ANÁLISE DE VARIÂNCIA

Celma de Oliveira Ribeiro/ Linda Lee Ho

Problema

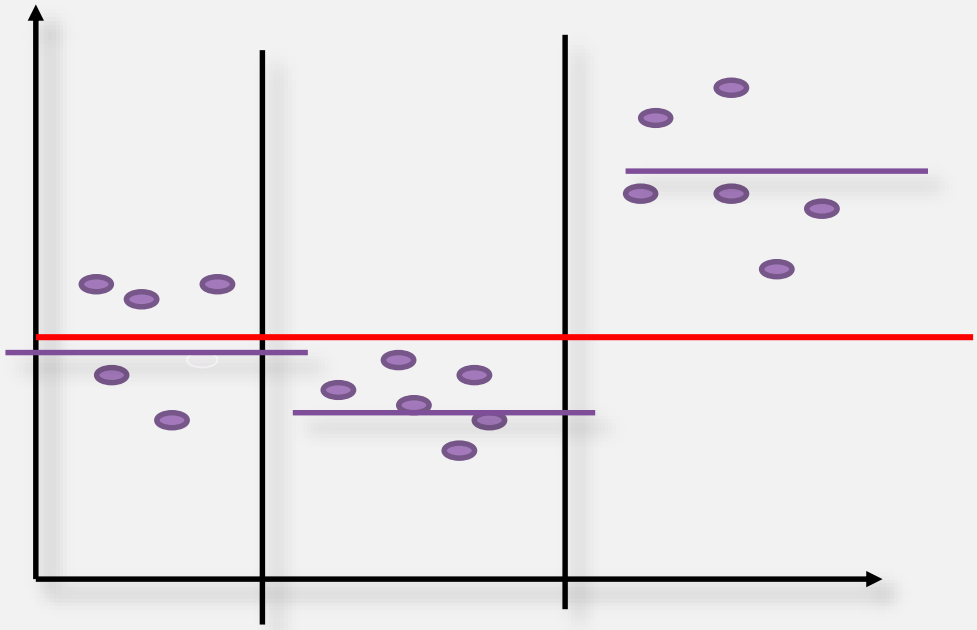
$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots \mu_k \\ H_1 : \exists (i, j) \mid \mu_i \neq \mu_j \end{cases}$$

K = 3 amostras

N = 100 observações (20 mec + 30 civ + 50 comp)

$$\overline{X} = 5{,}0$$

$$\overline{X}_{mec} = 5{,}2 \qquad \overline{X}_{civ} = 4{,}8 \qquad \overline{X}_{comp} = 4{,}9$$



Notação:

- K = número de amostras
- N = número total de observações
- $\overline{X}$ = média geral (todas observações)
- $\overline{X}_i$ = média da amostra i

Aplicação de análise de variâncias
Teste para igualdade de médias

Objetivo:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots \mu_k \\ H_1 : \exists (i, j) \mid \mu_i \neq \mu_j \end{cases}$$

- **Atenção: este teste não é equivalente a testar todos os pares de média, 2 a 2!**
- Ao rejeitar H0 estaremos concluindo que este fator (média) é importante para *explicar* a variável sendo analisada.
- **Hipótese:** homocedasticidade, ou seja, $\sigma_1^2 = \sigma_2^2 = \cdots \sigma_k^2 = \sigma^2$

- As variáveis X podem ser escritas como

$$X_{ij} = \mu_j + e_{ij}$$

onde eij é, por hipótese, um erro aleatório normal $N(0;\sigma^2)$ (note que estamos sob hipótese de homocedasticidade)

- **Idéia do teste:** Comparação de resíduos de modelos estatísticos (erros quadráticos).

S_i^2

a) Como cada um dos estima σ^2,

Variância da amostra i

$$S_P^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

também estima esta variância

(numerador é soma de quadrados)

$\bar{X}_i$ Cada um dos tem distribuição $N\left(\mu_i; \frac{\sigma^2}{n_i}\right)$

e se H0 é verdadeira, $\bar{X}_i = N\left(\mu; \frac{\sigma^2}{n_i}\right)$ (a mesma média).

Pode-se então verificar que $S_e^2 = \frac{\sum_{i=1}^{k} n_j (\bar{X}_j - \bar{X})^2}{k-1}$

é um estimador de σ^2 (numerador é soma de (k-1) quadrados)

- Assim, se vale H0, o quociente $\frac{S_e^2}{S_P^2}$

deve *estar próximo* de 1

- Teste de média pode ser substituído por

$$\begin{cases} H_0 : \sigma_e^2 = \sigma_P^2 \\ H_1 : \sigma_e^2 > \sigma_P^2 \end{cases}$$

- Sabe-se que

$$\frac{S_P^2 \times \left(\left(\sum_{i=1}^{k} n_i \right) - k \right)}{\sigma^2}$$ tem distribuição χ^2_{n-k}

e também $$\frac{S_e^2 \times (k-1)}{\sigma^2}$$ $\approx \chi^2_{k-1}$

Com isto $\frac{S_e^2}{S_P^2}$ tem distribuição F (k-1,n-k)

O teste fica:

- Da tabela F(k-1,n-k) obtenha Fc tal que $P(F \geq Fc) = \alpha$.
- Calcule, a partir da amostra o valor $F0 = \frac{S_e^2}{S_P^2}$

- Se F > Fc rejeite H0, ou seja, as médias são diferentes.

- Um consumidor pretende verificar se existe diferença no preço de aspirinas em diferentes cidades e em diferentes tipos de loja. Selecionou a seguinte amostra:

	Centro	Leste	Oeste	Sul
Drogaria	2,46	2,85	2,44	2,51
Farmácia	2,27	2,61	2,35	2,17
Supermercado	2,72	2,64	2,59	2,54

- Utilizando análise de variância de um fator é possível concluir que há diferença entre regiões? (5%)
- Utilizando análise de variância de um fator é possível concluir que há diferença entre regiões?

Tabela da ANOVA

Soma de quadrados

$$\sum_i \sum_j \left(X_{ij} - \bar{X}\right)^2 = \sum_i \sum_j \left(X_{ij} - \bar{X}_i\right)^2 + \sum_i \sum_j \left(\bar{X}_i - \bar{X}\right)^2$$

SQT	SQR	SQE
Total	**Residual**	**Entre amostras**

Tabela da ANOVA

Fonte de variação		Graus de Liberdade		F_0
Entre médias	SQE = $\sum_i \sum_j \left(\bar{X}_i - \bar{X}\right)^2$	K-1	$S_E^2 = \frac{SQE}{K-1}$	$\frac{S_E^2}{S_R^2}$
Residual	SQR= $\sum_i \sum_j \left(X_{ij} - \bar{X}_i\right)^2$	N-K	$S_R^2 = \frac{SQR}{N-K}$ (é o S^2p)	
Total	SQT= $\sum_i \sum_j \left(X_{ij} - \bar{X}\right)^2$	N-1		

Exercício

- Estudo sobre a presença de substâncias tóxicas no meio ambiente, associadas à utilização de um pesticida envolve análises efetuadas por quatro laboratórios diferentes.
- Há alguma preocupação de que os resultados destas análises diferem, devido à utilização de diferentes técnicas laboratoriais.
- Foram entregues a cada laboratório 6 pequenos contentores com solo recolhido aleatoriamente num terreno que antes fora tratado com o referido pesticida.
- Os resultados laboratoriais das análises químicas medem a concentração dum composto químico nocivo (em ppm).
- Os valores observados são indicados na tabela que se segue:

- Tabela

Laboratório	1	2	3	4
	53.2	51.0	47.4	51,0
	54.5	40,5	46,2	51,5
	52.8	50,8	46,0	48,8
	49.3	51,5	45,3	49,2
	50.4	52,4	48,2	48,3
	53.8	49,9	47,1	49,8
$\overline{X}_i$	52.33	49.35	46.70	49.77
S_i^2	4.151	19.480	1.120	1.587

- Qual a conclusão? (5% e 1%)

Quais médias são diferentes?

Teste de Schefee

$$\Delta_{ij} = \sqrt{S_p^2 \times F_c \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right) \times (K-1)}$$

Se $\left|X_i - X_j\right| > \Delta_{ij}$ as médias μ_i e μ_j são diferentes

Teste de homogeneidade de variância.

- Para testar se $\sigma_1^2 = \sigma_2^2 = \cdots \sigma_k^2 = \sigma^2$ pode-se utilizar a seguinte estatística:

$$D = \sum_{i=1}^{k} (n_i - 1) \ln\left(\frac{s_R^2}{s_i^2}\right)$$

- onde
  - k o número de classes,
  - ni o tamanho da amostra i
  - S_i^2 a variância amostral da i-ésima amostra
  - S_R^2 a variância residual ,

$$s_R^2 = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{\sum_{i=1}^{k} (n_i - 1)}$$

- Sabendo que a variável D tem distribuição χ^2 com K-1 graus de liberdade, verifique ao nível de 5% a hipótese de equivariância para o conjunto de dados abaixo

- **Exercício 3**
- **Um estudo sobre café robusta em Angola analisa se frequência da defeitos no grão de café para seis diferentes regiões.**
- **Em cada região foram escolhidos aleatoriamente 11 lotes.**
- **Na tabela são apresentados as médias e desvios padrão da percentagem do peso total de grãos sem defeito, assim como as variâncias e log variâncias**

Assim

Laboratório	Média	Desvio	Variância	Ln(var)
Cabinda	44,19	6,94	48.1636	3.8746
Ambriz	58,87	18,98	360.2404	5.8868
Encoje	51,80	13,92	193.7664	5.2667
Cazongo	60,61	13,29	176.6241	5.1740
Libolo	61,96	24,49	599.7601	6.3965
Amboim	42,11	21,31	454.1161	6.1184

a) Construa o quadro-resumo associado a esta ANOVA
b) Determine a variância da totalidade das n = 66 observações.
c) Teste a hipótese de a percentagem de grão sem defeito não depender da região de proveniência do grão. Utilize os níveis de significância 0.05 e 0.01 e comente. Calcule o valor p da estatística calculada.
d) Independentemente do resultado da alínea anterior, verifique quais as regiões cujas médias de grãos com defeito devem ser consideradas diferentes, e quais aquelas em que se pode admitir a igualdade, utilizando um teste de Schefeé, com um nível de significância 0.05.
e) Teste a validade da hipótese de homogeneidade de variâncias exigida no modelo que indicou na primeira alínea. Comente o resultado obtido