Introdução

Exemplos

- Para curar uma certa doença existem quatro tratamentos possíveis: A, B, C e D. Pretende-se saber se existem diferenças significativas nos tratamentos no que diz respeito ao tempo necessário para eliminar a doença.

- Comparar três lojas quanto ao volume médio de vendas.

- ...

Existem ***k*** populações de interesse, nas quais se estuda uma característica comum.

Sejam X_1, X_2, ..., X_k as variáveis aleatórias que representam tal característica nas populações **1**, **2**, ...,***k***, respectivamente.

Hipóteses a testar:

H_0: $\mu_1 = \mu_2 = ... = \mu_k$

H_1: $\mu_i \neq \mu_j$ para algum i e algum j tais que $i \neq j$.

As ***k* populações** podem ser vistas como ***k* níveis de um mesmo factor**.

A questão é saber se **o factor exerce alguma influência na variação da característica em estudo**.

Exemplo

Para curar uma certa doença existem quatro tratamentos possíveis: A, B, C e D.

Pretende-se saber se existem diferenças significativas nos tratamentos no que diz respeito ao tempo necessário para eliminar a doença.

Temos **apenas um factor**, **Tratamento**, que se apresenta em quatro níveis, A, B, C e D.

Através da aplicação da ***análise de variância com um factor*** ou "***one-way ANOVA***", podemos indagar se os tratamentos produzem os mesmos resultados no que diz respeito à característica em estudo.

Exemplo

Suponhamos agora que existe a suspeita de que uma estação quente é um factor determinante para uma cura rápida.

Então, o estudo deve ser conduzido tendo em conta este **segundo factor**, **Estação do Ano**.

Aqui, a técnica estatística apropriada será a ***análise de variância com dois factores***, também designada por "***two-way ANOVA***".

Neste caso, pode-se testar se existe diferença entre os tratamentos e também se existe diferença entre as estações do ano, no que respeita ao tempo de tratamento até à eliminação da doença.

Análise de Variância com Um Factor

Exemplo 1

O Sr. Fernando Estradas é dono de várias lojas que vendem todo o tipo de material para desportos radicais. Para uma determinada loja foram recolhidas três amostras aleatórias e independentes das vendas semanais (em u.m.); cada uma destas amostras constituída por cinco observações (vendas em 5 semanas, n=5).

Dados recolhidos:

	Amostra 1	Amostra 2	Amostra 3	
	49	52	55	
	55	51	51	
	51	55	52	
	52	58	52	
	48	49	50	
$\overline{X}$	51	53	52	→ 3 valores observados da v. a. $\overline{X}$

Exemplo 1

Naturalmente, obtivemos nas três amostras volumes de vendas médios diferentes, o que se deve, como sabemos, às **flutuações amostrais**.

A variação de $\overline{X}$, de amostra para amostra, pode ser medida pela sua variância:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}.$$

Em geral,

desconhece-se o valor de $\sigma_X \Rightarrow$ desconhece-se o valor de $\sigma^2_{\overline{X}}$

Mas, podemos obter uma estimativa deste parâmetro.

Exemplo 1

Calculamos a média dos valores observados de $\overline{X}$ – a **média das médias amostrais**:

$$\overline{\overline{x}} = \frac{51+53+52}{3} = 52 \qquad \text{(estimativa)}$$

Usámos o **estimador**: $\overline{\overline{X}} = \frac{1}{k}\sum_{i=1}^{k} \overline{X}_i$ (onde ***k*** é o número de amostras)

Finalmente, estimamos a variância de $\overline{X}$ por:

$$s^2_{\overline{X}} = \frac{1}{3-1}\left[(51-52)^2 + (53-52)^2 + (52-52)^2\right] = \frac{1}{2}(1+1+0) = 1 \quad \text{(estimativa)}$$

Usámos o **estimador**: $S^2_{\overline{X}} = \frac{1}{k-1}\sum_{i=1}^{k}\left(\overline{X}_i - \overline{\overline{X}}\right)^2$

Exemplo 2

Suponhamos agora, que o Sr Fernando Estradas pretende comparar **três** lojas quanto ao volume de vendas.

Para isso, para cada loja, ele selecciona aleatoriamente cinco semanas, onde observa o volume de vendas. Obtém assim uma amostra das vendas semanais para cada loja (as três amostras são independentes). Os dados estão registados na tabela seguinte.

	Loja 1	Loja 2	Loja 3	
	47	55	54	
	53	54	50	
	49	58	51	
	50	61	51	
	46	52	49	
$\overline{X}_i$ (médias amostrais)	$\overline{x}_1 = 49$	$\overline{x}_2 = 56$	$\overline{x}_3 = 51$	$\overline{\overline{x}} = 52$
$\left(\overline{X}_i - \overline{\overline{X}}\right)^2$	9	16	1	$\sum(\overline{x}_i - \overline{\overline{x}})^2 = 26$

Exemplo 2

Representemos por X_i o volume de vendas numa semana na loja i (i = 1,2,3) e por μ_i o valor médio de X_i.

Este exemplo tem apenas um factor de interesse, **o factor Loja**, e este apresenta três níveis ou grupos: **Loja 1**, **Loja 2** e **Loja 3**.

Cada nível do factor define uma população de média μ_i.

Pretende-se saber se as médias dos três níveis, ou populações, são iguais, isto é, pretende-se saber se é de rejeitar ou não a hipótese

H_0: $\mu_1=\mu_2=\mu_3$ (igualdade de vendas médias das três lojas).

Exemplo 2

Questão:

Serão as médias amostrais $\bar{x}_1$=49, $\bar{x}_2$=56 e $\bar{x}_3$=51 diferentes porque há diferenças entre as médias populacionais μ_1, μ_2 e μ_3?

Ou serão essas diferenças razoavelmente atribuídas a flutuações amostrais?

Podemos então formular as seguintes hipóteses:

***H*₀**: $\mu_1=\mu_2=\mu_3$ (não há diferença entre o volume médio de vendas das 3 lojas)

***H*₁**: $\mu_i \neq \mu_j$ para algum i e algum j tais que $i \neq j$ (há pelo menos duas lojas com diferentes volumes médios de vendas)

Não seria possível resolver a questão conduzindo três testes de hipóteses, cada um comparando duas médias populacionais, utilizando as técnicas vistas no capítulo anterior?

Suponhamos que, de facto, as vendas médias das três lojas são iguais, isto é $\mu_1=\mu_2=\mu_3$.

Admitindo a independência entre os três testes e fixando para cada teste um nível de significância de 0.05, o **nível de significância para o conjunto dos três testes**, isto é, a probabilidade de decidirmos erradamente que as três médias não são iguais quando de facto o são, seria aproximadamente 0.1426.

Pensemos nos 3 testes de hipóteses como 3 provas de Bernoulli.

sucesso ≡ "tomar a decisão errada de rejeitar H_0"

W ≡ "nº de decisões erradas (sucessos) nos três testes de hipóteses"

$W \sim B(3, 0.05)$

A probabilidade de concluirmos erradamente que as 3 médias não são iguais, é igual a

$$P(W \geq 1) = 1 - P(W = 0) = \binom{3}{0} 0.05^0 \ 0.95^3 = 0.1426.$$

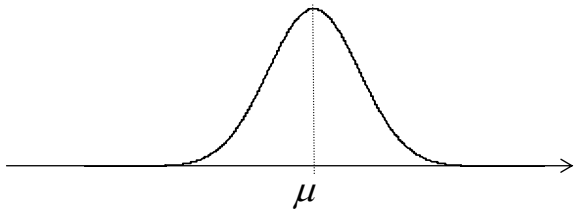
A aplicação da análise de variância pressupõe a verificação das seguintes **condições**:

1. As amostras devem ser aleatórias e independentes.
2. As amostras devem ser extraídas de populações normais.
3. As populações devem ter variâncias iguais ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$).

Temos então duas situações possíveis:

➢ **H_0 é verdadeira** – as diferenças observadas entre as médias amostrais são devidas a flutuações amostrais.

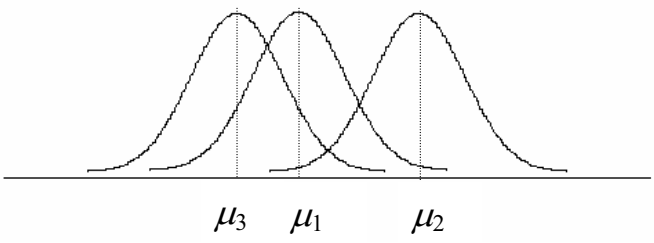
$\mu_1=\mu_2=\mu_3=\mu \Rightarrow$ todas as amostras provêm de populações com médias iguais. Como se supôs que todas as populações são normais e têm variâncias iguais, isto é o mesmo que extrair todas as amostras de uma única população (de uma única loja – como no Exemplo 1).



Distribuições populacionais quando H_0 é verdadeira ($\mu_1=\mu_2=\mu_3=\mu$).

➢ **H_0 é falsa** – as diferenças observadas entre as médias amostrais são demasiado grandes para serem devidas unicamente a flutuações amostrais.

As médias das populações não são iguais, ou seja pelo menos duas lojas têm volumes de vendas médios diferentes. As amostras recolhidas provêm de populações diferentes.



Distribuições populacionais quando H_0 é falsa (as médias não são todas iguais).

Note que é suposto que $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$.

A análise de variância vai estimar σ^2 por dois processos diferentes e comparar os valores obtidos.

1º PROCESSO – ***Estimativa "dentro" da variância***: s_p^2

Como todas as amostras são extraídas de populações com a mesma variância σ^2, então, para estimar este parâmetro, poderíamos utilizar qualquer uma das amostras. Assim, poderíamos obter *k* estimativas de σ^2, uma por cada amostra.

Exemplo 2

Temos as seguintes estimativas de σ^2:

$$s_1^2 = \frac{1}{5-1}\left[(47-49)^2 + (53-49)^2 + (49-49)^2 + (50-49)^2 + (46-49)^2\right] = 7.5$$

$$s_2^2 = \frac{1}{5-1}\left[(55-56)^2 + (54-56)^2 + (58-56)^2 + (61-56)^2 + (52-56)^2\right] = 12.5$$

$$s_3^2 = \frac{1}{5-1}\left[(54-51)^2 + (50-51)^2 + (51-51)^2 + (51-51)^2 + (49-51)^2\right] = 3.5.$$

Tomando a média destas estimativas obtemos outra estimativa para σ^2,

$$s_p^2 = \frac{s_1^2 + s_2^2 + s_3^2}{3} = 7.83.$$

O que fizemos foi combinar as três estimativas anteriores, de modo a produzir uma outra estimativa que use a informação contida nas três amostras recolhidas.

A fórmula geral para o cálculo da estimativa "dentro" da variância é:

$$s_p^2 = \frac{s_1^2 + s_2^2 + \cdots + s_k^2}{k}$$

onde,

s_i^2 → variância amostral da amostra i.

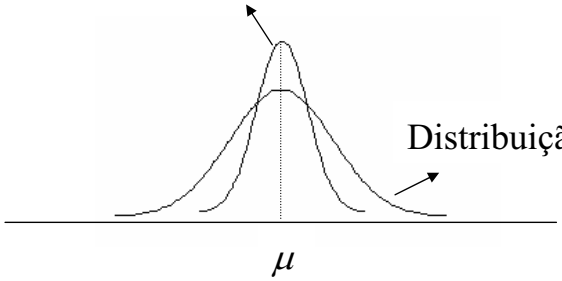
Note que esta estimativa não é afectada pela veracidade ou falsidade de H_0, o que já não acontece com a que iremos obter pelo processo seguinte.

2º PROCESSO – ***Estimativa "entre" da variância***: s_b^2

Já vimos anteriormente, que se ***H_0* é verdadeira** podemos encarar as três amostras como sendo provenientes da mesma população (X) (da mesma loja, como no Exemplo 1).

Admitindo que H_0 é verdadeira ($\mu_1=\mu_2=\mu_3=\mu$)

Distribuição da **média amostral**: $\bar{X} \sim N(\mu, \sigma^2/n)$



Distribuição populacional: $X \sim N(\mu, \sigma^2)$

Os valores médios observados nas três amostras, $\bar{x}_1$, $\bar{x}_2$ e $\bar{x}_3$, podem ser encarados como três valores observados de uma v. a. $\bar{X} \sim N(\mu, \sigma^2/n)$.

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \quad \Leftrightarrow \quad \sigma^2 = n.\sigma^2_{\bar{X}},$$

sugerindo que se estime σ^2 através de

$$s^2_b = n.s^2_{\bar{X}},$$

com

$$s^2_{\bar{X}} = \frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{x}_i - \bar{\bar{x}}\right)^2 \rightarrow \text{estimativa de } \sigma^2_{\bar{X}}.$$

Se H_0 for falsa

Pelo menos duas distribuições populacionais são diferentes. Isto é, as variáveis aleatórias X_i têm distribuições normais, com iguais variâncias, mas, pelo menos duas, têm médias diferentes.

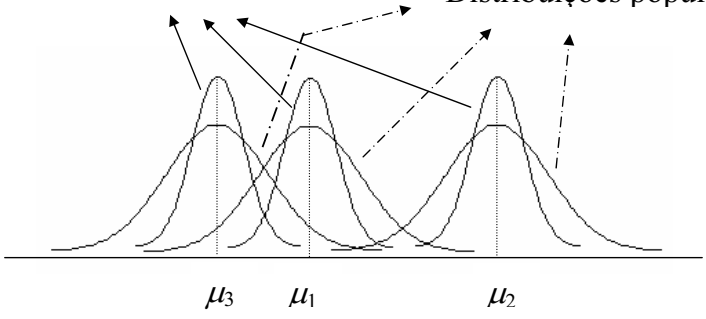
Então também $\bar{X}_1$, $\bar{X}_2$ e $\bar{X}_3$, vão ter distribuições diferentes:

$$\bar{X}_1 \sim N(\mu_1, \sigma^2/n) \quad , \quad \bar{X}_2 \sim N(\mu_2, \sigma^2/n) \quad \text{e} \quad \bar{X}_3 \sim N(\mu_3, \sigma^2/n),$$

onde $\mu_1 \neq \mu_2$ ou $\mu_1 \neq \mu_3$ ou $\mu_2 \neq \mu_3$.

Distribuições da **média amostral**

Distribuições populacionais



Assim, $\bar{x}_1$, $\bar{x}_2$ e $\bar{x}_3$ são valores observados de variáveis aleatórias com distribuições diferentes, o que se vai reflectir, eventualmente, numa maior dispersão desses valores, conduzindo a um maior valor de $s^2_{\bar{X}}$ e consequentemente a um maior valor de $s_b^2 = n.s^2_{\bar{X}}$.

Exemplo 2:

$$s^2_{\bar{X}} = \frac{1}{3-1}\left[(49-52)^2 + (56-52)^2 + (51-52)^2\right] = \frac{26}{2} = 13$$

logo a estimativa "entre" da variância é:

$$s_b^2 = n.s^2_{\bar{X}} = 5 \times 13 = 65.$$

Estatística de teste – F

A estimativa "dentro" da variância, s_p^2, não é afectada pela veracidade ou falsidade de H_0.

Ao contrário, a estimativa "entre" da variância, s_b^2, já o é, sendo aproximadamente igual a s_p^2 quando H_0 é verdadeira e maior do que esta se H_0 é falsa.

A estatística de teste é,

$$F = \frac{n.S^2_{\bar{X}}}{S_p^2} = \frac{S_b^2}{S_p^2}.$$

Se **H_0 é verdadeira**, σ^2 pode ser estimada pelos dois processos e como as duas estimativas serão aproximadamente iguais, a razão F será próxima de 1.

Se **H_0 for falsa**, as diferenças nas médias populacionais μ_1, μ_2 e μ_3 vão provocar maior variabilidade nas médias amostrais. Isto é, $s^2_{\bar{X}}$ será grande e consequentemente s^2_b será também grande comparativamente com s^2_p. A razão F tomará um valor maior que 1.

Sob o pressuposto de H_0 ser verdadeira, tem-se

$$F = \frac{n.S^2_{\bar{X}}}{S^2_p} = \frac{S^2_b}{S^2_p} \sim F^{k-1}_{k(n-1)}.$$

H_0 deve ser rejeitada se o valor observado de F se situar à direita do ponto crítico.

Isto é, rejeita-se H_0 se,

$$F_{\text{obs}} \geq p_c$$

onde, o ponto crítico p_c é dado por

$$P\left(F^{k-1}_{k(n-1)} \geq p_c\right) = \alpha = \text{nível de significância.}$$

O ponto crítico p_c é o **quantil de probabilidade 1-α da distribuição** $F^{k-1}_{k(n-1)}$ e é usualmente denotado por $F_{(1-\alpha)}$ ou por $F_{1-\alpha,k-1,k(n-1)}$.

Exemplo 2

Vamos ver o que podemos concluir ao nível de significância de 0.05.

Se a hipótese H_0 é verdadeira,

$$F = \frac{S_b^2}{S_p^2} \sim F_{12}^2.$$

$F_{1-\alpha,2,12} = 3.89$ (quantil de probabilidade 1-α da distribuição F_{12}^2)

R.C.=[3.89,+∞[

O valor observado da estatística F é: $F_{obs} = \frac{65}{7.83} = 8.3 \in$ R.C.

Então a hipótese H_0 é rejeitada ao nível de significância de 0.05, isto é, existem diferenças significativas entre as médias amostrais das vendas. Há portanto evidência de que existem pelo menos duas lojas com volumes médios de vendas diferentes. Por outras palavras, o factor Loja exerce uma influência significativa sobre o volume de vendas.

Tabela de análise de variância (ANOVA)

Os dados, usualmente, vêm representados da seguinte maneira:

	Amostra (j)					
	1	**2**	**3**	**...**	**k**	
Observações (i)	x_{11}	x_{12}	x_{13}	…	x_{1k}	
	x_{21}	x_{22}	x_{23}	…	x_{2k}	
	x_{31}	x_{32}	x_{33}	…	x_{3k}	
	⋮	⋮	⋮	⋱	⋮	
	x_{n1}	x_{n2}	x_{n3}	…	x_{nk}	
Médias amostrais	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	…	$\bar{x}_k$	$\bar{\bar{x}}$

Os cálculos para a análise de variância podem ser sumariados numa tabela chamada **Tabela ANOVA**:

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Variância (Soma Média de Quadrados)	Razão F
Entre grupos	$SS_A = n\sum_{j=1}^{k}\left(\bar{x}_j - \bar{\bar{x}}\right)^2$	k-1	$S_b^2 = MS_A = \frac{SS_A}{k-1}$	$F = \frac{S_b^2}{S_p^2} = \frac{MS_A}{MS_E}$
Dentro dos grupos ou residual	$SS_E = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$	k(n-1)	$S_p^2 = MS_E = \frac{SS_E}{k(n-1)}$	
Total	$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \bar{\bar{x}}\right)^2$	nk-1		

Note que:

$$s_p^2 = \frac{s_1^2 + s_2^2 + \cdots + s_k^2}{k} = \frac{\frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}{n-1} + \frac{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}{n-1} + \cdots + \frac{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}{n-1}}{k}$$

$$= \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}{k(n-1)} = \frac{SS_E}{k(n-1)} = MS_E$$

e,

$$s_b^2 = n.s_{\bar{X}}^2 = n \times \frac{\sum_{j=1}^{k}\left(\bar{x}_j - \bar{\bar{x}}\right)^2}{k-1} = \frac{SS_A}{k-1} = MS_A$$

$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ij} - \bar{\bar{x}})^2$ → é a **soma de quadrados total** e mede a variação total nos dados;

$SS_A = n\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2$ → é a **soma de quadrados entre os níveis**, ou **grupos**, **do factor** e mede a variação entre grupos (populações); é por vezes designada por "**variação explicada**", pois ela é explicada pelo facto de as amostras poderem provir de populações diferentes;

$SS_E = \sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ → é a **soma de quadrados dentro dos níveis**, ou **grupos**, **do factor** e mede a variação dentro dos grupos (populações); é por vezes designada por "**variação não explicada ou residual**", pois é atribuída a flutuações dentro do mesma população, portanto não pode ser explicada pelas possíveis diferenças entre os grupos (populações).

Pode-se provar que:

$$SS_T = SS_A + SS_E$$

o que permite verificar os cálculos da Tabela ANOVA.

Apresentamos a seguir a **Tabela ANOVA** relativa ao **Exemplo 2**.

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Variância (Soma Média de Quadrados)	Razão F
Entre grupos	SS_A=130	2	$MS_A = s_b^2 = 65$	8.3
Dentro dos grupos ou residual	SS_E=94	12	$MS_E = s_p^2 = 7.83$	
Total	SS_T=224	14		

Amostras de Tamanhos Diferentes

Se as amostras têm tamanhos diferentes, as fórmulas apresentadas anteriormente devem ser convenientemente modificadas.

- n_j – nº de observações na amostra j
- $\boldsymbol{k}$ – nº de amostras
- $N = \sum_{j=1}^{k} n_j$ (total de observações)
- $\bar{x}_j$ – média observada na amostra j
- $\bar{\bar{x}} = \dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_1 + n_2 + \cdots + n_k} = \dfrac{\sum_{j=1}^{k} n_j \bar{x}_j}{n_1 + n_2 + \cdots + n_k}$ – média ponderada das médias amostrais

Soma média de quadrados entre grupos

$$\mathbf{MS_A} = \frac{\sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2}{k-1} = \frac{SS_A}{k-1}.$$

Soma média de quadrados dentro dos grupos ou residual

$$\mathrm{MS_E} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2}{n_1 + n_2 + \cdots + n_k - k} = \frac{SS_E}{N-k}.$$

A **Tabela ANOVA** para amostras de tamanhos diferentes.

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Variância (Soma Média de Quadrados)	Razão F
Entre grupos	$\mathrm{SS_A} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2$	k-1	$MS_A = \dfrac{SS_A}{k-1}$	$\mathrm{F} = \dfrac{S_b^2}{S_p^2} = \dfrac{MS_A}{MS_E}$
Dentro dos grupos ou residual	$\mathrm{SS_E} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2$	$N-k$	$MS_E = \dfrac{SS_E}{N-k}$	
Total	$\mathrm{SS_T} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(x_{ij} - \bar{\bar{x}}\right)^2$	$N-1$		

Exemplo 3

Suponha que é director de marketing de uma empresa que pretende relançar um produto no mercado. Você estudou três campanhas de marketing diferentes, cada uma deles combina de modo diferente factores como o preço do produto, a apresentação do produto, promoções associadas, etc. Qualquer uma destas campanhas é levada a cabo no ponto de venda, não havendo qualquer publicidade nos meios de comunicação. Para saber se há diferença entre as três campanhas relativamente à sua eficácia, cada uma delas é feita num conjunto de lojas seleccionadas aleatoriamente, durante um período de duração limitada. Note que as lojas são seleccionadas de modo a que as três amostras sejam aleatórias e independentes entre si. As vendas (em unidades monetárias – u. m.) registadas durante este período constam da tabela seguinte.

	Campanha 1	Campanha 2	Campanha 3
	8	10	7
	6	8	5
	5	12	8
	6	7	6
	7	9	7
		10	5
		11	
Soma	32	67	38

Seja X_i a v.a. que representa o volume de vendas de uma loja sujeita à campanha i (i=1,2 ou 3).

Admitamos que X_1, X_2 e X_3 têm distribuição normal com iguais variâncias.

As hipóteses em teste são:

H_0: $\mu_1=\mu_2=\mu_3$

(não há diferença entre as campanhas de marketing relativamente ao volume médio de vendas a que conduzem)

H_1: $\mu_i \neq \mu_j$ para algum i e algum j tais que $i\neq j$

(pelo menos duas campanhas de marketing conduziram a volumes médios de vendas diferentes)

Fixemos o nível de significância em 0.01.

Sob o pressuposto de H_0 ser verdadeira,

$$F = \frac{MS_A}{MS_E} \sim F^2_{15}.$$

$F_{1-\alpha,2,15} = 6.36$ (quantil de probabilidade 1-α=0.99 da distribuição F^2_{15})

R.C.=[6.36,+∞[

Para as amostras recolhidas, tem-se:

- $\bar{x}_1 = 6.4$, $\bar{x}_2 = 9.5714$, $\bar{x}_3 = 6.3333$ e $\bar{\bar{x}} = 7.611$;
- $SS_A = 44.03$ e $MS_A = \frac{44.03}{2} = 22.015$;
- $SS_E = 30.2476$ e $MS_E = \frac{30.2476}{15} = 2.0165$.

O valor observado da estatística F é: $F_{obs} = \frac{22.015}{2.0165} = 10.9174 \in$R.C.

Ao nível de significância de 0.01, rejeita-se a hipótese H_0 de igualdade de médias, pois o valor observado da estatística de teste pertence à região crítica. Há, portanto, evidência estatística de que as três campanhas não são iguais relativamente ao volume médio de vendas a que conduzem. Isto é, **o tipo de campanha influencia significativamente o volume de vendas**.

A **Tabela ANOVA** para este exemplo é a seguinte.

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Variância (Soma Média de Quadrados)	Razão F
Entre grupos	SS_A=44.03	2	MS_A=22.015	10.9174
Dentro dos grupos ou residual	SS_E=30.247	15	MS_E=2.0165	
Total	SS_T=74.277	19		

Testes de Comparação Múltipla

Quando a aplicação da análise de variância conduz à rejeição da hipótese nula, temos evidência de que existem diferenças entre as médias populacionais. Mas, **entre que médias se registam essas diferenças?**

Os testes de comparação múltipla permitem responder à questão anterior, isto é, permitem investigar onde se encontram as diferenças possíveis entre *k* médias populacionais.

Existem muitos testes deste tipo, no entanto, aqui vamos abordar apenas dois:

- **teste HSD** (honestly significant difference) **de Tuckey**
- **teste de Scheffé**

Estes testes permitem examinar simultaneamente pares de médias amostrais para identificar quais os pares onde se registam diferenças significativas.

Pressupostos:

1. As amostras devem ser aleatórias e independentes.
2. As amostras devem ser extraídas de populações normais.
3. As populações devem ter variâncias iguais ($\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$).

Notação:

- $\mathrm{N} = \sum_{j=1}^{k} n_j$
- $\mathrm{MS_E} = \dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2}{n_1 + n_2 + \cdots + n_k - k} = \dfrac{SS_E}{N-k}$

Teste HSD de Tuckey

Quando **as amostras têm tamanhos iguais** este teste é mais adequado do que o teste de Scheffé.

O teste HSD de Tuckey foi originalmente desenvolvido para amostras de igual tamanho, no entanto, muitos estatísticos sustentam que este é um método robusto a desvios **moderados** deste pressuposto.

Neste teste, duas médias amostrais são comparadas usando

$$S_{T(1-\alpha)} \cdot \sqrt{\frac{MS_E}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

onde, $S_{T(1-\alpha)}$ é o quantil de probabilidade (1-α) da distribuição da ***"Studentized Range"*** com $(\mathrm{k}, \mathrm{N}-\mathrm{k})$ graus de liberdade – $\mathrm{S}_{\mathrm{T}(\mathrm{k}, N\text{-}\mathrm{k})}$:

$$P\left(W \le S_{T(1-\alpha)}\right) = 1-\alpha \ , \ \mathrm{W} \sim \mathrm{S}_{\mathrm{T}(\mathrm{k}, N\text{-}\mathrm{k})}.$$

A hipótese H_0: $\mu_i = \mu_j$ é rejeitada, isto é, as médias amostrais $\bar{x}_i$ e $\bar{x}_j$ são consideradas significativamente diferentes, se

$$\left|\bar{x}_i - \bar{x}_j\right| \geq S_{T(1-\alpha)} \cdot \sqrt{\frac{MS_E}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Exemplo 2

- $\left|\bar{x}_1 - \bar{x}_2\right| = \left|49 - 56\right| = 7$,
- $\left|\bar{x}_1 - \bar{x}_3\right| = \left|49 - 51\right| = 2$
- $\left|\bar{x}_2 - \bar{x}_3\right| = \left|56 - 51\right| = 5$

Usando um nível de significância igual a 0.05, vem:

$$S_{T(1-\alpha)}=3.77$$

$$S_{T(1-\alpha)} \cdot \sqrt{\frac{MS_E}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = 3.77\sqrt{\frac{7.83}{2} \times \frac{2}{5}} = 4.718$$

Como $\left|\bar{x}_1 - \bar{x}_2\right| = 7 > 4.718$, rejeita-se a hipótese H_0:$\mu_1 = \mu_2$.

Também, $\left|\bar{x}_2 - \bar{x}_3\right| = 5 > 4.718$, logo rejeita-se a hipótese H_0:$\mu_2 = \mu_3$.

Finalmente, como $\left|\bar{x}_1 - \bar{x}_3\right| = 2 < 4.718$, não se rejeita a hipótese H_0:$\mu_1 = \mu_3$.

Assim, há evidência de que a loja 2 tem um volume médio de vendas diferente das lojas 1 e 3. Isto é, a média observada para a loja 2 difere significativamente das médias observadas para as lojas 1 e 3, enquanto que, a diferença registada entre o volume de vendas da loja 1 e da loja 3 não é significativa.

Teste Scheffé

Neste teste a hipótese nula H_0: $\mu_i = \mu_j$ é rejeitada se

$$\left|\bar{x}_i - \bar{x}_j\right| \geq \sqrt{(\mathrm{k}-1)\mathrm{F}_{(1-\alpha)}} \cdot \sqrt{MS_E\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

onde, $F_{(1-\alpha)}$ é o quantil de probabilidade (1-α) da distribuição F_{N-k}^{k-1}:

$$P\left(F_{N-k}^{k-1} \leq F_{(1-\alpha)}\right) = 1-\alpha$$

Exemplo 3

- $\left|\bar{x}_1 - \bar{x}_2\right| = \left|6.4 - 9.5714\right| = 3.1714$
- $\left|\bar{x}_1 - \bar{x}_3\right| = \left|6.4 - 6.3333\right| = 0.0667$
- $\left|\bar{x}_2 - \bar{x}_3\right| = \left|9.5714 - 6.3333\right| = 3.2318$

Consideremos um nível de significância igual a 0.01.

- $\left|\bar{x}_1 - \bar{x}_2\right| = 3.1714 > \sqrt{(k-1)F_{(1-\alpha)}} \cdot \sqrt{MS_E\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

$= \sqrt{2 \times 6.36} \cdot \sqrt{2.0165\left(\frac{1}{5}+\frac{1}{7}\right)} = 2.97, \rightarrow$ rejeita-se a hipótese H_0: $\mu_1 = \mu_2$;

- $\left|\bar{x}_1 - \bar{x}_3\right| = 0.0667 < \sqrt{2 \times 6.36} \cdot \sqrt{2.0165\left(\frac{1}{5}+\frac{1}{6}\right)} = 3.0667 \rightarrow$ não se rejeita H_0: $\mu_1 = \mu_3$;

- $\left|\bar{x}_2 - \bar{x}_3\right| = 3.2318 > \sqrt{2 \times 6.36} \cdot \sqrt{2.0165\left(\frac{1}{6}+\frac{1}{7}\right)} = 2.8177, \rightarrow$ rejeita-se H_0: $\mu_2 = \mu_3$.

Assim, ao nível de significância de 0.01, há evidência de que à campanha de marketing 2 está associado um volume médio de vendas diferente dos volumes médios associados às campanhas 1 e 3. Isto é, a média observada para a campanha 2 difere significativamente das médias observadas para as campanhas 1 e 3, enquanto que, a diferença registada entre as campanhas 1 e 3 não é significativa.

Testes para a Comparação entre *k* Variâncias

Hipóteses a testar: $H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

$H_1: \sigma_i^2 \neq \sigma_j^2$ para algum i e algum j tais que com $i \neq j$

Teste de *Bartlett*.

Este teste tem como pressuposto que as populações tenham distribuição normal.

Além disso, só é aplicável quando as diferentes amostras envolvidas tenham dimensões $\mathbf{n_j}$ não inferiores a quatro ($n_j \geq 4$, para todo o j).

Estatística de teste: $B = \frac{1}{C}\left[(N-k)\ln\left(S_p^2\right) - \sum_{j=1}^{k}(n_j-1)\ln\left(S_j^2\right)\right] \overset{sob H_0}{\sim} \chi^2_{k-1}$

onde,

- $N = \sum_{j=1}^{k} n_j$
- $S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2$
- $S_p^2 = \frac{1}{N-k}\sum_{j=1}^{k}(n_j - 1)\, S_j^2$
- $C = 1 + \frac{1}{3(k-1)}\left[\sum_{j=1}^{k}\frac{1}{n_j - 1} - \frac{1}{N-k}\right]$

Trata-se de um **teste unilateral à direita**: rejeita-se H_0 se $B_{obs} \geq \chi^2_{1-\alpha,k-1}$, onde $\chi^2_{1-\alpha,k-1}$ é o quantil de probabilidade (1-α) da distribuição χ^2_{k-1}.

Exemplo 3

Vamos testar a hipótese H_0, de igualdade de variâncias das três variáveis consideradas, ao nível de significância de 0.01.

Sob o pressuposto de H_0 ser verdadeira,

$$B = \frac{1}{C}\left[(N-k)\ln\left(S_p^2\right) - \sum_{j=1}^{k}(n_j-1)\ln\left(S_j^2\right)\right] \sim \chi_2^2.$$

$\chi^2_{0.99,2} = 9.21$ (quantil de probabilidade 0.99 da distribuição χ_2^2)

R.C.= [9.21, +∞[.

Para as amostras recolhidas tem-se,

$$B_{obs} = \frac{1}{1.09167}\left[15\ln(2.0165) - 4\ln(1.3) - 6\ln(2.95) - 5\ln(1.4667)\right] = 0.971 \notin \text{R. C.}$$

Ao nível de significância de 0.01, não se pode rejeitar a hipótese de que as três variáveis populacionais tenham iguais variâncias.

Teste de *Levene*.

Pressuposto: normalidade das distribuições populacionais.

Estatística de teste: $W = \frac{N-k}{k-1} \cdot \frac{\sum_{j=1}^{k} n_j(\overline{Z}_j - \overline{Z})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Z_{ij} - \overline{Z}_j)^2} \overset{sob\ H_0}{\sim} F_{N-k}^{k-1}$

onde,

- $N = \sum_{j=1}^{k} n_j$
- $Z_{ij} = \left|X_{ij} - \overline{X}_j\right|$
- $\overline{Z}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Z_{ij}$ – média dos Z_{ij} na amostra j
- $\overline{Z} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} Z_{ij}$ – média global dos Z_{ij}.

Teste unilateral à direita: rejeita-se H_0 se $W_{obs} \geq F_{1-\alpha}$, onde $F_{1-\alpha}$ é o quantil de probabilidade (1-α) da distribuição F_{N-k}^{k-1}.

Note:

Se existirem suspeitas de que a distribuição populacional não é normal, é aconselhável tomar $Z_{ij} = \left| X_{ij} - \widetilde{X}_j \right|$, em que $\widetilde{X}_j$ é a mediana da amostra j. A fórmula de cálculo com recurso à mediana é particularmente robusta e potente para desvios à normalidade da variável em estudo