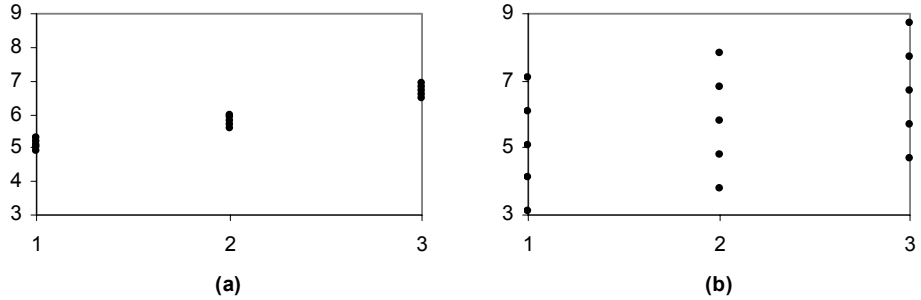
ANOVA - Análise da variância

1. Introdução

- Objectivos**:**
  - Testar a igualdade de médias para várias populações normais;
  - Atribuir a uma ou mais causas uma eventual diferença entre as referidas médias;
- Exemplo: Pretende-se aferir se o valor esperado de determinada variável aleatória é o mesmo nas diferentes regiões de um país.
- A hipótese em teste é então $H_0 : \mu_1 = \mu_2 = ... = \mu_m$. Para testar esta hipótese é necessário ter presente 2 aspectos fundamentais:
  - a forma como as *m* populações são definidas (com base num ou em mais critérios);
  - a variabilidade patenteada pelas amostras de cada uma das populações.

Exemplo:

- Pretende-se testar a igualdade de médias de três populações (A, B e C)
- Recolheu-se uma amostra casual simples de dimensão 5 em cada uma das populações e observou-se $\bar{x}_A = 5.7$, $\bar{x}_B = 5.8$, $\bar{x}_C = 6.7$.
- Observe-se como a nossa conclusão intuitiva seria diferente consoante se verificasse a situação (a) (populações homogéneas) ou (b) (populações heterogéneas)



Valores:

- Caso (a)
  - Pop 1 → 4.9; 5.0; 5.1; 5.2; 5.3
  - Pop 2 → 5.6; 5.7; 5.8; 5.9; 6.0
  - Pop 3 → 6.5; 6.6; 6.7; 6.8; 6.9
- Caso (b)
  - Pop 1 → 3.1; 4.1; 5.1; 6.1; 7.1
  - Pop 2 → 3.8; 4.8; 5.8; 6.8; 7.8
  - Pop 3 → 4.7; 5.7; 6.7; 7.7; 8.7

- **Ideia a reter:** A variabilidade dos dados observados, quer dentro de cada população quer entre as populações são aspectos fundamentais a ter em conta no teste da igualdade de médias. Assim sendo, parece razoável fundamentar o teste da hipótese H_0 na comparação entre estas variabilidades, técnica na qual se baseia a análise da variância.

2. Análise com um factor (one-way ANOVA) - Classificação simples

- A definição das *m* populações é feita considerando apenas **um** critério, designado por **factor**. Cada uma das populações corresponde a um nível do factor, que terá assim *m* níveis.
- Caso se rejeite a hipótese H_0, conclui-se, para a dimensão α escolhida, que as *m* populações não apresentam comportamento idêntico face ao critério ou factor que serviu para efectuar a classificação.

  Cuidado!!

  - Só é legitimo considerar este factor como a causa das diferenças entre as médias das populações, se se puder garantir a homogeneidade das populações face a todos os outros factores que poderiam ser relevantes para a explicação do fenómeno.
  - O contrário de todos iguais é pelo menos um diferentes e não todos diferentes.
- O processo que se vai estudar segue o modelo de efeitos fixos.
  - Recolhem-se *m* amostras casuais, independentes, cada um delas de dimensão n_i, isto é

$$\left(X_{i1}, X_{i2}, \ldots, X_{in_i}\right), \quad i = 1,2,\ldots,m$$

- As variáveis X_{ij} possuem distribuição normal com médias desconhecidas e **variância comum** também desconhecida,

$$X_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1,2,...,m, \quad j = 1,2,...,n_i.$$

Define-se μ tal que $\mu_i = \mu + \alpha_i$ $(i = 1,2,...,m)$, o que permite escrever, $X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

- Ao passar de $\mu_1 = \mu_2 = ... = \mu_m$, para $\mu_i = \mu + \alpha_i$ acrescentou-se uma variável, dispondo-se de um grau de liberdade para fixar um dos valores (μ ou um dos α_i). Opta-se, para simplificar as contas, por fixar $\sum_{i=1}^{m} n_i \alpha_i = 0$.

- O modelo de efeitos fixos assenta em cinco pressupostos que devem ser cuidadosamente ponderados antes de aplicar os resultados que se vão apresentar:

1. $E(\varepsilon_{ij}) = 0$ $(j = 1,2,...,n_i; i = 1,2,...,m)$.

2. ε_{ij} independentes, isto é, a amostragem é casual e as amostras são independentes entre si.

3. $\mathrm{V}(\varepsilon_{ij}) = \sigma^2$ $(j = 1,2,...,n_i; i = 1,2,...,m)$, isto é, verifica-se **homoscedasticidade** ou homogeneidade das variâncias.

   A violação deste pressuposto, isto é, a existência de **heteroscedasticidade**, pode ter sérias consequências no que diz respeito à validade das conclusões, consequências essas que podem ser minoradas se as amostras forem da mesma dimensão.

4. ε_{ij} tem distribuição normal.

   A violação deste pressuposto pode não ter consequências sérias se a dimensão das amostras for razoavelmente grande (teorema do limite central).

5. O efeito do nível *i* do factor é representado pelo parâmetro α_i $(i = 1,2,...,m)$ num modelo linear e aditivo.

- A hipótese a testar, $H_0 : \mu_1 = \mu_2 = ... = \mu_m$, é equivalente a, $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_m = 0$ ou ainda a $H_0 : \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_m^2 = \sum_{i=1}^{m} \alpha_i^2 = 0$, expressão que permite testar uma única restrição.
- O método para testar a hipótese H_0 baseia-se na construção de dois estimadores independentes da variância σ^2:
  - o primeiro é um estimador válido quer a hipótese seja verdadeira quer não;
  - o segundo é um estimador válido apenas quando hipótese é verdadeira e que **sobrestima** a variância no caso contrário.

A estatística de teste vai ser dada pelo quociente entre o segundo estimador e o primeiro. Caso H_0 seja verdadeira, os dois estimadores tenderão a produzir estimativas aproximadamente iguais e, portanto, a estatística de teste tende a assumir valores próximos de 1. Caso H_0 seja falsa, o segundo estimador tende a originar estimativas mais elevadas do que as do primeiro e, consequentemente, a estatística de teste tende a assumir valores maiores do que 1.

A hipótese H_0 é assim de rejeitar numa situação concreta (conhecidas as observações que compõem as m amostras) se o quociente entre o segundo e o primeiro estimador se apresenta significativamente elevado.

- Para estabelecer a distribuição por amostragem da estatística de teste parte-se da **Identidade fundamental da Análise da Variância** ou ANOVA:

Soma de **Q**uadrados **T**otal = **S**oma de **Q**uadrados **D**entro das Amostras +**S**oma de **Q**uadrados **E**ntre Amostras

i.e. **SQT=SQD+SQE**

sendo SQD independente de SQE, com

$$\text{SQT} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{\circ\circ}\right)^2$$

$$\text{SQD} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i\circ}\right)^2 \qquad \overline{X}_{i\circ} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \quad (i = 1,2,...,m)$$ Média da amostra da população i

$$\text{SQE} = \sum_{i=1}^{m} n_i\left(\overline{X}_{i\circ} - \overline{X}_{\circ\circ}\right)^2$$

$$\overline{X}_{\circ\circ} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{m} n_i} = \frac{\sum_{i=1}^{m} n_i\,\overline{X}_{i\circ}}{n}$$ Média global

- A primeira parte obtém-se fazendo

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_{\circ\circ}\right)^2=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\left(X_{ij}-\overline{X}_{i\circ}\right)+\left(\overline{X}_{i\circ}-\overline{X}_{\circ\circ}\right)\right)^2$$
$$=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_{i\circ}\right)^2+\sum_{i=1}^{m}n_i\left(\overline{X}_{i\circ}-\overline{X}_{\circ\circ}\right)^2$$

que resulta da anulação do termo cruzado.

- A demonstração da independência é mais complicada e **não será feita**.

- Alguns resultados auxiliares:

$$E(\overline{X}_{i\circ})=\frac{1}{n_i}\sum_{j=1}^{n_i}E(X_{ij})=\frac{1}{n_i}\sum_{j=1}^{n_i}(\mu+\alpha_i)$$
$$=\mu+\alpha_i \qquad (i=1,2,...,m)$$

$$\mathrm{var}(\overline{X}_{i\circ})=\frac{1}{n_i^{\,2}}\sum_{j=1}^{n_i}\mathrm{var}(X_{ij})=\frac{1}{n_i^{\,2}}\sum_{j=1}^{n_i}\sigma^2=\frac{\sigma^2}{n_i}$$

$$E(\overline{X}_{\circ\circ})=\frac{\sum_{i=1}^{m}n_i\,E(\overline{X}_{i\circ})}{n}=\frac{\sum_{i=1}^{m}n_i\left(\mu+\alpha_i\right)}{n}$$
$$=\mu+\frac{\sum_{i=1}^{m}n_i\,\alpha_i}{n}=\mu$$

$$,\ \mathrm{var}(\overline{X}_{\circ\circ})=\frac{\sum_{i=1}^{m}n_i^{\,2}\,\mathrm{var}(\overline{X}_{i\circ})}{n^2}=\frac{\sum_{i=1}^{m}n_i^{\,2}\,\sigma^2/n_i}{n^2}=\frac{\sum_{i=1}^{m}n_i}{n^2}\sigma^2=\frac{\sigma^2}{n}$$

Primeiro estimador de σ^2, independente da validade de $H_0 \rightarrow \dfrac{\text{SQD}}{n-m}$

Como $\overline{X}_{i\circ}$ é a média da *i*-ésima amostra ($i = 1,2,...,m$), vem,

$$\frac{\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i\circ}\right)^2}{\sigma^2} = \frac{(n_i - 1)S_i'^2}{\sigma^2} \sim \chi^2(n_i - 1),$$

onde,

$$S_i'^2 = \frac{\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i\circ}\right)^2}{n_i - 1},$$

é a variância corrigida da amostra *i*.

Somando para as *m* amostras independentes tem-se

$$\frac{\sum_{i=1}^{m}(n_i - 1)S_i'^2}{\sigma^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i\circ}\right)^2}{\sigma^2} = \frac{\text{SQD}}{\sigma^2} \sim \chi^2(n-m),$$

dado que $\sum_{i=1}^{m}(n_i - 1) = n - m$.

Ora $E\left(\frac{\text{SQD}}{\sigma^2}\right) = n - m$ (o valor esperado de uma χ^2 é igual aos seus graus de liberdade), logo $\frac{\text{SQD}}{n-m}$ é estimador centrado de σ^2, quer H_0 seja verdadeira quer não.

- **Segundo estimador** de $\sigma^2 \rightarrow \frac{\text{SQE}}{m-1}$

Assumindo H_0 verdadeira, não existe diferença entres as sub-populações e pode-se considerar que se dispõe de uma amostra de dimensão *n* referente a uma única população. Logo

$$\frac{(n-1)S'^2}{\sigma^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{\circ\circ}\right)^2}{\sigma^2} = \frac{\text{SQT}}{\sigma^2} \sim \chi^2(n-1).$$

Como SQT e SQD não são independentes, torna-se mais proveitoso procurar um estimador que envolva apenas SQE e não SQT. Ora como

$\frac{\text{SQD}}{\sigma^2} \sim \chi^2(n-m)$ e $\frac{\text{SQT}}{\sigma^2} \sim \chi^2(n-1)$

$SQT = SQD + SQE$ e SQD, SQE independentes

tem-se

$$\frac{\text{SQE}}{\sigma^2} \sim \chi^2(m-1),$$

e, portanto, $\frac{\text{SQE}}{m-1}$ é estimador centrado de σ^2 quando a hipótese H_0 é verdadeira.

- Mostre-se agora que este estimador sobre estima, em média, σ^2 quando H_0 é falsa. Para tal calcula-se $E(\text{SQE})$ e mostra-se que $E(\text{SQE}) > (m-1)\sigma^2$ quando H_0 é falsa.
  - Ora
    $$E\left(\overline{X}_{i\circ} - \overline{X}_{\circ\circ}\right) = E\left(\overline{X}_{i\circ}\right) - E\left(\overline{X}_{\circ\circ}\right) = \left(\mu + \alpha_i\right) - \mu = \alpha_i$$
  - Por outro lado $\text{Var}(\overline{X}_{i\circ} - \overline{X}_{\circ\circ}) = \sigma^2\left(\frac{1}{n_i} - \frac{1}{n}\right)$

$$\mathrm{Var}(\overline{X}_{i\circ} - \overline{X}_{\circ\circ}) = \mathrm{Var}(\overline{X}_{i\circ}) + \mathrm{Var}(\overline{X}_{\circ\circ}) - 2\,\mathrm{Cov}(\overline{X}_{i\circ}, \overline{X}_{\circ\circ})$$

$$= \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n} - 2\,\mathrm{Cov}\left(\overline{X}_{i\circ}, \frac{\sum_{k=1}^{m} n_k \overline{X}_{k\circ}}{n}\right)$$

$$= \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n} - \frac{2}{n}\sum_{k=1}^{m} n_k\,\mathrm{Cov}(\overline{X}_{i\circ}, \overline{X}_{k\circ})$$

já que as covariâncias entre $\overline{X}_{i\circ}$ e $\overline{X}_{k\circ}$ são

$$= \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n} - \frac{2}{n} n_i \mathrm{Var}\,(\overline{X}_{i\circ}) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n} - \frac{2}{n}\right)$$

$$= \sigma^2\left(\frac{1}{n_i} - \frac{1}{n}\right)$$

nulas para $i \neq k$.

- Logo $E(\mathrm{SQE}) = (m\text{-}1)\,\sigma^2 + \sum_{i=1}^{m} n_i \alpha_i^2$

$$
\begin{aligned}
E(\text{SQE}) &= E\left[\sum_{i=1}^{m} n_i (\overline{X}_{i\circ} - \overline{X}_{\circ\circ})^2\right] \\
&= \sum_{i=1}^{m} n_i \, E(\overline{X}_{i\circ} - \overline{X}_{\circ\circ})^2 \\
&= \sum_{i=1}^{m} n_i \left\{\text{Var}(\overline{X}_{i\circ} - \overline{X}_{\circ\circ}) + \left[E(\overline{X}_{i\circ} - \overline{X}_{\circ\circ})\right]^2\right\} \\
&= \sum_{i=1}^{m} n_i \left[\left(\frac{1}{n_i} - \frac{1}{n}\right)\sigma^2 + \alpha_i^2\right] \\
&= \sum_{i=1}^{m} \left[\left(1 - \frac{n_i}{n}\right)\sigma^2\right] + \sum_{i=1}^{m} n_i \alpha_i^2 \\
&= (m\text{-}1)\,\sigma^2 + \sum_{i=1}^{m} n_i \alpha_i^2,
\end{aligned}
$$

recordando que, existindo segundos momentos, $E(X^2) = \text{var}(X) + \left[E(X)\right]^2$.

- Assim $E\left(\dfrac{\text{SQE}}{m-1}\right) = \sigma^2 + \dfrac{\sum_{i=1}^{m} n_i \alpha_i^2}{m-1} > \sigma^2$ quando H_0 é falsa, uma vez que, nesta caso, $\Sigma_i n_i \alpha_i^2 > 0$.

- **Comparação entre os estimadores.**

  Para testar $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_m = 0$ vai utilizar-se como estatística de teste

$$F = \frac{\text{SQE}/(m-1)}{\text{SQD}/(n-m)} \sim F(m-1, n-m)$$

- Como definir a região de rejeição?

  Quando H_0 é verdadeira quer o numerador quer o denominador são estimadores centrados de σ^2.

  Quando H_0 é falsa o numerador sobre-estima σ^2 enquanto o denominador é estimador centrado.

  Logo a região de rejeição é unilateral direita.

- A região de rejeição é dada por

$$F = \frac{\text{SQE}/(m-1)}{\text{SQD}/(n-m)} > F_\alpha,$$

  onde, recorde-se, F_α é o limiar da distribuição $F(m-1, n-m)$ que, quando H_0 é verdadeira, tem uma probabilidade igual a α de ser excedido.

- **Apresentação habitual dos resultados**

Tabela da ANOVA

(classificação simples)

Origem da variação	**Soma de quadrados**	**Graus de Liberdade**	**Médias quadráticas**
Entre amostras	SQE	$m-1$	$\text{MQE} = \text{SQE}/(m-1)$
Dentro das amostras	SQD	$n-m$	$\text{MQD} = \text{SQD}/(n-m)$
Total	SQT	$n-1$	$F = \text{MQE}/\text{MQD}$

- Quando se rejeita H_0 (as médias não são todas iguais), é geralmente interessante procurar onde se situam as possíveis diferenças. Uma solução é construir IC para a média de cada uma das populações, e compará-los. Em alternativa à abordagem "clássica" pode utilizar uma estimativa combinada para σ^2 uma vez que se assumiu que este parâmetros é comum a todas as populações. Tem-se assim

$$\left[\bar{x}_{i\circ} - t_{\alpha/2}\sqrt{\frac{\text{MQD}}{n_i}}; \bar{x}_{i\circ} + t_{\alpha/2}\sqrt{\frac{\text{MQD}}{n_i}} \right]$$

- **Exemplo**

  Uma fábrica de papel produz, entre outros produtos, sacos para hipermercados. Uma das experiências que o departamento técnico resolveu fazer foi ver o efeito que o factor concentração de madeira de carvalho na polpa tinha sobre a resistência do papel (medida em libras por polegada quadrada). Os níveis relevantes do factor concentração são $m = 4$ (5, 10, 15 e 20%) e para cada nível foram feitas $n_i = 6$ $(i = 1,2,3,4)$ observações ou réplicas conforme se indica no quadro 8.16.

Concentração de carvalho (%)	Observações					
	1	2	3	4	5	6
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

Tabela da ANOVA, Efeito do factor concentração sobre a resistência

Origem da variação	Soma de quadrados	Graus de liberdade	Médias quadráticas
Entre amostras	SQE = 382.7917	3	MQE = 127.5972
Dentro das amostras	SQD = 130.1667	20	MQD = 6.5083
Total	SQT = 512.9584	23	F = 19.6052

Tomando $\alpha = 0.05$ para dimensão do teste, obtém-se $F_{0.05} = 3.10$ para uma distribuição com 3 e 20 graus de liberdade ou, em alternativa, um valor-p de 0.000004.

Rejeitada a igualdade das médias podem construir-se os intervalos de confiança

$$t_{\alpha/2}\sqrt{\frac{\text{MQD}}{n_i}} = 2.086\sqrt{\frac{6.5083}{6}} = 2.17,$$

$$\bar{x}_{1\circ} = 10;\ \bar{x}_{2\circ} = 15.67;\ \bar{x}_{3\circ} = 17;\ \bar{x}_{4\circ} = 21.17,$$

$$\mu_1:\ [7.83; 12.17]; \quad \mu_2:\ [13.5; 17.84]; \quad \mu_3:\ [14.83; 19.17]; \quad \mu_4:\ [19; 23.34].$$