regressão)
```



Correlação

Coeficiente de correlação (r)

O gráfico de dispersão pode indicar se a correlação linear é positiva, negativa ou a inexistência de correlação.

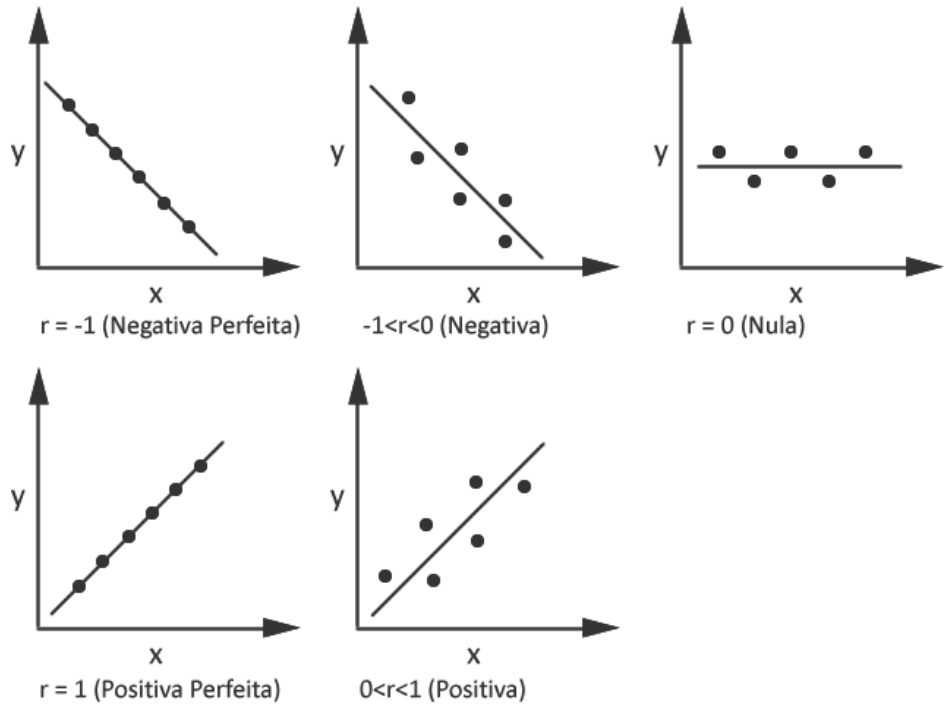
Através do coeficiente de correlação pode-se identificar o tipo de correlação, conforme a seguinte tabela:

r	Tipo	Descrição
r = -1	**Perfeita Negativa**	Os pontos estão perfeitamente alinhados, porém os valores crescentes de x associados a valores decrescentes de y
-1 < r < 0	**Negativa**	Os valores crescentes de x estão associados a valores decrescentes de y
r = 0	**Não há correlação**	Os valores de x e y ocorrem independentemente
0 < r < 1	**Positiva**	Os valores crescentes de x estão associados a valores crescentes de y
r = 1	**Positiva Perfeita**	Os pontos estão perfeitamente alinhados e os valores crescentes de x estão associados a valores crescentes de y

Coeficiente de correlação linear (r)

O coeficiente de correlação demonstra que a correlação será tanto mais forte quanto mais próximo o coeficiente estiver de –1 ou +1, e será tanto mais fraca quanto mais próximo o coeficiente estiver de zero.

r	Tipo
-1	Negativa perfeita
-1 < r < -0.7	Negativa forte
-0.7 < r -0.5	Negativa moderada
-0.5 < r < 0	Negativa fraca
0	Nula
0 < r < 0.5	Positiva fraca
0.5 < r < 0.7	Positiva moderada
0.7 < r < 1	Positiva forte
r = 1	Positiva perfeita



Coeficiente de correlação linear (r)

Exemplo:

```
> x <- c(205,225,305,380,560,600,685,735)
> y <- c(15,20,21,22,22,25,28,28)
> cor(x,y)
[1] 0.9155359
```

r = +0.916 indica uma correlação **positiva forte**, ou seja, a variável dependente **y** cresce quase na mesma proporção que a variável independente **x**.

Coeficiente de determinação (r^2)

O coeficiente de determinação explica o grau de ajuste do modelo, isto é, o percentual de variação de y que é explicado pela variabilidade de x. Seu valor varia de 0 a 1.

O valor do coeficiente de determinação corresponde ao valor do coeficiente de variação elevado ao quadrado.

Exemplo:

```
> x <- c(205,225,305,380,560,600,685,735)
> y <- c(15,20,21,22,22,25,28,28)
> cor(x,y)^2
[1] 0.8382059 #aproximadamente 84%
```

84% da variação da variável dependente **y** é explicada pela variável independente **x**. Os outros 16% possuem causas aleatórias desconhecidas (independentes de **x**).

Coeficiente de correlação linear (r)

```
> x <- c(-1,2,3,4)
> y <- c(-1,5,7,9)
> cor(x,y)
[1] 1
```

r = +1 indica uma **correlação perfeita positiva** entre as duas variáveis **x** e **y**.

$$y = \alpha + \beta x + \varepsilon$$

$$\beta = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\alpha = \frac{\sum y - \beta \sum x}{n}$$

$$n = 4$$

	x	y	xy	x^2	y^2
	-1	-1	1	1	1
	2	5	10	4	25
	3	7	21	9	49
	4	9	36	16	81
Σ	8	20	68	30	156

$$\beta = \frac{4 \cdot 68 - 8 \cdot 20}{4 \cdot 30 - (8)^2} = \frac{112}{56} = 2$$

$$\alpha = \frac{20 - 2 \cdot 8}{4} = \frac{4}{4} = 1$$

$$y = 1 + 2x + 0$$

Testes de Hipóteses

Testes de Hipóteses

Usar dados de uma amostra para tentar dizer algo sobre a população que lhes deu origem é um importante ramo da estatística, chamado **inferência estatística**, e o teste de hipótese é uma das suas principais ferramentas.

Um **teste de hipótese** é uma técnica estatística cujo **objetivo é verificar se uma dada amostra de dados é, ou não, compatível com uma hipótese feita sobre a população que lhe deu origem**.

O teste de hipótese coloca lado-a-lado duas hipóteses sobre a população que deu origem à amostra de dados: uma hipótese inicial, ou hipótese nula, e uma hipótese alternativa, designadas por $\mathbf{H_0}$ e $\mathbf{H_1}$ respectivamente.

É retirada uma amostra da população, cuja informação será tratada para encontrar evidência para se rejeitar, ou não a hipótese nula.

Caso ocorra a rejeição da hipótese nula, deve-se considerar, à partir desta constatação, a hipótese alternativa.

Teste t de Student

Teste t de Student

O test t de Student foi introduzido em 1908 por William Sealy Gosset, matemático e estatístico que trabalhava na cervejaria Guinness, em Dublin, na Irlanda.

Gosset desenvolveu o teste t, a princípio, para monitorar a qualidade da cerveja.

O teste t pode ser aplicado quando o tamanho da amostra é pequeno. Isso permite que se façam inferências usando um menor número de elementos, reduzindo os custos da pesquisa.

O uso de métodos estatísticos na fabricação da cerveja era considerado um segredo industrial. Portanto, quando Gosset publicou o artigo sobre o teste t na revista acadêmica Biometrika em 1908, teve de usar um pseudônimo "Student" e, por isso, o teste t passou a ser conhecido como **teste t de Student**.

Teste t de Student

A pluviosidade média em uma região é: 70.6

O vetor pluv apresenta mês-a-mês a pluviosidade em um determinado ano com relação a uma determinada região:

```
pluv = c(110,100,60,80,70,18,17,17,42,89,108,143)
```

Observando-se que a média mensal de pluviosidade é de 70.6 mm, terá o ano "pluv" sido excecionalmente chuvoso?

A hipótese nula contém sempre uma igualdade e a hipótese alternativa contém sempre uma desigualdade.

Testes com a alternativa diferente são denominados bilaterais. Testes com alternativa menor são designados unilaterais esquerdos, e com alternativa maior são designados unilaterais direitos.

As hipóteses podem ser formalmente descritas assim:

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0 \Rightarrow H_0 : \mu = 70.6 \text{ vs } H_1 : \mu \neq 70.6 \quad \text{(Bilateral)}$$

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu < \mu_0 \Rightarrow H_0 : \mu = 70.6 \text{ vs } H_1 : \mu < 70.6 \quad \text{(Unilateral esquerdo)}$$

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu > \mu_0 \Rightarrow H_0 : \mu = 70.6 \text{ vs } H_1 : \mu > 70.6 \quad \text{(Unilateral direito)}$$

Teste t de Student

Utilizando o R para resolver $H_0 : \mu = 70.6$ vs $H_1 : \mu \neq 70.6$ (teste bilateral)

```
t.test(pluv, mu = 70.6)
```

```
data:  pluv
t = 0.047327, df = 11, p-value = 0.9631
alternative hypothesis: true mean is not equal to 70.6
95 percent confidence interval:
 44.81350 97.51983
sample estimates:
mean of x
 71.16667
```

O **p-value** (valor de prova) de **96.3%** indica "não rejeição" de **H_0**, ou seja, não há razão para rejeitar a hipótese de que o ano "pluv" tenha sido um ano de pluviosidade normal.

Os limites entre não rejeição e rejeição situa-se entre 1% e 10% de **p-value**. Ou seja, para p-values menores que 1% rejeita-se a hipótese nula. Para valores maiores que 10% não se costuma rejeitar. O intervalo entre 1 e 10% é uma "zona cinzenta", pois o julgamento fica à critério do pesquisador.

Teste t de Student

Utilizando o R para resolver $H_0 : \mu = 70.6$ vs $H_1 : \mu < 70.6$ (teste unilateral esquerdo)

```
t.test(pluv, mu = 70.6, alternative = "less")

data:  pluv
t = 0.047327, df = 11, p-value = 0.5184
alternative hypothesis: true mean is less than 70.6
95 percent confidence interval:
     -Inf 92.66942
sample estimates:
mean of x
 71.16667
```

O **p-value** (valor de prova) de **51.8%** indica "não rejeição" de **H_0**.

Teste t de Student

Utilizando o R para resolver $H_0 : \mu = 70.6$ vs $H_1 : \mu > 70.6$ (teste unilateral direito)

```
t.test(pluv, mu = 70.6, alternative = "greater")
```

```
data:  pluv
t = 0.047327, df = 11, p-value = 0.4816
alternative hypothesis: true mean is greater than 70.6
95 percent confidence interval:
 49.66391      Inf
sample estimates:
mean of x
 71.16667
```

O **p-value** (valor de prova) de **48.2%** indica "não rejeição" de **H_0**.

Teste de Shapiro-Wilk

Existem alguns pressupostos teóricos que devem ser observados para que o resultado (p-value) seja confiável.

Para amostras pequenas, $N \leq 30$, a população em estudo deve apresentar uma distribuição normal (gaussiana). Distribuição normal é uma distribuição contínua, frequentemente associada a características físicas (peso, altura, temperatura, etc.).

Para amostras grandes, $N > 30$, não há restrições em relação à população.

A amostra "pluv" é pequena ($N = 12$) e para verificar se é uma distribuição normal deve-se realizar um teste de hipóteses preliminar, o teste de Shapiro-Wilk, disponível no R.

Teste de Shapiro-Wilk

O comando R para executar o teste Shapiro-Wilk é **shapiro.test**:

```
> shapiro.test(pluv)

  Shapiro-Wilk normality test

data:  pluv
W = 0.93775, p-value = 0.4694
```

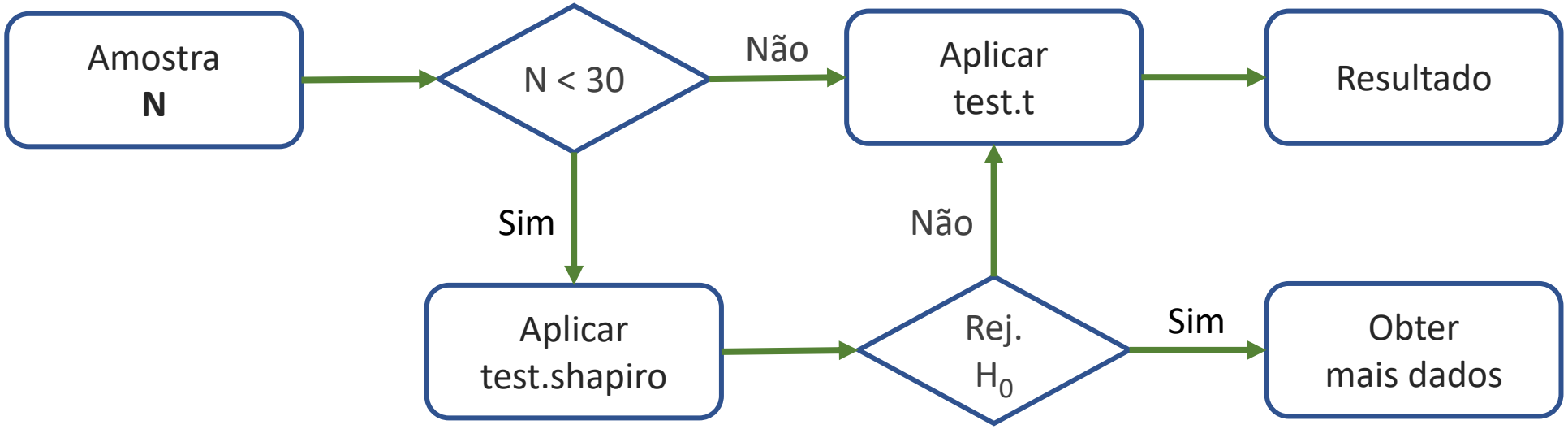
No teste Shapiro-Wilk a hipótese nula é: "a distribuição 'pluv' é normal" e a alternativa é "a distribuição 'pluv' não é normal".

O p-value 46.9% indica que não há evidência para rejeitar a hipótese nula, concluindo-se, portanto, que "pluv" provém de uma distribuição normal.

Desta forma os resultados obtidos anteriormente são válidos.

Caso o p-value do teste Shapiro-Wilk apresentasse valores baixos, (por exemplo, < 1%) seria necessário obter mais dados de "pluv" para chegar a uma amostra com pelo menos 30 valores.

Teste de Shapiro-Wilk



ANOVA - Análise de Variância

A análise de variância (ANOVA - analysis of variance) é um **teste de hipóteses aplicado para comparar médias de mais de duas amostras**. A técnica foi desenvolvida inicialmente para fins agrícolas.

Há diversos tipos de ANOVA. A seguir será aplicado a ANOVA de um fator de efeitos fixos (one-way fixed effects analisys of variance).

Pressupostos a serem observados:

1. Amostras devem ser aleatórias e mutuamente independentes;
2. Populações normalmente distribuída (aplicar o teste de Shapiro-Wilk para $N < 30$);
3. Populações apresentam variância homogênea (aplicar o teste de Bartlett).

ANOVA - Análise de Variância

Três áreas agrícolas foram submetidas a três tratamentos diferentes com adubos identificados por A, B e C. Foram recolhidas 5 colheitas de cada área, obtendo-se as seguintes produtividades por hectare para o mesmo produto:

```
A <- c(14,13,20,15,13)
B <- c(13,14,13,18,15)
C <- c(19,16,17,20,19)
```

Média dos três grupos:

```
> mean(A)
[1] 15
> mean(B)
[1] 14.6
> mean(C)
[1] 18.2
```

A diferença entre as médias amostrais é estatisticamente relevante? ou (em linguagem matemática): Os três grupos têm a mesma média de população, ou pelo menos um grupo tem média de população diferente dos outros?

ANOVA - Análise de Variância

1. **Amostras devem ser aleatórias e mutuamente independentes.**

Deve-se confirmar que não há interferência de um grupo no outro. Por exemplo: se as áreas fossem contíguas e a fertilização ocorresse por polinização, poderia haver mistura genética entre os grupos, invalidando a análise.

O pressuposto de independência não pode ser verificado por testes estatísticos preliminares.

No exemplo, assume-se que não há problemas de contaminação de uma área para outra e, portanto há independência.

ANOVA - Análise de Variância

2. **Populações normalmente distribuída (aplicar o teste de Shapiro-Wilk para N < 30)**

```
> shapiro.test(A)
  Shapiro-Wilk normality test
data:  A
W = 0.77559, p-value = 0.0505

> shapiro.test(B)
  Shapiro-Wilk normality test
data:  B
W = 0.84215, p-value = 0.171

> shapiro.test(C)
  Shapiro-Wilk normality test
data:  C
W = 0.91367, p-value = 0.4899
```

O grupo A (p-value 5.05%) está no borderline da normalidade, mas ainda assim pode ser considerado como seguindo uma distribuição normal.

ANOVA - Análise de Variância

3. Populações têm mesma variância (aplicar o teste de Bartlett).

No teste de Bartlett aceitar a hipótese H_0 indica que os grupos têm variância homogênea. Aceitar H_1 indica que não há variância homogênea (pelo menos um grupo não mantêm variância homogênea em relação aos outros).

```
> bartlett.test(list(A,B,C))

  Bartlett test of homogeneity of variances

data:  list(A, B, C)
Bartlett's K-squared = 1.2051, df = 2, p-value = 0.5474
```

O p-value 54.7% indica não rejeição de H_0, validando o pressuposto da homogeneidade da variância.

Validados os três pressupostos, pode-se preparar e aplicar, a seguir, o **teste ANOVA de um fator**.

ANOVA - Análise de Variância

Formatação dos dados

Para aplicar o teste ANOVA no R é preciso agregar os dados de produtividade das três áreas, A, B e C, em um único vetor.

```
> prod <- c(A,B,C)
> prod
 [1] 14 13 20 15 13 13 14 13 18 15 19 16 17 20 19
```

A seguir, cria-se o vetor com os elementos que correspondem aos grupos (áreas agrícolas).

```
> grupos <- c(rep("A",5),rep("B",5),rep("C",5))
> grupos
 [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C"
```

O próximo passo será juntar os dois vetores em um único data frame.

ANOVA - Análise de Variância

Formatação dos dados (continuação)

O data frame apresentado a seguir junta os dois vetores: **prod** e **grupos**.

```
> df <- data.frame(prod,grupos)
> df
   prod grupos
1    14      A
2    13      A
3    20      A
4    15      A
5    13      A
6    13      B
7    14      B
8    13      B
9    18      B
10   15      B
11   19      C
12   16      C
13   17      C
14   20      C
15   19      C
```

Esta visualização é importante para verificar se os dados foram inseridos corretamente.

ANOVA - Análise de Variância

Aplicando o teste ANOVA

O comando a seguir executa o ANOVA sobre um modelo linear (**lm**) em que cada valor de produtividade (**prod**) está associado ao respectivo grupo (**grupos**).

```
> anova(lm(prod ~ grupos))
Analysis of Variance Table

Response: prod
          Df Sum Sq Mean Sq F value  Pr(>F)
grupos     2 38.933 19.4667  3.7677 0.05372 .
Residuals 12 62.000  5.1667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

O p-value do ANOVA é o valor da coluna Pr(>F) = 0.05372, aproximadamente 5.3%.
É um valor no limite da rejeição de H_0, isto é, existe a suspeita de que pelo menos um tratamento tenha gerado diferença de produtividade.
Qual ou quais grupos são responsáveis pelas diferenças?
Um teste de comparação múltipla **post-hoc** tentará identificar o(s) grupo(s) responsáveis pela "quase" rejeição de H_0 no ANOVA.

Tukey HSD - Teste de comparações múltiplas

Aplicando o teste Tukey HSD

O R oferece recursos para realizar a maioria dos testes de comparações múltiplas. Um dos mais utilizados é o Tukey HSD (Honest Significant Difference).

O TukeyHSD atua sobre objetos da classe aov (Analysis Of Variance), portanto é preciso usá-la no comando a seguir:

```
> TukeyHSD(aov(lm(prod ~ grupos)))
 Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm(prod ~ grupos))

$grupos
    diff        lwr      upr     p adj
B-A -0.4 -4.2352956 3.435296 0.9583671
C-A  3.2 -0.6352956 7.035296 0.1068512
C-B  3.6 -0.2352956 7.435296 0.0665354
```

O p-value na comparação entre os grupos B-A (95.8%) indica clara não rejeição e H_0.
O p-value na comparação entre os grupos C-A (10.7%) não é suficientemente forte para rejeição de H_0.
O p-value na comparação entre os grupos C-B (6.7%) está no limite para não rejeição de H_0.
Se houve um tratamento que impactou na produtividade este foi aplicado no grupo C. O p-value próximo do limite de rejeição indica que é recomendável recolher mais amostras para testes.

Contato:

Prof. MSc. Marcos Alexandruk

E-mail: alexandruk@uninove.br