MAE217 Lista 3

Guilherme Marthe — NUSP:8661962

1. Considere os dados no arquivo empresa.xls, com informações sobre funcionários de certa empresa.

a) Construa uma tabela de dupla entrada com informações sobre estado civil e escolaridade.

Escolaridade	casado	solteiro	**Total**
ensino fundamental	5	7	**12**
ensino médio	12	6	**18**
superior	3	3	**6**
Total	**20**	**16**	**36**

b) Calcule frequências relativas adequadas para avaliar descritivamente se estado civil está associado com escolaridade. Interprete.

Escolaridade	casado	solteiro	**Total**
ensino fundamental	13.9%	19.4%	**33.3%**
ensino médio	33.3%	16.7%	**50.0%**
superior	8.3%	8.3%	**16.7%**
Total	**55.6%**	**44.4%**	**100.0%**

A divisão entre casados e solteiros está relativamente igual, com quase 55% casados e 45% solterios. Existe também uma concentração de funcionários cde 50% deles tendo completado o ensino médio. Esse número, o total de funcionários que completou o ensino médio, é maior entre os casados que os solteiros. Além disso, a concentração de funcionários que completaram o ensino superior é igual entre casados e solteiros.

c) Calcule a estatística de qui-quadrado de Pearson, coeficiente de contigência e coeficiente de Tschuprov. Você diria que existe associação entre estado civil e escolaridade?

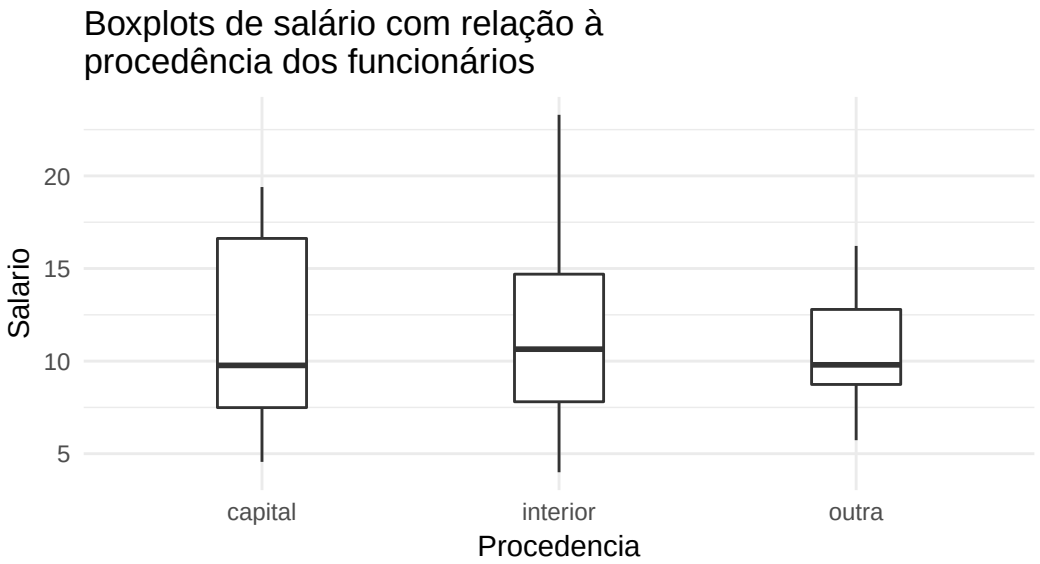
As estátisticas calculadas são:

- Chi-quadrado de pearson: $\chi^2 = 1.9125$
- Coeficiente de contingência: $C = 0.2246$
- Coeficiente de Tschuprov: $T = 0.326$

Com os coeficientes de contingência e de Tschuprov têm uma interpretação semelhante com o coeficiente de correlação de Pearson para dua variáveis quantitativas, podemos analisar que a associação entre estado civil e escolaridade é relativamente baixa quando olhamos para o valor de 0.326 do coeficiente de Tschuprov, que varia entre -1 e 1.

d) Considere agora as variáveis Procedência e Salário. Analise descritivamente (através de gráficos e medidas resumo) essas informações para verificar se o salário depende da procedência do funcionário.

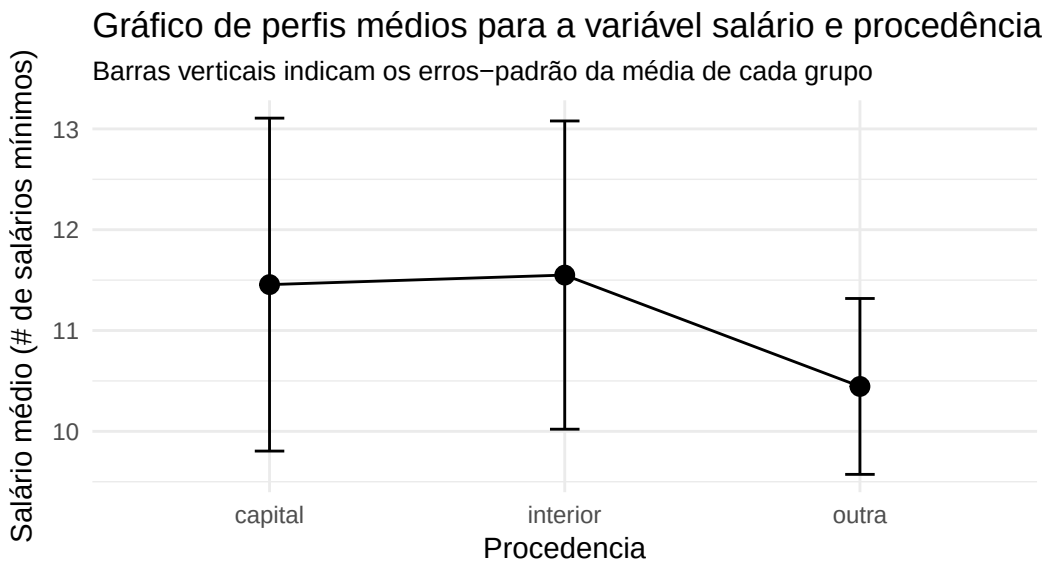
O boxplot a seguir nos mostra que funcionários do interior tem uma mediana reativamente mais alta que os outros grupos de procedência, e o comprimentos dos "bigodes" indicam a presença de valores mais altos nessa variável. O boxplot dos salários procedentes da capital possuem uma a amplitude da caixa (IQR) maior indicando uma maior variação nesse grupo.



Vamos observar então algumas medidas resumo por grupos. Como a tabela abaixo mostra, valores médios relativamente próximos, com apenas a amostra dos salários da procedência "outra" sendo ligeiramente mais baixos. Porém essa diferença parece tão pequena que indica que o valor dos salários não está fortemente relacionado com a procedência do funcionário do interior ou da capital. A procedência "outra" parece ter um salário menor.

Procedencia	Média	Desv. Padrão	Min	Q1	Mediana	Q3	Max	IQR
capital	11.45545	5.476653	4.56	7.490	9.770	16.625	19.40	9.135
interior	11.55000	5.296055	4.00	7.805	10.645	14.695	23.30	6.890
outra	10.44538	3.145453	5.73	8.740	9.800	12.790	16.22	4.050

O gráfico de perfis nos permite visualizar as conclusões que chegamos ao analisar a tabela de medidas resumo.



2. Os dados na tabela abaixo são provenientes de um estudo em que um dos objetivos é avaliar o efeito da dose de radiação gama (em centigrays) na formação de múltiplos micronúcleos em células de indivíduos normais.

dose_radiação_gama	frequência_celulas_multiplos_micronucleos	total_células_examinadas
0	1	2373
20	6	2662
50	25	1991
100	47	2047
200	82	2611
300	207	2442
400	254	2398
500	285	1746

a) Faça uma análise descritiva dos dados, calculando o risco relativo de ocorrência de micronúcleos para cada dose, tomando como base a dose nula.

Dose radiação gama	Prob. de células com micronúcleos	Risco relativo*
0	0.000	1.0
20	0.002	5.3
50	0.013	29.8
100	0.023	54.5
200	0.031	74.5
300	0.085	201.2
400	0.106	251.4
500	0.163	387.3

* Risco relativo calculado com relação ao caso de radiação nula.

b) Repita a análise do item anterior considerando agora razões de chances.

Dose radiação gama	Prob. de células com micronúcleos	Chance	Chance relativa*
0	0.000	0.000	1
20	0.002	0.002	5
50	0.013	0.013	30
100	0.023	0.024	56
200	0.031	0.032	77
300	0.085	0.093	220
400	0.106	0.118	281
500	0.163	0.195	463

* Chance reativa calculada com relação ao caso de radiação nula.

c) Considerando os dois itens anteriores, quais seriam suas conclusões?

Concluo que de fato, com um aumento da dose de radiação gama, o número de células com micronúcelos parece aumentar. Tanto o risco relativo quanto as chances relativas parece ser ~400X maior para uma pessoa que foi exposta a uma dose de radiação de 500 com relação à um idivíduo normal que não recebeu radiação. Esse aumento parece ainda ser condizente com o fato de que a chance de se ter células com micronúcleos no caso de radiação nula é praticamente 0, e essa chance aumenta gradativamente junto o aumento da radiação.

3. Um novo teste está sendo desenvolvido para a identificação do HIV. Em 200 pessoas estudadas, 100 têm HIV e 100 não têm. O teste deu positivo em 75 pessoas e negativo em 125, sendo 25 falsos-positivos e 50 falsos-negativos.

a) Construa uma tabela de dupla entrada com as informações do enunciado.

	Situação		
resultado do teste	Com HIV	Sem HIV	Total
positivo	50	25	75
negativo	75	50	125
Total	100	100	200

b) Encontre as medidas de sensibilidade e especificidade do teste.

As definições de sensibilidade e especificidade para o caso estudado são:

- **Sensibilidade**: probabilidade do teste dar positivo para pacientes com HIV
- **Especificidade**: probabilidade do teste dar negativo para pacientes sem HIV

Então, as estimativas para ambas são:

- Sensibilidade: $\frac{50}{100} = 0.5$
- Especificidade: $\frac{25}{100} = 0.25$

c) Calcule os valores preditivos positivo e negativo.

As definições de VPP (Valor preditivo positivo) e VPN (valor preditivo negativo) para o caso estudado são:

- **VPP**: probabilidade que o paciente tenha HIV dado que o teste foi positivo
- **VPN**: probabilidade que o paciente não tenha HIV que o teste é negativo

Assim, essas métricas estimadas são:

- VPP: $\frac{50}{75} = 0.67$
- VPN: $\frac{50}{125} = 0.4$

d) Qual é a acurácia do teste?

A acuracia do teste é basicamente a probabilidade de se obter resultados corretos, ou seja, a probabilidade de resutados verdadeiros positivos e negativos. No caso podemos estimá-la com o seguinte cálculo:

- Acurácia: $\frac{50+50}{200} = 0.5$

4. Um laboratório de pesquisa desenvolveu uma nova droga para febre tifóide com a mistura de duas substâncias químicas (A e B). Foi realizado um ensaio clínico com o objetivo de estabelecer as dosagens adequadas (baixa ou média para a substância A e baixa, média ou alta para a substância B) na fabricação da droga. Vinte e quatro voluntários foram aleatoriamente distribuídos em 6 grupos de 4 indivíduos e cada grupo foi submetido a um dos 6 tratamentos. A resposta observada foi o tempo para o desaparecimento dos sintomas (em dias). Os resultados obtidos estão na tabela a seguir.

dose_substancia_A	dose_B___baixa	dose_substancia_B___media	dose_substancia_B___alta
baixa	10.4	8.9	4.8
baixa	12.8	9.1	4.5
baixa	14.6	8.5	4.4
baixa	10.5	9.0	4.6
media	5.8	8.9	9.1
media	5.2	9.1	9.3
media	5.5	8.7	8.7
media	5.3	9.0	9.4

a) Faça uma análise descritiva dos dados com o objetivo de avaliar qual a combinação de dosagens das substâncias faz com que os sintomas desapareçam em menos tempo.

Primeiramente vamos checar as medidas resumo do tempo até se observar um sintoma agrupando, separadamente, para as doses de cada substância.

dose_A	Média	Desv. Padrão	Min	Q1	Mediana	Q3	Max	IQR
baixa	8.508	3.382	4.4	4.750	8.95	10.425	14.6	5.675
media	7.833	1.777	5.2	5.725	8.80	9.100	9.4	3.375

Como podemos ver, a dose média da substância A parece estar marginalmente relacionada com um menor tempo de desaparecimento de sintomas. Além disso ela parece ter menos variação que a dose baixa.

dose_B	Média	Desv. Padrão	Min	Q1	Mediana	Q3	Max	IQR
baixa	8.762	3.783	5.2	5.450	8.10	11.075	14.6	5.625
media	8.900	0.207	8.5	8.850	8.95	9.025	9.1	0.175
alta	6.850	2.443	4.4	4.575	6.75	9.150	9.4	4.575

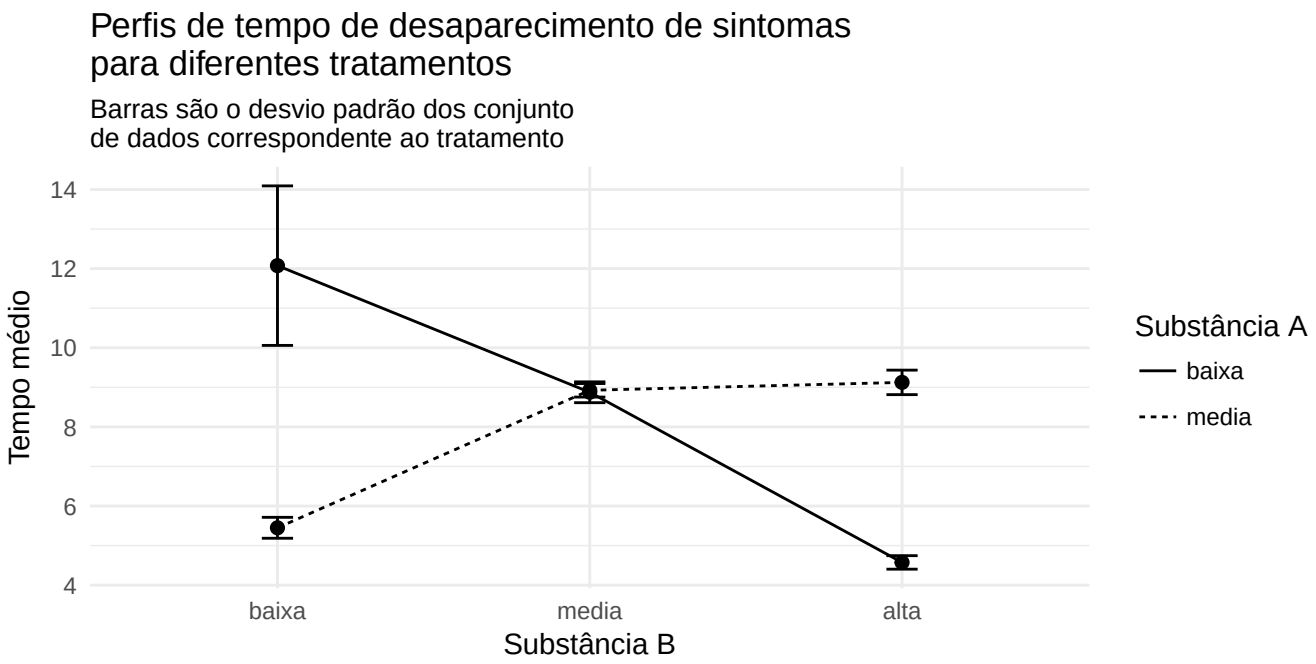
A medida de tempo de desaparecimento dos sintomas parece ser menor para doses altas da substância B. Além disso, podemos ver que o grupo tratado com a dose média da substância B possui uma baixa variação, olhando para o desvio padrão desse grupo.

b) Especifique o modelo para a comparação dos 6 tratamentos quanto ao tempo para o desaparecimento dos sintomas. Identifique os fatores e seus níveis.

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

Onde $i = 1, 2$ se a dose da substância a for baixa ou média, respectivamente e onde $j = 1, 2, 3$ se a dose da susbtância B for baixa, média ou alta respectivamente. O índice $k = 1, ..., 24$ representa uma observação do estudo, no caso um paciente diferente. Assim temos uma modelos com $2\beta = 6$ tratamentos, onde cada combinação de i e j é um tratamento. Por fim, y_{ijk} representa o tempo médio de desaparecimento para os sintomas de um passageiro tratado com as doses i e j de A e B.

c) Construa o gráfico de perfis médios e interprete-o. Com base nesse gráfico, você acha que existe interação entre os fatores? Justifique sua resposta.



Quando obsevamos o quando o tempo de desaparecimento dos sintomas da substância A em nível baixo quando aplicada junto com cada nível de sa substância B, podemos ver que ele decresce. Quando fazemos a mesma análise para o nível médio da substância A, apresenta-se uma tendência crescente do tempo de desaparecimento dos sintomas. Isso indica uma interação essencial entre os tratamentos, uma vez que quando analisamos as tendências formadas pelos níveis da substância A para cara nível de B, temos inclinações/tendências opostas de comportamento do tempo de desaparecimento dos sintomas.

5. Numa cidade do interior de São Paulo, exige-se a publicação de informações sobre proprietários inadimplentes com taxas públicas. A publicação lista o nome do proprietário, o valor da propriedade, quantia devida, avaliações e juros além das respectivas penalidades. O valor da propriedade e a quantia de taxas devidas para uma amostra de 10 propriedades são mostrados a seguir.

valor_mil_reais	quantia_devida
18.8	445
24.4	539
20.4	1212
35.8	2237
14.8	479
40.4	1181
49.0	4187
14.5	409
37.3	1002
54.7	2062

a) Calcule o coeficiente de correlação de Pearson e o coeficiente de correlação de Spearman para os dados apresentados. Comente.

correlação	valor
Pearson	0.7530873
Spearman	0.7939394

Os coeficientes de correlação variam entre -1 e 1, onde os extremos indicam uma associação alta, decrescente e crescente entre as variáveis estudadas. Se os coeficientes forem próximos de zero, a associação entre as variáveis é baixa.

No caso, temos valores altos para ambos os coeficientes, indicando que a associação entre eles é positiva. Todavia, o coeficiente de correlação de Spearman tem a vantagem de medir também a relação não-linear monotônica entre variáveis. Como esse valor é relativamente alto, pode se dizer que existe essa relação positiva e não-linear entre as variáveis.

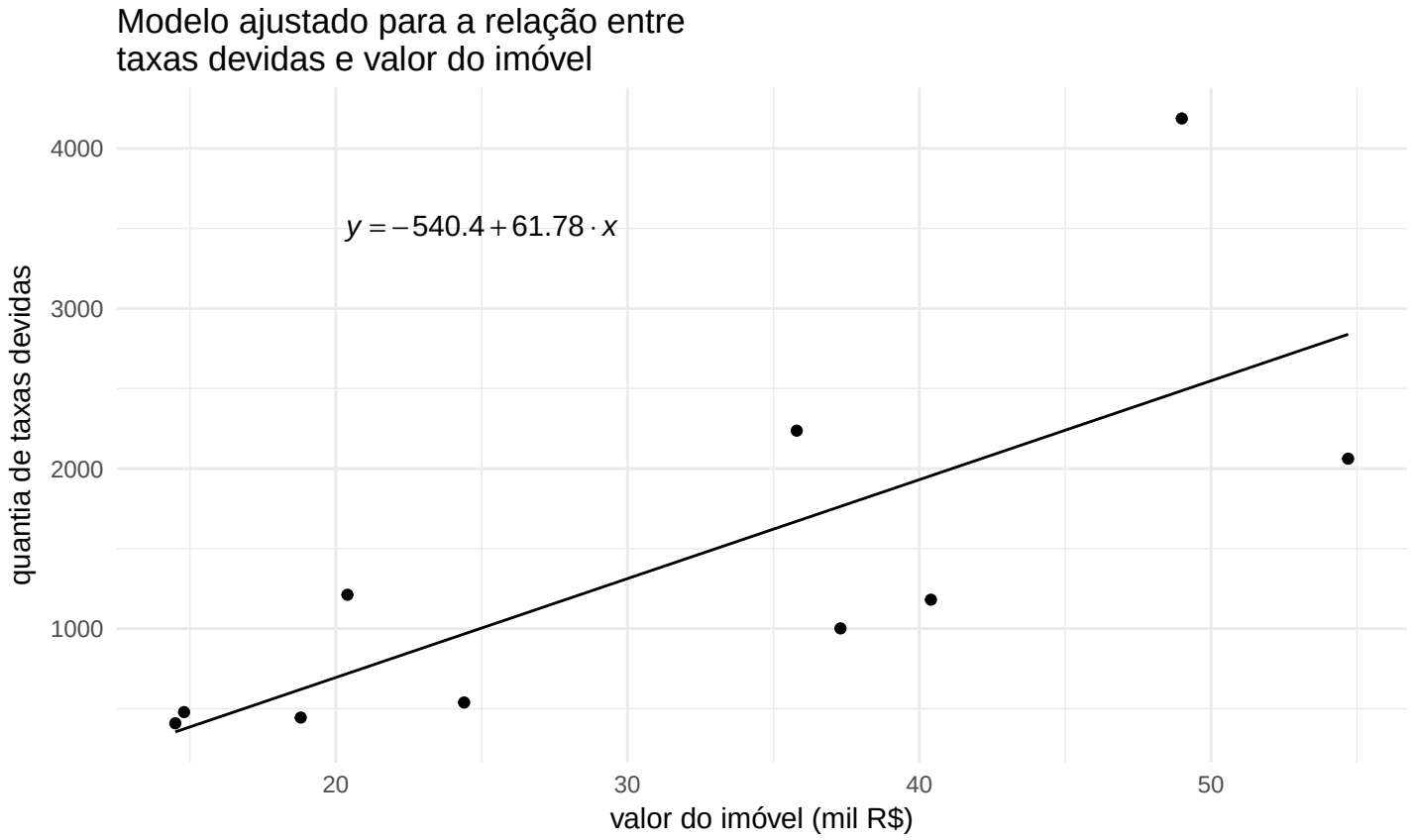
b) Especifique um modelo de regressão que poderia ser utilizado para estimar a quantia média devida em taxas dado o valor da propriedade. Ajuste-o, apresentando a reta estimada no diagrama de dispersão.

Neste caso podemos ajustar o seguinte modelo linear aos dados:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

Onde y_i é a quantia devida de taxas pelo proprietário e x_i é o valor da propreidade do devedor, e os outros termos são so parâmetros à serem estimados.

Abaixo temos o resultado do modelo ajustado junto com o diagrama de dispersão dos dados.



c) Use a equação estimada para a prever a taxa média devida de uma propriedade da cidade cujo valor é igual a R$ 42.400,00.

É estimado pelo modelo que o proprietário que possua uma propriedade no valor de **R$ 42.400,00** esteja devendo **2.079** em taxas.

Essa estimativa é dada pela seguinte equação: $\hat{y_i} = -540.4 + 61.78 \cdot 42.40 = 2078.939$

d) Calcule o coeficiente R2. Comente.

O R2 do modelo estimado é 0.57. Interpreta-se essa quantia como sendo 56% das variações da variável dependente, o valor das taxas devidas, são explicadas pela variáve explicativa, o valor da propriedade no caso. O restante das variações deve ser explicado por fatores ausentes no modelo.