

10 de maio de 2021

1 Associação entre variáveis

1.1 Coeficiente de Correlação Linear de Pearson

O coeficiente de correlação Linear de Pearson mensura a relação linear entre duas **variáveis quantitativas** X e Y (ρ_{xy}). Não implica, necessariamente, em causalidade. A relação de causa e efeito entre as variáveis é determinada com estudo detalhados.

O coeficiente ρ_{xy} mede se existe uma relação entre as variações de aumento ou diminuição entre duas variáveis X e Y . Por exemplo, existe alguma relação entre massa e altura? Isto é, pessoas mais altas tendem a ter maior massa? Essa medida amplamente utilizada nas pesquisas Filho e Júnior (2009).

Para o cálculo, considere $X = \{x_1, x_2, x_3, \dots, x_n\}$ e $Y = \{y_1, y_2, y_3, \dots, y_n\}$ duas variáveis que podem ser organizadas em pares $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. A medida de correlação é calculada por:

$$\rho_{xy} = \frac{COV(x, y)}{\sigma_x \times \sigma_y} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} \quad (1.1)$$

Considere $X = \{x_1, x_2, x_3, \dots, x_n\}$ e $Y = \{y_1, y_2, y_3, \dots, y_n\}$ duas variáveis que podem ser organizadas em pares $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Seja μ_x e μ_y a média aritmética de X e Y , respectivamente. Seja σ_x^2 e σ_y^2 a variância de X e Y , respectivamente. Mostre que:

$$\rho_{xy} = \frac{\sum_{i=1}^n \left[\left(\frac{(x_i - \mu_x)}{\sqrt{\sigma_x^2}} \right) \left(\frac{(y_i - \mu_y)}{\sqrt{\sigma_y^2}} \right) \right]}{n} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} \quad (1.2)$$

O coeficiente de assume valores entre -1 e 1, podendo ser classificado nesse intervalo de acordo com direção (positiva ou negativa) e intensidade (muito forte, forte, moderada, fraca e muito fraca). É uma medida sensível a valores extremos (*outliers*) e paramétrica que exige que os dados tenham comportamento de normalidade (Distribuição Normal). Falaremos disso nas próximas aulas quando falaremos de distribuições de probabilidade.

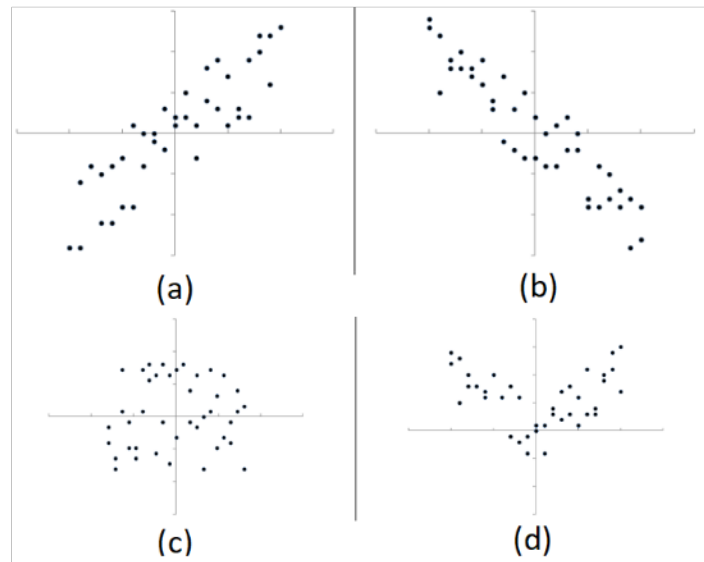


Figura 1 – Tipos de relação Correlação de Pearson

Intervalos	Classificação
1.0 a 0.8	Muito forte
0.8 a 0.6	Forte
0.6 a 0.4	Moderada
0.4 a 0.2	fraca
0.2 a 0.0	Muito fraca

Tabela 1 – Intervalos de ρ_{xy} e classificação

Existem variações da equação 1.2, considerando uma amostra. Também veremos isso em aulas futuras, quando falaremos de estimação.

Nesse momento estamos focados em realizar o cálculo de ρ_{xy} e compreender como interpretar a direção e intensidade. Para isso, vamos utilizar as imagens 1 com exemplos de gráficos de dispersão (bivariado), apresentadas no livro Morettin e Bussab (2017), que nos ajudam a compreender o direção do coeficiente de correlação. Cada ponto indica um par $\{(x_i, y_i)\}$.

A imagem (a) é um exemplo de relação positiva entre duas variáveis, isto é, quando há um aumento de uma variável também se observa o aumento da outra. Nesse caso o valor de ρ_{xy} será positivo. Por outro lado, a imagem (b) mostra uma relação negativa entre variáveis, isto é, quando há um aumento de uma variável se observa uma diminuição da outra variável. Nesse caso o valor de ρ_{xy} será negativo.

A imagem (c) mostra um padrão de dados sem relação evidente. Nesse caso o valor de ρ_{xy} será próximo de 0. Já a imagem (d) sugere que o padrão da relação é quadrático, o que vai além das capacidades da medida ρ_{xy} que mensura apenas a relação linear (reta) entre duas variáveis.

A intensidade de ρ_{xy} é medida de acordo com a proximidade com os extremos +1 e -1, isto é, quanto mais próximo, mais intenso a relação linear entre as variáveis. Valores próximos de zero indicam relação fraca entre as variáveis, parecida com a imagem (c) da figura 1

PA sistólica (mmHg)	PA diastólica(em mmHg)	Peso (kg)	Altura (cm)
126	76	71,7	176,7
95	52	49,3	166,8
121,5	68,5	66,1	170
122,5	72	67,4	160,5
132,6	90	80	178
107,4	64,4	84,2	169,2
112,5	67,5	48,6	166
104	63,5	55	177
113	69	55,4	165,3
112,5	72,5	57,6	184,6
112	63	58,9	182
103	69,5	67,2	167,1
108	61,5	70,3	182,6
116	77	72	170
106,5	60	102,4	182,5
101	60	46	167,8
109	63,5	49,1	158
119,5	80,5	50,2	164,6
135	86,5	52,3	164,3

Figura 2 – Dados coletados na Pesquisa Nacional de Saúde.

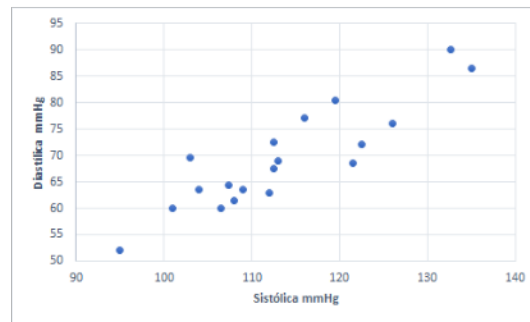


Figura 3 – Dispersão entre pressões arteriais.

1.1.0.1 Exemplo na medidas antropométricos

Para compreender e exemplificar esse cálculo, considere a Pesquisa Nacional de Saúde (PNS), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Entre outras informações coletadas, essa pesquisa investiga, entre outras, as características físicas, massa e altura, e relacionadas à saúde do brasileiro, pressão arterial sistólica e diastólica.

Para exemplificar o procedimento de cálculo, a tabela 2 a seguir apresenta os dados de alguns brasileiros para características como massa, altura, pressão sistólica e diastólica coletados pela PNS.

A figura 3 apresenta o gráfico de dispersão entre as medidas de pressão arterial sistólica e diastólica. Baseado nos dados, percebe-se que uma relação positiva entre essas variáveis. Isto é, a medida que aumenta-se o valor da pressão arterial Sistólica, observa-se maiores mensurações para pressão arterial Diastólica.

A figura 4 apresenta o gráfico de dispersão entre a massa e a altura. Baseado nos dados, percebe-se que se parece uma relação positiva entre essas variáveis, no entanto, não é tão evidente quando o outro par de variáveis.

A tabela 5 apresenta os somatórios apresentados na equação 1.2 para o cálculo da correlação linear de Pearson.

A primeira linha indica o número de observações (n), que é igual para todas as variáveis. A segunda linha indica o somatório de cada uma das variáveis ($\sum x$). A terceira linha indica o somatório do quadrado de cada uma das variáveis ($\sum x^2$), isto é, somando cada um dos valores apresentados na tabela 2 elevados ao quadrado. Esses dois

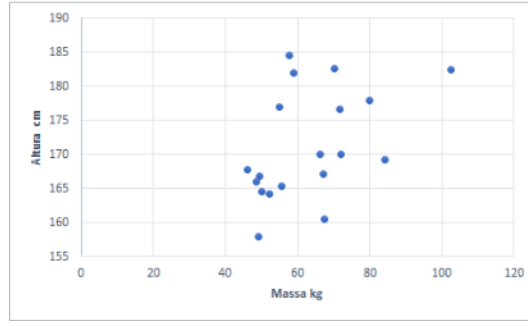


Figura 4 – Dispersão entre massa e altura.

Estatística	PA sistólica (mmHg)	PA diastólica(em mmHg)	Peso (kg)	Altura (cm)
n	19	19	19	19
Σx	2157,0	1316,9	1203,7	3253,0
Σx^2	246887,0	92914,6	80125,9	558108,2
Σxy	151117,6		207073,3	

Figura 5 – Somatórios para cálculo da correlação.

somatórios também são utilizados para o cálculo da média e da variância, apresentados anteriormente.

A última medida necessária para o cálculo do coeficiente de correlação de pearson, é a multiplicação de $(x_i \times y_i)$ de cada par (x_i, y_i) . Como estamos investigando a correlação entre a par de pressões e a massa e altura, apresentada na tabela 2, a quarta medida (Σxy) indica o produto entre os pares. Por exemplo, para o primeiro par, o valor de 151117,6 é resultado da soma:

$$151117,6 = (126 \times 76) + (95 \times 52) + (121,5 \times 68,5) + \dots (135 \times 86,5)$$

Dessa forma, para calcular a correlação entre o gasto com energia elétrica e o número de funcionários, precisamos saber que $n = 19$, $\sum_{i=1}^n x_i = 2157,0$, $\sum_{i=1}^n x_i^2 = 246887,0$, $\sum_{i=1}^n y_i = 1316,9$, $\sum_{i=1}^n y_i^2 = 92914,6$, $\sum_{i=1}^n x_i y_i = 151117,6$, então:

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{151117,6}{19} - \frac{(2157,0)}{19} \frac{(1316,9)}{19} \right)}{\sqrt{\frac{246887,0}{19} - \left(\frac{2157,0}{19} \right)^2} \sqrt{\frac{92914,6}{19} - \left(\frac{1316,9}{19} \right)^2}} = +0,889 \quad (1.3)$$

Dessa forma, a correlação entre pressão diastólica e sistólica é positiva e muito forte. Isto é, baseado nos dados, a medida em que aumentamos observamos um valor elevado de pressão sistólica também observamos aumento da pressão diastólica.

De maneira análoga, para calcular a correlação entre massa e altura, precisamos saber que $n = 19$, $\sum_{i=1}^n x_i = 1203,7$, $\sum_{i=1}^n x_i^2 = 80125,9$, $\sum_{i=1}^n y_i = 3253,0$, $\sum_{i=1}^n y_i^2 = 558108,2$, $\sum_{i=1}^n x_i y_i = 207073,3$, então:

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{207073,3}{19} - \frac{(1203,7)}{19} \frac{(3253,0)}{19} \right)}{\sqrt{\frac{80125,9}{19} - \left(\frac{1203,7}{19} \right)^2} \sqrt{\frac{558108,2}{19} - \left(\frac{3253,0}{19} \right)^2}} = +0,466 \quad (1.4)$$

Dessa forma, a correlação entre massa e altura é positiva e moderada. Isto é, baseado nos dados, a medida em que aumentamos a altura, também observamos um aumento da massa. No entanto, esse padrão de concordância não é tão forte e evidente quanto a relação linear das pressões.

Esses resultados são baseados nos resultados apresentados nesse estudo para fins didáticos. Para compreender melhor as relações e característica desse tema é necessário aprofundar o estudo, assim como em Peixoto, Benício e Jardim (2006). Além disso, é importante destacar que não é uma relação de causalidade.

1.1.0.2 Exemplo na Engenharia

Para compreendermos isso, considere os dados da pesquisa de Yeh (2007) para elaboração de concreto. Essa base de dados investiga de que maneira se pode obter o melhor resistência à compressão de um concreto? Para isso são investigados os seguintes parâmetros em uma pesquisa com coleta de dados (informações) baseado em um experimento com parâmetros controlados. Os pesquisadores controlam os seguintes parâmetros: cimento, escória, água, areia grossa, areia fina e tempo de cura.

Para exemplificar o procedimento de cálculo, a tabela 15 a seguir apresenta as 25 primeiras observações da base de resistência a compressão da UCI. Clique aqui e leia sobre elaboração de concreto.

A figura 16 apresenta o gráfico de dispersão entre a resistência do concreto e a quantidade de água. Baseado nos dados, percebe-se que não há uma relação clara entre a resistência do concreto e a água.

A figura 17 apresenta o gráfico de dispersão entre a resistência do concreto e a quantidade de escória. Baseado nos dados, percebe-se que uma relação negativa entre a resistência do concreto e escória. Isto é, a medida que aumenta a quantidade de escória, observa-se menor resistência do concreto.

A figura 13 apresenta o gráfico de dispersão entre a resistência do concreto e a quantidade de cimento. Baseado nos dados, percebe-se que uma relação positiva entre a resistência do concreto e cimento. Isto é, a medida que aumenta a quantidade de cimento, observa-se maior resistência do concreto.

A tabela 14 apresenta os somatórios apresentados na equação 1.2 para o cálculo da correlação linear de pearson.

A primeira linha indica o número de observações (n), que é igual para todas as variáveis. A segunda linha indica o somatório de cada uma das variáveis ($\sum x$). A terceira linha indica o somatório do quadrado de cada uma das variáveis ($\sum x^2$), isto é, somando cada um dos valores apresentados na tabela 15 elevados ao quadrado. Esses dois somatórios também são utilizados para o cálculo da média e da variância, apresentados anteriormente.

A última medida necessária para o cálculo do coeficiente de correlação de pearson, é a multiplicação de $(x_i \times y_i)$ de cada par (x_i, y_i) . Como estamos investigando a correlação de cada variável apresentada na tabela 15 e a variável resistência do concreto, a quarta medida ($\sum xy$) indica o produto entre cada variável e a variável resistência do concreto. Por exemplo, para a segunda coluna (“cement (kg in m^3 mixture)”), o valor de 369.905,6 é resultado da soma:

$$369.905,6 = (79,99 \times 540) + (61,89 \times 540) + (40,27 \times 332,5) + \dots (52,52 \times 380)$$

Concrete compressive strength(MPa)	Cement (kg in a m³ mixture)	Blast Furnace Slag(kg in a m³ mixture)	Water (kg in a m³ mixture)	Coarse Aggregate(kg in a m³ mixture)	Fine Aggregate (kg in a m³ mixture)	Age (day)
79,99	540,0	0,0	162,0	1040,0	676,0	28
61,89	540,0	0,0	162,0	1055,0	676,0	28
40,27	332,5	142,5	228,0	932,0	594,0	270
41,05	332,5	142,5	228,0	932,0	594,0	365
44,30	198,6	132,4	192,0	978,4	825,5	360
47,03	266,0	114,0	228,0	932,0	670,0	90
43,70	380,0	95,0	228,0	932,0	594,0	365
36,45	380,0	95,0	228,0	932,0	594,0	28
45,85	266,0	114,0	228,0	932,0	670,0	28
39,29	475,0	0,0	228,0	932,0	594,0	28
38,07	198,6	132,4	192,0	978,4	825,5	90
28,02	198,6	132,4	192,0	978,4	825,5	28
43,01	427,5	47,5	228,0	932,0	594,0	270
42,33	190,0	190,0	228,0	932,0	670,0	90
47,81	304,0	76,0	228,0	932,0	670,0	28
52,91	380,0	0,0	228,0	932,0	670,0	90
39,36	139,6	209,4	192,0	1047,0	806,9	90
56,14	342,0	38,0	228,0	932,0	670,0	365
40,56	380,0	95,0	228,0	932,0	594,0	90
42,62	475,0	0,0	228,0	932,0	594,0	180
41,84	427,5	47,5	228,0	932,0	594,0	180
28,24	139,6	209,4	192,0	1047,0	806,9	28
8,06	139,6	209,4	192,0	1047,0	806,9	3
44,21	139,6	209,4	192,0	1047,0	806,9	180
52,52	380,0	0,0	228,0	932,0	670,0	365

Figura 6 – Dados das 25 primeiras observações do experimento de concreto

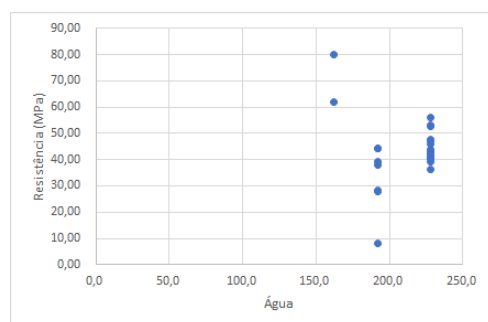


Figura 7 – Dispersão entre Resistência e água.

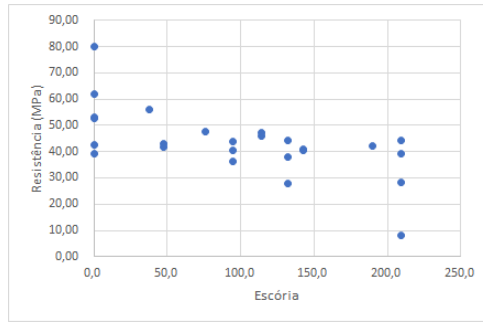


Figura 8 – Dispersão entre Resistência e escória.

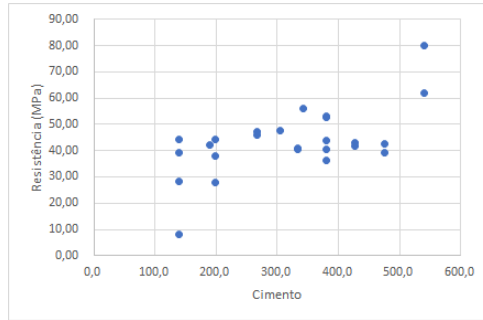


Figura 9 – Dispersão entre Resistência e cimento.

	Concrete compressive strength(MPa)	Cement (kg in a m³ mixture)	Blast Furnace Slag(kg in a m³ mixture)	Water (kg in a m³ mixture)	Coarse Aggregate(kg in a m³ mixture)	Fine Aggregate (kg in a m³ mixture)	Age (day)
<u>n</u>	25	25	25	25	25	25	25
<u>Σx</u>	1.085,5	7.972,2	2.431,8	5.316,0	24.130,2	17.092,1	3.667,0
<u>Σx²</u>	51.042,9	2.926.345,5	369.494,7	1.142.280,0	23.349.244,7	11.880.477,2	960.381,0
<u>Σxy</u>	51.042,9	369.905,6	91.200,3	229.844,6	1.046.858,3	733.694,4	165.925,9

Figura 10 – Somatórios para cálculo da correlação.

Dessa forma, para calcular a correlação entre a quantidade de cimento e resistência do concreto, fazemos:

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right) \left(\frac{\sum_{i=1}^n y_i}{n} \right) \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{369905,6}{25} - \frac{(1085,5)}{25} \frac{(7972,2)}{25} \right)}{\sqrt{\frac{51042,9}{25} - \left(\frac{1085,5}{25} \right)^2} \sqrt{\frac{2926345,5}{25} - \left(\frac{7972,2}{25} \right)^2}} = +0,61$$

Dessa forma, a correlação entre o cimento e resistência do concreto é positiva e forte. Isto é, baseado nos dados, a medida em que aumentamos a quantidade de cimento, também observamos um aumento da resistência do concreto.

De maneira análoga, temos que a correlação entre escória e resistência do cimento é:

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right) \left(\frac{\sum_{i=1}^n y_i}{n} \right) \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{91200,3}{25} - \frac{(1085,5)}{25} \frac{(2431,8)}{25} \right)}{\sqrt{\frac{51042,9}{25} - \left(\frac{1085,5}{25} \right)^2} \sqrt{\frac{369494,7}{25} - \left(\frac{2431,8}{25} \right)^2}} = -0,63$$

Dessa forma, a correlação entre a escória e resistência do concreto é negativa e forte. Isto é, baseado nos dados, a medida em que aumentamos a quantidade de escória, também observamos uma diminuição da resistência do

Mês	Energia Elétrica(R\$ mil)	Funcionários	Produção (R\$ mil)
1	94	250	506
2	89	249	437
3	130	253	500
4	130	258	512
5	153	258	562
6	158	264	600
7	169	267	607
8	182	318	645
9	132	320	596
10	168	320	603
11	182	322	649
12	149	314	554

Figura 11 – Dados de despesas indiretas com energia da empresa Alpha

concreto.

Esses resultados são baseados em uma pequena quantidade de observações desse estudo para fins didáticos. Para compreender melhor as relações e característica é necessário aprofundar o estudo Velten et al. (2006).

1.1.1 Exemplo na Contabilidade

Conhecer os custos, no entanto, é fundamental para se ter a capacidade mínima e necessária para conduzir os negócios neste competitivo ambiente, conforme aponta o trabalho de Elias et al. (2009).

Neste campo, a Contabilidade tem um papel fundamental no gerenciamento dos custo diretos e indiretos e na contribuição para a tomada de decisão. Seja no desenvolvimento de métodos de custeio que facilitem o conhecimento e a administração dos custos de produção, ou no entendimento dos processos produtivos, bem como na adequação destes processos sempre com o objetivo de aperfeiçoar o processo de produção

Nos processos industriais, os custos claramente evidenciados (Custos Diretos) são efetivamente alocados aos produtos de maneira objetiva e direta.

O quadro começa a ficar complexo na medida em que existem custos de difícil mensuração (Custos Indiretos) e precisam ser efetivamente alocados aos produtos fabricados por meio de rateios.

Entre eles:

- Mão-de-obra indireta: supervisores, controle de qualidade, etc.
- Materiais indiretos: graxas e lubrificantes, lixas etc.
- Outros custos indiretos: depreciação, seguros, manutenção de equipamentos, etc.

Para compreendermos isso, considere os dados da pesquisa de Elias et al. (2009) que investiga critérios de apropriação dos custos indiretos são mais adequados para avaliar a realidade do custo final do produto. Para isso, são investigados, por exemplo qual a melhor variável entre número de funcionários e produção está relacionada aos custos associados ao custo com Energia elétrica.

Para exemplificar o procedimento de cálculo, a tabela 11 a seguir apresenta os dados de 12 meses dos gastos em milhares com energia elétrica, o número de funcionários e o valor de produção em milhares para a Empresa Alpha.

A figura 17 apresenta o gráfico de dispersão entre o valor gasto com energia elétrica e o número de funcionários. Baseado nos dados, percebe-se que uma relação positiva entre essas variáveis. Isto é, a medida que aumenta-se o número de funcionários, observa-se maior gasto com energia.

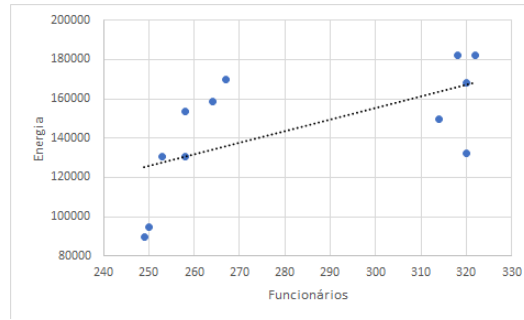


Figura 12 – Dispersão entre gasto com energia e número de funcionários.

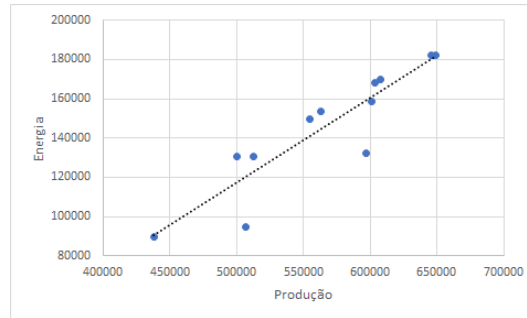


Figura 13 – Dispersão entre gasto com energia e produção produzido.

	Energia Elétrica(R\$ mil)	Funcionários	Produção (R\$ mil)
n	12	12	12
Σx	1.736	3.393	6.771
Σx^2	261.588	970.827	3.866.409
Σxv	261.588	497.666	999.416

Figura 14 – Somatórios para cálculo da correlação.

A figura 13 apresenta o gráfico de dispersão entre o valor gasto com energia elétrica e o valor produzido. Baseado nos dados, percebe-se que uma relação positiva entre essas variáveis. Isto é, a medida que aumenta-se a produção, observa-se maior gasto com energia.

A tabela 14 apresenta os somatórios apresentados na equação 1.2 para o cálculo da correlação linear de pearson.

A primeira linha indica o número de observações (n), que é igual para todas as variáveis. A segunda linha indica o somatório de cada uma das variáveis (Σx). A terceira linha indica o somatório do quadrado de cada uma das variáveis (Σx^2), isto é, somando cada um dos valores apresentados na tabela 11 elevados ao quadrado. Esses dois somatórios também são utilizados para o cálculo da média e da variância, apresentados anteriormente.

A última medida necessária para o cálculo do coeficiente de correlação de pearson, é a multiplicação de $(x_i \times y_i)$ de cada par (x_i, y_i) . Como estamos investigando a correlação de número de funcionários e produção, apresentada na tabela 15, com a variável gasto com energia elétrica, a quarta medida (Σxy) indica o produto entre cada variável e a variável gasto com energia elétrica. Por exemplo, para a segunda coluna (“Funcionários”), o valor de 497666 é resultado da soma:

$$497666 = (94 \times 250) + (89 \times 249) + (130 \times 253) + \dots (149 \times 314)$$

Dessa forma, para calcular a correlação entre o gasto com energia elétrica e o número de funcionários, precisamos saber que $n = 12$, $\sum_{i=1}^n x_i = 3393$, $\sum_{i=1}^n x_i^2 = 970827$, $\sum_{i=1}^n y_i = 1736$, $\sum_{i=1}^n y_i^2 = 261588$, $\sum_{i=1}^n x_i y_i = 497666$, então:

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right) \left(\frac{\sum_{i=1}^n y_i}{n} \right) \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{497666}{12} - \left(\frac{3393}{12} \right) \left(\frac{1736}{12} \right) \right)}{\sqrt{\frac{970827}{12} - \left(\frac{3393}{12} \right)^2} \sqrt{\frac{261588}{12} - \left(\frac{1736}{12} \right)^2}} = +0,623$$

Dessa forma, a correlação entre gasto com energia elétrica e número de funcionários é positiva e forte. Isto é, baseado nos dados, a medida em que aumentamos a quantidade de funcionários, também observamos um aumento do gasto com energia elétrica.

De maneira análoga, para calcular a correlação entre o gasto com energia elétrica e a produção, precisamos saber que $n = 12$, $\sum_{i=1}^n x_i = 6771$, $\sum_{i=1}^n x_i^2 = 3866409$, $\sum_{i=1}^n y_i = 1736$, $\sum_{i=1}^n y_i^2 = 261588$, $\sum_{i=1}^n x_i y_i = 999416$, então:

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right) \left(\frac{\sum_{i=1}^n y_i}{n} \right) \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{999416}{12} - \left(\frac{6771}{12} \right) \left(\frac{1736}{12} \right) \right)}{\sqrt{\frac{3866409}{12} - \left(\frac{6771}{12} \right)^2} \sqrt{\frac{261588}{12} - \left(\frac{1736}{12} \right)^2}} = +0,867$$

Dessa forma, a correlação entre gasto com energia elétrica e a produção é positiva e muito forte. Isto é, baseado nos dados, a medida em que aumentamos o a produção, também observamos um aumento do gasto com energia elétrica.

O custos indiretos com Energia para a empresa Alpha estão mais associados com a produção, correlação muito forte, do que com o número de funcionários, forte. Isto é, saber sobre a produção nos fornece mais informação sobre o custo com energia. Para além desse resultados, podemos ter evidências de que o funcionamento da empresa Alpha depende mais da produção e menos da mão-de-obra. Esse resultado faz sentido em empresas que funcionam com um nível maior de mecanização. Esse resultado não seria esperado, por exemplo, em empresas do setor de serviços que dependam mais dos funcionários.

Esses resultados são baseados nos resultados apresentados nesse estudo para fins didáticos. Para compreender melhor as relações e característica desse tema é necessário aprofundar o estudo. Além disso, é importante destacar que não é uma relação de causalidade.

1.1.2 Exemplo na Química

Para compreendermos isso, considere os dados da pesquisa de Panero et al. (2009) no estudo da composição química de quiabos. Essa base de dados investiga as característica dos quiabos nos estados do Rio Grande do Norte e Pernambuco. Para isso são investigados os seguintes parâmetros em uma pesquisa com coleta de dados (informações) baseado em um experimento amostral: Cu,Zn,Na,Fe,K,Ca,Mn,Mg,P04,SO4 e Cl. A tabela 15 a seguir apresenta os valores das observações e as medidas descritivas dos quiabos.

A figura 16 apresenta o gráfico de dispersão entre a CL e Na. Baseado nos dados, percebe-se que há uma relação clara entre a quantidade de CL e Na, isto é, quiabos que apresentam um desses elementos tendem a apresentar o outro.

A figura 17 apresenta o gráfico de dispersão entre a Mg e Na. Baseado nos dados, percebe-se que há uma relação clara entre a quantidade de Mg e Na, isto é, quiabos que apresentam elevados níveis de um desses elementos tendem a apresentar baixos níveis do outro.

A tabela 15 apresenta os somatórios apresentados na equação 1.2 para o cálculo da correlação linear de pearson.

Amostra	Na (X)	Cl (Y)	Mg (Z)	P04 (W)
XPE1	5,98	12,80	676,59	114,07
XPE2	5,86	12,60	677,36	115,17
XPE3	6,10	12,85	676,82	114,72
VSAPE1	1,50	12,04	935,33	92,46
VSAPE2	1,41	12,12	938,57	94,44
VSAPE3	1,58	12,09	937,73	93,56
CRN1	11,81	37,63	300,13	93,50
CRN2	12,23	37,75	302,17	94,55
CRN3	11,95	37,69	303,37	94,12
MRN1	9,87	34,41	506,76	85,49
MRN2	9,02	34,57	507,19	87,76
MRN3	9,36	34,48	506,57	87,12
ERN1	18,01	39,90	364,81	102,46
ERN2	17,82	40,12	366,22	100,83
ERN3	17,23	40,07	365,72	101,23
Σx	139,73	411,12	8.365,34	1.471,48
Σx^2	1751,07	13562,6	5454803	145711
Σxv	1751,07	4737,63	60932,97	13698,93

Figura 15 – Dados da composição química dos quiabos

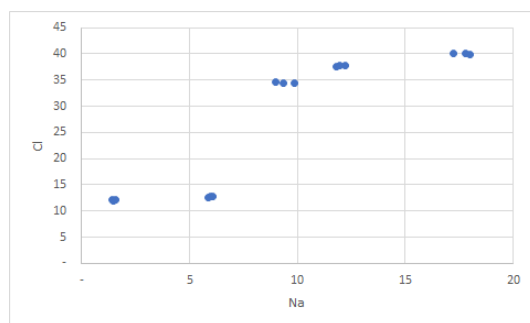


Figura 16 – Dispersão CL e Na.

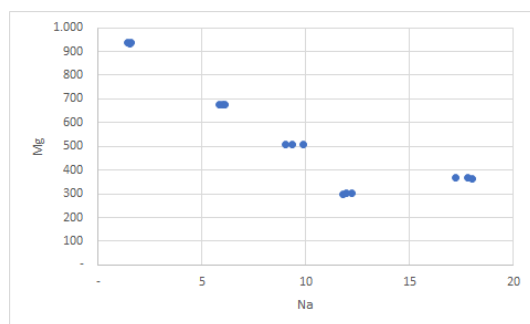


Figura 17 – Dispersão Mg e Na.

A primeira linha indica o somatório de cada uma das variáveis ($\sum x$). A segunda linha indica o somatório do quadrado de cada uma das variáveis ($\sum x^2$), isto é, somando cada um dos valores apresentados na tabela 15 elevados ao quadrado. Esses dois somatórios também são utilizados para o cálculo da média e da variância, apresentados anteriormente.

A última medida necessária para o cálculo do coeficiente de correlação de pearson, é a multiplicação de $(x_i \times y_i)$ de cada par (x_i, y_i) . Como estamos investigando a correlação de cada variável apresentada na tabela 15 e a variável Na do, a quarta medida ($\sum xy$) indica o produto entre cada variável e a variável Na. Por exemplo, para a segunda coluna ("Cl"), o valor de 4737,63 é resultado da soma:

$$4737,63 = (5,98 \times 12,8) + (5,86 \times 12,60) + \dots (17,23 \times 40,0)$$

Dessa forma, para calcular a correlação entre a quantidade de Na e Cl, fazemos:

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{4737,6}{16} - \frac{(139,7)}{16} \frac{(411)}{16} \right)}{\sqrt{\frac{1751}{16} - \left(\frac{139,7}{16} \right)^2} \sqrt{\frac{13562,6}{16} - \left(\frac{411,1}{16} \right)^2}} = +0,89 \quad (1.5)$$

Dessa forma, a correlação entre o Cl e Na é positiva e muito forte. Isto é, baseado nos dados, a medida em que observamos quiabos com elevada quantidade de Cl, também observamos níveis mais altos de Na.

De maneira análoga, temos que a correlação entre Mg e Na é:

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{60932,9}{16} - \frac{(139,7)}{16} \frac{(8365,3)}{16} \right)}{\sqrt{\frac{1751}{16} - \left(\frac{139,7}{16} \right)^2} \sqrt{\frac{5454803}{16} - \left(\frac{8365,3}{16} \right)^2}} = -0,90$$

Dessa forma, a correlação entre Mg e Na é negativa e muito forte. Isto é, baseado nos dados, a medida em que observamos quiabos com elevadas quantidades de Mg, observamos níveis menores de Na.

1.2 Qui-Quadrado

A medida de Qui-quadrado permite avaliar mensurar a associação entre variáveis X e Y do tipo qualitativa. Isto é, existem n_1 categorias para X e n_2 categorias para Y. A análise de associação é feita fazendo a contagem de frequência de ocorrência de cada um dos pares. Para ilustrar esse processo, considere a tabela 18 de frequência, também conhecida como tabela de contigência:

Nesse formato, as categorias de i de X estão relacionadas às posições das linhas e as j de Y às posições nas colunas da tabela, podendo ser representado de maneira geral por O_{ij} , i sendo a linha e j a coluna.

Para a combinação de categorias $i = 1$ e $j = 1$, temos a frequência de O_{11} , isto é, há ocorrência de O_{11} valores para as categorias $i = 1$ e $j = 1$, linha 1 e coluna 1. De maneira análoga, para a combinação de categorias $i = 3$ e $j = 2$, temos a frequência de O_{32} , isto é, há ocorrência de O_{32} valores para as categorias $i = 3$ e $j = 2$, linha 3 e coluna 2.

Ao final de cada coluna j , temos os totais de colunas ($O_{.j}$) somando os n_1 valores das linhas $O_{.j} = O_{1j} + O_{2j} + O_{3j} + \dots + O_{n_1j}$. Perceba que o valor de j é fixo, variando apenas o valor de i . Dessa forma, podemos representar tal soma por meio de somatórios. Veja como é mais fácil essa representação:

$$O_{.j} = O_{1j} + O_{2j} + O_{3j} + \dots + O_{n_1j} = \sum_{i=1}^{n_1} O_{ij} \quad (1.6)$$

Tabela de contingência		Categorias de Y				Total da linha
		j=1	j=2	j=3	j	j=n2
Categorias de X	i=1	O ₁₁	O ₁₂	O ₁₃	O _{1j}	O _{1.}
	i=2	O ₂₁	O ₂₂	O ₂₃	O _{2j}	O _{2.}
	i=3	O ₃₁	O ₃₂	O ₃₃	O _{3j}	O _{3.}
	i	O _{i1}	O _{i2}	O _{i3}	O _{ij}	O _{i.}
Total da coluna	i=n1	O _{.1}	O _{.2}	O _{.3}	O _{.j}	O _{..}

Figura 18 – Tabela de contingência

Ao final de cada linha i , temos os totais de linhas ($O_{i.}$) somando os n_2 valores das colunas $O_{i.} = O_{i1} + O_{i2} + O_{i3} + \dots + O_{in_2}$. Perceba que o valor de i é fixo, variando apenas o valor de j . Dessa forma, podemos representar tal soma por meio de somatórios. Veja como é mais fácil essa representação:

$$O_{i.} = O_{i1} + O_{i2} + O_{i3} + \dots + O_{in_2} = \sum_{j=1}^{n_2} O_{ij} \quad (1.7)$$

Por fim, temos o total geral ($O_{..}$), somando todos os valores de linhas e colunas.

$$O_{..} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} O_{ij} \quad (1.8)$$

A ideia básica da medida de Qui-Quadrado é comparar os valores observados (O_{ij}) com os valores caso não houvesse relação entre as variáveis, denominado valor esperado (e_{ij}). A determinação de e_{ij} é feita por meio do seguinte cálculo:

$$e_{ij} = \frac{O_{i.} \times O_{.j}}{O_{..}} \quad (1.9)$$

Para o cálculo da medida Qui-Quadrado, vamos comparar cada valor observado O_{ij} com cada valor esperado e_{ij} .

$$\chi^2 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \quad (1.10)$$

Para determinar se existe relação entre as variáveis, comparamos o valor χ^2 com uma distribuição de probabilidade Qui-Quadrado com $gl = (n_1 - 1) \times (n_2 - 1)$ graus de liberdade. Com isso, calculamos o **p-valor** de um teste de hipótese. Voltaremos a falar de χ^2 no segundo módulo da disciplina.

Outras medidas calculadas por meio do χ^2 é o coeficiente de contingência C .

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad (1.11)$$

Coeficiente T é no máximo 1 quando $n_1 = n_2$:

$$T = \sqrt{\frac{\chi^2/n}{(n_1 - 1) \times (n_2 - 1)}} \quad (1.12)$$

Para exemplificar o processo de cálculo, vamos considerada a tabela 19 apresenta os valores informados no trabalho de Carvalho et al. (2001) para estabelecer relação sexo e as categorias de Índice de Massa Corpórea (IMC).

	HOMEM	MULHER	Total
NORMAL	110	148	258
MAGRO	5	5	10
SOBREPESO	38	28	66
Total	153	181	334

Figura 19 – Tabela de contingência apresentada em Carvalho et al. (2001)

Para ficar claro a eventual relação entre sexo e as categorias de IMC, a tabela a seguir apresenta a mesma tabela de contingência utilizando os percentuais para cada linha.

	HOMEM	MULHER	Total
NORMAL	71,9%	81,8%	77,2%
MAGRO	3,3%	2,8%	3,0%
SOBREPESO	24,8%	15,5%	19,8%
Total	100,0%	100,0%	100,0%

Figura 20 – Tabela de percentual

Percebe-se que o percentual de homens em condição de normalidade é menor do que o de mulher. Para a condição de sobrepeso, temos um percentual mais elevado de homens. O gráfico 21 apresenta de forma mais direta as diferenças de percentuais observadas.

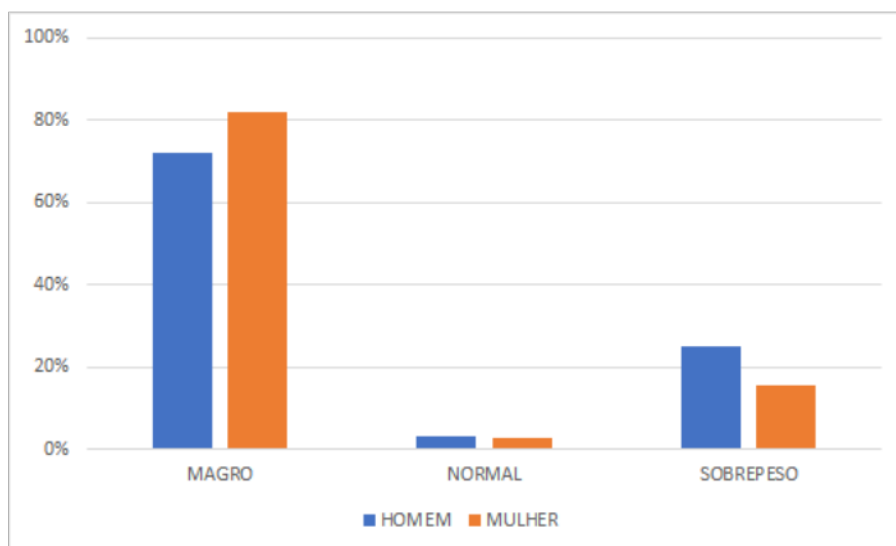


Figura 21 – Gráfico de colunas para dados de Carvalho et al. (2001)

Dessa forma, uma pergunta surge é: existe relação entre sexo e a categoria de IMC? A medida de Qui-Quadro nos ajuda a responder essa pergunta. Para isso, a primeira etapa é calcular os totais de linha e coluna. Esses totais já são apresentados na tabela 19.

A segunda etapa é calcular os valores esperados, utilizando a equação 1.9. O valor esperado para a primeira linha e primeira coluna e_{11} é:

$$e_{11} = \frac{O_{1.} \times O_{.1}}{O_{..}} = \frac{258 \times 153}{334} = 118,18 \quad (1.13)$$

Dessa forma, o valor de frequência esperado homens normais, caso não houvesse relação entre sexo e classe de IMC, seria de 118,18. No entanto, o observado foi de 110. Isto é, supondo não haver relação, o percentual de homens em normalidade deveria ser levemente maior.

O valor esperado para a terceira linha e segunda coluna e_{32} é:

$$e_{32} = \frac{O_{3.} \times O_{.2}}{O_{..}} = \frac{66 \times 181}{334} = 35,76 \quad (1.14)$$

Dessa forma, o valor esperado para mulheres em sobrepeso, caso não houvesse relação entre sexo e classe de IMC, seria de 35,76. No entanto, o observado foi de 28. Isto é, supondo não haver relação, o percentual de mulheres em sobrepeso deveria ser maior.

Fazendo para todas as combinações de linhas e colunas, temos os valores apresentados na tabela 22:

	HOMEM	MULHER	Total
NORMAL	118,19	139,81	258
MAGRO	4,58	5,42	10
SOBREPESO	30,23	35,77	66
Total	153	181	334

Figura 22 – Valores esperados para frequência.

A medida χ^2 é baseada nessas diferenças. Aplicando a fórmula, temos as expressões apresentadas na tabela 23:

	HOMEM	MULHER
NORMAL	$(110 - 118,2)^2/118,2$	$(148 - 139,8)^2/139,8$
MAGRO	$(5 - 4,6)^2/4,6$	$(5 - 5,4)^2/5,4$
SOBREPESO	$(38 - 30,2)^2/30,2$	$(28 - 35,8)^2/35,8$

Figura 23 – Valores para as diferenças quadráticas entre os valores esperados e os valores observados.

Com isso, resolvendo esses cálculos, temos os valores apresentados na tabela 24:

	HOMEM	MULHER
NORMAL	0,567	0,479
MAGRO	0,038	0,032
SOBREPESO	1,995	1,686

Figura 24 – Valores para as diferenças quadráticas entre os valores esperados e os valores observados.

A medida χ^2 é determinada pela soma dessas diferenças. Temos assim o valor de $\chi^2 = 4.79$, valor de contingência $C = 0,119$ e valor $T = 0,085$. Como valor de contingência está próximo de 0, temos indícios que não há relação entre essas variáveis. Como mencionado, para determinar se existe relação entre essas variáveis o valor $\chi^2 = 4.79$ é comparado em uma distribuição de probabilidade Qui-Quadrado, por meio do **teste do Qui-quadrado**, calculando o $p - \text{valor}$. Falaremos desse teste em aulas futuras, mas você já pode procurar mais informações para esse teste em Morettin e Bussab (2017). Para esses dados, temos o $p - \text{valor}$ de $p = 0.09$, apresentado em Carvalho et al. (2001), o que revela a inexistência de evidências de associação entre o estado nutricional (classe IMC) e o sexo dos adolescentes ($p > 0.05$).

Dessa forma, apesar da diferença percentual apresentada na figura 21, não conseguimos estabelecer relação entre sexo e estado nutricional segundo IMC, isto é, a situação nutricional dos adolescentes de cada sexo pode ser **considerada semelhante**.

1.3 Introdução à análise de variância

Também é possível mensurar a associação entre variáveis quantitativas e qualitativas. Para isso, comparamos utilizamos as medidas descritivas de média e variância.

Compreender o processo do cálculo e de interpretação da associação entre variáveis quantitativas e qualitativas te ajudará a compreender, em aulas futuras, o processo de análise de variância (ANOVA), que é amplamente utilizada nos estudos científicos. Um exemplo disso é o estudo de MACEDO et al., 2015

Para mensurar a relação entre uma variável quantitativa e qualitativa, faz-se MORETTIN; BUSSAB, 2017

$$\bar{\sigma}_{1:k}^2 = \frac{\sum_{i=1}^k n_i \times \sigma_i^2}{\sum_{i=1}^k n_i} \quad (1.15)$$

Sendo σ_i^2 a variância da variável quantitativa para a categoria i e n_i o número de observações para a categoria i . Dessa forma, $\bar{\sigma}_{1:k}^2$ representa uma variância média das categorias. Para realizar a medida associativa compara-se essa variância média com a variância caso os dados não fossem divididos por essas categorias.

A medida de associação é determinada por:

$$R^2 = 1 - \frac{\bar{\sigma}_{1:k}^2}{\sigma^2} \quad (1.16)$$

Sendo σ^2 a variância total dos dados, juntando as informações de todas as categorias. O valor de R^2 está entre 0 e 1, pois $\bar{\sigma}_{1:k}^2 < \sigma^2$.

1.3.1 Aplicação na Engenharia

Para compreender essa medida de correlação vamos considerar os dados do estudo de Jiao et al. (2018). Os dados também podem ser encontrados na plataforma mendeley. Uma das perguntas do estudo é: existe relação entre tempo de cura e a resistência do concreto? Para investigar isso, o trabalho de Jiao et al. (2018) compara a resistência do concreto com diferentes concretos com baixo tempo de cura (3 dias) e elevado tempo de cura (28 dias). Dessa forma, a variável tempo de cura pode ser considerada qualitativa, pois está categorizada em baixo (3D) e elevado (28D).

A tabela 25 apresenta as medidas descritivas de média, variância (VAR), desvio padrão (DESVIO), assimetria (AS3), Limite inferior do boxplot (LI) e superior do boxplot (LS) para os dados de 21 concretos elaborados com

Estatística	3D	28D	Todos
n	21	21	42
Σx	3.254,3	8.589,7	11.844,0
Σx^2	558.849,3	3.633.641,9	4.192.491,2
Média	155,0	409,0	282,0
Var	2.597,8	5.721,1	20.297,4
Desvio	51,0	75,6	142,5
Q1	112,6	379,9	140,8
Q2	141,4	425,8	258,8
Q3	172,7	471,8	426,6
AS3	0,042	0,002	0,174
LS	262,9	609,7	855,4
LI	22,4	242,1	-288,0

Figura 25 – Dados do experimento de tempo de cura de concreto.

3 dias de cura (3D) e os dados de 21 concretos elaborados com 28 dias de cura (28D). Além disso, essas mesmas medidas são apresentadas juntando todos os 42 dados.

Nesse dos dados de JIAO et al., 2018 há dois níveis da variável categórica : baixo e alto tempo de cura. A variável quantitativa é a resistência à compressão. Então, deseja-se saber se o tempo de cura muda o padrão da resistência à compressão?

$$\bar{\sigma}_{1:2}^2 = \frac{21 \times 2.597,8 + 21 \times 5.721,1}{21 + 21} = 4.159,5 \quad (1.17)$$

$$R^2 = 1 - \frac{4.159,5}{20.297,4} = 0.79507 \quad (1.18)$$

Dizemos que 79.507% da variação total da resistência à compressão é explicada pela variável “tempo de cura”. Isto é, saber o tempo de cura do concreto altera muito a resistência do concreto, pois as variâncias considerando o grupo de observações 3D (2597,8) e 28D (5721,1) separadamente **REDUZEM** drasticamente a variância total dos dados (20297,4). Por isso, esse procedimento será chamada **análise de variância**. Veremos com maiores detalhes a ANOVA em capítulos futuros.

Essa análise da variância pode ser feita visualmente utilizando o boxplot dos três conjuntos de dados, apresentado na figura 26. A variabilidade da resistência à compressão das observações agregadas é elevada, percebida pela distância entre os quartis (caixa). Isto é, caso não se saiba como o concreto foi elaborado (TODOS) há muita incerteza do nível de resistência do produto.

No entanto, quando controlamos a variável categórica tempo de cura baixa e alto, essa variabilidade é dramaticamente reduzida, percebida pelos boxplots em vermelho (28D) e (3D). Os concretos elaborados com 28 dias de tempo de cura possuem níveis elevados de resistência, enquanto os concretos elaborados com 3 dias de cura possuem níveis mais baixos de resistência. Dessa forma, dizemos que existe relação entre o tempo de cura e a resistência do concreto, isto é, reduzimos nossa incerteza (variabilidade) da resistência quando temos a informação de como o concreto foi elaborado.

1.3.2 Aplicação na Química

Para compreender essa medida de correlação vamos considerar os dados do estudo de Oliveira e Santos (2011). Uma das perguntas do estudo é: existe relação entre tipo de abelha (africanizada e nativa) e os parâmetros físico-químicos entre os méis produzidos? Para investigar isso, o trabalho de Oliveira e Santos (2011) investiga o padrão da Hidroximetilfurfural em abelhas do tipo nativa e africanizada.

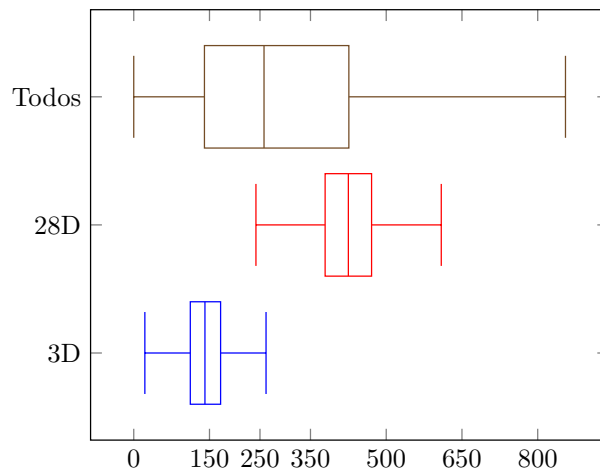


Figura 26 – Distribuição dos dados segundo tempo de cura.

Estatística	AFRICANA	NATIVA	TOTAL
n	8	8	16
Σx	15,04	37,21	52,25
Σx^2	39,05	177,37	216,42
Média	1,88	4,65	3,27
Var	1,35	0,54	2,86
Desvio	1,16	0,73	1,69
Q1	0,83	4,02	1,44
Q2	1,49	4,98	3,61
Q3	3,34	5,23	4,99
LI	0,00	2,21	0,00
LS	7,11	7,04	10,32

Figura 27 – Dados dos físico-químicas das abelhas.

A tabela 27 apresenta as medidas descritivas de média, variância (VAR), desvio padrão (DESVIO), Limite inferior do boxplot (LI) e superior do boxplot (LS) para os dados das abelhas. Além disso, essas mesmas medidas são apresentadas juntando todos os dados das 16 abelhas.

Nesse dos dados de Oliveira e Santos (2011) , há dois níveis da variável categórica : nativa e africanizada. A variável quantitativa é nível de sacarose. Então, deseja-se saber se existe relação entre o nível de sacarose e o tipo de abelha?

$$\bar{\sigma}_{1:2}^2 = \frac{8 \times 1,35 + 8 \times 0,54}{8 + 8} = 0,94 \quad (1.19)$$

$$R^2 = 1 - \frac{0,94}{2,86} = 0,6709 \quad (1.20)$$

Dizemos que 67.09% da variação total do nível de sacarose é explicada pela variável “tipo de abelha”. Isto é, saber o tipo de abelha altera muito o nível de sacarose do mel produzido, pois as variâncias considerando o grupo de observações nativa (0,54) e africanizada (1,35) separadamente **REDUZEM** drasticamente a variância total dos dados (2,86). Por isso, esse procedimento será chamada **análise de variância**. Veremos com maiores detalhes a ANOVA em capítulos futuros.

Essa análise da variância pode ser feita visualmente utilizando o boxplot dos três conjuntos de dados, apresentado na figura 28. A variabilidade do nível de sacarose das observações agregadas é elevada, percebida pela distância entre os quartis (caixa). Isto é, caso não se saiba o tipo de abelha (TODOS) há muita incerteza do nível de sacarose do produto.

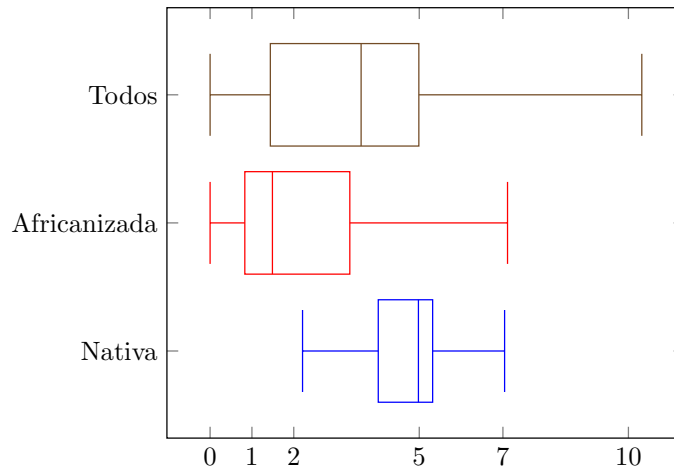


Figura 28 – Distribuição dos dados segundo tempo de cura.

No entanto, quando controlamos a variável categórica tipo de abelha, essa variabilidade é drasticamente reduzida, percebida pelos boxplots em vermelho (nativa) e azul (africanizada). As abelhas do tipo africanizada possuem níveis elevados de sacarose, enquanto as abelhas nativas possuem níveis mais baixos de sacarose. Dessa forma, dizemos que existe relação entre o tipo de abelha e a sacarose do mel produzido, isto é, reduzimos nossa incerteza (variabilidade) da sacarose quando temos a informação do tipo de abelha que produz o mel.

1.4 Correlação de Spearman

Uma medida não paramétrica de associação entre variáveis quantitativas pode ser feita por meio do coeficiente de correlação de Spearman. Essa medida de correlação é indicada quando os dados tem padrão assimétrico ou uma curtose muito diferente de 0.26. Veremos que esse é o padrão de comportamentos diferentes da distribuição Normal.

Esse coeficiente baseia-se nas diferença entre o ordenamento da variável X e Y (d_i). Por exemplo, suponha que tenhamos 4 pares: $(x_1 = 6, y_1 = 1)$, $(x_2 = 5, y_2 = 2)$, $(x_3 = 7, y_1 = 3)$ e $(x_4 = 8, y_4 = 0)$.

Inicialmente vamos ordenar os dados. O menor elemento entre os valores X é o $x_2 = 5$, então ele fica com o valor “1”. O segundo valor é o x_1 . De maneira análoga, o valor de y_4 está na posição 1 e y_1 na posição 2.

Com isso, o valor d_1 será a diferença entre a posição de x_1 e a posição de y_1 . Nesse caso, $d_1 = 2 - 2 = 0$. O mesmo é feito para os demais. A medida de $Spearman_{xy}$ é feita calculando:

$$Spearman_{xy} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{(n^3 - n)} \quad (1.21)$$

Sendo d_i a diferença entre a posição da observação i .

Referências

CARVALHO, C. et al. Consumo alimentar de adolescentes matriculados em um colégio particular de teresina, piauí, brasil. 2001.

- ELIAS, Z. dos S. et al. Rateio dos custos indiretos: aplicação da análise de correlação e de regressão. *Revista de contabilidade do mestrado em ciências contábeis da UERJ*, v. 14, n. 2, p. 50–66, 2009.
- FILHO, D. B. F.; JÚNIOR, J. A. d. S. Desvendando os mistérios do coeficiente de correlação de pearson (r). Universidade Federal de Pernambuco, 2009.
- JIAO, D. et al. Mixture design of concrete using simplex centroid design method. *Cement and Concrete Composites*, v. 89, p. 76 – 88, 2018. ISSN 0958-9465. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0958946517306467>.
- MACEDO, L. d. et al. Propriedades físicas de painéis aglomerados de madeira produzidos com adição de película de polipropileno biorientado. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 19, n. 7, p. 674–679, 2015.
- MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. São Paulo: Editora Saraiva, 2017.
- OLIVEIRA, E. N. A. d.; SANTOS, D. d. C. Análise físico-química de méis de abelhas africanizada e nativa. *Revista do Instituto Adolfo Lutz (Impresso)*, Instituto Adolfo Lutz, v. 70, n. 2, p. 132–138, 2011.
- PANERO, F. d. S. et al. Aplicação da análise exploratória de dados na discriminação geográfica do quiabo do rio grande do norte e pernambuco. *Eclética Química*, SciELO Brasil, v. 34, n. 3, p. 33–40, 2009.
- PEIXOTO, M. d. R. G.; BENÍCIO, M. H. D.; JARDIM, P. C. B. V. Validade do peso e da altura auto-referidos: o estudo de goiânia. *Revista de Saúde Pública*, SciELO Public Health, v. 40, p. 1065–1072, 2006.
- VELTEN, R. Z. et al. Caracterização mecânica de misturas solo-escória de alto-forno granulada moída para aplicações em estradas florestais. *Revista Árvore*, SciELO Brasil, v. 30, n. 2, p. 235–240, 2006.
- YEH, I.-C. *Concrete Compressive Strength Data Set*. 2007. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>.