Análise de Variância simples (*One way ANOVA*)

Análise de experiências com vários grupos de observações classificados através de um só factor (por exemplo grupos de indivíduos sujeitos a diferentes tratamentos para uma mesma doença).

Muitas vezes também se utiliza a palavra **tratamento** em vez de grupo e diz-se que a experiência tem tantos **níveis** ou **efeitos** quantos tratamentos (ou grupos) distintos.

Se os grupos são pré-determinados à partida temos uma experiência com **efeitos fixos**.

Se os grupos forem escolhidos aleatoriamente entre um conjunto alargado de possibilidades temos uma experiência com **efeitos aleatórios**.

Um planeamento diz-se **completamente aleatorizado** se os indivíduos são escolhidos aleatoriamente e a distribuição pelos grupos também é aleatória.

Exemplo:

Um biólogo está interessado em estudar 3 variedades diferentes de trigo. O objectivo é averiguar se o tamanho médio dos grão se pode considerar igual para as três variedades. Para tal foram seleccionados 15 campos considerados homogéneos (mesmo tipo de solo e de condições climáticas) que foram divididos em três grupos de 5, de forma aleatória. As 3 variedades foram atribuídas aleatoriamente a cada um dos grupos de campos e ao fim de 3 meses de crescimento foi feita uma colheita de grãos de cada campo e calculado o peso médio da cada colheita.

Planeamento equilibrado

Quando o número de observações em cada grupo é igual diz-se que temos um planeamento equilibrado. Por razões de simplicidade na notação iremos apenas apresentar o modelo resultante de um planeamento equilibrado. Refira-se no entanto, que os resultados são equivalentes para outros planeamentos.

No que se segue iremos utilizar a seguinte notação:

Temos

- g grupos;
- n observações em cada grupo (planeamento equilibrado);
- total de $N = gn$ observações.

Análise de Variância simples - Efeitos fixos

As observações são designadas por Y_{ij} onde $i = 1, \ldots, g$ identifica o grupo e $j = 1, \ldots, n$ identifica a posição de cada observação dentro do seu grupo.

$$Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij},$$

onde

- μ_i representa a média de cada grupo,
- μ representa a média de todos os grupos,
- τ_i representa a diferença entre a média total e a média de cada grupo $(\sum_{i=1}^{g} \tau_i = 0)$, e
- ϵ_{ij} representa um erro aleatório de cada observação sendo estes erros independentes entre si.

Pressupõe-se que

$$\epsilon_{ij} \frown N(0, \sigma^2), \quad \text{pelo que} \quad Y_{ij} \frown N(\mu_i, \sigma^2)$$

Isto significa que cada grupo provém de uma população Normal com uma certa média μ_i, mas todos com a mesma variância σ^2.

Hipóteses a testar

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_g = \mu \quad vs \quad H_1: \ \mu_i \neq \mu \ \text{ pelo menos para um } i$$

ou equivalentemente

$$H_0: \ \tau_1 = \tau_2 = \ldots = \tau_g = 0 \quad vs \quad H_1: \ \tau_i \neq 0 \ \text{pelo menos para um } i$$

Resumindo:

Pressupostos exigidos:

1. Temos g grupos de observações independentes (g amostras aleatórias) sendo os grupos independentes entre si.
2. Cada grupo de observações deve provir de uma distribuição Normal.
3. A variância das g populações deve ser a mesma.

Hipóteses a testar

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_g = \mu \quad vs \quad H_1: \ \mu_i \neq \mu \ \text{ pelo menos para um } i$$

Modelo:

$$Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \frown N(0, \sigma^2)$$

Ideia de base

Para testar estas hipóteses recorre-se a uma análise das variâncias dos vários grupos e daí o nome ANOVA. A ideia de base é a seguinte: Vamos estimar a variância σ^2 por dois métodos diferentes, um que não depende da veracidade de H_0 e outro que sim. Depois comparamos as duas estimativas. Se os grupos tiverem todos a mesma média (H_0 verdadeiro) as duas estimativas deverão ser próximas, senão deverão diferir significativamente.

Uma forma de estimar σ^2, sem depender da veracidade de H_0, consiste em calcular para cada grupo a variância amostral corrigida (estimativa de σ^2) e tomar a média das várias estimativas que se obtêm.

Se pensarmos agora que as médias são todas iguais (H_0 verdadeiro) estamos perante um conjunto de g amostras todas da mesma população. Sabemos que $Var[\bar{X}] = \sigma^2/n$ e podemos obter uma "amostra"de g médias amostrais (uma para cada grupo). Calculando a variância amostral desta "amostra"de médias amostrais temos uma estimativa de σ^2/n. Multiplicando por n temos uma estimativa de σ^2.

Mas esta última estimativa só é boa se H_0 for verdadeira. Senão fica muito inflaccionada. Assim, ao dividir a última estimativa pela primeira devemos obter um valor próximo de 1 se H_0 for verdadeiro e muito maior que 1 caso contrário.

Partição da soma de quadrados

Seja

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{i\cdot} = \frac{y_{i\cdot}}{n}$$

$$y_{\cdot\cdot} = \sum_{i=1}^{g}\sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{N}$$

$$SS_T = \sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2.$$

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2}_{SS_T} = \underbrace{n\sum_{i=1}^{g}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}_{SS_G} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2}_{SS_E}$$

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2}_{SS_T} = \underbrace{n\sum_{i=1}^{g}(\bar{y}_{i.}-\bar{y}_{..})^2}_{SS_G} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}_{SS_E}$$

Seja

$$MS_G = \frac{SS_G}{g-1}, \quad \text{e} \quad MS_E = \frac{SS_E}{g(n-1)}.$$

Então,

sob H_0	sob H_1
$E[MS_G] = \sigma^2$	$E[MS_G] = \sigma^2 + \frac{n\sum_{i=1}^{g}\tau_i^2}{g-1}$
$E[MS_E] = \sigma^2$	$E[MS_E] = \sigma^2$

SS_T tem $N-1 = gn-1$ graus de liberdade.

SS_G tem $g-1$ graus de liberdade.

SS_E tem $g(n-1)$ graus de liberdade.

Pode-se mostrar que sob H_0

$$\frac{SS_G}{\sigma^2} \frown \chi^2_{g-1} \quad \text{e} \quad \frac{SS_E}{\sigma^2} \frown \chi^2_{g(n-1)},$$

sendo estas variáveis independentes.

Assim, sob H_0

$$\frac{MS_G}{MS_E} \frown F_{g-1,g(n-1)}$$

e podemos efectuar um teste com base nesta estatística.

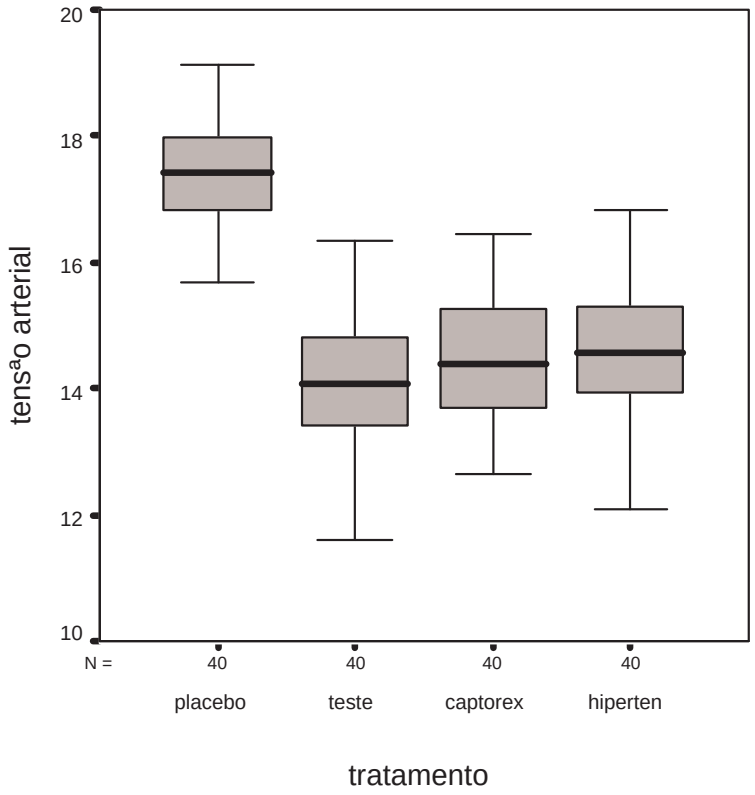
Tabela de ANOVA

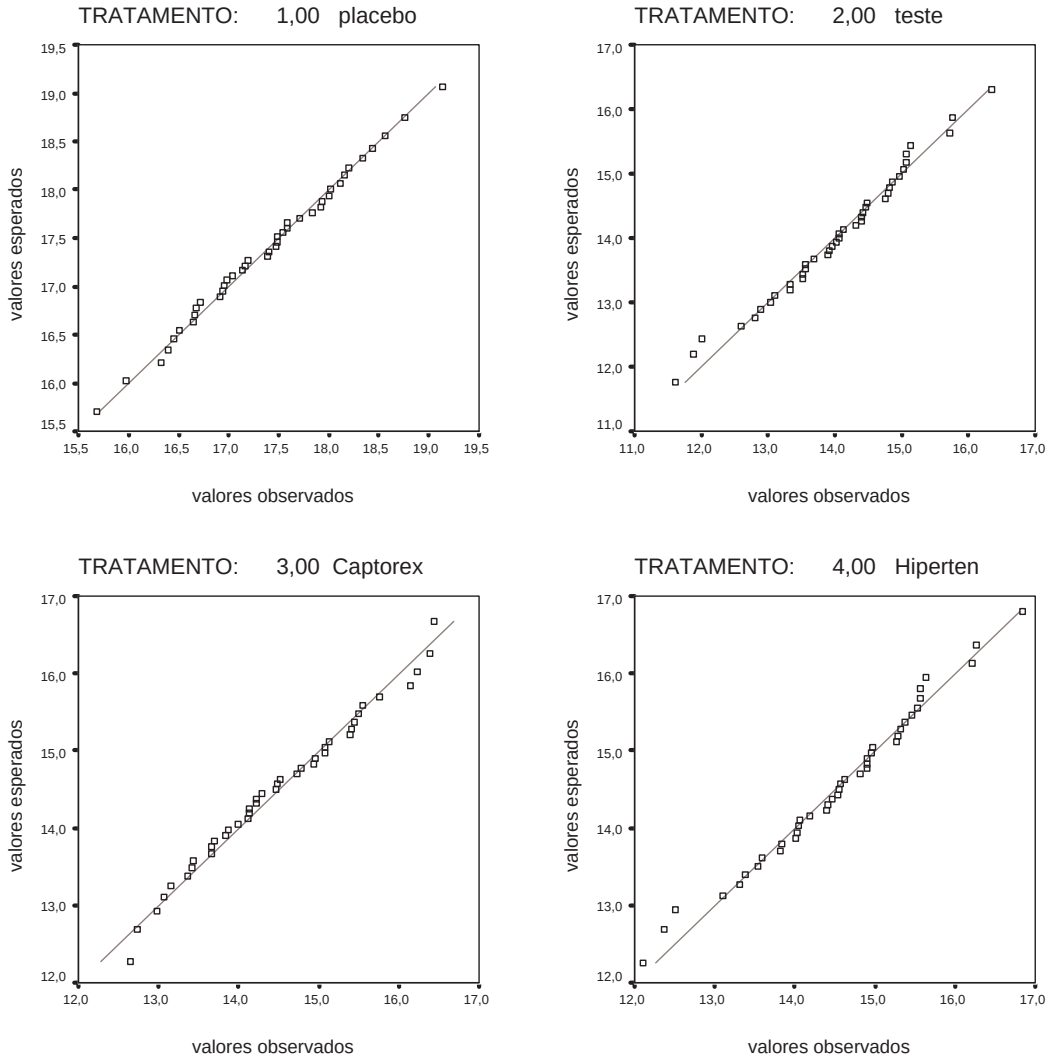
Fonte de Variação	Soma de quadrados	g.l.	Média de quadrados	F_{obs}	p
Entre Grupos	SS_G	$g-1$	MS_G	$\frac{MS_G}{MS_E}$	$(\cdot)$
Dentro dos grupos	SS_E	$g(n-1)$	MS_E		
Total	SS_T	$gn-1$			

F_{obs} é o valor observado da estatística de teste F.
p é o *p-value* do teste.

Exemplo:

160 indivíduos hiper-tensos divididos em 4 grupos de 40.
4 tratamentos: hiperten, captorex, novo medicamento e placebo.




TRATAMENTO: 1,00 placebo
valores esperados
valores observados
TRATAMENTO: 2,00 teste
valores esperados
valores observados
TRATAMENTO: 3,00 Captorex
valores esperados
valores observados
TRATAMENTO: 4,00 Hiperten
valores esperados
valores observados

Test of Homogeneity of Variances

tensão arterial

Levene Statistic	df1	df2	Sig.
1,182	3	156	,318

ANOVA

tensão arterial

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	283,126	3	94,375	97,550	,000
Within Groups	150,923	156	,967		
Total	434,049	159			

A forma habitual de apresentar os resultados de uma ANOVA num trabalho científico consiste em apresentar características amostrais de cada grupo (médias e desvios padrões) e depois indicar o valor observado da estatística de teste F e o valor do *p-value* da ANOVA. A tabela de ANOVA propriamente dita poderá vir em anexo.

Análise de Variância simples - Efeitos aleatórios

Modelo:

$$Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij},$$

onde τ_i e ϵ_{ij} são variáveis aleatórias independentes.

$$\epsilon_{ij} \frown N(0, \sigma^2), \quad \tau_i \frown N(0, \sigma^2_\tau).$$

$$Y_{ij} \frown N(\mu_i = \mu + \tau_i, \sigma^2 + \sigma^2_\tau).$$

Hipóteses a testar

$$H_0 : \ \sigma^2_\tau = 0 \quad vs \quad H_1 : \sigma^2_\tau > 0.$$

Mantém-se a relação

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2}_{SS_T} = \underbrace{n\sum_{i=1}^{g}(\bar{y}_{i.}-\bar{y}_{..})^2}_{SS_G} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}_{SS_E}$$

Agora

sob H_0	sob H_1
$E[MS_G] = \sigma^2$	$E[MS_G] = \sigma^2 + n\sigma^2_\tau$
$E[MS_E] = \sigma^2$	$E[MS_E] = \sigma^2$

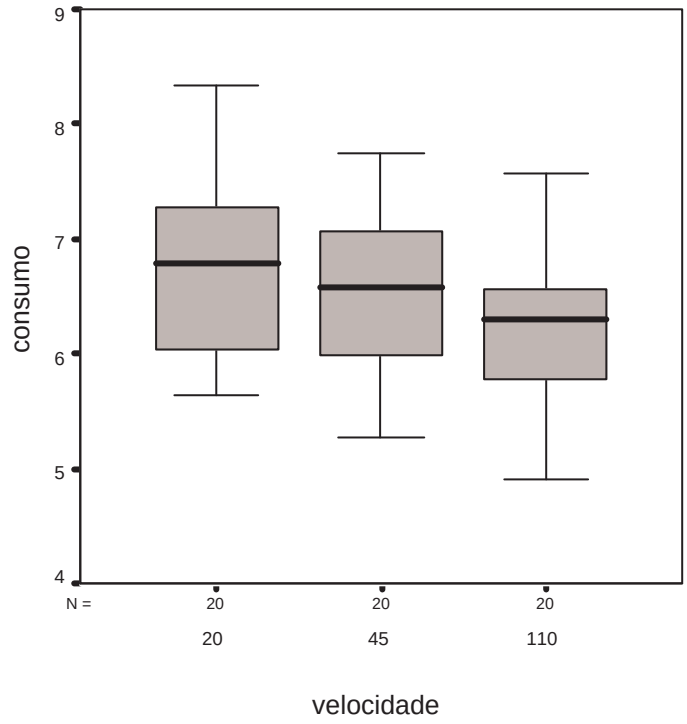
Sob H_0

$$F = \frac{MS_G}{MS_E} \frown F_{g-1,g(n-1)}.$$

$$\hat{\sigma}^2_\tau = \frac{MS_G - MS_E}{n}$$

Exemplo:

Pretende-se estudar se o consumo de combustível de um automóvel da Marca A depende da velocidade com que o automóvel se desloca. Para tal seleccionaram-se aleatoriamente 3 valores de velocidade e efectuou-se uma experiência envolvendo 60 automóveis distribuídos aleatoriamente em 3 grupos homogéneos.



Descriptives

consumo

		Std. Deviation	Std. Error	95% Confidence Interval for Mean		Between-Component Variance
				Lower Bound	Upper Bound	
Model	Fixed Effects	,69847	,09017	6,3366	6,6977	
	Random Effects		,15237	5,8615	7,1727	,04526

ANOVA

consumo

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2,786	2	1,393	2,855	,066
Within Groups	27,808	57	,488		
Total	30,594	59			

Comparações múltiplas

Uma vez rejeitada H_0 o que fazer para procurar identificar quais os grupos que causam as diferenças?

Considerar todas as comparações de pares de médias envolvidos na ANOVA para procurar detectar quais os grupos que provocam a rejeição de H_0 na tabela de ANOVA. Em n grupos há $\frac{n!}{2!(n-1)!}$ comparações de pares de médias distintos.

Dois problemas:

1. Cálculo do nível de significância de cada comparação e do nível de significância do conjunto de comparações que se está a efectuar em simultâneo.
2. As comparações não são todas independentes.

Se uma comparação individual tiver tamanho α_m, um conjunto de m comparações (independentes) tem tamanho $\alpha = 1 - (1 - \alpha_m)^m$. Por exemplo, em 20 comparações, se cada comparação tiver tamanho 5%, o tamanho total é 64% que é inaceitável.

Importante lembrar:

A análise de comparações múltiplas não faz sentido nos modelos de efeitos aleatórios e só deve ser utilizada nos modelos de efeitos fixos.

A análise de comparações múltiplas só deve ser efectuada quando se rejeita H_0 na tabela da ANOVA.

Existem muitos métodos para efectuar comparações múltiplas. Iremos apenas referir alguns, nomeadamente o método de Bonferroni, o método de Tuckey e o método de Dunnett.

Método de Bonferroni

α — tamanho total das comparações múltiplas,

α_m — tamanho de cada comparação individual

$R_i = \{$a i-ésima hipótese nula é rejeitada quando é verdadeira$\}$.

$$\alpha = P\{R_1 \text{ ou } R_2 \text{ ou } \ldots \text{ ou } R_m\} \leq m\alpha_m,$$

O método de Bonferroni consiste em considerar para cada comparação individual um nível de significância $\alpha_m = \alpha/m$ por forma a garantir que o nível total não ultrapassa α.

Aplicando este método alguns dos pares que eventualmente acusavam diferenças significativas podem deixar de o fazer.

No SPSS a tabela que é produzida para este método fornece $p - values$ para cada comparação que resultam da multiplicação dos p-values dos testes por m. Assim, em vez de compararmos os p-values com α/m, comparamos os produtos $m \times p\text{-}value$ com α.

Exemplo:

Multiple Comparisons

Dependent Variable: tensão arterial

	(I) tratamento	(J) tratamento	Mean Difference (I-J)	Std. Error	Sig.
LSD	placebo	teste	3,3540*	,21994	,000
		captorex	2,9099*	,21994	,000
		hiperten	2,8540*	,21994	,000
	teste	placebo	-3,3540*	,21994	,000
		captorex	-,4440*	,21994	,045
		hiperten	-,5000*	,21994	,024
	captorex	placebo	-2,9099*	,21994	,000
		teste	,4440*	,21994	,045
		hiperten	-,0560	,21994	,800
	hiperten	placebo	-2,8540*	,21994	,000
		teste	,5000*	,21994	,024
		captorex	,0560	,21994	,800
Bonferroni	placebo	teste	3,3540*	,21994	,000
		captorex	2,9099*	,21994	,000
		hiperten	2,8540*	,21994	,000
	teste	placebo	-3,3540*	,21994	,000
		captorex	-,4440	,21994	,271
		hiperten	-,5000	,21994	,146
	captorex	placebo	-2,9099*	,21994	,000
		teste	,4440	,21994	,271
		hiperten	-,0560	,21994	1,000
	hiperten	placebo	-2,8540*	,21994	,000
		teste	,5000	,21994	,146
		captorex	,0560	,21994	1,000

Método de Tuckey

Construção de intervalos de confiança para todos os pares de comparações de tal forma que o conjunto de todos os intervalos tenha uma certa confiança, $1-\alpha$.

$$\max_{i,j} \frac{|(\bar{Y}_{i.} - \mu_i) - (\bar{Y}_{j.} - \mu_j)|}{\sqrt{MS_E}}$$

onde o máximo é calculado para todos os pares i, j. A distribuição desta variável é denominada ***studentized range distribution*** com parâmetros g e $g(n-1)$.

No SPSS após a tabela de comparações múltiplas é produzida uma tabela de grupo homogéneos. Trata-se de uma tabela que subdivide os g grupos de observações em sub-grupos dentro dos quais podemos considerar que as médias não apresentam diferenças significativas.

Exemplo:

Multiple Comparisons

Dependent Variable: tensão arterial
Tukey HSD

(I) tratamento	(J) tratamento	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
placebo	teste	3,3540*	,21994	,000	2,7828	3,9252
	captorex	2,9099*	,21994	,000	2,3388	3,4811
	hiperten	2,8540*	,21994	,000	2,2828	3,4252
teste	placebo	-3,3540*	,21994	,000	-3,9252	-2,7828
	captorex	-,4440	,21994	,185	-1,0152	,1271
	hiperten	-,5000	,21994	,109	-1,0712	,0712
captorex	placebo	-2,9099*	,21994	,000	-3,4811	-2,3388
	teste	,4440	,21994	,185	-,1271	1,0152
	hiperten	-,0560	,21994	,994	-,6271	,5152
hiperten	placebo	-2,8540*	,21994	,000	-3,4252	-2,2828
	teste	,5000	,21994	,109	-,0712	1,0712
	captorex	,0560	,21994	,994	-,5152	,6271

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

ANOVA simples não paramétrica — Teste de Kruskal-Wallis

Temos

- g grupos;
- n_i observações no grupo i;
- total de $N = \sum_{i=1}^{g} n_i$ observações.

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

$i = 1, \ldots, g$, $j = 1 \ldots, n_j$ onde os erros ϵ_{ij} são v.a.'s contínuas com a mesma distribuição, e μ_i representa a mediana do grupo i.

Pressupostos exigidos:

1. Temos g grupos de observações independentes (g amostras aleatórias) sendo os grupos independentes entre si.
2. As observações são medidas numa escala pelo menos ordinal.
3. Cada grupo de observações deve provir de uma população contínua.
4. As populações apenas diferem na localização (portanto têm a mesma forma).

Hipótese a testar

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_g = \mu \quad vs \quad H_1: \ \mu_i \neq \mu \text{ pelo menos para um } i,$$

onde μ_i representa a mediana do grupo i.

Procedimento:

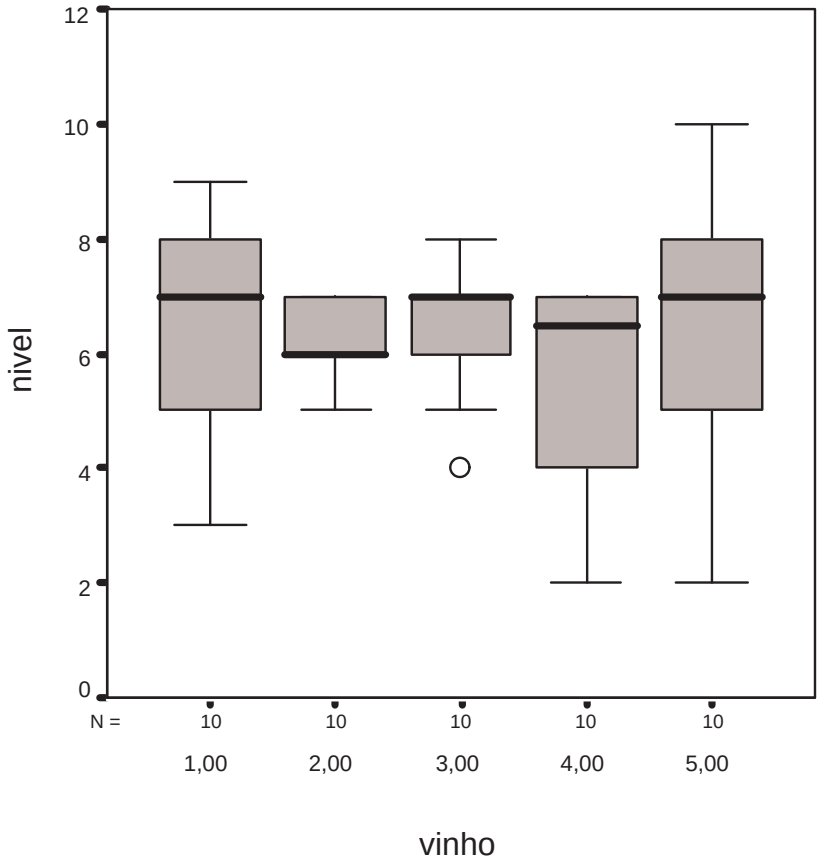
Ordenar o total das N observações em conjunto, e atribuir *ranks* às observações.

Seja R_{ij} o *rank* da observação Y_{ij}. Denote-se por $R_{i\cdot}$ e $\bar{R}_{i\cdot}$ a soma e a média dos *ranks* do grupo i, respectivamente. A Estatística de teste é dada por

$$T = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \left(\bar{R}_{i\cdot} - \frac{N+1}{2} \right)^2 = \frac{12}{N(N+1)} \sum_{i=1}^{g} \frac{R_{i\cdot}^2}{n_i} - 3(N-1).$$

T tem distribuição aproximadamente χ^2 com $g-1$ graus de liberdade, sob H_0. Portanto rejeita-se H_0 se $T > \chi_{1-\alpha,g-1}$ ao nível de significância α.

Exemplo:



Ranks

	VINHO	N	Mean Rank
NIVEL	1,00	10	28,75
	2,00	10	22,00
	3,00	10	26,85
	4,00	10	20,90
	5,00	10	29,00
	Total	50	

Test Statistics[a,b]

	NIVEL
Chi-Square	2,901
df	4
Asymp. Sig.	,575

a. Kruskal Wallis Test

b. Grouping Variable: VINHO