Comparação das médias de duas populações

Vicente G. Cancho
garibay@icmc.usp.br

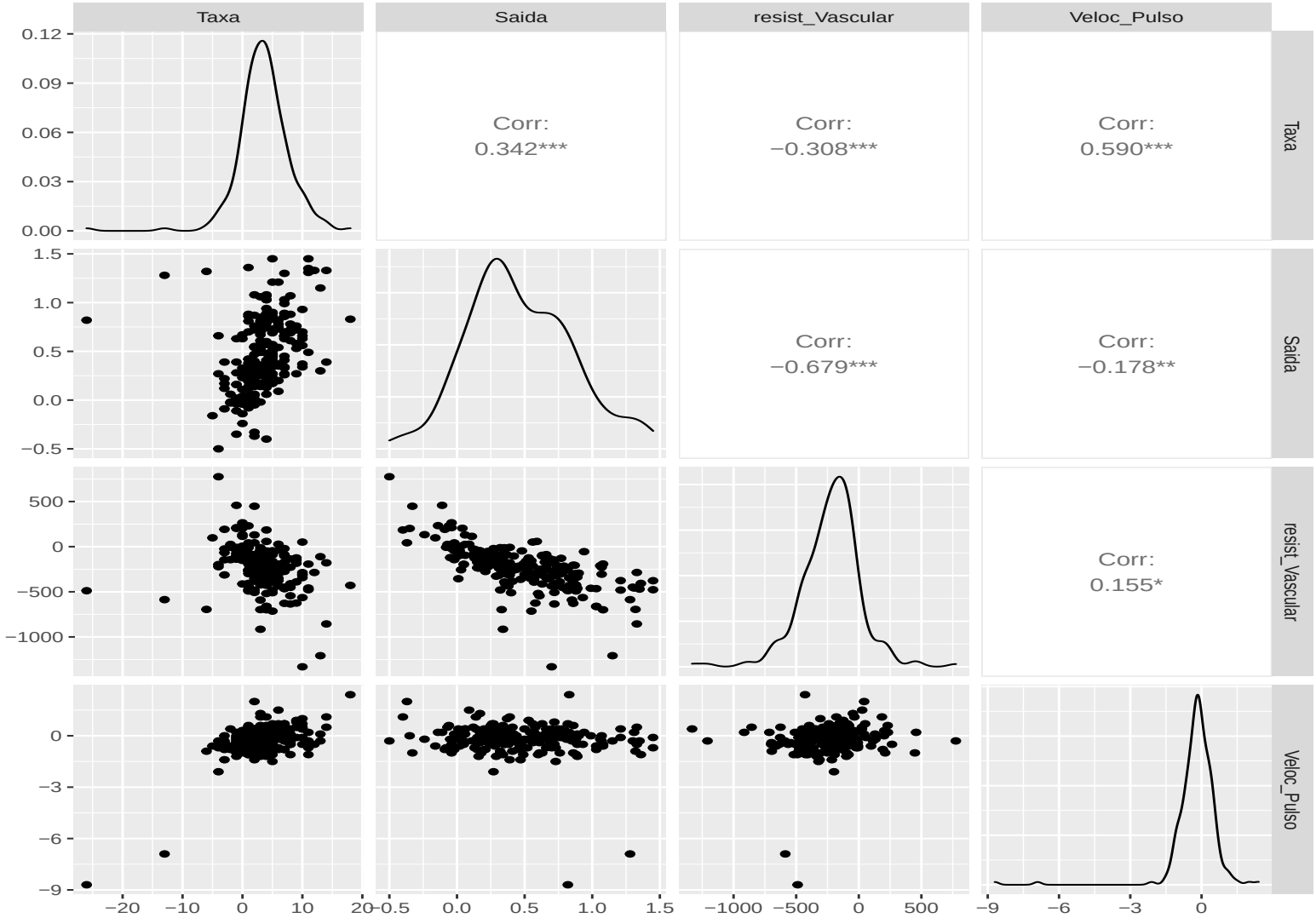
Exemplo: Risco Cardiovascular

- Num estudo do risco cardiovascular foram considerados variáveis referentes a medições cardíacas de 223 indivíduos.
- Para cada um dos indivíduos, as variáveis foram coletadas em duas posições: (1) Supina e (2) Inclinada
- Seja $\boldsymbol{X}_i$ o vetor aleatório de 4 medições na posição i, $i = 1, 2$, com componentes representando as seguintes medições:
  - X_{i1} taxa cardíaca
  - X_{i2} saída cardíaca
  - X_{i3} resistência vascular sistêmica
  - X_{i4} velocidade do pulso de onda
- Note que $\boldsymbol{X}_1^\top = (X_{11}, X_{12}, X_{12}, X_{14})$ e $\boldsymbol{X}_2^\top = (X_{21}, X_{22}, X_{22}, X_{24})$ são correlacionados, pois são mensurações no mesmo indivíduo (dados pareados).

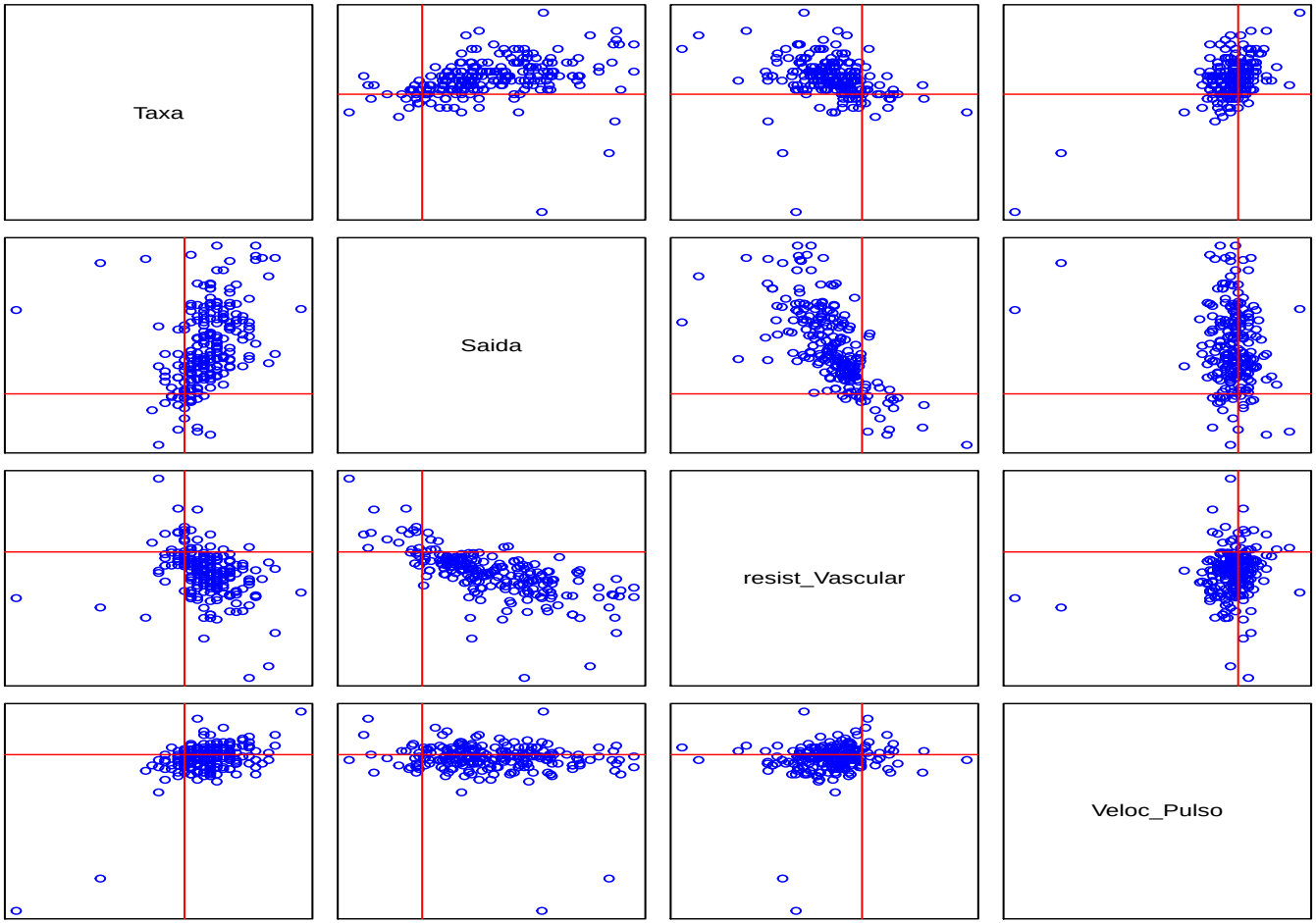
Exemplo: Cardiovascular Risk

1. Um grande interesse nesse tipo de estudo é avaliar se as medições cardíacas são iguais nas duas posições.
2. Nesse caso, é necessário avaliar se as médias dos vetores aleatórios $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ são iguais.
3. Isso é equivalente a testar a a diferença de medições cardíacas nas duas diferentes posições.
4. Para representar essa diferença, seja $\boldsymbol{D} = \boldsymbol{X}_1 - \boldsymbol{X}_2$.
5. Uma amostra desse vetor de diferenças pode ser obtida em função das amostras das medições dos indivíduos nas duas diferentes posições.

Exemplo: Cardiovascular Risk



Exemplo: Cardiovascular Risk



Exemplo: Cardiovascular Risk

- As densidades marginais apresentam uma distribuição aproximadamente simétrica;
- A dispersão do pontos nos gráficos apresentam approximadamente, um formato elíptico (normal bivariado)
- correlações entre as diferenças parecem variar entre fracas e moderadas
- na maioria dos casos, a origem do gráfico (0, 0) parece estar próxima do centro de massa dos dados.
- dados mostram que é razoável testar a hipótese da igualdade do vetor de médias das medições cardíacas na posições supino e inclinado

Comparações pareadas

Sejam

- $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n}$ vetores aleatórios $p \times 1$ referentes a uma população normal multivariada antes de um tratamento com $\mathrm{E}(\boldsymbol{X}_{1j}) = \boldsymbol{\mu}_1$ para $j = 1, \ldots, n$,
- $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n}$ vetores aleatórios $p \times 1$ referentes a uma população normal multivariada após de um tratamento com $\mathrm{E}(\boldsymbol{X}_{2j}) = \boldsymbol{\mu}_2$ para $j = 1, \ldots, n$,

sendo que $\boldsymbol{X}_{1j}$ e $\boldsymbol{X}_{2j}$ são correlacionadas (por exemplo, vetores aleatórios de medições antes e após um tratamento).

Comparações pareadas

- Sejam $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ os vetores de médias em situações 1 e 2, respectivamente.
- Deseja-se testar se não há diferença entre as situações 1 e 2 para verificar, por exemplo, que o tratamento não produz nenhum efeito, ou seja, se $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.
- Hipóteses de interesse:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$
$$H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

- Considera-se as diferenças:

$$\boldsymbol{D}_j = \boldsymbol{X}_{1j} - \boldsymbol{X}_{2j},\ j = 1, \ldots, n$$

- Supor, $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n$ é uma amostra aleatória de uma população $N_p(\boldsymbol{\mu}_D, \Sigma_D)$, com $\boldsymbol{\mu}_D = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

- A hipóteses anterior é equivalente a,

$$\begin{aligned} H_0 &: \boldsymbol{\mu}_D = \mathbf{0} \\ H_1 &: \boldsymbol{\mu}_D \neq \mathbf{0} \end{aligned} .$$

- Pode-se mostrar que o teste da razão de verosimilhança tem distribuição T^2 de Hotelling, ou seja,

$$T^2 = n(\bar{\boldsymbol{D}} - \mathbf{0})^\top S_D^{-1}(\bar{\boldsymbol{D}} - \mathbf{0}) \overset{sob\ H_0}{\sim} \frac{(n-1)p}{n-p} F_{p,n-p},$$

em que $\bar{\boldsymbol{D}}$ e S_D são o vetor de médias e a matriz de variâncias e covariâncias amostrais de $\boldsymbol{D}$.

Um teste análogo poderia ser desenvolvido para avaliar

$$\begin{array}{l} H_0 : \boldsymbol{\mu}_D = \boldsymbol{\delta}_0 \\ H_1 : \boldsymbol{\mu}_D \neq \boldsymbol{\delta}_0 \end{array} ,$$

onde δ_0 é conhecido.
A estatística de teste é

$$T^2 = n(\bar{\boldsymbol{D}} - \boldsymbol{\delta}_0)^\top S_D^{-1}(\bar{\boldsymbol{D}} - \boldsymbol{\delta}_0) \overset{sob\ H_0}{\sim} \frac{(n-1)p}{n-p} F_{p,n-p}.$$

- Região de confiança

$$\{\boldsymbol{\mu}_D; (\bar{\boldsymbol{d}} - \boldsymbol{\mu}_D)^\top S_D^{-1} (\bar{\boldsymbol{d}} - \boldsymbol{\mu}_D) \leq \frac{(n-1)p}{(n-p)n} q_{F_{p,n-p,\alpha}}\}$$

onde $q_{F_{p,n-p,\alpha}}$ é o quantil α superior da distribuição F com p e n-p graus de liberdade.

Quando $H_0 : \mu_D = 0$ é rejeitado

- Intervalos de confiança simultâneos T^2 de Hotelling para cada componente da diferença

$$\bar{D}_i \mp \sqrt{\frac{(n-1)p}{(n-p)n} q_{F_{p,n-p,\alpha}}} \sqrt{\frac{s_{d_i}}{n}}, \ i = 1, \ldots, p$$

onde $\bar{D}_i$ é a diferença média da i-ésima variável e s_{d_i} é o i-ésimo elemento diagonal de S_D.

Comparações pareadas

- Intevalos de confiança de Bonferroni

$$\bar{D}_i \mp F^{-1} T_{(n-1)}(1 - \frac{1}{2m})\sqrt{\frac{s_{d_i}}{n}},\ i = 1, \ldots, p$$

onde m é numero de intervalos (comparações).

- Para (n - p) grande (ou seja, $\boldsymbol{D}_j$ não precisa ser normal multivariado)

$$n(\bar{\boldsymbol{D}} - \boldsymbol{\delta}_0)^\top S_D^{-1}(\bar{\boldsymbol{D}} - \boldsymbol{\delta}_0) \overset{sob\ H_0}{\approx} \chi^2_{(p)}$$

Comparações pareadas: Exemplo

Das 223 observações do dados de *Cardiovascular Risk*, o vetor medias e covariâncias amostrais das diferenças das medições são resultaram respectivamente

$$\bar{\boldsymbol{d}} = \begin{pmatrix} 3.529 \\ 0.467 \\ -224.354 \\ -0.212 \end{pmatrix}, \text{ e } S_D = \begin{pmatrix} 18.800 & 0.569 & -325.713 & 2.401 \\ 0.569 & 0.147 & -63.375 & -0.064 \\ -325.713 & -63.375 & 59301.266 & 35.396 \\ 2.401 & -0.064 & 35.396 & 0.881 \end{pmatrix}$$

As hipóteses de interesse: $H_0 : \boldsymbol{\mu}_D = \mathbf{0}$ vs $H_1 : \boldsymbol{\mu}_D \neq \mathbf{0}$ (as medições cardíacas em médias são iguais nas posições supino e inclinado).
A estatística observada

$$T^2_{obs} = n(\bar{\boldsymbol{d}})^\top S_D^{-1}(\bar{\boldsymbol{d}}) = 412.2676$$

Comparações pareadas: Exemplo

O valor critico: $\frac{(n-1)p}{n-p} q_{F_{p,n-p},0.05} = \frac{(223-1)4}{223-4} \times 2.41287 = 9,783 < T^2_{obs}$
Rejeita-se H_0
Similarmente $T^2 \frac{n-p}{(n-1)p} = 101.6741$ e

$$p - valor = P(F_{p,n-p} > 101,6741) = 0.$$

Comparações pareadas: Exemplo

O valor critico: $\frac{(n-1)p}{n-p} q_{F_{p,n-p,0.05}} = \frac{(223-1)4}{223-4} \times 2.41287 = 9,783 < T^2_{obs}$
Rejeita-se H_0
Similarmente $T^2 \frac{n-p}{(n-1)p} = 101.6741$ e

$$p - valor = P(F_{p,n-p} > 101,6741) = 0.$$

```
library(ICSNP)
data(LASERI)
cardio <- LASERI[,c("HRT1T4","COT1T4","SVRIT1T4","PWVT1T4")]
#renomeando variáveis
names(cardio) <-c("Taxa","Saida","resist_Vascular","Veloc_Pulso")
> HotellingsT2(cardio)
Hotelling's one sample T2-test
data:  cardio
T.2 = 101.67, df1 = 4, df2 = 219, p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(0,0,0,0)
```

Comparações pareadas: Exemplo

- Intervalo simultâneo de 95% de confiança

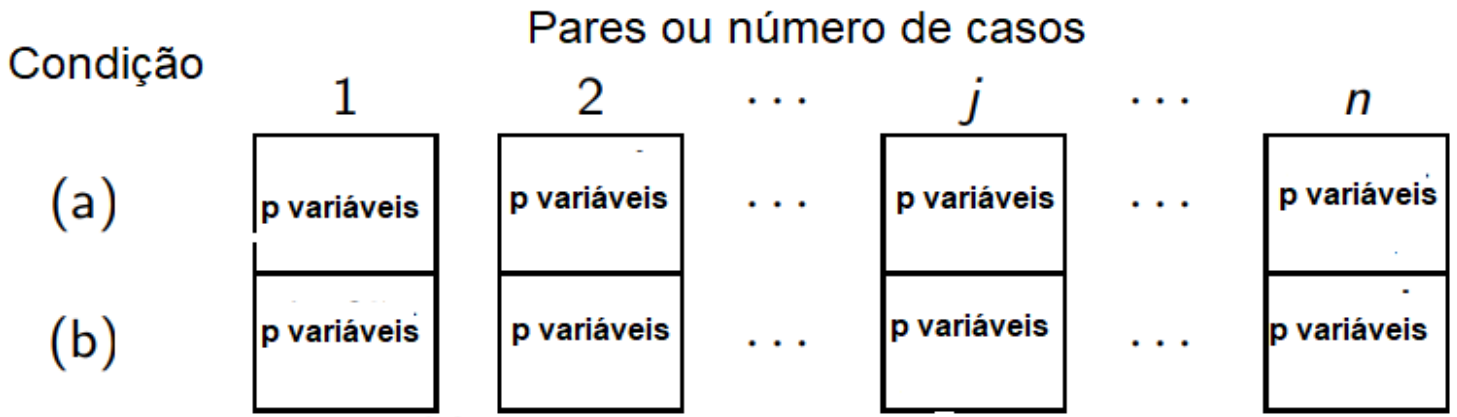
	LI	LS
Taxa Cardíaca	2.621	4.437
Saida Cardíaca	0.387	0.547
Resistência Vascular sistêmica	-275.361	-173.347
Velocidade de Pulso de onde	-0.409	-0.016

- Intervalo de 95% de confiança de Bonferroni

	LI	LS
Taxa Cardíaca	2.798	4.260
Saida Cardíaca	0.402	0.531
Resistência Vascular sistêmica	-265.419	-183.289
Velocidade de Pulso de onde	-0.370	-0.054

Abordagem: Dados completos

- Para as comparações emparelhas determinar as diferenças das medições repetidas, ou seja $\boldsymbol{D} = \boldsymbol{X}_1 - \boldsymbol{X}_2$.
- Agora vamos considerar o método de "Amostra Completa" que considera cada caso como um par e cada um com p medidas em cada membro do par.



- Portanto, temos 2p variáveis mensurados para cada caso (par).
- Em uma situação experimental, presume-se que as condições foram atribuídas aleatoriamente aos membros dos pares.

Matriz de dados completa:

$$\boldsymbol{X} = \left(\begin{array}{cccc|cccc} X_{111} & X_{112} & \cdots & X_{11p} & X_{121} & X_{122} & \cdots & X_{12p} \\ X_{211} & X_{212} & \cdots & X_{21p} & X_{221} & X_{222} & \cdots & X_{22p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n11} & X_{n12} & \cdots & X_{n1p} & X_{n21} & X_{n22} & \cdots & X_{n2p} \end{array}\right)_{n\times 2p}$$

$$= \left(\begin{array}{c|c} \underbrace{\boldsymbol{X}_1}_{n\times p} & \underbrace{\boldsymbol{X}_2}_{n\times p} \end{array}\right).$$

Vetor de médias de dados completos

$$\bar{\boldsymbol{X}}^\top = \left(\begin{array}{cccc|cccc} \bar{\boldsymbol{X}}_{11} & \bar{\boldsymbol{X}}_{12} & \cdots & \bar{\boldsymbol{X}}_{1p} & \bar{\boldsymbol{X}}_{21} & \bar{\boldsymbol{X}}_{22} & \cdots & \bar{\boldsymbol{X}}_{2p} \end{array}\right) = \left(\begin{array}{c|c} \bar{\boldsymbol{X}}_1^\top & \bar{\boldsymbol{X}}_2^\top \end{array}\right).$$

Abordagem: Dados completos

Matriz de covariância amostral de dados completa:

$$\boldsymbol{S}_{2p\times 2p} = \left(\begin{array}{c|c} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \hline \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{array}\right)$$

onde

- $\boldsymbol{S}_{11}$ é a matriz de covariâncias amostral de $\boldsymbol{X}_1$, de ordem $p \times p$
- $\boldsymbol{S}_{22}$ é a matriz covariância amostral de $\boldsymbol{X}_2$ de ordem $p \times p$.
- $\boldsymbol{S}_{12} = \boldsymbol{S}_{21}^{\top}$ é a matriz covariância amostral de $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ de ordem $p \times p$.

Defina a matriz de constantes

$$\boldsymbol{C}_{p\times 2p} = \left(\begin{array}{cccc|cccc} 1 & 0 & \dots & 0 & -1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & o & \dots & 1 & 0 & 0 & \dots & -1 \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{I}_p & -\boldsymbol{I}_p \end{array}\right)$$

Abordagem: Dados completos

Sejam

- $\underbrace{\boldsymbol{x}_j}_{2p\times 1}$ é j-ésima linha da matrix $\boldsymbol{X}_{n\times 2p}$ é escrita como um vetor coluna.
- $\boldsymbol{d}_j = \boldsymbol{C}\boldsymbol{x}_j$ (diferença das 2 medições repetidas)
- $\bar{\boldsymbol{d}} = \boldsymbol{C}\bar{\boldsymbol{x}} = \boldsymbol{C}(n^{-1}\sum_{j=1}^{n}\boldsymbol{x}_j)$.
- $E(\bar{\boldsymbol{d}}) = E(\boldsymbol{X}_1) - E(\boldsymbol{X}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

Daí tem-se sob $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{0}$

$$\begin{aligned} T^2 &= n(\boldsymbol{C}\bar{\boldsymbol{x}})^\top(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\bar{\boldsymbol{x}}) \\ &= n\bar{\boldsymbol{x}}^\top\boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1}\boldsymbol{C}\bar{\boldsymbol{x}} \sim \frac{(n-1)p}{n-p}F_{p,n-p} \end{aligned}$$

Com este método, não temos que dividir o conjunto de dados e calcular as diferenças

Abordagem: Dados completos

- Esta é outra generalização do teste t para dados emparelhado univariado.
- Situação: q condições são comparadas com relação a um resposta variável.

  Cada caso recebe cada tratamento uma vez em períodos sucessivos de tempo.

  A ordem dos tratamentos deve ser aleatorizada (e balançado, se possível).

Abordagem: Dados completos:Exemplo

Um experimento planejado foi realizado,a fim de verificar se há diferenças entre a escrita informal e formal, onde 15 alunos escreveram uma redação formal e informal. As variáveis registradas foram número de palavras e número de verbos.

- x_{11}: palavras em ensaio informal
- x_{12}: verbos em ensaio informal
- x_{21}: palavras em ensaio formal
- x_{22}: verbos em ensaio formal

Abordagem: Dados completos:Exemplo

Um experimento planejado foi realizado,a fim de verificar se há diferenças entre a escrita informal e formal, onde 15 alunos escreveram uma redação formal e informal. As variáveis registradas foram número de palavras e número de verbos.

- x_{11}: palavras em ensaio informal
- x_{12}: verbos em ensaio informal
- x_{21}: palavras em ensaio formal
- x_{22}: verbos em ensaio formal

	Informal($\boldsymbol{x}_1$)		Formal($\boldsymbol{x}_2$)	
Estudante	x_{11}	x_{12}	x_{21}	x_{22}
1	148	20	137	15
2	159	24	164	25
3	144	19	224	27
4	103	18	208	33
5	121	17	178	24
6	89	11	128	20
7	119	17	154	18
8	123	13	158	16
9	76	16	102	21
10	217	29	214	25
11	148	22	209	24
12	151	21	151	16
13	83	7	123	13
14	135	20	161	22
15	178	15	175	23

Abordagem: Dados completos:Exemplo

- Seja $E(\boldsymbol{X}_j) = \boldsymbol{\mu}_j$ o vetor de médias para j-ésima condição (informal ou formal).
- Hipótese de interesse: $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$,
- Dos dados tem-se o vetor de médias amostrais

$$\bar{\boldsymbol{x}}^\top = (132.933 \quad 17.933 \quad | \quad 165.733 \quad 21.467)$$

e a matriz de covariância amostral resultou

$$\boldsymbol{S}_{4\times 4} = \left(\begin{array}{cc|cc} 1405.78 & 153.71 & 804.77 & 43.10 \\ 153.71 & 28.64 & 108.05 & 12.53 \\ \hline 804.77 & 108.05 & 1299.78 & 137.35 \\ 43.10 & 12.53 & 137.35 & 27.98 \end{array}\right)$$

A matriz de contraste é

$$\boldsymbol{C} = \left(\begin{array}{c|c} \boldsymbol{I}_2 & -\boldsymbol{I}_2 \end{array} \right)$$

Abordagem: Dados completos:Exemplo

A estatística T^2 observada resulta em

$$T^2_{obs} = n\bar{\boldsymbol{x}}^\top \boldsymbol{C}^\top (\boldsymbol{CSC}^\top)^{-1} \boldsymbol{C}\bar{\boldsymbol{x}} = 15,19123.$$

o valor crítico

$$\frac{(15-1)2}{15-2} qf(0.95, p, n-p) = 8.196 > T^2_{obs},$$

portanto, rejeita-se H_0.
Alternativamente $\frac{15-2}{(15-1)2} T^2 = 7,053$ que tem uma distribuição F com p e $n-p$ graus de liberdade. Dai o p-valor resultou 0,0084,
Conclusão: Os dados apoiam a conclusão de que o o número médio de palavras e verbos em ensaios informais não é igual ao número de ensaios formais..

Code in R

```
T2=function(x,C,alpha){
p=2
n=nrow(x)
xbar=colMeans(x)
Sc=cov(x)
dbar=C%*%xbar
T=n*t(dbar)%*%solve(C%*%Sc%*%t(C))%*%dbar
vc=(n-1)*p/(n-p)*qf(0.05,p,n-p,lower.tail = F)
f=(n-p)/((n-1)*p)*T
pvalor=1-pf(f,p, n-p)
saida=list(T2=T,T_critico=vc, pvalor=pvalor)
saida
}
x=dados[,-1]
C=cbind(diag(2),-diag(2))
T2(x,C,0.05)
$T2
[1,] 15.19123
$T_critico
[1] 8.196602
$pvalor
[1,] 0.008427354
```

Medidas repetidas

- É uma generalização do teste t pareado univariado.
- Situação: q condições são comparadas com relação a um variável de resposta
- Cada caso recebe cada tratamento uma vez em períodos sucessivos de tempo.
- A ordem dos tratamentos deve ser randomizada (balanceada, se possível).

Exemplo

:
Existe quatro modelos de calculadora a cada pessoa faz cálculos específicos em cada uma delas e a velocidade é registrada. A ordem de uso da calculadora foi atribuída aleatoriamente, os dados são mostrados a continuação:

Pessoa	Modelos			
	1	2	3	4
1	30	21	21	14
2	22	13	22	5
3	29	13	18	17
4	12	7	16	14
5	23	24	23	8

O interesse é verificar se os quatro modelos tem a mesma velocidade de cálculo.

Medidas repetidas

- Seja a j-ésima observação igual a

$$\boldsymbol{x}_j = \begin{pmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jq} \end{pmatrix}, \quad j = 1, \ldots, n$$

onde x_{ji} é a resposta ou mensuração do i-ésimo tratamento no j-ésimo caso.

- Pergunta (hipótese): Existe um efeito do tratamento?

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_q$, H_1 : Ao menos um para é diferente

Mesmo teste de hipótese em medidas univariadas e repetidas ANOVA.

Medidas repetidas

- Para testar isso como um vetor médio multivariado, precisamos usar contrastes dos componentes de $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = E(\boldsymbol{X}_j) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}$$

- Supor $\boldsymbol{X}_j \sim N(\boldsymbol{\mu}, \Sigma)$
- A matriz de contraste

$$\underbrace{\begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{pmatrix}}_{(q-1)\times 1} = \underbrace{\begin{pmatrix} 1 & -1 & 0\ldots & 0 \\ 1 & 0 & -1\ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0\ldots & -1 \end{pmatrix}}_{(q-1)\times q} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}}_{q\times 1} = \boldsymbol{C}\boldsymbol{\mu}$$

- Assim $H_0 : \boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{0}$ (efeito de tratamento nulo)

Matriz de constrastes

- Qualquer matriz de contraste de ordem $(q-1) \times q$ pode ser considerado
- Por exemplo

$$\boldsymbol{C}_1\boldsymbol{\mu} = \underbrace{\begin{pmatrix} 1 & -1 & 0\ldots & 0 & 0 \\ 0 & 1 & -1\ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0\ldots & 1 & -1 \end{pmatrix}}_{(q-1)\times q} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}}_{q\times 1}$$

- Para ser uma matriz de contraste
  - As linhas são linearmente independentes.
  - Cada linha é um vetor de contraste.

Teste de Hipótese em Medidas Repetidas

A hipótese de nenhum efeito devido ao tratamento em um delineamento de medidas repetidas.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_q,$$

é o mesmo que realizar o teste T^2 de Hotelling para

$$H_0 : \boldsymbol{C}\boldsymbol{\mu} = \mathbf{0},$$

onde $\boldsymbol{C}$ é uma matriz de contrastes de ordem $(q-1) \times q$.

Teste de Hipótese em Medidas Repetidas

Dado $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ e uma matriz de contraste $\boldsymbol{C}$, o estatística de teste T^2 é igual

$$T^2 = n\bar{\boldsymbol{x}}^\top \boldsymbol{C}^\top (\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1} \boldsymbol{C}\bar{\boldsymbol{x}}$$

Rejeita H_0 se

$$T^2_{obs} > \frac{(n-1)(q-1)}{n-q+1} F_{q-1,n-q+1}(\alpha) = T^2_{q-1,n-q+1}(0,05)$$

onde $F_{q-1,n-q+1}(\alpha)$ é o quantil α superior da distribuição F com q-1 e n-q+1 graus de liberdade.

Teste de Hipótese em Medidas Repetidas: Exemplo

No Exemplo anterior o interesse é testar $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, a qual é equivalente a testar $H_0 : \boldsymbol{C}\boldsymbol{\mu} = \mathbf{0}$, onde e matriz de contrastes é

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}$$

Dos dados tem-se

$$\bar{\boldsymbol{x}} = \begin{pmatrix} 23.2 \\ 15.6 \\ 20.0 \\ 11.6 \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} 51.70 & 29.85 & 9.25 & 7.35 \\ 29.85 & 46.80 & 16.25 & -8.70 \\ 9.25 & 16.25 & 8.50 & -10.50 \\ 7.35 & -8.70 & -10.50 & 24.30 \end{pmatrix}$$

$$T^2_{obs} = n\bar{\boldsymbol{x}}^\top \boldsymbol{C}^\top (\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1} \boldsymbol{C}\bar{\boldsymbol{x}} = 29.73605 < T^2_{q-1,n-q+1}(\alpha) = 114.9858$$

A hipoteses nula não rejeitado ao nível de significância de 5%, indicando que os quatro modelos de calculadores tem em média a mesma velocidade de processamento.

T^2 e Medidas Repetidas

- Uma rigião de $100(1-\alpha)\%$ de confiança para $\boldsymbol{C}\boldsymbol{\mu}$

$$n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top \boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1}\boldsymbol{C}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) \leq \frac{(n-1)(q-1)}{n-q+1}F_{q-1,n-q+1}(\alpha)$$

- Intervalos de confiança simultânea de T^2 Hotelling para um único contraste $\boldsymbol{c}_i^\top\boldsymbol{\mu}$ onde $\boldsymbol{c}_i'$ é o i-ésima linha da matriz de contraste, $\boldsymbol{C}$

$$\boldsymbol{c}_i^\top\bar{\boldsymbol{x}} \mp \underbrace{\sqrt{\frac{(n-1)(q-1)}{n-q+1}F_{q-1,n-q+1}(\alpha)}}\sqrt{\frac{\boldsymbol{c}_i^\top\boldsymbol{S}\boldsymbol{c}_i}{n}}$$

- Para intervalos de confiança de Bonferroni (ou um de cada vez), substitua a estatística acima da chave pelo valor apropriado da distribuição t_{n-1}.
- Para n grande use $\chi^2_{(q-1)}$.

Comparação de vetor médias de duas populações independentes

Situação: Duas amostras, cada uma tendo p medições onde nós ter uma amostra aleatória de tamanho n_1 da população 1 e uma amostra aleatória de tamanho n_2 da população 2.

- $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ vetores aleatórios $p \times 1$ referentes a uma população com $\mathrm{E}(\boldsymbol{X}_{1j}) = \boldsymbol{\mu}_1$ para $j = 1, \ldots, n_1$,
- $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ vetores aleatórios $p \times 1$ referentes a uma população com $\mathrm{E}(\boldsymbol{X}_{2j}) = \boldsymbol{\mu}_2$ para $j = 1, \ldots, n_2$,

	Média amostral	Matriz de covariâncias amostrais
Pop. 1	$\overline{\boldsymbol{X}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \boldsymbol{X}_{1j}$	$S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1} (\boldsymbol{X}_{1j} - \overline{\boldsymbol{X}}_1)(\boldsymbol{X}_{1j} - \overline{\boldsymbol{X}}_1)^\top$
Pop. 2	$\overline{\boldsymbol{X}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \boldsymbol{X}_{2j}$	$S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2 - 1} (\boldsymbol{X}_{2j} - \overline{\boldsymbol{X}}_2)(\boldsymbol{X}_{2j} - \overline{\boldsymbol{X}}_2)^\top$

Comparação de médias de duas populações independentes

Suposições:

1. A população 1 é independente da população 2.
2. Ambas as populações têm distribuição normal multivariada, ou seja $\boldsymbol{X}_j \sim N_p(\boldsymbol{\mu}, \Sigma_j)$, $j = 1, 2$.

Deseja-se tetar as hipóteses

$$\begin{array}{l} H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \\ H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \end{array} \iff \begin{array}{l} H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0 \\ H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0 \end{array}$$

onde $\boldsymbol{\delta}_0 = \mathbf{0}$

Caso I: Σ_1 e Σ_2 são conhecidos

A estatística de teste é

$$(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \left(\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2} \right)^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) \overset{sob\ H_0}{\sim} \chi^2_{(p)}$$

Pois

$$\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 \sim N_p(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2})$$

Caso II: $\Sigma_1 = \Sigma_2 = \Sigma$ desconhecido

Um estimador para Σ, é

$$S = S_{pooled} = \frac{\sum_{j=1}^{n_1}(\boldsymbol{X}_{1j} - \overline{\boldsymbol{X}}_1)(\boldsymbol{X}_{1j} - \overline{\boldsymbol{X}}_1)^\top + \sum_{j=1}^{n_2}(\boldsymbol{X}_{2j} - \overline{\boldsymbol{X}}_2)(\boldsymbol{X}_{2j} - \overline{\boldsymbol{X}}_2)^\top}{n_1 + n_2 - 2}$$

ou seja

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

Note que

$$\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 \sim N_p(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$$

e

$$(n_1 + n_2 - 2)S_{pooled} \sim W_p(n_1 + n_2 - 2, \Sigma).$$

Da definição da distribuição T^2 de Hotteling tem-se

$$(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pooled}\right]^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$$

a qual tem distribuição

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1+n_2-p-1}.$$

Então, reescrevemos as hipóteses de interesse na forma mais geral

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$$
$$H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0.$$

e rejeitamos H_0 ao nível de significância α se

$$T^2_{obs} = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - \boldsymbol{\delta}_0)^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S \right]^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - \boldsymbol{\delta}_0) > c^2$$

com $c^2 = \dfrac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} q_{F_{p, n_1+n_2-p-1, \alpha}}$.

Comparação de médias de duas populações independentes: Exemplo

Os dados a seguir referem-se à produtividade e altura de plantas de duas variedades de milho (A e B).

A		B	
Produtividade	Altura de Planta	Produtividade	Altura de Planta
5,70	2,10	4,4	1,80
8,90	1,90	7,5	1,76
6,20	1,98	5,40	1,78
5,80	1,92	4,60	1,89
6,80	2,00	5,90	1,90
6,20	2,01		

Verifique se a produtividade e altura de planta são as mesma das duas variedades de milho.

Comparação de médias de duas populações independentes: Exemplo

Seja

- X_{j1}: produtividade da j-ésima variedade;
- X_{j2}: Altura da planta da j-ésima variedade.

Supor

$$\boldsymbol{X}_j = \begin{pmatrix} X_{j1} \\ X_{j2} \end{pmatrix} \sim N_2 \left(\begin{bmatrix} \mu_{1j} \\ \mu_{2j} \end{bmatrix}, \Sigma \right), \; j = 1, 2$$

O interesse é testar as seguintes hipóteses

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \begin{pmatrix} \mu_{11} - \mu_{12} \\ \mu_{21} - \mu_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \; vs \; H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \begin{pmatrix} \mu_{11} - \mu_{12} \\ \mu_{21} - \mu_{22} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Comparação de médias de duas populações independentes: Exemplo

Dos dados tem-se $n_1 = 6$, $n_2 = 5$, p=2,

$$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 6.600 \\ 1.985 \end{pmatrix}; \; S_1 = \begin{pmatrix} 1.420 & -0.050 \\ -0.050 & 0.005 \end{pmatrix}$$

$$\bar{\boldsymbol{x}}_2 = \begin{pmatrix} 5.560 \\ 1.826 \end{pmatrix}; \; S_2 = \begin{pmatrix} 1.543 & -0.032 \\ -0.032 & 0.004 \end{pmatrix}$$

A matriz de covariâncias da amostra combinada

$$S_{pooled} = \begin{pmatrix} 1.475 & -0.042 \\ -0.042 & 0.005 \end{pmatrix} \text{ e } S_{pooled}^{-1} = \begin{pmatrix} 0.911 & 8.165 \\ 8.165 & 286.091 \end{pmatrix}$$

$$T_{obs}^2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - \boldsymbol{\delta}_0)^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - \boldsymbol{\delta}_0) = 29,778$$

Para $\alpha = 0,05$, $\frac{(9)2}{8} q_{F_{2,8,0,05}} = 10.03268 < T_{obs}$, rejeita-se H_O

Região de 100(1-α)% de confiança para $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. É o conjunto de pontos de $\boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \begin{pmatrix} \mu_{11} - \mu_{12} \\ \mu_{21} - \mu_{22} \end{pmatrix}$ tal que

$$\frac{n_1 n_2}{n_1 + n_2} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - \boldsymbol{\delta})^\top S_{pooled}^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - \boldsymbol{\delta}) \leq c^2,$$

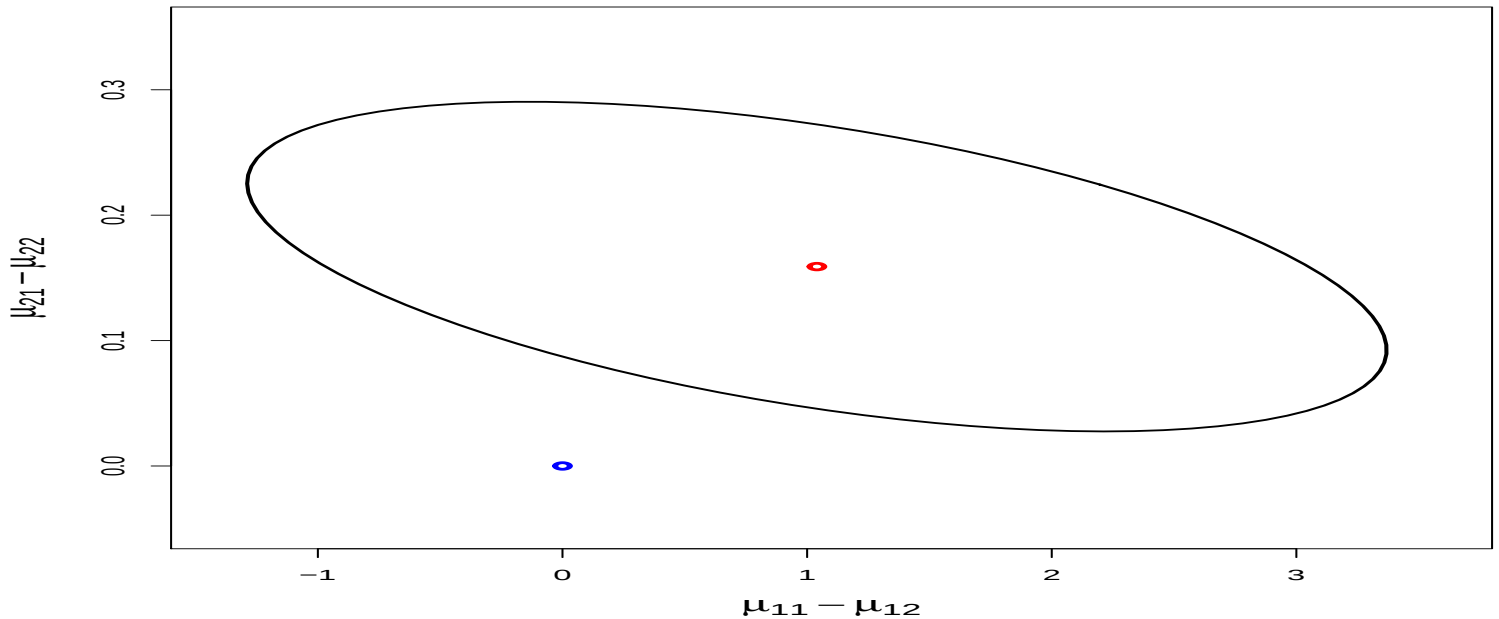
onde

$$c^2 = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} q_{F_{p, n_1+n_2-p-1, \alpha}}.$$

Comparação de médias de duas populações independentes: Exemplo

Uma região 95% de confiança para $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ é dada por

$$\frac{30}{11}\begin{pmatrix}1,01-\delta_1 & 0,17-\delta_2\end{pmatrix}\begin{pmatrix}0.911 & 8.165\\ 8.165 & 286.091\end{pmatrix}\begin{pmatrix}1,01-\delta_1\\ 0,17-\delta_2\end{pmatrix} \leq 10.03268$$



Verifica-se da Figura, que a origem (0,0), não pertence a região de confiança, indicando que as duas variedades diferem na produtividade média ou tamanho da planta média.

Intervalos de confiança simulatânes T^2

Seja $c^2 = \dfrac{(n_1+n_2-2)p}{n_1+n_2-p-1} q_{F_{p,n_1+n_2-p-1,\alpha}}$.

- Um intervalo com $100(1-\alpha)\%$ de confiança para $\boldsymbol{a}^\top(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ é dada por

$$\boldsymbol{a}^\top(\bar{\boldsymbol{x}}_1-\bar{\mu}_2) \mp c\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sqrt{\boldsymbol{a}^\top S_{poleed}\boldsymbol{a}}$$

- Por meio de escolhas apropriadas para $\boldsymbol{a}$, podemos obter intervalos das componentes:

$$\boldsymbol{a}_1 = \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix},\ \boldsymbol{a}_2 = \begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix},\ \ldots,\ \boldsymbol{a}_p = \begin{pmatrix}0\\0\\\vdots\\1\end{pmatrix}$$

Portanto, os intervalos de componentes são

$$\bar{x}_{11} - \bar{x}_{12} \mp c\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sqrt{S_{poleed,11}}$$

$$\bar{x}_{21} - \bar{x}_{22} \mp c\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sqrt{S_{poleed,22}}$$

$$\vdots$$

$$\bar{x}_{p1} - \bar{x}_{p2} \mp c\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sqrt{S_{poleed,pp}}$$

Intervalos de confiança simulatânes T^2

Dos dados e $1-\alpha = 0,95$,tem-se

$$c^2 = \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} q_{F_{p,n_1+n_2-p-1,\alpha}} = \frac{(9)2}{8} q_{F_{2,8,0,05}} = 10.03268$$

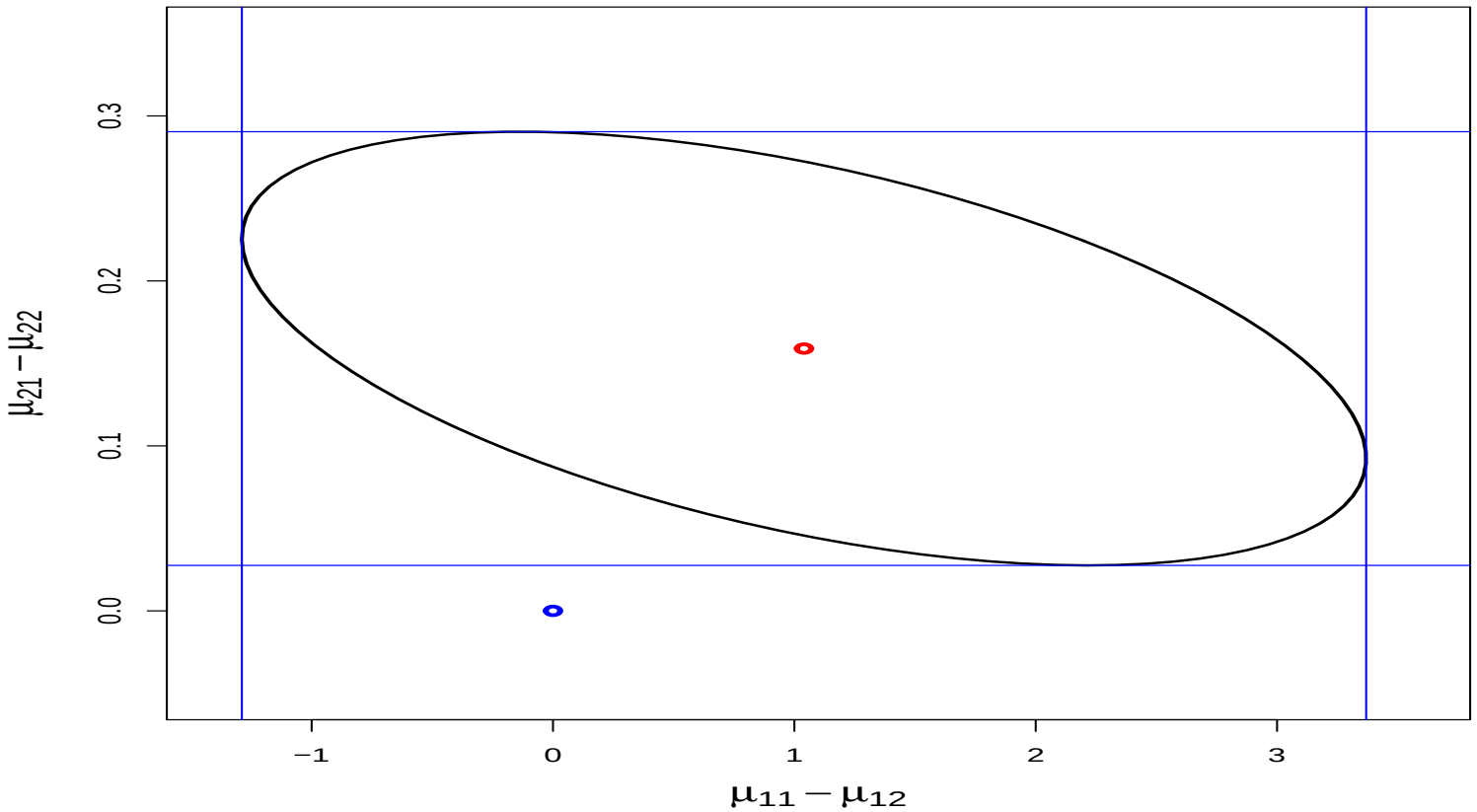
- O intervalo de 95% de confiança para $\mu_{11}-\mu_{12}$ é dado por

$$(6,6-5,56) \mp \sqrt{10,033}\sqrt{\frac{11}{6\times 5}}\sqrt{1,475} = (-1.289, 3.369)$$

- O intervalo de 95% de confiança para $\mu_{21}-\mu_{22}$ é dado por

$$(1,985-1,826) \mp \sqrt{10,033}\sqrt{\frac{11}{6\times 5}}\sqrt{0,004} = (0,028, 0,290)$$

Intervalos de confiança simulatânes T^2



Intervalos de Bonferroni e um por vez

Para intervalos de Bonferroni e Um por vez (ou seja, método univariado), basta alterar simplesmente o valor de c.

- Bonferroni

$$c = t_{n_1+n_2-2}(\alpha/2m)$$

- Um por vez

$$c = t_{n_1+n_2-2}(\alpha/2)$$

Intervalos de Bonferroni:Exemplo

Dos dados e $1-\alpha = 0,95$,tem-se
$c = t_{n_1+n_2-2}(\alpha/2(m)) = qt(\alpha/4, 6+5-2) = 2.685011$

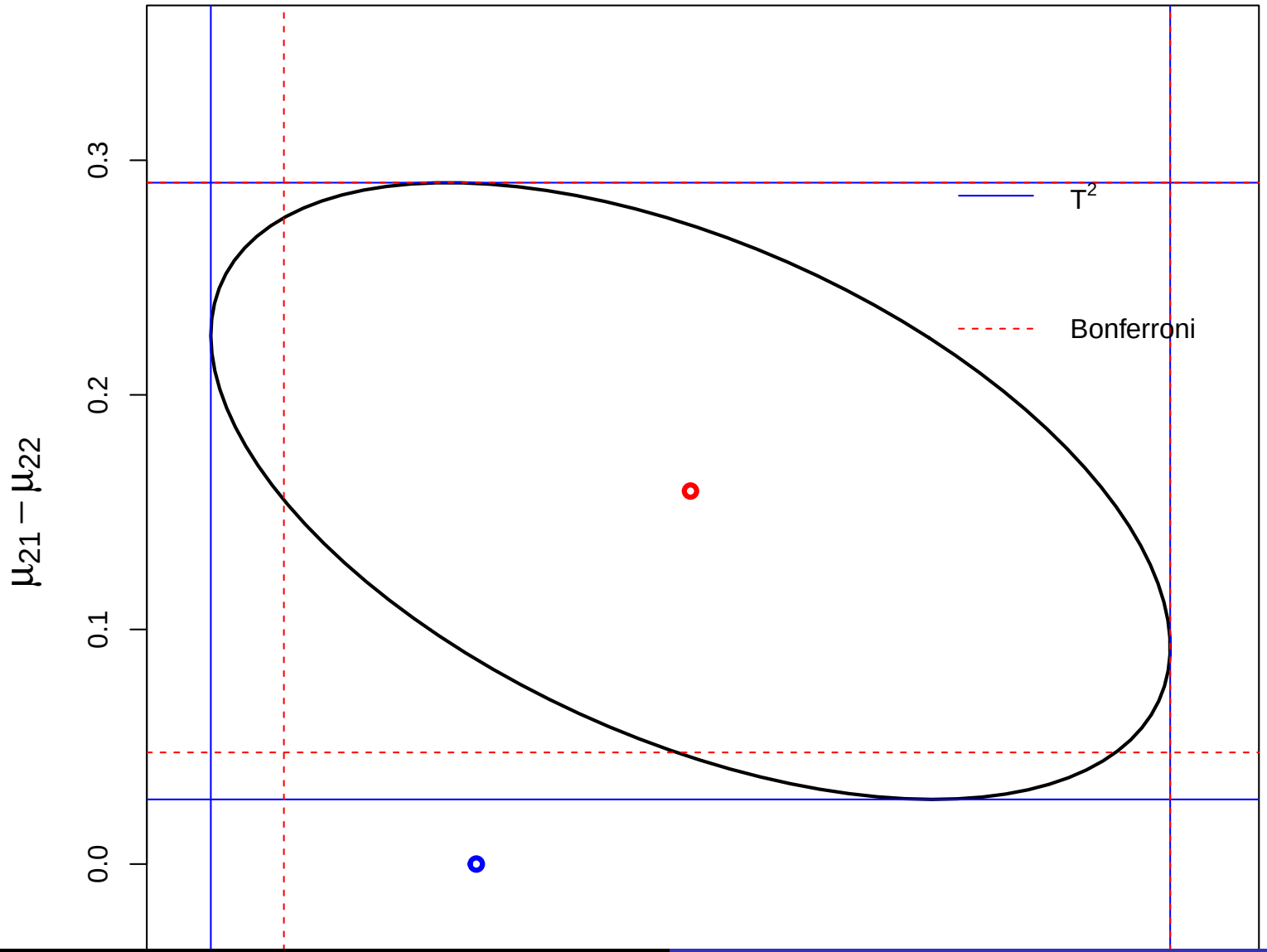
- O intervalo de 95% de confiança para $\mu_{11} - \mu_{12}$ é dado por

$$(6,6-5,56) \mp 2.685\sqrt{\frac{11}{6\times 5}}\sqrt{1,475} = (-0.934, 3.014))$$

- O intervalo de 95% de confiança para $\mu_{21} - \mu_{22}$ é dado por

$$(1,985-1,826) \mp 2.685\sqrt{\frac{11}{6\times 5}}\sqrt{0,004} = (0,048, 0,270)$$

Intervalo de Bonferroni e T^2



Caso III: $n_1 - p$ e $n_2 - p$ são grandes

Se $n_1 - p$ e $n_2 - p$ são grandes, não é necessário supor:

- $\Sigma_1 = \Sigma_2$
- X_{1j} tem distribuição normal multivariada;
- X_{2j} tem distribuição normal multivariada.

Supor:

- As populações são independentes;
- $X_{11}, \ldots, x_{1n_1}$ é uma amostra da população 1 com vetor de médias $\boldsymbol{\mu}_1$ e covariâncias Σ_1
- $X_{21}, \ldots, x_{2n_2}$ é uma amostra da população 2 com vetor de médias $\boldsymbol{\mu}_2$ e covariâncias Σ_2.

Se $n_1 - p$ e $n_2 - p$ são amostrras grandes, então distribuição para a estatística de teste T^2 é aproximadamente $\chi^2_{(p)}$.

Caso III: $n_1 - p$ e $n_2 - p$ são grandes

- A matriz de covariância das diferenças do vetor de médias

$$\begin{aligned} Cov(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) &= Cov(\bar{\boldsymbol{X}}_1) + Cov(\bar{\boldsymbol{X}}_2) \\ &= \frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2} \end{aligned}$$

pode ser estimado por

$$\frac{S_1}{n_1} + \frac{S_2}{n_2}$$

- A estatística de teste para testar $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$ é

$$T^2 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta_0})^\top \left(\frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta_0}) \; \widetilde{H}_0 \; \chi^2_{(p)}.$$

Comparação de duas populações independentes

- Uma região (elipsoide) de 100(1-α)% de confiança para $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ é dado por

$$(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta})^\top \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta}) \leq \chi^2_{(p)}(\alpha).$$

- Um intervalo de 100$(1-\alpha)$% de confiança para $\boldsymbol{a}^\top \boldsymbol{\mu}$ é dado por

$$\boldsymbol{a}^\top \bar{\boldsymbol{x}} \mp \sqrt{\chi^2(\alpha)} \sqrt{\boldsymbol{a}^\top \left[\frac{S_1}{n_1} + \frac{S_2}{n_2}\right] \boldsymbol{a}}$$

Exemplo Usando Amostras Grandes

Dos dados tem-se

$$\frac{S_1}{n_1} + \frac{S_2}{n_2} = \frac{1}{6}\begin{pmatrix} 1.420 & -0.050 \\ -0.050 & 0.005 \end{pmatrix} + \frac{1}{5}\begin{pmatrix} 1.543 & -0.032 \\ -0.032 & 0.004 \end{pmatrix}$$
$$= \begin{pmatrix} 0.545 & -0.0147 \\ -0.015 & 0.0017 \end{pmatrix}$$

A estatística de teste para testar $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0 = \mathbf{0}$ é

$$T^2_{obs} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = 29.141$$

Para $\alpha = 0,05$, tem-se o valor crítico $c = \chi^2_p(\alpha) = 5.991 < T^2_{obs}$ rejeita-se H_0. O nível descritivo,

$$pvalor = P(\chi^2_2 > 29,141) = 4.699823e - 07$$

Exemplo Usando Amostras Grandes

- Usando os mesmos vetores da combinação linear acima:

$$\boldsymbol{a}_1^\top = (1, 0), \implies \boldsymbol{a}_1^\top \boldsymbol{\mu} = \mu_{11} - \mu_{12}.$$
$$\boldsymbol{a}_2^\top = (0, 1), \implies \boldsymbol{a}_2^\top \boldsymbol{\mu} = \mu_{21} - \mu_{22}$$

  Para $1 - \alpha = 0,95$ de nível de confiança tem-se $c = \chi^2_{(2)} = 5.991$

- O intervalo de 95% de confiança para $\mu_{11} - \mu_{12}$ é dado por

$$(6,6 - 5,56) \mp \sqrt{5.991}\sqrt{0.545} = (-1,28, 3.369))$$

- O intervalo de 95% de confiança para $\mu_{21} - \mu_{22}$ é dado por

$$(1,985 - 1,826) \mp \sqrt{5.991}\sqrt{0.0017} = (0,028, 0,290)$$