Análise de Variância

ANOVA - Analysis of Variance

ISPV - ESCOLA SUPERIOR DE TECNOLOGIA DE VISEU
DepMat
departamento de matemática

Nos testes de hipótese fala-se em testar a igualdade de duas médias. E se for necessário testar a comparação de várias médias?

$\boldsymbol{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_k$

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

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

Sejam $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, ..., $\boldsymbol{X_k}$ as variáveis aleatórias que representam tal característica nas populações **1**, **2**, ...,***k***, respetivamente.

Será que podemos comparar as médias duas a duas através de um teste *t*-Student? **Não!…**

- Este procedimento é possível mas não é válido…pois a estatística do teste *t*-Student e o valor crítico só são válidos para comparar médias de duas, e apenas duas população (a partir das quais foram extraídas duas amostras aleatórias). Se utilizarmos este teste para comparar mais do que duas populações, duas a duas, a probabilidade do erro tipo I (*i.e.* a probabilidade de concluir erradamente que existem diferenças significativas) será aproximadamente $1-(1-\alpha)^k \times 100\%$, sendo k o número de amostras (populações) a comparar.

Por exemplo se k=3, para α=0,05 (relembre: α é o nível de significância do teste *i.e.* P(rej H_0 sendo H_0 verdadeira)= P(erro tipo I) a probabilidade de concluir erradamente que existem diferenças significativas fazendo comparações duas a duas é aproximadamente 14,3%!....

Técnica estatística para descobrir fatores que produzam mudanças sistemáticas em alguma variável de interesse

Podem ser quantitativos ou atributos

Quantitativa

- Os fatores são também designados de variáveis exógenas ou independentes
- A variável de interesse é designada de variável dependente

Exemplos

- Consumo de gasolina dos automóveis → variável de interesse

Marca

Idade veículo

Potência ….

exemplo de fatores que podemos pensar que influenciam consumo de gasolina

Numa determinada doença está-se a estudar o tempo que demora a recuperação → variável de interesse

Tipo de tratamento/medicação,
Idade do paciente
Sexo
Estação do ano …

exemplos de fatores

- Volume de vendas de lojas "iguais" → variável de interesse

Cidade onde está instalada
Equipa de vendedores …

exemplos de fatores

Será que o fator W 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 fator, *Tratamento*, que se apresenta em quatro níveis, A, B, C e D.

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

★ Suponhamos agora que existe a suspeita de que uma estação quente é um fator determinante para uma cura rápida. Então, o estudo deve ser conduzido tendo em conta este **segundo fator**, *Estação do Ano*.

Aqui, a técnica estatística apropriada será a ***análise de variância com dois fatores***, 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. (isto assumindo um "modelo aditivo" simples – onde não existe interação entre os dois fatores), ou até a existência de **interação** entre os dois fatores:

Será que pelo facto de estarmos no inverno um determinado tipo de tratamento afeta o tempo de tratamento até à eliminação da doença?

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

1. As amostras devem ser aleatórias e independentes.

2. As observações dentro de cada grupo têm distribuição Normal, *ié*, as *k* amostras devem ser extraídas de populações normais. Usar teste Kolmogorov-Smirnov[1] com a correção de Lilliefors, ou teste Shapiro Wilk (é preferível ao K-S nas amostras de pequena

---

[1] **Teste K-S** – teste *não paramétrico* já abordado na Unidade II, usados para decidir se distribuição de uma variável em estudo numa determinada amostra provém de uma população com uma distribuição específica. Neste caso estamos interessados apenas na distribuição Normal. Quando os verdadeiros valores de σ e μ são desconhecidos, o que acontece frequentemente, temos que recorrer à correção de Lilliefors (ver unidade II)

dimensão (n<30). Ambos usados no SPSS. O SPSS produz o p-value para este teste sempre que a dimensão da amostra ≤ 50.

3. As *k* populações devem ter variâncias iguais/homogéneas ($\sigma_1=\sigma_2= \ . \ . \ . \ =\sigma_k$), *ié*, há homocedasticidade.

- O teste de Levene - *teste paramétrico* já abordado - tem o pressuposto da normalidade das distribuições populacionais. Se existirem suspeitas de que o pressuposto da Normalidade falha deve fazer-se o teste de Levene usando a mediana (particularmente robusto e potente no caso de distribuições fortemente enviesadas).

Este teste é usado na ANOVA do SPSS .

Exemplo nº 1 a resolver na aula (Análise de Variância – 1 fator))

Suponha que se pretende comparar 3 lojas (uma de Aveiro, outra em Viseu e outra em Coimbra) quanto ao volume de vendas. Para tal em cada loja selecionou-se aleatoriamente 5 semanas onde se regista o volume de vendas. Obtém assim uma amostra das vendas semanais para cada loja (as três amostras são independentes). Supõe-se as três populações normais de igual variância.

Dados recolhidos: Lojas

	Aveiro	Viseu	Coimbra	
	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	$\sum\left(\overline{x}_i - \overline{\overline{x}}\right)^2 = 26$

fator de interesse → o fator **localização da Loja**, e este apresenta três níveis ou grupos: **Loja Aveiro**, **Loja Viseu** e **Loja Coimbra** (cada nível do fator define uma população de média μ_i).

Variável de interesse → Volume de vendas

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?

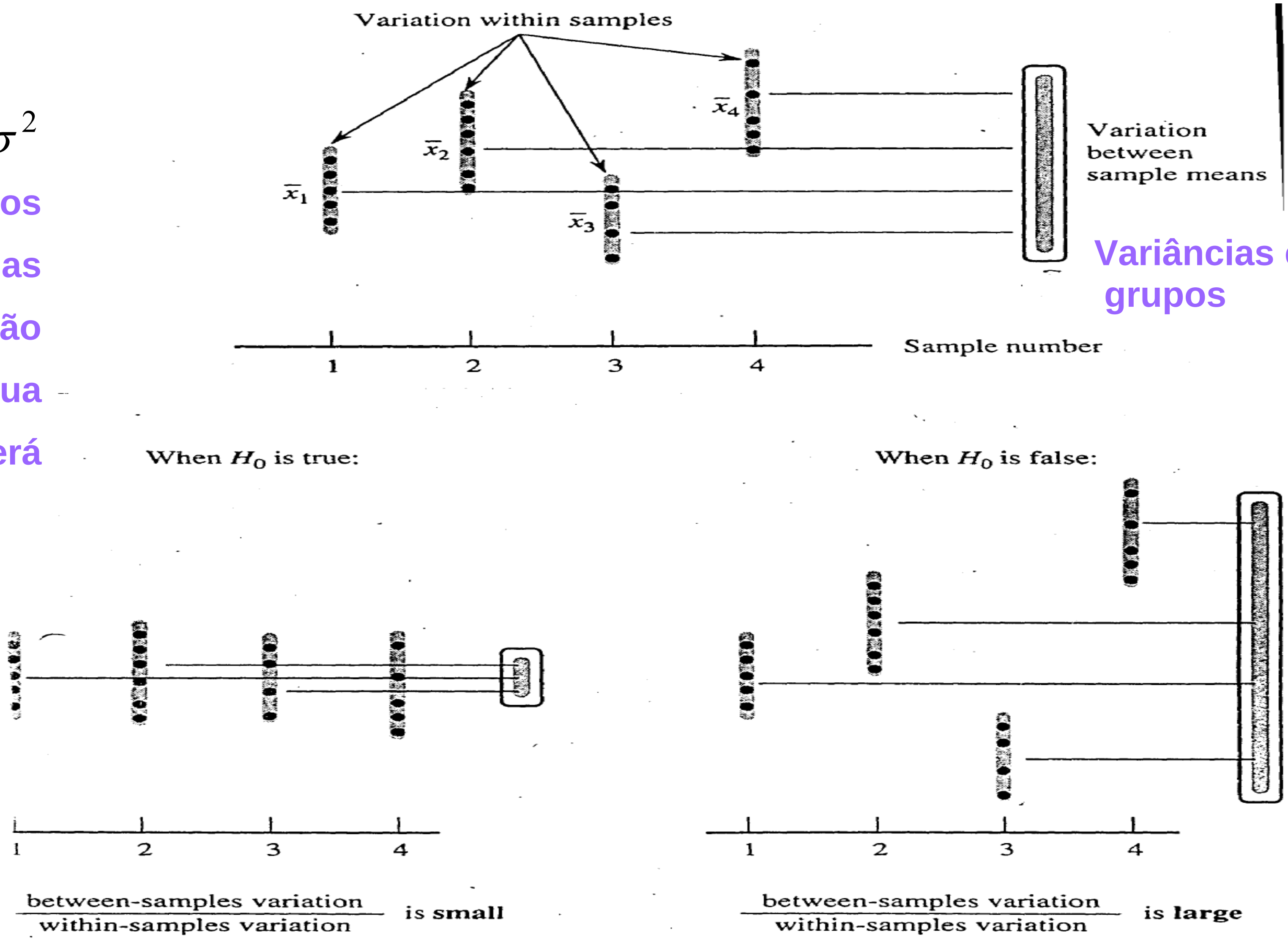
Podemos então formular as seguintes hipóteses:

***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)

Variâncias Dentro grupos, residual, ou dos erros

Se H_0 é verdadeira, σ^2 pode ser estimada pelos dois processos e as duas estimativas serão semelhantes, logo a sua razão, **F**, será aproximadamente 1



Variâncias entre grupos

How an ANOVA test works

Se a variância residual (dentro) for significativamente inferior à variância entre os grupos ou amostras (que será devido ao efeito fator em estudo), então as médias populacionais estimadas a partir das amostras são significativamente diferentes.

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.

Temos as seguintes estimativas de σ^2 para o nosso exemplo:

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

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

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

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:

$$\boxed{s_p^2 = \frac{s_1^2 + s_2^2 + \cdots + s_k^2}{k}}$$ onde, $s_i^2 \rightarrow$ variância amostral da amostra i.

Note que esta estimativa não é afetada 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

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. $\bar{X} \sim N(\mu, \sigma^2 / n)$.

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \qquad \Leftrightarrow \qquad \sigma^2 = n.\sigma_{\bar{X}}^2,$$

sugerindo que se estime σ^2 através de $\boxed{s_b^2 = n.s_{\bar{X}}^2}$, com $s_{\bar{X}}^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(\bar{x}_i - \bar{\bar{x}}\right)^2$ → estimativa de $\sigma_{\bar{X}}^2$.

O nosso exemplo viria:

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

logo a estimativa "entre" da variância é: $s_b^2 = n.s_{\bar{X}}^2 = 5 \times 13 = 65.$

Estatística de teste – F

Se a variância residual (dentro) for significativamente inferior à variân
entre os grupos ou amostras (que será devido ao efeito fator em estu
então as médias populacionais estimadas a partir das amostras
significativamente diferentes.

Sob o pressuposto de H_0 ser verdadeira, tem-se

$$F = \frac{\text{variação entre grupos}}{\text{variação dentro grupos}} \sim F_{N-k}^{k-1} \text{ i.é } F = \frac{n.S_{\overline{X}}^2}{S_p^2} = \frac{S_b^2}{S_p^2} \sim F_{N-k}^{k-1}$$, onde N é o total de observações

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

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

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

No nosso exemplo viria:

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

Se a hipótese H_0 é verdadeira, $F = \dfrac{S_b^2}{S_p^2} \sim F_{12}^{2}$.

$F_{1-\alpha,2,12} = 3.89$ (quantil de probabilidade 1-α da distribuição F_{12}^{2}) ; R.C.=[3.89,+∞[

O valor observado da estatística F é: $F_{obs} = \dfrac{65}{7.83} = 8.3 \in$ 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 fator Loja **exerce** uma influência significativa sobre o volume de vendas.

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

Na ANOVA one-way o comportamento da variável de medida é, supostamente, influenciado apenas por um fator (uma variável dependente). Os dados, usualmente, vêm representados da seguinte maneira:

	Amostra ou Grupos(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}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$x_{n_1 1}$	$x_{n_2 1}$	$x_{n_3 1}$	...	$x_{n_k 1}$

Onde

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

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

Tabela ANOVA:

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

$$SS_T=SS_A+SS_E$$

Assim para o exemplo 1 resolvido na aula, fica:

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

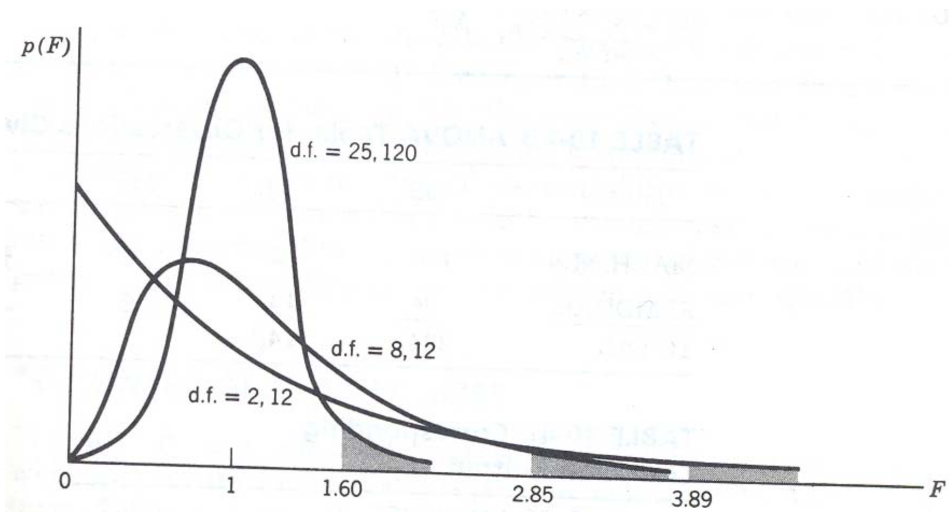
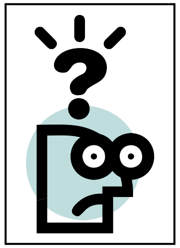
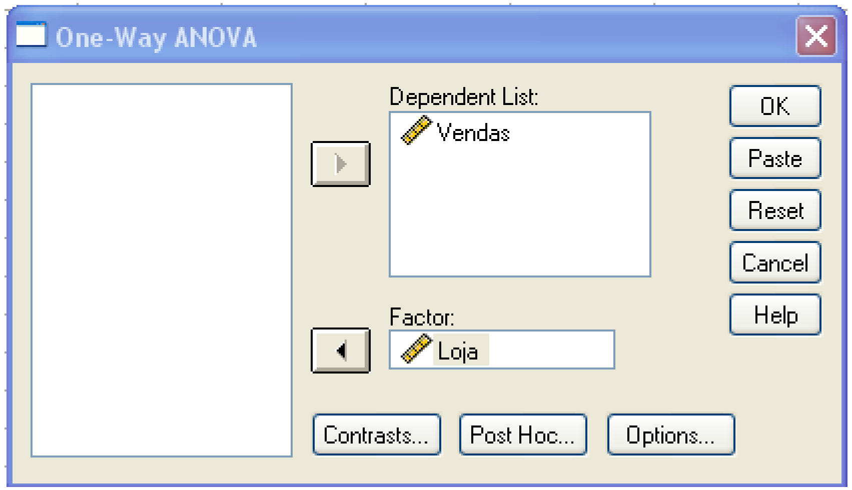


Figura 2: Ilustração de algumas distribuições F-Snedecor, com diferentes graus de liberdade no numerador e denominador. Note como o ponto crítico para 5% se move para a esquerda com o aumento dos graus de liberdade.

Resolução no SPSS:

`Analyse/Compare Means/One-way ANOVA/`





ANOVA

Vendas

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	130,000	2	65,000	8,298	,005
Within Groups	94,000	12	7,833		
Total	224,000	14			

O SPSS não testa *a priori* os pressupostos necessários a verificar em determinados testes. Na ANOVA não testa a normalidade da variável em estudo nem a igualdade das variâncias. Estes pressupostos têm que ser testados, antes ou depois da análise principal!

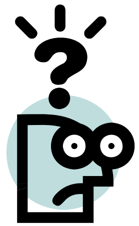
Teste à igualdade da dispersão (variâncias) das vendas

`Analyse/Compare Means/One-way ANOVA/Options/homogeneity of variance test`

Test of Homogeneity of Variances

Vendas

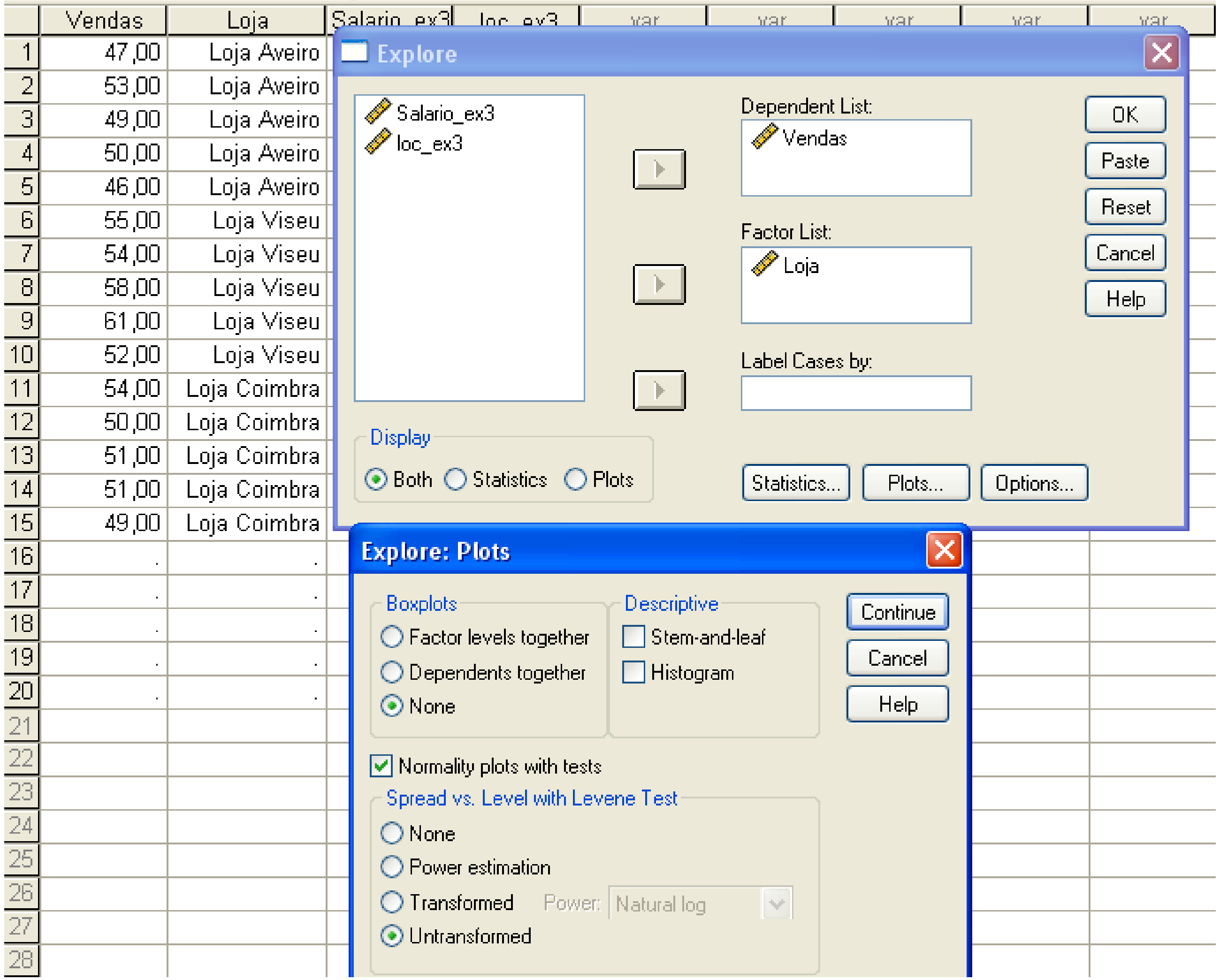
Levene Statistic	df1	df2	Sig.
1,391	2	12	,286



Aula...

O teste de Kolmogorov-Smirnov, e (novamente) o teste de Levene encontram-se no menu

Analyse/Descriptive Statistics/Explore:



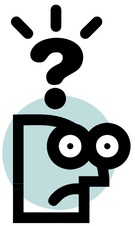
Aparecendo 4 quadros, sendo os dois primeiros um resumo descritivo da variável em estudo. O terceiro o teste de K-S com a correção de Lilliefors (pois μ e σ) são desconhecidos.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Loja	Statistic	df	Sig.	Statistic	df	Sig.
Vendas	Loja Aveiro	,167	5	,200*	,964	5	,833
	Loja Viseu	,211	5	,200*	,965	5	,844
	Loja Coimbra	,300	5	,161	,908	5	,453

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



E o quarto quadro apresenta o teste de Levene calculado a partir da média (Based on Mean).

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Vendas	Based on Mean	1,391	2	12	,286
	Based on Median	,779	2	12	,481
	Based on Median and with adjusted df	,779	2	9,696	,486
	Based on trimmed mean	1,353	2	12	,295

Exemplo 2 a resolver na aula

Pretende-se comparar a altura de determinada espécie de árvores com igual número de anos de três localidades distintas (localidade A, localidade B e localidade C), pois acredita-se que a altura das mesmas difere de localidade para localidade. Recolheram-se dados aleatoriamente das três localidades e de forma independente.

	Localidade A	Localidade B	Localidade C
	29	27	30
	27	27	30
	31	30	31
	29	28	27
	32		29
	30		
Soma	…	…	…

Fazendo o teste de Normalidade, e (novamente) o teste de Levene que se encontram-se no menu: `Analyse/Descriptive Statistics/Explore:`

ANOVA

Altura_ex2

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7,200	2	3,600	1,415	,281
Within Groups	30,533	12	2,544		
Total	37,733	14			

Test of Homogeneity of Variances

Altura_ex2

Levene Statistic	df1	df2	Sig.
,180	2	12	,837

Não há evidência estatística para rejeitar H0, logo assume-se a igualdade das variâncias nos 3 grupos.

Tests of Normality

	localid_ex2	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Altura_ex2	local A	,185	6	,200*	,974	6	,918
	local B	,260	4	.	,827	4	,161
	local C	,254	5	,200*	,914	5	,492

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

- Não há evidência estatística para rejeitar H_0assumindo-se a igualdade das alturas médias das árvores nas 3 localidades

  Para um erro de tipo 1 de 0.05, temos uma potência de 0.246.

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?**

> Estamos interessados em testar *a posteriori* (do que resulta a designação de "testes `Post-Hoc`") qual, ou quais os pares de médias diferentes.

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, que fazem basicamente o mesmo tipo de análise, não existindo consenso sobre qual destes é o mais apropriado, no entanto, aqui vamos abordar com maior detalhe o teste HSD (honestly significant difference) de Tuckey.

teste HSD de Tuckey

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*** - $n_{i\ max} < 2(n_{i\ min})$ - deste pressuposto.

Também se considera um dos testes mais robusto a desvios à normalidade e homogeneidade de variâncias (condições de aplicabilidade do teste) para amostras grandes.

$$[IC]_{(1-\alpha),(\mu_A-\mu_B)} = (\bar{x}_A - \bar{x}_B) \pm q_{K,N-K}\sqrt{\frac{MS_E}{2}\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}$$

quantil de prob. (1-α) da distribuição tabelada *Studentized Range*, com (*K*,*N*-*K*) graus de liberdade

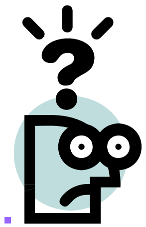
Observações

- O teste Bonferroni é considerado dos mais potentes para amostras pequenas;
- O testes Least Significant Difference - LSD podem ser utilizados especialmente quando se compara um nº reduzido de grupos, de um modo geral 4 ou menos, assim como o teste de Scheffé;
- Os testes de Tuckey e REGWQ, recomendado quando o número de observações em cada grupo (os n's) são iguais e variâncias populacionais iguais,
- Se os n´s forem muito diferentes e as variâncias populacionais iguais pode usar-se o testes Hochberg GT2. O SNK também é aplicado quando os n's não são iguais.
- De um modo geral os testes de Tuckey e de Scheffé são os mais usados. O de Tuckey quando as amostras são de tamanho igual ou de diferenças moderadas o de Scheffé quando as dimensões dos grupos são diferentes.

Continuação da resolução do exercício 1 – a resolver na aula (Resolução SPSS)

`Analyse/Compare Means/One-way ANOVA/Post Hoc (Multiple comparisons); Tukey, significance level=0.05`

Multiple Comparisons



Ver Aula....

Dependent Variable: Vendas
Tukey HSD

(I) Loja	(J) Loja	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Loja Aveiro	Loja Viseu	-7,00000*	1,77012	,005	-11,7224	-2,2776
	Loja Coimbra	-2,00000	1,77012	,515	-6,7224	2,7224
Loja Viseu	Loja Aveiro	7,00000*	1,77012	,005	2,2776	11,7224
	Loja Coimbra	5,00000*	1,77012	,038	,2776	9,7224
Loja Coimbra	Loja Aveiro	2,00000	1,77012	,515	-2,7224	6,7224
	Loja Viseu	-5,00000*	1,77012	,038	-9,7224	-,2776

*. The mean difference is significant at the .05 level.

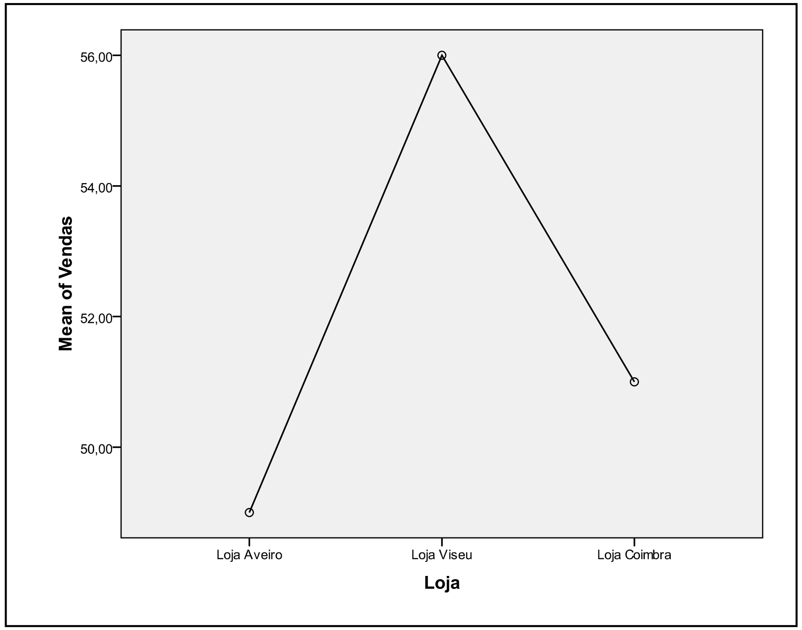
Vendas

Tukey HSD[a]

Loja	N	Subset for alpha = .05	
		1	2
Loja Aveiro	5	49,0000	
Loja Coimbra	5	51,0000	
Loja Viseu	5		56,0000
Sig.		,515	1,000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 5,000.



Options/Means Plots

Observações:

- É possível, se bem que pouco provável, que a ANOVA e os testes de Comparação Múltipla cheguem a conclusões diferentes. Isto deve-se ao facto de que a ANOVA é um teste mais potente (ié a prob. de rejeitar H_0, corretamente ou a confiança com que se rejeita a hipótese nula; `Observed Power`) do que as Comparações Múltiplas. Pode-se repetir o estudo com amostras de maior dimensão.