Análise de Variância Simples (One-Way ANOVA)

Pretende-se testar se determinado factor, quando aplicado a várias populações, tem um efeito significativo sobre determinada variável dependente, ou seja, se faz com que as médias populacionais da variável dependente sejam diferentes para os diferentes níveis do factor independente.

Consideremos k amostras independentes das populações $X_1, X_2, \ldots, X_k$ (ou k grupos de uma mesma população):

Amostra 1	Amostra 2	...	Amostra k
x_{11}	x_{12}	...	x_{1k}
x_{21}	x_{22}		x_{2k}
...	...		...
x_{n11}	x_{n22}		x_{nkk}

Sendo

x_{ij} – valor observado no i-ésimo elemento ($i = 1, \ldots, n_j$) da j-ésima amostra ($j = 1, \ldots, k$)

n_j – nº de elementos no grupo ou amostra j

Admitamos que as populações de onde se retiraram as amostras seguem distribuição normal com variâncias desconhecidas mas iguais, i.e., $X_j \cap N(\mu_j, \sigma)$ com $j = 1, \ldots, k$.

A Análise de Variância é um método estatístico que permite testar se existem diferenças entre 2 ou mais grupos de uma mesma população ou de populações diferentes. Assim, as hipóteses a testar são:

H0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H1: $\mu_r \neq \mu_j$, para algum par (r, j), $r \neq j$ $(r, j = 1, \ldots, k)$, isto é, existe pelo menos 2 grupos cujas médias sejam diferentes entre si.

Para rejeitar H0 basta que apenas 2 médias sejam diferentes.

Catarina Marques
ISCTE

A Análise de Variância envolve alguns **pressupostos**, pelo que só pode ser aplicada quando se verificam os seguintes requisitos:

a) As k populações têm distribuição normal;

b) As k populações têm a mesma variância;

c) As k amostras recolhidas são aleatórias e independentes entre si.

Embora o nome do método seja Análise de Variância, as hipóteses a testar, como vimos, respeitam às médias dos k grupos e não às variâncias. Estas últimas são utilizadas para definir a estatística de teste.

Para se encontrar esta estatística é necessário começar por decompor a variância total ou a variação total (soma total de quadrados) numa soma de 2 parcelas: a variação explicada pelo factor independente e a variação não explicada por esse factor (que é devida ao erro). Pretende-se, assim, **avaliar a importância de cada componente na variação total** e verificar se as diferenças encontradas entre as médias de cada grupo de observações são diferenças reais devidas a fontes controláveis de variação (factor independente) ou se se devem a fontes aleatórias e devem ser desprezadas.

Média amostral de cada grupo j $\quad \overline{X}_j = \dfrac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$

Média global – média de todos os valores observados não considerando a divisão por grupos

$$\overline{X} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \ldots + n_k\overline{x}_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{j=1}^{k} n_j\overline{x}_j}{\sum_{j=1}^{k} n_j} = \frac{\sum_{j=1}^{k} n_j\overline{x}_j}{n}, \qquad n = \sum_{j=1}^{k} n_j$$

Indicador da variabilidade dentro dos grupos (devido ao erro)

$$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{x}_j\right)^2$$

Sum of squares ***within*** *groups* (soma dos quadrados dentro dos grupos)

$$MSSW = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{x}_j\right)^2}{n-k}$$

Mean sum of squares ***within*** *groups* (média quadrática dentro dos grupos)

Catarina Marques
ISCTE

Indicador da variabilidade entre os grupos (devido ao factor independente)

$$SSB = (\overline{x}_1 - \overline{x})^2 n_1 + (\overline{x}_2 - \overline{x})^2 n_2 + \ldots + (\overline{x}_k - \overline{x})^2 n_k = \sum_{j=1}^{k} (\overline{x}_j - \overline{x})^2 n_j$$

Sum of squares ***between*** *groups* (soma dos quadrados entre grupos)

$$MSSB = \frac{\sum_{j=1}^{k} (\overline{x}_j - \overline{x})^2 n_j}{k-1}$$

Mean sum of squares ***between*** *groups* (média quadrática entre grupos)

Estatística de Teste

$$T = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}} = \frac{MSSB}{MSSW} \cap F_{(k-1;n-k)}$$

Para um dado α, rejeita-se H0 para valores da estatística de teste superiores ou igual ao quantil de probabilidade (1- α) da distribuição $F_{(k-1;n-k)}$. Isto porque só faz sentido rejeitar a hipótese de igualdade das k médias populacionais para valores elevados de T, valores esses que ocorrem quando a variação entre grupos (devido ao factor independente) for relativamente elevada quando comparada com a variação dentro dos grupos (devido ao erro).

É comum organizar os dados de uma Análise de Variância num quadro onde se apresenta as somas dos quadrados, o nº de graus de liberdade a elas associado e as médias quadráticas correspondentes a cada fonte de variação:

Fontes de Variação	**Soma dos quadrados**	**Graus de Liberdade**	**Médias quadráticas**	**Estatística de Teste**
Entre (explicada pelo factor indep.)	SSB	k - 1	$MSSB = \frac{SSB}{k-1}$	$T = \frac{MSSB}{MSSW} \cap F_{(k-1;n-k)}$
Dentro (devido ao erro	SSW	n - k	$MSSW = \frac{SSW}{n-k}$	
Total	SST = SSB+SSW	n - 1		

Catarina Marques
ISCTE

Como T representa $\frac{\text{variabilidade entre os grupos}}{\text{variabilidade dentro dos grupos}}$,

- Quanto maior for T, com maior probabilidade as médias são diferentes;
- Quanto menor for T, com maior probabilidade as médias são iguais;
- Quando as médias da população forem iguais, T tende para zero.

Se o valor do teste for grande, suspeita-se da veracidade de H0, porque é mais provável que para grandes valores do teste H0 seja falsa.

Catarina Marques
ISCTE

Consideremos três amostras, cujas distribuições estão representadas a vermelho, a azul e a verde. A média de cada amostra está identificada pela linha vertical e o Sw respectivo pela linha horizontal.
Consideremos as três amostras em conjunto, a média geral e a respectiva medida de dispersão SB (representadas a cinzento).

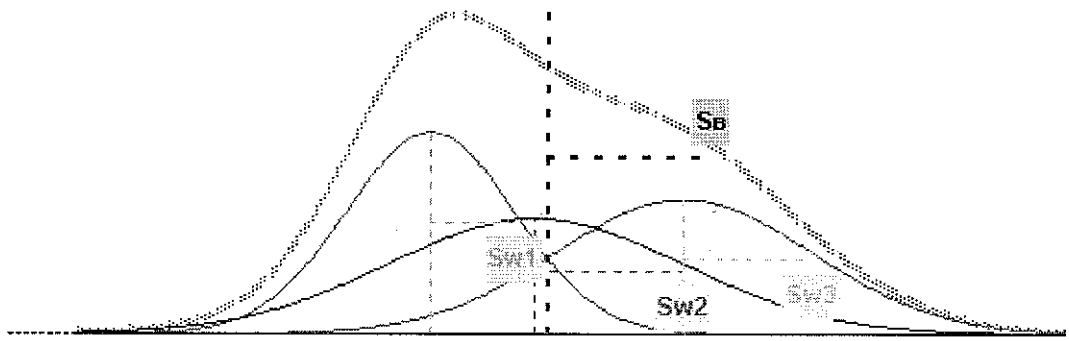

Fig.1

Analisemos alguns cenários:

I – Caso a dispersão dentro dos grupos (SW#) se mantenha mas as médias de cada amostra estejam mais afastadas entre si, aumenta a dispersão entre os grupos (SB).

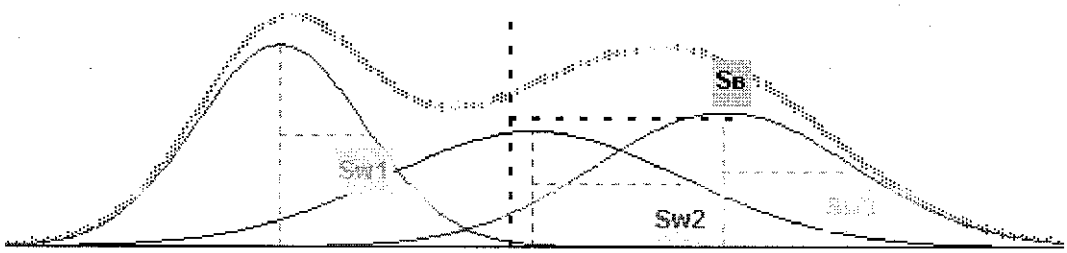

Fig.2

$$F = \frac{\text{variabilidade entre as médias dos grupos}}{\text{variabilidade dentro dos grupos}}$$

F aumenta, logo é maior a probabilidade de rejeitar Ho, isto é, de as médias serem diferentes.

Catarina Marques
ISCTE

II – Mantendo a média das amostras e diminuindo a dispersão dentro dos grupos (SW#),

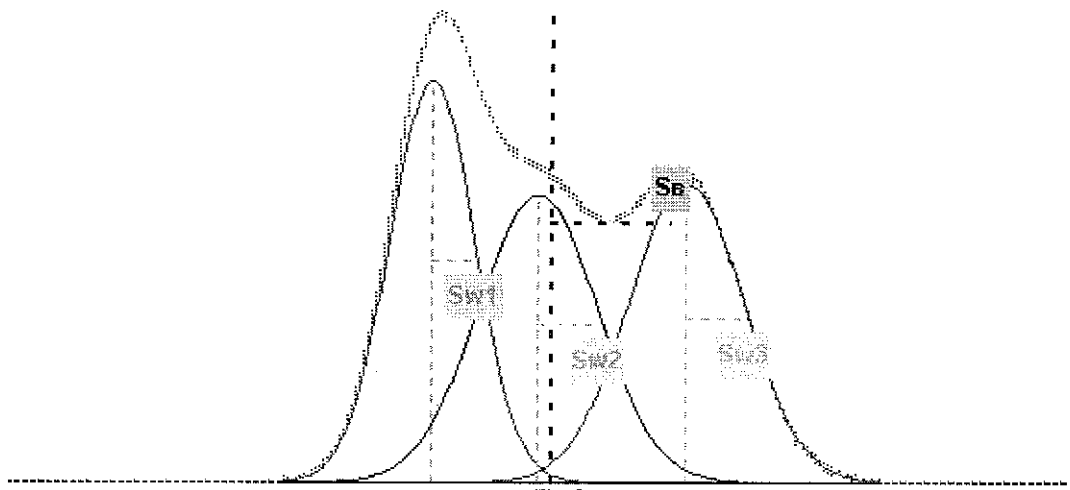

Fig.3

$$F = \frac{\text{variabilidade entre as médias dos grupos}}{\text{variabilidade dentro dos grupos}}$$

F aumenta, logo é maior a probabilidade de rejeitar Ho, isto é, de as médias serem diferentes.

III –

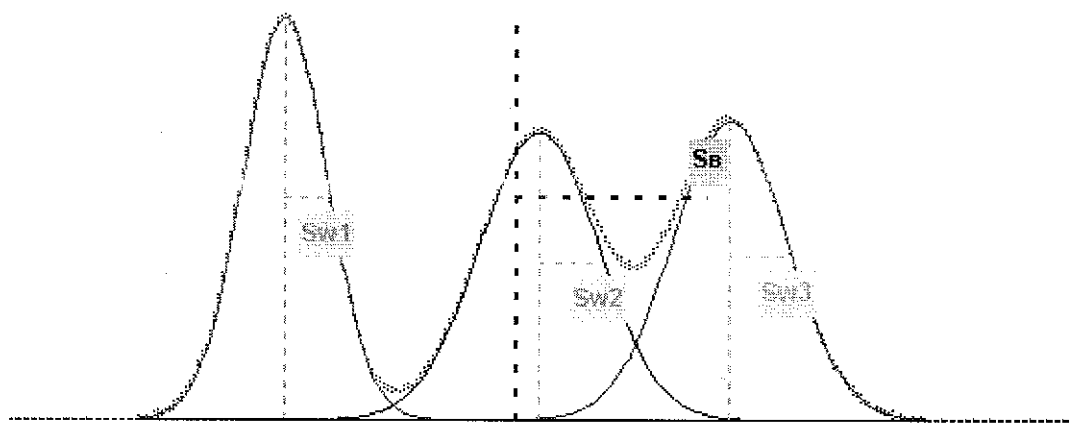

Fig 4

Comparemos as figuras 2, 3 e 4. O valor de F será maior nesta última?

Quanto menor a "sobreposição" entre as amostras, maior será o valor de F, logo mais provavelmente o teste indicará significância estatística.

Catarina Marques
ISCTE