5 Inferências sobre o vetor média

5.1. Introdução

Este capítulo é o primeiro deste material a apresentar inferências, utilizando as técnicas, os conceitos e os resultados apresentados nos capítulos prévios. Este capítulo, por estar intimamente relacionado à inferência estatística, ou seja, é voltado para obtenção de conclusões válidas para a população com base nas informações amostrais. As inferências realizadas neste capítulo são relativas a vetor populacional de médias e nos seus componentes. Umas das mensagens centrais da análise multivariada, que deverá ser abordada neste e nos próximos capítulos, é que p variáveis correlacionadas devem ser analisadas simultaneamente.

5.2. Inferências sobre média de uma população normal

Nesta seção serão abordados os testes de significância e a obtenção de intervalos de confiança (IC) para a média de uma população normal.

Inicialmente será abordado o problema de verificar se um determinado valor $\underset{\sim}{\mu}_0$ é um possível valor (plausível) para a verdadeira média populacional desconhecida. Do ponto de vista dos testes de hipóteses este problema pode ser abordado através do teste:

$$H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0 \quad \text{vs} \quad H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$$

aqui, H_0 é a hipótese nula e H_1 é a hipótese (bilateral) alternativa. Considerando o caso univariado, e se X_1, X_2, ..., X_n representam uma amostra aleatória extraída de uma população normal, o teste estatístico apropriado para esta hipótese, quando p é igual a 1, é:

$$t = \frac{(\overline{X} - \mu_0)}{S/\sqrt{n}}, \text{ em que, } \overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \text{ e } S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2 .$$

O teste em questão segue a distribuição de t-student com n-1 graus de liberdade. A hipótese H_0 será rejeitada se o valor observado de |t| exceder um ponto percentual especificado da distribuição de t-student com n-1 graus de liberdade (GL).

Analogamente, considerando agora a distância quadrada da média amostral $\overline{X}$ para o valor a ser testado, pode-se rejeitar H_0 a um nível de significância α, se

$$t^2 = n(\bar{X}-\mu_0)(S^2)^{-1}(\bar{X}-\mu_0) \geq t^2_{n-1}(\alpha/2) \qquad (5.1)$$

em que, $t^2_{n-1}(\alpha/2)$ representa o quantil quadrático superior 100(α/2) da distribuição de t-student com n-1 GL.

Se H_0 não é rejeitada, então se conclui que μ_0 é um valor plausível para representar a média populacional normal. No entanto, uma pergunta natural pode surgir: existem outros valores de μ que são consistentes com os dados? A resposta é sim. De fato, existe um conjunto de valores plausíveis que serviriam como média para a população normal estudada. Da bem conhecida correspondência entre a região de aceitação dos testes de hipóteses e o intervalo de confiança para μ tem-se:

$\left|\dfrac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| < t_{n-1}(\alpha/2)$ (não rejeitar H_0) é equivalente a:

$$\bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \qquad (5.2)$$

Antes de a amostra ser retirada, o intervalo de confiança de 100(1-α)% de (5.2) é um intervalo aleatório, pois seus limites dependem das variáveis aleatórias $\bar{X}$ e S. A probabilidade do intervalo conter μ é 100(1-α)% e entre um grande número independentes de tais intervalos, 100(1-α)% deles conterão μ.

É considerada agora a generalização do caso univariado para o multivariado. O problema de determinar se um dado vetor $\underset{\sim}{\mu}_0$ (p x 1) é um valor plausível da média de uma distribuição normal multivariada. Uma generalização da distância quadrada apresentada em (5.1) é:

$$T^2 = n\left(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}_0\right)' S^{-1}\left(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}_0\right) \tag{5.3}$$

em que,

$$\underset{\sim}{\bar{X}} = \frac{1}{n}\sum_{j=1}^{n} \underset{\sim}{X}_j \,,\; S = \frac{1}{n-1}\sum_{j=1}^{n}\left(\underset{\sim}{X}_j - \underset{\sim}{\bar{X}}\right)\left(\underset{\sim}{X}_j - \underset{\sim}{\bar{X}}\right)' \quad \text{e} \quad \underset{\sim}{\mu}_0 = \begin{bmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{bmatrix}$$

A estatística T^2 é chamada de chamada de T^2 de Hotelling, em honra a Harold Hotelling (Bock, 1975), um pioneiro da estatística multivariada, que pela primeira vez obteve a sua distribuição. Felizmente, tabelas especiais dos pontos percentuais para a distribuição T^2 não são necessárias na realização dos testes de hipóteses, devido à estatística:

T^2 ser distribuído como $\frac{(n-1)p}{n-p} F_{p,n-p}$ (5.4)

em que, $F_{p,n-p}$ representa uma variável com distribuição F com p e n-p GL.

De uma forma geral a distribuição de T^2 considerando ν graus de liberdade e dimensão p é dada por:

$$T^2 = F_{p,\nu+1-p} \times \frac{\nu p}{\nu + 1 - p} \qquad (5.4.1)$$

Desta forma para se testar a hipótese $H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ versus $H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$, no valor nominal α de significância, deve-se rejeitar H_0 em favor de H_1 se

$$T^2 = n\left(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}_0\right)' S^{-1}\left(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}_0\right) > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \qquad (5.5)$$

Infelizmente, é raro nas situações multivariadas, estar contente com o teste da hipótese $H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$, em que todos os componentes do vetor média são especificados sob a hipótese de nulidade. Em geral é preferível encontrar regiões de valores de $\underset{\sim}{\mu}$ que são plausíveis na luz dos dados observados.

Exemplo 5.1

A matriz X, apresentada a seguir, representa uma amostra de n=3 observações retiradas de uma distribuição normal bivariada.

$$X = \begin{bmatrix} 11 & 2 \\ 10 & 4 \\ 9 & 3 \end{bmatrix}$$

Teste a hipótese de que $\underset{\sim}{\mu}_0' = [9\ 2]$ seja um valor plausível para representar a média populacional.

A estatísticas amostrais são:

$$\underset{\sim}{\bar{X}} = \begin{bmatrix} 10 \\ 3 \end{bmatrix} \text{ e } S = \begin{bmatrix} 1{,}0 & -0{,}5 \\ -0{,}5 & 1{,}0 \end{bmatrix}$$

Então,

$$S^{-1} = \frac{1}{3}\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

E o valor de T^2 será obtido da seguinte forma:

$$T^2 = 3\ \begin{bmatrix} 10-9 & 3-2 \end{bmatrix} \frac{1}{3} \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 10-9 \\ 3-2 \end{bmatrix} = 12$$

O valor de $F_{2,1}$ ao nível de 5% é 199,5. então H_0 será rejeitada se o valor observado de T^2 superar

$$\frac{(n-1)p}{n-p} F_{2,1} = \frac{4}{1} 199,5 = 798,0$$

Como neste caso, o valor de T^2 observado (12,0) foi inferior ao valor crítico (798,0), então H_0 não deve ser rejeitada. É importante salientar neste ponto, que a hipótese H_0 será rejeitada se um ou mais dos componentes do vetor média amostral, ou alguma combinação de médias, diferir muito do valor hipotético $\underset{\sim}{\mu}_0'$ = [9 2]. Neste estágio, não se tem idéia de quais os valores hipotéticos não são suportados pelos dados.

5.3. Região de confiança e Comparações simultâneas de componentes de média

Será inicialmente, generalizado o conceito univariado de intervalo de confiança para o multivariado de região de confiança, R(**X**). A região de confiança conterá 100(1-α)% se antes de a amostra ser selecionada,

$$P[R(X) \text{ cobrir o verdadeiro } \underset{\sim}{\theta}] = 1 - \alpha \qquad (5.6)$$

em que $\underset{\sim}{\theta}$, representa um vetor de parâmetros desconhecidos (Krzanowski, 1993). No caso, a região de confiança para $\underset{\sim}{\mu}$ de uma distribuição normal p variada, será todos os valores de $\underset{\sim}{\mu}$ tais que:

$$P\left[n\left(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu}\right)' S^{-1}\left(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu}\right) \le \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \right] \quad (5.7)$$

Para determinar se um dado valor $\underset{\sim}{\mu}_0$ é um valor plausível de $\underset{\sim}{\mu}$, basta calcular a distância quadrada generalizada $n(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu})'S^{-1}(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu})$ e comparar com $(n-1)pF_{p,n-p}(\alpha)/(n-p).$ Se a distância quadrada for maior que $(n-1)pF_{p,n-p}(\alpha)/(n-p)$, então $\underset{\sim}{\mu}_0$ não pertence a região de confiança. Isto é equivalente a testar a hipótese H_0: $\underset{\sim}{\mu}=\underset{\sim}{\mu}_0$ contra a H_1: $\underset{\sim}{\mu}\neq\underset{\sim}{\mu}_0$, a qual possibilita afirmar que a região de confiança constitui-se em todos os valores de $\underset{\sim}{\mu}_0$ cujo teste T^2 não rejeitaria a hipótese nula a favor da alternativa, em um nível de significância α.

Para p≥4 não se pode fazer o gráfico da região de confiança para $\underset{\sim}{\mu}$. Pode se, no entanto, calcular os eixos da elipsóide de confiança e seus tamanhos relativos, os quais são determinados pelos autovalores λ_i e autovetores $\underset{\sim}{e}_i$ de **S**. Os tamanhos dos semi-eixos de

$$n(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu})'S^{-1}(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu}) \le c^2 = \frac{p(n-1)}{n-p}F_{p,n-p}(\alpha)$$

são determinados por

$\frac{\sqrt{\lambda_i}c}{\sqrt{n}} = \sqrt{\lambda_i}\sqrt{[p(n-1)F_{p,n-p}(\alpha)]/[n(n-p)]}$ unidades ao longo de $\underset{\sim}{e}_i$.

Começando do centro, determinado por $\underset{\sim}{\bar{X}}$, os eixos da elipsóide são:

$$\pm\sqrt{\lambda_i}\sqrt{[p(n-1)F_{p,n-p}(\alpha)]/[n(n-p)]}\ \underset{\sim}{e}_i$$

Exemplo 5.2

A partir dos dados do exemplo 5.1, obter a região de confiança de 95%, e verificar se o ponto $\underset{\sim}{\mu}_0'$=(13, 4) pertence a mesma.

$$\underset{\sim}{\bar{X}} = \begin{bmatrix} 10 \\ 3 \end{bmatrix},\ S = \begin{bmatrix} 1{,}0 & -0{,}5 \\ -0{,}5 & 1{,}0 \end{bmatrix} \text{ e } S^{-1} = \frac{1}{3}\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

Os autovalores e autovetores de **S**, são:

$$\lambda_1 = 1{,}5 \quad \underset{\sim}{e}_1' = \begin{bmatrix} 0{,}707107 & -0{,}707107 \end{bmatrix}$$

$$\lambda_2 = 0{,}5 \quad \underset{\sim}{e}_2' = \begin{bmatrix} 0{,}707107 & 0{,}707107 \end{bmatrix}$$

A elipse de confiança 95% para $\underset{\sim}{\mu}$ consiste de todos os valores (μ_1, μ_2) que satisfazem:

$$3\,[10-\mu_1, 3-\mu_2]\,\frac{1}{3}\begin{bmatrix}4 & 2\\ 2 & 4\end{bmatrix}\begin{bmatrix}10-\mu_1\\ 3-\mu_2\end{bmatrix} \le \frac{2\,(2)}{1}\times 199,5$$

ou, $4(10-\mu_1)^2 + 4(10-\mu_1)(3-\mu_2) + 4(3-\mu_2)^2 \le 798$

Para verificar se o ponto $\underset{\sim}{\mu}_0'$=(13, 4) pertence à elipse, calcula-se:

$$4(10-13)^2 + 4(10-13)(3-4) + 4(3-4)^2 = 52 \le 798,0$$

o que permite que se conclua que o ponto testado está na região de confiança. O gráfico da elipse obtida pode ser visualizado na Figura 5.1. com a análise gráfica, pode-se confirmar que o ponto em questão pertence à região de confiança.

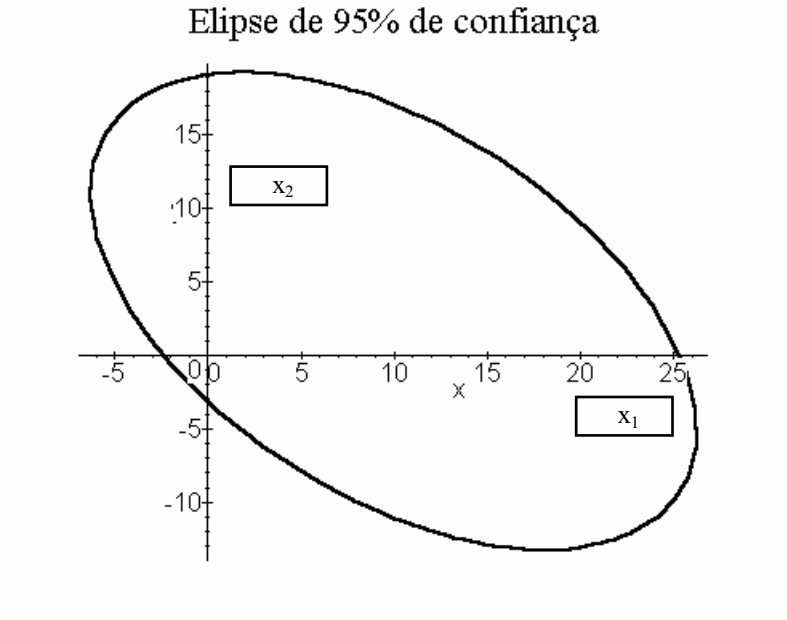

Figura 5.1. Elipse de 95% de confiança para o vetor populacional de médias, obtido a partir dos dados do exemplo 5.1.

Exemplo 5.3

Para exemplificar a região tridimensional para a média populacional, os dados de produção comercial (t/ha), produção de tubérculos graúdos (t/ha) e peso médio de tubérculos graúdos (g) de 15 clones de batata selecionados em Maria da Fé e Lavras (Momenté, 1994), foram utilizados e encontram-se no quadro a seguir.

Obter a região de 95% de confiança para o vetor média populacional. Verificar se o ponto $\underset{\sim}{\mu}_0' = (16,89 \;\; 8,76 \;\; 109,23)$ pertence a região de confiança (ponto referente a cultivar Achat). Traçar a região de confiança.

Clones	Produção comercial	Produção de tubérculos graúdos	Peso médio de tubérculos graúdos
1	47,82	40,40	146,30
2	42,40	26,96	94,58
3	41,82	27,33	143,66
4	40,77	21,81	127,29
5	40,27	33,06	115,17
6	39,84	22,31	99,32
7	38,36	32,81	150,13
8	38,15	26,02	131,17
9	37,55	21,69	152,04
10	36,19	25,65	154,83
11	36,15	23,46	95,43
12	35,17	25,29	105,97
13	34,90	22,92	113,59
14	34,57	16,25	86,39
15	34,15	21,75	119,50

Fonte: Momenté, 1994

O vetor de médias e a matriz de covariância amostrais são:

$$\underset{\sim}{\bar{X}} = \begin{bmatrix} 38,541 \\ 25,854 \\ 122,358 \end{bmatrix} \qquad S = \begin{bmatrix} 13,8195 & 15,8284 & 24,7250 \\ 15,8284 & 34,8769 & 63,0215 \\ 24,7250 & 63,0215 & 540,1553 \end{bmatrix}$$

Os autovalores e autovetores de S são:

$\lambda_1 = 549,208 \quad \underset{\sim}{e}_1' = (0,049 \;\; 0,123 \;\; 0,991)$

$\lambda_2 = 34,460 \quad \underset{\sim}{e}_2' = (0,500 \;\; 0,856 \;\; -0,131)$

$\lambda_3 = 5,185 \quad \underset{\sim}{e}_3' = (0,865 \;\; -0,502 \;\; 0,019)$

A região de confiança fica determinada por:

$$n(\overline{\underset{\sim}{X}}-\underset{\sim}{\mu})'S^{-1}(\overline{\underset{\sim}{X}}-\underset{\sim}{\mu}) \le c^2 = \frac{p(n-1)}{n-p}F_{p,n-p}(\alpha)$$

$$15\begin{bmatrix}38,541-\mu_1 & 25,854-\mu_2 & 122,358-\mu_3\end{bmatrix}\begin{bmatrix}0,15149 & \text{Sim.} & \\ -0,07124 & 0,06983 & \\ 0,00138 & -0,00489 & 0,002358\end{bmatrix}\begin{bmatrix}38,541-\mu_1\\ 25,854-\mu_2\\ 122,358-\mu_3\end{bmatrix} \le$$

$$\frac{3\times 14}{12}\times 3,49 = 12,215$$

$$\begin{aligned}&= 2,27(38,541-\mu_1)^2 - 2,14(38,541-\mu_1)(25,854-\mu_2) + 0,04(38,541-\mu_1)(122,358-\mu_3) + \\ &+1,05(25,854-\mu_2)^2 - 0,15(25,854-\mu_2)(122,358-\mu_3) + 0,04(122,358-\mu_3)^2 \le 12,215\end{aligned}$$

Para verificar se o ponto $\underset{\sim}{\mu}_0' = (16,89\ \ 8,76\ \ 109,23)$ pertence a região de confiança, basta substituir os valores de μ_1 por 16,89, de μ_2 por 8,76 e o de μ_3 por 109,23. O valor encontrado de 563,4964 é superior a 12,215, o que indica que a média da Cultivar Achat, não pertence a região de 95% de confiança para média das 15 famílias clonais estudadas.

Utilizando o programa Maple, através da seguinte macro, foi traçado o gráfico, elipsóide de confiança (Figura 5.2), da região de 95% de confiança para $\underset{\sim}{\mu}$. Pode-se visualizar também que o ponto em questão não pertence a elipsóide de confiança.

$$\begin{aligned}&\text{Implicitplot3d}\big(2.27\,(38.541-x1)^2-2.14\,(38.541-x1)\,(25.854-x2)\\&\quad+.04\,(38.541-x1)\,(122.358-x3)+1.05\,(25.854-x2)^2\\&\quad-.15\,(25.854-x2)\,(122.358-x3)+.04\,(122.358-x3)^2=12.125,\\&x1=34\,..\,48,\ x2=17\,..\,37,\ x3=105\,..\,152,\ grid=[16,16,23],\\&numpoints=10000,\ labels=[x1,x2,x3],\ axes=boxed,\\&style=PATCHCONTOUR,\ thickness=2,\ linestyle=0\big)\end{aligned}$$

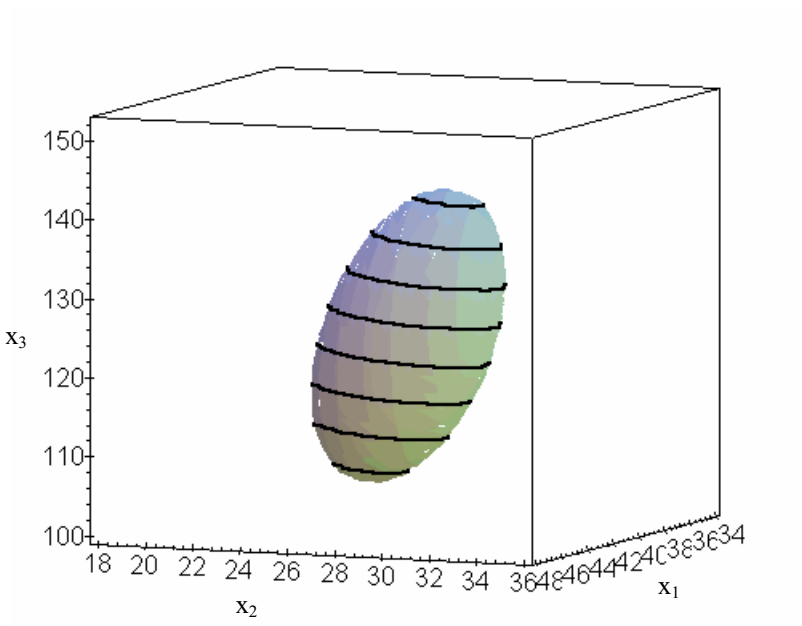

Figura 5.2. Elipsóide de 95% de confiança para o vetor de médias populacional, obtida a partir dos dados do exemplo 5.3.

Intervalos de confiança simultâneos

Enquanto a região de confiança fornece corretamente o conjunto de valores plausíveis para a média de uma população normal, qualquer resumo de conclusões, em geral, inclui intervalos de confiança sobre médias individuais. Assim, adota-se que todos os intervalos de confiança sejam verdadeiros simultaneamente com uma alta probabilidade específica. Isto garante que

qualquer afirmação não seja incorreta com alta probabilidade, o que conduz ao termo intervalo de confiança simultâneo (Jonhson e Wichern, 1998).

Considerando uma combinação linear das médias amostrais,

$$\underset{\sim}{\ell}'\overline{\underset{\sim}{X}} = \ell_1\overline{X}_1 + \ell_2\overline{X}_2 + \cdots + \ell_p\overline{X}_p$$

cuja distribuição amostral possui estimador da covariância dada por:

$$\frac{\underset{\sim}{\ell}'S\underset{\sim}{\ell}}{n}$$

Dessa forma poderia se pensar em se obter intervalos de confiança de 95% baseados na distribuição de t-student,

$$\ell'\overline{\underset{\sim}{X}} \pm t_{n-1}(\alpha/2)\frac{\sqrt{\underset{\sim}{\ell}'S\underset{\sim}{\ell}}}{\sqrt{n}} \qquad (5.8)$$

O intervalo de (5.8) pode ser interpretado como intervalos sobre componentes do vetor média, assim por exemplo, fazendo-se $\underset{\sim}{\ell}' = [1 \quad 0 \ldots . 0]$, (5.8) se torna o intervalo clássico para a média de uma população normal univariada. Neste caso tem-se uma série de inferências sobre os componentes de $\underset{\sim}{\mu}$, cada um associado com o coeficiente de confiança de 1-α, através de diferentes escolhas de $\underset{\sim}{\ell}$. No entanto o coeficiente de confiança para todos os intervalos

tomados simultaneamente não é 1-α. Para corrigir esta imperfeição demonstra-se (Johnson e Wichern, 1988; Anderson, 1984) que para garantir o nível de confiança simultâneo de 1-α é necessário recorrer à distribuição de T^2. Este resultado está apresentado a seguir:

$$\underset{\sim}{\ell}'\bar{\underset{\sim}{X}} \pm \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha)\underset{\sim}{\ell}'S\underset{\sim}{\ell}} \qquad (5.9)$$

Método de Bonferroni para Comparações múltiplas

Muitas vezes um pequeno número de intervalos de confiança é requerido. Nestas situações pode-se ter uma melhor opção do que as comparações simultâneas, proposta em (5.9), obtendo intervalos de confiança mais curtos (mais precisos) do que o intervalo simultâneo de T^2. Esta alternativa de intervalo é conhecida por método de Bonferroni.

A seguir será apresentado o método para obtenções de intervalo de confiança para os componentes de média. Se as m=p médias forem consideradas, então, o método de Bonferroni é:

$$\bar{X}_i \pm t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{S_{ii}}{n}} \qquad i=1,2,...,p=m \qquad (5.10)$$

Exemplo 5.4

Utilizando os dados do exemplo 5.2, obter os intervalos clássicos de t-student, T^2 e Bonferroni, para os componentes individuais do vetor média, e compará-los entre si, quanto ao comprimento.

O vetor de médias e a matriz de covariância amostral são:

$$\overline{\underset{\sim}{X}} = \begin{bmatrix} 10 \\ 3 \end{bmatrix} \text{ e } S = \begin{bmatrix} 1,0 & -0,5 \\ -0,5 & 1,0 \end{bmatrix}$$

1. Intervalo T^2

- $$IC_{\mu_1(0,95)} = \bar{X}_1 \pm \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)} \sqrt{\frac{S_{11}}{n}}$$

$$IC_{\mu_1(0,95)} = 10 \pm \sqrt{\frac{2(3-1)}{3-2} 199,5} \sqrt{\frac{1}{3}}$$

$$IC_{\mu_1(0,95)} = 10 \pm 16,31 = [-6,31; 26,31]$$

- $$IC_{\mu_2(0,95)} = 3 \pm \sqrt{\frac{2(3-1)}{3-2} 199,5} \sqrt{\frac{1}{3}}$$

$$IC_{\mu_2(0,95)} = 3 \pm 16,31 = [-13,31; 19,31]$$

Observa-se que os limites dos intervalos de confiança múltiplos representam os limites da elipse de confiança de 95% (Figura 5.1), projetados nos respectivos eixos.

2. Intervalo de Bonferroni

Neste caso, m=p=2, portanto $\alpha/2m=0,0125$. O valor de t-student correspondente, com n-1=2 GL é 6,21. Então,

- $$IC_{\mu_1(0,95)} = 10 \pm 6,21\sqrt{\frac{1}{3}}$$

$$IC_{\mu_1(0,95)} = [6,41; 13,39]$$

- $$IC_{\mu_2(0,95)} = 3 \pm 6,21\sqrt{\frac{1}{3}}$$

$$IC_{\mu_2(0,95)} = [-0,59; 6,59]$$

Observa-se nesta situação que os intervalos são bem mais estreitos que o seu correspondente em 1.

3. Intervalo t de Student

Neste caso $\alpha/2=0,025$ e o valor de t-student correspondente com 2 GL é 4,30. Então,

- $$IC_{\mu_1(0,95)} = 10 \pm 4,30\sqrt{\frac{1}{3}}$$

$$IC_{\mu_1(0,95)} = [7,52; 12,48]$$

- $$IC_{\mu_2(0,95)} = 3 \pm 4,30\sqrt{\frac{1}{3}}$$

$$IC_{\mu_2(0,95)} = [0,52; 5,48]$$

Apesar destes últimos intervalos individualmente garantir com 95% de probabilidade que as médias populacionais estão contidas nos mesmos, não há garantia de que simultaneamente eles contenham as médias populacionais no mesmo valor nominal de probabilidade, diga-se 95%. Na melhor das hipóteses, variáveis não correlacionadas, o valor real do coeficiente de confiança é $(1-\alpha)^p=0,95^2=0,9025$.

5.4. Inferências sobre proporções de grandes amostras

Freqüentemente, algumas características de interesse na população estão na forma de atributos. Cada indivíduo nesta população pode ser descrito em termos dos atributos que possui, os quais são codificados, pela sua presença e ausência. Na população, com q característica, a proporção de elementos possuindo os respectivos atributos são p_1, p_2, ..., p_q. Considerando q atributos mutuamente exclusivos e características exaustivas, então, $p_q=1-(p_1+p_2+...+p_{q-1})$.

Numa grande amostra de tamanho n, pelo teorema do limite central, $\hat{\underset{\sim}{p}}$ possui distribuição aproximadamente normal, com

$$E(\hat{\underset{\sim}{p}}) = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_q \end{bmatrix} \quad \text{e} \quad Cov(\hat{\underset{\sim}{p}}) = \frac{1}{n}\begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_q \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_q \\ \vdots & \vdots & \ddots & \vdots \\ -p_qp_1 & -p_qp_2 & \cdots & p_q(1-p_q) \end{bmatrix}.$$

Para grandes amostras, a aproximação continua válida se um estimador de $\mathrm{Cov}\left(\hat{\underset{\sim}{p}}\right)$, $(1/n)\hat{\Sigma}$, for utilizado.

Desde que cada elemento da população pode ter apenas um atributo, então, $p_q=1-(p_1+p_2+...+p_{q-1})$, o que trás como conseqüência que o posto de $\hat{\Sigma}$ é igual a q-1, portanto sua inversa não existe. Apesar disso, pode-se desenvolver intervalos de confiança simultâneos de $100(1-\alpha)\%$, para qualquer combinação $\underset{\sim}{\ell}'\underset{\sim}{p}$.

Para uma amostra de tamanho n, considerando q categorias da distribuição multinomial, o intervalo aproximado de confiança simultâneo de 100(1-α)%, para qualquer combinação $\underset{\sim}{\ell}'\underset{\sim}{p} = \ell_1 p_1 + \ell_2 p_2 + ... + \ell_q p_q$, é dado por:

$$\underset{\sim}{\ell}'\hat{\underset{\sim}{p}} \pm \sqrt{\chi^2_{q-1}(\alpha)}\sqrt{\frac{\underset{\sim}{\ell}'\hat{\Sigma}\underset{\sim}{\ell}}{n}} \qquad (5.11)$$

garantindo que n-1-q seja grande. Segundo Johnson e Wichern (1988), o valor grande de n-q-1, significa que $n\hat{p}_k$ deve estar em torno de 20 para cada categoria k=1, 2, ..., q.

Exemplo 5.5

Numa amostra de n=35 cochonilhas, obtida na região de Jacuí, MG, em fevereiro de 1995, em plantas de pessegueiro tratadas, Diniz (1996) obteve os seguintes resultados:

Fêmeas adultas	Ninfa móvel	Ninfa fêmea	Ninfa macho	Total
5	11	15	4	35

Obter os intervalos de confiança simultâneos de 95% usando a aproximação de grandes amostras para proporções de insetos em cada categoria.

O vetor de proporções e a matriz de covariância amostral são:

$$\hat{\underset{\sim}{p}} = \begin{bmatrix} 0,1429 \\ 0,3143 \\ 0,4286 \\ 0,1142 \end{bmatrix} \text{ e } \hat{\Sigma} = \begin{bmatrix} 0,1225 & & \text{Sim.} & \\ -0,0449 & 0,2155 & & \\ -0,0612 & -0,1347 & 0,2449 & \\ -0,0163 & -0,0359 & -0,0489 & 0,1012 \end{bmatrix}$$

O valor de $\chi^2_3(0,05)$ é 7,815, e os intervalos são:

$$p_1: 0,1429 \pm \sqrt{7,815}\sqrt{\frac{0,1225}{35}} = 0,1429 \pm 0,1654 = [-0,0225;\ 0,3083]$$

$$p_2: 0,3143 \pm \sqrt{7,815}\sqrt{\frac{0,2155}{35}} = [0,0949;\ 0,5337]$$

$$p_3: 0,4286 \pm \sqrt{7,815}\sqrt{\frac{0,2449}{35}} = [0,1948;\ 0,6624]$$

$$p_4: 0,1142 \pm \sqrt{7,815}\sqrt{\frac{0,1012}{35}} = [-0,0361;\ 0,2645]$$

5.5. Comparações pareadas

Em muitas situações experimentais deseja-se testar o efeito ou eficácia de um tratamento. Para isso, medidas são tomadas nas unidades experimentais antes e após a aplicação do tratamento. Uma outra situação em

que esta situação pode ser de interesse, é nas situações em que na mesma unidade amostral ou experimental dois tratamentos são aplicados. Estas respostas são denominadas medidas pareadas, e podem ser analisadas calculando-se suas diferenças, eliminando a influência da variação entre as unidades experimentais ou amostrais.

Será, inicialmente, abordado o caso univariado, e, em seguida, a sua respectiva generalização para o caso multivariado. Denotando X_{1j} a resposta do tratamento 1 (ou resposta antes do tratamento) e X_{2j} a resposta do tratamento 2 (ou resposta após o tratamento) para a j-ésima unidade amostral ou experimental, em que (X_{1j}, X_{2j}) são medidas tomadas na mesma unidade amostral ou experimental, então as n diferenças:

$$D_j = X_{2j} - X_{1j} \text{ , } j=1, 2, ..., n \qquad (5.12)$$

devem refletir somente o efeito diferencial entre os tratamentos.

Assumindo que as diferenças D_j são observações independentes de uma distribuição normal $N(\delta, \sigma_D^2)$, a variável

$$t = \frac{\overline{D} - \delta}{\frac{S_D}{\sqrt{n}}} \qquad (5.13)$$

segue a distribuição de t-student com n-1 graus de liberdade, em que:

$$\overline{D} = \frac{1}{n}\sum_{j=1}^{n} D_j \quad \text{e} \quad S_D^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(D_j - \overline{D}\right)^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} D_j^2 - \frac{(\sum_{j=1}^{n} D_j)^2}{n}\right] \tag{5.14}$$

Consequentemente, para um coeficiente de confiança de 1-α, o teste para a hipótese:

$$H_0: \delta = 0 \quad (\text{efeito nulo de tratamento})$$
$$H_1: \delta \neq 0$$

pode ser realizado comparando-se | t | com $t_{n-1}(\alpha/2)$, o quantil 100(α/2) superior da distribuição de t-student com n-1 graus de liberdade.

O intervalo de confiança de 100(1-α)% para o efeito do tratamento (ou diferença de efeitos dos tratamentos) é dado pela maneira usual e apresentada a seguir.

$$\overline{D} \pm t_{n-1}(\alpha/2)\frac{S_D}{\sqrt{n}} \tag{5.15}$$

Para extensão multivariada dos procedimentos adotados no caso univariado, a seguinte notação é utilizada, pois existe a necessidade de distinguir entre os índices para os dois tratamentos (1º índice), a resposta da j-ésima unidade experimental ou amostral (2º índice) e as p variáveis (3º índice). Neste caso, X_{1jk} representa a resposta do tratamento 1 (ou medida antes de se aplicar o tratamento) na k-ésima variável tomadas na j-ésima unidade e, X_{2jk} representa a

resposta do tratamento 2 (ou medida após se aplicar o tratamento) na k-ésima variável tomadas na j-ésima unidade, sendo que k=1, 2, ..., p; j=1, 2, ..., n.

As diferenças têm a mesma notação com exceção do primeiro índice, do efeito do tratamento, que deve desaparecer. Isto se deve ao fato de as diferenças refletir o efeito diferencial dos tratamentos. Assim, D_{jk} reflete a diferença entre os tratamentos na k-ésima variável obtida na j-ésima unidade amostral ou experimental. Fazendo $\underset{\sim}{D}_j' = \begin{bmatrix} D_{j1} & D_{j2} & \cdots & D_{jp} \end{bmatrix}$, e assumindo que são distribuídos normal e independentemente $N_p(\underset{\sim}{\delta},\ \Sigma_D)$, a estatística T^2 se aplica para se realizar inferências sobre o vetor média das diferenças. Os seguintes resultados podem ser obtidos, a partir destas pressuposições assumidas.

Dadas as diferenças observadas $\underset{\sim}{D}_j' = \begin{bmatrix} D_{j1} & D_{j2} & \cdots & D_{jp} \end{bmatrix}$, j=1, 2, ..., n, um teste de a hipótese $H_o : \underset{\sim}{\delta} = \underset{\sim}{\delta}_0 \ \ Vs \ \ H_o : \underset{\sim}{\delta} \neq \underset{\sim}{\delta}_0$ deve rejeitar H_0 se o valor observado

$$T^2 = n\left(\underset{\sim}{\overline{D}} - \underset{\sim}{\delta}_0\right)' S_d^{-1}\left(\underset{\sim}{\overline{D}} - \underset{\sim}{\delta}_0\right) > \frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha) \qquad (5.16)$$

em que,

$$\underset{\sim}{\overline{D}} = \frac{1}{n}\sum_{j=1}^{n} \underset{\sim}{D}_j \quad \text{e} \quad S_D = \frac{1}{n-1}\sum_{j=1}^{n}\left(\underset{\sim}{D}_j - \underset{\sim}{\overline{D}}\right)\left(\underset{\sim}{D}_j - \underset{\sim}{\overline{D}}\right)'$$

A região de confiança de 100(1-α)% para $\underset{\sim}{\delta}$ consiste em todos os valores de $\underset{\sim}{\delta}$ tais que

$$T^2 = n(\underset{\sim}{\bar{D}} - \underset{\sim}{\delta})' S_D^{-1} (\underset{\sim}{\bar{D}} - \underset{\sim}{\delta}) \leq \frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha) \qquad (5.17)$$

Os intervalos de confiança simultâneos 100(1-α)% para as diferenças de médias individuais δ_i são dados por:

$$IC_{\delta_i}(1-\alpha): \bar{D}_i \pm \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{S_{D(ii)}}{n}} \qquad (5.18)$$

em que, $\bar{D}_i$ é o i-ésimo elemento de $\underset{\sim}{\bar{D}}$ e $S_{D(ii)}$ é i-ésimo elemento da diagonal de S_d.

Para n-p grande, $[(n-1)p/(n-p)]F_{p,n-p}(\alpha) \cong \chi_p^2(\alpha)$, e a normalidade não precisa ser assumida.

O intervalo simultâneo de Bonferroni 100(1-α)% para as médias individuais das diferenças δ_i é:

$$IC_{\delta_i}(1-\alpha): \bar{D}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right) \sqrt{\frac{S_{D(ii)}}{n}} \qquad (5.19)$$

Exemplo 5.6

Em uma amostra de n=4 fazendas em Marechal Cândido Rondon foram mensurados os valores da produção leiteira diária média por animal (X_1) e a renda total diária da produtividade de leite (X_2) antes da aplicação do plano governamental “panela cheia” e após a aplicação. Testar a hipótese de que o plano foi ineficiente em aumentar a média dos dois índices zootécnicos. Os dados da amostra são:

Antes		Após	
X_{1j1}	X_{1j2}	X_{2j2}	X_{2j2}
10	80	13	90
11	80	15	92
9	60	16	88
8	60	19	90

A hipótese a ser testada é:

$$H_0 : \underset{\sim}{\delta} = \underset{\sim}{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

As diferenças foram obtidas e são dadas por:

D_{j1}	D_{j2}
3	10
4	12
7	28
11	30

As estimativas amostrais são:

$$\underset{\sim}{\overline{D}} = \begin{bmatrix} 6,25 \\ 20,00 \end{bmatrix} \text{ e } S_D = \begin{bmatrix} 12,9167 & 34,6667 \\ 34,6667 & 109,3333 \end{bmatrix}$$

O valor da estatística T^2 pode ser computado por:

$$T^2 = 4\begin{bmatrix} 6,25 & 20 \end{bmatrix}\begin{bmatrix} 0,5195 & -0,1647 \\ -0,1647 & 0,0614 \end{bmatrix}\begin{bmatrix} 6,25 \\ 20,00 \end{bmatrix} = 14,6515$$

O valor crítico é:

$$\frac{p(n-1)}{(n-p)}F_{p,n-p}(5\%) = \frac{2\times(4-1)}{(4-2)}F_{2,4-2}(5\%) = 3\times 19 = 57$$

Como T^2=14,6515<57, então, H_0 não pode ser falseada para o valor nominal de 5% de significância.

Os intervalos de confiança simultâneos são:

$$IC_{\delta_1}(0,95): \overline{D}_1 \pm \sqrt{\frac{2(4-1)}{(4-2)}F_{2,4-2}(0,05)}\sqrt{\frac{12,9167}{4}} = 6,25 \pm 13,57 = \left[-7,32; 19,82\right]$$

$$IC_{\delta_2}(0,95): \overline{D}_2 \pm \sqrt{\frac{2(4-1)}{(4-2)}F_{2,4-2}(0,05)}\sqrt{\frac{109,3333}{4}} = 20 \pm 39,47 = \left[-19,47; 59,47\right]$$

5.6. Comparações de vetores médias de duas populações

O teste T^2 para testar a igualdade de vetores média de duas populações pode ser desenvolvido por analogia ao procedimento univariado. Este teste T^2 é apropriado para comparar a resposta média de um grupo experimental (população 1) com a resposta média "independente" de outro grupo experimental (população 2). Se possível, as unidades experimentais devem ser sorteadas para cada conjunto de observações de ambas as populações, o que abrandará o efeito da variabilidade entre unidades na comparação entre tratamentos. Apesar disto, este tipo de comparação, é em geral, menos preciso do que o caso de comparações pareadas.

Considerando uma amostra aleatória de tamanho n_1 da população 1 e uma amostra n_2 da população 2. As observações das p variáveis podem ser organizadas como:

Amostra	Estatísticas amostrais	
(População 1) $\underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, ..., \underset{\sim}{X}_{1n_1}$	$\bar{\underset{\sim}{X}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\underset{\sim}{X}_{1j}$	$S_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}\left(\underset{\sim}{X}_{1j}-\bar{\underset{\sim}{X}}_1\right)\left(\underset{\sim}{X}_{1j}-\bar{\underset{\sim}{X}}\right)'$
(População 2) $\underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, ..., \underset{\sim}{X}_{2n_2}$	$\bar{\underset{\sim}{X}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\underset{\sim}{X}_{2j}$	$S_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(\underset{\sim}{X}_{2j}-\bar{\underset{\sim}{X}}_2\right)\left(\underset{\sim}{X}_{2j}-\bar{\underset{\sim}{X}}_2\right)'$

Subscritos 1 e 2, denotam a população.

Deseja-se realizar inferência a respeito da diferença de médias populacionais ($\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$), para verificar se esta diferença é nula, o que eqüivale a afirmar que não existe efeito dos tratamentos. De forma equivalente, pode-se fazer tal inferência, testando a hipótese de igualdade dos vetores médias

populacionais ($H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$). Algumas pressuposições devem ser obedecidas para a validade dos testes e da inferência realizada. Entre as pressuposições destaca-se a necessidade de que sejam realizadas amostras aleatórias, de tamanho n_1 e n_2, de ambas as populações (população 1 com média $\underset{\sim}{\mu}_1$ e covariância Σ_1, e população 2 com média $\underset{\sim}{\mu}_2$ e covariância Σ_2); além disso, supõe-se que as observações da amostra 1 são independentemente obtidas em relação aquelas da amostra 2. Se n_1 e n_2 são pequenos, então é necessário assumir que ambas as populações sejam normais que a matriz de covariância amostral seja a mesma ($\Sigma_1 = \Sigma_2 = \Sigma$).

As matrizes de covariância S_1 e S_2 são estimadores de Σ_1 e de Σ_2, respectivamente. Consequentemente, pode-se combinar as informações de ambas as amostras para estimar a variância comum Σ, que pode ser obtida da seguinte forma.

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{5.20}$$

Para se testar a hipótese $H_0 : \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 = \underset{\sim}{\delta}_0$, considera-se os seguintes resultados:

$$E\left(\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2\right) = \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \tag{5.21}$$

$$\mathrm{Cov}\left(\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma \qquad (5.22)$$

Devido ao resultado (5.20), em que S_p é um estimador de Σ, então,

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S_p$$

é um estimador de $\mathrm{Cov}\left(\overline{X}_1 - \overline{X}_2\right)$.

Demonstra-se que o teste da razão de verossimilhança para a hipótese,

$$H_0 : \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 = \underset{\sim}{\delta}_0$$

é dado pela distância quadrada T^2. Rejeita-se H_0 se

$$T^2 = [\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2 - \underset{\sim}{\delta}_0]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p\right]^{-1} [\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2 - \underset{\sim}{\delta}_0] > \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p,n_1+n_2-p-1}(\alpha)$$

Exemplo 5.7

Os dados a seguir referem-se à produtividade e altura de plantas de duas variedades de milho (A e B). Determinar a região de 95% de confiança para diferença $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$.

A		B	
Produtividade	Altura da planta	Produtividade	Altura da planta
5,7	2,10	4,4	1,80
8,9	1,90	7,5	1,75
6,2	1,98	5,4	1,78
5,8	1,92	4,6	1,89
6,8	2,00	5,9	1,90
6,2	2,01		

As estatísticas amostrais são:

$$\bar{\underset{\sim}{X}}_1 = \begin{bmatrix} 6,57 \\ 1,99 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 1,4587 & -0,0514 \\ -0,0514 & 0,0051 \end{bmatrix}$$

$$\bar{\underset{\sim}{X}}_2 = \begin{bmatrix} 5,56 \\ 1,82 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 1,5430 & -0,0366 \\ -0,0366 & 0,0045 \end{bmatrix}$$

A matriz de variância e covariância amostral combinada é:

$$S_p = \begin{bmatrix} 1,4962 & -0,0448 \\ -0,0448 & 0,0048 \end{bmatrix}$$

Os autovalores e autovetores de S_p são:

$$\lambda_1 = 1,4975 \quad \underset{\sim}{e}_1' = [0,9995 \quad -0,0300]$$

$$\lambda_2 = 0,0035 \quad \underset{\sim}{e}_2' = [0,0300 \quad 0,9995]$$

O valor de $F_{2,8}(0,05)=4,459$. A região de confiança é dada por:

$$T^2 = [\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2 - \underset{\sim}{\delta}_0]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_p \right]^{-1} [\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2 - \underset{\sim}{\delta}_0] \leq \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p,n_1+n_2-p-1}(\alpha)$$

em que, $\underset{\sim}{\delta}_0 = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \begin{bmatrix} \mu_{11} - \mu_{21} \\ \mu_{12} - \mu_{22} \end{bmatrix}$

Desta forma com os valores amostrais, tem-se:

$$[1,01 - \delta_1 \quad 0,17 - \delta_2] \times \frac{30}{11} \begin{bmatrix} 0,9276 & 8,6575 \\ 8,6575 & 289,1364 \end{bmatrix} \times \begin{bmatrix} 1,01 - \delta_1 \\ 0,17 - \delta_2 \end{bmatrix} \leq 10,0328$$

Esta equação foi implementada no programa Maple, para se obter a elipse de 95% de confiança, apresentada na Figura 5, cujos comandos estão apresentados a seguir:

$$\text{Implicitplot}\Big(2.5298\,(1.01-d1)^2+47.2226\,(1.01-d1)\,(.17-d2)+788.5538\,(.17-d2)^2=10.0328,\ d1=\text{-}1.5\ ..\ 3.5,\ d2=.01\ ..\ 1.1,\ numpoints=12400,\ resolution=12600,\ title=\textit{Elipse de 95\% de confiança},\ thickness=2\Big)$$

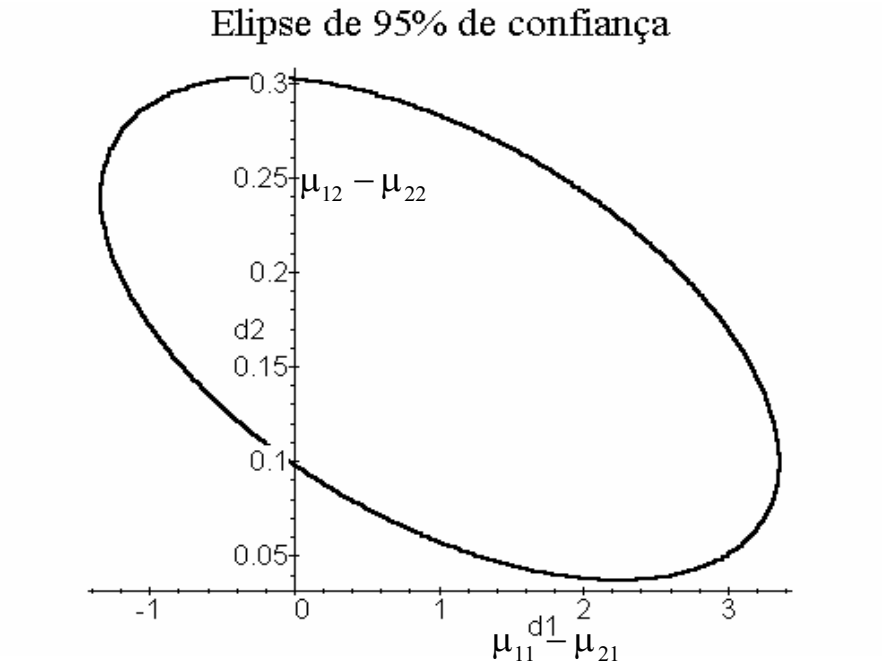

Figura 5.3. Elipse de 95% de confiança para diferença do vetor média de ambas as variedades de milho.

Verifica-se pela Figura 5.3 que a origem $\underset{\sim}{0}'$ =[0, 0], não pertence a região de confiança, indicando que as duas variedades diferem quanto ao vetor média.

Intervalos de confiança simultâneos

Para desenvolver intervalos de confiança simultâneos para um componente de $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$, adota-se o vetor $\underset{\sim}{\ell}$ tal que a combinação $\underset{\sim}{\ell}'(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$, será abrangida com probabilidade 1-α, para qualquer escolha de $\underset{\sim}{\ell}$, por

$$\underset{\sim}{\ell}'\left(\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2\right) \pm \sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}(\alpha)}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\underset{\sim}{\ell}'S_p\underset{\sim}{\ell}} \qquad (5.23)$$

Método de Bonferroni para comparações múltiplas

O intervalo de confiança simultâneo de 100(1-α)% de Bonferroni para as p diferenças entre duas médias populacionais é dado por:

$$\mu_{1i} - \mu_{2i} : (\overline{X}_{1i} - \overline{X}_{2i}) \pm t_{n_1+n_2-2}\left(\alpha/2p\right)\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{ii}} \qquad (5.24)$$

Comparações entre vetores médias quando $\Sigma_1 \neq \Sigma_2$

Quando $\Sigma_1 \neq \Sigma_2$, a distribuição das estatísticas dependem de uma medida de distância que não são independentes das covariâncias populacionais desconhecidas. Por serem desconhecidas as covariâncias populacionais, o teste de Bartlett pode ser usado para testar H_0: $\Sigma_1 \neq \Sigma_2$. No entanto, este teste é fortemente afetado se a pressuposição de normalidade for violada. O teste em

questão não pode diferenciar entre a ausência de normalidade e a heterogeneidade das covariâncias. Quando ambos n_1-p e n_2-p são grandes, pode-se evitar as complicações da desigualdade de variâncias, utilizando a elipsóide de 100(1-α)% de confiança aproximada, dada por (5.25). O problema de covariâncias heterogêneas, quando as amostras são provenientes de populações normais é conhecido como problema de Behrens-Fisher multivariado.

$$[\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2 - \underset{\sim}{\delta}_0]' \left[\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right]^{-1} [\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2 - \underset{\sim}{\delta}_0] \leq \chi^2_p(\alpha) \tag{5.25}$$

O intervalo de confiança simultâneo aproximado é dado por:

$$\underset{\sim}{\ell}'\left(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2\right) \pm \sqrt{\chi^2_p(\alpha)} \sqrt{\underset{\sim}{\ell}'\left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)\underset{\sim}{\ell}} \tag{5.26}$$

Sete soluções para o problema multivariado de Behrens-Fisher foram estudadas por Christensen e Rencher (1997) por meio de simulação Monte Carlo, comparando as taxas de erro tipo I e o poder destas soluções. Algumas dessas soluções estudadas por estes autores são apresentadas a seguir.

a) Aproximação de Bennett

A primeira dessas alternativas é àquela estudada por Bennett (1951), a qual assume que $n_2 \geq n_1$, o que não é limitante. Para contornar o problema, caso essa condição não seja atendida, basta trocar os nomes das amostras, isto é, a amostra 1 passa ser a amostra 2 e vice-versa. Inicialmente é necessário calcular os vetores $\underset{\sim}{Z}_j,\ j = 1, 2, \cdots, n_1$ da seguinte forma.

$$\underset{\sim}{Z}_j = \underset{\sim}{X}_{1j} - \sqrt{\frac{n_1}{n_2}}\underset{\sim}{X}_{2j} + \frac{1}{\sqrt{n_1 n_2}}\sum_{j=1}^{n_1}\underset{\sim}{X}_{2j} - \frac{1}{n_2}\sum_{k=1}^{n_2}\underset{\sim}{X}_{2k} \qquad (5.27)$$

Em seguida calcula-se a média ($\overline{\underset{\sim}{Z}}$) e a covariância ($\mathbf{S}_Z$) a partir das n_1 observações amostrais p-variadas obtidas em 5.27. A estatística

$$T^2 = n_1 \overline{\underset{\sim}{Z}}' S_Z^{-1} \overline{\underset{\sim}{Z}} \qquad (5.28)$$

possui distribuição T^2 de Hotelling com dimensão p e $\nu = n_1 - 1$ graus de liberdade, que pode ser dada pela expressão geral 5.4.1.

b) Aproximação de James

A aproximação de James (1954) envolve uma correção do valor de χ^2 quando se utiliza a estatística T^{*2}, definida por:

$$T^{*2} = [\underset{\sim}{\overline{X}}_1 - \underset{\sim}{\overline{X}}_2]' \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]^{-1} [\underset{\sim}{\overline{X}}_1 - \underset{\sim}{\overline{X}}_2] \dot{\sim} \chi^2_p \tag{5.29}$$

James (1954) propõe valores críticos ajustados ao invés de utilizar a distribuição aproximada de qui-quadrado diretamente. Os valores críticos propostos por James (1954) são dados em 5.30.

$$\chi^2_p(\alpha) \times \left(A + B\chi^2_p(\alpha) \right) \tag{5.30}$$

em que $\chi^2_p(\alpha)$ é o quantil superior α da distribuição de qui-quadrado e A e B são dados em 5.31 e 5.32.

$$A = 1 + \frac{1}{2p} \sum_{i=1}^{2} \left\{ \frac{1}{n_i - 1} \left[tr\left(S_e^{-1} \frac{S_i}{n_i} \right) \right]^2 \right\} \tag{5.31}$$

$$B = \frac{1}{2p(p+2)} \sum_{i=1}^{2} \frac{1}{n_i - 1} \left\{ tr\left[2\left(S_e^{-1} \frac{S_i}{n_i} \right)^2 \right] + \left[tr\left(S_e^{-1} \frac{S_i}{n_i} \right) \right]^2 \right\} \tag{5.32}$$

em que

$$S_e = \frac{S_1}{n_1} + \frac{S_2}{n_2} \tag{5.33}$$

c) Aproximação de Yao

A aproximação de Yao (1965) é uma extensão da aproximação de Welch para os graus de liberdade. A estatística (T^{*2}) apresentada em 5.29 é aproximada por uma T^2 de Hotelling com dimensão p e graus de liberdade ν dados por 5.34.

$$\frac{1}{\nu} = \frac{1}{\left(T^{*2}\right)^2} \sum_{i=1}^{2} \left\{ \frac{1}{n_i - 1} \left[\left(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2 \right)' S_e^{-1} \frac{S_i}{n_i} S_e^{-1} \left(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2 \right) \right]^2 \right\} \tag{5.34}$$

d) Aproximação de Johansen

A aproximação de Johansen (1980) usa a estatística T^{*2} de (5.29) dividida por uma constante C para que a estatística resultante tenha distribuição aproximada pela distribuição F com ν_1=p e ν_2=ν graus de liberdade. Assim, os valores necessários para calcular a estatística F_c de Johansen (1980) são:

$$F_c = \frac{T^{*2}}{C} \tag{5.35}$$

$$C = p - \frac{2D + 6D}{p(p-1) + 2} \tag{5.36}$$

$$D = \sum_{i=1}^{2} \frac{1}{2(n_i - 1)} \left\{ \left[tr\left(I - V^{-1}V_i\right)^2 \right] + \left[tr\left(I - V^{-1}V_i\right) \right]^2 \right\} \tag{5.37}$$

$$\nu = \frac{p(p+2)}{3D} \tag{5.38}$$

com $V_i=(S_i/n_i)^{-1}$ para i=1 ou 2 e $V=V_1+V_2$.

e) Aproximação de Nel e Van der Merwe

A aproximação de Nel e Van der Merwe (1986) usa a estatística T^{*2} de 5.29, a qual é aproximada pela T^2 de Hotelling com dimensão p e graus de liberdade ν, em que

$$\nu = \frac{tr\left(S_e\right)^2 + \left[tr\left(S_e\right) \right]^2}{\frac{1}{n_1 - 1}\left\{ tr\left(\frac{S_1}{n_1}\right)^2 + \left[tr\left(\frac{S_1}{n_1}\right) \right]^2 \right\} + \frac{1}{n_2 - 1}\left\{ tr\left(\frac{S_2}{n_2}\right)^2 + \left[tr\left(\frac{S_2}{n_2}\right) \right]^2 \right\}} \tag{5.39}$$

É conveniente chamar a atenção para o fato de que nas expressões anteriormente apresentadas aparece um termo como: $tr(A)^2$. Esse termo significa que é necessário calcular tr(A*A). Em outras ocasiões os termos eram $[tr(A)]^2$, o

que significa que o traço da matriz A deve ser calculado e o seu quadrado é a resposta almejada.

f) Aproximação de Kim

A aproximação de Kim (1992) é a mais elaborada de todas e também se refere a uma extensão da aproximação dos graus de liberdade de Welch, como acontece com o procedimento de Yao (1965). O procedimento de Kim requer a maximização de um par de formas quadráticas dado por:

$$d = \frac{\underset{\sim}{q}' \frac{S_1}{n_1} \underset{\sim}{q}}{\underset{\sim}{q}' \frac{S_2}{n_2} \underset{\sim}{q}}$$

A maximização desse par de formas quadráticas resulta na solução do sistema de equações homogêneas dado por 5.40.

$$\left[\frac{S_1}{n_1} - d_k \frac{S_2}{n_2} \right] \underset{\sim}{q}_k = \underset{\sim}{0} \qquad (5.40)$$

A solução desse sistema pode ser obtida conforme descrito no capítulo 2. O autovalores d_k e os autovetores $\underset{\sim}{q}_k$ (k=1, 2, ..., p) são utilizados para

definir a matriz D=diag(d_1, d_2, ..., d_p) e $Q = \left[\underset{\sim}{q}_1 \quad \underset{\sim}{q}_2 \quad \cdots \quad \underset{\sim}{q}_p \right]$. A partir dessas matrizes definem-se as seguintes quantidades:

$$\underset{\sim}{w} = Q'\left(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2 \right) \tag{5.41}$$

$$r = \left(\prod_{k=1}^{p} d_k \right)^{\frac{1}{2p}} \tag{5.42}$$

$$\ell_k = \frac{d_k + 1}{\left(\sqrt{d_k} + r \right)^2} \tag{5.43}$$

$$c = \frac{\sum_{k=1}^{p} \ell_k^2}{\sum_{k=1}^{p} \ell_k} \tag{5.44}$$

$$f = \frac{\left(\sum_{k=1}^{p} \ell_k \right)^2}{\sum_{k=1}^{p} \ell_k^2} \tag{5.45}$$

O próximo passo é calcular a estatística do teste que tem uma aproximação F (5.47) com ν_1=f e ν_2=ν-p+1 graus de liberdade. O valor ν é definido em 5.48.

$$G = \underset{\sim}{w}'\left(D^{1/2} + rI\right)^{-1}\left(D^{1/2} + rI\right)^{-1}\underset{\sim}{w} \tag{5.46}$$

$$F_c = \frac{(\nu - p + 1)G}{cf\nu} \tag{5.47}$$

$$\frac{1}{\nu} = \frac{1}{n_1 - 1}\left[\frac{\underset{\sim}{w}'D(D+I)^{-2}\underset{\sim}{w}}{\underset{\sim}{w}'(D+I)^{-1}\underset{\sim}{w}}\right]^2 + \frac{1}{n_2 - 1}\left[\frac{\underset{\sim}{w}'(D+I)^{-2}\underset{\sim}{w}}{\underset{\sim}{w}'(D+I)^{-1}\underset{\sim}{w}}\right]^2 \tag{5.48}$$

Teste de Bartlett para igualdade de matrizes de covariâncias

O teste da razão de verossimilhança para igualdade de matrizes de covariâncias de populações Wishart foi apresentado por Bartlett (1947). Este autor demonstrou que sob a hipótese

$$H_o : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_n = \Sigma$$

a estatística 5.49 tem distribuição assintótica de qui-quadrado com $\nu=(n-1)p(p+1)/2$ graus de liberdade. Em que, n é o número de grupos ou subpopulações amostradas, p é a dimensão das matrizes.

$$\chi_c^2 = -\left[1 - \left(\sum_{j=1}^{n}\frac{1}{n_j - 1} - \frac{1}{N - n}\right)\left(\frac{2p^2 + 3p - 1}{6(p+1)(n-1)}\right)\right] \times \left[\sum_{j=1}^{n}\left(n_j - 1\right)\ln\left|S_j\right| - (N - n)\ln\left|S_p\right|\right] \tag{5.49}$$

em que: S_j é o estimador não viesado da covariância da sub-população j, baseado em n_j observações multivariadas de dimensão p; $N = \sum_{j=1}^{n} n_j$; j=1, 2, ..., n, e

$$S_p = \frac{\sum_{j=1}^{n} (n_j - 1) S_j}{\sum_{j=1}^{n} n_j - n}$$

Exemplo 5.8. Testar a hipótese de igualdade das covariâncias de 2 populações. Uma amostra de 11 observações foi obtida da primeira população e outra de 15 da segunda. Duas variáveis foram mensuradas, sendo as estimativas amostrais apresentadas a seguir.

$$S_1 = \begin{bmatrix} 0,51964 & 0,44700 \\ 0,44700 & 0,47600 \end{bmatrix} \text{ com } n_1=11 \text{ e } S_2 = \begin{bmatrix} 0,85143 & 0,73786 \\ 0,73786 & 1,54828 \end{bmatrix} \text{ com } n_2=15$$

O valor de N=11+15=26 e de n=2 (populações). A hipótese a ser testada é:

$$H_o : \Sigma_1 = \Sigma_2 = \Sigma$$

Os demais valores necessários para a realização do teste de hipótese são:

$$\ln|S_1| = -3,0692181 ; \ \ln|S_2| = -0,2564228 ; \text{ e } \ln|S_p| = -0,9031351$$

Logo,

$$\chi_c^2 = -\left[1-\left(\frac{1}{10}+\frac{1}{14}-\frac{1}{24}\right)\left(\frac{2\times 2^2+3\times 2-1}{6\times 3\times 1}\right)\right]\times$$

$$\times\left[\left(10\times\left(-3,0692181\right)+14\times\left(-0,2564228\right)\right)-24\times\left(-0,9031351\right)\right]=$$

$$=11,43$$

Os graus de liberdade são ν=1x2x3/2=3 e os valores críticos 5% e 1% da distribuição de qui-quadrado são $\chi_3^2(0,05)=7,8147$ e $\chi_3^2(0,01)=11,3448$. Como o valor calculado (11,43) é superior aos valores críticos, rejeita-se H_0 com P<0,01. Portanto, existem evidências de que as covariâncias das duas populações não sejam iguais.

5.7. Exercício

5.7.1. A matriz X, apresentada a seguir, representa uma amostra de n=4 observações retiradas de uma distribuição normal bivariada.

$$X = \begin{bmatrix} 11 & 2 \\ 10 & 4 \\ 9 & 3 \\ 10 & 6 \end{bmatrix}$$

a) Teste a hipótese de que $\underset{\sim}{\mu}'_0$ = [9 2] seja um valor plausível para representar a média populacional.

b) Obtenha a região de 95% de confiança e esboce graficamente a mesma, destacando o valor hipotético nessa região.

5.7.2. Com os dados do exercício 5.7.1, determine os intervalos de confiança simultâneo para os componentes de média individual por:

a) T^2 de Hotteling

b) Procedimento de Bonferroni

c) Teste de t-student univariado.

5.7.3. Com os dados do exemplo 5.3, utilizando as duas primeiras variáveis, teste a pressuposição de normalidade univariada (marginal) e bivariada, utilizando os procedimentos apresentados no capítulo 4.

5.7.4. Utilizando os dados do exemplo 5.5, faça o IC simultâneo para proporções de 90% de confiança.

5.7.5. Os dados abaixo se referem ao peso e ao teor de proteína medidos em 6 animais antes e após um período de dieta balanceada. Teste a hipótese de que não houve efeito da dieta. Determinar a região de confiança e o esboço da região de confiança, o intervalo de confiança simultâneo e de Bonferroni, no nível de 5% de probabilidade.

Antes		Após	
Peso	Teor de proteína (%)	Peso	Teor de proteína (%)
250	10	280	12
300	12	320	16
350	13	360	13
320	15	380	18
400	9	410	15
320	11	350	12

5.7.6. Com os dados do exemplo 5.7, reapresentados a seguir, obter os intervalos de confiança de 95% simultâneos e de Bonferroni, para as diferenças de médias marginais. Compare os resultados com a Figura 5.3, e obtenha conclusões de interesse.

A		B	
Produtividade	Altura da planta	Produtividade	Altura da planta
5,7	2,10	4,4	1,80
8,9	1,90	7,5	1,75
6,2	1,98	5,4	1,78
5,8	1,92	4,6	1,89
6,8	2,00	5,9	1,90
6,2	2,01		

5.8. Referências

ANDERSON, T.W. **An introduction to multivariate statistical analysis.** 2th edition. John Wiley & Sons. New York, 1984. 675p.

BENNETT, B.M. Note on a solution of the generalized Behrens-Fisher problem, **Annals of the Institute of Statistical Mathematics**, v.2, p.97-90, 1951.

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

CHRISTENSEN, W.F.; RENCHER, A.C. A comparison of type I rates and power levels for seven solutions to the multivariate Behrens-Fisher problem. **Communication in Statistics-Simula.,** v.26, n.4, p.1251-1273, 1997.

DINIZ, L de C. **Dinâmica populacional do piolho de são josé** *Quadraspidiotus perniciosus* **(Comostock, 1881) (Homoptera: Dispididae) em pessegueiro, no município de Jacuí - Minas Gerais.** UFLA, Lavras, MG, 1996. 61p. (dissertação de mestrado).

JAMES, G.S. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown, **Biometrika**, v.41, p.19-43, 1954.

JOHANSEN, S. The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regresion, **Biometrika**, v.67, n.1, p.85-92, 1980.

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

KIM, S. A pratical solution to the multivariate Behrens-Fisher problem, **Biometrika**, v.79, n.1, p.171-176, 1992.

KRZANOWSKI, W.J. **Principles of multivariate analysis. A user's perspective.** Oxford, 3th edition, 1993. 563p.

MOMENTÉ, V.G. **Comparações entre diferentes tipos de famílias clonais para o melhoramento genético da batata (***Solanum tuberosum* L**.)**. ESAL, Lavras, MG, 1994. 83p. (dissertação de mestrado).

NEL, D.G. VAN DER MERWE, C.A. A solution to the multivariate Behrens-Fisher problem. **Communications in Statistics: Theory and Methods**, v.15, p.3719-3735, 1986.

YAO, Y. An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. **Biometrika**, v.52, n.1, p.139-147, 1965.