4

Distribuição Normal Multivariada

4.1. Introdução

A generalização da densidade normal univariada para duas ou mais dimensões desempenha um papel fundamental na análise multivariada. De fato, a maioria das técnicas multivariadas parte do pressuposto de que os dados foram gerados de uma distribuição normal multivariada. Apesar dos dados originais não serem quase nunca “exatamente” normal multivariados, a densidade normal se constitui muitas vezes numa aproximação adequada e útil da verdadeira distribuição populacional.

A distribuição normal, além da sua atratividade pela sua facilidade de tratamento matemático, possui duas razões práticas que justificam a sua utilidade. A primeira, diz que a distribuição normal é a mais adequada para modelos populacionais em várias situações; e a segunda refere-se ao fato da distribuição amostral de muitas estatísticas multivariadas ser aproximadamente normal, independentemente da forma da distribuição da população original, devido ao efeito do limite central.

4.2. Pressuposições das análises multivariada

É importante compreender que as análises estatísticas de modelos com erros aditivos baseiam-se na pressuposição de normalidade. A distribuição normal requerida refere-se, não a variação dos dados, mas a variação residual entre as observações e o modelo ajustado. A variação sistemática dos dados deve-se presumidamente aos efeitos fixos dos modelos e o restante da variação aleatória é devida à pequenas influências independentes, as quais produzem resíduos com distribuição normal (Bock, 1975).

Um segundo ponto, muitas vezes negligenciado nas discussões das pressuposições sobre a distribuição, refere-se ao fato de que as afirmações probabilísticas dos testes de significância e dos intervalos de confiança, dizem respeito a estatísticas tais como médias amostrais ou diferenças entre médias, e não a distribuição das observações individuais. É conhecido que a distribuição destas estatísticas torna-se tipicamente normal quando a amostra aumenta de tamanho. Este resultado se deve ao teorema do limite central.

Do ponto de vista prático existe consideráveis vantagens de se trabalhar com grandes amostras. Nestes casos, a violação da pressuposição de que a população seja normal é menos crítica para os testes estatísticos e intervalos de confiança, e a precisão da estimação de parâmetros desconhecidos é “melhor”.

4.3. Densidade normal multivariada e suas propriedades

A densidade normal multivariada é uma generalização da densidade normal univariada. Para a distribuição normal univariada com média μ e variância σ^2, a função de densidade de probabilidade é bem conhecida e é dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad x \in]-\infty, +\infty[\qquad (4.1)$$

O gráfico da função (4.1) tem a forma de sino e está apresentado na Figura 4.1. As probabilidades são áreas sob a curva entre dois valores da variável **X**, limitada pela abcissa. É bem conhecido o fato de que as áreas entre ± 1 desvio padrão da média e ± 2 desvios padrões da média são respectivamente 68,3% e 95,4%, como ilustrado na Figura 4.1.

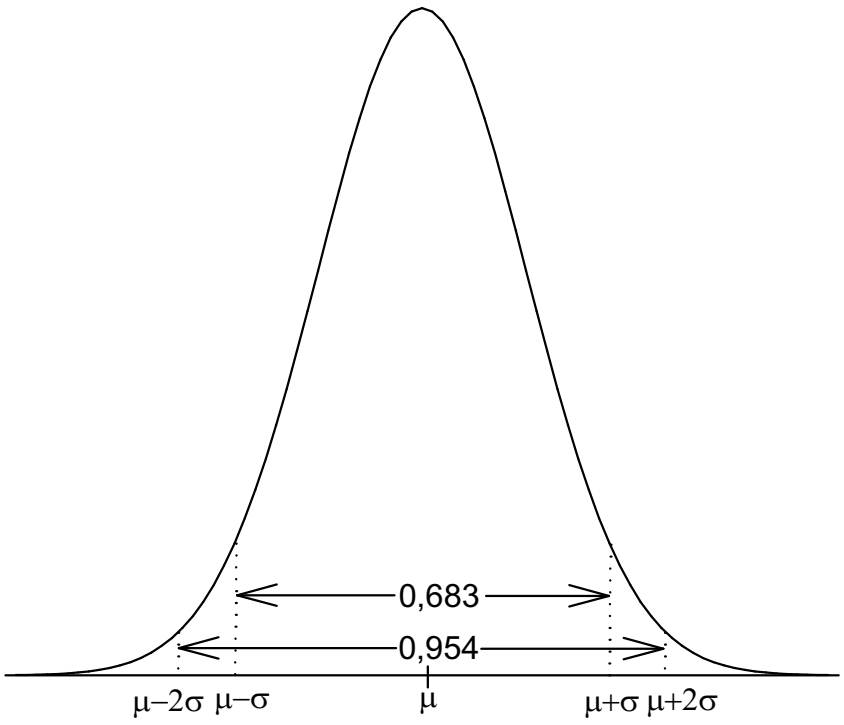


Figura 4.1. Densidade normal univariada com média μ e variância σ^2, destacando-se as áreas entre $\mu \pm \sigma$ **e** $\mu \pm$ **2**σ.

O expoente da função de densidade normal univariada:

$$\frac{(\mathbf{x}-\mu)^2}{\sigma^2} = (\mathbf{x}-\mu)(\sigma^2)^{-1}(\mathbf{x}-\mu) \qquad (4.2)$$

mede a distância quadrada de x em relação à μ em unidade de desvio padrão. Esta distância pode ser generalizada para o caso multivariado, com um vetor $\underset{\sim}{X}$ de observações (px1), dada por,

$$\left(\underset{\sim}{X}-\underset{\sim}{\mu}\right)'(\Sigma)^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu}) \qquad (4.3)$$

Nesta expressão (4.3) o vetor $\underset{\sim}{\mu}$ (px1) representa o valor esperado do vetor $\underset{\sim}{X}$ e a matriz Σ (pxp) representa a sua covariância. Então, (4.3) representa a distância generalizada de $\underset{\sim}{X}$ para $\underset{\sim}{\mu}$.

Substituindo a expressão (4.3) na função de densidade (4.1), a constante univariada de normalização $\sqrt{2\pi\sigma^2}$ deve ser trocada de modo a fazer com que o volume sob a superfície da função de densidade multivariada obtida, seja igual a unidade para qualquer p. Pode-se demonstrar (Anderson, 1984) que esta constante é $(2\pi)^{-p/2}\left|\Sigma\right|^{-1/2}$, sendo a densidade dada por:

$$f(\underset{\sim}{X}) = \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} e^{-(1/2)(\underset{\sim}{x}-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})} \qquad (4.4)$$

Propriedades da distribuição normal multivariada

Seja um vetor $\underset{\sim}{X}$ tendo distribuição normal multivariada, então:

1. Combinações lineares dos componentes de $\underset{\sim}{X}$ serão normalmente distribuídos: seja a combinação linear $\underset{\sim}{a}'\underset{\sim}{X} = a_1\mathbf{X_1} + a_2\mathbf{X_2} + ... + a_p\mathbf{X_p}$, então, $\underset{\sim}{a}'\underset{\sim}{X}$ terá distribuição $N(\underset{\sim}{a}'\underset{\sim}{\mu}, \underset{\sim}{a}'\Sigma\underset{\sim}{a})$;

2. Todos os subconjuntos de $\underset{\sim}{X}$ tem distribuição normal (multivariada). Pelos resultados da propriedade 1, fazendo alguns a_i's iguais a zero, isto se torna evidente;

i) Fazendo $\underset{\sim}{a}'\underset{\sim}{X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = X_1$ a propriedade 2 se torna evidente. Assim, $X_1 \sim N(\underset{\sim}{a}'\underset{\sim}{\mu} = \mu_1, \underset{\sim}{a}'\Sigma\underset{\sim}{a} = \sigma_{11})$. De uma forma mais geral pode-se afirmar que todo componente X_i tem distribuição $N(\mu_i, \sigma_{ii})$.

ii) A distribuição de várias combinações lineares é:

$$_qA_p \; _p\underset{\sim}{X}_1 = \begin{bmatrix} a_{11}X_1 + & ... & a_{1p}X_p \\ \vdots & \ddots & \vdots \\ a_{q1}X_1 + & ... & a_{qp}X_p \end{bmatrix} \sim N_q\left(A\mu; A\Sigma A'\right)$$

iii) Todos os subconjuntos de $\underset{\sim}{X}$ tem distribuição normal (multivariada)

Tomando-se uma partição: $_{p}\underset{\sim}{X}_{1}=\begin{bmatrix}\frac{_{q}\underset{\sim}{X}_{1}}{_{(p-q)}\underset{\sim}{X}_{1}}\end{bmatrix}=\begin{bmatrix}\frac{\underset{\sim}{X}_{1}}{\underset{\sim}{X}_{2}}\end{bmatrix}$ e suas correspondentes partições no vetor de média e de covariância, dadas por:

$$_{p}\underset{\sim}{\mu}_{1}=\begin{bmatrix}\frac{_{q}\underset{\sim}{\mu}_{1}}{_{(p-q)}\underset{\sim}{\mu}_{1}}\end{bmatrix}=\begin{bmatrix}\frac{\underset{\sim}{\mu}_{1}}{\underset{\sim}{\mu}_{2}}\end{bmatrix} \text{ e } \Sigma=\begin{bmatrix}_{q}\Sigma_{11_{q}} & _{q}\Sigma_{12_{(p-q)}}\\ _{(p-q)}\Sigma_{21_{q}} & _{(p-q)}\Sigma_{22_{(p-q)}}\end{bmatrix}$$

Logo,

$$\underset{\sim}{X}_{1}\sim N_{q}\left(\underset{\sim}{\mu}_{1};\Sigma_{11}\right)$$

Prova: Basta fazer $_{q}A_{p}=[_{q}I_{q}\mid {}_{q}0_{(p-q)}]$ e aplicar (ii).

3. Componentes de covariância zero entre dois subconjuntos de $\underset{\sim}{X}$ implica em dizer que eles são independentemente distribuídos. Esta propriedade só é valida se $\underset{\sim}{X}$ tiver distribuição normal multivariada; e

4. A distribuição condicional de componentes de $\underset{\sim}{X}$ é normal (multivariada).

Dada a partição $_{p}\underset{\sim}{X}_{1}=\begin{bmatrix}\frac{_{q}\underset{\sim}{X}_{1}}{_{(p-q)}\underset{\sim}{X}_{1}}\end{bmatrix}=\begin{bmatrix}\frac{\underset{\sim}{X}_{1}}{\underset{\sim}{X}_{2}}\end{bmatrix}$, logo a distribuição condicional de $\underset{\sim}{X}_{1}/\underset{\sim}{X}_{2}=\underset{\sim}{x}_{2}$ é normal e têm média e covariância dados por:

$$\underset{\sim}{\mu}_{c}=\underset{\sim}{\mu}_{1}+\Sigma_{12}\Sigma_{22}^{-1}\left(\underset{\sim}{x}_{2}-\underset{\sim}{\mu}_{2}\right) \text{ e } \Sigma_{c}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

4.4. Distribuição normal bivariada

Sejam $\mathbf{X}_1$ e $\mathbf{X}_2$ duas variáveis com parâmetros $E(\mathbf{X}_1)=\mu_1$, $E(\mathbf{X}_2)=\mu_2$, $Var(\mathbf{X}_1)=\sigma_{11}$, $Var(\mathbf{X}_2)=\sigma_{22}$ e $\rho_{12}=\frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}=\mathrm{Corr}(X_1,X_2)$.

A matriz de covariância é

$$\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}$$

Cuja inversa é,

$$\Sigma^{-1}=\frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\begin{bmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{21} & \sigma_{11}\end{bmatrix}$$

Fazendo $\sigma_{12}=\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$, obtém-se $|\Sigma|=\sigma_{11}\sigma_{22}-\sigma_{12}^2=\sigma_{11}\sigma_{22}(1-\rho_{12}^2)$, e a distância generalizada de (4.3) será:

$$\frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}\begin{bmatrix}X_1-\mu_1 & X_2-\mu_2\end{bmatrix}\begin{bmatrix}\sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11}\end{bmatrix}\begin{bmatrix}X_1-\mu_1\\ X_2-\mu_2\end{bmatrix}= \tag{4.5}$$

$$=\frac{1}{1-\rho_{12}^2}\left[\left(\frac{X_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{X_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{X_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{X_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]$$

Desde que, $|\Sigma|=\sigma_{11}\,\sigma_{22} - (\sigma_{12})^2 = \sigma_{11}\,\sigma_{22}\,(1-\rho_{12}^2)$, pode-se substituir Σ^{-1} e $|\Sigma|$ em (4.4) para se ter a expressão da densidade normal bivariada, apresentada a seguir.

$$f(x_1,x_2)=\frac{1}{2\pi\sqrt{\sigma_{11}\,\sigma_{22}(1-\rho_{12}^2)}}$$

$$\exp\left\{\frac{-1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\} \qquad (4.6)$$

Se X_1 e X_2 não são correlacionadas, $\rho_{12}=0$, a densidade conjunta pode ser escrita como produto das densidades normais univariadas, ambas com a forma de (4.1), ou seja, $f(x_1,x_2)= f(x_1)\,f(x_2)$, além do que X_1 e X_2 são ditas independentes, como comentado na propriedade número 3 da seção 4.3. Duas distribuições normais bivariadas com variâncias iguais são mostradas nas Figuras 4.2. e 4.3. A Figura 4.2 mostra o caso em que X_1 e X_2 são independentes ($\rho_{12}=0$) e a Figura 4.3 o caso de $\rho_{12}=0.8$. Observa-se que a presença de correlação faz com que as probabilidades se concentrem ao longo de uma linha.

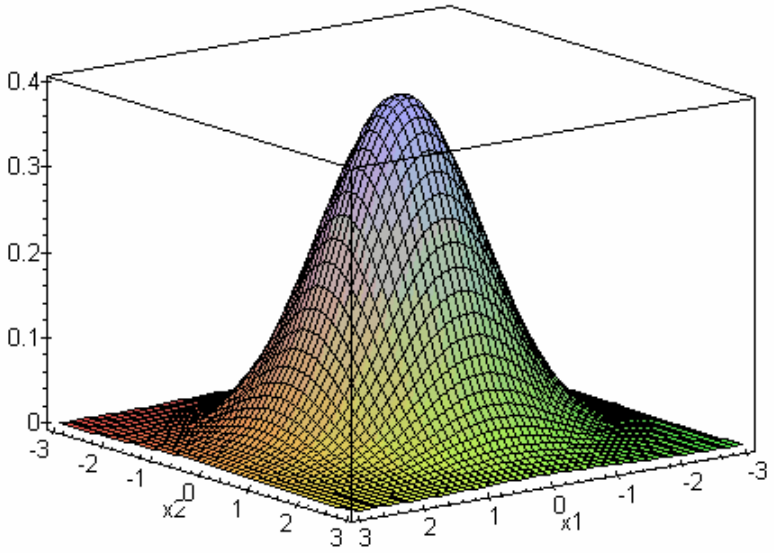
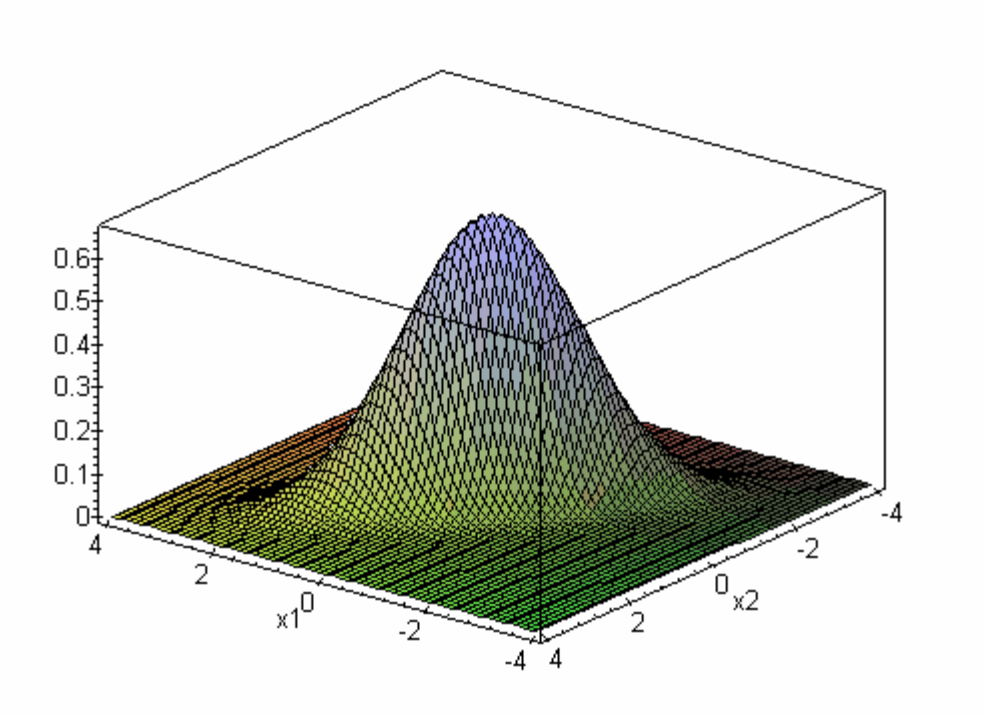


Figura 4.2. Distribuição normal bivariada com $\sigma_{11} = \sigma_{22}$ e ρ_{12}=0.

Figura 4.3. Distribuição normal bivariada com $\sigma_{11} = \sigma_{22}$ e ρ_{12}=0.8.



Da análise da expressão (4.4), relativa à densidade de p-variáveis normais, fica claro que alguns valores padrões de $\underset{\sim}{X}$ fornecem alturas constantes para as densidades elipsóides. Isto significa que a densidade normal é constante em superfícies cujas distâncias quadráticas $(\underset{\sim}{X}-\underset{\sim}{\mu})'(\Sigma)^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu})$ são constantes. Esses padrões são chamados de contornos ou curvas de nível.

Contornos={todo $\underset{\sim}{X}$ tal que $(\underset{\sim}{X}-\underset{\sim}{\mu})'(\Sigma)^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu})=c^2$} (4.7)

A expressão (4.7) é uma superfície de uma elipsóide centrada em $\underset{\sim}{\mu}$, cujos eixos possuem direção dos autovetores de Σ^{-1} e seus comprimentos são proporcionais ao recíproco da raiz quadrada dos seus autovalores. Demonstra-se que se λ_i e $\underset{\sim}{e}_i$ são os autovalores e autovetores, respectivamente, de Σ, então a elipsóide $(\underset{\sim}{X}-\underset{\sim}{\mu})'(\Sigma)^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu})=c^2$ é centrada em $\underset{\sim}{\mu}$ e tem eixos na direção de $\pm c\sqrt{\lambda_i}\,\underset{\sim}{e}_i$ (i=1, 2, ..., p).

Considerando como ilustração a densidade normal bivariada com $\sigma_{11}=\sigma_{22}$, os eixos da elipsóide dados por (4.7) são fornecidos pelos autovalores e autovetores de Σ. Portanto para obtê-los a equação $|\Sigma-\lambda\mathbf{I}|=0$ deve ser resolvida.

$$\begin{vmatrix}\sigma_{11}-\lambda & \sigma_{12}\\ \sigma_{12} & \sigma_{11}-\lambda\end{vmatrix}=(\sigma_{11}-\lambda)^2-\sigma_{12}^2=0$$

$$=(\lambda-\sigma_{11}-\sigma_{12})(\lambda-\sigma_{11}+\sigma_{12})$$

Conseqüentemente os autovalores são:

$$\lambda_1 = \sigma_{11} + \sigma_{12} \quad \textbf{e} \quad \lambda_2 = \sigma_{11} - \sigma_{12}$$

Os autovetores são determinados por:

$$\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$$

Para i=1, tem-se:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = (\sigma_{11} + \sigma_{12}) \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

ou,

$$\sigma_{11} e_1 + \sigma_{12} e_2 = (\sigma_{11} + \sigma_{12}) e_1$$
$$\sigma_{12} e_1 + \sigma_{11} e_2 = (\sigma_{11} + \sigma_{12}) e_2$$

Essas equações levam ao resultado de que $e_1=e_2$, e após normalização, o primeiro autovetor é:

$$\underset{\sim}{e}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

De forma similar foi obtido o segundo autovetor, o qual é:

$$\underset{\sim}{e}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

Se a covariância é positiva, $\lambda_1 = \sigma_{11} + \sigma_{12}$ é o maior autovalor e seu autovetor associado se posiciona ao longo de uma linha de 45^0 através do ponto $\mu' = [\mu_1 \quad \mu_2]$, para qualquer $\sigma_{12} > 0$. Os eixos são fornecidos por $\pm c\sqrt{\lambda_i}\ \underset{\sim}{e}_i$ (i=1, 2) e estão representados na Figura 4.4.

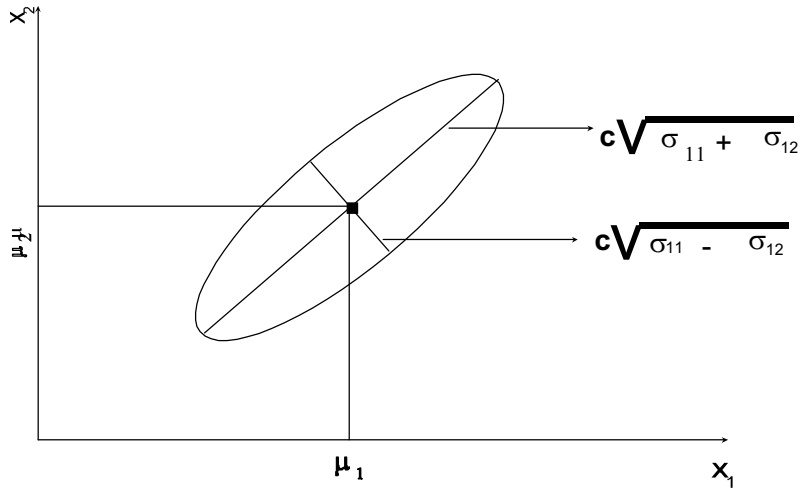


Figura 4.4. Curva de nível de densidade constante para a distribuição normal bivariada com $\sigma_{11} = \sigma_{22}$ e $\sigma_{12} > 0$.

Anderson (1984) demonstra que a escolha de $c^2 = \chi_p^2(\alpha)$, em que $\chi_p^2(\alpha)$ é o percentil (100α) superior da distribuição de Qui-quadrado com p graus de liberdade,

leva a contornos que contém (1-α)x100% de probabilidade. Para a distribuição normal multivariada (p variada), a elipsóide dos valores de $\underset{\sim}{X}$ satisfazendo,

$$(\underset{\sim}{X}-\underset{\sim}{\mu})'(\Sigma)^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu}) \leq \chi_p^2(\alpha) \qquad (4.8)$$

tem probabilidade 1-α.

Os contornos contendo 95% e 99% de probabilidade sob a densidade normal bivariada das Figuras 4.2 e 4.3, estão representados nas Figuras 4.5 e 4.6.

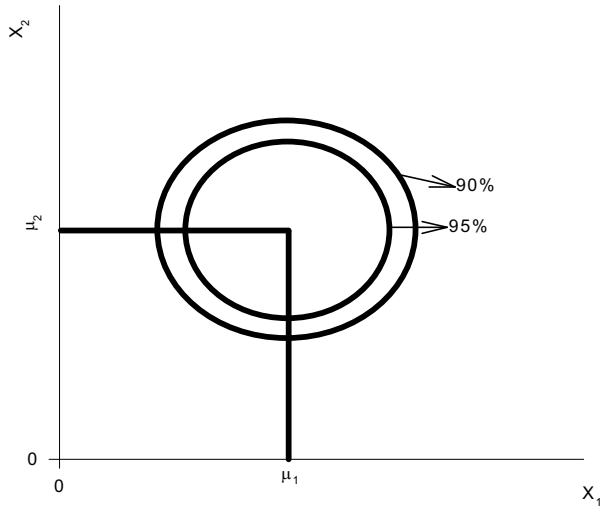


Figura 4.5. Curvas de níveis de 95% e 99% de probabilidade para a distribuição normal bivariada apresentada na Figura 4.2, $\sigma_{11} = \sigma_{22}$ e ρ_{12}=0.

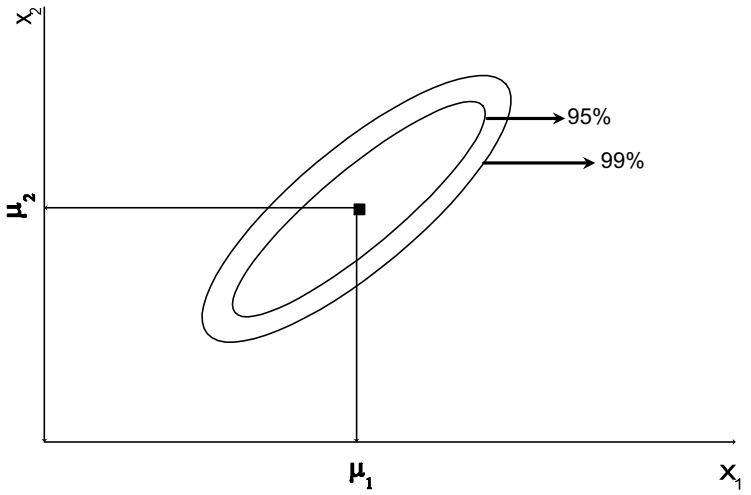


Figura 4.6. Curvas de níveis de 95% e 99% de probabilidade para a distribuição normal bivariada apresentada na Figura 4.3, $\sigma_{11} = \sigma_{22}$ e ρ_{12}=0.8.

A densidade (4.4) possui máximo quando $\underset{\sim}{X} = \underset{\sim}{\mu}$. Portanto, $\underset{\sim}{\mu}$ é o ponto de máxima densidade ou moda, bem como o valor esperado de $\underset{\sim}{X}$, ou média.

4.5. Distribuição amostral de $\overline{\underset{\sim}{X}}$ e S

Se a pressuposição de que as linhas de

$$\underset{n \times p}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \cdots & \mathbf{X}_{1p} \\ \mathbf{X}_{21} & \mathbf{X}_{22} & \cdots & \mathbf{X}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_{n1} & \mathbf{X}_{n2} & \cdots & \mathbf{X}_{np} \end{bmatrix}$$

se constituem numa amostra aleatória de uma população normal com média $\underset{\sim}{\mu}$ e covariância Σ for verdadeira**,** então este fato é suficiente para completamente definir a distribuição amostral de $\underset{\sim}{\bar{X}}$ e de **S**. Será apresentado a seguir estas distribuições amostrais, fazendo-se um paralelo com a distribuição amostral univariada que já é familiar e bem conhecida.

No caso univariado (p = 1), sabe-se que $\bar{\mathbf{X}}$ possui distribuição normal com média μ (média populacional) e variância

$$\frac{\sigma^2}{\mathbf{n}} = \frac{\textbf{Variância populacional}}{\textbf{tamanho da amostra}}$$

O resultado para o caso multivariado ($p \geq 2$) é similar a este, no sentido que $\underset{\sim}{\bar{X}}$ possui distribuição normal com média $\underset{\sim}{\mu}$ e matriz de covariância $(1/n)\Sigma$.

Para a variância amostral, caso univariado, sabe-se que a distribuição de $(n-1)S^2 \big/ \sigma^2$ possui distribuição de Qui-quadrado com n - 1 graus de liberdade. Para o caso multivariado, a distribuição da matriz de covariância é chamada de distribuição de Wishart, após sua descoberta, com (n – 1) graus de liberdade. Os resultados a seguir resumem detalhes destas distribuições:

Sendo $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ uma amostra aleatória de tamanho n de uma população normal p-variada com média $\underset{\sim}{\mu}$ e matriz de covariância Σ. Então,

1. $\overline{\underset{\sim}{X}}$ possui distribuição normal com média $\underset{\sim}{\mu}$ e matriz de covariância (1/n)Σ.
2. (n-1)**S** possui distribuição de uma matriz aleatória de Wishart com n-1 gl.
3. $\overline{\underset{\sim}{X}}$ e **S** são independentes.

Devido a Σ não ser conhecida, a distribuição de $\overline{\underset{\sim}{X}}$ não pode ser usada diretamente para se fazer inferência sobre $\underset{\sim}{\mu}$. Felizmente, **S** fornece informação independente sobre Σ e a distribuição de **S** não depende de $\underset{\sim}{\mu}$. Isto permite que se construa estatísticas para fazer inferência sobre $\underset{\sim}{\mu}$, como será abordado no capítulo 5.

Densidade da distribuição de Wishart

Seja **S** uma matriz positiva definida, com n>p, então se pode definir,

$$w_{n-1}(S/\Sigma) = \frac{|S|^{(n-p-2)/2} e^{-tr(S\Sigma^{-1})/2}}{2^{p(n-1)/2} \pi^{p(p-1)/4} |\Sigma|^{(n-1)/2} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n-i)\right)} \qquad (4.9)$$

em que, Γ(.) representa a função gama.

Retornando ao caso da distribuição das médias amostrais, o resultado 4.1, sintetiza um importante teorema em estatística.

Resultado 4.1. (teorema do limite central) Sendo $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ uma amostra aleatória de n independentes observações de uma população qualquer com média $\underset{\sim}{\mu}$ e matriz de covariância Σ**,** finita e não singular. Então,

$\sqrt{n}\left(\overline{\underset{\sim}{X}} - \underset{\sim}{\mu}\right)$ possui distribuição aproximadamente normal $N_p(\underset{\sim}{0}, \Sigma)$ para grandes amostras. Aqui n deve ser também bem maior do que p (número de variáveis).

Como já foi comentado quando n é grande, **S** converge em probabilidade para Σ, consequentemente, a substituição de Σ por **S** causa efeitos apenas negligíveis nos cálculos de probabilidades. Desta forma, utilizando a expressão (4.8), pode-se obter o importante resultado, apresentado a seguir.

Resultado 4.2. (teorema do limite central) Sendo $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ uma amostra aleatória de n independentes observações de uma população qualquer com média $\underset{\sim}{\mu}$ e matriz de covariância Σ**,** finita e não singular. Então,

$\sqrt{n}\left(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}\right)$ possui distribuição aproximadamente normal $N_p(\underset{\sim}{0}, \Sigma)$

e

$n\left(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}\right)'\Sigma^{-1}\left(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}\right)$ se distribui aproximadamente como χ^2_p para n - p grande.

Para a distribuição normal univariada, se μ e σ são conhecidos, as probabilidades sob a curva para a distribuição de $\overline{X}$, podem ser obtidos das tabelas da distribuição normal, ou da integral da função apresentada em (4.1) nos intervalos apropriados, com μ=0 e σ=1, sendo

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (4.10)$$

Alternativamente, pode-se obter a aproximação de Hasting (1955) citado por Bock (1975), com erro máximo de 10^{-6}, dada por

$$\Phi(z) \cong \begin{cases} G \text{ se } z \leq 0 \\ 1 - G \text{ se } z > 0 \end{cases} \tag{4.11}$$

em que,

$\Phi(z)$ é a probabilidade acumulada sob a curva da distribuição normal de $-\infty$ a z;

$$G = (a_1\eta + a_2\eta^2 + a_3\eta^3 + a_4\eta^4 + a_5\eta^5)\,\phi(z);$$

$$\eta = \frac{1}{1 + 0{,}2316418|z|};$$

$$\phi(z) = (2\pi)^{-1/2}\, e^{-z^2/2};$$

a_1=0,319381530

a_2=-0,356563782

a_3=1,781477937

a_4=-1,821255978

a_5=1,330274429

4.6. Distribuições amostral derivada da distribuição normal multivariada

Teoria da Distribuição das grandes amostras e distribuição exata

Na análise dos dados freqüentemente são utilizadas funções das observações chamadas estatísticas, as quais servem como estimadores dos parâmetros ou como critério para os testes de hipóteses. A importância de tais estatísticas muitas vezes depende do conhecimento da (1) distribuição assumida para as observações, (2) do método de amostragem, e (3) da natureza da função das observações. A dois tipos de teoria amostral avaliada para derivar a distribuição amostral. A teoria das grandes amostras, a qual fornece a distribuição aproximada à medida que o tamanho amostral cresce indefinidamente, e a teoria das pequenas amostras ou teoria exata, a qual é válida para qualquer tamanho amostral.

As distribuições derivadas assumindo o tamanho amostral indefinidamente grande são chamadas de distribuições assintóticas ou “limitante”. A teoria assintótica é especialmente simples, como conseqüência do teorema do limite central que demonstra que muitas estatísticas têm distribuição normal como limite. Para tais estatísticas é necessário somente obter a média e a variância para ter a distribuição assintótica.

A distribuição amostral sem considerar os argumentos da teoria assintótica, geralmente depende do tamanho da amostra e pode ser não-normal para pequenas amostras, mesmo se a forma limite for normal. Se este for o caso, algum indicativo de qual tamanho amostral é necessário para uma dada acurácia na teoria

assintótica é extremamente útil para trabalhos práticos. Como exemplo, pode citar que a distribuição de F, de razões de variâncias, com ν_1 graus de liberdade do numerador e ν_2 do denominador, se aproxima de qui-quadrado dividido por ν_1 quando ν_2 cresce sem limite.

$$\lim_{\upsilon_2 \to \infty} F(\upsilon_1, \upsilon_2) = \frac{\chi^2_{(\upsilon_1)}}{\upsilon_1}$$

Comparando as tabelas de F e qui-quadrado dividido por ν_1, pode-se concluir que ao nível de 0,05, com erro de duas unidades na segunda casas decimal, quando ν_2 for maior que 40, haverá boa concordância. Semelhantemente, ao nível de 0,01 a concordância com a mesma precisão se dá quando o valor de ν_2 excede 100.

Distribuição da soma de quadrados de n desvios normais aleatórios

Seja **Z** um vetor ν x 1 de ν observações normais N(0,1) padronizadas. A estatística

$$\chi^2_{(\nu)} = Z' Z = z_1^2 + z_2^2 + \ldots + z_\upsilon^2 \qquad (4.12)$$

é distribuída como uma variável qui-quadrado com ν graus de liberdade. Foi obtida em 1876 por Helmert e independentemente em 1900 por Karl Pearson. A função de distribuição de qui-quadrado pode ser expressa pela função gamma incompleta.

$$P(\chi^2 \le \chi / \upsilon) = \frac{1}{2^{\frac{\upsilon}{2}}\Gamma(\upsilon/2)} \int_0^{\chi} t^{(\upsilon/2)-1} e^{-t/2} dt \qquad (4.13)$$

A função de distribuição (4.13) pode ser aproximada para aplicações em computadores pela série convergente apresentada a seguir.

$$P(\chi^2 \le \chi / \upsilon) = \frac{e^{-\chi}}{\chi^{-\upsilon}} \sum_{n=0}^{\infty} \frac{\chi^n}{\Gamma(\upsilon+n+1)} \qquad (4.14)$$

quando $\frac{1}{2}\chi < \max(\frac{1}{2}\upsilon, 13)$, e caso contrário pela expansão assintótica:

$$P(\chi^2 \le \chi / \upsilon) \approx \chi^{\upsilon-1} e^{-\chi} \left[1 + \frac{\upsilon-1}{\chi} + \frac{(\upsilon-1)(\upsilon-2)}{\chi^2} + \ldots \right] \qquad (4.15)$$

Os valores de $\Gamma(a)$ pode ser obtida pela fórmula de Stirling:

$$\Gamma(a) = (a-1)! = e^{-a} a^{a-1/2} (2\pi)^{1/2} \left[1 + \frac{1}{12a} + \frac{1}{288a^2} - \frac{139}{51840a^3} - \frac{571}{2488320a^4} \right] \qquad (4.16)$$

A forma recursiva $\Gamma(a+1)=a\Gamma(a)$ e $\Gamma(2)=\Gamma(1)$ pode ser usada quando "a" for pequeno. Sabe-se que a média da distribuição de qui-quadrado, E(χ^2), é ν e que sua variância é 2ν. Para $\nu>30$, as probabilidades podem ser obtidas usando a aproximação normal assintótica usando $\sqrt{2\chi^2}-\sqrt{2\upsilon-1}$ como um desvio normal unitário.

Razão entre independentes χ^2 (F de Fisher)

Sejam χ_1^2 e χ_2^2, dois χ^2 independentes com ν_1 e ν_2 graus de liberdade, respectivamente. Então,

$$F=\frac{\chi_1^2/\upsilon_1}{\chi_2^2/\upsilon_2}$$

possui distribuição de uma variável F com ν_1 e ν_2 graus de liberdade. A distribuição de F foi derivada por R. A. Fisher (1924). A função de distribuição de F pode ser aproximada pela série convergente da função beta incompleta:

$$I_x(a,b)=\frac{x^a(1-x)^b}{aB(a,b)}\left[1+\sum_{n=0}^{\infty}\frac{B(a+1,n+1)}{B(a+b,n+1)}x^{n+1}\right] \qquad (4.17)$$

em que, $B(a,b)=\dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

Então,

$$P(F,\upsilon_1,\upsilon_2) = 1 - I_x(\frac{\upsilon_2}{2},\frac{\upsilon_1}{2})$$

em que, $x = \frac{\upsilon_2}{\upsilon_2 + \upsilon_1 F}$

4.7. Verificando a normalidade

A pressuposição de que cada vetor de observação $\underset{\sim}{X}_j$ veio de uma distribuição normal multivariada será requerida nas técnicas estatísticas que serão abordadas nos capítulos subsequentes. Por outro lado, nas situações em que a amostra é grande e as técnicas dependem apenas do comportamento de $\underset{\sim}{\overline{X}}$, ou distâncias envolvendo $\underset{\sim}{\overline{X}}$ da forma $n\left(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}\right)' S^{-1}\left(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}\right)$, a pressuposição de normalidade das observações individuais $\underset{\sim}{X}_j$ é menos crucial. Isto devido à aproximação da distribuição normal assintótica das principais estatísticas. No entanto, melhor será a qualidade da inferência quanto mais próxima à população parental se assemelhar da forma da distribuição normal multivariada. É imperativo que exista procedimentos para detectar os caso onde os dados exibam desvios de moderados a extremos do esperado sob normalidade multivariada.

Baseado na distribuição normal, sabe-se que todas as combinações lineares de variáveis normais são normais e que contornos da densidade normal são elipsóides. Devido às dificuldades de avaliação de um teste conjunto em todas as dimensões, os testes para checar a normalidade serão concentrados em uma ou duas dimensões. Obviamente se paga um preço por estas simplificações, como não revelar algumas características que só podem ser observadas em dimensões maiores. É possível por exemplo, construir uma distribuição não normal bivariada com marginais normais. No entanto, muitos tipos de não normalidade são revelados em geral nas distribuições marginais, e para aplicações práticas será suficiente checar a normalidade em uma ou duas dimensões.

Verificando a validade da normalidade por meio da distribuição marginal

Textos elementares muitas vezes recomendam que a normalidade univariada seja investigada, examinando o histograma de freqüência amostral para avaliar discrepâncias entre as freqüências observadas e esperadas pelo ajuste da distribuição normal. Usualmente, sugere-se também que as discrepâncias sejam submetidas ao teste de aderência de qui-quadrado. Um χ^2 significativo ($P<0,05$) é tido como evidência contra a normalidade da população.

Apesar deste método ter a virtude da simplicidade de computação e ser livre do tipo de desvios da normalidade que esteja sendo testado (curtose, assimetria, etc.), tem a desvantagem, quando aplicados a dados contínuos, de depender da

arbitrariedade da escolha dos intervalos de agrupamento dos dados. Essa escolha determina a resolução do histograma e o número de termos a ser somado para obter a estatística de χ^2. Uma escolha errada pode levar a resultados não consistentes. Se a escolha dos intervalos for muito estreitas, o histograma pode ser irregular e a acurácia do χ^2 pode ser grandemente afetada devido aos pequenos valores esperados. Se os intervalos são largos, desvios de normalidade podem ser obscurecidos tanto no histograma quanto no teste de χ^2.

Uma melhor aproximação, evitando todas essas dificuldades, é conseguida fazendo uso de métodos que não requerem agrupamento de escores. Felizmente, excelentes procedimentos gráficos e computacionais existem para este propósito.

a) Distribuição de proporções

A distribuição normal univariada possui probabilidade de 0,683 para o intervalo $\left[\mu_i - \sqrt{\sigma_{ii}};\ \mu_i + \sqrt{\sigma_{ii}}\right]$ e probabilidade de 0,954 para o intervalo $\left[\mu_i - 2\sqrt{\sigma_{ii}};\ \mu_i + 2\sqrt{\sigma_{ii}}\right]$ (Figura 4.1). Consequentemente, para grandes amostras de tamanho n, é esperado que a proporção de $\hat{P}_{i1}$ observações contidas no intervalo $\left[\overline{X}_i - \sqrt{s_{ii}};\ \overline{X}_i + \sqrt{s_{ii}}\right]$ seja de cerca de 0,683, e de forma semelhante, espera-se que a proporção $\hat{P}_{i2}$ de observações em $\left[\overline{X}_i - 2\sqrt{s_{ii}};\ \overline{X}_i + 2\sqrt{s_{ii}}\right]$ seja de cerca de 0,954. Usando a aproximação normal da distribuição de $\hat{P}_i$, então se

$$| \hat{P}_{i1} - 0{,}683 | > 3\sqrt{\frac{0{,}683 \times 0{,}317}{n}} = \frac{1{,}396}{\sqrt{n}}$$

$$| \hat{P}_{i2} - 0{,}954 | > 3\sqrt{\frac{0{,}954 \times 0{,}046}{n}} = \frac{0{,}628}{\sqrt{n}}$$

devem indicar desvios da distribuição normal para i-ésima característica (Johnson & Wichern, 1988).

b) Processos gráficos

Os gráficos são em geral úteis para avaliar desvios da normalidade. Dois processos gráficos serão considerados neste capítulo.

i) Q-Q plot

Esses gráficos são obtidos da distribuição marginal das observações de cada variável. Consiste em plotar em um plano cartesiano os percentis amostrais versus os percentis esperados pelo ajuste de uma distribuição normal. Se os pontos pertencem a uma linha reta, a pressuposição de normalidade deve ser aceita.

Sejam x_1, x_2, ..., x_n as n observações de uma variável X. Sejam $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$ essas observações ordenadas crescentemente, ou seja, $x_{(1)}$ é a menor observação e $x_{(n)}$ é a maior. Quando os $x_{(j)}$ são distintos, exatamente j observações são menores ou iguais a $x_{(j)}$ (isto é teoricamente verdadeiro quando as observações são do tipo

contínuo, o que em geral será assumido). A proporção amostral j/n é aproximada por (j-½)/n, onde ½ é usado para correção de descontinuidade.

Os percentis esperados sob normalidade são dados por ($q_{(j)}$):

$$\frac{j-½}{n} = \int_{-\infty}^{q_{(j)}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \qquad (4.18)$$

Os percentis $q_{(j)}$ podem ser obtidos, como pode ser visto em (4.18) pela inversão da função de distribuição de probabilidade da normal, em rotinas apropriadas em computadores ou através de tabelas da distribuição normal. (Tabela A.1).

Os percentis $q_{(j)}$ e $x_{(j)}$ são plotados em um sistema cartesiano com $q_{(j)}$ na abcissa e $x_{(j)}$ na ordenada. Desvios da normalidade podem ser observados pela inspeção deste tipo de gráfico, cujos pontos, quando da normalidade devem pertencer a uma linha reta (de qualquer inclinação). O exemplo 4.1, ilustrará os cálculos necessários para obtenção dos Q-Q plots.

Exemplo 4.1

Seja uma amostra (n=10) obtida de uma população normal N(3; 4) apresentada a seguir. Neste caso, a observação 4 constitui-se um “outlier”, propositadamente gerado.

{3,74; 2,91; 4,79; 8,65; 2,06; 4,59; 4,02; 0,46; 1,79; 3,30}

Dessa forma para se obter o Q-Q plot é necessário os seguintes passos:

1) ordenar a amostra: $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$ e obter os seus valores correspondentes de probabilidade acumulada (j-½)/n.

j	$x_{(j)}$	(j-½)/n	$q_{(j)}$
1	0,46	0,05	-1,645
2	1,79	0,15	-1,036
3	2,06	0,25	-0,675
4	2,91	0,35	-0,385
5	3,30	0,45	-0,126
6	3,74	0,55	0,126
7	4,02	0,65	0,385
8	4,59	0,75	0,675
9	4,79	0,85	1,036
10*	8,65	0,95	1,645

2) calcular os percentis da distribuição normal padrão.

Ex. Para a observação 1 tem-se: $\frac{j-½}{n} = \frac{1-½}{10} = 0,05 = \int_{-\infty}^{q_{(1)}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$

Portanto, $q_{(1)}$ = -1,645, e assim sucessivamente.

3) plotar $(q_{(1)}, x_{(1)})$, $(q_{(2)}, x_{(2)})$, ..., $(q_{(n)}, x_{(n)})$ e examinar os resultados

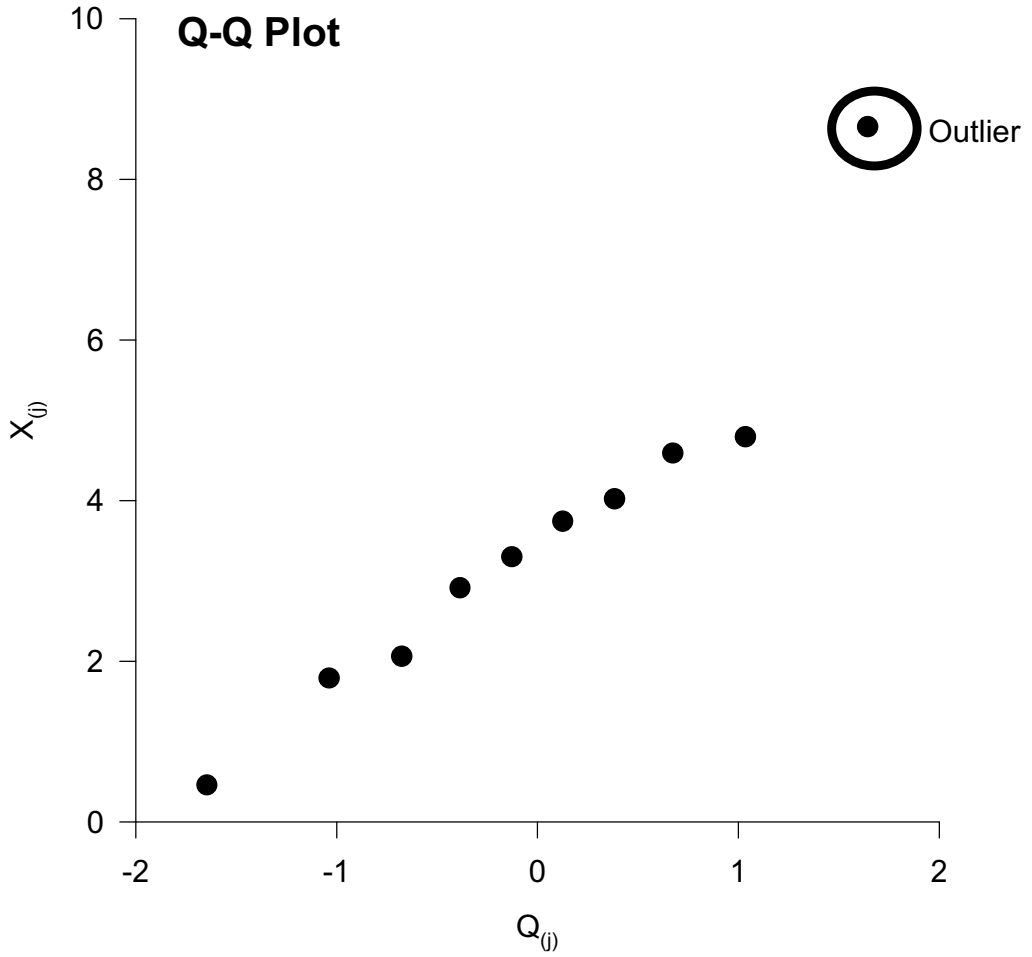


Figura 4.7. Q-Q plot para os dados do exemplo 4.1, destacando a presença de um outlier.

Observa-se que os pontos amostrais se situam praticamente em uma linha reta imaginária, com exceção da presença de um outlier, destacado na Figura 4.6. O procedimento adequado seria de eliminar esta observação e refazer a análise para os dados amostrais remanescentes, o que fica a cargo do leitor.

Este processo gráfico, embora bastante poderoso para se verificar desvios da normalidade, não se constitui num teste formal deste propósito. Para contornar esta limitação, Johnson & Wichern (1988) apresentam um teste complementar a este processo gráfico, o qual mede o ajuste dos pontos do Q-Q Plot a linha reta imaginária, através de uma medida de um coeficiente de correlação, apresentado a seguir.

$$r_Q = \frac{\sum_{j=1}^{n}\left(x_{(j)} - \bar{x}\right)\left(q_{(j)} - \bar{q}\right)}{\sqrt{\sum_{j=1}^{n}\left(x_{(j)} - \bar{x}\right)^2}\sqrt{\sum_{j=1}^{n}\left(q_{(j)} - \bar{q}\right)^2}} \qquad (4.19)$$

Um poderoso teste de normalidade pode ser construído, baseado neste coeficiente de correlação (4.19). Formalmente, rejeita-se a hipótese de normalidade se o valor calculado for menor que os valores críticos para um determinado nível de significância (Tabela 4.1).

Tabela 4.1. Valores críticos para o teste para normalidade baseado no coeficiente de correlação Q-Q plot.

Tamanho amostral	Nível de significância (α)		
n	0,01	0,05	0,10
5	0,8299	0,8788	0,9032

10	0,8801	0,9198	0,9351
15	0,9126	0,9389	0,9503
20	0,9269	0,9508	0,9604
25	0,9410	0,9591	0,9665
30	0,9479	0,9652	0,9715
40	0,9599	0,9726	0,9771
50	0,9671	0,9768	0,9809
60	0,9720	0,9801	0,9836
75	0,9771	0,9838	0,9866
100	0,9822	0,9873	0,9895
150	0,9879	0,9913	0,9928
200	0,9905	0,9931	0,9942
300	0,9935	0,9953	0,9960

Fonte: Johnson & Wichern (1998)

Exemplo 4.1 (continuação)

Calculando a correlação amostral, através de (4.19), obteve-se:

$$r_Q = \frac{18,77109}{\sqrt{44,15849}\sqrt{8,798094}} = 0,9523$$

Como, o valor tabelado ao nível de 5% de probabilidade (0,918) é inferior ao valor calculado (0,9523), então, não existe razão para duvidar da hipótese de normalidade.

ii) Gráfico das probabilidades acumuladas

Um segundo processo gráfico, bastante utilizado, refere-se aos gráficos em que são plotados as probabilidades amostrais acumuladas versus a probabilidades acumuladas da distribuição normal (Bock, 1975). O algoritmo é:

1) ordenar a amostra: $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$ e obter os seus valores correspondentes de probabilidade acumulada $p_j = (j-½)/n$, amostrais.

2) Calcular a média amostral e o desvio padrão viesado

$$S_n = \sqrt{\frac{\sum_{j=1}^{n} x_j^2 - \frac{(\sum_{j=1}^{n} x_j)^2}{n}}{n}} \qquad (4.20)$$

3) Obter as probabilidades normais acumuladas utilizando (4.11) ou tabelas da distribuição normal, através de:

$$z_j = \frac{x_j - \overline{x}}{S_n}$$

$$P_j = \Phi(z_j)$$

4) Plotar P_j (abcissa) contra p_j (na ordenada)

Exemplo 4.2

Com os dados do exemplo 4.1, o algoritmo apresentado no item (ii) foi executado, resultando nos seguintes valores:

j	$x_{(j)}$	$p_j = (j-½)/n$	P_j
1	0,46	0,05	0,066
2	1,79	0,15	0,189
3	2,06	0,25	0,227
4	2,91	0,35	0,367
5	3,30	0,45	0,436
6	3,74	0,55	0,520
7	4,02	0,65	0,575
8	4,59	0,75	0,677
9	4,79	0,85	0,709
10*	8,65	0,95	0,992

Na Figura 4.8 estão plotados os pontos P_j (abcissa) contra p_j (na ordenada).

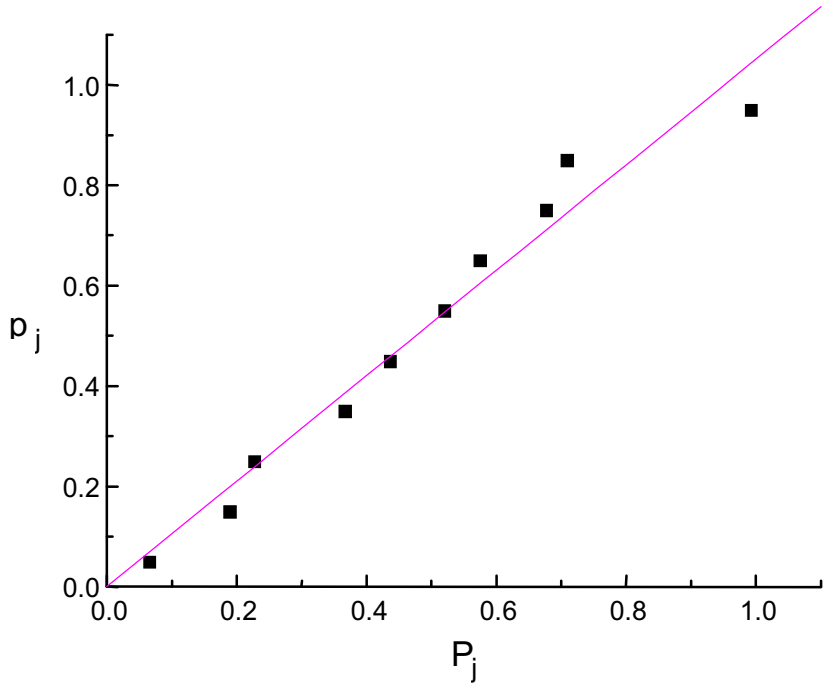


Figura 4.8. Gráfico normal acumulado da amostra simulada no exemplo 4.1.

Se a população for normal, os pontos tendem a cair em uma linha definida pela reta $P_j=p_j$. Uma vez que o gráfico apresenta efeitos cumulativos, os pontos não são independentes e ainda pode-se afirmar que sucessivos pontos não tenderão a se situar aleatoriamente em ambos os lados da linha. Em outras palavras, um grupo de pontos sucessivos poderá estar de um lado da reta ou de outro, sem ser um indicativo de desvio da normalidade. Alguma familiaridade com este tipo de gráfico, indicarão a forma da distribuição e os desvios da normalidade que possam ocorrer.

De maneira geral, as situações mais comuns devem se enquadrar nos seguintes tipos de gráficos. Distribuições assimétricas à esquerda tenderão a ter seus pontos de extremos no lado superior da reta, e os pontos intermediários no lado inferior da mesma. Para distribuições assimétricas à direita, o oposto deve ocorrer, ou seja, pontos extremos no lado inferior da reta e pontos intermediários no lado superior.

Os achatamentos da distribuição, conhecidos por curtose, também podem ser detectados. Nas distribuições leptocúrticas, os pontos de menor densidade acumulada se concentram no lado inferior da reta, vindo a cruzá-la no centro. Os pontos de maior densidade, se concentram no lado superior da reta, a partir do centro. Nas distribuições platicúrticas, o oposto se dá, ou seja, pontos de menor densidade acumulada se concentram no lado superior, e os pontos de maior densidade no lado inferior da reta, vindo a cruzá-la no centro. Distribuições bimodais possuem gráficos que representam os casos extremos da distribuição platicúrtica.

c) Uso dos momentos

Os momentos não centrados para a média, podem ser calculados a partir dos dados amostrais, fazendo 1/n como densidade para cada ponto amostral. Desta forma, pode-se definir, o r-ésimo momento amostral não centrado para média por:

$$\tilde{m}_r = \frac{1}{n}\sum_{j=1}^{n} x_j^r \tag{4.21}$$

Pode-se então, definir a média amostral, e o segundo, terceiro e quarto momentos centrados na média, em função dos momentos não centrados por:

Média: $$\tilde{\mu}_1 = 0 \tag{4.22}$$

Variância: $$\tilde{\mu}_2 = \tilde{m}_2 - \tilde{m}_1^2 \tag{4.23}$$

Assimetria $$\tilde{\mu}_3 = \tilde{m}_3 - 3\tilde{m}_1\tilde{m}_2 + 2\tilde{m}_1^3 \tag{4.24}$$

Curtose $$\tilde{\mu}_4 = \tilde{m}_4 - 4\tilde{m}_1\tilde{m}_3 + 6\tilde{m}_1^2\tilde{m}_2 - 3\tilde{m}_1^4 \tag{4.25}$$

Os valores amostrais de o coeficiente de assimetria e curtose são, respectivamente:

$$\sqrt{b_1} = \frac{\tilde{\mu}_3}{\tilde{\mu}_2 \sqrt{\tilde{\mu}_2}} \tag{4.26}$$

$$b_2 = \frac{\tilde{\mu}_4}{\tilde{\mu}_2^2} \tag{4.27}$$

Os coeficientes de assimetria populacional, para a distribuição normal, é $\beta_1=0$ e o coeficiente de curtose é $\beta_2=3$. Se $\beta_1<0$, então, a distribuição é assimétrica a esquerda, caso contrário, $\beta_1>0$, a distribuição é assimétrica a direita. Distribuições com $\beta_2<3$ são platicúrticas (menos pontudas com caudas mais baixas do que a normal), e aquelas com $\beta_2>3$ são leptocúrticas (mais pontudas e com caudas mais altas do que a normal).

Exemplo 4.3

Utilizando os dados do exemplo 4.1 calcular os momentos e os coeficientes de assimetria e curtose amostrais.

x	x^2	x^3	x^4
0,46	0,2116	0,0973	0,0448
1,79	3,2041	5,7353	10,2663
2,06	4,2436	8,7418	18,0081
2,91	8,4681	24,6422	71,7087
3,30	10,8900	35,9370	118,5921

3,74	13,9876	52,3136	195,6530
4,02	16,1604	64,9648	261,1585
4,59	21,0681	96,7026	443,8648
4,79	22,9441	109,9022	526,4317
8,65	74,8225	647,2146	5598,4070
36,31	176,0001	1046,2520	7244,1350

Têm-se:

$\tilde{m}_1$=36,31/10=3,631

$\tilde{m}_2$=176,0001/10=17,6000

$\tilde{m}_3$=1046,2520/10=104,6252

$\tilde{m}_4$=7244,135/10=724,4135

$\tilde{\mu}_1$ = 3,631

$\tilde{\mu}_2$ = 17,6 - $(3,631)^2$ = 4,4158

$\tilde{\mu}_3$ = 104,6252 - 3 x 3,631 x 17,6 + 2 x $(3,631)^3$ = 8,6518

$\tilde{\mu}_4$ = 724,4135 - 4 x 3,631 x 104,6252 + 6 x $(3,631)^2$ x 17,6 - 3 x $(3,631)^4$ = 75,6182

$\sqrt{b_1}$ = 8,6518/(4,4158 x $4,4158^{1/2}$) = 0,9324

$b_2 = 75{,}6182/(4{,}4158)^2 = 3{,}8780$

c.1) Uso do coeficiente de assimetria

Para se avaliar o grau de assimetria da distribuição, um teste baseado no coeficiente de assimetria (4.26), pode ser realizado. Níveis críticos para a estatística $\sqrt{b_1}$, podem ser encontrados em Pearson e Hartley (1966) para n>24, e em D'Agostino e Tietjen (1973) para n variando de 5 a 35. A assimetria será a esquerda se $\sqrt{b_1}$ for negativo, e a direita se $\sqrt{b_1}$ for positivo, significativamente. Em grandes amostras, os valores críticos de $\sqrt{b_1}$ podem ser obtidos com boa aproximação usando como desvio da normal padrão a estatística:

$$z_1 = \sqrt{b_1}\sqrt{\frac{(n+1)(n+3)}{6(n-2)}} \qquad (4.28)$$

c.2) Uso do coeficiente de curtose

Valores críticos para o coeficiente de curtose (4.27), podem ser encontrados em Pearson e Hartley (1966) para n>49 e D'Agostino e Tietjen (1971) para n variando de 7 a 50. Em grandes amostras, os valores críticos para o teste de

achatamento da curva, podem ser aproximados usando como desvio normal a seguinte estatística:

$$z_2 = (b_2 - 3 + \frac{6}{n+1}) \sqrt{\frac{(n+1)^2 (n+3)(n+5)}{24\, n(n-2)(n-3)}} \qquad (4.29)$$

Valores de b_2 maiores que 3 indicam que a distribuição é mais pontuda com caldas mais altas do que a normal; valores menores que 3 indicam uma distribuição achatada no centro e com caudas mais baixas do que a distribuição normal.

Exemplo 4.3 (continuação)

Os valores de z_1 e z_2, para o teste de assimetria e curtose foram:

z_1=1,609 com P(Z>|z_1|)=0,1074

z_2=1,886 com P(Z>|z_2|)=0,0592

Desta forma, ao nível de 5% de probabilidade se aceita a hipótese de simetria e de não achatamento da curva, demonstrando não se ter desvio da normalidade.

Verificando a normalidade por meio da distribuição bivariada

Em geral se deseja verificar a normalidade para dimensões superiores a 1, ou seja, para a distribuição p-variada, p≥2. Como já comentado anteriormente, é suficiente para propósitos práticos, avaliar apenas as distribuições univariadas e bivariadas. O caso bivariado será enfocado nesta seção.

Pelo resultado 4.2, dado vetor $\underset{\sim}{X}$ com distribuição normal bivariada, tem-se que,

$$(\underset{\sim}{x}-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}) \leq \chi^2_2(1-\alpha)$$

Através deste resultado, pode-se então, generalizar o processo gráfico conhecido como Q-Q plot. Dada uma amostra bivariada com n observações, o algoritmo seguinte pode ser usado para generalizar o processo gráfico mencionado. É importante salientar que este processo não é limitado apenas ao espaço bidimensional.

O algoritmo será apresentado, utilizando os dados do exemplo 1.1, com X_1 representando a quantidade de reais pela venda de ração, e X_2 sendo o número de sacos de rações vendidos, por n=4 firmas de Minas Gerais.

Exemplo 4.4

1) Calcular a distância quadrada generalizada amostral $d_{(j)}$ de cada observação em relação à média amostral, dada por:

$$d_j^2 = (\underset{\sim}{x}_j - \underset{\sim}{\bar{x}})' S^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\bar{x}}) \text{ , j=1, 2, ..., n}$$

Os valores da média e da matriz de covariância amostrais foram apresentados no exemplo 1.2, e são:

$$\underset{\sim}{\bar{X}} = \begin{bmatrix} 100 \\ 9 \end{bmatrix} \text{ e } S = \begin{bmatrix} 333{,}333 & 20{,}000 \\ 20{,}000 & 6{,}667 \end{bmatrix}$$

A matriz inversa de S é:

$$S^{-1} = \begin{bmatrix} 0{,}0037 & -0{,}0110 \\ -0{,}0110 & 0{,}1829 \end{bmatrix}$$

A distância generalizada para primeira observação é:

$$d_1 = \begin{bmatrix} 80-100 & 10-9 \end{bmatrix} \begin{bmatrix} 0{,}0037 & -0{,}0110 \\ -0{,}0110 & 0{,}1829 \end{bmatrix} \begin{bmatrix} 80-100 \\ 10-9 \end{bmatrix} = 2{,}0853$$

E assim sucessivamente, para as demais observações:

d_2 = 1,7926

d_3 = 1,3536

d_4 = 0,7683

2) ordenar as distâncias quadráticas amostrais do menor para o maior $d^2_{(1)} \leq d^2_{(2)} \leq ... \leq d^2_{(n)}$.

3) Obter os valores correspondentes, percentis, de probabilidade acumulada $q_{(j)} = \chi^2_p((j-½)/n)$, da distribuição de qui-quadrado. Estes percentis dependem da inversa da função de distribuição de qui-quadrado, e podem ser obtidos em vários softwares estatísticos.

J	$d^2_{(j)}$	(j-½)/n	$q_{(j)}$
1	0,7683	0,125	0,2671
2	1,3536	0,375	0,9400
3	1,7926	0,625	2,2479
4	2,0853	0,875	4,1589

4) Plotar ($d^2_{(j)}$; $q_{(j)}$) e examinar os resultados

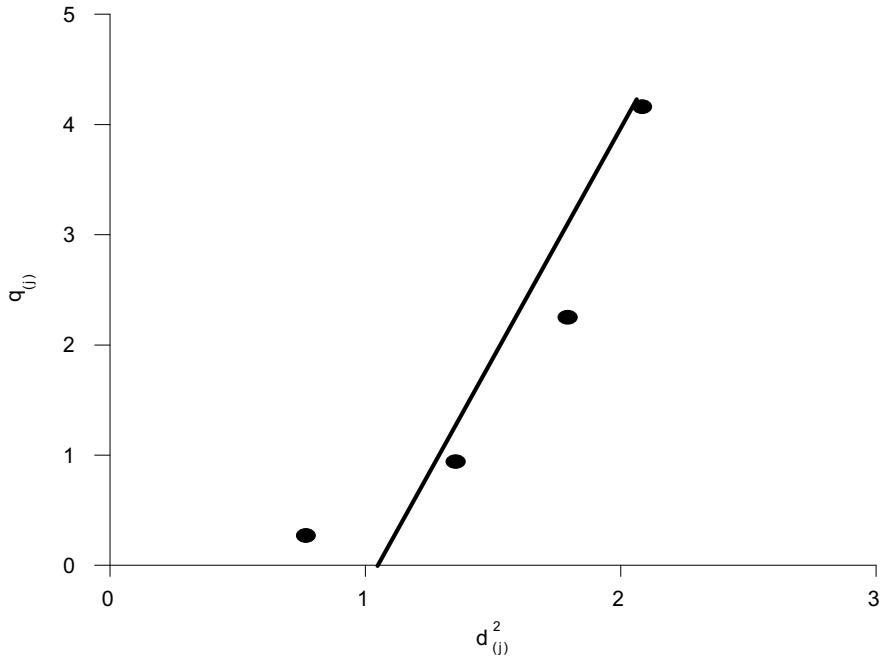


Figura 4.9. Q-Q plot para os dados do exemplo 1.1, destacando a possibilidade de utilização deste processo para os casos de dimensões superiores ou iguais a 2.

Pela Figura 4.9, verifica-se que não existem razões para duvidar de que a distribuição do número de sacos de rações vendidos e o montante de dinheiro arrecadado pelas firmas de rações em Minas Gerais, não seja normal bivariada, apesar do pequeno tamanho de amostras.

Verificando a normalidade multivariada por meio da curtose e assimetria de Mardia

Os coeficientes de assimetria e curtose de uma distribuição multivariada qualquer são definidos por:

$$\beta_{1,p} = E\left\{\left(\underset{\sim}{X} - \underset{\sim}{\mu}\right)' \Sigma^{-1} \left(\underset{\sim}{Y} - \underset{\sim}{\mu}\right)\right\}^3 \qquad (4.30)$$

em que $\underset{\sim}{X}$ é independente de $\underset{\sim}{Y}$, mas tem a mesma distribuição; e

$$\beta_{2,p} = E\left\{\left(\underset{\sim}{X} - \underset{\sim}{\mu}\right)' \Sigma^{-1} \left(\underset{\sim}{X} - \underset{\sim}{\mu}\right)\right\}^2 \qquad (4.31)$$

Essas esperanças para a distribuição normal multivariada são:

$$\beta_{1,p} = 0 \text{ e } \beta_{2,p} = p(p+2)$$

Para uma amostra de tamanho n, os estimadores de $\beta_{1,p}$ e $\beta_{2,p}$ são:

$$\hat{\beta}_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij}^3$$

$$\hat{\beta}_{2,p} = \frac{1}{n} \sum_{i=1}^{n} g_{ii}^2 = \frac{1}{n} \sum_{i=1}^{n} d_i^4$$

em que,

$$g_{ij} = \left(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}}\right)' \; S_n^{-1} \; \left(\underset{\sim}{X}_j - \bar{\underset{\sim}{X}}\right) \text{ e } d_i = \sqrt{g_{ii}}$$

Os estimadores $\hat{\beta}_{1,p}$ (quadrado do coeficiente de assimetria quando p=1) e $\hat{\beta}_{2,p}$ (igual ao coeficiente de curtose univariado quando p=1) são não-negativos. Sob distribuição normal multivariada espera-se que a E($\hat{\beta}_{1,p}$) seja próxima de zero. O estimador $\hat{\beta}_{2,p}$ é muitas vezes usado para avaliar observações que estão a grandes distâncias da média amostral.

Mardia (1970) mostra que para grandes amostras,

$$k_1 = \frac{n\hat{\beta}_{1,p}}{6}$$

segue a distribuição de χ^2 com p(p+1)(p+2)/6 graus de liberdade, e

$$k_2 = \frac{\left\{\hat{\beta}_{2,p} - p(p+2)\right\}}{\left[\frac{8p(p+2)}{n}\right]^{1/2}}$$

segue a distribuição normal padrão. Para pequenos valores de n, as tabelas de valores críticos para testar a hipótese multivariada de normalidade são fornecidas por Mardia (1974).

Exemplo 4.5

Usando o exemplo das rações testar a normalidade multivariada pelo teste dos desvios de assimetria e curtose. Os valores amostrais são:

Obs	Reais	Vendas
1	80	10

2	120	12
3	90	6
4	110	8

As estatísticas amostrais são:

$$\underset{\sim}{\bar{X}} = \begin{bmatrix} 100 \\ 9 \end{bmatrix} \quad S_n = \begin{bmatrix} 250 & 15 \\ 15 & 5 \end{bmatrix} \quad S_n^{-1} = \begin{bmatrix} 0{,}004878 & -0{,}014634 \\ -0{,}014634 & 0{,}243902 \end{bmatrix} \text{ ou } S_n^{-1} = \frac{1}{1025}\begin{bmatrix} 5 & -15 \\ -15 & 250 \end{bmatrix}$$

Os desvios de cada observação da média amostral ($\underset{\sim}{\varepsilon}_i$):

1. $\underset{\sim}{\varepsilon}_1' = \begin{bmatrix} -20 & 1 \end{bmatrix}$ 2. $\underset{\sim}{\varepsilon}_2' = \begin{bmatrix} 20 & 3 \end{bmatrix}$ 3. $\underset{\sim}{\varepsilon}_3' = \begin{bmatrix} -10 & -3 \end{bmatrix}$ 4. $\underset{\sim}{\varepsilon}_4' = \begin{bmatrix} 10 & -1 \end{bmatrix}$

i) Teste baseado no coeficiente de assimetria

É necessário calcular os valores de g_{ij} para todos os pares de i e j, obtidos da seguinte forma:

Para i=1 e j=1, $g_{11} = \begin{bmatrix} -20 & 1 \end{bmatrix} S_n^{-1} \begin{bmatrix} -20 \\ 1 \end{bmatrix} = 2{,}7805$

Para i=1 e j=2, $g_{12} = \begin{bmatrix} -20 & 1 \end{bmatrix} S_n^{-1} \begin{bmatrix} 20 \\ 3 \end{bmatrix} = -0{,}6341$

Para as demais combinações, têm-se: g_{13}=-0,4878, g_{14}=-1,6585, g_{22}=2,3902, g_{23}=-1,8537, g_{24}=0,0976, g_{33}=1,8049, g_{34}=0,5366 e g_{44}=1,0244.

Logo,

$$\hat{\beta}_{1,2} = \frac{\left(2,7805^3 + 2(-0,6341)^3 + \cdots + 1,0244^3\right)}{16} = 1,2766$$

então,

$$k_1 = \frac{n\hat{\beta}_{1,2}}{6} = \frac{4 \times 1,2766}{6} = 0,8511$$

Como $k_1 \sim \chi^2$ com p(p+1)(p+2)/6=4 graus de liberdade, e sabendo que $\chi^2_{0,05;\,4} = 9,488$, então H_0 não deve ser falseada, ou seja, não existe razões para suspeitar da violação da simetria da distribuição multivariada.

ii) Teste baseado no coeficiente de curtose

Inicialmente, estima-se o coeficiente de curtose da seguinte forma:

$$\hat{\beta}_{2,p} = \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2 = \frac{1}{4}\left(2,7805^2 + 2,3902^2 + 1,8049^2 + 1,0244^2\right) = \frac{17,7513}{4} = 4,4378$$

e em seguida, estima-se o valor estimado da normal (0, 1):

$$k_2 = \frac{4,4378 - 2(2 \times 2)}{\left(\frac{8 \times 2 \times 4}{4}\right)^{1/2}} = \frac{-3,5621}{4} = -0,8905$$

Não existem razões para duvidar de que a distribuição multivariada tenha algum desvio de curtose, uma vez que $|k_2| < z_{0,025} = 1,96$.

iii) Programa SAS para o teste de normalidade

A seguir são apresentados um programa SAS usando o Proc Calis para o teste da curtose e um programa em IML, para ambos parâmetros. O programa fornece as estatísticas amostrais e os valores das significâncias exatas.

```
Data FR;
Input Reais Vendas;
cards;
 80 10
120 12
 90  6
110  8
;
Proc Calis data=FR Kurtosis;
 Title1 j=1 "Uso do Calis para testar a
normalidade";
 Title2 "pela Curtose de Mardia";
 Lineqs
 Reais=e1,
 vendas=e2;
 std
 e1=eps1, e2=eps2;
 Cov
 e1=eps1, e2=eps2;
Run;
```

```
Proc IML;
  use FR;
  read next 4 into X;  /* lendo n observacoes dentro de X  */
  n=nrow(X);p=ncol(X);
  dfchi=p*(p+1)*(p+2)/6; /*definindo GL para B1,p          */
  q=i(n) - (1/n)*j(n,n,1); /* criando q=I-1/nJ, auxiliar                */
  S=(1/n)*x`*q*x;        /* matriz de covariancias viesada              */
 S_inv=inv(S);          /* inversa de S                                 */
  print s s_inv;
  g=q*x*s_inv*x`*q;      /* matriz com gij                              */
  print g;
  beta1=(sum(g#g#g))/(n*n); /*produto elem. a elem. E sua soma/n^2 */
  beta2=trace(g#g)/n;    /* idem com tomada do traco/n                  */
  print beta1 beta2;
  k1=n*beta1/6;     /* definindo k1 e k2, transformacoes de b1,p e b2,p   */
  k2=(beta2-p*(p+2))/sqrt(8*p*(p+2)/n);
  pvalskew=1-probchi(k1,dfchi); /* calculo dos p_values respectivos       */
  pvalkurt=2*(1-probnorm(abs(k2)));
  print k1 pvalskew;
  print k2 pvalkurt;
Quit; /* abandonando IML                                                */
```

Finalmente é apresentado a seguir um programa SAS para orientar os leitores na simulação de dados com distribuição normal multivariada com média e covariância especificada. O exemplo apresentado gera uma distribuição normal trivariada.

```
Proc IML;
  SIG={8   4    1,
         4 10   3,
         1   3 18};
  St=Root(sig);
  mu={1, 10, 8};
  x=j(100,3,0);
  do i=1 to 100;
        zi=j(3,1,0);
        do ii=1 to 3;
              zi[ii]=rannor(0);
        end;
```

```
          xi=st`*zi+mu;
          do ii=1 to 3;
              x[l,ii]=xi[ii];
          end;
   end;
   print x;
   create dtnorm from x;
   append from x;
quit;
proc print data=dtnorm;
run;quit;
```

4.8. Exercícios

4.8.1. Com os dados do exemplo 4.4, tendo como hipótese que os mesmos seguem a distribuição normal bivariada, utilize o resultado 4.2, ao nível de 50%, de que as distâncias generalizadas seguem a distribuição qui-quadrado. Utilizando então a distribuição de proporções, item (a), verifique a normalidade bivariada dos dados, contando a proporção observada ($\hat{P}_i$) de distâncias que pertencem a elipse, e comparando com a estatística abaixo.

$$| \hat{P}_i - 0,5 | > 3\sqrt{\frac{0,5 \times 0,5}{n}} = \frac{1,5}{\sqrt{n}}$$

4.8.2. Utilizando os dados deste exemplo (1.1), realize todos os testes univariados, propostos, neste capítulo, para ambas variáveis.

4.8.3. Utilizando os dados climáticos, obtidos por Diniz (1996), na fazenda Cooparaíso-EPAMIG, Jacuí, MG, de agosto de 1994 a janeiro de 1995, teste a pressuposição de normalidade tridimensional dos mesmos. Utilize para isso, o processo gráfico apresentado, e o teste do exercício número 4.8.1 e o teste baseado nos desvios de assimetria e curtose de Mardia.

Temperatura	Umidade Relativa (%)	Precipitação (mm)
22,7	64,1	7,9
23,7	56,1	1,5
24,3	54,9	0,0
24,4	58,2	0,0
24,5	62,8	8,7
25,2	70,3	22,5
25,5	75,2	57,0
24,7	81,4	75,7
24,3	79,3	123,2
24,7	74,6	124,4
24,9	78,0	148,0

4.8.4. Utilize os dados de uma amostra de 24 cochonilhas, fêmeas adultas, de *Quadraspidiotus perniciosus* (Comst.), por ramo de pessegueiro, na região de Jacuí-MG, e teste a pressuposição de normalidade dos dados, utilizando os procedimentos apresentados univariados na seção 4.7.

0,8 1,0 0,6 0,6 0,2 0,8 2,5 1,5 0,3 1,7 1,9 2,5 1,1 5,0 0,9 1,7 2,6 4,5 1,8 1,0 0,5 0,4 1,8 0,7

4.9. Referências

ANDERSON, T.W. **An introduction to multivariate statistical analysis**. 2nd ed. New York, John Wiley, 1984, 675p.

BOCK, R.D. **Multivariate statistical methods in behavioral research.** McGrawn Hill, 1975.

D'AGOSTINO, R.B.;TIÉTJEN, G.L. Simulation probability points of b_2 in small samples, **Biometrika**, v.58, p.:669-672, 1971.

_____.;_____ Approaches to the null distribution of $\sqrt{b_1}$, **Biometrica**, v.60, p.:169-173, 1973.

DINIZ, L. de C. **Dinâmica populacional do piolho-de-são josé *Quadraspidiotus perniciosus* (Comstock, 1881) (Homoptera: Diaspididae) em pessegueiro, no município de Jacuí-Minas Gerais**. Lavras, Universidade Federal de Lavras, 1996. 61p. (tese Ms)

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 4th edition. Prentice Hall, New Jersey, 1998. 816p.

MARDIA, K.V. Measures of multivariate skewness and kurtosis with applications. **Biometrika**, p.519-530, 1970.

MARDIA, K.V. Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. **Sankyã**. A36, p.115-128, 1974.

PEARSON, E.S.; HARTLEY, H.O. **Biometrika Tables for Statisticians** Vol. 1 ed ed., Cambridge University Press, New York, 1966