Técnicas Multivariadas em Saúde

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

1. Introdução
2. Distribuições de Probabilidade Multivariadas
3. Representação de Dados Multivariados
4. Testes de Significância c/ Dados Multivariados
5. Análise de Componentes Principais
6. Análise Fatorial
7. Análise de Correlação Canônica
8. Análise de Conglomerados
9. Análise Discriminante
10. Análise de Correspondência
11. Referências

Comparações de Médias Multivariadas

Testes de Significância

- Amostra de dados multivariados:
  - √ É sempre possível aplicar testes de significância para as variáveis isoladamente
    - A diferença entre as médias para dois grupos pode ser testada separadamente para cada variável

      H_0: $\mu_1 = \mu_2$. ou $\mu_1 - \mu_2 = 0$
    - Cada teste tem uma certa probabilidade de levar a uma conclusão errada

      (nível de significância)

- Importante:
  - √ O uso repetido de testes de significância individuais pode aumentar a probabilidade de se encontrar falsamente pelo menos uma diferença significante.
- Há maneiras de ajustar níveis de significância para permitir a aplicação simultânea de muitos testes
  - √ Pode ser preferível conduzir um único teste usando a informação conjunta de todas as variáveis.

Comparação de Valores Médios – Duas Amostras

Exemplo

- Bumpus (1898)
- Pardais sobreviventes de tempestade
  - √ Dados de 1/Fev/1898
  - √ Medidas morfológicas e peso de 49 pássaros fêmeas
  - √ 28 morreram, 21 não morreram
- Dados: *birds.csv* (*birds.txt*)

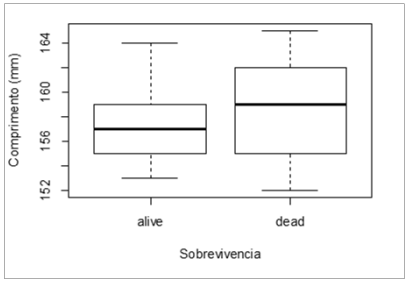
- Variável de interesse:
  - √ tl: comprimento total (bico à ponta da cauda), em mm
- Amostras:
  - √ Grupo 1: 21 sobreviventes
  - √ Grupo 2: 28 não-sobreviventes
  - √ Assume-se que sejam efetivamente amostras aleatórias de populações muito maiores de sobreviventes e não sobreviventes

- Questão (sob o ponto de vista das populações):
  - √ As duas médias populacionais são significativamente diferentes?
- Questão (sob o ponto de vista das amostras):
  - √ A diferença média observada é tão grande que é improvável que ela tenha ocorrido caso as médias populacionais sejam iguais?

- Estatísticas descritivas da amostra:

Sobreviventes			Não sobreviventes		
n_1	$\bar{x}_1$	s_1	n_2	$\bar{x}_2$	s_2
21	157,38	3,32	28	158,43	3,88

- Análise gráfica:



- Comandos em R:
  - √ Estatísticas descritivas

```
> pardal <-  read.csv2("pardal2.csv", header=TRUE)
> attach(pardal)
> descritiva<-aggregate(comp, by=list(sobrev),
            FUN=function(x) c(mean = mean(x), sd = sd(x)))
> linhas<- c("alive", "dead")
> colunas <- c("média", "desvio")
> rownames(descritiva) <- linhas
> colnames(descritiva) <- colunas
> descritiva
      média desvio.mean  desvio.sd
alive     1  157.380952   3.323796
dead      2  158.428571   3.881853
```

  - √ Box-plot

```
> # Box-plot
>
> dev.new(width=5, height=4)
> boxplot(comp~sobrev, names = c("alive", "dead"),
+ xlab="Sobrevivencia", ylab="Comprimento (mm)")
```


Teste t Combinado para $\mu_1 - \mu_2$

- Uma única variável X, considerando que a mesma variação nas duas populações
  - √ Caso Homocedástico
- Hipótese nula:
  - √ H_0: $\mu_1 - \mu_2 = 0$.
- Estatística de teste:
  - √ S_p: desvio-padrão combinado

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Distribuição amostral:

$$T_0 \overset{H_0}{\sim} t_{(n_1+n_2-2)}$$

Estimador Combinado de σ^2

- Se a variância populacional é a mesma para as duas populações, parece razoável combinar as duas variâncias amostrais S_1^2 e S_2^2.
  - √ Como?
- Estimador combinado de σ^2:

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

  - √ S_p^2 pode ser escrito como uma média ponderada:

$$S_p^2 = \frac{(n_1 - 1)}{n_1 + n_2 - 2}S_1^2 + \frac{(n_2 - 1)}{n_1 + n_2 - 2}S_2^2 = wS_1^2 + (1 - w)S_2^2$$

    - Os pesos dependem do tamanho das amostras

- Comentários:
  - √ O teste é robusto para desvios moderados de normalidade das populações
    (particularmente para tamanhos amostrais ≥ 20)
  - √ A suposição de homocedasticidade não é crucial se a razão das duas variâncias verdadeiras (populacionais) estiver entre 0,4 e 2,5)
    - O teste é particularmente robusto se os tamanhos amostrais forem iguais ou quase iguais.
  - √ Caso não haja considerações de não normalidade, mas as variâncias populacionais forem bastante desiguais
    - Teste t modificado (Teste de Welch)

Exemplo

- Pardais sobreviventes:
  - √ Consideradas variâncias populacionais iguais

```
> t.test(comp ~ sobrev, var.equal=TRUE)

        Two Sample t-test

data:  comp by sobrev
t = -0.993, df = 47, p-value = 0.3258
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.170113  1.074874
sample estimates:
mean in group 1 mean in group 2
       157.3810        158.4286
```

  - √ Diferença não é significativamente diferente de zero ao nível de 5%
    - Não há evidência de uma diferença na média populacional entre sobreviventes e não-sobreviventes

Teste t Modificado (Welch)

- Supondo variâncias populacionais diferentes:
  - √ H_0: $\mu_1 - \mu_2 = 0$ verdadeira:
  - √ Estatística de teste:

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

  - √ Distribuição amostral:

$$T_0^* \overset{H_0}{\sim} t_\nu$$

    - Graus de liberdade:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

    - Para υ não inteiro, arredonde para menor inteiro mais próximo (mais conservativo para rejeitar H_0)
      (pacotes estatísticos calculam com valores não inteiros)

- Se há indícios de não-normalidade e de variâncias desiguais:
  - √ Pode não ser possível testar confiavelmente a diferença entre as médias populacionais
  - √ Pode ocorrer um número excessivo de resultados significantes

Exemplo

- Pardais sobreviventes:
  - √ Consideradas variâncias populacionais diferentes

```
> teste.comp<-t.test(comp ~ sobrev, var.equal=FALSE) # heterocedasticidade
> teste.comp

        Welch Two Sample t-test

data:  comp by sobrev
t = -1.0155, df = 46.108, p-value = 0.3152
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.124040  1.028801
sample estimates:
mean in group 1 mean in group 2
       157.3810        158.4286
```

  - √ Diferença não é significativamente diferente de zero ao nível de 5%
    - Não há evidência de uma diferença na média populacional entre sobreviventes e não-sobreviventes

Comparação de Médias para 2 Amostras – Caso Multivariado

- Exemplo:
  - √ Teste t dos pardais para as variáveis:
    - tl: comprimento total (bico à ponta da cauda), em mm.
    - ae: extensão alar, (ponta a ponta de asas), em mm.
    - bh: comprimento bico e cabeça, em mm.
    - hl: comprimento do úmero (osso braço), em polegadas.
    - kl: comprimento da quilha do esterno, em polegadas

- Questão:
  - √ Algumas dessas variáveis parecem ter valores médios diferentes para as duas populações?
    (sobreviventes e não sobreviventes)

- Teste t combinado:
  - √ Supondo-se mesma variância para duas populações
    (graus de liberdade: 47)

	Sobreviventes		Não-sobreviventes			
Variável	$\bar{x}_1$	s_1^2	$\bar{x}_2$	s_2^2	t	p-valor
Comprimento total	157,38	11,05	158,43	15,07	–0,99	0,327
Extensão alar	241,00	17,50	241,57	32,55	–0,39	0,698
Comprimento do bico e cabeça	31,43	0,53	31,48	0,73	–0,20	0,842
Comprimento do úmero	18,50	0,18	18,45	0,43	0,33	0,743
Comprimento da quilha do esterno	20,81	0,58	20,84	1,32	–0,10	0,921

  - √ Em nenhum caso há qualquer evidência de uma diferença na média populacional entre sobreviventes e não-sobreviventes

- Comentários:
  - √ Se considerarmos as 5 variáveis conjuntamente, há sugestão de uma diferença entre a população dos dois grupos?
    - – O total aponta para diferenças de médias entre populações de pardais sobreviventes e não sobreviventes?
- Teste multivariado para verificar a diferença entre os dois vetores de médias populacionais:
  - √ Teste T^2 de Hotteling
    (generalização da estatística t^2)

Teste T^2 de Hotteling

- Sejam duas amostras com tamanhos n_1 e n_2 de p variáveis $X_1, X_2, \ldots X_p$.
  - √ Vetores de médias e matrizes de covariâncias amostrais
    - – Amostra 1: $\bar{\mathbf{X}}_1$ e $\mathbf{S}_1$.
    - – Amostra 2: $\bar{\mathbf{X}}_2$ e $\mathbf{S}_2$.
  - √ Assumindo-se igualdades das matrizes de covariâncias populacionais ($\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$)
    - – Estimativa combinada da matriz de covariâncias

$$\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}.$$

- Estatística T^2 de Hotteling:

$$T_0^2 = \frac{n_1 n_2 (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{n_1 + n_2}.$$

  - √ Valores grandes de T^2 oferecem evidência de diferença entre os dois vetores de médias populacionais
- Distribuição amostral sob H_0:

$$F_0 = \frac{(n_1 + n_2 - p - 1)\, T_0^2}{(n_1 + n_2 - 2)p} \overset{H_0}{\sim} F_{p,(n_1+n_2-p-1)}.$$

- Pressupostos para aplicação do teste:
  - √ Normalidade multivariada
    - Amostras provenientes de populações normalmente distribuídas
  - √ Homocedasticidade:
    - Matrizes de covariâncias populacionais iguais

- Violações dos pressupostos
  - √ Em geral, desvios moderados de normalidade multivariada não afetam seriamente a conclusão do teste (Carter et al., 1979)
  - √ Se os tamanhos das amostras são iguais (ou quase iguais), diferenças moderadas entre as matrizes de covariâncias populacionais não são importantes.
  - √ Pode-se usar teste modificado se as duas matrizes de covariâncias populacionais e os tamanhos amostrais são muito diferentes
    - Apoia-se na suposição de normalidade multivariada

Exemplo

- √ Pardais: Teste T^2 para as variáveis:
  - tl: comprimento total (bico à ponta da cauda), em mm.
  - ae: extensão alar, (ponta a ponta de asas), em mm.
  - bh: comprimento bico e cabeça, em mm.
  - hl: comprimento do úmero (osso braço), em polegadas.
  - kl: comprimento da quilha do esterno, em polegadas

- Comandos em R:
  - √ Vetores de médias amostrais:

```
> mean.list <- lapply(unique(sobrev),
+                     function(x)colMeans(pardal[pardal$sobrev==x,-1], na.rm=T))
> mean.list
[[1]]
     comp       ext    cabeca     umero   esterno
157.38095 241.00000  31.43333  18.50000  20.80952

[[2]]
     comp       ext    cabeca     umero   esterno
158.42857 241.57143  31.47857  18.44643  20.83929
```

  - √ Matrizes de covariâncias amostrais:

```
> cov.list <- lapply(unique(sobrev),
+                    function(x)cov(pardal[pardal$sobrev==x,-1],
+                                   use="na.or.complete"))
> cov.list
[[1]]
             comp   ext    cabeca  umero   esterno
comp    11.047619  9.10 1.5566667 0.8700 1.2861905
ext      9.100000 17.50 1.9100000 1.3100 0.8800000
cabeca   1.556667  1.91 0.5313333 0.1890 0.2396667
umero    0.870000  1.31 0.1890000 0.1760 0.1325000
esterno  1.286190  0.88 0.2396667 0.1325 0.5749048

[[2]]
             comp       ext    cabeca     umero   esterno
comp    15.068783 17.190476 2.2428571 1.7460317 2.9306878
ext     17.190476 32.550265 3.3978836 2.9502646 4.0656085
cabeca   2.242857  3.397884 0.7284127 0.4695503 0.5590212
umero    1.746032  2.950265 0.4695503 0.4344312 0.5058862
esterno  2.930688  4.065608 0.5590212 0.5058862 1.3209921
```

- T^2 de Hotteling – Comandos em R:
  (Pacote Hotelling)
  √ T^2 de Hotteling

```
> library(Hotelling)
> split.dados = split(pardal[,-1],pardal$sobrev)
> x = split.dados[[1]]
> y = split.dados[[2]]
> T2<-hotelling.stat(x, y)
> T2$statistic
[1] 2.823698
```

  √ Teste T^2 de Hotteling

```
> ajuste <- hotelling.test(.~sobrev, data = pardal)
> ajuste
Test stat:  0.51668
Numerator df:  5
Denominator df:  43
P-value:  0.7622
```

Técnicas Multivariadas em Saúde - 2015

Testes Multivariados vs. Testes Univariados

- Testes univariados não significantes e teste multivariado significante:
  √ Teste global acumula evidências das variáveis individuais
- Teste multivariado não significante e alguns testes univariados significantes
  √ Evidência de diferença fornecida pelas variáveis significantes é superada pela evidência de não diferença fornecida pelas outras variáveis

Técnicas Multivariadas em Saúde - 2015

- Erro tipo I (univariado):
  √ Encontrar resultado significante quando, em realidade, as duas amostras vêm de populações com mesma média.
  √ Um teste univariado, com $\alpha = 5\%$
    – 95% de resultado não significante quando as médias populacionais são as mesmas
  √ p testes univariados (supostos independentes):
    – $0{,}95^p$: probabilidade de se obter nenhum resultado significante quando as médias populacionais são as mesmas
    – $1 - 0{,}95^p$: probabilidade de pelo menos um teste significante quando as médias populacionais são as mesmas
    – Se p = 5, então $1 - 0{,}95^5 = 0{,}23$ (se testes independentes)

Técnicas Multivariadas em Saúde - 2015

- Quanto mais testes individuais são feitos, maior é a probabilidade de se obter, ao acaso, pelo menos um resultado significante
  √ Um teste multivariado (por exemplo, o T^2 de Hotteling com $\alpha = 5\%$) oferece uma probabilidade de 0,05 de erro tipo I, independentemente do número de variáveis envolvidas

Técnicas Multivariadas em Saúde - 2015

Ajuste de Bonferroni

- Controla a probabilidade total de erro tipo I, quando são aplicados vários testes univariados:
- Exemplo: Deseja-se nível de significância global de 5%
  - √ Condução de p testes individuais (univariados) com nível de significância $\alpha = 5/p$ %
  - √ Probabilidade de se obter pelo menos um resultado significante quando as médias populacionais são iguais é menor ou igual a 5%

- Restrição ao uso do ajuste de Bonferroni:
  - √ Se p é grande, os níveis de significância aplicados aos testes individuais se tornam extremos
  - √ Para $\alpha_{global} = 5\%$ e p = 10, tem-se $\alpha_{individual} = 0{,}5\%$
- Em geral, o uso de um único teste multivariado é mais adequado do que conduzir um grande número de testes univariados
  - √ O teste multivariado considera apropriadamente a correlação entre as variáveis

Comparação de Variabilidade – Duas Amostras

Comparação Variâncias – Duas Amostras

- Comparação de variação no caso univariado
- Teste F:
  - √ Estatística de teste: $F_0 = \frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F_{(n_1-1),(n_2-1)}.$
    - n_j: tamanho da j-ésima amostra
    - s^2_j: variância da j-ésima amostra
  - √ Razões significantivamente diferentes de 1 evidenciam que as amostras são de duas populações com variâncias diferentes

- Restrição ao uso do teste F:
  - √ Ele é bastante sensível à suposição de normalidade
  - √ Resultado significante pode ser devido ao fato de uma variável não ser normalmente distribuída e não pela desigualdade entre as variâncias
  - √ Alguns autores não recomendam o uso do teste F para comparar variâncias

- Teste de Levene
  - √ Alternativa robusta para comparar variâncias
  - √ Princípio do procedimento:
    - Transforma os dados originais em cada amostra em desvios absolutos da média amostral
    - Teste t para verificar um diferença significativa entre os desvios médios nas duas amostras
- Teste Schultz, 1983
  - √ Procedimento:
    - Similar ao teste de Levene, utilizando desvios absolutos da mediana amostral

- Variável `comp`:
  - √ Estatísticas descritivas:

```
> descritiva<-aggregate(pardal$comp, by=list(pardal$sobrev),
+ FUN=function(x) c(mediana = median(x), variancia = var(x)))
> linhas<- c("alive", "dead")
> rownames(descritiva$x) <- linhas
> descritiva$x
      mediana variancia
alive     157  11.04762
dead      159  15.06878
```

  - √ Teste de Levene – medianas:

```
> library(lawstat)
> ajuste <- levene.test(pardal$comp,sobrev, location="median")
> ajuste

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  pardal$comp
Test Statistic = 1.447, p-value = 0.235
```

  - √ Valor observado não evidencia heterocedasticidade
    - Se ocorresse seleção estabilizadora, não-sobreviventes seriam mais variáveis

- Mediana – Demais variáveis

```
> medianas$sobrev <- factor(medianas$sobrev)
> levels(medianas$sobrev) <- c("alive", "dead")
> medianas<-aggregate(. ~ sobrev, data=pardal, FUN=median)
> medianas$sobrev <- factor(medianas$sobrev)
> levels(medianas$sobrev) <- c("alive", "dead")
> medianas
  sobrev comp ext cabeca umero esterno
1  alive  157 240   31.4  18.5    20.6
2   dead  159 242   31.5  18.5    20.7
```

- Teste de Levene – demais variáveis:
  - √ `extensao`

```
> ajuste.ext <- levene.test(pardal$ext,sobrev, location="median")
> ajuste.ext

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  pardal$ext
Test Statistic = 1.403, p-value = 0.2422
```

  - √ `cabeca`

```
> ajuste.cabeca <- levene.test(pardal$cabeca,sobrev, location="median")
> ajuste.cabeca

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  pardal$cabeca
Test Statistic = 0.6638, p-value = 0.4193
```

  - √ `Umero`

```
> ajuste.umero <- levene.test(pardal$umero, sobrev, location="median")
> ajuste.umero

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  pardal$umero
Test Statistic = 3.6559, p-value = 0.06198
```


- Teste de Levene – demais variáveis:
  - √ `esterno`

```
> ajuste.esterno <- levene.test(pardal$esterno,sobrev, location="median")
> ajuste.esterno

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  pardal$esterno
Test Statistic = 1.984, p-value = 0.1655
```


- Teste de Levene – Variáveis individuais:
  - √ Transformação para desvios das medianas amostrais (graus de liberdade: 47)

	Sobreviventes		Não-sobreviventes			
Variável	$\tilde{x}_1$	s_1^2	$\tilde{x}_2$	s_2^2	Estatística	p-valor
Comprimento total	157	11,05	159	15,07	1,447	0,235
Extensão alar	240	17,50	242	32,55	1,403	0,242
Comprimento do bico e cabeça	31,4	0,53	31,5	0,73	0,664	0,419
Comprimento do úmero	18,5	0,18	18,5	0,43	3,656	0,062
Comprimento da quilha do esterno	20,6	0,58	20,7	1,32	1,984	0,166

  - √ Comprimento do úmero próximo de nível de significância de 5%
    - Pode ser indício de seleção estabilizadora

Comparação de Variação – Caso Multivariado

- Comparação de variação no caso univariado
- Teste F:
  - √ Estatística de teste: $F_0 = \frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F_{(n_1-1),(n_2-1)}.$
    - n_j: tamanho da j-ésima amostra
    - s_j^2: variância da j-ésima amostra
  - √ Razões significativamente diferentes de 1 evidenciam que as amostras são de duas populações com variâncias diferentes

Comparação de Variação – Caso Multivariado

- Procedimento usando o princípio do teste de Levene:
  - √ Procedimento robusto
  - √ Transformação dos dados em desvios absolutos das médias (ou medianas amostrais)

Duas amostras apresentam quantidades de variação significativamente diferentes ⇨ Vetores de médias dos valores transformados são significativamente diferentes

⇩

Teste T^2 de Hotteling

- Teste sugerido por Van Valen (1978)

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ijk} - \bar{x}_{jk})^2}$$

  - x_{ijk}: valor da variável X_k para o i-ésimo indivíduo na amostra j
  - $\bar{x}_{jk}$: média da variável X_k na amostra j
  - √ Comparação por um teste t das médias dos valores d_{ij}
  - √ Se uma amostra é mais variável que outra, então a média dos valores d_{ij} tenderá a ser mais alta na amostra mais variável
  - √ Os desvios são calculados com as variáveis padronizadas
    - Assegura o mesmo peso para todas as variáveis

- Versão mais robusta:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ijk} - \tilde{x}_{jk})^2}$$

  - √ Utiliza a mediana amostral ($\tilde{x}_{jk}$)

- Pressuposto dos testes:
  - √ Supõe que se duas amostras diferem, então uma amostra será mais variável do que a outra para todas as variáveis
    - Não se espera um resultado significante quando, por exemplo, X_1 e X_2 são mais variáveis na amostra 1, mas X_3 e X_4 são mais variáveis na amostra 2.
  - √ O teste de Van Halen não é apropriado para situações em que não se espera que mudanças no nível de variação sejam consistentes para todas as variáveis

√ Exemplo: Pardais – Teste de Van Valen:

```
> library(matrixStats)
>
> dados<-split(pardal[,2:6], sobrev)
> dados.1<-as.matrix(dados$'1')
> dados.2<-as.matrix(dados$'2')
>
> mediana.1 <- colMedians(dados.1)
> mediana.2 <- colMedians(dados.2)
>
> desvio.1 <- apply(dados.1, 2, sd)
> desvio.2 <- apply(dados.2, 2, sd)
>
> scaled.1 <- sweep(dados.1, 2, mediana.1, FUN="-")
> scaled.1 <- sweep(scaled.1, 2, desvio.1, FUN="/")
> d.1 <- sqrt(rowSums(scaled.1^2))
>
> scaled.2 <- sweep(dados.2, 2, mediana.2, FUN="-")
> scaled.2 <- sweep(scaled.2, 2, desvio.2, FUN="/")
> d.2 <- sqrt(rowSums(scaled.2^2))
>
> ajuste.d <- t.test(d.1, d.2, var.equal=F)
> ajuste.d

        Welch Two Sample t-test

data:  d.1 and d.2
t = -2.0573, df = 45.282, p-value = 0.0455
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9990010  -0.0124436
sample estimates:
mean of x mean of y
 1.760287  2.265566
```


Comentários

- Diferença significante de variação conjunta das 5 variáveis nas duas amostras a um nível de 5%
  - √ Sobreviventes são menos variáveis do que os não-sobreviventes
  - √ Todas as variáveis mostram menos variação na amostra 1 do que na amostra 2
- Teste de Levene não é direcional
  - √ Teste de Levene não enfatizou fato de diferença de variação consideradas todas as dimensões

Comparação de Valores Médios – Várias Amostras

Comparação de Médias para Várias Amostras – Caso Univariado

- Teste F de Anova de um fator:

Fonte de variação	Soma de Quadrados	Graus de liberdade	Quadrado médio
Entre amostras	$SQ_B = SQ_T - SQ_W$	m – 1	$s_B^2 = \frac{SQ_B}{m-1}$
Dentro de amostras	SQ_W	n – m	$s_W^2 = \frac{SQ_W}{n-m}$
Total	SQ_T	n – 1	

$$F = \frac{s_B^2}{s_W^2}$$

$$SQ_W = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \quad SQ_T = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

- n_j: tamanho da amostra j $n = n_1 + n_2 + \cdots + n_m$
- x_{ij}: i-ésima observação da amostra j
- $\bar{x}_j$: média da amostra j
- $\bar{x}$: média de todas as observações

Comparação de Médias para Várias Amostras – Caso Multivariado

- Hipótese:
  - √ Todas amostras vêm de populações com mesmo vetor médio
- Estatística de teste:
  - √ Há 4 alternativas
  - √ Todos os procedimentos de teste envolvem a suposição de normalidade multivariada, com mesma matriz de covariâncias em todas m populações

- Estatística lambda de Wilks

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|}$$

  - |**W**|: determinante da matriz das somas de quadrados de produtos cruzados dentro da amostra
  - |**T**|**:** determinante da matriz das somas totais de quadrados e produtos cruzados
  - √ Essencialmente, compara a variação dentro das amostras com a variação em ambos, dentro e entre amostras

- Construção das matrizes de produtos cruzados:

$$\mathbf{W} = [w_{rc}]_{p\times p} = \left[\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ijr} - \bar{x}_{jr})(x_{ijc} - \bar{x}_{jc})\right]_{p\times p}.$$

$$\mathbf{T} = [t_{rc}]_{p\times p} = \left[\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ijr} - \bar{x}_{r})(x_{ijc} - \bar{x}_{c})\right]_{p\times p}.$$

  - x_{ijk}: valor da variável X_k para o i-ésimo indivíduo da amostra j
  - (i = 1, 2, ..., n_j; j = 1, 2, ..., m; k = 1, 2, ..., p)
  - $\bar{x}_{jk}$: média da variável X_k na amostra j
  - $\bar{x}_{k}$: média global variável X_k de todas as observações

- Comportamento da estatística:
  - √ Λ é pequeno
    - Indica que a variação dentro das amostras é baixa com a variação total
    - Evidencia que as amostras não vêm de populações com o mesmo vetor de médias

- Estatística lambda de Wilks – Forma alternativa
  - √ Sejam $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ os autovalores da matriz $\mathbf{W}^{-1}\,\mathbf{B}$, em que $\mathbf{B} = \mathbf{T} - \mathbf{W}$ é chamada a matriz entre amostras de somas de quadrados e produtos cruzados.

$$\Lambda = \prod_{i=1}^{p} \frac{1}{1+\lambda_i}$$

  - √ Autovalores são também chamados de raízes latentes

- Distribuição amostral da estatística:

$$F_0 = \frac{1-\Lambda}{\Lambda}\frac{\mathrm{gl}_2}{\mathrm{gl}_1} \overset{H_0}{\sim} F_{\mathrm{gl}_1,\mathrm{gl}_2}.$$
$$\mathrm{gl}_1 = p(m-1)$$
$$\mathrm{gl}_2 = wt - \frac{\mathrm{gl}_1}{2}$$
$$w = (n-1)\left(\frac{p+m}{2}\right)$$
$$t = \frac{\mathrm{gl}_1^2 - 4}{\sqrt{p^2+(m-1)^2-5}}.$$

  - √ Se $\mathrm{gl}_1 = 2$, faça t = 1

- Teste da maior raiz de Roy

$$\lambda_1, \text{ maior autovalor de } \mathbf{W}^{-1}\mathbf{B}$$

  - √ Motivação:
    - λ_1 é a razão máxima entre soma dos quadrados entre amostras e dentro das amostras, obtida por combinação linear entre as variáveis.
    - Este autovalor deve ser uma boa estatística para testar se a variação entre amostras é significativamente grande
      (evidência de que as amostras consideradas não vêm de populações com mesmo vetor de médias)
  - √ Alguns pacotes denominam como estatística da maior raiz de Roy a quantidade $\frac{\lambda_1}{1-\lambda_1}$

- Distribuição amostral da estatística:

$$F_0 = \lambda_1\frac{\mathrm{gl}_2}{\mathrm{gl}_1} \overset{H_0}{\sim} F_{\mathrm{gl}_1,\mathrm{gl}_2}.$$
$$\mathrm{gl}_1 = d$$
$$\mathrm{gl}_2 = n - m - d - 1$$
$$d = \max\{p, m-1\}.$$

- Estatística traço de Pillai

$$V = \sum_{i=1}^{p} \frac{\lambda_i}{1+\lambda_i}$$

√ Valores grandes para esta estatística fornecem evidência de que as amostras vêm de populações com vetores de médias diferentes.

- Distribuição amostral da estatística:

$$F_0 = V\frac{n-m-p+s}{d(s-V)} \overset{H_0}{\sim} F_{\mathrm{gl}_1,\mathrm{gl}_2}.$$
$$\mathrm{gl}_1 = sd$$
$$\mathrm{gl}_2 = s(n-m-p+s)$$
$$s = \min\{p, m-1\}$$
$$d = \max\{p, m-1\}.$$

- Estatística traço de Hotelling-Lawley:

$$U = \sum_{i=1}^{p} \lambda_i$$

√ Valores grandes fornecem evidência contra H_0.

- Distribuição amostral da estatística:

$$F_0 = U\frac{\mathrm{gl}_2}{\mathrm{gl}_1} \overset{H_0}{\sim} F_{\mathrm{gl}_1,\mathrm{gl}_2}.$$
$$\mathrm{gl}_1 = s(2A+s+1)$$
$$\mathrm{gl}_2 = 2(sB+1)$$
$$s = \min\{p, m-1\}$$
$$A = \frac{|m-p-1|-1}{2}$$
$$A = \frac{n-m-p-1}{2}.$$

Comentários

- Espera-se que os 4 testes apresentem níveis de significância similares
- Todos os testes supõem normalidade multivariada
- São considerado bastante robustos se os tamanhos das m amostras são aproximadamente iguais.
- A estatística traço de Pillai pode ser a mais robusta se há questões sobre a normalidade multivariada ou a igualdade das matrizes de covariâncias

Exemplo

- Thomson e Randall-Maciver (1905)
- Medidas em crânios masculinos da área de Tebas
  - √ 5 amostras de 30 crânios cada uma:
    - Período pré-dinástico primitivo (~ 4.000 aC)
    - Período pré-dinástico antigo (~3.300 aC)
    - 12ª e 13ª dinastias (~ 1.850 aC)
    - Período Ptolemaico (~ 200 aC)
    - Período romano (~150 dC)
  - √ Dados: *skulls*{*ade4*} ou *skulls2.csv* (*skulls.txt*)

- Variáveis:
  - √ MB (X_1): Largura máxima do crânio
  - √ BH (X_2): Altura do basibregmático do crânio
  - √ BL (X_3): Comprimento do basialveolar do crânio
  - √ NH (X_4): Altura nasal do crânio
  - √ Ano: Ano aproximado de formação do crânio

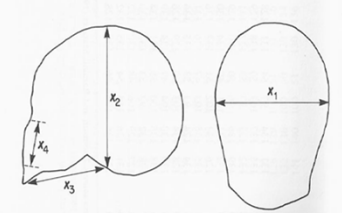


Figure 1.1 Measurements on Egyptian skulls.

- Comandos em R:
  - √ Carregamento do conjunto de dados:

```
> cranios <- read.csv("skulls2.csv",header=T,skip=4)
> colunas <- c("X1", "X2", "X3", "X4", "Período", "Per")
> colnames(cranios) <- colunas
> attach(cranios)
```


- Anova – Variável individual:
  - √ Variável X_1: Amplitude máxima do crânio

```
> ajuste.X1 <- aov(X1 ~ Per, data=cranios)
> summary(ajuste.X1)
             Df Sum Sq Mean Sq F value    Pr(>F)
Per           4  502.8  125.71   5.955 0.000183 ***
Residuals   145 3061.1   21.11
```

  - √ Variável X_2: Altura basilobregmática do crânio

```
> ajuste.X2 <- aov(X2 ~ Per, data=cranios)
> summary(ajuste.X2)
             Df Sum Sq Mean Sq F value Pr(>F)
Per           4    230   57.48   2.447  0.049 *
Residuals   145   3405   23.48
```

  - √ Variável X_3: Comprimento basiloalveolar do crânio

```
> ajuste.X3 <- aov(X3 ~ Per, data=cranios)
> summary(ajuste.X3)
             Df Sum Sq Mean Sq F value    Pr(>F)
Per           4    803  200.82   8.306 4.64e-06 ***
Residuals   145   3506   24.18
```


  - √ Variável X_4: Altura nasal do Crânio

```
> ajuste.X4 <- aov(X4 ~ Per, data=cranios)
> summary(ajuste.X4)
             Df Sum Sq Mean Sq F value Pr(>F)
Per           4   61.2   15.30   1.507  0.203
Residuals   145 1472.1   10.15
```

- Conclusão:
  - √ Existe clara evidência de que a mudaram com o tempo as média populacionais da largura máxima (X_1), da altura basibregmático (X_2), do comprimento do basibregmático (X_3).
  - √ Há evidência de que a média da altura nasal (X_4) não mudou com o tempo

- Comparação conjunta das médias das 4 variáveis:
  - √ Comandos em R – objeto Manova

```
> variaveis <- cbind(X1, X2, X3, X4)
> ajuste<- manova(variaveis ~ Per)
>
```

  - √ Estatística lambda de Wilks:

```
> ajuste.wilks <- summary(ajuste, test = "Wilks")
> ajuste.wilks
          Df   Wilks approx F num Df den Df    Pr(>F)
Per        4 0.66359   3.9009     16 434.45 7.01e-07 ***
Residuals 145
```

  - √ Estatística da maior raiz de Roy:

```
> ajuste.roy <- summary(ajuste, test = "Roy")
> ajuste.roy
          Df    Roy approx F num Df den Df    Pr(>F)
Per        4 0.4251    15.41      4    145 1.588e-10 ***
Residuals 145
```


- Comparação conjunta das médias das 4 variáveis:
  - √ Estatística traço de Pillai:

```
> ajuste.pillai <- summary(ajuste, test = "Pillai")
> ajuste.pillai
          Df  Pillai approx F num Df den Df    Pr(>F)
Per        4 0.35331    3.512     16    580 4.675e-06 ***
Residuals 145
```

  - √ Estatística traço de Hotteling-Lawley:

```
> ajuste.hotelling <- summary(ajuste, test = "Hotelling-Lawley")
> ajuste.hotelling
          Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
Per        4          0.48182    4.231     16    562 8.278e-08 ***
Residuals 145
```

- Conclusão:
  - √ Há uma evidência muito forte de que os valores médios mudam com o tempo para as 4 variáveis.

Comparação de Variação para Várias Amostras – Caso Multivariado

- Comparação de variação no caso multivariado
- Teste M de Box:
  - √ Estatística de teste: $M_0 = \frac{\prod_{i=1}^{m} |\mathbf{S}_i|^{\frac{n_i-1}{2}}}{|\mathbf{S}|^{\frac{n-m}{2}}}.$
    - $\mathbf{S}_i$: matriz de covariâncias para a i-ésima amostra
    - $\mathbf{S}$: matriz de covariâncias combinada $\mathbf{S} = \frac{\sum_{i=1}^{m}(n_i-1)\mathbf{S}_i}{n-m}.$
  - √ Grandes valores de M fornecem evidência de que as amostras não provêm de populações com mesma matriz de covariâncias.

- Teste F aproximado para M:
  - √ Estatística de teste: $F_0 = -2b\ln(M).$
  - √ Distribuição amostral da estatística: $F_0 \overset{H_0}{\sim} F_{\nu_1,\nu_2}.$

$$\nu_1 = \frac{p(p+1)(m-1)}{2}$$

$$\nu_2 = \frac{\nu_1+2}{c_2-c_1^2}.$$

  - √ Outras quantidades:

$$b = \frac{1-c_1-\frac{\nu_1}{\nu_2}}{\nu_1}$$

$$c_1 = \frac{(2p^2+3p-1)\left(\sum_{i=1}^{m}\frac{1}{n_i-1}-\frac{1}{n-m}\right)}{6(p+1)(m-1)}$$

Teste válido quando $c_2 > c_1^2$

$$c_2 = \frac{(p-1)(p+2)\left(\sum_{i=1}^{m}\frac{1}{(n_i-1)^2}-\frac{1}{(n-m)^2}\right)}{6(m-1)}$$

- Teste aproximado alternativo

  (Quando $c_2 < c_1^2$)

  √ Estatística de teste: $F_0^* = \frac{-2b_1\nu_2 \ln(M)}{\nu_1 + 2b_1 \ln(M)}.$

  $b_1 = \frac{1 - c_1 - \frac{2}{\nu_2}}{\nu_1}.$

  √ Distribuição amostral: $F_0^* \overset{H_0}{\sim} F_{\nu_1,\nu_2}.$

- O teste de Box é sensível a desvios da normalidade multivariada

- Generalização de sugestão robusta (2 amostras):
  - √ Desvios absolutos de medianas amostrais para os dados em m amostras
  - √ Procedimento de teste
    - Univariado
      - Anova
    - Mais de uma variável
      - Aplicação dos 4 testes sugeridos nos dados transformados
  - √ Resultado significante indica que a matriz de covariâncias não é constante para as m populações
  - √ Variáveis podem ser padronizadas

Exemplo – Crânios Egípcios

- Comparação da variação conjunta das 4 variáveis:
  - √ Teste M de Box: $M_0 = \frac{\prod_{i=1}^{m} |\mathbf{S}_i|^{\frac{n_i-1}{2}}}{|\mathbf{S}|^{\frac{n-m}{2}}} = 2,869 \times 10^{-11}.$

    $b = 0,0235$

    $\nu_1 = 40$

    $\nu_2 = 46,479.$

    $F_0 = -2b\ln(M) = 1,14.$

    $\text{p-valor} = \text{P}\{F_{40;46,379} > 1,14\} = 0,250.$

  - √ Estatística de teste não é significante
    - Teste não mostra evidência de que a matriz de covariâncias mudou com o tempo

- Comparação da variação conjunta das 4 variáveis:
  - √ Comandos em R
    - Apresenta apenas função com aproximação χ^2, que deve ser usada apenas quando ni > 20, p < 6 e m < 6.
    - Caso estas condições não sejam atendidas, deve-se usar a aproximação F

```
> library(biotools)
> variaveis <- cbind(X1, X2, X3, X4)
> boxM(variaveis, Per)

        Box's M-test for Homogeneity of Covariance Matrices

data:  variaveis
Chi-Sq (approx.) = 45.6672, df = 40, p-value = 0.2483
```

  - √ Estatística de teste não é significante
    - Teste não mostra evidência de que a matriz de covariâncias mudou com o tempo.

- Conclusão:
  - √ Há uma evidência muito forte de que os valores médios mudam com o tempo para as 4 variáveis, embora não haja evidência de que a variação tenha mudado.

Referências

Bibliografia Recomendada

- MANLY, B. J. F. *Métodos Estatísticos Multivariados: uma Introdução.* Bookman, 2008.
- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- EVERITT, B.; HOTHORN, T. *An Introduction to Applied Multivariate Analysis with R*. Springer, 2011.