Estatística aplicada

Licenciatura em Engenharia Informática

Lino Costa

Departamento de Produção e Sistemas
Escola de Engenharia
lac@dps.uminho.pt

Revisto em 2022/2023

Sumário

1. Planeamento de experiências
   - tipos de variáveis, fator, níveis, tratamentos e réplicas
   - planeamentos equilibrados e desequilibrados
   - amostras independentes e dependentes
2. Análise de variância
3. Planeamento completamente aleatório (PCA)
   - modelo
   - tabela ANOVA e teste de hipótese
   - comparações múltiplas: intervalo de confiança e teste de hipótese para a diferença entre dois tratamentos
4. Planeamento com blocos aleatórios (PBA)
   - modelo
   - tabela ANOVA e testes de hipóteses
   - comparações múltiplas: intervalo de confiança e teste de hipótese para a diferença entre dois tratamentos

Planeamento de experiências

Tipos de variáveis

- **variável dependente (resposta)** - a resposta de interessa no estudo
- **variável independente (fator)** - a variável cujo efeito na resposta pretende ser estudado
- **variável externa (fator "ruído")** - a variável cujo efeito na resposta não está ser considerado no estudo mas pode afetar a resposta

Terminologia

- **Tratamento (nível de fator)** - o termo tratamento refere-se a uma condição individual do fator.
- **Réplica** - o termo réplica refere-se a uma medição repetida nas mesmas condições experimentais.
- **Planeamento equilibrado** - um planeamento equilibrado corresponde a um planeamento em que o número de réplicas por tratamento é igual. Caso contrário, o planeamento será um **planeamento desequilibrado**.
- **Amostras independentes** - as observações das amostras são independentes (não existe heterogeneidade entre os valores observados). Caso contrário, as amostras serão amostras dependentes ou relacionadas, existindo **blocos**.

Planeamento de experiências

Exemplo 1

Um estudo for realizado para estudar o desenvolvimento de moscas medido pelo comprimento das asas (em $mm \times 10^{-1}$). O procedimento experimental consistiu na criação de moscas em três meio de cultura diferentes. Os resultados experimentais obtidos são indicados na seguinte tabela onde se apresenta o comprimento das asas de 5 moscas recolhidas aleatoriamente de cada meio de cultura.

Meio 1	Meio 2	Meio 3
36	50	45
39	42	53
43	51	56
38	40	52
37	43	56

Identifique qual a variável resposta, a variável independente (fator), os tratamentos e o número de réplicas. Diga se as amostras são independentes e se o planeamento é equilibrado.

- variável resposta: comprimento das asas das moscas (em $mm \times 10^{-1}$)
- variável independente (fator): meio de cultura
- tratamentos: meio 1, meio 2 e meio 3 (3 níveis do fator meio de cultura)
- réplicas: 5 réplicas por tratamento (5 observações por amostra)
- as amostras são independentes pois não existe heterogeneidade entre as observações das amostras
- o planeamento é equilibrado pois o número de réplicas por tratamento é igual (o número de observações por amostra é igual)

Planeamento de experiências

Exemplo 2

Foi realizado um estudo sobre consumo de combustível dos automóveis quando estes utilizam 3 tipos de gasolina sem chumbo. Para o efeito foram selecionados 5 automóveis idênticos mas conduzidos por diferentes pilotos. Cada automóvel percorreu o mesmo percurso nas mesmas condições com cada um dos tipos de gasolina, tendo-se registado o consumo de combustível (em $l/100km$).

Piloto	Gasolina A	Gasolina B	Gasolina C
P1	8.9	9.5	8.9
P2	7.9	8.0	8.0
P3	9.0	8.8	8.9
P4	9.1	9.0	9.2
P5	7.7	8.1	8.0

Identifique qual a variável resposta, a variável independente (fator), os tratamentos e o número de réplicas. Diga se as amostras são independentes.

- variável resposta: consumo de combustível (em $l/100km$)
- variável independente (fator): tipo de gasolina
- tratamentos: gasolina 1, gasolina 2 e gasolina 3 (3 níveis do fator tipo de gasolina)
- réplicas: 5 réplicas por tratamento (5 observações por amostra)
- as amostras não são independentes pois existe heterogeneidade entre as observações das amostras devido aos pilotos (existem 5 blocos correspondentes a cada um dos pilotos)

Análise de variância

ANOVA

ANOVA é um método que é usado para testar a significância do efeito de um fator (variável independente) numa resposta (variável dependente).
São exemplos de análise de variância:

- **Planeamento Completamente Aleatório (PCA)** - análise de um fator com amostras independentes (extensão a mais do que duas amostras do teste t-student à diferença de médias para duas amostras independentes)
- **Planeamento com Blocos Aleatórios (PBA)** - análise de um fator com amostras relacionadas (extensão a mais do que duas amostras do teste t-student para duas amostras relacionadas)

A variabilidade da variável resposta é particionada em diversas componentes (fontes de variação). Por exemplo, no PCA consideram-se fontes de variação:

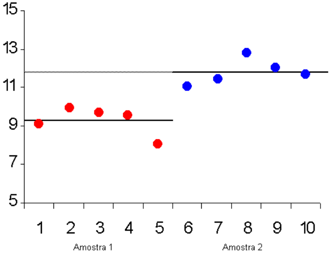
- a variabilidade explicada pelo fator
- a variabilidade devida a erros aleatórios

Análise de variância

Exemplo 3

(CASO 1) Considere os seguintes dados de duas amostras independentes:

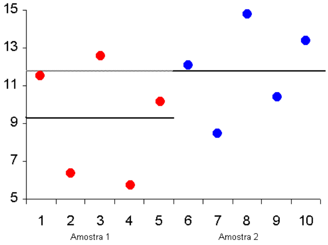
Amostra 1	Amostra 2
9.1	11.1
10	11.5
9.7	12.8
9.6	12.1
8.1	11.7
$\bar{x}_1 = 9.3$	$\bar{x}_2 = 11.8$
$s_1 = 0.75$	$s_2 = 0.65$



Exemplo 3

(CASO 2) Considere agora os seguintes dados de duas outras amostras independentes (com as médias amostrais idênticas às das amostras do **(CASO 1)**):

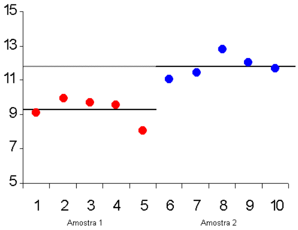
Amostra 1	Amostra 2
11.5	12.1
6.4	8.5
12.6	14.8
5.8	10.4
10.2	13.4
$\bar{x}_1 = 9.3$	$\bar{x}_2 = 11.8$
$s_1 = 3.05$	$s_2 = 2.74$

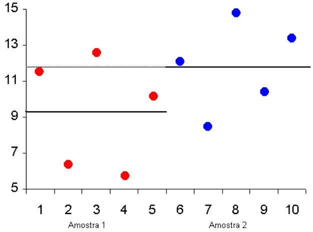


Análise de variância

Exemplo 3

Comparando a variabilidade, parece haver diferenças entre as médias no **(CASO 1)** e não haver diferenças entre as médias no **(CASO 2)**.





Análise de variância

Exemplo 3

Consideremos o **(CASO 2)**, como $(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$, a variabilidade total pode ser particionada em dois componentes:

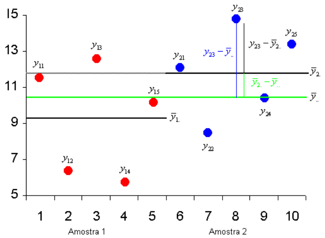
$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

em que k é o número de amostras ($k = 2$) e n é o número de observações por amostra ($n = 5$).

Logo, a Soma Total dos Quadrados (STQ) é particionada em

- Soma dos Quadrados dos Tratamentos (SQT)
- Soma dos Quadrados dos Resíduos (SQR)

isto é, $STQ = SQT + SQR$.



Planeamento Completamente Aleatório (PCA)

Contexto

Análise de variância com um fator com k amostras independentes (tratamentos) cada uma com n_j observações (réplicas).[1]

Modelo populacional

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} = \mu_j + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, n_j \\ j = 1, 2, \ldots, k \end{cases}$$

onde

- Y_{ij} é a variável resposta
- μ é a média global
- α_j é o efeito do tratamento j
- μ_j é a média do tratamento j
- ε_{ij} são erros aleatórios com $\varepsilon_{ij} \sim N(0, \sigma^2)$

---

[1] No caso de amostras equilibradas, $n_1 = n_2 = \ldots = n_k = n$.

Planeamento Completamente Aleatório (PCA)

Pressupostos

- variável resposta normalmente distribuída, i.e., $Y_{ij} \sim N(\mu_j, \sigma^2)$
- homogeneidade das variâncias dos tratamentos

Hipóteses

$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
(Não existem diferenças significativas entre os tratamentos)
$H_1 : \alpha_j \neq 0, \exists j = 1, \ldots, k$
(existem diferenças significativas entre os tratamentos)

Região de rejeição

$R.R. : F > F_{\alpha,k-1,N-k}$ onde $N = \sum_{j=1}^{k} n_j$ e α é o nível de significância.

Planeamento Completamente Aleatório (PCA)

Tabela ANOVA e estatística de teste F

Fonte	SQ	gl	MQ	F
Tratamentos	SQT	$k-1$	$MQT=\frac{SQT}{k-1}$	$F=\frac{MQT}{MQR}$
Resíduos	SQR	$N-k$	$MQR=\frac{SQR}{N-k}$	
Total	STQ	$N-1$		

- $SQT=\sum_{j=1}^{k}\frac{T_{.j}^2}{n_j}-\frac{T_{..}^2}{N}$ com $T_{.j}=\sum_{i=1}^{n_j}y_{ij}$ e $T_{..}=\sum_{j=1}^{k}T_{.j}$
- $STQ=\sum_{j=1}^{k}\sum_{i=1}^{n_j}y_{ij}^2-\frac{T_{..}^2}{N}$
- $SQR=STQ-SQT$

Planeamento Completamente Aleatório (PCA)

Intervalo de confiança para a diferença entre dois tratamentos

$$(\bar{y}_{.i} - \bar{y}_{.j}) \pm t_{\alpha/2,N-k}\sqrt{MQR\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Teste de hipótese para a diferença entre dois tratamentos

$$H_0 : \mu_i - \mu_j = 0$$
$$H_1 : \mu_i - \mu_j \neq 0$$

$$E.T. : T = \frac{(\bar{y}_{.i} - \bar{y}_{.j}) - (\mu_i - \mu_j)}{\sqrt{MQR\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t_{N-k}$$

$$R.R. : |T| > t_{\alpha/2,N-k}$$

Planeamento Completamente Aleatório (PCA)

Exemplo 4

Um estudo for realizado para estudar o desenvolvimento de moscas medido pelo comprimento das asas (em $mm \times 10^{-1}$). O procedimento experimental consistiu na criação de moscas em três meio de cultura diferentes. Os resultados experimentais obtidos são indicados na seguinte tabela onde se apresenta o comprimento das asas de 5 moscas recolhidas aleatoriamente de cada meio de cultura.

Meio 1	Meio 2	Meio 3
36	50	45
39	42	53
43	51	56
38	40	52
37	43	56

Verifique se existem diferenças significativas para $\alpha = 5\%$ entre os comprimentos das asas das moscas recolhidas de cada meio.

- fator: meio de cultura; $k = 3$ (3 tratamentos); amostras independentes, logo PCA
- pressupostos: variável resposta (comprimento das asas) normalmente distribuída, i.e., $Y_{ij} \sim N(\mu_j, \sigma^2)$ e homogeneidade das variâncias dos tratamentos
- $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$
  (Não existem diferenças significativas no comprimentos das asas das moscas dos 3 meios de cultura)
  $H_1 : \alpha_j \neq 0, \exists j = 1, 2, 3$
  (existem diferenças significativas no comprimentos das asas das moscas dos 3 meios de cultura)

Planeamento Completamente Aleatório (PCA)

Exemplo 4

	Meio 1	Meio 2	Meio 3	
$T_{.j}$	193	226	262	$681 = T_{..}$
n_j	5	5	5	$15 = N$

- $\sum_{j=1}^{3}\sum_{i=1}^{n_j} y_{ij}^2 = 31603$
- $SQT = \sum_{j=1}^{3} \frac{T_{.j}^2}{n_j} - \frac{T_{..}^2}{N} = \frac{193^2}{5} + \frac{226^2}{5} + \frac{262^2}{5} - \frac{681^2}{15} = 476.4$
- $STQ = \sum_{j=1}^{3}\sum_{i=1}^{n_j} y_{ij}^2 - \frac{T_{..}^2}{N} = 31603 - \frac{681^2}{15} = 685.6$
- $SQR = STQ - SQT = 685.6 - 476.4 = 209.2$
Fonte	SQ	gl	MQ	F
Tratamentos	$SQT = 476.4$	2	$MQT = \frac{476.4}{2} = 238.2$	$F = \frac{238.2}{17.433} = 13.663$
Resíduos	$SQR = 209.2$	12	$MQR = \frac{209.2}{12} = 17.433$	
Total	$STQ = 685.6$	14		

- $R.R. : F > F_{0.05,2,12} = 3.89$ (Tabela 8)
- Decisão: Como $13.633 > 3.89$, rejeita-se H_0 para $\alpha = 5\%$, pelo que existem diferenças estatisticamente significativas entre os valores médios de crescimento das asas nos 3 meios de cultura.

Planeamento com Blocos Aleatórios (PBA)

Contexto

Análise de variância com um fator com k amostras dependentes (tratamentos) e b blocos de observações relacionadas.

Modelo populacional

$$Y_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij} = \mu_j + \beta_i + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, b \\ j = 1, 2, \ldots, k \end{cases}$$

onde

- Y_{ij} é a variável resposta
- μ é a média global
- α_j é o efeito do tratamento j
- β_i é o efeito do bloco i
- μ_j é a média do tratamento j
- ε_{ij} são erros aleatórios com $\varepsilon_{ij} \sim N(0, \sigma^2)$

Planeamento com Blocos Aleatórios (PBA)

Pressupostos

- variável resposta normalmente distribuída, i.e., $Y_{ij} \sim N(\mu_j, \sigma^2)$
- homogeneidade das variâncias dos tratamentos e blocos

Hipóteses

$H_{01} : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
(Não existem diferenças significativas entre os tratamentos)
$H_{11} : \alpha_j \neq 0, \exists j = 1, \ldots, k$
(existem diferenças significativas entre os tratamentos)

$H_{02} : \beta_1 = \beta_2 = \ldots = \beta_b = 0$
(Não existem diferenças significativas entre os blocos)
$H_{12} : \beta_i \neq 0, \exists i = 1, \ldots, b$
(existem diferenças significativas entre os blocos)

Regiões de rejeição

$R.R. : F_1 > F_{\alpha, k-1, (k-1)(b-1)}$ e $R.R. : F_2 > F_{\alpha, b-1, (k-1)(b-1)}$ onde α é o nível de significância.

Planeamento com Blocos Aleatórios (PBA)

Tabela ANOVA e estatísticas de teste F_1 e F_2

Fonte	SQ	gl	MQ	F
Tratamentos	SQT	$k-1$	$MQT = \frac{SQT}{k-1}$	$F_1 = \frac{MQT}{MQR}$
Blocos	SQB	$b-1$	$MQB = \frac{SQB}{b-1}$	$F_2 = \frac{MQB}{MQR}$
Resíduos	SQR	$(k-1)(b-1)$	$MQR = \frac{SQR}{(k-1)(b-1)}$	
Total	STQ	$kb-1$		

- $SQT = \frac{\sum_{j=1}^{k} T_{.j}^2}{b} - \frac{T_{..}^2}{kb}$ com $T_{.j} = \sum_{i=1}^{b} y_{ij}$ e $T_{..} = \sum_{j=1}^{k} T_{.j} = \sum_{i=1}^{b} T_{i.}$
- $SQB = \frac{\sum_{i=1}^{b} T_{i.}^2}{k} - \frac{T_{..}^2}{kb}$ com $T_{i.} = \sum_{j=1}^{k} y_{ij}$
- $STQ = \sum_{j=1}^{k} \sum_{i=1}^{b} y_{ij}^2 - \frac{T_{..}^2}{kb}$
- $SQR = STQ - SQT - SQB$

Planeamento com Blocos Aleatórios (PBA)

Intervalo de confiança para a diferença entre dois tratamentos

$$(\bar{y}_{.i} - \bar{y}_{.j}) \pm t_{\alpha/2,(k-1)(b-1)} \sqrt{MQR \left(\frac{2}{b} \right)}$$

Teste de hipótese para a diferença entre dois tratamentos

$$H_0 : \mu_i - \mu_j = 0$$
$$H_1 : \mu_i - \mu_j \neq 0$$

$$E.T. : T = \frac{(\bar{y}_{.i} - \bar{y}_{.j}) - (\mu_i - \mu_j)}{\sqrt{MQR \left(\frac{2}{b} \right)}} \sim t_{(k-1)(b-1)}$$

$$R.R. : |T| > t_{\alpha/2,(k-1)(b-1)}$$

Planeamento com Blocos Aleatórios (PBA)

Exemplo 5

Foi realizado um estudo sobre consumo de combustível dos automóveis quando estes utilizam 3 tipos de gasolina sem chumbo. Para o efeito foram selecionados 5 automóveis idênticos mas conduzidos por diferentes pilotos. Cada automóvel percorreu o mesmo percurso nas mesmas condições com cada um dos tipos de gasolina, tendo-se registado o consumo de combustível (em $l/100km$).

Piloto	Gasolina A	Gasolina B	Gasolina C
P1	8.9	9.5	8.9
P2	7.9	8.0	8.0
P3	9.0	8.8	8.9
P4	9.1	9.0	9.2
P5	7.7	8.1	8.0

O que pode concluir para $\alpha = 5\%$?

- fator: tipo de gasolina; $k = 3$ (3 tratamentos); amostras dependentes com $b = 5$ (5 blocos), logo PBA
- pressupostos: variável resposta (consumo de combustível) normalmente distribuída, i.e., $Y_{ij} \sim N(\mu_j, \sigma^2)$ e homogeneidade das variâncias dos tratamentos e dos blocos
- $H_{01} : \alpha_1 = \alpha_2 = \alpha_3 = 0$
  (Não existem diferenças significativas nos consumos de combustível com os diferentes tipos de gasolina)
  $H_{11} : \alpha_j \neq 0, \exists j = 1, 2, 3$
  (existem diferenças significativas nos consumos de combustível com os diferentes tipos de gasolina)
- $H_{02} : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
  (Não existem diferenças significativas nos consumos de combustível com os diferentes pilotos)
  $H_{12} : \beta_i \neq 0, \exists i = 1, 2, 3, 4, 5$
  (existem diferenças significativas nos consumos de combustível com os diferentes pilotos)

Planeamento com Blocos Aleatórios (PBA)

Exemplo 5

-

Piloto	Gasolina A	Gasolina B	Gasolina C	$T_{i.}$
P1	8.9	9.5	8.9	27.3
P2	7.9	8.0	8.0	23.9
P3	9.0	8.8	8.9	26.7
P4	9.1	9.0	9.2	27.3
P5	7.7	8.1	8.0	23.8
$T_{.j}$	42.6	43.4	43.0	$129.0 = T_{..}$

- $\sum_{j=1}^{3}\sum_{i=1}^{5} y_{ij}^2 = 1114.08$
- $SQT = \frac{\sum_{j=1}^{3} T_{.j}^2}{5} - \frac{T_{..}^2}{3\times 5} = \frac{42.6^2+43.4^2+43.0^2}{5} - \frac{129.0^2}{15} = 0.064$
- $SQB = \frac{\sum_{i=1}^{5} T_{i.}^2}{3} - \frac{T_{..}^2}{3\times 5} = \frac{27.3^2+23.9^2+26.7^2+27.3^2+23.8^2}{3} - \frac{129.0^2}{15} = 4.307$
- $STQ = \sum_{j=1}^{3}\sum_{i=1}^{5} y_{ij}^2 - \frac{T_{..}^2}{3\times 5} = 1114.08 - \frac{129.0^2}{15} = 4.680$
- $SQR = STQ - SQT - SQB = 4.68 - 0.064 - 4.307 = 0.309$
-

Fonte	SQ	gl	MQ	F
Tratamentos	$SQT = 0.064$	2	$MQT = \frac{0.064}{2} = 0.032$	$F_1 = \frac{0.032}{0.039} = 0.828$
Blocos	$SQB = 4.307$	4	$MQB = \frac{4.307}{4} = 1.077$	$F_2 = \frac{1.077}{0.039} = 27.845$
Resíduos	$SQR = 0.309$	8	$MQR = \frac{0.309}{8} = 0.039$	
Total	$STQ = 4.680$	14		

Planeamento com Blocos Aleatórios (PBA)

Exemplo 5

- $R.R.: F_1 > F_{0.05,2,8} = 4.46$ (Tabela 8)
- Decisão: Como $0.828 \leq 4.46$, não se rejeita H_{01} para $\alpha = 5\%$, pelo que poderão não existir diferenças estatisticamente significativas entre os consumos médios para os diferentes tipos de gasolina.
- $R.R.: F_2 > F_{0.05,4,8} = 3.84$ (Tabela 8)
- Decisão: Como $27.845 > 3.84$, rejeita-se H_{02} para $\alpha = 5\%$, pelo que existem diferenças estatisticamente significativas entre os consumos médios para os diferentes pilotos.