

MAE116 – Noções de Estatística

Grupos B e D – II Semestre de 2020

Lista de exercícios 2 – Estatística Descritiva III – CASA (gabarito)

Exercício 1.

Por elasticidade de um queijo, entende-se a capacidade do queijo derretido esticar, sendo que, na avaliação desta capacidade envolvem-se ainda fatores como a resistência, ausência de rompimento do fio e a aderência (a tendência é exigir-se um queijo com boa elasticidade e que, ao se esticar, não arrebente e mantenha ainda certa aderência, por exemplo, à superfície de uma pizza).

Amostra	Temperatura (X)	Elasticidade (Y)
1	59	178
2	63	182
3	68	207
4	72	208
5	74	197
6	78	215
7	83	212

(a) Construa o diagrama de dispersão de Y em função de X.

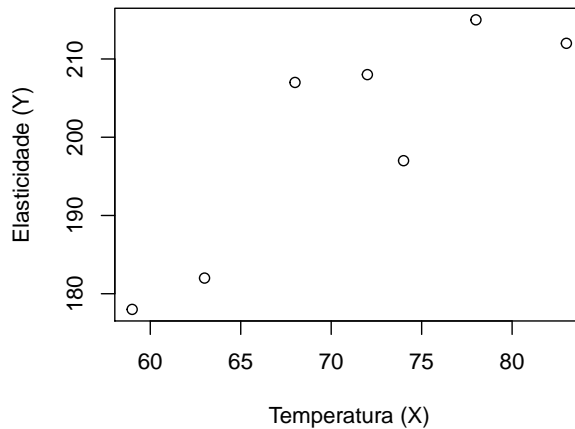


Figura 1: Diagrama de dispersão da Elasticidade (Y) em função da temperatura (X).

(b) Obtenha o coeficiente de correlação entre X e Y. Com base nesse valor, faça um comentário sobre a associação entre as variáveis.

O coeficiente de correlação entre as variáveis X e Y é dado por

$$r = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{(n-1) S_X S_Y},$$

em que

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_X^2 = \frac{1}{n-1} \sum_{i=1}^n X_i^2 - n \bar{X}^2 \quad \text{e} \quad S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n Y_i^2 - n \bar{Y}^2.$$

Calculando as quantidades necessárias, obtemos que

$$\bar{X} = \frac{59 + 63 + \dots + 78 + 83}{7} = 71 \quad \text{e} \quad \bar{Y} = \frac{178 + 182 + \dots + 215 + 212}{7} = 199,8571;$$

e que

$$\sum_{i=1}^7 X_i^2 = 59^2 + 63^2 + \dots + 78^2 + 83^2 = 35707, \quad \sum_{i=1}^7 Y_i^2 = 178^2 + 182^2 + \dots + 215^2 + 212^2 = 280899,$$

$$\sum_{i=1}^7 X_i Y_i = 59 \times 178 + 63 \times 182 + \dots + 78 \times 215 + 83 \times 212 = 99964.$$

Logo,

$$S_X^2 = \frac{\sum_{i=1}^7 X_i^2 - 7\bar{X}^2}{6} = \frac{35707 - 7.(71)^2}{6} = 70,$$
$$S_Y^2 = \frac{\sum_{i=1}^7 Y_i^2 - 7\bar{Y}^2}{6} = \frac{280899 - 7.(199,8571)^2}{6} = 216,4762,$$

e, portanto, o coeficiente de correlação r é dado por

$$r = \frac{\sum_{i=1}^7 X_i Y_i - 7\bar{X}\bar{Y}}{6S_X S_Y} = \frac{99964 - 7.(71).(199,8571)}{6.(\sqrt{70}).(\sqrt{216,4762})} = 0,8597$$

O coeficiente de correlação $r = 0,8597$ indica correlação positiva entre a temperatura e a elasticidade de ruptura do queijo. Este valor sugere forte relação linear crescente entre as variáveis, de forma que aumentando-se a temperatura, aumenta-se a elasticidade do queijo.

(c) Obtenha a reta de regressão.

Recorde que a reta de regressão ajustada é dada por

$$\hat{Y} = a + bX,$$

em que

$$a = \bar{Y} - b\bar{X} \quad \text{e} \quad b = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{(n-1)S_X^2}.$$

Utilizando os valores já calculados no item (b), obtemos que $a = 92,512$ e $b = 1,512$. Logo, a reta de regressão ajustada é dada por

$$\hat{Y} = 92,512 + 1,512 \times X$$

(d) Qual é o significado do coeficiente b obtido?

Para um aumento da temperatura em 1 °C, estima-se que a elasticidade do queijo aumenta, em média, 1,512 unidades.

(e) Com base na reta do item (c), obtenha uma estimativa da elasticidade média de um queijo sujeito a uma temperatura de 70 °C.

Com base no item (c), a reta de regressão ajustada para este conjunto de dados é dada por

$$\hat{Y} = 92,512 + 1,512 X,$$

em que X é a temperatura em °C.

Assim, para uma temperatura de 70°, estima-se que a elasticidade média do queijo seja de

$$\hat{Y} = 92,512 + 1,512 \times 70 = 198,352 \text{ unidades.}$$

Exercício 2.

O sanduíche Big Mac, presente em todos os países onde a rede McDonalds opera, foi utilizado como um índice econômico. Em 1986, a revista britânica The Economist criou o índice Big Mac, com o objetivo de comparar o valor do Big Mac em diversos países do mundo. Os dados do arquivo `bigmac.xls` representam as variáveis

Y - número médio de minutos de trabalho necessários para comprar um Big Mac

X - número médio de minutos de trabalho necessários para comprar um quilo de pão

Z - salário médio anual do professor do ensino fundamental em milhares de dólares, para uma amostra de 45 países, em um determinado instante de tempo.

(a) Construa os diagramas de dispersão de Y em função de X e Y em função de Z.

Para fazer o gráfico com o pacote `Rcmdr`, siga os seguintes passos:

(1) Ler o conjunto de dados: Clique na guia Dados \Rightarrow Importar arquivos de dados \Rightarrow do Excel. Em seguida defina um nome para o seu conjunto de dados (por exemplo, `bigmac`) e clique em OK. Navegue até o diretório onde o seu conjunto de dados está localizado e clique em Abrir.

(2) Diagrama de dispersão: Após a leitura do banco de dados, vá até a guia Gráficos \Rightarrow Diagrama de dispersão; escolha as variáveis de interesse e aperte em OK.

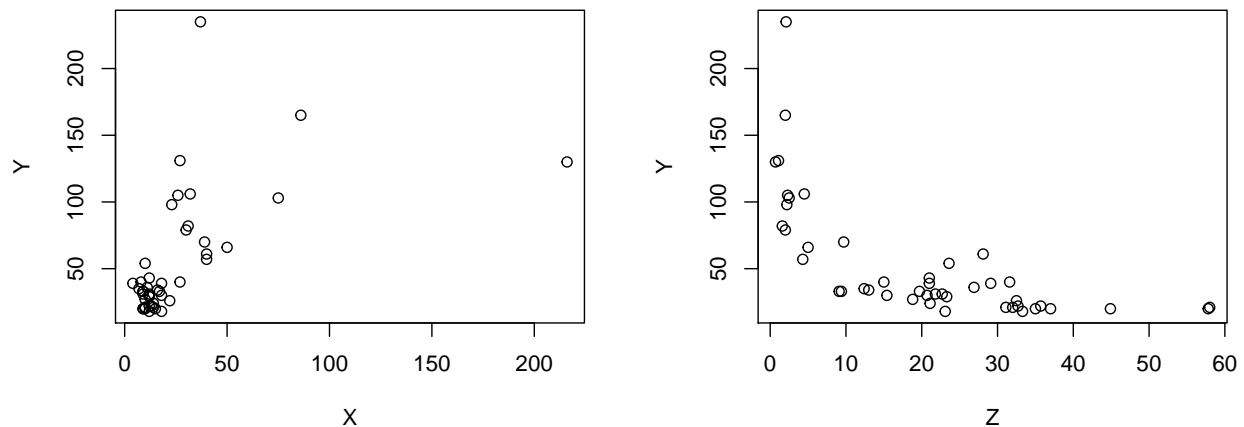


Figura 2: Diagrama de dispersão entre o número médio de minutos de trabalho necessários para comprar um *Big Mac* (Y) em função de (a) número médio de minutos de trabalho para comprar um quilo de pão e (b) salário médio anual do professor do ensino fundamental em milhares de dólares.

(b) Calcule o coeficiente de correlação linear de Pearson entre X e Y e entre Z e Y.

Para calcular a correlação linear entre as variáveis no Rcmdr, utilizamos a sequência de comandos:

Estatísticas \Rightarrow Resumos \Rightarrow Matriz de correlação. Selecione as variáveis X, Y e Z (mantendo a tecla `Ctrl` pressionada para selecionar mais de uma variável).

Para as variáveis X e Y temos que $r = 0,55$.

Para as variáveis Z e Y obtemos $r = -0,67$.

(c) Com base nos itens (a) e (b) comente sobre a associação entre as variáveis Y e X e entre as variáveis Y e Z.

No diagrama de dispersão da variável Y em função de X observa-se que com o aumento da variável X, tem-se um aumento na variável Y e a associação parece ser linear. O valor 0,55 para o coeficiente de correlação linear de Pearson indica uma relação positiva e moderada entre as duas variáveis.

Por outro lado, no diagrama de dispersão da variável Y em função de Z, observa-se que com o aumento da variável Z, tem-se um decréscimo na variável Y, mas a relação não parece ser linear.

Embora o valor $-0,67$ para o coeficiente de correlação linear de Pearson indique uma relação negativa e moderada entre as duas variáveis, este valor deve ser interpretado com cautela em vista do aspecto do gráfico de dispersão.

(d) Obtenha a reta de regressão de Y em função de X. Qual é a estimativa do acréscimo médio no número de minutos de trabalho necessários para comprar um Big Mac com o aumento de uma unidade do número de minutos de trabalho necessários para comprar um quilo de pão?

A reta de regressão pode ser obtida utilizando o Rcmdr seguindo os seguintes comandos:

Estatísticas \Rightarrow Ajuste de Modelos \Rightarrow Regressão linear. Selecione as variáveis informando qual a variável resposta e a variável explicativa.

A reta de regressão ajustada é dada por

$$\hat{Y} = 34,48 + 0,74 X.$$

Dessa forma, a estimativa do acréscimo médio em Y com o aumento de uma unidade em X é $0,74$.

(e) Obtenha uma previsão do valor de Y para um país não observado na amostra para o qual X é igual a 50.

Para prever o número de minutos de trabalho necessários para comprar um Big Mac (Y) em função do número de minutos de trabalho necessários para comprar um quilo de pão (X) podemos utilizar a reta de regressão de Y em função de X.

Utilizando a reta de regressão obtida temos que para $X = 50$

$$\hat{Y} = 34,48 + 0,74 \times 50 = 71,48$$

Exercício 3.

Uma amostra aleatória de 352 pessoas foi entrevistada, sendo anotados de cada participante o gênero (Masculino; Feminino), a faixa etária (Menos de 40 anos; 40 anos ou mais) e o resultado da preferência por um de dois adoçantes (A; B). Os resultados estão apresentados no quadro a seguir.

Gênero e Faixa etária	Preferência	
	Adoçante A	Adoçante B
Feminino e Menos de 40 anos	20	40
Masculino e Menos de 40 anos	30	70
Feminino e 40 anos ou mais	50	16
Masculino e 40 anos ou mais	100	26

(a) Verifique se há indicação de associação entre Preferência pelo adoçante e Gênero. Comente.

Para determinar se há indícios de associação entre a preferência pelo adoçante e o gênero, vamos calcular as porcentagens segundo os totais das linhas da tabela de contingência apresentada como segue:

Gênero	Preferência		Total
	Adoçante A	Adoçante B	
Feminino	70 (55,6%)	56 (44,4%)	126 (100%)
Masculino	130 (57,5%)	96 (42,5%)	226 (100%)
Total	200 (56,8%)	152 (43,2%)	352 (100%)

Independentemente do gênero, observou-se na amostra que 56,8% dos participantes preferem o adoçante A e que 43,2% preferem o adoçante B. Note que porcentagens similares a estas foram observadas em cada um dos gêneros (entre os indivíduos do gênero feminino 55,6% preferem A e 44,4% preferem B; enquanto que para indivíduos do gênero masculino 57% preferem A e 42,5% preferem B). Aparentemente não há indicação de associação entre a preferência pelo adoçante e gênero.

Observação: Podemos também fixar os totais das colunas. Neste caso, a tabela é dada por

Gênero	Preferência		Total
	Adoçante A	Adoçante B	
Feminino	70 (35%)	56 (36,8%)	126 (35,8%)
Masculino	130 (65%)	96 (63,2%)	226 (64,2%)
Total	200 (100%)	152 (100%)	352 (100%)

Independentemente da preferência, observou-se na amostra que 35,8% dos participantes são do gênero feminino e que 64,2% são do gênero masculino. Note que porcentagens similares a estas foram observadas para cada preferência (para indivíduos que preferem A, 35% são do gênero feminino e 65% são do gênero masculino; enquanto que para indivíduos que preferem B, 36,8% são do gênero

feminino e 63,2% são do gênero masculino. Assim, não há indicação de associação entre a preferência pelo adoçante e gênero.

(b) Verifique se há indicação de associação entre Preferência pelo adoçante e Faixa etária. Comente.

Análogo ao item (a), vamos calcular as porcentagens segundo os totais das linhas da tabela de contingência apresentada como segue:

Faixa etária	Preferência		Total
	Adoçante A	Adoçante B	
Menos de 40 anos	50 (31,2%)	110 (68,8%)	160 (100%)
40 anos ou mais	150 (78,1%)	42 (21,9%)	192 (100%)
Total	200 (56,8%)	152 (43,2%)	352 (100%)

Independentemente da faixa etária, observou-se na amostra que 56,8% dos participantes preferem o adoçante A e que 43,2% preferem o adoçante B. Note que estas porcentagens diferem consideravelmente entre as faixas etárias (para indivíduos com menos de 40 anos temos que 31,2% preferem A e 68,8% preferem B; enquanto que para indivíduos com 40 anos ou mais temos 78,1% preferem A e 21,9% preferem B). Aparentemente, há indicação de associação entre a preferência pelo adoçante e faixa etária.

Observação: Calculando as porcentagens segundo os totais das colunas, a tabela de contingência é dada por

Faixa etária	Preferência		Total
	Adoçante A	Adoçante B	
Menos de 40 anos	50 (25%)	110 (72,4%)	160 (45,5%)
40 anos ou mais	150 (75%)	42 (27,6%)	192 (54,5%)
Total	200 (100%)	152 (100%)	352 (100%)

Independentemente da preferência, observou-se na amostra que 45,5% dos participantes têm menos de 40 anos e que 54,5% têm 40 anos ou mais. Note que estas porcentagens diferem consideravelmente dependendo da preferência (para indivíduos que preferem A, 25% têm menos de 40 anos e 75% têm 40 anos ou mais; enquanto que para indivíduos que preferem B, 72,4% têm menos de 40 anos e 27,6% têm 40 anos ou mais. Aparentemente, há indicação de associação entre a preferência pelo adoçante e faixa etária.

Exercício 4.

Considere o arquivo `aeusp.xls`.

(a) Usando recursos computacionais, construa tabelas de contingência para Renda familiar (Renda) versus cada uma das seguintes variáveis: sexo (Sexo), Estado civil (Ecivil) e Tipo de trabalho (Ttrab).

Antes da construção das tabelas de contingência, segue a codificação utilizada no banco de dados

- `Sexo` : Sexo, 1 = masculino; 2 = feminino;
- `Ecivil` : Estado civil, 1 = solteiro; 2 = casado; 3 = divorciado; 4 = viúvo; 5 = outro;
- `Ttrab`: Tipo de trabalho, somente para os que trabalham, 1 = empregado com carteira; 2 = empregado sem carteira; 3 = profissional liberal; 4 = autônomo; 5 = rural;
- `Renda` : Renda familiar em faixas de reais; 1 = [0, 150); 2 = [150, 300); 3 = [300, 450); 4 = [450, 900); 5 = [900, 1500); 6 = [1500, 3000).

Toda a análise pode ser feita via `Rcmdr`. Podemos utilizar os seguintes passos:

(1) Ler o conjunto de dados: Clique na guia `Dados` \Rightarrow Importar arquivos de dados \Rightarrow do Excel. Em seguida defina um nome para o seu conjunto de dados (por exemplo, `aeusp`) e clique em OK. Navegue até o diretório onde o seu conjunto de dados está localizado e clique em Abrir.

(2) Converta as variáveis em fator: Para que o R entenda que as codificações representam níveis do fator, e não observações numéricas, é necessário converter as variáveis da classe `numeric` para `factor`. No `Rcmdr` clique em `Dados` \Rightarrow Modificar variáveis do conjunto de dados ativo \Rightarrow Converter variável numérica em fator. Selecione as variáveis `Ecivil`, `Renda`, `Sexo` e `Ttrab` pressionando a tecla `Ctrl` para selecionar mais de uma. Selecione a opção `Utilizar números` e clique em `Ok`. Sobreescreva todas as variáveis.

(3) Tabelas de contingência: Vá até a guia `Estatísticas` \Rightarrow Tabelas de contingência \Rightarrow Tabela de dupla entrada; escolha as variáveis de interesse (somente duas por vez). Clique em OK.

Observação: Segundo os dados, na variável Tipo de trabalho (`Ttrab`), só foram contabilizadas 241 respostas em razão de ser observada apenas para indivíduos que trabalham.

Tabela 1: Tabela de contingência para as variáveis Renda e Sexo.

Sexo	Renda						Total
	1	2	3	4	5	6	
1	12	30	39	51	27	9	168
2	25	46	34	69	21	21	216
Total	37	76	76	120	48	30	384

Tabela 2: Tabela de contingência para as variáveis Renda e Estado civil.

Ecivil	Renda						Total
	1	2	3	4	5	6	
1	18	34	32	50	20	10	164
2	10	29	28	56	28	15	166
3	1	1	3	3	0	2	10
4	5	2	2	1	0	2	12
5	3	10	8	10	0	1	32
Total	37	76	73	120	48	30	384

Tabela 3: Tabela de contingência para as variáveis Renda e Tipo de trabalho.

Ttrab	Renda						Total
	1	2	3	4	5	6	
1	1	11	19	30	8	9	78
2	3	11	9	21	9	3	56
3	3	0	1	3	5	0	12
4	4	23	19	23	18	8	95
Total	11	45	48	77	40	20	241

(b) Dentre as famílias cujo respondente era do sexo feminino, qual é a porcentagem que tinham renda entre 300 e 450 reais?

Segundo o resultado da Tabela 1, as famílias cujo respondente foi do sexo feminino (Sexo = 2) foram 216, e dessas, 34 tinham renda entre 300 e 450 reais (Renda = 3). Assim, a porcentagem solicitada é:

$$\frac{34}{216} = 0,1574,$$

ou seja, 15,74% tinham renda entre 300 e 450 reais.

(c) Dentre as famílias com renda entre 300 e 450 reais, qual é a porcentagem que tiveram como respondente uma pessoa casada?

Segundo o resultado da tabela 2, as famílias com renda entre 300 e 450 reais (Renda = 3) foram 73, e dessas, 28 tiveram como respondente uma pessoa casada (Ecivil = 2). Assim, a porcentagem solicitada é:

$$\frac{28}{73} = 0,3836,$$

ou seja, 38,36% tiveram como respondente uma pessoa casada.

(d) Qual é a porcentagem de famílias com renda entre 300 e 450 reais e cujo respondente estava empregado com carteira?

Segundo o resultado da tabela 3, o número de famílias com renda entre 300 e 450 reais (Renda = 3) e cujo respondente estava empregado com carteira (Ttrab = 1) foi 19. Assim, a porcentagem solicitada é a razão:

$$\frac{19}{241} = 0,0788,$$

ou seja, 7,88%.

(e) Você diria que há indicação de associação entre Renda familiar e Sexo? Justifique.

A tabela com os percentuais em relação aos totais das linhas nos ajudará a perceber uma possível associação entre as variáveis Renda e Sexo. No pacote Rcmdr, repita os comandos descritos no passo (3) **Tabelas de Contingência**, com as variáveis Sexo e Renda, mas antes de pressionar o OK, vá até a aba Estatísticas e selecione Percentual nas linhas. As tabelas geradas pelo comando são resumidas a seguir.

Tabela 4: Tabela de contingência para as variáveis Renda e Sexo.

Sexo	Renda						Total
	1	2	3	4	5	6	
1	12 (7,1%)	30 (17,91%)	39 (23,2%)	51 (30,4%)	27 (16,1%)	9 (5,4%)	168 (100%)
2	25 (11,6%)	46 (21,3%)	34 (15,7%)	69 (31,9%)	21 (9,7%)	21 (9,7%)	216 (100%)
Total	37 (9,6%)	76 (19,8%)	76 (19,0%)	120 (31,3%)	48 (12,5%)	30 (7,8%)	384 (100%)

Segundo os resultados da Tabela 4, podemos observar que não há diferenças expressivas nas porcentagens das linhas com relação à porcentagem total. Aparentemente, não há indicação de associação entre a Renda familiar e o Sexo do respondente.