UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DE RIBEIRÃO PRETO
DEPARTAMENTO DE ADMINISTRAÇÃO

RAD1507 – Estatística Aplicada à Administração I

Prof. Dr. Evandro Marcos Saidel Ribeiro

RESUMO

REVISÃO DE INTRODUÇÃO À ESTATÍSTICA

Medidas de centro

Média (média aritmética)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mediana (primeiramente: ordenar os dados)

A mediana é o valor que divide o conjunto de dados ao meio, ou seja, o valor que separa 50% dos valores inferiores dos 50% dos valores superiores, em outras palavras é o percentil 50.

$$\tilde{x} = P_{50}$$

Moda

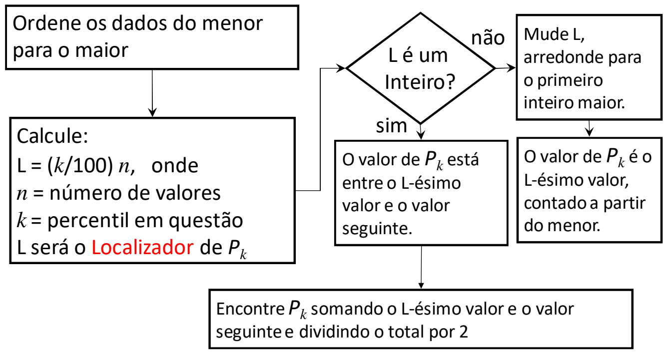
A moda é o valor mais frequente no conjunto de dados. Mais utilizados para variáveis discretas.

Ponto Médio

O ponto médio é a média entre o valor máximo e mínimo:

$$PtoMed = \frac{x_{\max} + x_{\min}}{2}$$

Como localizar o percentil



Medidas de dispersão

Desvio Padrão Amostral

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Variância s^2

Distribuição Binomial

A probabilidade de se obter exatamente x sucessos em n tentativas, sendo a p a probabilidade de sucesso em uma tentativa, é dada por

$$P(x) = \frac{n!}{(n-x)!\,x!}p^x q^{n-x}$$

e $q = 1 - p$

Distribuição Normal

Densidade de probabilidade

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

No Excel a densidade é obtida por:

DIST.NORM(x; μ;σ)

ou DIST.NORMP(z)

No Excel a área acumulada até um valor de z é:

INV.NORM(x; μ;σ)

ou INV.NORMP(z)

Score-z ou padronização dos dados

Encontre a média e o desvio padrão. Em seguida utilize a expressão:

$$z = \frac{(x - \bar{x})}{s}$$

Intervalo de confiança (estimativa intervalar)

Estimativa para proporção

$\hat{p} - E < p < \hat{p} + E$ sendo $\hat{p}$ a proporção de sucessos observada na amostra: $\hat{p} = x/n$, onde x é o número de sucessos observados.

O erro é dado por: $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}$.

Se for necessário estimar o tamanho da amostra para um erro máximo, considere:

$$n = \frac{N\hat{p}\hat{q}(z_{\alpha/2})^2}{\hat{p}\hat{q}(z_{\alpha/2})^2 + (N-1)E_{\max}^2}$$

Observe que se a população puder ser considerada infinita as expressões acima não dependem de N.

Estimativa para média (σ conhecido)

$\bar{x} - E < \mu < \bar{x} + E$ sendo $\bar{x}$ a média amostral.

O erro é dado por: $E = z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N-n}{N-1}}$.

Se for necessário estimar o tamanho da amostra para um erro máximo, considere:

$$n = \frac{N\sigma^2 (z_{\alpha/2})^2}{\sigma^2 (z_{\alpha/2})^2 + (N-1)E_{\max}^2}$$

Estimativa para média (σ não-conhecido)

$\bar{x} - E < \mu < \bar{x} + E$ sendo $\bar{x}$ a média amostral.

O erro é dado por: $E = t_{\alpha/2} \dfrac{s}{\sqrt{n}}$. Observe que o desvio padrão é o amostral e utiliza-se a distribuição t de Student para obter o valor crítico $t_{\alpha/2}$.

Estimativa para desvio padrão ou variância

O intervalo é obtido com base na distribuição qui-quadrado (χ^2). A distribuição qui-quadrado não é simétrica e os valores críticos são sempre positivos. Para fazer o intervalo de confiança utiliza-se χ_E^2 (valor crítico da esquerda) e χ_D^2 (valor crítico da direita). as estimativas intervalares são:

Para o desvio padrão,

$$\sqrt{\frac{(n-1)s^2}{\chi_D^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_E^2}}.$$

Para a variância,

$$\frac{(n-1)s^2}{\chi_D^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_E^2}.$$

ESTATÍSTICA APLICADA À ADM. I

TESTE DE HIPÓTESE: Introdução

Ter em mente a **"Regra do evento raro":**
Se, sob uma dada suposição, a probabilidade de um evento observado particular é excepcionalmente pequena, concluímos que a suposição provavelmente não é correta.

Procedimento para Teste de Hipótese:

1 – Escreva a afirmativa original na forma simbólica.

2 – Escreva o oposto da afirmativa original na forma simbólica.

3 – Expresse H_0 e H_1. (H_0 é sempre a igualdade)

4 – Selecione o nível de significância α.

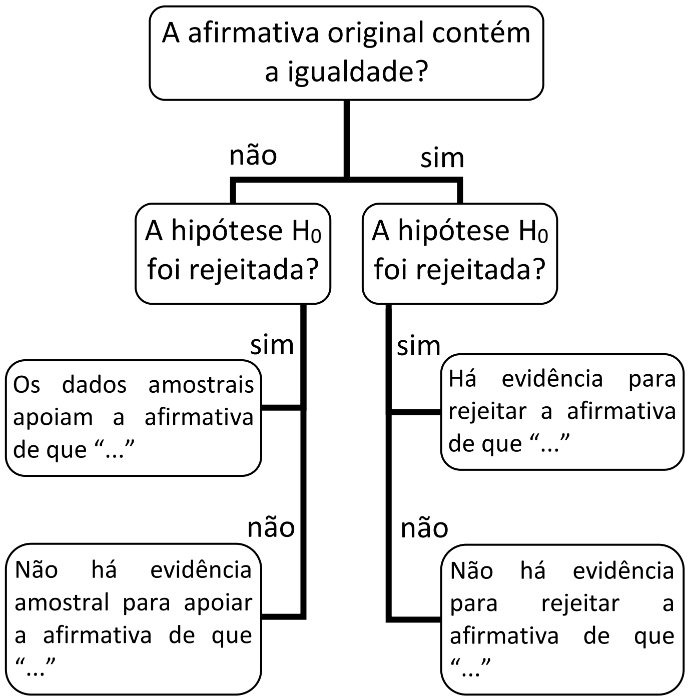
5 – Verifique a distribuição a ser utilizada.

6 – Calcule a estatística teste.

7 – A partir da estatística teste verifique se H_0 é rejeitada ou não, para tanto e utilize um método:
Método tradicional: Compare a estatística teste com o valor crítico, se a estatística teste está na região crítica, Rejeite H_0.
Método valor-P: a partir da estatística teste determine o valor-P. Se valor-P < α, Rejeite H_0.

8 – Estabeleça a conclusão (fraseado final abaixo).

Fraseado Final para o Teste de Hipótese:



1 TESTE DE HIPÓTESE - uma amostra

Inferência sobre uma proporção: $H_0 : p = p_1$

Estatística teste: $z_{teste} = \dfrac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$,

sendo $q = 1 - p$.

Inferência sobre uma média: Desvio padrão Conhecido $H_0 : \mu = \mu_1$

Estatística teste: $z_{teste} = \dfrac{\bar{x} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$

Inferência sobre uma média: Desvio padrão Desconhecido $H_0 : \mu = \mu_1$

Estatística teste: $t_{teste} = \dfrac{\bar{x} - \mu}{\left(\dfrac{s}{\sqrt{n}}\right)}$,

grau de liberdade: $gl = (n-1)$

Inferência sobre um desvio padrão ou uma variância: $H_0 : \sigma = \sigma_1$ **ou** $H_0 : \sigma^2 = \sigma_1^2$

Estatística teste: $\chi^2_{teste} = \dfrac{(n-1)s^2}{\sigma^2}$,

grau de liberdade: $gl = (n-1)$

2 TESTE DE HIPÓTESE - duas amostras

Inferências sobre duas proporções:

$H_0 : p_1 = p_2$

Estatística teste: $z_{teste} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\bar{p}\,\bar{q}}{n_1} + \dfrac{\bar{p}\,\bar{q}}{n_2}}}$,

sendo $\hat{p}_i = \dfrac{x_i}{n_i}$, $\bar{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ e $\bar{q} = 1 - \bar{p}$

Inferências sobre duas médias: Amostras Independentes $H_0 : \mu_1 = \mu_2$

Estatística teste: $t_{teste} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$,

grau de liberdade:

$gl =$ escolha o menor entre $(n_1 - 1)$ e $(n_2 - 1)$

Inferências sobre duas médias: Amostras Emparelhadas $H_0 : \mu_d = 0$

Estatística teste: $t_{teste} = \dfrac{\bar{d}}{\left(\dfrac{s_d}{\sqrt{n}}\right)}$,

sendo $\bar{d}$ a média de todas as diferenças amostrais $d_i = x_{i1} - x_{i2}$; e s_d é o desvio padrão das diferenças amostrais, ou seja para amostras 1 e 2:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{e} \quad s_d = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}}$$

grau de liberdade: $gl = (n-1)$

Comparação da variação em duas amostras: Teste F para comparação de variância

$H_0 : \sigma_1^2 = \sigma_2^2$

Estatística teste: $F_{teste} = \dfrac{s_1^2}{s_2^2}$,

sendo s_1^2 a maior das duas variâncias amostrais.

- grau de liberdade do numerador:

$gl_1 = n_1 - 1$

- grau de liberdade do denominador:

$gl_2 = n_2 - 1$

3 ANOVA (ANalysis Of VAriance) - Inferência a partir de mais de duas amostras

Hipóteses: $H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : pelo menos uma das médias é diferente das outras

- **ANOVA de um fator** -

  - **Cálculos com tamanhos iguais** -

    Variância entre amostras = $n\,s_{\bar{x}}^2$, sendo $s_{\bar{x}}^2$ a variância das médias amostrais.

    Variância dentro das amostras = s_p^2, sendo s_p^2 a média das variâncias amostrais.

    Estatística Teste: $F_{teste} = \dfrac{n\,s_{\bar{x}}^2}{s_p^2}$

    - grau de lib. do numerador: $gl_{Num} = k - 1$

    - grau de lib. do denominador: $gl_{Den} = k\,(n-1)$

	A	B	C	D	E	F
1						
2		**Tam 1**	**Tam 2**	**Tam 3**		
3		79945,00	65040,00	53685,00		
4		51490,00	67750,00	60975,00		
5		89430,00	69105,00	90785,00		
6		65040,00	81300,00	82655,00		
7	n_i	4	4	4	=CONT.VALORES(D3:D6)	
8	$\bar{x}_i$	71476,25	70798,75	72025,00	=MÉDIA(D3:D6)	
9	s_i^2	278310790	51867706,25	307797533,3	=VAR(D3:D6)	
10						
11	$s_{\bar{x}}^2$	377303,646	=VAR(B8:FD8)			
12						
13	s_p^2	212658676	=MÉDIA(B9:D9)		n =	4
14						
15		Variância entre amostras:		1509214,583	=F13*B11	= $n\,s_{\bar{x}}^2$
16		Variância dentro da am.:		212658676	=B13	= s_p^2
17						
18		Estatística F =	0,0071	=D15/D16	= $F_{teste} = \dfrac{n s_{\bar{x}}^2}{s_p^2}$	
19						

  - **Cálculos com tamanhos amostrais diferentes** -

    Estatística Teste: $F_{teste} = \dfrac{\left[\dfrac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\dfrac{\sum (n_i - 1)s_i^2}{N-k}\right]}$, sendo $\bar{\bar{x}}$ a média de todos os valores amostrais e

    - grau de liberdade do numerador: $gl_{Num} = k - 1$; grau de liberdade do denominador: $gl_{Den} = N - k$

Exemplo para ANOVA de um fator com tamanhos amostrais diferentes:

ANOVA.xls

	A	B	C	D	E	F	G	H	I	J
4			PL (em R$ mil) de Empresas							
5			Tam 1	Tam 2	Tam 3		i	$n_i(\bar{x}_i - \bar{\bar{x}})^2$	$(n_i - 1)s_i^2$	
6			79945	84010	92140		1	4,51E+08	2,89E+08	
7			67750	67750	73170		2	3,01E+07	2,08E+08	
8			56910	69105	90785		3	5,11E+08	7,00E+08	
9			65040	81300	82655		Soma	9,92E+08	1,20E+09	
10			71815	74525	78590		k -1 =	2		
11					105690		N-k =	14		
12					81300		Variância Entre		4,96E+08	=H9/H10
13		n_i	5	5	7		Variância Dentro		8,55E+07	=I9/H11
14		$\bar{x}_i$	68292,00	75338,00	86332,86					
15		s_i^2	7,22E+07	5,20E+07	1,17E+08			Estatística teste =	5,80218	=I12/I13
16		N	17	=SOMA(C13:E13)						
17		k	3	=CONT.VALORES(C13:E13)				valor-P =	0,01461	
18		$\bar{\bar{x}}$	77792,94	=MÉDIA(C6:E12)						

4 CORRELAÇÃO

Considere pares de dados (y,x)

y	x
y_1	x_1
y_2	x_2
...	...
y_n	x_n

Coeficiente de Correlação Linear de Pearson:

$$r = \frac{SQ_{xy}}{\sqrt{SQ_{xx}}\sqrt{SQ_{yy}}}$$

sendo

$$SQ_{wz} = \left(\sum wz\right) - \frac{\left(\sum w\right)\left(\sum z\right)}{n} \qquad \text{e}$$

$$SQ_{zz} = \left(\sum z^2\right) - \frac{\left(\sum z\right)^2}{n}$$

O teste de hipótese:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$ gl = $n-2$

$t_{teste} = r\sqrt{\frac{n-2}{1-r^2}}$. **Ou** valor crítico $r_\alpha = \frac{t_\alpha}{\sqrt{t_\alpha^2 + n - 2}}$

Na HP:
Digite os dados:
y , [enter] , *x* , Σ+
r: [g], 2, [xy]

Registros	Valor
R_1	n
R_2	Σ x
R_3	Σ x^2
R_4	Σ y
R_5	Σ y^2
R_6	Σ xy

No Excel: Ferramentas
> Análise de Dados
> Correlação

5 REGRESSÃO LINEAR SIMPLES

- **Equação da reta:** $\hat{y}_i = b_0 + b_1 x_i$ sendo

$$b_1 = \frac{SQ_{xy}}{SQ_{xx}} \qquad \text{e} \qquad b_0 = \bar{y} - b_1\bar{x}$$

- **Para avaliar o ajuste:**

$$SQ_{Tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\sum y^2\right) - \frac{\left(\sum y\right)^2}{n},$$

$$SQ_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b_0\left(\sum y\right) + b_1\left(\sum xy\right) - \frac{\left(\sum y\right)^2}{n}$$

$$SQ_{Erro} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \left(\sum y^2\right) - b_0\left(\sum y\right) - b_1\left(\sum xy\right)$$

$$R^2 = 1 - \frac{SQ_{Erro}}{SQ_{Tot}}$$

Significado do R^2: proporção das variações de y que são explicadas pelas variações de x.

Erro Padrão: $S_{xy} = \sqrt{\frac{SQ_{Erro}}{n-2}}$ $\qquad S_{b_1} = \frac{S_{xy}}{\sqrt{SQ_{xx}}}$

$$S_{b_0} = S_{xy}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SQ_{xx}}}$$

Teste de Hipótese

$H_0 : \beta_i = 0$ $\qquad t_{teste} = \dfrac{b_i}{S_{b_i}}$

$H_1 : \beta_i \neq 0$ $\qquad gl = n - 2$

No Excel: > Ferramentas > Análise de Dados
> Regressão

Na HP:

b_0 : 0, [g], 2

b_1 : 1, [g], 2, [x y], R↓, [xy], [-]

5.1 Intervalo de confiança na regressão

Para obter o erro deve-se calcular hi em cada valor de xi.

$$\mathrm{h_i} = \frac{1}{\mathrm{n}} + \frac{(\mathrm{X_i} - \overline{\mathrm{X}})^2}{\mathrm{SSX}} = \frac{1}{\mathrm{n}} + \frac{(\mathrm{X_i} - \overline{\mathrm{X}})^2}{\sum(\mathrm{X_i} - \overline{\mathrm{X}})^2}$$

hi será utilizado para o cálculo do intervalo de confiança, como visto nas expressões seguintes

5.1.1 Estimativas para média

Intervalo de Confiança para $\mu_{\mathrm{Y|X=X_i}}$:

$$\hat{Y} \pm t_{\alpha/2} \mathrm{S_{YX}} \sqrt{h_i}$$

5.1 Estimativas para valor

Intervalo de Confiança para $\mathrm{Y_{X=X_i}}$:

$$\hat{Y} \pm t_{\alpha/2} \mathrm{S_{YX}} \sqrt{1 + h_i}$$

6 MATRIZ DE CORRELAÇÃO

A matriz de correlação é escrita na forma:

$$\mathbf{R} = \begin{bmatrix} 1 & \mathrm{r}_{12} & \cdots & \mathrm{r}_{1k} \\ \mathrm{r}_{21} & 1 & \cdots & \mathrm{r}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{r}_{k1} & \mathrm{r}_{k2} & \cdots & 1 \end{bmatrix}. \qquad (1)$$

O cálculo de cada elemento desta matriz pode ser obtido como mostrado na seção 4.

A matriz R é simétrica, ou seja, $\mathrm{r_{ij}} = \mathrm{r_{ji}}$.

7 MATRIZ DE COVARIÂNCIA (a partir da correlação)

O cálculo de cada elemento da matriz de covariância pode ser feito a partir da matriz de correlação. Assim, a partir da correlação r_{12}, entre duas variáveis x_1 e x_2, a covariância é dada por:

$$\mathrm{S}_{ij} = \mathrm{r}_{ij} s_i s_j \qquad (2)$$

Na expressão (2) s_i e s_j são os desvios-padrão das variáveis i e j, respectivamente.

8 REGRESSÃO LINEAR MÚLTIPLA

Considere mais de duas variáveis (Ex: y,x1,x2)

y	x_1	x_2
y_1	x_{11}	x_{12}
y_2	x_{21}	x_{22}
...	...	...
y_n	x_{n1}	x_{n2}

O objetivo é obter o modelo de regressão linear múltipla, considerando k variáveis independentes:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Obter o modelo significa encontrar os valores de b_0, b_1, b_2, ..., b_k.

Etapas para o estudo de regressão linear múltipla:

A resposta do problema (obtenção de b's) pode ser dada através de matrizes:

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \boldsymbol{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Y}) \qquad (3)$$

A seguir os passos para obtenção dos b's:

- Gráficos de dispersão $y - x_1$, $y - x_2$, ..., $y - x_k$
- A partir da tabela de dados considere as matrizes
- $\mathbf{Z}$ e $\mathbf{Y}$ dadas por:

$$\mathbf{Z} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \text{e} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Obtenha $\mathbf{Z}'$, ou seja, a transposta da matriz $\mathbf{Z}$:

$$\mathbf{Z}' = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix}$$

- Multiplique a matriz transposta $\mathbf{Z}'$ pela matriz $\mathbf{Z}$:

$$\mathbf{Z}'\mathbf{Z}$$

- Considere o resultado anterior e inverta obtendo a matriz

$$\mathbf{M} = (\mathbf{Z}'\mathbf{Z})^{-1} \qquad (4)$$

Para inverter a matriz veja: 9 MATRIZ INVERSA.

- Faça a multiplicação das matrizes $\mathbf{Z}'$ e $\mathbf{Y}$ obtendo a matriz $\mathbf{A}$ dada por

$$\mathbf{A} = \mathbf{Z}'\mathbf{Y} \qquad (5)$$

- Multiplique as matrizes obtidas em (4) e (5) obtendo assim os valores de b's definidos em (3), ou seja:

$$\boldsymbol{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Y})$$

- Com os valores de b's escreva a equação linear:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \qquad (6)$$

- A partir da expressão (6) calcule estimativas para cada caso em análise.
- Calcule os resíduos ε_i para cada estimativa:

$$\varepsilon_i = y_i - \hat{y}_i$$

- Faça um gráfico dos resíduos (no eixo-y) como função das estimativas (no eixo-x).

9 MATRIZ INVERSA

A matriz inversa de uma matriz $\mathbf{A}$ (se existir) é a matriz $\mathbf{A}^{-1}$ tal que $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\,\mathbf{A} = \mathbf{1}$.

Determinação da Matriz Inversa pela matriz Adjunta:

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\operatorname{adj}(\mathbf{A}) \qquad (7)$$

Para uma matriz **3x3**, considere os elementos nomeados de acordo com a expressão (8) abaixo:

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \qquad (8)$$

A matriz adjunta de $\mathbf{A}$ é obtida pela matriz dos cofatores transposta, que no caso da matriz $\mathbf{A}$ da expressão (8) resulta em:

$$\operatorname{adj}(\mathbf{A}) = \begin{bmatrix} +\begin{vmatrix} e & f \\ h & i \end{vmatrix} & -\begin{vmatrix} b & c \\ h & i \end{vmatrix} & +\begin{vmatrix} b & c \\ e & f \end{vmatrix} \\ -\begin{vmatrix} d & f \\ g & i \end{vmatrix} & +\begin{vmatrix} a & c \\ g & i \end{vmatrix} & -\begin{vmatrix} a & c \\ d & f \end{vmatrix} \\ +\begin{vmatrix} d & e \\ g & h \end{vmatrix} & -\begin{vmatrix} a & b \\ g & h \end{vmatrix} & +\begin{vmatrix} a & b \\ d & e \end{vmatrix} \end{bmatrix}$$

O determinante de $\mathbf{A}$ é obtido por:

$$\det(\mathbf{A}) = a\,e\,i + b\,f\,g + d\,h\,c - g\,e\,c - a\,h\,f - b\,d\,i$$