Testes de Hipóteses

Introdução

Apresentaremos, neste capítulo, os testes de hipóteses mais utilizados do ponto de vista paramétrico e não-paramétrico. Os testes paramétricos exigem que seja verificada a pressuposição de que os dados coletados sejam normalmente distribuídos enquanto que os testes não-paramétricos não fazem essa exigência e por isso são considerados menos consistentes, sendo, porém, uma alternativa a ser usada caso os pressupostos de normalidade não sejam observadas ou, ainda, quando o tamanho da amostra não é suficientemente grande. No caso paramétrico, como o nome já diz, o objetivo é testar hipóteses acerca de parâmetros, com base em dados amostrais. No caso não-paramétrico, as hipóteses não são formuladas em termos de parâmetros, já que não há preocupação com a distribuição que os dados seguem. Para cada tipo de plano experimental existem testes específicos a serem utilizados. Nos preocuparemos aqui com os seguintes planos: a) comparação de duas amostras independentes; b) comparação de duas amostras relacionadas; c) comparação de três ou mais amostras independentes; d) teste de aderência.

Comparação de duas amostras independentes

Neste caso estamos interessados em comparar duas populações, representadas cada uma por suas respectivas amostras. Não necessariamente as duas amostras têm o mesmo tamanho. Os principais testes são:

- Teste t de Student para médias;
- Teste Z para proporções;
- Teste Mann-Whitney (não-paramétrico)

Teste t de Student para comparação de médias

A média de uma população é uma de suas características mais importantes. É muito comum desejarmos tomar decisões a seu respeito, por exemplo,

quando são comparadas duas amostras ou dois tratamentos. Considere as seguintes hipóteses:

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 < \mu_2$$

ou

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 > \mu_2$$

ou ainda

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

As duas primeiras situações definem os chamados testes *unilaterais*, por que a região de rejeição está somente em uma das caudas da distribuição. A última situação define os testes *bilaterais*, no qual a região de rejeição se distribui igualmente em ambas as caudas da distribuição.

Assim, se estivermos interessados em mostrar que um parâmetro é significativamente superior ou inferior a um determinado valor, teremos que realizar um teste unilateral e teremos uma única região de rejeição, do tamanho do nível de significância fixado. Mas se, no entanto, estivermos interessados em mostrar que um determinado parâmetro é diferente de um determinado valor (sem especificar se inferior ou superior) teremos que realizar um teste bilateral e a região de rejeição será dividida em duas partes iguais, nas extremidades da curva do teste, em que cada região de rejeição terá metade do nível de significância.

Dessa forma, para realização do teste, deveremos primeiramente estimar a média e o desvio padrão de cada uma das amostras envolvidas e calcular a estatística do teste:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad \textbf{(1)}$$

a qual tem distribuição t de Student com $n_1 + n_2 - 2$ graus de liberdade. Nesse caso, supõe-se que *as variâncias amostrais são diferentes*. Caso as variâncias *não sejam diferentes*, devemos usar:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \textbf{(2)}$$

onde:

– $\overline{X}_1$ e $\overline{X}_2$ são as médias amostrais do grupo 1 e 2 respectivamente;

– S_1 e S_2 são os desvios padrões do grupo 1 e 2 respectivamente;

– n_1 e n_2 são os tamanhos de amostra do grupo 1 e 2 respectivamente;

$$S_p^2 = \frac{(n_1-1).S_1^2+(n_2-1).S_2^2}{n_1+n_2-2}$$

A tabela abaixo resume o procedimento a ser seguido:

Tabela 1. Decisão nos testes de comparação de médias

Hipóteses	Decisão
$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 < \mu_2$	rejeita H_0 se, $t < -t(\alpha)_{n_1+n_2-2}$
$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 > \mu_2$	rejeita H_0 se, $t > t(\alpha)_{n_1+n_2-2}$
$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$	rejeita H_0 se, $\lvert t \rvert > t(\alpha/2)_{n_1+n_2-2}$

Exemplo: Um teste de resistência a ruptura feito em seis cabos usualmente utilizados acusou resistência média de 3 530kg com variância de 660kg. Um novo cabo foi testado e verificou-se uma resistência média de 3 560kg e variância de 600kg em uma amostra de tamanho 8. Compare as médias dos dois cabos, ao nível de significância α = 5%. E se a variância do cabo novo fosse 850kg?

Assim, queremos testar se $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$. O teste é bilateral pois se deseja verificar se os dois cabos diferem em relação à resistência média, sem especificar para que lado. Usaremos a expressão (2), pois vamos considerar as variâncias "iguais" (ou seja, muito próximas). Rigorosamente, essa verificação deveria ser feita através da aplicação do teste F para razão de variâncias. Considerando válida essa suposição de igualdade das variâncias, teremos:

$$S_p^2 = \frac{(6-1).660+(8-1).600}{6+8-2} = 625 \quad \text{e} \quad t = \frac{(3530-3560)}{25\sqrt{\frac{1}{6}+\frac{1}{8}}} = -2,22.$$

O valor crítico $t(\alpha/2)_{n_1+n_2-2}$ para α = 5% é dado por 2,179. Este valor é encontrado na tabela t de Student consultando a coluna 0,025 (pois o teste é bilateral) e a linha 12 ($n_1 + n_2 - 2$). Assim, teremos 2 valores críticos, –2,179 e

+2,179. Como t < –2,179, rejeitamos a hipótese nula e afirmamos que existe diferença significativa entre os dois tipos de cabo. Os dois cabos diferem significativamente em relação à resistência média.

Agora, considerando que $S_2^2 = 850$kg teremos,usando a expressão (1):

$$t = \frac{(3530 - 3560)}{\sqrt{\frac{660}{6} + \frac{850}{8}}} = -2{,}04$$

e, neste caso, a nossa decisão será exatamente o contrário do que obtivemos, ou seja, como t > –2,179 não rejeitamos a hipótese nula e não observamos diferença entre os cabos.

Teste Z para comparação de proporções

Em alguns estudos, o interesse está em comparar duas proporções provenientes de amostras distintas. Nesse caso, obtém-se n_1 observações da população 1 e n_2 observações da população 2. Verifica-se em cada uma das amostras o total x_1 e x_2, respectivamente, de "sucessos" e calculam-se as proporções amostrais $p_1 = \frac{x_1}{n_1}$ e $p_2 = \frac{x_2}{n_2}$. As hipóteses testadas são as seguintes:

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 < P_2$$

ou

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 > P_2$$

ou ainda

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 \neq P_2$$

A estatística do teste é dada por:

$$Z = \frac{p_1 - p_2}{S_p} \quad \textbf{(3)}$$

Onde $S_p = \sqrt{\frac{p.(1-p)}{n_1} + \frac{p.(1-p)}{n_2}}$ **(4)** e $p = \frac{n_1.p_1 + n_2.p_2}{n_1 + n_2}$ **(5)**

Exemplo: Em uma cidade do interior realizou-se uma pesquisa eleitoral com 200 eleitores, na qual o candidato a presidente X aparece com 35%

das intenções de voto. A mesma pesquisa também foi realizada na cidade vizinha, com 500 eleitores, e o mesmo candidato surge com 28% das intenções de voto. Podemos afirmar estatisticamente que na primeira cidade o candidato X apresenta uma maior intenção de voto? (nível de significância $\alpha = 0,05$)

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 > P_2$$

É um teste unilateral pois está claramente verificado se na primeira pesquisa foi encontrada uma proporção maior do que na segunda cidade.

Pela expressão (5) temos $p = \frac{(200.0,35)+(500.0,28)}{200+500} = 0,3$ e pela expressão (4)

$$S_p = \sqrt{\frac{0,3.(1-0,3)}{200}+\frac{0,3.(1-0,3)}{500}} = 0,038 \text{ e finalmente:}$$

$$Z = \frac{0,35-0,28}{0,038} = 1,84$$

Ao nível de significância de 5% temos Z (α) = 1,64. Este valor crítico é obtido na tabela da distribuição normal padrão, considerando uma área marcada em cinza de tamanho 0,45, ou seja, 0,5 – 0,05. Localizando o valor 0,45 no corpo da tabela (ou o valor mais próximo), veremos que ele se localiza na linha 1,6 e na coluna 0,04. Então, somamos os dois valores e obtemos 1,64.

Como a estatística Z calculada é superior ao valor crítico, rejeitamos a hipótese nula. Existem evidências para admitir que na primeira cidade o candidato X apresenta uma proporção significativamente superior de intenção de voto.

Teste não-paramétrico de Mann-Whitney

Esse teste se aplica na comparação de dois grupos independentes, para se verificar se pertencem ou não à mesma população. É a alternativa a ser usada quando as suposições de normalidade não são verificadas. Considere, portanto, duas amostras de tamanho n_1 e $n_{2,}$ respectivamente. O teste consiste basicamente na substituição dos dados originais pelos seus respectivos postos ordenados (*ranks*) e cálculo da estatística do teste. Além disso, o

procedimento de teste depende do tamanho das amostras. Considere o grupo 2 aquele com o maior número de observações:

- Quando $9 \leq n_2 \leq 20$, calcula-se:

  $U = n_1.n_2 + \frac{n_1.(n_1+1)}{2} - R_1$, onde R_1 é a soma dos postos atribuídos aos valores do grupo 1.

- $n_2 > 20$

  Utiliza-se nesse caso a aproximação normal dada por:

$$\mu_U = \frac{n_1.n_2}{2} \qquad \sigma_U = \sqrt{\frac{n_1.n_2.(n_1+n_2+1)}{12}} \qquad z = \frac{U-\mu_U}{\sigma_U}$$

Os valores da estatística calculada são comparados com os valores críticos obtidos a partir de uma tabela (Mann Whitney). Caso a estatística U calculada seja inferior ao valor crítico deveremos rejeitar a hipótese nula.

Exemplo: Dois tipos de solução química, A e B, foram ensaiadas para determinação de Ph. As análises de amostras de cada solução estão apresentadas na tabela que segue. Verifique se há diferença entre elas.

$H_0: Ph_A = Ph_B$

$H_a: Ph_A > Ph_B$

A	Posto (A)	B	Posto (B)
7,49	13	7,28	2
7,35	4,5	7,35	4,5
7,54	19	7,52	17,5
7,48	11	7,50	14,5
7,48	11	7,38	7
7,37	6	7,48	11
7,51	16	7,31	3
7,50	14,5	7,22	1
7,52	17,5	7,41	8
		7,45	9
	R_A = 112,5		**R_B =77,5**

$$U = (9.10) + \frac{(9.10)}{2} - 112,5 = 22,5$$

O valor crítico para $n_1 = 9$ e $n_2 = 10$ em que $\alpha = 0{,}05$ (teste unilateral) será $U_c = 24$. Como o valor calculado da estatística é inferior ao valor crítico então iremos rejeitar H_0. Assim, temos evidências suficientes para afirmar que a solução química A apresenta Ph superior à solução química B.

Comparação de duas amostras relacionadas

Neste caso estamos interessados em comparar uma amostra extraída em dois momentos distintos. Deseja-se verificar se a diferença observada entre os dois momentos (efeito do tratamento) é significativa. Os principais testes são:

- Teste t de Student para dados pareados;
- Teste de Wilcoxon (não-paramétrico)

Teste t para dados pareados

Para observações pareadas, o teste apropriado para a diferença entre as médias das duas amostras consiste em primeiro determinar a diferença **d** entre cada par de valores e então testar a hipótese nula de que a média das diferenças na população é zero. Então, do ponto de vista de cálculo, o teste é aplicado a uma única amostra de valores **d**.

A diferença média para um conjunto de observações pareadas é $\bar{d} = \frac{\sum d}{n}$

e o desvio padrão das diferenças das observações pareadas é dado por:

$$S_d = \sqrt{\frac{\sum d^2 - n\bar{d}^2}{n-1}}$$

e a estatística do teste será: $t = \frac{\bar{d}}{S_d / \sqrt{n}}$ **(6)**

Essa estatística deve ser comparada com o valor crítico do teste t de Student para determinado nível de significância α e n–1 graus de liberdade.

Exemplo: Considere o experimento realizado com 10 automóveis de certa fábrica. Os veículos foram avaliados com dois tipos de combustíveis. Primeiramente, um combustível sem aditivo e em seguida o mesmo combustível com aditivo. Deseja-se verificar se os automóveis conseguem uma quilo-

metragem maior com a utilização do combustível com aditivo. Seguem os dados abaixo:

Automóvel	Quilometragem sem aditivo (B)	Quilometragem com aditivo (A)	d (A–B)
1	26,2	26,7	0,5
2	25,2	25,8	0,6
3	22,3	21,9	-0,4
4	19,6	19,3	-0,3
5	18,1	18,4	0,3
6	15,8	15,7	-0,1
7	13,9	14,2	0,3
8	12,0	12,6	0,6
9	11,5	11,9	0,4
10	10,0	10,3	0,3
Total	**174,6**	**176,8**	**2,2**

$$H_0: \mu_A = \mu_B \quad vs \quad H_a: \mu_A < \mu_B$$

Pelos dados da tabela temos $\bar{d}$ =0,22 e Sd = 0,361

Assim, $t = \dfrac{0,22}{\dfrac{0,361}{\sqrt{10}}} = 1,927$ e comparando com o valor crítico t (0,05) com 9 graus de liberdade que é 1,833, podemos concluir que o valor calculado se encontra dentro da região de rejeição, ou seja, existe diferença significativa entre as quilometragens obtidas com e sem aditivo. A quilometragem obtida com aditivo é significativamente superior.

Note que o valor crítico 1,833 foi encontrado na tabela t de Student na coluna 0,05 (pois o teste é unilateral) e linha 9.

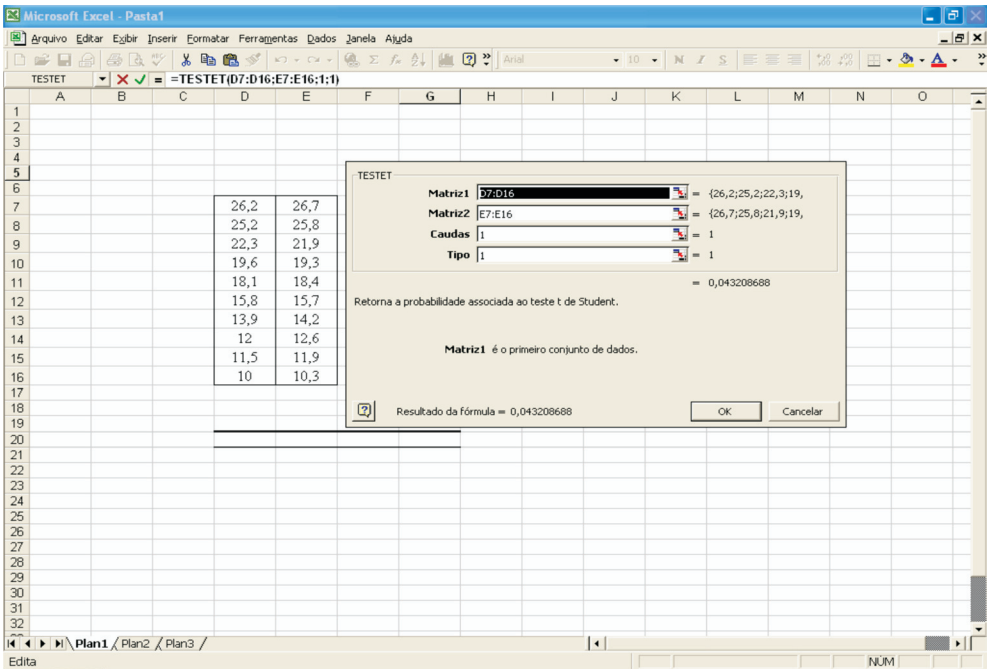
Com a planilha *Excel*, é possível realizar diversos testes de significância estatística, desde que se possuam os dados brutos. Para resolver esse exemplo, usaríamos a função TESTET, considerando:

Matriz 1: conjunto de dados referente ao primeiro grupo;

Matriz 2: conjunto de dados referente ao segundo grupo;

Caudas: indica se o teste é unilateral (1) ou bilateral (2). No caso, aqui o teste é unilateral;

Tipo: indica o tipo do teste, se é pareado (1) ou de amostras independentes (2 ou 3). No caso, aqui o teste é pareado.



Observe que a planilha irá fornecer p–valor = 0,0432, que, comparado com o nível de significância de 0,05, indica a existência de diferença significativa.

Teste de Wilcoxon

Neste teste não-paramétrico, devemos considerar as diferenças di's, onde $di = Y_i - X_i$. Devemos ordenar os di's, atribuindo postos do menor para o maior, sem considerar o sinal da diferença (em módulo). A continuação do teste, a partir daqui, depende do tamanho da amostra:

- $n < 25$

Considere T sendo a menor soma dos postos de mesmo sinal. Compara-se então o valor de T calculado com aqueles tabelados. O objetivo é testar se a mediana é nula, ou seja,

H_0: Mediana = 0

H_a: Mediana > 0

Mediana < 0

Mediana ≠ 0

Iremos rejeitar a hipótese nula quando o valor calculado de T for inferior ao valor crítico definido pelo nível de significância.

- $n \geq 25$

Nesse caso, T tem distribuição aproximadamente normal e podemos usar a aproximação considerando:

$$\mu_T = \frac{N.(N+1)}{4} \qquad e \qquad \sigma_T = \sqrt{\frac{N.(N+1).(2N+1)}{24}}$$

Calcula-se assim a estatística $z = \frac{T-\mu_T}{\sigma_T}$ e compara-se com os valores tabelados da distribuição de Z (Normal Padrão).

Podem ocorrer alguns empates. Nesse caso, deveremos considerar duas situações:

- Quando $X_i = Y_i$, ou seja, a informação pré equivale à informação pós para um mesmo indivíduo, descarta-se esse par da análise e redefinimos **n** como sendo o número de pares, tais que $X_i \neq Y_i$ para i = 1, 2, 3, ... , n.
- Quando duas ou mais di's tem o mesmo valor, atribui-se como posto a média dos postos que seriam atribuídos a eles caso não ocorresse empate.

Exemplo:

Di	|di|	Postos	Cálculo para Empates
-5	5	2*	$\rightarrow \frac{1+2+3}{3}$
5	5	2*	
5	5	2*	
7	7	4	
10	10	5	
-13	13	6,5**	$\rightarrow \frac{6+7}{2} = 6,$
13	13	6,5**	
15	15	8	

Exemplo: Numa pesquisa realizada em dois momentos distintos em 11 empresas operadoras de telefonia celular, investigou-se o % de clientes que avaliaram positivamente cada uma delas:

	% de avaliação positiva		di	|di|	p
Operadora	1° momento	2° momento			
1	8,7	7,7	1,0	1,0	4
2	18,6	9,6	9,0	9,0	9
3	8,0	16,0	–8,0	8,0	6
4	12,9	13,4	–0,5	0,5	2
5	10,9	9,6	1,3	1,3	5
6	13,4	13,0	0,4	0,4	1
7	11,9	23,7	–11,8	11,8	11
8	14,3	6,2	8,1	8,1	7
9	20,0	9,6	10,4	10,4	10
10	14,4	13,8	0,6	0,6	3
11	6,6	15,1	–8,5	8,5	8

Aplicando o teste de Wilcoxon, testaremos as seguintes hipóteses:

$$H_0 : \mu_T = 0 \text{ vs } H_a : \mu_T \neq 0$$

Somando-se os postos associados a diferenças negativas, teremos T = 6 + 2 + 11 + 8 = 27. O valor crítico, consultando a linha n = 11 e $\alpha = 0,05$ é igual a 13 (na verdade, o nível de significância aqui acaba sendo um valor próximo de 0,05, mais precisamente, 0,0471). Assim, não podemos rejeitar H_0, ou seja, a porcentagem de avaliação positiva não se modificou nos dois momentos.

Comparação de 3 ou mais amostras independentes

Esse tipo de plano é uma extensão do caso em que duas amostras independentes estão sendo comparadas, mas agora para o caso de 3 ou mais amostras. Se houver pelo menos um par de amostras diferentes, o teste irá apontar diferença significativa. No caso paramétrico, a opção é o teste F de Snedecor, também chamado de Análise de variância ou Anova. Mais uma vez aqui não há necessidade de os grupos que estarão sendo comparados terem tamanhos de amostras iguais. Consideremos, então, a seguinte estrutura de dados:

Tratamentos				
1	**2**	**3**	**...**	**k**
X_{11}	X_{21}	X_{31}	...	X_{K1}
X_{12}	X_{22}	X_{32}	...	X_{K2}
X^{13}	X_{23}	X_{33}	...	X_{K3}
..	...	...	...	...
X_{1n_1}	X_{2n_2}	X_{3n_3}	...	X_{Kn_K}

Análise de Variância

Uma análise de variância permite que vários grupos sejam comparados a um só tempo, utilizando variáveis contínuas. O teste é paramétrico (a variável de interesse deve ter distribuição normal) e os grupos têm que ser independentes. As hipóteses testadas são as seguintes:

$H_0 : \mu_1 = \mu_2 = ... = \mu_k$ vs H_1 : pelo menos um par $\mu_i \neq \mu_j$, para $i \neq j$

Os elementos que compõem o cálculo da Anova são sumarizados na tabela abaixo:

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Entre grupos	SQA	k – 1	QMA = $SQA/k-1$	QMA/QME
Erro amostral	SQE	N – k	QME = $SQE/N-k$	
Total	**SQT**	**N – 1**		

$$SQA = \sum \frac{T_k^2}{n_K} - \frac{T^2}{N} \quad \textbf{(7)} \quad \text{e} \quad SQT = \sum_{i=1}^{n}\sum_{k=1}^{k} X^2 - \frac{T^2}{N} \quad \textbf{(8)} \quad \text{e} \quad SQE = SQT - SQA$$

- T_k é a soma dos valores de um certo tratamento k;
- n_k é o número de observações no tratamento k;
- T^2 é a soma de todos os valores amostrados elevada ao quadrado;
- N é o número total de observações;
- X é cada observação amostrada.

O valor calculado de F é comparado com o valor crítico, definido pelo nível de significância e pelos graus de liberdade k – 1 e N – k. Caso $F_{cal} > F_{crit}$, devemos rejeitar a hipótese nula.

Exemplo: Quinze pessoas que participaram de um programa de treinamento são colocadas, de forma aleatória, sob três diferentes tipos de ensino. Os graus obtidos no exame de conclusão do treinamento são apresentados abaixo. Teste a hipótese de que não existe diferença significativa entre os 3 métodos de instrução, a um nível de significância de 5%.

Métodos de instrução		
A_1	A_2	A_3
86	90	82
79	76	68
81	88	73
70	82	71
84	89	81

$H_0 : \mu_1 = \mu_2 = \mu_3$ vs H_1 : pelo menos um par $\mu_i \neq \mu_j$, para $i \neq j$ $i, j = 1, 2, 3$.

Analisando a tabela acima, obtemos as seguintes informações:

$n_1 = n_2 = n_3 = 5$

$T_1 = 400$ $T_2 = 425$ $T_3 = 375$ $T = 1\,200$

$T_1^2 = 160\,000$ $T_2^2 = 180\,625$ $T_3^2 = 140\,625$ $T = 1\,440\,000$

Calculando as expressões (7) e (8):

$$SQA = \sum \frac{T_k^2}{n_K} - \frac{T^2}{N} = \left(\frac{160\,000}{5} + \frac{180\,625}{5} + \frac{140\,625}{5} \right) - \frac{1\,440\,000}{15} = 250$$

$$SQT = \sum_{i=1}^{n} \sum_{k=1}^{k} X^2 - \frac{T^2}{N} = 96\,698 - 96\,000 = 698$$

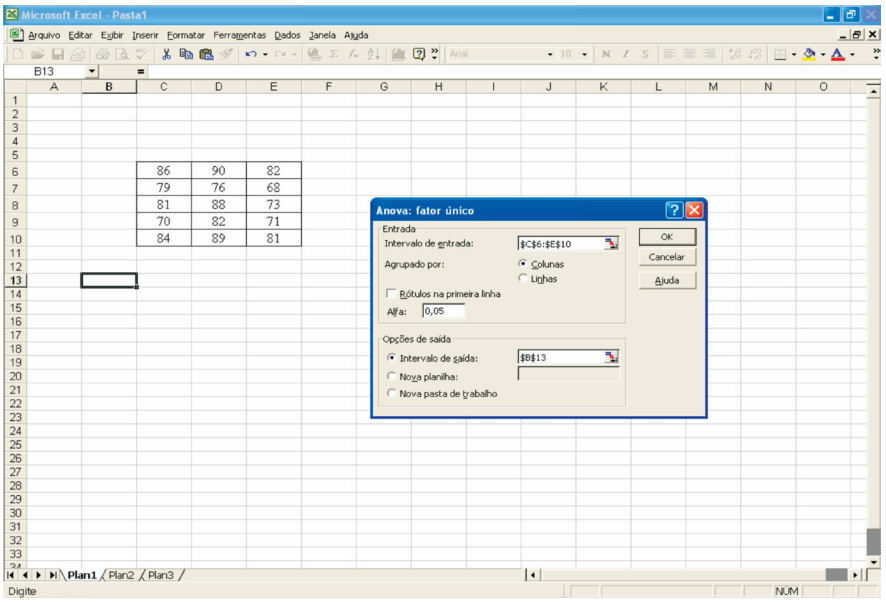
$SQE = 698 - 250 = 448$

A tabela da Anova fica então:

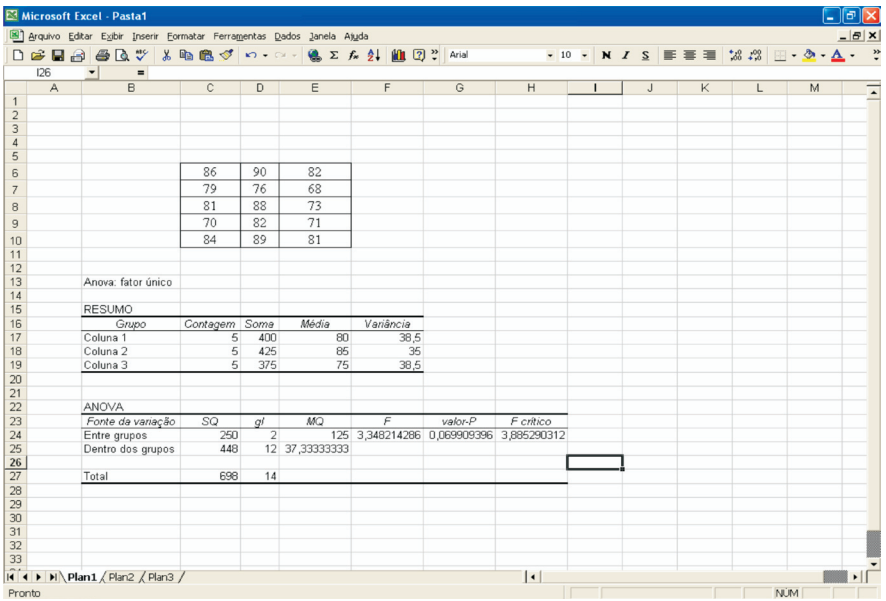
Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Entre grupos	250	2	125	3,35
Erro amostral	448	12	37,33	
Total	**698**	**14**		

Comparando o valor de F calculado com o valor crítico de 3,89, que é obtido considerando-se $\alpha = 0,05$ e cruzando a coluna $n_1 = 2$ e linha $n_2 = 12$ (graus de liberdade), podemos concluir que não há diferença significativa entre os métodos de instrução.

Com a planilha *Excel*, selecionamos FERRAMENTAS E ANÁLISE DE DADOS e selecionamos a opção: Anova: fator único.



A planilha nos fornecerá o seguinte resultado:



Teste de Kruskal-Wallis

Outro teste útil na comparação de **k** tratamentos independentes é o teste de Kruskal-Wallis. Ele nos indica se há diferença entre pelo menos dois deles. É na verdade uma extensão do teste de Wilcoxon para duas amostras independentes e se utiliza dos postos atribuídos aos valores observados.

Primeiramente, deve-se atribuir um posto a cada valor observado, sempre atribuindo o menor posto ao menor valor e o maior posto ao maior valor. Após se efetuar a soma dos postos para cada tratamento (R_j) calcula-se a estatística H:

$$H = \frac{12}{N.(N+1)} \cdot \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3 \cdot (N+1)$$

onde $\mathbf{n_j}$ é o número de observações do j-ésimo tratamento, **N** é o total de observações e $\mathbf{R_j}$ é a soma de postos do tratamento j.

Compara-se o valor calculado H com o valor crítico, que é definido pelo nível de significância e pelos tamanhos de amostra n_1, n_2, ..., n_k. Caso o valor de H calculado seja superior ao valor crítico, rejeita-se H_0.

Exemplo: Numa pesquisa sobre qualidade de vinho, foram provados três tipos por cinco degustadores. Cada degustador provou 12 amostras (4 de cada tipo) e atribuiu a cada uma delas uma nota de zero a dez. As médias das notas atribuídas pelos 5 degustadores a cada uma das amostras foram:

Tipo 1	Posto	Tipo 2	Posto	Tipo 3	Posto
5,0	1	8,3	7	9,2	11
6,7	2	9,3	12	8,7	9
7,0	4	8,6	8	7,3	5
6,8	3	9,0	10	8,2	6

Vamos verificar se há preferência dos degustadores por algum dos tipos de vinho.

H_0: não existe preferência por algum tipo de vinho

H_1: existe pelo menos uma diferença nas comparações realizadas entre os vinhos.

Calculando-se a estatística do teste, considerando $R_1 = 10$, $R_2 = 37$ e $R_3 = 31$

$$H = \frac{12}{12.13} \cdot 607,5 - 3.(12+1) = 7,73$$

O valor crítico ao nível de significância de 5% é 5,6923. Este valor é obtido na tabela fazendo $n_1 = 4$, $n_2 = 4$ e $n_3 = 4$. O nível de significância é precisamente 0,049. Desta forma, rejeitamos a hipótese nula. Certamente o vinho tipo 1 é considerado inferior pelos degustadores.

Testes de aderência

Estes testes são úteis para verificar se determinada amostra pode provir de uma população ou distribuição de probabilidade especificada. São usualmente conhecidos como testes de aderência ou bondade do ajuste. Nesse caso, retira-se uma amostra aleatória e compara-se à distribuição amostral com a distribuição de interesse.

Teste Qui-quadrado

É um teste amplamente utilizado em análise de dados provenientes de experimentos, em que o interesse está em observar freqüências em diversas categorias (pelo menos duas).

É uma prova de aderência útil para comprovar se a freqüência observada difere significativamente da freqüência esperada. Está geralmente especificada por uma distribuição de probabilidade.

Para utilizar o teste, não devemos ter mais de 20% das freqüências esperadas abaixo de 5 e nenhuma freqüência esperada igual a zero. Para evitar freqüências esperadas pequenas, devem-se combinar as categorias até que as exigências sejam atendidas.

Após definirmos a hipótese nula, testamos se as freqüências observadas diferem muito das freqüências esperadas da seguinte forma:

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \text{ em que}$$

- k = número de categorias (classes)
- o_i = freqüência observada na categoria i
- e_i = freqüência esperada na categoria i

Quanto maior o valor de X^2, maior será a probabilidade de as freqüências observadas estarem divergindo das freqüências esperadas.

A estatística do teste X^2 tem distribuição Qui-Quadrado com k – 1 graus de liberdade. Depois de calculada a estatística do teste, deve-se compará-la com o seu respectivo valor crítico, definido pelo nível de significância e graus de liberdade.

Exemplo: Deseja-se testar se a posição de largada de um cavalo (por dentro ou por fora) influencia o resultado de uma corrida de cavalos.

Posição	1	2	3	4	5	6	7	8
Número de Vitórias	29	19	18	25	17	10	15	11
	18*	18*	18*	18*	18*	18*	18*	18*

* Resultado esperado pela hipótese nula

$$H_0 : f_1 = f_2 = \cdots = f_8 \quad \text{versus} \quad H_a : f_1 \neq f_2 \neq \cdots \neq f_8$$

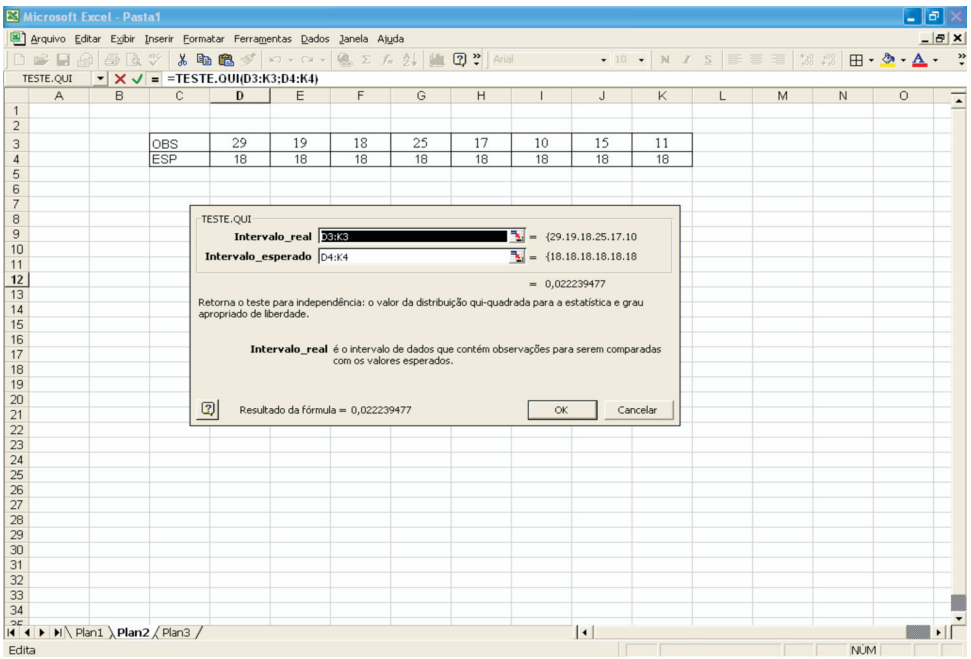
$$X^2 = \sum_{k=1}^{8} \frac{(o_i - e_i)^2}{e_i} = \frac{(29-18)^2}{18} + \frac{(19-18)^2}{18} + \cdots + \frac{(11-18)^2}{18} = 16,3$$

A tabela Qui-quadrado com 7 graus de liberdade indica que o valor 14,06 está associado a um nível de significância de 5%. Este valor é obtido na tabela, cruzando as informações da coluna 0,05 e linha 7. Nota-se que o valor calculado do qui-quadrado é superior ao valor crítico, o que nos leva a rejeitar a hipótese nula. Portanto, temos evidência de que a posição de largada dos cavalos influencia no resultado da corrida.

Com a planilha *Excel*, usaríamos a função TESTE.QUI, considerando:

Intervalo_real: posição das freqüências observadas na planilha;

Intervalo_esperado: posição das freqüências esperadas na planilha;



Observe que a planilha irá fornecer o p–valor = 0,022 que sendo menor que o nível de significância (0,05) nos leva à rejeição da hipótese nula.

Ampliando seus conhecimentos

Mineração de dados

(GONÇALVES, 2001)

Mineração de dados, ou *data mining*, é definida como uma etapa na descoberta do conhecimento em bancos de dados que consiste no processo de analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis. Para isso são utilizadas técnicas que envolvem métodos estatísticos que visam descobrir padrões e regularidades entre os dados pesquisados.

Em um mundo globalizado, sem fronteiras geográficas, onde as empresas competem mundialmente, a informação torna-se um fator crucial na busca pela competitividade. O fato de uma empresa dispor de certas informações possibilita-lhe aumentar o valor agregado de seu produto ou reduzir seus custos em relação àquelas que não possuem o mesmo tipo de informação. As informações e o conhecimento compõem um recurso estratégico essencial para o sucesso da adaptação da empresa em um ambiente de concorrência. Toda empresa tem informações que proporcionam sustentação para suas decisões, entretanto apenas algumas conseguem otimizar o seu processo decisório e aquelas que estão nesse estágio evolutivo seguramente possuem vantagem empresarial.

As ferramentas de mineração de dados, por definição, devem trabalhar com grandes bases de dados e retornar, como resultado, conhecimento novo e relevante; porém devemos ser céticos quanto a essa afirmação, pois esse tipo de ferramenta irá criar inúmeras relações e equações, o que pode tornar impossível o processamento desses dados.

A grande promessa da mineração de dados resume-se na afirmação de que ela 'vasculha' grandes bases de dados em busca de padrões escondidos, que extrai informações desconhecidas e relevantes e as utiliza para tomar decisões críticas de negócios. Outra promessa em relação a essa tecnologia de informação diz respeito à forma como elas exploram as inter-relações entre os dados. As ferramentas de análise disponíveis dispõem de um método basea-

do na verificação, isto é, o usuário constrói hipóteses sobre inter-relações específicas e então verifica ou refuta essas hipóteses por meio do sistema. Esse modelo torna-se dependente da intuição e habilidade do analista em propor hipóteses interessantes, em manipular a complexidade do espaço de atributos e em refinar a análise, baseado nos resultados de consultas potencialmente complexas ao banco de dados. Já o processo de mineração de dados, para o autor, seria responsável pela geração de hipóteses, garantindo mais rapidez, acurácia e completude dos resultados.

A cada ano, companhias acumulam mais e mais dados em seus bancos de dados. Esses dados muitas vezes são mantidos mesmo depois de esgotados seus prazos legais de existência, como no caso de notas fiscais. Com o passar do tempo, esse volume de dados passa a armazenar internamente o histórico das atividades da organização. Como conseqüência, esses bancos de dados passam a conter verdadeiros 'tesouros' de informação sobre vários procedimentos dessas companhias. Toda essa informação pode ser usada para melhorar os procedimentos da empresa, permitindo que ela detecte tendências e características disfarçadas e reaja rapidamente a um evento que ainda pode estar por vir. No entanto, apesar do enorme valor desses dados, a maioria das organizações é incapaz de aproveitar totalmente o que está armazenado em seus arquivos.

Essa informação está implícita, escondida sob uma montanha de dados, e não pode ser descoberta utilizando-se sistemas de gerenciamento de banco de dados convencionais. A quantidade de informação armazenada está explodindo e ultrapassa a habilidade técnica e a capacidade humana na sua interpretação.

Por isso, diversas ferramentas têm sido usadas para examinar os dados que as empresas possuem, no entanto, a maioria dos analistas tem reconhecido que existem padrões, relacionamentos e regras escondidos nesses dados, os quais não podem ser encontrados por meio da utilização de métodos tradicionais. A resposta é usar softwares de mineração de dados que utilizam algoritmos matemáticos avançados para examinar grandes volumes de dados detalhados.

A necessidade de transformar a 'montanha' de dados armazenados em informações significativas é óbvia, entretanto, sua análise ainda é demorada, dispendiosa, pouco automatizada e sujeita a erros, mal entendidos e falta de precisão. A automatização dos processos de análise de dados, com a utilização de softwares ligados diretamente à massa de informações, tornou-se uma necessidade. Esse motivo deve ser o responsável pelo crescimento do mercado de tecnologias de informação.

Atividades de aplicação

1. Um experimento foi realizado em 115 propriedades para verificar a eficácia de um novo adubo para plantações de milho. As produções médias das propriedades com o novo adubo encontram-se tabuladas abaixo. Compare com as produções médias garantidas pelo fabricante nas especificações técnicas do produto. Considere α = 0,05.

Classes (sacas/hectare)	f_i	e_i
2 700 |— 3 000	13	12
3 000 |— 3 300	18	20
3 300 |— 3 600	24	25
3 600 |— 3 900	32	25
3 900 |— 4 200	17	20
4 200 |— 4 500	11	13
Total	**115**	**115**

2. Em um exame a que se submeteram 117 estudantes de escolas públicas, a nota média foi 74,5 e o desvio padrão 8. Em uma escola particular, em que 200 estudantes foram submetidos a esse mesmo exame, a nota média foi de 75,9 com desvio padrão 10. A escola particular apresenta um melhor rendimento no exame? Considere α = 0,05.

3. Um médico-cientista imagina ter inventado uma droga revolucionária que baixa a febre em 1 minuto. Quinze voluntários foram selecionados (pacientes de uma clínica, com febre acima de 37°c) e os resultados foram os seguintes (em graus Celsius):

Paciente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Diferença*	1	0	3	4	3	2	1	1	4	1	0	0	2	3	3

* diferença de temperatura: o quanto a temperatura baixou em 1 minuto.

A droga inventada pelo médico é verdadeiramente eficiente?

4. Um criador verificou em uma amostra do seu rebanho (500 cabeças) 50 animais com verminose. Em seguida, avaliou outras 100 cabeças de

gado, mas antes solicitou ao veterinário uma solução para o problema. O veterinário alterou a dieta dos animais e acredita que a doença diminuiu de intensidade. Um exame nesse grupo de 100 cabeças do rebanho, escolhidas ao acaso, indicou 4 delas com verminose. Ao nível de significância de 1%, há indícios de que a proporção é menor?

5. Queremos comparar três hospitais, com relação à satisfação demonstrada por pacientes quanto ao atendimento durante o período de internação. Para tanto, foram selecionados, aleatoriamente, pacientes com grau de enfermidade semelhante. Cada paciente preencheu um questionário e as respostas geraram índices variando de 0 a 100, indicando o grau de satisfação. Os resultados foram:

	Hospital		
Pacientes	A	B	C
1	93	60	70
2	86	58	75
3	85	47	77
4	90	62	72
5	91	58	78
6	82	61	78
7	88	63	70
8	86	64	71
9	87	68	68
10	85	58	73
11		57	74
12		67	80
13		61	68
14		56	
15		58	

Baseando-se nos dados apresentados, teste se as médias populacionais são iguais. Qual sua conclusão? Use $\alpha = 0,05$.

Análise de Correlação e Medidas de Associação

Introdução

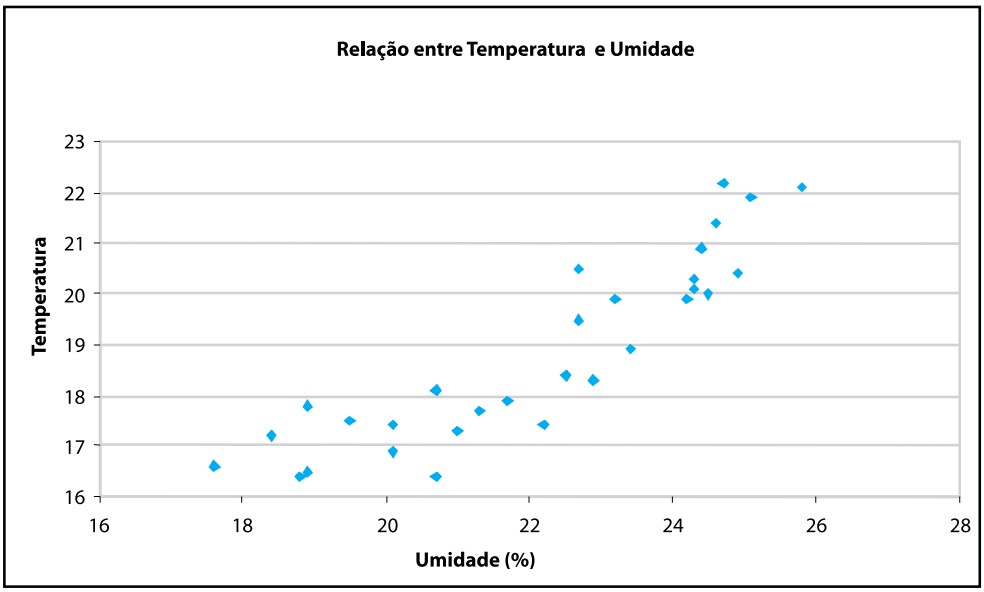
Muitas vezes, precisamos avaliar o grau de relacionamento entre duas ou mais variáveis. É possível descobrir, com precisão, o quanto uma variável interfere no resultado de outra. As técnicas associadas à Análise de Correlação representam uma ferramenta fundamental de aplicação nas Ciências Sociais e do comportamento, da Engenharia e das Ciências Naturais. A importância de se conhecer os diferentes métodos e suas suposições de aplicação é exatamente pelo cuidado que se deve ter para não se utilizar uma técnica inadequada. Existem diversos critérios de avaliação dessa relação, alguns próprios para variáveis que seguem uma distribuição normal e outros para variáveis que não seguem uma distribuição teórica conhecida. É comum a utilização do Coeficiente de Correlação de Pearson. No entanto, existem situações em que o relacionamento entre duas variáveis não é linear, ou uma delas não é contínua ou as observações não são selecionadas aleatoriamente. Nesses casos, outras alternativas de coeficientes devem ser aplicadas. Entre as diversas alternativas, veremos aqui algumas das mais importantes: Coeficiente de Spearman e Coeficiente de Contingência.

Segundo o dicionário Aurélio, *correlação* significa *relação mútua entre dois termos*, qualidade de correlativo, correspondência. Correlacionar, significa estabelecer relação ou correlação entre; ter correlação. Enquanto que a palavra *regressão* significa *ato ou efeito de regressar*, de voltar, retorno, regresso; dependência funcional entre duas ou mais variáveis aleatórias. A palavra regredir significa ir em marcha regressiva, retroceder.

Mas, onde e como surgiram os termos correlação e regressão? Foi Francis Galton (1822-1911), primo de Charles Darwin, quem usou pela primeira vez esses termos, cujo trabalho influenciou a Estatística e a Psicologia. Galton publicou o livro *Gênio Hereditário*, em 1869, no qual aplicou conceitos estatísticos a problemas da hereditariedade. O primeiro relato em que Galton usou o termo "co-relações" foi em 1888.

Diagramas de Dispersão

Um dos métodos mais usados para a investigação de pares de dados é a utilização de diagramas de dispersão cartesianos (ou seja, os conhecidos diagramas x-y). Geometricamente, um diagrama de dispersão é simplesmente uma coleção de pontos num plano cujas duas coordenadas cartesianas são os valores de cada membro do par de dados. E para quê fazemos um diagrama de dispersão? Este é o melhor método de examinar os dados no que se refere à ocorrência de tendências (lineares ou não), agrupamentos de uma ou mais variáveis, mudanças de espalhamento de uma variável em relação à outra e verificar a ocorrência dos valores discrepantes. Observe o exemplo a seguir:



Podemos notar pela análise da figura acima, a relação linear entre as duas variáveis. Os coeficientes apresentados a seguir nos auxiliam na quantificação do grau de relacionamento entre as variáveis de interesse.

A Covariância e o Coeficiente de Correlação de Pearson

Quando estudamos a relação entre duas variáveis X e Y, devemos primeiramente compreender o conceito de covariância. Se a variância é uma estatística por meio da qual chegamos ao desvio padrão que é uma medida de dispersão, da mesma maneira a covariância é uma estatística pela qual che-

gamos ao coeficiente de correlação que mede o grau de associação "linear" entre duas variáveis aleatórias X e Y.

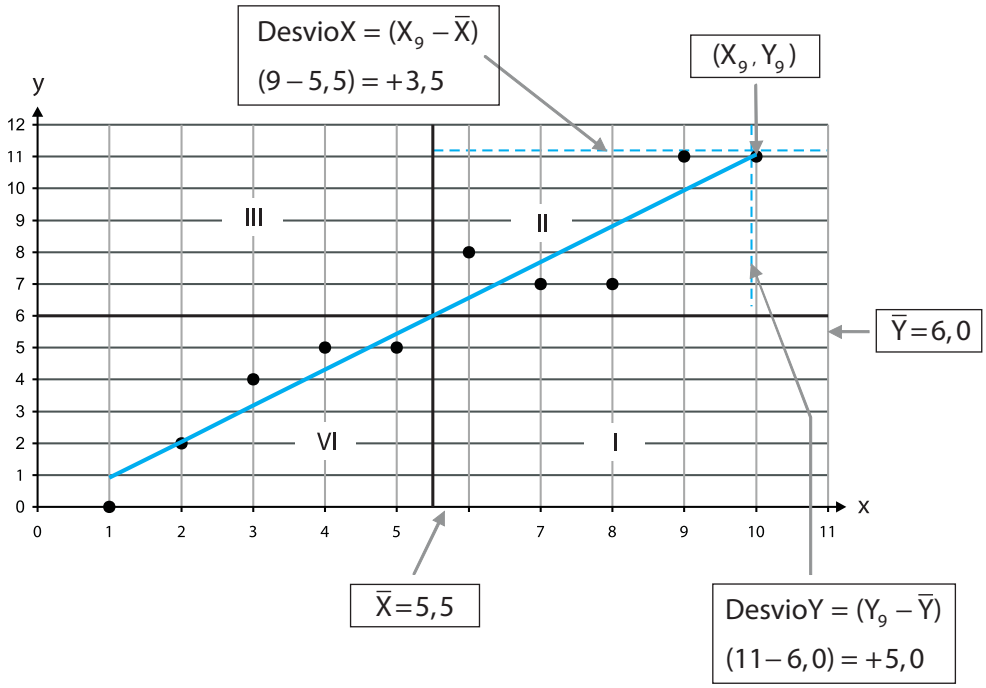
Observe o exemplo abaixo. Sejam X e Y duas variáveis aleatórias quaisquer, que tomam os seguintes valores:

Tabela 1. Cálculo do Coeficiente de Correlação de Pearson

X	Y	DesvioX $(X_i - \bar{X})^2$	DesvioY $(Y_i - \bar{Y})$	D X D Y $(X_i - \bar{X}).(Y_i - Y)$	Desvio X^2 $(X_i - \bar{X})^2$	Desvio Y^2 $(Y_i - \bar{Y})^2$	PRE_1 Y=a+bX
1	0	–4,50	–6,00	27,00	20,25	36,00	0,92727
2	2	–3,50	–4,00	14,00	12,25	16,00	2,05455
3	4	–2,50	–2,00	5,00	6,25	4,00	3,18182
4	5	–1,50	–1,00	1,50	2,25	1,00	4,30909
5	5	–0,50	–1,00	0,50	0,25	1,00	5,43636
6	8	0,50	2,00	1,00	0,25	4,00	6,56364
7	7	1,50	1,00	1,50	2,25	1,00	7,69091
8	7	2,50	1,00	2,50	6,25	1,00	8,81818
9	11	3,50	5,00	17,50	12,25	25,00	9,94545
10	11	4,50	5,00	22,50	20,25	25,00	11,07273
55	**60**	**0**	**0**	**93,00**	**82,50**	**114,00**	**60,0000**

Na tabela anterior está uma ilustração dos cálculos dos componentes da covariância e correlação.

A figura a seguir mostra a relação entre as duas variáveis X e Y, bem como a linha ajustada a esses valores pelo método de mínimos quadrados. Observe que a média de X é 5,5 e a média de Y é 6,0, e que elas estão formadas pelas linhas paralelas ao eixo Y e ao eixo X respectivamente. Vejamos agora o que significa os desvios de cada ponto em relação à média. Observe que cada ponto está formado pelo par ordenado (X_i,Y_i), onde X_i indica o valor da variável X e Y_i o valor da variável Y naquele ponto.



Tome, agora, por exemplo,

$$\text{DesvioX} = (X_9 - \bar{X}) = (9 - 5{,}5) = +3{,}5 \text{ e DesvioY} = (Y_9 - Y) = (11 - 6{,}0) = +5{,}0$$

O produto dos desvios:

$$\text{DesviosX . DesvioY} = (X_9 - \bar{X}).(Y_9 - \bar{Y}) = (9 - 5{,}5).(11 - 6{,}0) = (+3{,}5).(+5{,}0) = 17{,}5$$

Se calcularmos esses produtos para todos os valores de X e Y e somarmos temos o numerador da covariância de X e Y:

$$C(X,Y) = \frac{\sum (X_i - \bar{X}).(Y_i - \bar{Y})}{n} = \frac{93}{10} = 9{,}3 \qquad \textbf{(1)}$$

Logo, covariância significa co-variação, como as duas variáveis variam de forma conjunta. Agora, vejamos o que acontece se os pontos estivessem no quadrante I. Neste caso, os desvios de X seriam todos positivos, enquanto que os desvios de Y seriam todos negativos, logo, os produtos tomam valores negativos. O mesmo vai acontecer com os pontos do quadrante III, nele os desvios de X tomam valores negativos e os desvios de Y, valores positivos, logo, os produtos tomam valores negativos. Assim, se a maioria dos pontos caem nos quadrantes I e III, a covariância toma valores negativos, indicando que essas duas

variáveis se relacionam de forma negativa ou inversa, ou seja, quando uma cresce a outra diminui e vice-versa.

Quando os pontos se distribuem nos quatro quadrantes, haverá valores positivos e negativos, logo a soma tende para zero, e nesse caso, afirmamos que não existe relação linear entre essas variáveis. Observamos que esta estatística tende para zero, mesmo havendo uma relação que não seja linear, por exemplo se os dados tivessem o formato de uma parábola, ou relação quadrática.

Apesar de a covariância ser uma estatística adequada para medir relação linear entre duas variáveis, ela não é adequada para comparar graus de relação entre variáveis, dado que ela está influenciada pelas unidades de medida de cada variável, que pode ser metros, quilômetro, quilogramas, centímetros etc. Para evitar a influência da ordem de grandeza e unidades de cada variável, dividimos a covariância pelo desvio padrão de X e de Y, dando origem ao coeficiente de correlação de Pearson:

Notação:

Coeficiente de correlação amostral: **r**

Coeficiente de correlação populacional: **ρ**

$$r = \frac{C(X,Y)}{S_Y . S_X} \quad \textbf{(2)}$$

$$r = \frac{9,3}{2,8723 \,.\, 3,3764} = 0,95896$$

Onde: $S^2_x = 82,5 / 10 = 8,25 \rightarrow S_x = 2,8723$

$S_y^2 = 114,0 / 10 = 11,4 \rightarrow S_y = 3,3764$

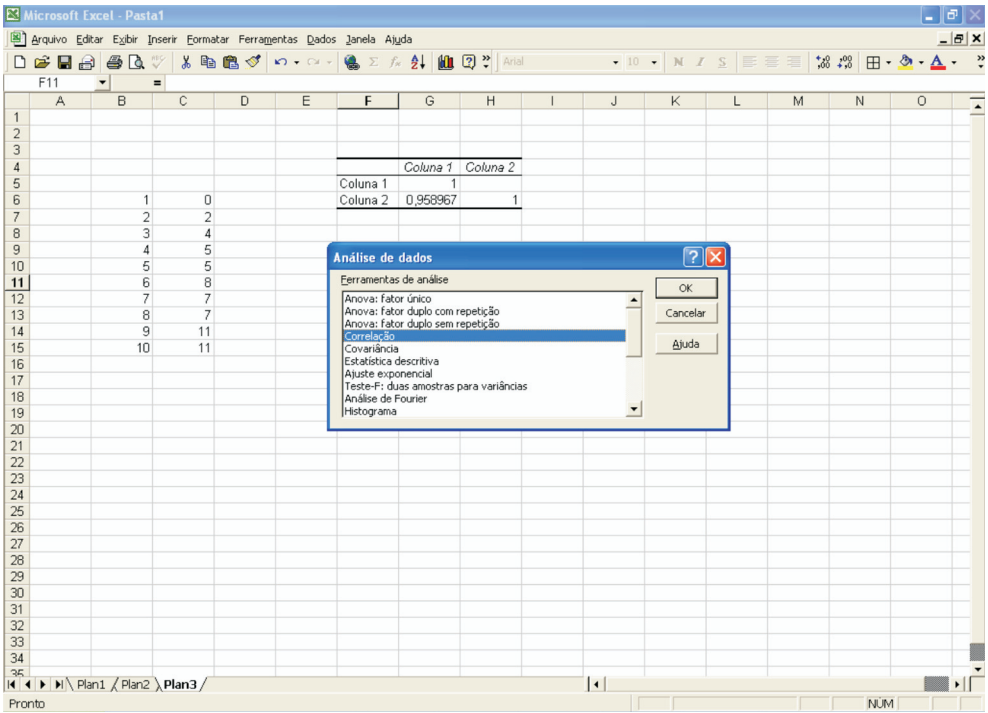
Como o coeficiente de correlação está isento de unidades e da ordem de grandeza das variáveis, este toma valores entre –1 e 1.

Relação positiva ➔ **r** tomará o valor 1 quando a relação é perfeita.

Relação negativa ➔ **r** tomará o valor –1 quando a relação é perfeita.

Relação difusa ou não linear ➔ **r** será igual a 0.

No *Excel*, usando a opção Correlação em "Análise de dados", obtemos:



O coeficiente de Determinação

Outro coeficiente amplamente utilizado para mensurar o grau de correlação entre duas variáveis é o *coeficiente de determinação*. É definido elevando o valor do coeficiente de Pearson ao quadrado e denotado por $\mathbf{r^2}$. Pode ser interpretado como a proporção da variação de Y que é explicada pela variável X (e vice versa).

Muito embora o coeficiente de determinação seja relativamente fácil de interpretar, ele não pode ser testado estatisticamente. Contudo, a raiz quadrada do coeficiente de determinação, que é o coeficiente de correlação (r), pode ser testada estatisticamente, pois está associada a uma estatística de teste que é distribuída segundo uma distribuição t de Student, quando a correlação populacional $\rho = 0$.

O coeficiente de correlação para dados populacionais é:

População: $\rho = \sqrt{\rho^2}$

O coeficiente de correlação para dados amostrais é:

Amostra: $r = \sqrt{r^2}$

Significância do coeficiente de correlação

Para comprovarmos se o coeficiente de correlação é significativo, devemos realizar o seguinte teste de hipóteses:

Hipóteses:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

A estatística de teste é $t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

com n-2 graus de liberdade na tabela t de Student. Caso o valor de t_c seja superior ao valor crítico de t, devemos rejeitar a hipótese nula. Se a hipótese nula, ao nível de significância α, for rejeitada podemos concluir que efetivamente existe uma relação significativa entre as variáveis.

Exemplo 1: Para estudar a poluição de um rio, um cientista mediu a concentração de um determinado composto orgânico (Y) e a precipitação pluviométrica na semana anterior (X):

X	Y
0,91	0,10
1,33	1,10
4,19	3,40
2,68	2,10
1,86	2,60
1,17	1,00

Existe alguma relação entre o nível de poluição e a precipitação pluviométrica? Teste sua significância, ao nível de 5%.

Calculando a média de X e de Y temos $\overline{X} = 2{,}023$ e $\overline{Y} = 1{,}717$.

Calculando a covariância entre X e Y pela expressão (1),

$$C(X,Y) = \frac{(0{,}91-2{,}023).(0{,}10-1{,}717)+(1{,}33-2{,}023).(1{,}10-1{,}717)+...+(1{,}17-2{,}023).(1{,}00-1{,}717)}{6}$$

$$C(X,Y) = 1{,}0989$$

Calculando os desvios padrões de X e Y temos: $S_x = 1,125$ e $S_y = 1,10$

E assim, pela expressão (2),

$$r = \frac{C(X,Y)}{S_y.S_x} = \frac{1,0989}{1,125.1,1} = 0,888$$

Testando a significância do coeficiente,

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,888\sqrt{6-2}}{\sqrt{1-(0,888)^2}} = 3,86$$

O valor crítico de t para n– 2 = 4 graus de liberdade e 5% de nível de significância é 2,78. Note que o teste de significância do coeficiente será sempre bilateral.

Como o valor calculado de t é superior ao valor crítico, podemos concluir que existem evidências suficientes para afirmar que o composto orgânico (Y) e a precipitação pluviométrica (X) estejam correlacionados.

Exemplo 2: Procurando quantificar os efeitos da escassez de sono sobre a capacidade de resolução de problemas simples, um agente tomou ao acaso 10 sujeitos e os submeteu a experimentação. Deixou-os sem dormir por diferentes números de horas, após o que solicitou que os mesmos resolvessem os itens "contas de adicionar" de um teste. Obteve, assim, os seguintes dados:

Nº de erros - Y	Horas sem dormir - X
8	8
6	8
6	12
10	12
8	16
14	16
14	20
12	20
16	24
12	24

Calcule o coeficiente de correlação linear de Pearson e teste a sua significância ao nível de 1%.

Calculando a média de X e de Y temos $\overline{X} = 16$ e $\overline{Y} = 10,6$.

Calculando a covariância entre X e Y pela expressão (1),

$$C(X,Y) = \frac{(8-16).(8-10,6)+(8-16).(6-10,6)+...+(24-16).(12-10,6)}{10} = 15,2$$

Calculando os desvios padrões de X e Y temos:

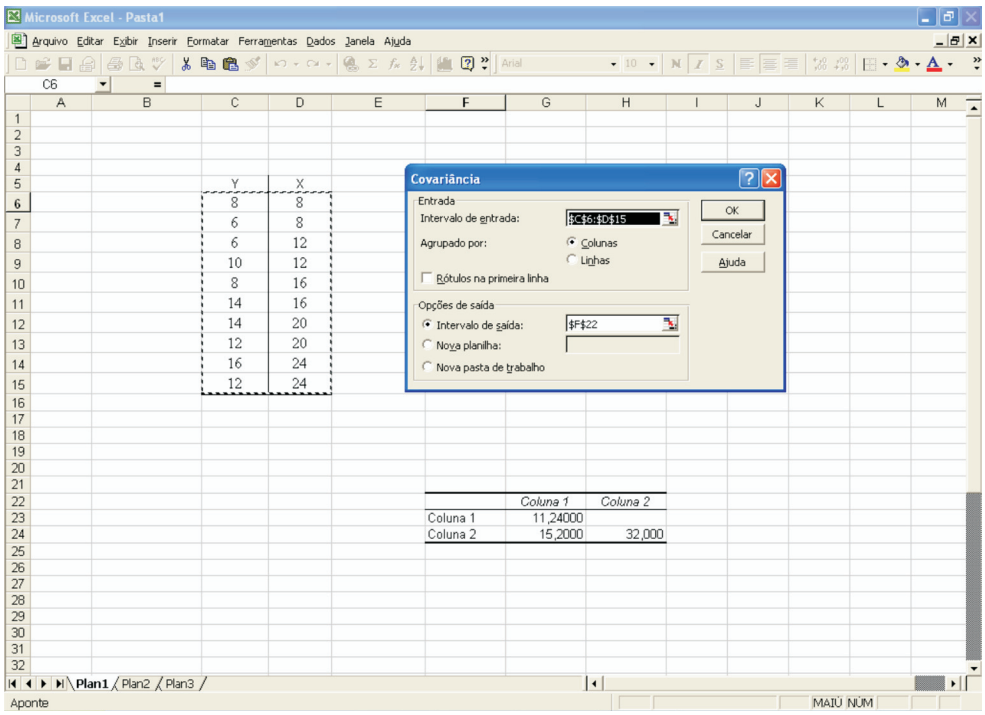
$S_x = 5,656854$ e $S_y = 3,352611$

E assim, pela expressão (2),

$$r = \frac{C(X,Y)}{S_y.S_x} = \frac{15,2}{5,656854.3,352611} = 0,801467$$

Observação: procure sempre usar o maior número de casas decimais possível.

Usando a planilha *Excel* poderemos também obter uma matriz de covariância, que nos fornece a covariância entre X e Y além da variância de X e de Y.



Y	X
8	8
6	8
6	12
10	12
8	16
14	16
14	20
12	20
16	24
12	24

Agora testando a significância do coeficiente,

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}801467\sqrt{10-2}}{\sqrt{1-(0{,}801467)^2}} = 3{,}79$$

O valor crítico de t para n–2 = 8 graus de liberdade e 1% de nível de significância é 3,355 (bilateral).

Como o valor calculado de t é superior ao valor crítico, podemos concluir que existem evidências suficientes para afirmar que o número de horas sem dormir (X) influencia significativamente o número de erros (Y).

Medidas de Associação

Freqüentemente, estamos interessados em verificar a existência de associação entre dois conjuntos de escores e também o grau desta associação. No caso paramétrico, a medida usual é o coeficiente de correlação **r** de Pearson que exige mensuração dos escores no mínimo ao nível intervalar. Ainda, se estivermos interessados em comprovar a significância de um valor observado de **r** de Pearson deveremos supor que os escores provenham de uma distribuição normal. Quando estas suposições não são atendidas, podemos utilizar um dos coeficientes de correlação não-paramétricos e suas respectivas provas de significância.

Coeficiente de Contingência C

Este coeficiente mede a associação entre dois conjuntos de atributos quando um ou ambos os conjuntos são medidos em escala nominal.

Considere uma tabela de contingência k x r, que representa as freqüências cruzadas dos escores A (divididos em k categorias) e escores B (divididos em r categorias). O grau de associação entre dois conjuntos de atributos é calculado por:

$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$ onde χ^2 é a estatística Qui-quadrado.

O p-valor associado ao valor da estatística Qui-quadrado com (r-1) x (k-1) graus de liberdade é a prova de significância do coeficiente de contingência **C**.

O coeficiente **C** se caracteriza por assumir valor zero quando há inexistência de associação porém nunca será igual à 1. O limite superior do coeficiente é dado por $\sqrt{\frac{k-1}{k}}$ (quando k = r). Note que para calcular o coeficiente **C**, a tabela de contingência deve satisfazer as restrições do teste Qui-quadrado.

Exemplo: Estudantes de escolas particulares e de escolas públicas selecionados aleatoriamente foram submetidos a testes padronizados de conhecimento, e produziram os resultados abaixo. Verifique o grau de associação entre as variáveis mensuradas e teste a significância ao nível de 5%.

	Escores			
Escola	**0 – 275**	**276 – 350**	**351 – 425**	**426 – 500**
Particular	6	14	17	9
Pública	30	32	17	3

Queremos aqui verificar o grau de associação entre as variáveis "Escola" e "Escore de conhecimento". A variável Escola é mensurada em nível nominal, o que inviabiliza a utilização do coeficiente **r** de Pearson.

Obtendo então o coeficiente de Contingência, necessitamos inicialmente calcular o valor da estatística χ^2:

Freq. Obs.	6	14	17	9
	30	32	17	3
Freq. Esp.	12,94	16,53	12,22	4,31
	23,06	29,47	21,78	7,69

$$\chi^2 = \frac{(6-12,94)^2}{12,94} + \frac{(14-16,53)^2}{16,53} + \ldots + \frac{(3-7,69)^2}{7,69} = 17,28$$

O coeficiente de contingência é:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{17,28}{128+17,28}} = 0,345$$

Para testar a significância do coeficiente, precisamos verificar o valor crítico de χ^2 considerando α=0,05 e (r–1) x (k–1) = 3 graus de liberdade. Esse valor é igual a 7,81. Comparando com o valor calculado de 17,28, podemos admitir a existência de associação significativa entre a escola e o escore de

conhecimento. Analisando atentamente, poderíamos acrescentar que o fato de um estudante pertencer a uma escola particular faz com que ele obtenha um escore de conhecimento mais alto.

Coeficiente de correlação de Spearman

É uma medida de associação que exige que ambas as variáveis se apresentem em escala de mensuração pelo menos ordinal. Basicamente, equivale ao coeficiente de correlação de Pearson aplicado a dados ordenados. Assim,

$$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}} = r_s$$

ou seja, o coeficiente de correlação de Spearman se utiliza da expressão do coeficiente de Pearson, porém calculado com postos. Esta expressão equivale à

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$ onde $d_i = x_i - y_i$ a diferença de postos dos escores X e Y.

Para verificar a significância do valor observado de r_s, podemos usar a expressão de t de Student

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$ onde **t** tem n–2 graus de liberdade.

Exemplo: As notas obtidas por 10 estudantes de Administração e o seu QI (quociente de inteligência) são apresentadas no quadro abaixo:

Notas	8	9,5	10	9,1	6,5	9	9,5	5,2	9,1	9,3
QI	127	149	150	135	122	129	142	100	136	139

Utilize o coeficiente de Spearman para verificar se as variáveis estão associadas e qual o seu grau de associação.

Inicialmente, ordenamos os valores originais, transformando-os em postos. Aqui então substituímos os valores originais pelos seus respectivos postos, ou seja, o menor valor da variável em questão será substituído pelo valor 1 e assim por diante. Em seguida, calculamos as diferenças de postos:

Notas	3	8,5	10	5,5	2	4	8,5	1	5,5	7
QI	3	9	10	5	2	4	8	1	6	7
di	0	–0,5	0	0,5	0	0	0,5	0	–0,5	0
(di)2	0	0,25	0	0,25	0	0	0,25	0	0,25	0

Calculando o coeficiente:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6.\left(0^2 + 0{,}25^2 + \ldots + 0^2\right)}{10^3 - 10} = 1 - \frac{6.0{,}25}{990} = 0{,}998$$

Verificando a significância estatística do coeficiente:

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} = 0{,}998\sqrt{\frac{8}{1-(0{,}998)^2}} = 0{,}998\sqrt{\frac{8}{0{,}004}} = 44{,}63$$

O valor crítico da estatística t de Student é obtido definindo-se n–2 = 8 graus de liberdade e o nível de significância, que admitiremos igual a 1%. Este valor é igual a 3,36. Mais uma vez temos aqui um teste bilateral pois estamos verificando se o coeficiente é diferente de zero.

Assim, podemos comprovar que o coeficiente de associação é altamente significativo, ou seja, existem fortes indícios que apontam para notas altas obtidas por aqueles que possuem maiores quocientes de inteligência.

Ampliando seus conhecimentos

Teste de Kappa

(LANDIS; KOCH, 1977)

O Teste de Kappa é uma medida de concordância interobservador e mede o grau de concordância, além do que seria esperado tão-somente pelo acaso.

Para descrevermos se há ou não concordância entre dois ou mais avaliadores, ou entre dois métodos de classificação, utilizamos a medida Kappa que é baseada no número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre os avaliadores. Esta medida de concordância assume valor máximo igual a 1, que representa total concordância ou, ainda,

pode assumir valores próximos e até abaixo de 0, os quais indicam nenhuma concordância.

O coeficiente Kappa é calculado a partir da seguinte fórmula:

$$\text{Kappa} = \frac{P_0 - P_E}{1 - P_E}$$

$$\text{onde} P_0 = \frac{\textit{número de concordâncias}}{\textit{número de concordâncias} + \textit{número de discordâncias}}$$

$$\text{e } P_E = \sum_{i=1}^{n} (p_{i1}.p_{i2}) \text{ sendo que:}$$

- n é o número de categorias;
- i é o índice da categoria (que vale de 1 a n);
- p_{i1} é a proporção de ocorrência da categoria i para o avaliador 1;
- p_{i2} é a proporção de ocorrência da categoria i para o avaliador 2.

Para avaliar se a concordância é razoável, Landis, JR e Koch, GG (1977) sugerem a seguinte interpretação:

Valores obtidos de Kappa	Interpretação
<0	Nenhuma concordância
0 – 0,19	Concordância pobre
0,20 – 0,39	Concordância leve
0,40 – 0,59	Concordância moderada
0,60 – 0,79	Concordância substancial
0,80 – 1,00	Concordância quase perfeita

Fonte: Landis JR, Koch GG. The measurement of observer agreement for categorical data. **Biometrics** 1977; **33**: 159-174

Exemplo: Em certo órgão de financiamento, em cada edital aberto, se apresentam diversos pesquisadores que enviam projetos, solicitando recursos para desenvolvê-los. Estes projetos recebem uma avaliação, muitas vezes subjetiva, baseada na opinião de um consultor.

Considere a tabela a seguir, que resume as avaliações feitas por dois avaliadores a 30 projetos que concorrem ao financiamento. O interesse deste estudo é saber qual é a concordância entre estes dois profissionais e se há alguma classificação com concordância maior do que as demais.

		AVALIADOR 2			
		A	B	C	**Total**
AVALIADOR 1	A	14 (0,47)	1 (0,03)	1 (0,03)	**16 (0,53)**
	B	3 (0,10)	3 (0,10)	2 (0,07)	**8 (0,27)**
	C	0 (0,00)	1 (0,03)	5 (0,17)	**6 (0,20)**
	Total	**17 (0,57)**	**5 (0,16)**	**8 (0,27)**	**30 (1,00)**

* entre parênteses as proporções

Calculando o coeficiente Kappa:

$$P_0 = \frac{14+3+5}{30} = \frac{22}{30} = 0,7333$$

$$P_E = \sum_{i=1}^{n} (p_{i1}.p_{i2}) = (0,57 \,.\, 0,53) + (0,16 \,.\, 0,27) + (0,27 \,.\, 0,20) = 0,3021 + 0,0432$$

$$+ 0,054 = 0,3993$$

$$\text{Kappa} = \frac{0,733 - 0,3993}{1 - 0,3993} = 0,556$$

Note que a concordância geral pode ser considerada apenas moderada. Avaliando cada uma das três classificações, notamos que a concordância é alta quando os avaliadores atribuem o conceito A e o conceito C. No entanto, para atribuir o conceito B, um conceito intermediário, a concordância já não é tão satisfatória.

Atividades de aplicação

1. Foi tomada uma amostra aleatória de 10 carregamentos recentes feitos por caminhão de uma companhia, anotada a distância em quilômetros e o tempo de entrega. Os dados seguem abaixo:

Carregamento	1	2	3	4	5	6	7	8	9	10
Distância em Km (X)	825	215	1 070	550	480	920	1 350	325	670	1 215
Tempo de entrega em dias (Y)	3,5	1,0	4,0	2,0	1,0	3,0	4,5	1,5	3,0	5,0

a) Construa o diagrama de dispersão.

b) Calcule o coeficiente de correlação de Pearson para os dados desta amostra.

c) Calcule o coeficiente de determinação.

d) Verifique se o coeficiente de correlação é significativo (α=0,05).

2. Para uma amostra de n = 10 tomadores de empréstimos em uma companhia financeira, o coeficiente de correlação entre a renda familiar média e débitos a descoberto de curto prazo foi calculado r = 0,50.

Teste a hipótese de que não existe correlação entre as duas variáveis, usando um nível de significância de 5%.

3. Para avaliar a relação entre habilidade verbal e habilidade matemática, escores de 8 estudantes foram obtidos, gerando a tabela abaixo:

	Estudantes							
Escore	1	2	3	4	5	6	7	8
Matemática	80	50	36	58	72	60	56	68
Verbal	65	60	35	39	48	44	48	61

Calcule o coeficiente de correlação e teste sua significância.

4. Em um estudo conduzido com 10 pacientes, estes foram colocados sob uma dieta de baixas gorduras e altos carboidratos. Antes de iniciar a dieta, as medidas de colesterol e de triglicerídeos foram registradas para cada indivíduo .

a) Construa um gráfico de dispersão para esses dados.

b) Há alguma evidência de relação linear entre os níveis de colesterol e de triglicerídeos?

c) Calcule o coeficiente de correlação de Spearman e teste sua significância.

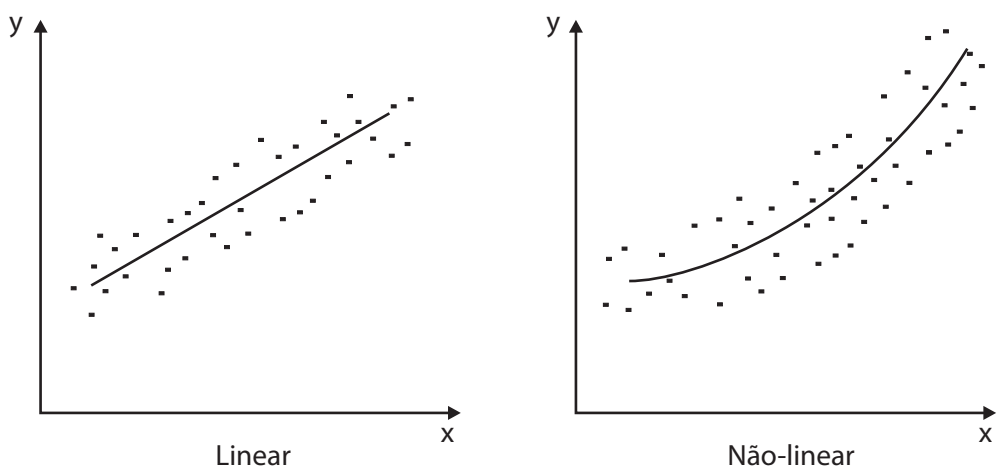
Paciente	Colesterol (mmol/l)	Triglicerídeos (mmol/l)
1	5,12	2,30
2	6,18	2,54
3	6,77	2,95
4	6,65	3,77
5	6,36	4,18
6	5,90	5,31
7	5,48	5,53
8	6,02	8,83
9	10,34	9,48
10	8,51	14,20

Análise de Regressão

Introdução

Os modelos de regressão são largamente utilizados em diversas áreas do conhecimento, tais como: computação, administração, engenharias, biologia, agronomia, saúde, sociologia etc. O principal objetivo desta técnica é obter uma equação que explique satisfatoriamente a relação entre uma variável resposta e uma ou mais variáveis explicativas, possibilitando fazer predição de valores da variável de interesse. Este relacionamento pode ser por uma equação linear ou uma função não-linear, conforme figura abaixo:

Figura 1: Formas lineares e não lineares de relação entre pares de variáveis



Regressão linear simples

Se uma relação linear é válida para sumarizar a dependência observada entre duas variáveis quantitativas, então a equação que descreve esta relação é dada por:

$$Y = a + b.X$$

Esta relação linear entre X e Y é determinística, ou seja, ela "afirma" que todos os pontos caem exatamente em cima da reta de regressão. No entanto este fato raramente ocorre, ou seja, os valores observados não caem todos

exatamente sobre esta linha reta. Existe uma diferença entre o valor observado e o valor fornecido pela equação. Esta diferença, denominada erro e representada por ε, é uma variável aleatória que quantifica a falha do modelo em ajustar-se aos dados exatamente. Tal erro pode ocorrer devido ao efeito, dentre outros, de variáveis não consideradas e de erros de medição. Incorporando esse erro à equação acima temos:

$$Y = a + b.X + \varepsilon$$

que é denominado modelo de regressão linear simples. *a* e *b* são os parâmetros do modelo.

A variável X, denominada variável regressora, explicativa ou independente, é considerada uma variável controlada pelo pesquisador e medida com erro desprezível. Já Y, denominada variável resposta ou dependente, é considerada uma variável aleatória, isto é, existe uma distribuição de probabilidade para Y em cada valor possível de X. É muito freqüente, na prática, encontrarmos situações em que Y tenha distribuição normal. Este é um dos principais pressupostos para aplicação desta técnica.

Exemplo 1: O preço de aluguel de automóveis de uma agência é definido pela seguinte equação: $Y = 8 + 0{,}15.X$, onde Y = Taxa de aluguel (R$); X = distância percorrida (km).

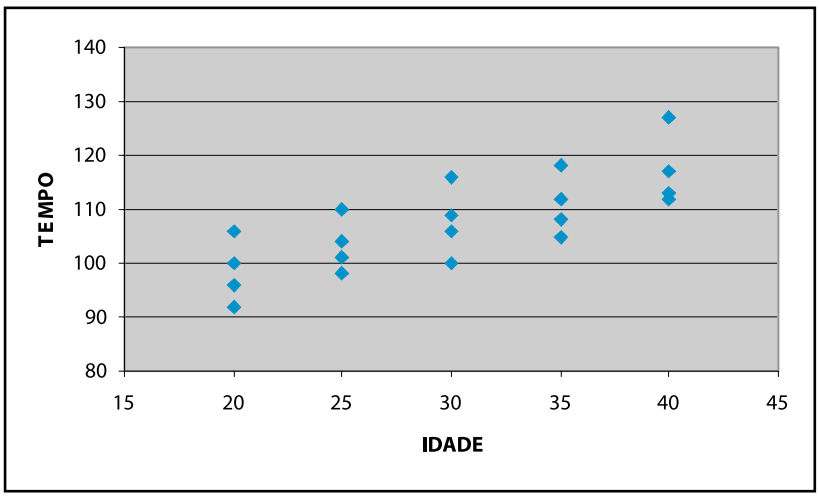
Assim, a taxa de aluguel inicia com o preço de R$ 8,00 e vai aumentando à medida que a distância percorrida aumenta. Assim, se fosse percorrida uma distância de 100 km, a taxa de aluguel seria de $8 + 0{,}15 \times 100 =$ R$ 23,00. No entanto, como essa equação foi obtida baseada em dados de automóveis de diversas marcas, certamente haverá uma variação no preço, por causa de diversos outros fatores. Assim, essa equação terá uma margem de erro, que é devida a esses inúmeros fatores que não foram controlados.

Exemplo 2: Um psicólogo investigando a relação entre o tempo que um indivíduo leva para reagir a um certo estímulo e sua idade obteve os seguintes resultados:

Tabela 1: Idade (em anos) e tempo de reação à um certo estímulo (em segundos)

Y - Tempo de reação (segundos)	X - Idade (em anos)
96	20
92	20
106	20
100	20
98	25
104	25
110	25
101	25
116	30
106	30
109	30
100	30
112	35
105	35
118	35
108	35
113	40
112	40
127	40
117	40

Figura 2: Diagrama de dispersão entre a idade (X) e o tempo de reação (Y)



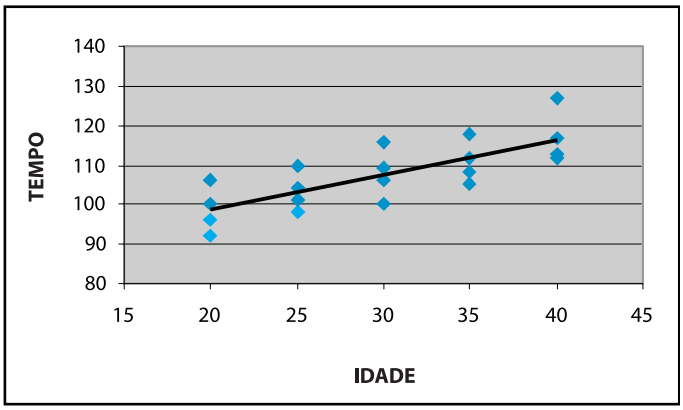
A partir da representação gráfica desses dados, mostrada na figura 2, é possível visualizar uma relação linear positiva entre a idade e o tempo de reação. O coeficiente de correlação de Pearson para esses dados resultou em r = 0,768, bem como seu respectivo teste de significância em t_{cal} = 5,09, que comparado ao valor tabelado $t_{tab,5\%}$ = 2,1 , fornece evidências de relação linear entre essas duas variáveis, ou seja, há evidências de considerável relação linear positiva entre idade e tempo de reação.

Podemos, então, usar um modelo de regressão linear simples para descrever essa relação. Para isso, é necessário estimar, com base na amostra observada, os parâmetros desconhecidos *a e b* deste modelo. O método de estimação denominado Mínimos Quadrados Ordinários (MQO) é freqüentemente utilizado em regressão linear, para esta finalidade, e será apresentado mais adiante.

Continuando a análise dos dados do exemplo, é possível obter o seguinte modelo de regressão linear simples ajustado:

Y = 80,5 + 0,9.X

Figura 3: Reta de regressão ajustada aos dados



Como a variação dos dados em X não inclui x = 0, não há interpretação prática do coeficiente *a* = 80,5. Por outro lado, *b* = 0,9 significa que a cada aumento de 1 ano na idade das pessoas, o tempo de reação médio (esperado) aumenta em 0,9 segundos.

Assim, se: X = 20 anos, teremos Y = 98,5 seg.

Para X = 21 anos, Y = 99,4 seg.

X = 22 anos, Y = 100,3 seg.

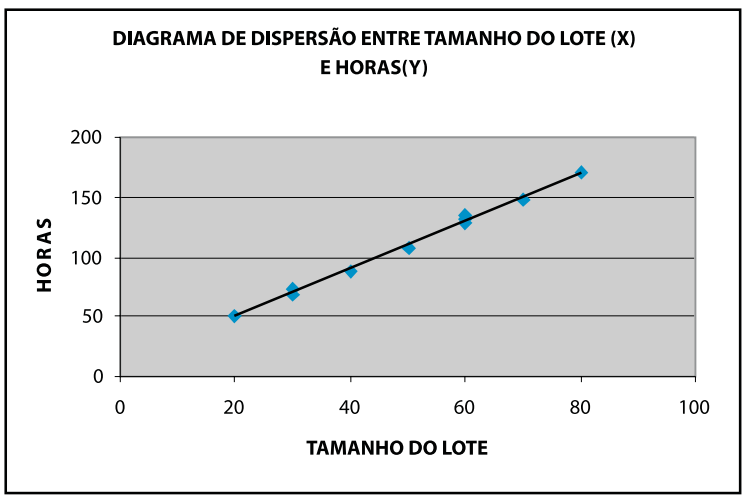
Dessa maneira, de ano para ano, o aumento no tempo de reação esperado é de 0,9 segundos.

Exemplo 3: Uma certa peça é manufaturada por uma companhia, uma vez por mês, em lotes, que variam de tamanho de acordo com as flutuações na demanda. A tabela abaixo contém dados sobre tamanho do lote e número de horas gastas na produção de 10 recentes lotes produzidos sob condições similares. Estes dados são apresentados graficamente na Figura 4, tomando-se horas-homem como variável *dependente* ou variável *resposta* (Y) e o tamanho do lote como variável *independente* ou *preditora* (X).

Tabela 2: Tamanho de lote e número de horas gastas na produção de cada lote.

Lote (i)	Horas (Y_i)	Tamanho do lote (X_i)
1	73	30
2	50	20
3	128	60
4	170	80
5	87	40
6	108	50
7	135	60
8	69	30
9	148	70
10	132	60

Figura 4: Relação estatística entre Y e X, referente aos dados da Tabela 2.



A Figura 4 sugere claramente que há uma relação linear positiva entre o tamanho do lote e o número de horas, de modo que, maiores lotes tendem a corresponder a maiores números de horas-homem consumidas. Porém, a relação não é perfeita, ou seja, há uma dispersão de pontos sugerindo que alguma variação no número de horas não é dependente do tamanho do lote. Por exemplo, dois lotes de 30 unidades (1 e 8) demandaram quantidades um pouco diferentes de horas. Na Figura 4, foi traçada uma linha (reta) de relacionamento descrevendo a relação estatística entre horas e tamanho do lote. Ela indica a tendência geral da variação em horas-homem quando há trocas no tamanho do lote.

Observa-se que grande parte dos pontos da figura não cai diretamente sobre a linha de relacionamento estatístico. A dispersão dos pontos em torno da linha de relacionamento representa a variação em horas que não é associada ao tamanho do lote, e que é usualmente considerada aleatória. Relações estatísticas são geralmente úteis, mesmo não tendo uma relação funcional exata.

Método dos mínimos quadrados ordinários (MQO)

Para estimar os parâmetros do modelo, é necessário um método de estimação. O método estatístico utilizado e recomendado pela sua precisão é o método dos mínimos quadrados que ajusta a melhor "equação" possível aos dados observados.

Com base nos *n* pares de observações (y_1, x_1), (y_2, x_2), ..., (y_n, x_n), o método de estimação por MQO consiste em escolher *a* e *b* de modo que a soma dos quadrados dos erros, ε_i (i=10 ,..., n), seja mínima.

Para minimizar esta soma, que é expressa por:

$$SQ = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{I-1}^{n} .(y_i - a - b.x_i)$$

devemos, inicialmente, diferenciar a expressão com respeito a "a" e "b" e, em seguida, igualar a zero as expressões resultantes. Feito isso, e após algumas operações algébricas, os estimadores resultantes são:

$$b = \frac{\sum x_i.y_i - n.\overline{y}.\overline{x}}{\sum x_i^2 - n.\overline{x}^2}$$

$$a = \overline{y} - b.\overline{x}$$

onde $\overline{Y}$ é a média amostral dos y_i's e $\overline{x}$ a média amostral dos x_i's.

Logo, $E(Y|x) = a + b.x$ é o modelo de regressão linear simples ajustado, em que $E(Y|x)$, denotado também $\hat{Y}$ por simplicidade, é o valor médio predito de Y para qualquer valor $X = x$ que esteja na variação observada de X.

No exemplo 2, as estimativas dos parâmetros resultaram em $a = 80,5$ e $b = 0,9$. Veja como esses valores foram obtidos:

$$\sum X_i = 2\,150 \qquad \sum Y_i = 600 \qquad n = 20 \qquad \sum X_i Y_i = 65\,400$$

$$\overline{X} = 30 \qquad \overline{Y} = 107,5 \qquad \sum X_i^2 = 19\,000$$

$$b = \frac{\sum x_i y_i - n.\overline{y}.\overline{x}}{\sum x_i^2 - n.\overline{x}^2} = \frac{65\,400 - 20.107,5.30}{19\,000 - 20.(30)^2} = \frac{900}{1\,000} = 0,9$$

$$a = \overline{y} - b.\overline{x} = 107,5 - 0,9.30 = 80,5$$

No exemplo 3, as estimativas dos parâmetros *a* e *b* são:

$$\sum X_i = 500 \qquad \sum Y_i = 1\,100 \qquad n = 10 \qquad \sum X_i Y_i = 61\,800$$

$$\overline{X} = 50 \qquad \overline{Y} = 110 \qquad \sum X_i^2 = 28\,400$$

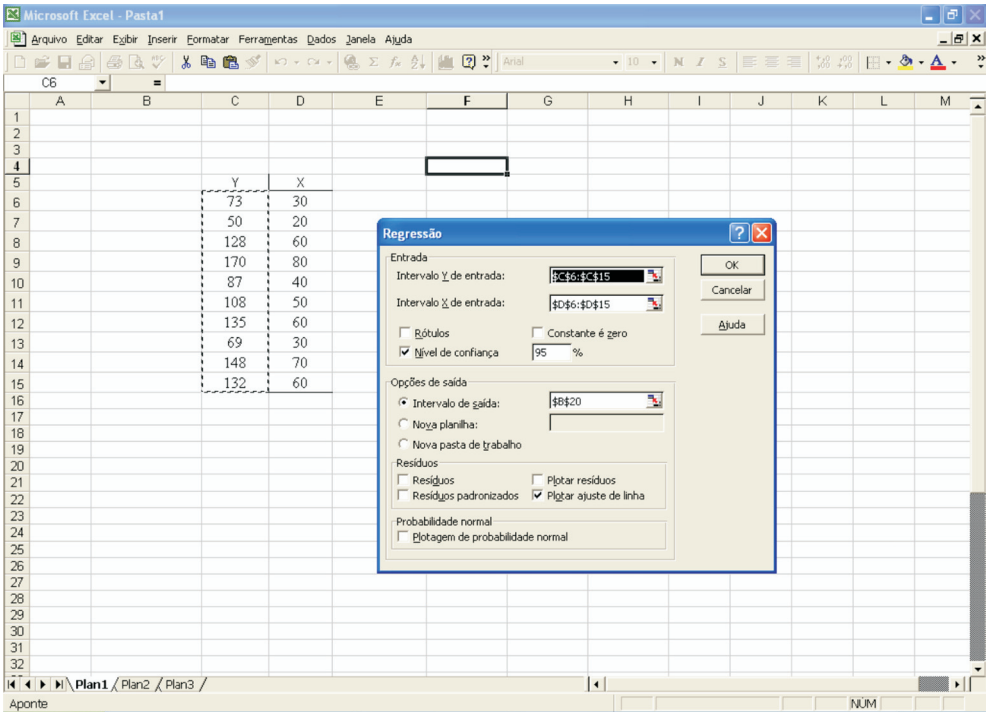
$$b = \frac{\sum x_i.y_i - n.\overline{y}.\overline{x}}{\sum x_i^2 - n.\overline{x}^2} = \frac{61\,800 - 10.110.50}{28\,400 - 10.(50)^2} = \frac{6\,800}{3\,400} = 2$$

Assim, a equação de regressão linear entre X e Y será dada por:

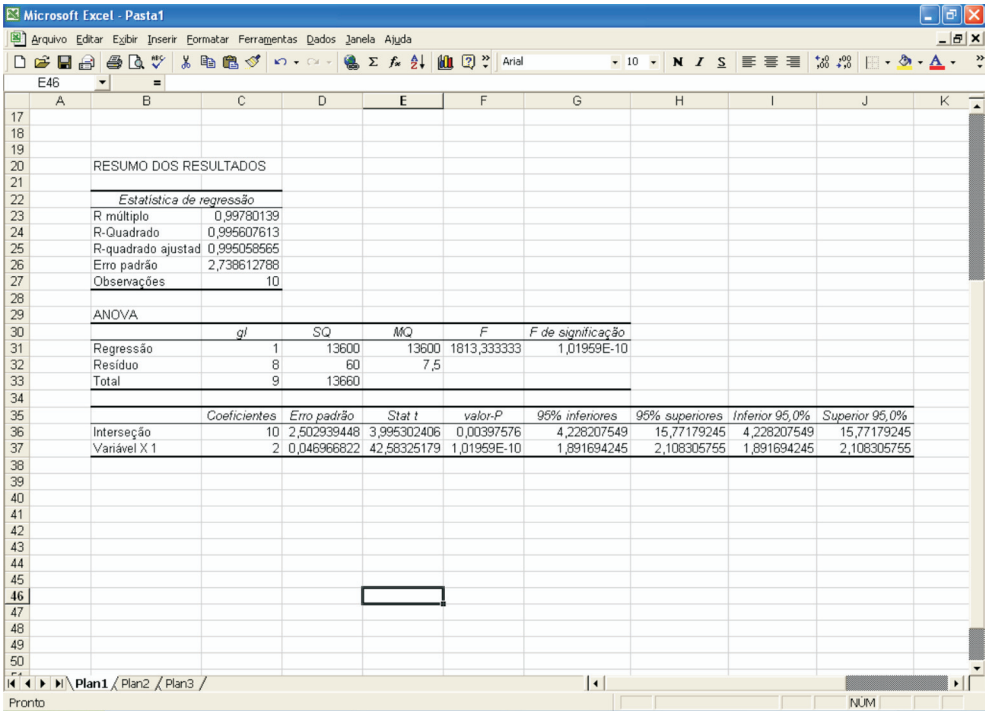
$$Y = 10 + 2.X + \varepsilon$$

Interpretando o modelo acima, poderemos observar que, aumentando o tamanho do lote em uma unidade, o número de horas gastas na produção será aumentado em 2 horas.

Obtendo a reta de regressão com ajuda da planilha *Excel*, teremos que selecionar a opção REGRESSÃO no módulo de Análise de dados (em ferramentas):



A saída fornecida pela planilha é a seguinte:



Observe que o *Excel* fornece, além dos coeficientes de correlação, a Anova da regressão para testar a sua significância e os coeficientes estimados com seus respectivos testes de significância.

Análise de Variância da Regressão

Para verificar a adequação do modelo aos dados, algumas técnicas podem ser utilizadas. A "análise de variância da Regressão" é uma das técnicas mais usadas. Assim, podemos analisar a adequação do modelo pela ANOVA da regressão a qual é geralmente apresentada como na tabela abaixo:

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	p-1	SQreg	SQreg/p-1	QMreg/QMres	
Resíduos	n-p	SQres	SQres/n-p		
Total	**n-1**	**SQtotal**	**Sqtotal/n-1**		

Onde:

- SQreg = soma dos quadrados devido à regressão:

$$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2$$

- SQres = soma dos quadrados devido aos erros:

$$SQres = SQtotal - Sqreg = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2$$

- SQtotal = soma dos quadrados totais:

$$SQtotal = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- p = número de variáveis do modelo
- n = numero de observações.

Caso o *p-valor* seja inferior ao nível de significância estabelecido, então consideramos a regressão como significativa.

Uma maneira auxiliar de medir o "ganho" relativo introduzido pelo modelo é usar o coeficiente de determinação o qual é definido por R^2 que é calculado por SQreg/SQtotal.

Para os exemplos 2 e 3, a tabela da Anova seria construída de seguinte forma:

Exemplo 2:

$$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2 = \sum_{i=1}^{n} (80,5 + 0,9x_i - 107,5)^2 = 810$$

Para obter a soma de quadrados acima, deveremos substituir em X_i todos os valores de idade da Tabela 1.

$$SQtotal = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - 107,5)^2 = 1\ 373$$

Para obter a soma de quadrados acima, deveremos substituir em Y_i todos os valores de tempo de reação da Tabela 1.

$$SQres = 1\ 373 - 810 = 563$$

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	1	810	810	25,90	< 0,01
Resíduos	18	563	31,27		
Total	**9**	**1 373**	**72,26**		

O que indica que a regressão entre X e Y é significativa. O modelo Y = 80,5 +0,9.X pode ser considerado adequado para realizar predições de Y. O coeficiente r^2 de determinação para esse modelo é de 0,59 o que representa um poder apenas razoável de explicação dos valores de tempo de reação pela idade. Muito provavelmente outras variáveis estejam influenciando o tempo de reação.

Exemplo 3:

$$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2 = \sum_{i=1}^{n} (10 + 2x_i - 110)^2 = 13\ 600$$

Para obter a soma de quadrados acima, deveremos substituir em X_i todos os valores do tamanho do lote da Tabela 2.

$$SQtotal = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - 107,5)^2 = 13\ 660$$

Para obter a soma de quadrados acima, deveremos substituir em Y_i todos os valores de números de horas gastas da Tabela 2.

$$SQres = 13\,660 - 13\,600 = 60$$

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	1	13 600	13 600	1 813,33	< 0,01
Resíduos	8	60	7,5		
Total	**9**	**13 660**	**1 517,78**		

O que indica que a regressão entre X e Y é significativa. O modelo Y = 10 + 2.X pode ser considerado de boa qualidade para realizar predições de Y. O coeficiente r^2 de determinação para esse modelo é de 0,996.

Erro padrão de estimação e intervalos de predição

O erro padrão da estimação é um desvio padrão condicional, na medida em que indica o desvio padrão da variável dependente Y, dado um valor específico da variável dependente X. O erro padrão baseado em dados amostrais é dado por:

$$\hat{\sigma}_u = \sqrt{\frac{\sum (y - \hat{Y})^2}{n-2}}$$

Para fins de cálculo, é mais conveniente uma versão alternativa da fórmula:

$$\hat{\sigma}_u = \sqrt{S_y^2 . \left(1 - r^2\right)}$$

onde $S_y^2 = \frac{\sum_{i=1}^{n} \left(y - \overline{y}\right)^2}{n}$

O erro padrão pode ser usado para estabelecer um intervalo de predição para a variável dependente, dado um valor específico da variável independente.

Uma vez que o erro padrão de estimação está baseado em dados de amostra, é apropriado o uso da distribuição *t* de Student com n-2 graus de liberdade. Assim, um intervalo de predição para a variável dependente Y, em análise de regressão simples é:

$$\left[Y \pm t_{n-2;\alpha/2} . \hat{\sigma}_u\right]$$

Para os dados do exemplo 2, teríamos o erro padrão da estimação dado por:

Dado que $S_y^2 = 68{,}65$ e $r^2 = 0{,}59$ então

$$\hat{\sigma}_u = \sqrt{S_y^2 . \left(1 - r^2\right)} = \sqrt{68{,}65 . \left(1 - 0{,}59\right)} = 5{,}30$$

E o intervalo de predição, com 95% de confiança, para um valor de Y=112 seria:

$$[\overline{Y} \pm t_{n-2;\alpha/2} . \hat{\sigma}_u] = [112 \pm 2{,}10 . 5{,}30] = [100{,}87 , 123{,}13]$$

Ou seja, para uma pessoa com 35 anos, o tempo de reação predito estaria entre 100,87 e 123,13 segundos, com 95% de confiança.

Para os dados do exemplo 3 teríamos o erro padrão da estimação dado por:

Dado que $S_y^2 = 1\,366$ e $r^2 = 0{,}996$ então

$$\hat{\sigma}_u = \sqrt{S_y^2 . \left(1 - r^2\right)} = \sqrt{1\,366 . \left(1 - 0{,}996\right)^2} = 2{,}34$$

E o intervalo de predição, com 95% de confiança, para um valor predito de Y = 110 seria:

$$[\overline{Y} - t_{n-2;\alpha/2} . \hat{\sigma}_u] = [110 - 2{,}31 . 2{,}34] = [104{,}59; 115{,}41]$$

Ou seja, para um lote de tamanho 50, seriam necessárias de 104,59 a 115,41 horas, com 95% de confiança.

Análise de Resíduos

Os desvios $e_i = y_i - \hat{y}_i$ ($i = 1, ..., n$) são denominados resíduos e são considerados uma amostra aleatória dos erros. Por este fato, uma análise gráfica dos resíduos é, em geral, realizada para verificar as suposições assumidas para os erros ε_i.

Para verificação dos pressupostos necessários para ajuste de um modelo de regressão é necessário realizar uma Análise de Resíduos. Os 3 tipos de resíduos mais comumente utilizados são:

- Resíduos brutos;
- Resíduos padronizados;
- Resíduos estudentizados.

Ampliando seus conhecimentos

Análise de Regressão Múltipla

A regressão múltipla envolve três ou mais variáveis, ou seja, uma única variável dependente, porém duas ou mais variáveis independentes (explicativas).

A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples. Mesmo quando estamos interessados no efeito de apenas uma das variáveis, é aconselhável incluir as outras capazes de afetar Y, efetuando uma análise de regressão múltipla, por 2 razões:

a) Para reduzir os resíduos. Reduzindo-se a variância residual (erro padrão da estimativa), aumenta a força dos testes de significância;

b) Para eliminar a tendenciosidade que poderia resultar se simplesmente ignorássemos uma variável que afeta Y substancialmente.

Uma estimativa é tendenciosa quando, por exemplo, numa pesquisa em que se deseja investigar a relação entre a aplicação de fertilizante e o volume de safra, atribuímos erroneamente ao fertilizante os efeitos do fertilizante, mais a precipitação pluviométrica.

O ideal é obter o mais alto relacionamento explanatório com o mínimo de variáveis independentes, sobretudo em virtude do custo na obtenção de dados para muitas variáveis e também pela necessidade de observações adicionais para compensar a perda de graus de liberdade decorrente da introdução de mais variáveis independentes.

A equação da regressão múltipla tem a forma seguinte:

$Y = a + b_1x_1 + b_2x_2 + \ldots + b_k x_k + e_i$, onde:

- a = intercepto do eixo y;
- b_i = coeficiente angular da i-ésima variável;
- k = número de variáveis independentes.

Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis resulta um plano, e um problema de k variáveis resulta um hiperplano.

Também na regressão múltipla, as estimativas dos mínimos quadrados são obtidas pela escolha dos estimadores que minimizam a soma dos quadrados dos desvios entre os valores observados Y_i e os valores ajustados $\hat{Y}$.

Na regressão simples:

b = aumento em Y, decorrente de um aumento unitário em X.

Na regressão múltipla:

b_i = aumento em Y se X_i for aumentado de 1 unidade, mantendo-se constantes todas as demais variáveis X_j.

Atividades de aplicação

1. Os encargos diários com o consumo de gás propano (Y) de uma empresa dependem da temperatura ambiente (X). A tabela seguinte apresenta o valor desses encargos em função da temperatura exterior:

Temperatura (°C)	5	10	15	20	25
Encargos (dólares)	20	17	13	11	9

Seja $Y = \beta_0 + \beta_1 X + \varepsilon$ o correspondente modelo de regressão linear.

a) Determine, usando o método dos mínimos quadrados, a respectiva reta de regressão e represente-a no diagrama de dispersão.

b) Quantifique a qualidade do ajuste obtido e interprete.

c) Determine um intervalo de confiança a 95% para os encargos médios com gás propano num dia em que a temperatura ambiente é de 17ºC.

2. Suponha que um analista toma uma amostra aleatória de 9 carregamentos feitos recentemente por caminhões de uma companhia. Para cada carregamento, registra-se a distância percorrida em km (X) e o respectivo tempo de entrega (Y). Obteve-se:

$$\sum x_i = 6\,405, \sum y_i = 23,5,\ \sum x_i^2 = 56\,280\,75, \sum y_i^2 = 74,75, \sum x_i y_i = 20\,295.$$

a) Estime, usando o modelo de regressão linear, o tempo esperado de entrega para uma distância de 1 050km.

b) Comente a afirmação "o tempo de entrega é explicado em aproximadamente 94% pela distância percorrida".

3. Seja Y o número de chamadas telefônicas atendidas num determinado serviço de atendimento a clientes decorridos X minutos após as 8h30. Em determinado dia da semana observaram-se os seguintes pares de valores:

Tempo após 8h30(min)	1	3	4	5	6
Número de chamadas atendidas	2	5	10	11	12

Seja $Y = \beta_0 + \beta_1 X + \varepsilon$ o correspondente modelo de regressão linear.

a) Estime β_0 e β_1 usando o método dos mínimos quadrados e represente a correspondente reta de regressão no diagrama de dispersão.

b) Determine o correspondente coeficiente de determinação, bem como o coeficiente de correlação; como você interpreta os valores obtidos?

c) Estime a variância do erro.

d) Seja E [Y (2)] = E [Y | x = 2]. Estime E [Y (2)]; determine um intervalo de confiança para E [Y (2)] com 95% de confiança.

Gabaritos

Capítulo 1 – Conceitos e Aplicações

1.

a) É uma estratégia adequada. Se a amostra coletada for representativa da população, os resultados serão bastante confiáveis.

b) Também pode ser considerada uma estratégia adequada. A pesquisa atingirá, nos locais de venda, o público-alvo do novo produto e apresentará resultados confiáveis.

c) Esta é uma estratégia mais qualitativa, denominada discussão em grupo (grupo focal). Os resultados obtidos apresentam muitas informações em profundidade, porém sem muita representatividade, pelo número reduzido da amostra.

2.

a) Esta é uma estratégia adequada, pois compara dois grupos de pacientes homogêneos e possibilita avaliar o efeito do novo medicamento. É preciso, no entanto, garantir que o número de pacientes escolhidos seja em número satisfatório.

b) Não é uma estratégia adequada. Não se devem disponibilizar medicamentos novos no mercado sem que antes tenham sido avaliados em laboratório e outros experimentos controlados. E nada garante que será atingida a população alvo de interesse do estudo.

c) É uma estratégia parcialmente adequada. Deve-se avaliar se os pacientes deste hospital representam de forma satisfatória a população alvo ou se é apenas uma escolha por conveniência. Pode ser que os pacientes hospitalizados sejam pacientes em estado mais grave, o que poderá viesar os resultados do estudo.

3.

a) É uma estratégia adequada. Escolhendo uma amostra representativa do lote conseguiremos, com uma boa margem de confiança, avaliar a qualidade do lote.

b) Não é adequado. Não devemos liberar mercadorias para o comércio sem que antes a sua qualidade tenha sido avaliada.

c) Não é adequado. Avaliar 10% do lote pode ser exaustivo ou insuficiente, dependendo do tamanho do lote. Existem maneiras definidas de calcular o número de amostras que vão representar satisfatoriamente a população.

Capítulo 2 – Análise Exploratória de Dados

1. Construindo-se a tabela de freqüência dos dados considerando 5 classes:

$k = 1 + 3{,}3.\log(n)$ $\quad h_i = \frac{AT}{k}$ $\quad AT = 119 - 50$

$k = 1 + 3{,}3.\log(20)$ $\quad h_i = \frac{69}{5}$ $\quad AT = 69$

$k = 1 + 3{,}3 \,.\, 1{,}30103$ $\quad \mathbf{h_i = 13{,}80}$

k = 5,29

Para facilitar a construção da tabela de freqüências, utilizaremos classe igual a 5 e intervalo de classe igual a 15.

Classe	Freqüência	%
50 |— 65	8	40
65 |— 80	7	35
80 |— 95	4	20
95 |— 110	0	0
110 |— 125	1	5

Podemos observar que a grande maioria das instituições (75%) apresentou lucro de até 80 milhões de dólares enquanto que uma delas apresentou um lucro muito superior às demais (119 milhões de dólares).

2. Construindo a tabela com os dados do problema obteremos:

i	Pesos (kg)	f_i	Pm_i	fr_i	%
1	48 |— 53	10	50,5	0,20	20
2	53 |— 58	7	55,5	0,14	14
3	58 |— 63	5	60,5	0,10	10
4	63 |— 68	7	65,5	0,14	14
5	68 |— 73	5	70,5	0,10	10
6	73 |— 78	6	75,5	0,12	12
7	78 |— 83	6	80,5	0,12	12
8	83 |— 88	1	85,5	0,02	2
9	88 |— 93	1	90,5	0,02	2
10	93 |— 98	2	95,5	0,04	4
–	**TOTAL**	**50**		**1**	**100**

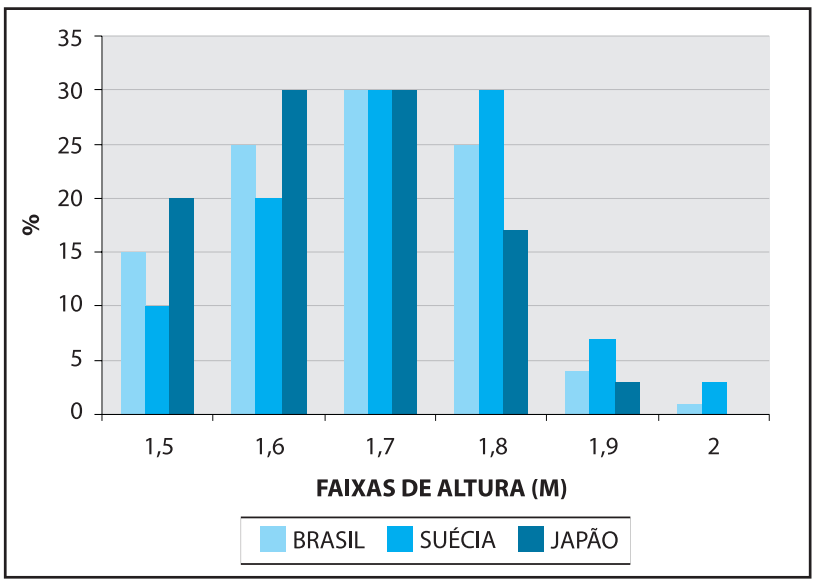
Fazendo a leitura da tabela:

a) 58 **b)** 68 **c)** 5 **d)** 50

e) 65,5 **f)** 10 **g)** 29 **h)** 16

i) 23 **j)** 4% **k)** 34% **l)** 20%

3. Um possível gráfico para representar a distribuição de altura da população dos 3 países poderia ser um histograma:



4. Podemos observar, pela interpretação dos ramos-e-folhas, que as duas corretoras apresentam porcentagens médias de lucros semelhantes, por volta de 5,0%. Por outro lado, a corretora B apresenta uma variabilidade muito menor que a corretora A. A corretora B, portanto apresenta um desempenho muito mais homogêneo que a corretora A.

Capítulo 3 – Medidas de Posição e Variabilidade

1. A. O mais provável seria ganhar menos, pois se o terceiro quartil é de R$ 5.000,00, significa que 75% dos salários são inferiores a este valor, a despeito da média ser de R$ 10.000,00 muito provavelmente influenciada por salários muito elevados dos altos cargos desta empresa.

B. Apresentaria-me na empresa Y, pois lá é praticamente certo que meu salário seria muito próximo da média de R$ 7.000,00 dado que os salários praticamente não apresentam variabilidade; quase todos recebem o mesmo salário.

2. B. O somatório dos valores e o número deles.

3. B. 60.

4. C. a mediana.

5. C. zero.

6. B. a média e a mediana.

7. A. moda.

8. A. desvio padrão e média.

9. D. A dispersão absoluta da turma 1 é maior que a turma 2, mas em termos relativos as duas turmas não diferem quanto ao grau de dispersão das notas.

10. A. R$ 1.050,00

11. A. média

12. D. zero

13. B. ao desvio padrão de X, multiplicado pela constante 5

$$\overline{X}_x = \frac{-2-1+0+1+2}{5} = 0$$

$$\overline{X}_Y = \frac{220+225+230+235+240}{5} = \frac{1\,150}{5} = 230$$

$\overline{X}_x = 0$

x_i	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2 \cdot f_i$
−2	−2	4	4
−1	−1	1	1
0	0	0	0
1	1	1	1
2	2	4	4
TOTAL			**10**

$$S^2 = \frac{10}{4} \rightarrow S^2 = 2{,}5$$

$$S = \sqrt{2{,}5} \rightarrow S = 1{,}58$$

$\overline{X}_Y = 230$

x_i	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2 \cdot f_i$
220	−10	100	100
225	−5	25	25
230	0	0	0
235	5	25	25
240	10	100	100
TOTAL			**25**

$$S^2 = \frac{250}{4} \rightarrow S^2 = 62{,}5$$

$$S = \sqrt{62{,}5} \rightarrow S = 7{,}905$$

$$\frac{7{,}905}{1{,}58} = 5 \text{ (constante)}$$

Capítulo 4 – Introdução à Probabilidade

1.

a) S={KKK, KKC, KCK, CKK, KCC, CKC, CCK, CCC}

b) S={MMM, MMF, MFM, FMM, MFF, FMF, FFM, FFF}

c) S={(1,1), (1,2), …, (1,6), (2,1), …, (2,6), ...,(6,1), ..., (6,6)}

d) S={DD, DV, VD, VV}

e) S={BB, BA, AB, AA}

2.

a) A={(3,6), (4,5), (5,4), (6,3)}

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

c) P(A) = 4/36

d) P(B) = 6/36

e) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/36 + 6/36 - 0 = 10/36$

f) $P(A \cap B) = 0$

3.

a) P(retirar uma bola branca da urna "A") = 5/10

b) P(retirar uma bola branca ou uma vermelha da urna "A") = 8/10

c) P(retirar uma bola branca e uma vermelha da urna "A") = 0

d) P(retirar duas bolas vermelhas da urna "A", com reposição) = (3/10).(3/10) = 9/100

e) P(retirar duas bolas pretas da urna "A", sem reposição) = (2/10).(1/10) = 2/100

4.

$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y) = 3/5 + 4/7 - (3/5 \,.\, 4/7) = 29/35 = 82{,}86\%$

5.

a) P(H) = 60/100 = 0,6 ou 60%.

b) $P(M \cap NE) = 26/100 = 0{,}26$ ou 26%.

c) P(NE) = 65/100 = 0,65 ou 65%

d) $P(H \cap NE) = 39/100 = 0{,}39$ ou 39%.

e) P(M/E) = 14/35 = 0,4 ou 40%

f) P(NE/H) = 39/60 = 0,65 ou 65%

6.

a) $P((B1 \cap B2) \cup (A1 \cap A2) \cup (P1 \cap P2)) = (4/15 \cdot 5/13) + (5/15 \cdot 6/13) + (6/15 \cdot 2/13) = 62/195$

b) $P(A1 \cap P2) = 5/15 \cdot 2/13 = 10/195$

c) $P((A1 \cap P2) \cup (P1 \cap A2)) = (5/15 \cdot 2/13) + (6/15 \cdot 6/13) = 46/195$

d) $P(B1\ C\ B2^{C}) = 4/15 \cdot 8/13 = 32/195$

7.

$P(W) = (1/10 \cdot 3/4) + (3/5 \cdot 1/6) + (3/10 \cdot 1/20) = 3/40 + 3/30 + 3/200 = 0,19$

a) $P(A/W) = P(W \cap A)/ P(W) = P(A) \cdot P(W/A) / P(W) = (1/10 \cdot 3/4)/0,19 = 0,3947$

b) $P(B/W) = P(W \cap B)/ P(W) = P(B) \cdot P(W/B) / P(W) = (3/5 \cdot 1/6)/0,19 = 0,5263$

c) $P(C/W) = P(W \cap C)/ P(W) = P(C) \cdot P(W/C) / P(W) = (3/10 \cdot 1/20)/0,19 = 0,0789$

8.

$P(D) = (0,4 \cdot 0,03) + (0,5 \cdot 0,05) + (0,1 \cdot 0,02) = 0,012 + 0,025 + 0,002 = 0,039$

a) $P(M_1/D) = P(M_1 \cap D)/ P(D) = P(M_1) \cdot P(D/M_1) / P(D) = (0,4 \cdot 0,03)/0,039 = 0,3077$

b) $P(M_2/D) = P(M_2 \cap D)/ P(D) = P(M_2) \cdot P(D/M_2) / P(D) = (0,5 \cdot 0,05)/0,039 = 0,6410$

c) $P(M_3/D) = P(M_3 \cap D)/ P(D) = P(M_3) \cdot P(D/M_3) / P(D) = (0,1 \cdot 0,02)/0,039 = 0,0513$

9.

a) Sabemos que $\Sigma_i\ p(x_i) = 1$, assim: $k/2 + 0,15 + 3k + 0,1 + 0,05 = 1$, ou seja, $3,5k + 0,30 = 1$ e isto implica que $k = 0,2$

b) $P(X>22) = P(X=23) + P(X=24) = 0,15$ ou 15%

c) $P(20<X<24) = P(X=21) + P(X=22) + P(X=23) = 0,85$ ou 85%

d) Pela definição de esperança de uma variável aleatória discreta:

$E(X) = \sum_{i=1}^{\infty} x_i.p.(x_i)$.

Assim,

$E(X) = (20 \,.\, 0,1) + (21 \,.\, 0,15) + (22 \,.\, 0,6) + (23 \,.\, 0,1) + (24 \,.\, 0,05) = 21,85$ dias

e) Pela definição de variância, temos que: $Var(X) = E(X^2) - [E(X)]^2$

Temos que $E(X^2) = (20^2 \,.\, 0,1) + (21^2 \,.\, 0,15) + (22^2 \,.\, 0,6) + (23^2 \,.\, 0,1) + (24^2 \,.\, 0,05) = 478,25$ e assim $Var(X) = 478,25 - (21,85^2) = 0,8275$

f) Custo da obra: 16.000 + (750 . 21,85) = 32.387,50 euros.

Custo da obra + lucro = 34.887,50 euros.

Capítulo 5 – Distribuição Binomial, Distribuição Poisson e Distribuição Normal

1.

a) $P(X \leq 8) = \sum_{x=0}^{8} \binom{10}{x}.0,3^x.0,7^{10-x} = 0,999$

b) $P(X=7) = \binom{10}{7}.0,3^7.0,7^3 = 0,009$

c) $P(X \geq 6) = \sum_{x=7}^{10} \binom{10}{x}.0,3^x.0,7^{10-x} = 0,047$

2.

a) $0,9^5 = 0,59$

3. P(no máximo duas peças defeituosas) =

$P(X=0) + P(X=1) + P(X=2) = \sum_{x=0}^{2} \binom{10}{x}.0,05^x.0,95^{10-x} = 0,9885$ ou 98,85%

4. O número de navios petroleiros que chegam a determinada refinaria, a cada dia, tem distribuição de Poisson, com parâmetro $\lambda = 2$. As atuais instalações do porto podem atender a três petroleiros por dia. Se mais de 3 navios aportarem por dia, os excedentes devem seguir para outro porto.

a) $P(X > 3) = 1 - \sum_{x=0}^{3} \frac{e^{-\lambda}.\lambda^{x}}{x!} = 1 - 0,857 = 0,143$

b) Se as instalações forem ampliadas para permitir mais um petroleiro, teremos:

$P(X \leq 4) = \sum_{x=0}^{4} \frac{e^{-\lambda}.\lambda^{x}}{x!} = 0,947$

c) $E(X) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}.\lambda^{x}}{x!} = \sum_{x=0}^{\infty} x \frac{e^{-2}.2^{x}}{x!} = 2$

d) 1 ou 2 petroleiros. $P(X=1) = P(X=2) = 0,2707$

e) Qual é o número esperado de petroleiros a serem atendidos diariamente?

Se chegarem 0, 1, 2 ou 3 petroleiros todos serão atendidos. Se vierem mais de 3 petroleiros, somente 3 serão atendidos. Dessa forma:

Número esperado:

$0.P(X=0) + 1.P(X=1) + 2.P(X=2) + 3.P(X \geq 3) = 1,78$

f) Se vierem 0,1, 2 ou 3 petroleiros nenhum precisará ir a outros portos. Caso mais de 3 petroleiros cheguem, apenas 3 podem ser recebidos. Assim:

Número esperado:

$1.P(X=4) + 2.P(X=5) + 3.P(X=6) + 4.P(X=7) + ... = 0,22$

5.

c) $P(X=0) = \frac{e^{-5}.5^{0}}{0!} = 0,0067$

6.

a) –9,6 e 29,6

Para obtermos o valor padronizado 1,96, faremos: $\frac{X-10}{10} = 1,96$

Assim, $X = 29,6$

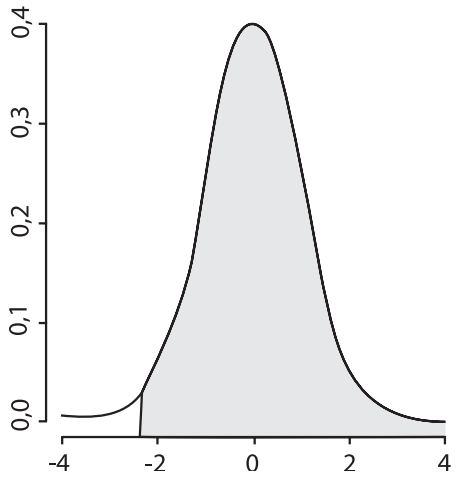
Para obtermos o valor padronizado –1,96, faremos: $\frac{X-10}{10} = -1,96$

Assim, $X = -9,6$

7. $P(X < 5\,000) = P\left(Z < \dfrac{5\,000 - 15\,000}{2\,000}\right) = P(Z < -5) = 0{,}0000002871$

8. P(X≥772N)

$$= P\left(Z \geq \frac{772 - 800}{\sqrt{144}}\right) = P(Z \geq -2{,}33) = 1 - P(Z \leq -2{,}33) = 1 - 0{,}0098 = 0{,}99$$



Capítulo 6 – Estimação de Parâmetros

1.

a) derivando a função de verossimilhança.

2. $\mu_1 = 2$

$\mu_1 = 1$

$\mu_3 = \bar{x} = \sum \dfrac{x}{n} = \dfrac{21}{15} = 1{,}4$

μ_3 é o melhor estimador porque leva em consideração todos os valores da amostra, proporcionando um resumo de dados e por isso pode ser considerado mais confiável.

3. Os limites do intervalo são obtidos a partir da seguinte expressão:

$$\left[\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = \left[78{,}3 - 2{,}58 \cdot \frac{2}{\sqrt{25}}; 78{,}3 + 2{,}58 \cdot \frac{2}{\sqrt{25}}\right] = [77{,}27; 79{,}33]$$

4.

a) 95%

$$\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right]$$

$$= \left[0,298 - 2,78 \cdot \frac{0,024}{\sqrt{5}}; 0,298 + 2,78 \cdot \frac{0,024}{\sqrt{5}}\right] = [0,268; 0,328]$$

b) 99%

$$\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right] =$$

$$= \left[0,298 - 4,60 \cdot \frac{0,024}{\sqrt{5}}; 0,298 + 4,60 \cdot \frac{0,024}{\sqrt{5}}\right] = [0,248; 0,348]$$

5. $$\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}\right) =$$

$$= \left(0,80 - 2,58.\sqrt{\frac{0,80.(0,20)}{200}}; \quad 0,80 + 2,58.\sqrt{\frac{0,80.(0,20)}{200}}\right)$$

$$= (0,723 ; 0,873)$$

O valor 0,90 declarado pelo fabricante, não está incluído no intervalo. Portanto, não temos evidências de que a declaração do fabricante seja legítima, ao nível de significância de 1%.

6.

a) $$\left[\bar{X} - z_{0,05} . \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{0,05} . \frac{\sigma_0}{\sqrt{n}}\right] =$$

$$\left[4,52 - 1,64 . \frac{4}{\sqrt{100}}; 4,52 + 1,64 . \frac{4}{\sqrt{100}}\right] = (3,864; 5,176)$$

b) Sim, a probabilidade do verdadeiro valor da média (valor populacional) estar incluído nos limites do intervalo encontrado é de 90%.

7.

a) O verdadeiro valor do salário inicial médio estará entre 8 e 10 salários mínimos com probabilidade de 95%.

b) Quanto maior o tamanho da amostra, menor é o erro de estimativa e portanto a média amostral estará mais próxima da média populacional. Veja, por exemplo em

$\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right]$ o erro de estimativa $z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}$ é menor a medida que se aumenta o valor de n.

8. Queremos obter uma amostra para estimar a média de uma distribuição normal que respeite a seguinte probabilidade:

$$P\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right] = 0{,}92$$

O valor de Z na tabela será obtido encontrando a área 0,5 – $\alpha/2$ = 0,5 – 0,04 = 0,46. Este valor é 1,75.

Assim, $P\left[\bar{X} - 1{,}75 \cdot \frac{\sqrt{30}}{\sqrt{n}}; \bar{X} + 1{,}75 \cdot \frac{\sqrt{30}}{\sqrt{n}}\right] = 0{,}92$

Como o erro de estimativa, segundo o enunciado, não deve ser superior a 3 unidades, então:

$1{,}75 \cdot \frac{\sqrt{30}}{\sqrt{n}} = 3$. Isolando n, teremos que ele será maior que 10,28.

9. Neste problema, o nível de confiança fixado é de 90% e conseqüentemente, o nível de significância é de 10%.

a) Como não temos uma estimativa prévia da proporção amostral, consideramos p=0,05. Desta forma, teremos:

$$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 \cdot \frac{1}{4} = \left(\frac{z_{\alpha/2}}{2e}\right)^2 \rightarrow n = \left(\frac{1{,}64}{2.0{,}05}\right)^2 = 268{,}96$$

b) Agora temos uma informação prévia sobre a proporção amostral (0,8) e assim o cálculo da amostra será:

$$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 . p_0 . (1 - p_0) = \left(\frac{1{,}64}{0{,}05}\right)^2 . 0{,}20 . 0{,}80 = 172{,}13$$

Capítulo 7 – Testes de Hipóteses: conceitos

1.

a) A população é a totalidade de alunos do Curso X. A amostra é composta pelos 80 alunos do Curso, selecionados aleatoriamente. O parâmetro de interesse é a proporção de alunos favoráveis a eliminação da disciplina de Estatística do currículo. O teste adequado seria para testar a proporção de uma amostra.

b) A população é a totalidade de pessoas obesas com certa idade. A amostra é composta pelas 20 pessoas obesas daquela faixa etária, selecionadas aleatoriamente. O parâmetro de interesse é a média de perda de peso, ou seja peso antes – peso depois (do curso). O teste adequado seria para comparar amostras relacionadas.

c) A população é a totalidade de moradores fumantes da cidade. A amostra é composta pelas 100 pessoas fumantes, selecionadas aleatoriamente. Um dos parâmetros de interesse pode ser a média de cigarros consumidos. O teste adequado seria para testar a média de uma amostra.

2.

a) H_0 = opinião antes = opinião depois

H_a = opinião antes ≠ opinião depois

b) Embora a maioria das pessoas tenha se manifestado mais favorável ao candidato, não seria prudente afirmarmos que este resultado possa ser considerado estatisticamente significativo.

c) Com este tamanho de amostra já é possível realizar um teste de significância. Muito provavelmente, iremos rejeitar a hipótese nula, de igualdade das opiniões. Poderemos, se o teste comprovar, inferir os resultados para toda a população e afirmar com um certo nível de confiança, que se passou a ter melhor impressão sobre o candidato após a apresentação.

d) Um teste para comparação da proporção de duas amostras relacionadas (antes e depois da apresentação).

3.

a) H_0 = vendas sem brinde = vendas com brinde

H_a = vendas sem brinde ≠ vendas com brinde

b) Com exceção de uma loja, todas as 5 demais apresentaram maiores índices de venda ao oferecer o brinde. É um forte indicativo de maiores vendas com oferta do brinde, embora o número de lojas participantes deste experimento possa ser considerado baixo.

c) O tipo de teste mais adequado seria um teste para comparação de médias de duas amostras independentes, embora pudesse ser utilizado também um teste para comparação de médias de duas amostras relacionadas, desde que bem justificado o critério de pareamento das unidades observadas.

4.

a) H_0 = eficácia relativa comerciais de 15 segundos = eficácia relativa comerciais de 30 segundos

H_a = eficácia relativa comerciais de 15 segundos < eficácia relativa comerciais de 30 segundos

b) Caso o tamanho de amostra seja satisfatório e a suposição de normalidade seja comprovada, pode ser aplicado um teste paramétrico para comparação de duas amostras independentes. Caso os pressupostos para aplicação de um teste paramétrico não sejam atendidos, podemos recorrer a um teste não paramétrico para comparação de duas amostras independentes. O nível de significância mais indicado seria de 1% ou 5%.

c) Nas 4 variáveis avaliadas podemos observar que os comerciais de 30 segundos apresentaram uma melhor avaliação em relação aos comerciais de 15 segundos.

Capítulo 8 – Testes de Hipóteses

1. As hipóteses a serem testadas são:

H_0: As produções médias de milho estão de acordo com a especificação do fabricante;

H_a: A produção média de milho não se ajusta à distribuição especificada pelo fabricante.

Aplicando o teste Qui-quadrado para testar a aderência dos dados à distribuição especificada pelo fabricante, temos:

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \frac{(13-12)^2}{12} + \frac{(18-20)^2}{20} + \ldots + \frac{(11-13)^2}{13} = 3{,}04$$

Consultando a tabela de valores críticos, considerando k–1 = 5 graus de liberdade e $\alpha = 0{,}05$, temos $x^2 = 11{,}1$. Como o valor calculado é inferior ao valor crítico, não rejeitamos a hipótese nula e podemos concluir que os dados se ajustam satisfatoriamente à distribuição especificada pelo fabricante.

2. As hipóteses a serem testadas são:

H_0: a nota média dos estudantes de escola pública não difere da nota média dos estudantes da escola particular;

Ha: a nota média dos estudantes de escola pública difere da nota média dos estudantes da escola particular.

Aplicando o teste t de Student para comparação de duas amostras independentes, temos que verificar primeiramente se as variâncias podem ser consideradas iguais. Construindo o intervalo de confiança para a razão de variâncias temos:

$$\left[\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_2}; \frac{S_1^2}{S_2^2}\cdot\frac{1}{F_1}\right] = \left[\frac{64}{100}\cdot\frac{1}{1{,}4833}; \frac{64}{100}\cdot 1{,}4833\right] = (0{,}43\,;\,0{,}94)$$

Desta forma as variâncias não são iguais.

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(75{,}9 - 74{,}5)}{\sqrt{\frac{64}{117} + \frac{100}{200}}} = 1{,}3682$$

Consultando a tabela de valores críticos, considerando $n_1 + n_2 - 2 = 315$ graus de liberdade e $\alpha = 0{,}05$, temos $t_c = 1{,}96$. Como o valor calculado é inferior ao valor crítico, não rejeitamos a hipótese nula e podemos concluir que as notas médias das duas escolas não diferem.

3. As hipóteses a serem testadas são:

H_0: a nova droga não baixa a febre, ou seja, Diferença = 0;

H_a: a nova droga baixa a febre, ou seja, Diferença ≠ 0.

Aplicando o teste t de Student para comparação de duas amostras relacionadas, temos:

$$S_d = \sqrt{\frac{\sum d^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{80 - \left(15.(1{,}866)^2\right)}{14}} = 1{,}408$$ e a estatística do teste será:

$$t = \frac{1{,}866}{1{,}408 / \sqrt{15}} = 5{,}131$$

Consultando a tabela de valores críticos, considerando n–1 = 14 graus de liberdade e $\alpha = 0{,}05$ (bilateral), temos $t_c = 2{,}14$. Como o valor calculado é superior ao valor crítico, rejeitamos a hipótese nula e podemos concluir que a nova droga baixa a febre significativamente.

4. As hipóteses a serem testadas são:

H_0: a proporção de animais com verminose é igual nos dois grupos;

H_a: a proporção de animais com verminose é inferior no grupo que teve alteração da dieta.

O teste, portanto, é unilateral e aplicando o teste Z para proporção, temos:

$$p = \frac{n_1.p_1 + n_2.p_2}{n_1 + n_2} = \frac{(500.0{,}10) + (100.0{,}04)}{600} = 0{,}09$$

$$S_p = \sqrt{\frac{p.(1-p)}{n_1} + \frac{p.(1-p)}{n_2}} = \sqrt{\frac{0{,}09.0{,}91}{500} + \frac{0{,}09.0{,}91}{100}} = 0{,}031$$

$$Z = \frac{p_1 - p_2}{S_p} = \frac{0{,}10 - 0{,}04}{0{,}031} = 1{,}93$$

Consultando a tabela de valores críticos da distribuição normal padrão, considerando $\alpha = 0{,}01$, temos $Z_c = 2{,}33$. Como o valor calculado é inferior ao valor crítico, não rejeitamos a hipótese nula e podemos concluir que a doença não diminuiu significativamente de intensidade.

5. As hipóteses a serem testadas são:

H_0: não existe diferença de satisfação entre os 3 hospitais;

H_a: existe pelo menos uma diferença entre os hospitais, com relação à média de satisfação.

Realizando o Teste F, de Análise de Variâncias, temos:

$$SQA = \sum \frac{T_k^2}{n_K} - \frac{T^2}{N} = \frac{(873)^2}{10} + \frac{(898)^2}{15} + \frac{(954)^2}{13} - \frac{(2725)^2}{38} =$$
$$= 76\,212,9 + 53\,760,267 + 70\,008,92 - 195\,411,1842 = 4\,570,9$$

$$SQT = \sum_{i=1}^{n}\sum_{k=1}^{k} X^2 - \frac{T^2}{N} = 200\,623 - 195\,411,1842 = 5\,211,82$$

e SQE = SQT – SQA = 5 211,82 – 4 570,9 = 640,92

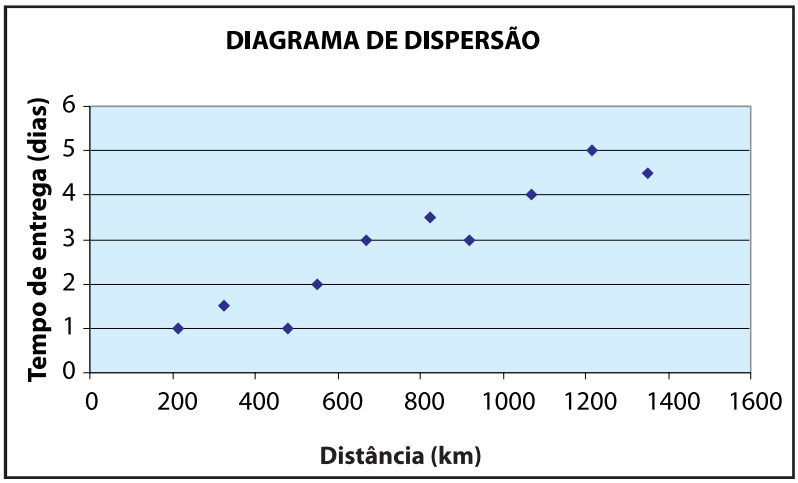
Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Entre grupos	4 570,90	2	2 285,450	124,8
Erro amostral	640,92	35	18,312	
Total	**5 211,82**	**37**		

O valor crítico de F, definido pelo nível de significância ($\alpha = 0,05$) e pelos graus de liberdade 2 e 35 é igual a 3,30. Como $F_{cal} > F_{crit}$ devemos rejeitar a hipótese nula. Os hospitais diferem em relação à satisfação média.

Capítulo 9 – Análise de Correlação e Medidas de Associação

1.

a)



b) $C(X,Y)=\frac{\sum(X_i-\bar{X}).(Y_i-\bar{Y})}{n}=\frac{4\,653}{10}=465,3$

$r=\frac{C(X,Y)}{S_Y.S_X}=\frac{465,3}{360,26.1,36}=0,9497$

c) $r^2=(r)^2=(0,9497)^2=0,9019$

d) $t_c=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{0,9497\sqrt{8}}{\sqrt{1-0,9019}}=8,576$

Comparando o valor calculado com o valor crítico, considerando 8 graus de liberdade e 5% de significância temos $t_{crítico} = 2,31$. Assim, podemos considerar o coeficiente de correlação altamente significativo.

2. $t_c=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{0,50\sqrt{8}}{\sqrt{1-0,25}}=1,63$

Comparando o valor calculado com o valor crítico, considerando 8 graus de liberdade e 5% de significância temos $t_{crítico} = 2,31$. Assim, não podemos considerar o coeficiente de correlação significativo. Não existe correlação entre a renda familiar e os débitos a descoberto de curto prazo.

3. $C(X,Y)=\frac{\sum(X_i-\overline{X}).(Y_i-\overline{Y})}{n}=\frac{654}{8}=81{,}75$

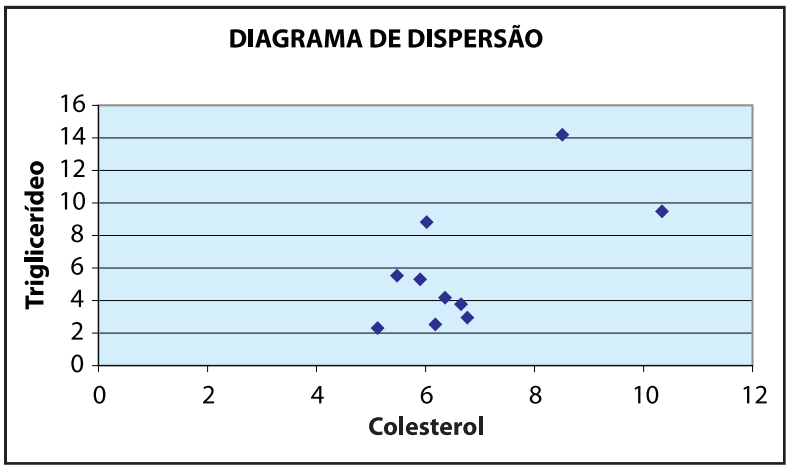
$$r=\frac{C(X,Y)}{S_Y.S_X}=\frac{81{,}75}{12{,}77.10{,}22}=0{,}626$$

$$t_c=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{0{,}626\sqrt{6}}{\sqrt{1-0{,}392}}=1{,}967$$

Comparando o valor calculado com o valor crítico, considerando 6 graus de liberdade e 5% de significância temos $t_{crítico}$ = 2,45. Assim, podemos considerar o coeficiente de correlação não significativo, ou seja, não existem evidências de correlação significativa entre habilidade verbal e habilidade matemática.

4.

a)



b) baseado no diagrama acima, não está muito clara a existência de relação linear entre colesterol e triglicerídeos.

Paciente	Colesterol (mmol/l)	Triglicerídeos (mmol/l)	Postos Colesterol	Postos Triglicerídeos	d_i	d_i^2
1	5,12	2,30	1	1	0	0
2	6,18	2,54	5	2	3	9
3	6,77	2,95	8	3	5	25
4	6,65	3,77	7	4	3	9

Paciente	Colesterol (mmol/l)	Triglicerídeos (mmol/l)	Postos Colesterol	Postos Triglicerídeos	d_i	d_i^2
5	6,36	4,18	6	5	1	1
6	5,90	5,31	3	6	–3	9
7	5,48	5,53	2	7	–5	25
8	6,02	8,83	4	8	–4	16
9	10,34	9,48	10	9	1	1
10	8,51	14,20	9	10	–1	1
Soma						**96**

c) $$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6.96}{1000 - 10} = 0{,}418$$

Para verificar a significância do valor observado de r_s podemos usar a expressão de t de Student

$$t = r_s . \sqrt{\frac{n-2}{1-r_s^2}} = 0{,}418 . \sqrt{\frac{8}{1-0{,}1748}} = 1{,}30$$

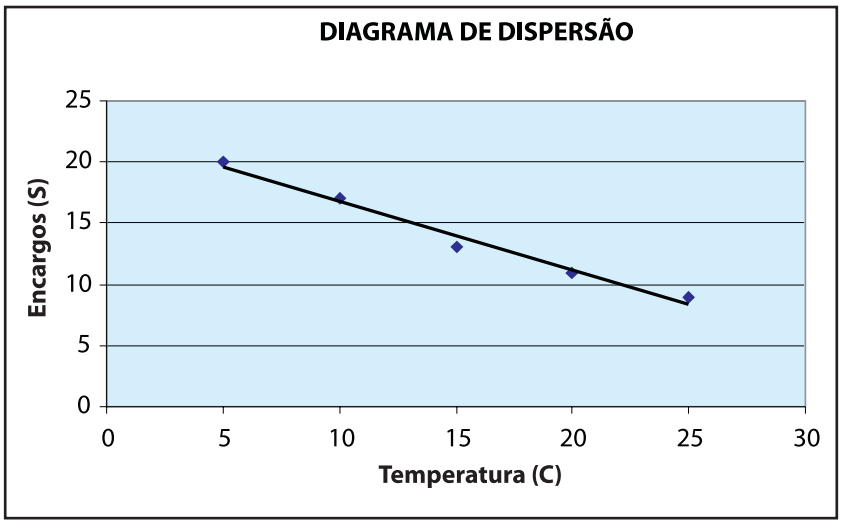
Comparando o valor calculado com o valor crítico, considerando 8 graus de liberdade e 5% de significância temos $t_{crítico} = 2{,}31$. Assim, podemos considerar o coeficiente de associação significativo, ou seja, existem evidências de correlação significativa entre colesterol e triglicerídeos.

Capítulo 10 – Análise de Regressão

1. $\hat{\beta}_1 = \dfrac{\sum x_i.y_i - n.\bar{y}.\bar{x}}{\sum x_i^2 - n.\bar{x}^2} = \dfrac{910 - 5.14.15}{1375 - 5.225} = -0{,}56$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1.\bar{x} = 14 - (-0{,}56).15 = 22{,}4$

Então $\hat{Y} = 22{,}4 - 0{,}56X$.



b) Dado que $\bar{y} = \dfrac{70}{5} = 14$

$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2 = \sum_{i=1}^{n} (22{,}4 - 0{,}65x_i - 14)^2 = 78{,}4$

$SQres = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (y_i - 22{,}4 - 0{,}65x_i)^2 = 1{,}6$

SQtotal = 78,4 + 1,6 = 80

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	1	78,4	78,4	147	< 0,001
Resíduos	3	1,6	0,53		
Total	**4**	**80**	**20**		

A regressão pode ser considerada altamente significativa ($p < 0,001$). O coeficiente de determinação calculado a partir dos dados da Anova, $r^2 = 78,4/80 = 0,98$. Pode se considerar bastante satisfatória a qualidade do ajuste.

c) $S_y^2 = \frac{\sum_{i=1}^{n}(y-\bar{y})^2}{n} = \frac{80}{5} = 16$

$\hat{\sigma} = \sqrt{S_y^2.(1-r^2)} = \sqrt{16.(1-0,98)} = 0,565$

$\hat{} = 22,4 - 0,56 \,.\, 17 = 12,88$

$[\hat{Y} \pm t_{n-2;\,\alpha/2}.\hat{\sigma}_u] = [12,88 \pm 3,18\,.\,0,565] = [11,08; 14,68]$

2.

a) $\hat{\beta}_1 = \frac{\sum x_i.y_i - n.\bar{y}.\bar{x}}{\sum x_i^2 - n.\bar{x}^2} = \frac{20\,295 - 9.2,61.711,67}{5\,628\,075 - 9.(711,66)^2} = \frac{3\,577,87}{106\,993,4} = 0,00334$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1.\bar{x} = 2,611 - 0,00334.711,66 = 0,234$

Então $\hat{Y} = 0,234 + 0,00334.X = 0,234 + 0,00334 \,.\, 1\,050 = 3,741$ dias

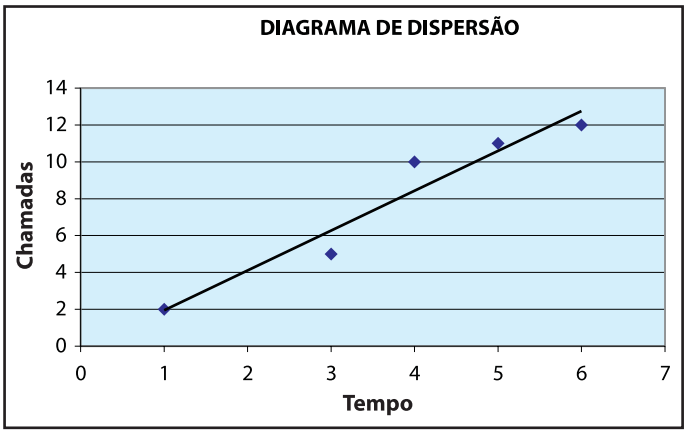
b) Isto significa que 94% da variação do tempo de entrega está associada à distância a ser percorrida e outras variáveis como: região urbana ou rural, clima durante o percurso, treinamento do motorista etc, são responsáveis pelos demais 6%. No entanto, essas variáveis não foram observadas nesse estudo.

3.

a) $\hat{\beta}_1 = \frac{\sum x_i.y_i - n.\bar{y}.\bar{x}}{\sum x_i^2 - n.\bar{x}^2} = \frac{184 - 5.8.3,8}{87 - 5.(3,8)^2} = \frac{32}{14,8} = 2,16$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1.\bar{x} = 8 - 2,16.3,8 = -0,21$

Então $\hat{Y} = -0,21 + 2,16.X$



b) $SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2 = \sum_{i=1}^{n} (-0{,}21 + 2{,}16x_i - 8)^2 = 69{,}05$

$SQres = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (y_i + 0{,}21 - 2{,}16x_i)^2 = 4{,}8109$

$SQtotal = 69{,}05 + 4{,}8109 = 73{,}8609$

Assim $r^2 = \dfrac{SQres}{SQtotal} = \dfrac{69{,}05}{73{,}86} = 0{,}9349$ e $r = \sqrt{r^2} = 0{,}9668$

O coeficiente de determinação calculado nos indica que é bastante satisfatória a qualidade do ajuste. A relação entre as duas variáveis pode ser considerada bastante forte, pela análise do coeficiente de correlação.

c) $\hat{\sigma}_u = \sqrt{\dfrac{\sum (y - \hat{Y})^2}{n-2}} = \sqrt{\dfrac{4{,}8109}{3}} = 1{,}266$

d) $E[Y(2)] = -0{,}21 + 2{,}16 \cdot 2 = 4{,}11$

$[\hat{Y} \pm t_{n-2;\,\alpha/2} . \hat{\sigma}_u] = [4{,}11 \pm 3{,}18 . 1{,}266] = [0{,}08;\ 8{,}13]$

Anexo I

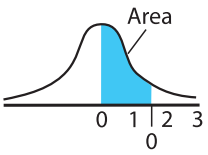


Tabela de valores críticos – Normal

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3112	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.1965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0..4982	0.4982	0.4983	0.4983	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4988	0.4989	0.4989	0.4990	0.4990
3.1	0.49903									

Anexo II

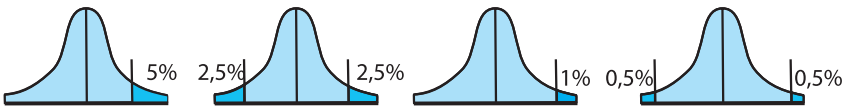


Tabela de valores críticos – t de Student

df	0.05	0.025	0.01	0.005
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	.3.55
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
50	1.676	2.009	2.403	2.678
100	1.660	1.984	2.364	2.626
∞	1.645	1.960	2.326	2.576

Anexo III

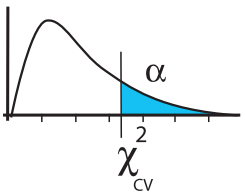


Tabela de valores críticos – Qui-quadrado				
df	**0.05**	**0.025**	**0.01**	**0.005**
1	3.84	5.02	6.63	7.88
2	5.99	7.38	9.21	10.60
3	7.82	9.35	11.35	12.84
4	9.49	11.14	13.28	14.86
5	11.07	12.83	15.09	16.75
6	12.59	14.45	16.81	18.55
7	14.07	16.01	18.48	20.28
8	15.51	17.54	20.09	21.96
9	16.92	19.02	21.66	23.59
10	18.31	20.48	23.21	25.19
11	19.68	21.92	24.72	26.75
12	21.03	23.34	26.21	28.30
13	22.36	24.74	27.69	29.82
14	23.69	26.12	29.14	31.31
15	25.00	27.49	30.58	32.80
16	26.30	28.85	32.00	34.27
17	27.59	30.19	33.41	35.72
18	28.87	31.53	34.81	37.15
19	30.14	32.85	36.19	38.58
20	31.41	34.17	37.56	40.00
21	32.67	35.48	38.93	41.40
22	33.93	36.78	40.29	42.80
23	35.17	38.08	41.64	44.18
24	36.42	39.37	42.98	45.56
25	37.65	40.65	44.32	46.93
26	38.89	41.92	45.64	48.29
27	40.11	43.20	46.96	49.64
28	41.34	44.46	48.28	50.99
29	42.56	45.72	49.59	52.34
30	43.77	46.98	50.89	53.67
40	55.75	59.34	63.71	66.80
50	67.50	71.42	76.17	79.52
100	124.34	129.56	135.82	140.19

Anexo IV

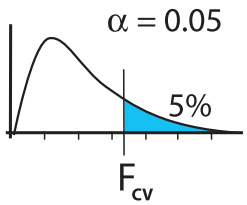


Tabela de valores críticos – F de Snedecor

Degrees of Freedom for the F-Ratio denominator	*Degrees of Freedom for the F-Ratio numerator*									
	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1
2	18.51	19.00	19.16	19.25	19.30	19.36	19.35	19.37	19.38	19.40
3	10.13	9.55	9.328	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85
1000	3.85	3.01	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84

Anexo V

Tabela de valores críticos – Mann Whitney																				
1- tail test at $\alpha = 0.025$ or 2- tail test at $\alpha = 0.05$																				
	N_1																			
N_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	21	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	95	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Anexo V – Continuação

1- tail test at $\alpha = 0.05$ or 2- tail test at $\alpha = 0.10$																				
	N_1																			
N_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2					0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3			0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4			0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5		0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6		0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7		0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
8		1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
9		1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
10		1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
11		1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
12		2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
13		2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
14		2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
15		3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
16		3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
17		3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
18		4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

$N_1 < N_2$

Anexo VI

Tabela de valores críticos – Lilliefors		
n	**α= 0,05**	**α=0,01**
5	0,337	0,405
10	0,258	0,294
15	0,220	0,257
20	0,190	0,231
25	0,173	0,200
30	0,161	0,187
>30	$0,886/\sqrt{n}$	$1,031/\sqrt{n}$

Anexo VII

Tabela de valores críticos – Wilcoxon								
Number of pairs N	T	.05 α	T	.025 α	T	.01 α	T	.005 α
5	0	.0313						
	1	.0625						
6	2	.0469	0	.0156				
	3	.0781	1	.0313				
7	3	.0391	2	.0234	0	.0078		
	4	.0547	3	.0391	1	.0156		
8	5	.0391	3	.0195	1	.0078	0	.0039
	6	.0547	4	.0273	2	.0117	1	.0078
9	8	.0488	5	.0195	3	.0098	1	.0039
	9	.0645	6	.0273	4	.0137	2	.0059
10	10	.0420	8	.0244	5	.0098	3	.0049
	11	.0527	9	.0322	6	.0137	4	.0068
11	13	.0415	10	.0210	7	.0093	5	.0049
	14	.0508	11	.0269	8	.0122	6	.0068
12	17	.0461	13	.0212	9	.0081	7	.0046
	18	.0549	14	.0261	10	.0105	8	.0061
13	21	.0471	17	.0239	12	.0085	9	.0040
	22	.0549	18	.0287	13	.0107	10	.0052
14	25	.0453	21	.0247	15	.0083	12	.0043
	26	.0520	22	.0290	16	.0101	13	.0054
15	30	.0473	25	.0240	19	.0090	15	.0042
	31	.0535	26	.0277	20	.0108	16	.0051
16	35	.0467	29	.0222	23	.0091	19	.0046
	36	.0523	30	.0253	24	.0107	20	.0055
17	41	.0492	34	.0224	27	.0087	23	.0047
	42	.0544	35	.0253	28	.0101	24	.0055
18	47	.0494	40	.0241	32	.0091	27	.0045
	48	.0542	41	.0269	33	.0104	28	.0052
19	53	.0478	46	.0247	37	.0090	32	.0047
	54	.0521	47	.0273	38	.0102	33	.0054
20	60	.0487	52	.0242	43	.0096	37	.0047
	61	.0527	53	.0266	44	.0107	38	.0053

Anexo VIII

Tabela de valores críticos – Kruskal Wallis														
n1	n2	n3	H	P	n1	n2	n3	H	P	n1	n2	n3	H	P
2	1	1	2,7000	0,500	4	4	1	6,6667	0,010	5	4	1	6,9545	0,008
2	2	1	3,6000	0,200				6,1667	0,022				6,8400	0,011
2	2	2	4,5714	0,067				4,9667	0,048				4,9855	0,044
			3,7143	0,200				4,8667	0,054				4,8600	0,056
3	1	1	3,2000	0,300				4,1667	0,082				3,9873	0,098
3	2	1	4,2857	0,100				4,0667	0,102				3,9600	0,102
			3,8571	0,133	4	4	2	7,0364	0,006	5	4	2	7,2045	0,009
3	2	2	5,3572	0,029				6,8727	0,011				7,1182	0,010
			4,7143	0,148				5,4545	0,046				5,2727	0,049
			4,5000	0,067				5,2364	0,052				5,2682	0,050
			4,4643	0,105				4,5545	0,098				4,5409	0,098
3	3	1	5,1429	0,043				4,4455	0,103				4,5182	0,101
			4,5714	0,100	4	4	3	7,1439	0,010	5	4	3	7,4449	0,010
			4,0000	0,129				7,1364	0,011				7,3949	0,011
3	3	2	6,2500	0,011				5,5985	0,049				5,6564	0,049
			5,3611	0,032				5,5758	0,051				5,6308	0,050
			5,1389	0,061				4,5455	0,099				4,5487	0,099
			4,5556	0,100				4,4773	0,102				4,5231	0,103
			4,2500	0,012	4	4	4	7,6538	0,008	5	4	4	7,7604	0,009
3	3	3	7,2000	0,004				7,5385	0,011				7,7440	0,011
			6,4889	0,011				5,6923	0,049				5,6571	0,049
			5,6889	0,029				5,6538	0,054				5,6176	0,050
			5,6000	0,050				4,6539	0,097				4,6187	0,100
			5,0667	0,086				4,5001	0,104				4,5527	0,102
			4,6222	0,100	5	1	1	3,8571	0,143	5	5	1	7,3091	0,009
4	1	1	3,5714	0,200	5	2	1	5,2500	0,036				6,8364	0,011
4	2	1	4,8214	0,057				5,0000	0,048				5,1273	0,046
			4,5000	0,076				4,4500	0,071				4,9091	0,053
			4,0179	0,114				4,2000	0,095				4,1091	0,086
4	2	2	6,0000	0,014				4,0500	0,119				4,0364	0,105
			5,3333	0,033	5	2	2	6,5333	0,008	5	5	2	7,3385	0,010
			5,1250	0,052				6,1333	0,013				7,2692	0,010
			4,4583	0,100				5,1600	0,034				5,3385	0,047
			4,1667	0,105				5,0400	0,056				5,2462	0,051
4	3	1	5,8333	0,021				4,3733	0,090				4,6231	0,970
			5,2083	0,050				4,2933	0,122				4,5077	0,100
			5,0000	0,057	5	3	1	6,4000	0,012	5	5	3	7,5780	0,010
			4,0556	0,093				4,9600	0,048				7,5429	0,010
			3,8889	0,129				4,8711	0,052				5,7055	0,046
4	3	2	6,4444	0,008				4,0178	0,095				5,6264	0,510
			6,3000	0,011				3,8400	0,123				4,5451	0,100
			5,4444	0,046	5	3	2	6,9091	0,009				4,5363	0,102
			5,4000	0,051				6,8218	0,010	5	5	4	7,8229	0,100
			4,5111	0,098				5,2509	0,049				7,7914	0,010
			4,4444	0,102				5,1055	0,052				5,6657	0,049
								4,6509	0,091				5,6429	0,050
								4,4945	0,101				4,5229	0,099
					5	3	3	7,0788	0,009				4,5200	0,101
								6,9818	0,011	5	5	5	8,0000	0,009
								5,6485	0,049				7,9800	0,010
													5,7800	0,049

Referências

BUSSAB, W. O.; MORETIN, P. A. **Estatística Básica**. 4. ed. São Paulo: Saraiva, 2003.

BARROS, Emilio. **Aplicações e Simulações Monte Carlo e *Bootstrap***. Monografia (Bacharelado em Estatística) – Universidade Estadual de Maringá, Maringá, 2005. Disponível em: <http://www.des.uem.br/graduacao/Monografias/Monografia_Emilio.pdf.>. Acesso em: 23 nov. 2007.

CAMPOS, G. M. **Estatística Prática para Docentes e Pós-Graduados**. Disponível em: <http://www.forp.usp.br/restauradora/gmc/gmc_livro/gmc_livro_cap14.html>. Acesso em: 23 nov. 2007.

COSTA NETO, P. L. de O. **Estatística Básica**. 2. ed. São Paulo: Edgard Blücher, 2002.

GONÇALVES, Lóren Pinto Ferreira. **Avaliação de Ferramentas de Mineração de Dados como Fonte de Dados Relevantes para a Tomada de Decisão**: aplicação na Rede Unidão de Supermercados. Dissertação (Mestrado Interinstitucional em Administração) – Universidade da Região da Campanha (Urcamp), São Leopoldo, 2001. Disponível em: <http://volpi.ea.ufrgs.br/teses_e_dissertacoes/td/000410.pdf>

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. **Análise Exploratória de Dados** – Técnicas Robustas. Lisboa: Edições Salamandra, 1983.

HOEL, PORT & STONE. **Introdução à Teoria da Probabilidade**. Rio de Janeiro: Editora Interciência ,1981.

KAZMIER, L. J. **Estatística Aplicada à Economia e Administração**. 4. ed. São Paulo: Bookman 2007.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. et al. **Estatística**: Teoria e Aplicações – Usando Microsoft Excel. 3. ed. Rio de Janeiro: LTC, 2005.

MATTAR, F. N. **Pesquisa de Marketing**. São Paulo: Atlas, 2001.

_____. São Paulo: Atlas, 1996. (Edição compacta).

MEYER, P. L. **Probabilidade**: Aplicações à Estatística. 2. ed. Rio de Janeiro: LTC, 2000.

SIEGEL, S.; CASTELLAN JR., N. J. **Estatística Não-Paramétrica para Ciências do Comportamento**. Porto Alegre: Artmed, 2006.

TRIOLA, M. F. **Introdução à Estatística**. 9. ed. Rio de Janeiro: LTC, 2005.

VIEIRA, S., WADA, R. **O que é Estatística?** 3. ed. São Paulo: Brasiliense, 1991.

WONNACOT, T. H. WONNACOTT, R. J. **Estatística Aplicada à Economia e à Administração**. Rio de Janeiro: LTC, 1981.