Análise de Variância a um factor

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.

Temos

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

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.

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 um acerta média μ_i, mas todos com a mesma variância σ^2.

Hipótese a testar

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_g = \mu \quad vs \quad H_1: \ \mu_i \neq \mu$$

ou equivalentemente

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

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.

Seja

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

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

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

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

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

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.

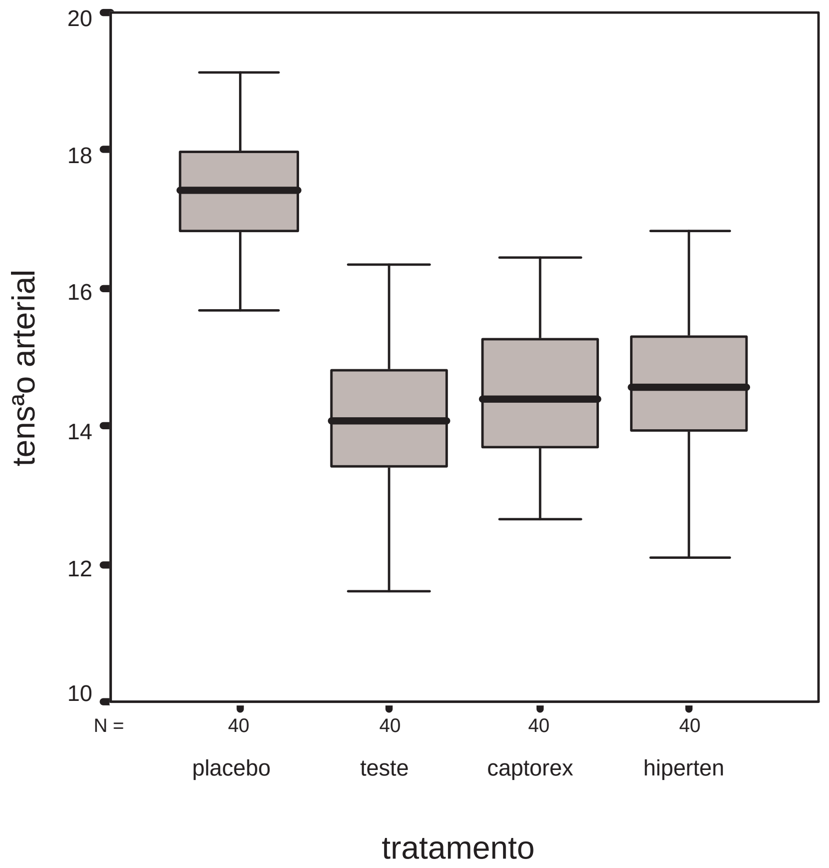
Tabela de ANOVA

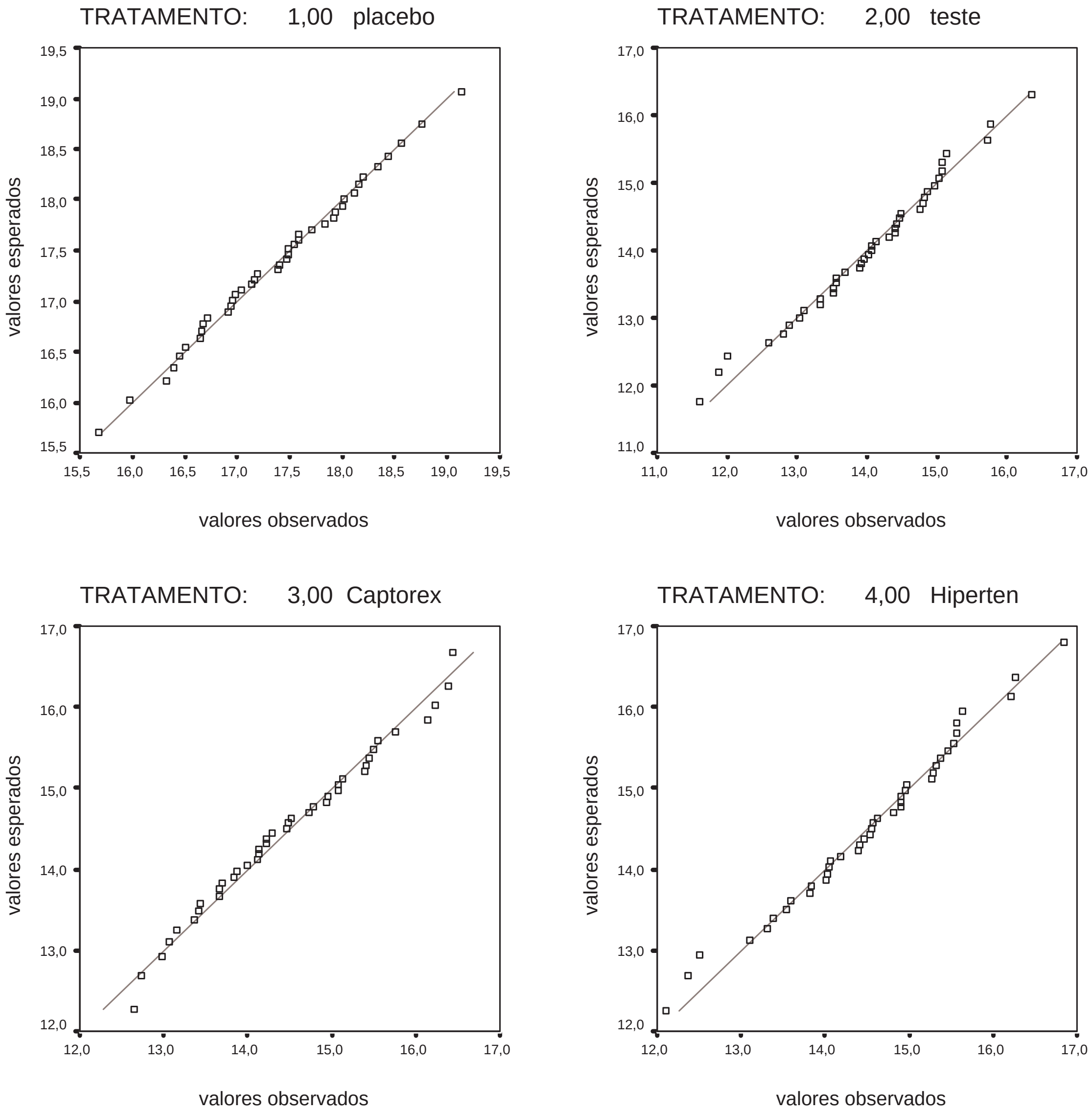
Fonte de Variação	Soma de quadrados	g.l.	Média de quadrados	F_0	p
Entre Grupos	SS_G	$g-1$	MS_G	$\frac{MS_G}{MS_E}$	$(\cdot)$
Dentro dos grupos	SS_E	$g(n-1)$	MS_E		
Total	SS_T	$gn-1$			

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
15,5
16,0
16,5
17,0
17,5
18,0
18,5
19,0
19,5
TRATAMENTO: 2,00 teste
valores esperados
valores observados
11,0
12,0
13,0
14,0
15,0
16,0
17,0
TRATAMENTO: 3,00 Captorex
valores esperados
valores observados
12,0
13,0
14,0
15,0
16,0
17,0
TRATAMENTO: 4,00 Hiperten
valores esperados
valores observados
12,0
13,0
14,0
15,0
16,0
17,0

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 reportar os resultados de uma ANOVA num trabalho da área de Ciências de fala e audição consiste em apresentar uma tabela com características amostrais de cada grupo (médias e desvios padrões) e depois indicar o valor da estatística de teste F e o valor do p-value da tabela de ANOVA. Regra geral não se apresenta a tabela de ANOVA.

Efeitos aleatórios

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

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

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

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

Hipóteses a testar

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

Mantém-se a relação

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

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

Agora

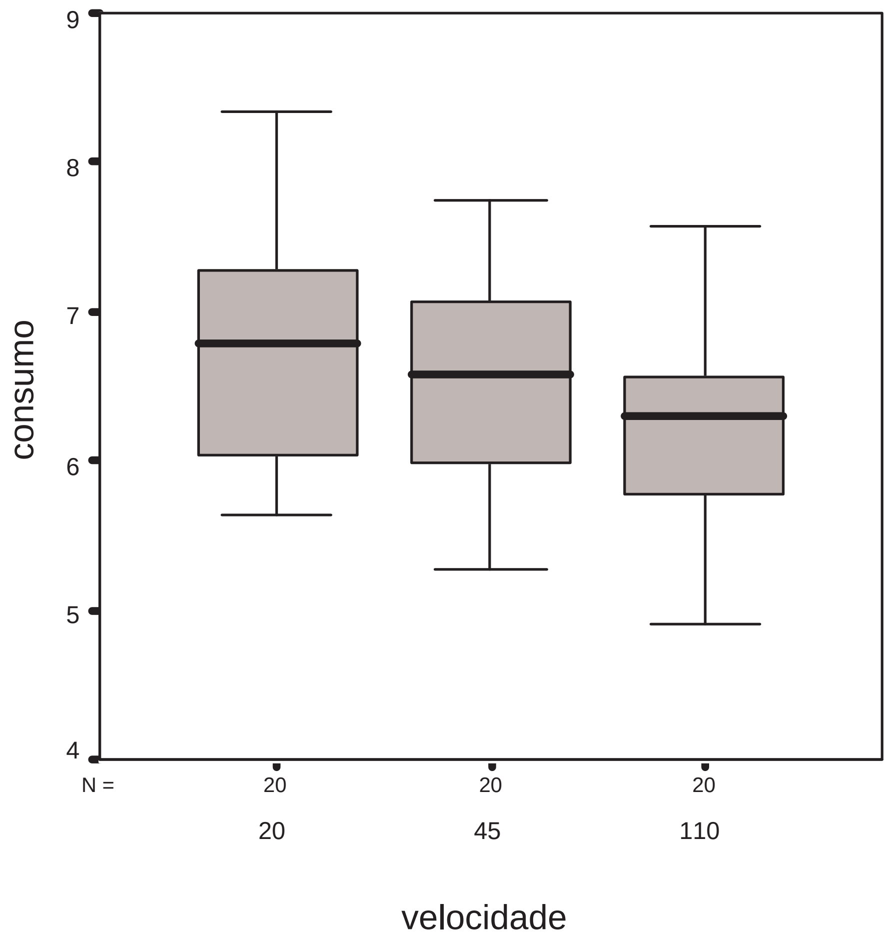
sob H_0	sob H_1
$E[MS_G] = \sigma^2$	$E[MS_G] = \sigma^2 + n\sigma_\tau^2$
$E[MS_E] = \sigma^2$	$E[MS_E] = \sigma^2$

Sob H_0

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

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

Exemplo



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			

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

Comparações múltiplas — métodos de Bonferroni e Tuckey

Vamos 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.

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.

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

Tukey

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, $1 - \alpha$.

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

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

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.

Nota: A aplicação de contrastes ou de comparações múltiplas não faz sentido nos modelos de efeitos aleatórios e só deve ser utilizada nos modelos de efeitos fixos.

ANOVA 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.

Hipóteses a testar

$$H_0 : \ \mu_1 = \mu_2 = \ldots = \mu_g = 0 \quad vs \quad H_1 : \ \mu_i \neq 0 \text{ pelo menos para um } i.$$

O teste pressupõe apenas que: as distribuições dos grupos são contínuas e apenas diferem na localização (portanto têm a mesma forma); todas as observações são independentes.

Procedimento:

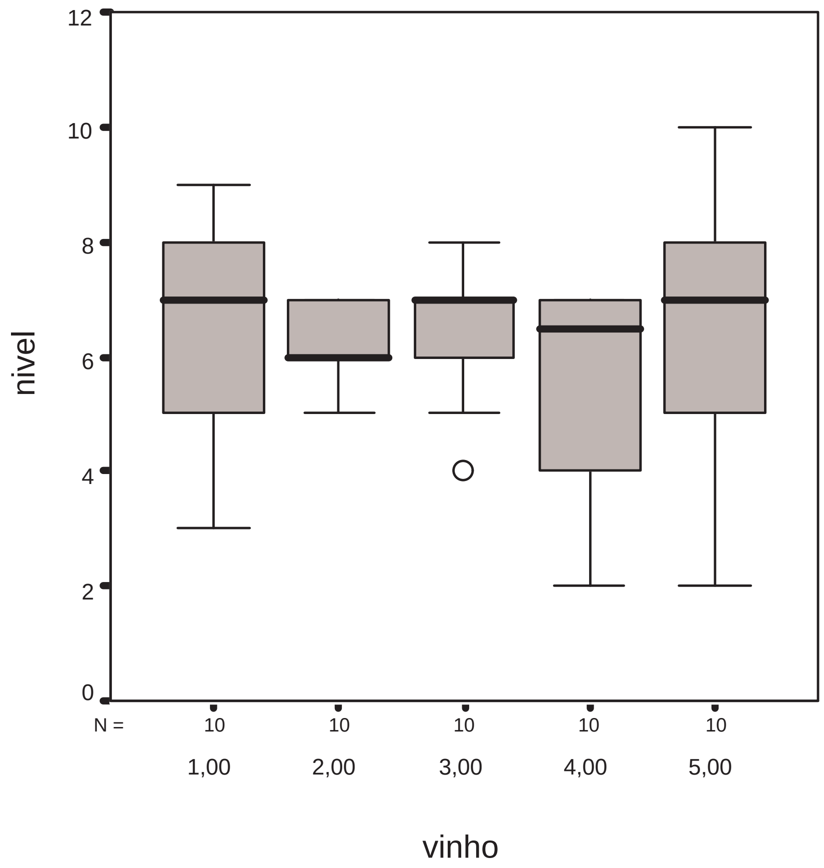
Ordenar o total das N observações em conjunto, e atribuir *ranks* às observações.

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

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

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

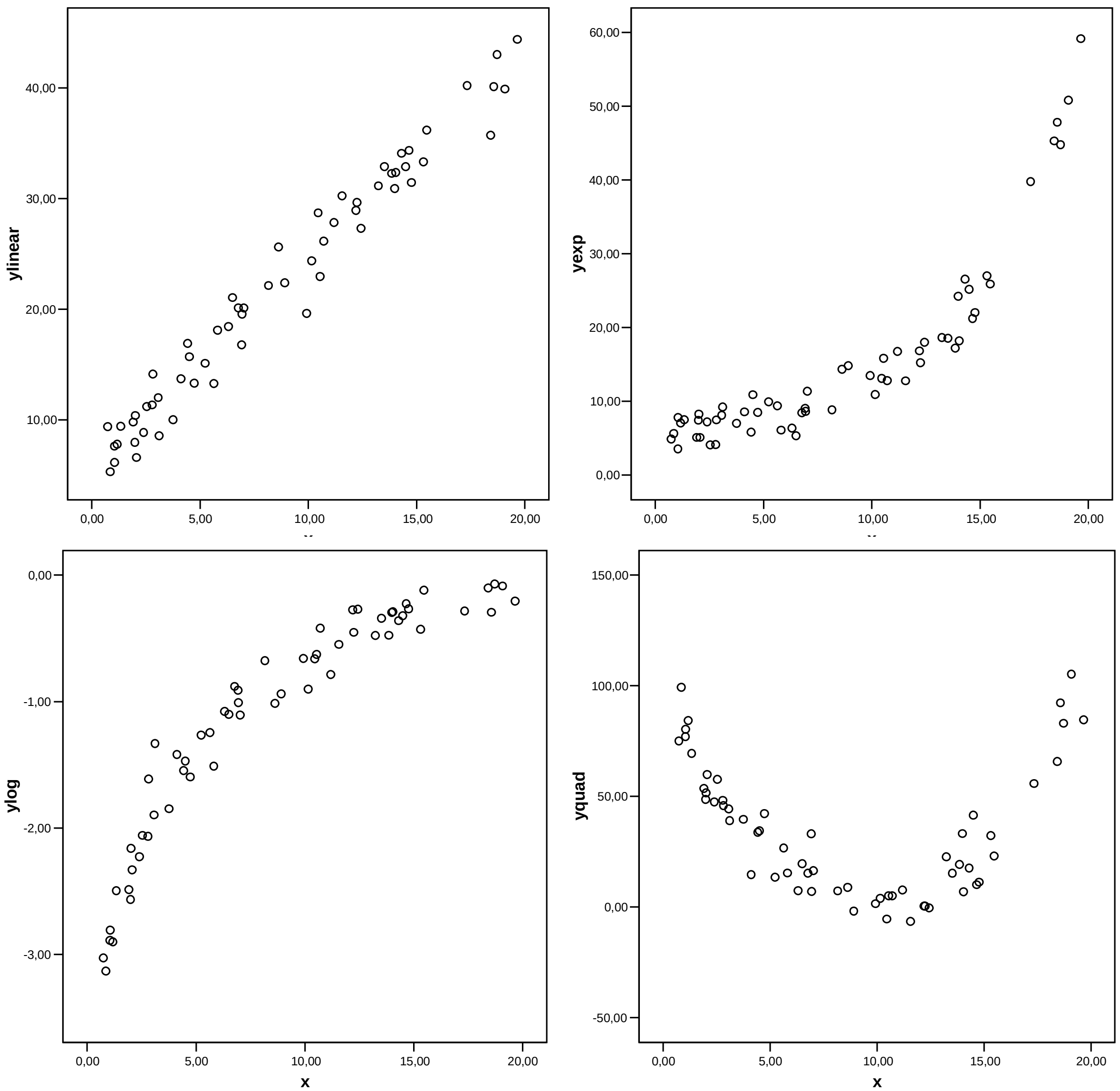
Associação entre variáveis

Questões de interesse:

Será que duas variáveis são independentes ou pelo contrário dependentes? E se forem dependentes, qual o tipo e grau de dependência?

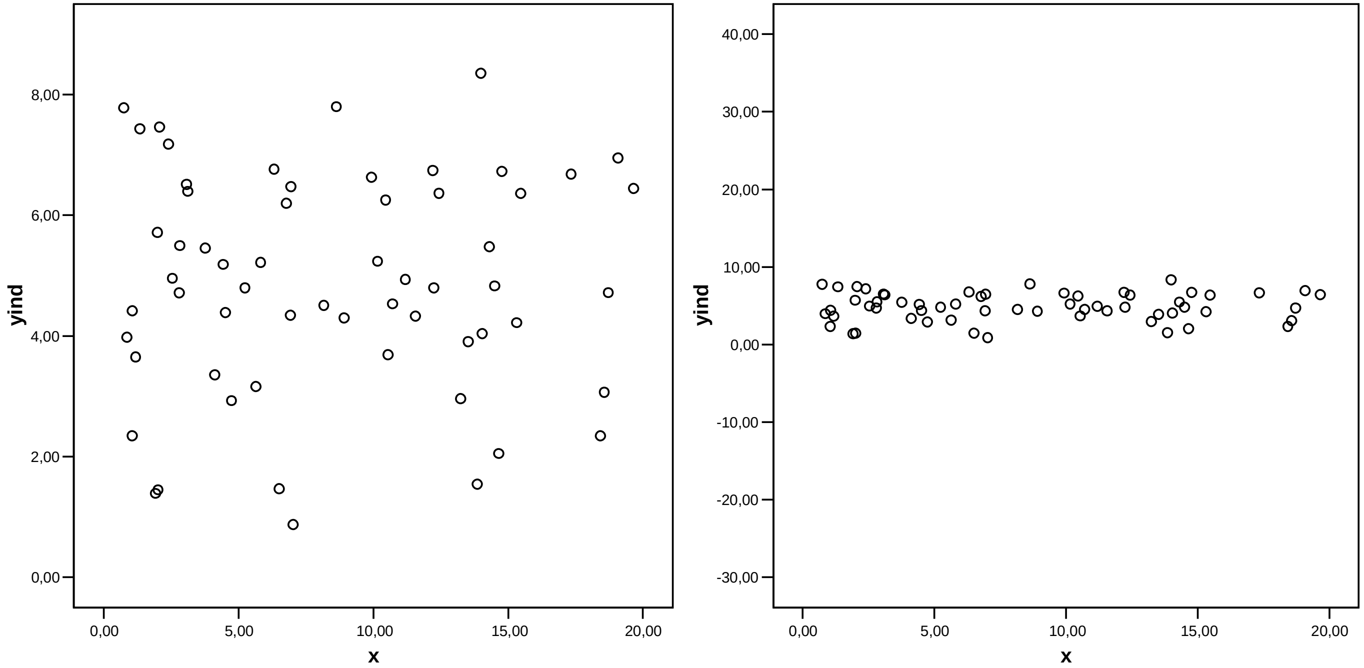
Medir o grau de dependência é mais ambicioso do que simplesmente testar a existência de alguma associação entre variáveis. É obviamente de interesse poder medir o grau de associação entre dois conjuntos de observações obtidos a partir de um dado conjunto de unidades experimentais (indivíduos por exemplo). Mas, talvez seja mais importante podermos dizer se uma certa associação observada nos dados indica ou não uma associação na população de onde foram retirados.

Formas de associação entre variáveis numéricas: lineares, exponenciais, logarítmicas ou quadráticas.

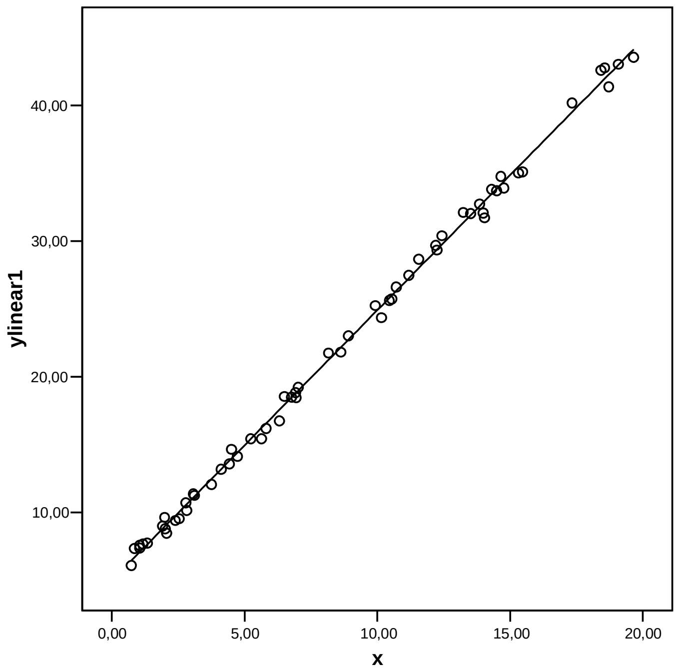

ylinear
40,00
30,00
20,00
10,00
0,00
5,00
10,00
15,00
20,00
yexp
60,00
50,00
40,00
30,00
20,00
10,00
0,00
ylog
0,00
-1,00
-2,00
-3,00
x
yquad
150,00
100,00
50,00
0,00
-50,00
x

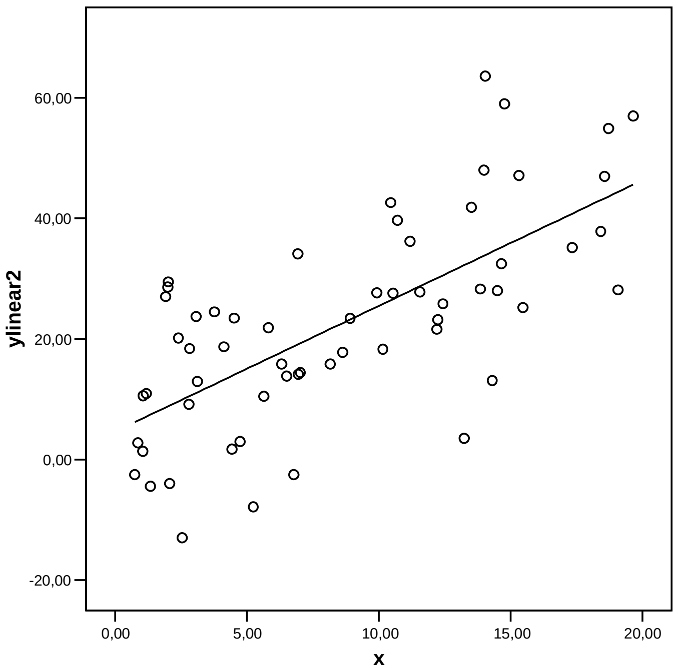
Primeiro passo: construção de diagramas de dispersão.

Quando duas variáveis são independentes, o diagrama de dispersão respectivo apresenta uma mancha de pontos aleatória (ou quando muito) um conjunto de pontos dispostos sobre uma recta horizontal.



Se a relação entre duas variáveis for linear, ao confrontarmos duas amostras num diagrama de dispersão devemos esperar observar um conjunto de pontos que se dispõem aproximadamente sobre uma recta. Por vezes os desvios em relação à recta são mínimos, mas noutras os pontos apresentam bastante dispersão tornando difícil a identificação da dita relação linear.



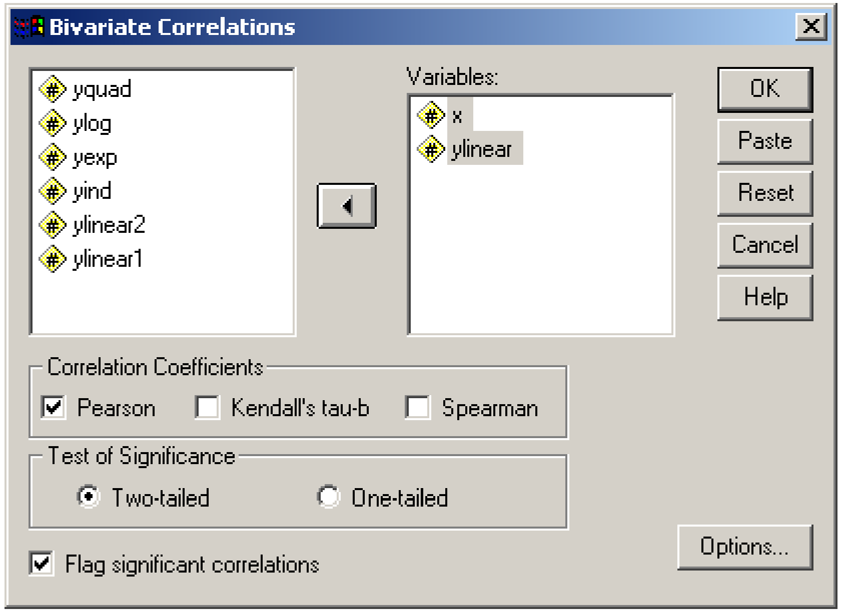


Segundo passo: calcular medidas de associação.

Último passo: realizar um teste de hipóteses para averiguar se os valores das medidas de associação observados nos dados são significativos, ou seja se podemos estatisticamente concluir a favor de uma associação na população.

Medidas de associação para dados numéricos ou ordinais

No SPSS os coeficientes de associação (correlação) para dados numéricos ou ordinais podem ser obtido através do menu `Analyse / Correalte / Bivariate`.



Neste menu podem-se seleccionar mais do que duas variáveis, caso em que o SPSS fornece uma tabela de correlações para todas as combinações de pares de variáveis. O SPSS fornece também o p-value dos testes ao significado dos coeficientes, para cada par de variáveis.

1 - O coeficiente de correlação de Pearson (*Pearson product-moment correlation coefficient*)

Dadas duas amostras de observações medidas numa escala de intervalos ou razões, podemos medir o grau de associação **linear** através da estatística

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

r pertence ao intervalo $[-1, 1]$. Se $r = 1$ temos uma recta perfeita com declive positivo. Se $r = -1$ temos uma recta perfeita com declive negativo. Se as variáveis são independentes $r \simeq 0$.

Uma interpretação usual: r^2 mede a percentagem de variabilidade de uma das variáveis explicada pela outra.

Podemos testar se duas variáveis são correlacionadas através das hipóteses:

$$H_0: \ \rho = 0 \quad vs \quad H_1: \ \rho \neq 0$$

onde ρ representa o coeficiente de correlação da população onde foram retirados os dados.

Pressupostos do teste

1. ambas as populações de onde foram retirados as amostras têm distribuição Normal,
2. a relação entre as variáveis é de forma linear, caso exista.

No SPSS o coeficiente de Pearson pode ser obtido através do menu **`Analyse / Correalte / Bivariate`**.

2 - O coeficiente de correlação de Spearman (*Spearman rank-order coefficient*)

Aplica-se a duas variáveis medidas apenas numa escala ordinal, ou que apresentam uma relação não linear mas monótona (se uma aumenta a outra tem sempre tendência a aumentar (ou a diminuir)). Aplica-se ainda quando não são satisfeitos os requisitos to teste ao coeficiente de Pearson (variáveis não Normais).

Dadas duas amostras de observação ordenáveis, substitui-se cada um dos seus valores pela sua ordem de ordenação, em inglês *rank*. O coeficiente de Spearman não é mais do que o coeficiente de Pearson aplicado aos *ranks*.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

onde d_i representa a diferença de *ranks* correspondentes a cada par de observações x_i, y_i.

Tal como no caso do coeficiente de Pearson é possível testar as hipóteses

$$H_0:\ \rho = 0 \quad vs \quad H_1:\ \rho \neq 0.$$

Tal como para o coeficiente de Pearson, no SPSS o coeficiente de Spearman pode ser obtido através do menu `Analyse / Correalte / Bivariate`.

3- O coeficiente de correlação de Kendall

Uma alternativa ao coeficiente de Spearman é o coeficiente de Kendall (*Kendall's tau coeffficient*) que se aplica nas mesmas condições.

Uma diferença muito importante entre os dois coeficientes (Kendall e Spearman) reside na sua interpretação e na impossibilidade de comparar directamente valores provenientes de ambos. Embora o objectivo comum seja o de medir associação, a forma de o fazer é distinta.

O coeficiente de Kendall é muitas vezes descrito como uma medida de concordância entre dois conjuntos de classificações relativas a um conjunto de objectos ou experiências.

$$T = \frac{\#\text{concordâncias} - \#\text{discordâncias}}{\text{número total de pares}}$$

Tal como para os coeficientes de Pearson e Spearman é possível efectuar um teste de hipóteses para averiguar se a associação é significativa.

No SPSS o coeficiente de Kendall pode ser obtido através do menu **`Analyse / Correalte / Bivariate`**.

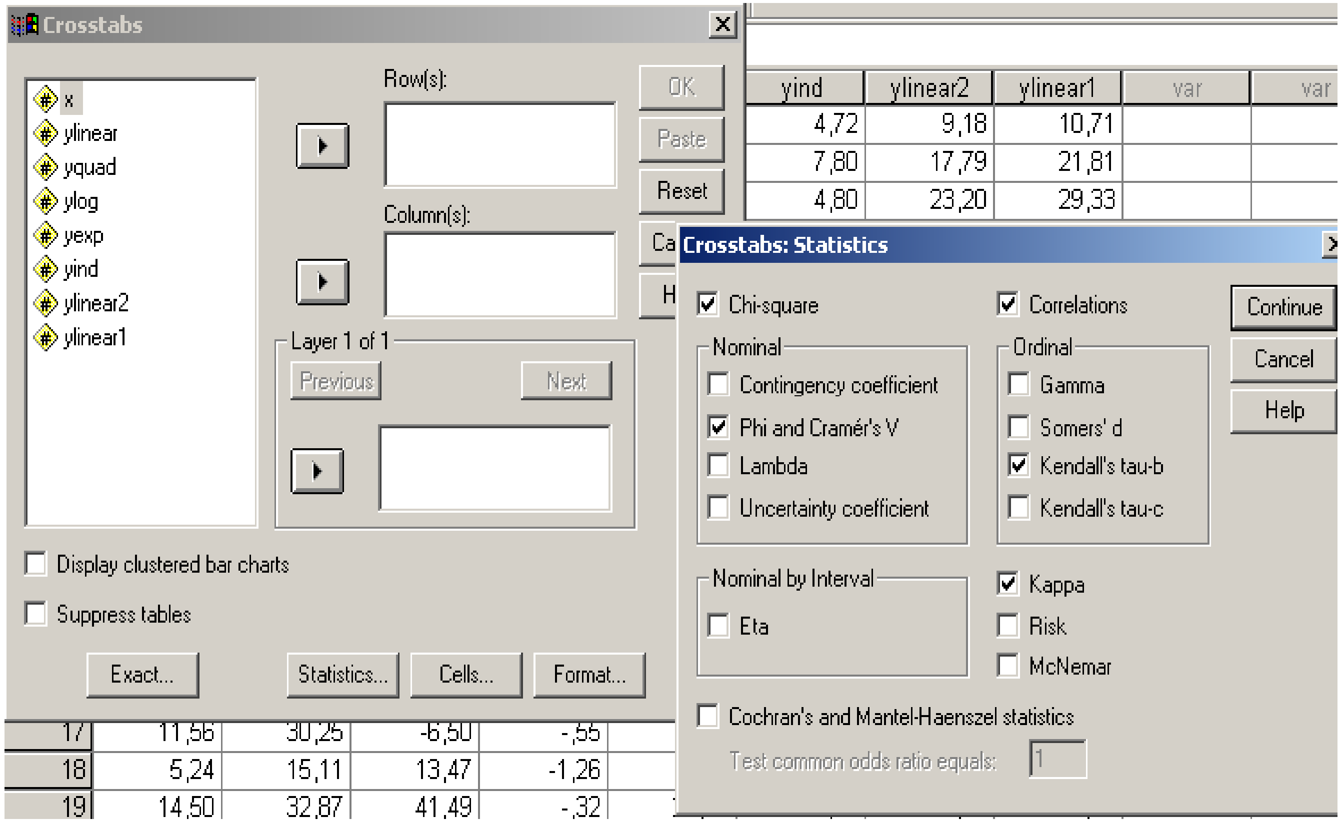
Medidas de associação para dados categóricos

Dados apresentados em **tabelas de contingência**. Por exemplo:

	Patologia		
Sexo	Presente	Ausente	Total
Feminino	30	20	50
Masculino	15	35	50
Total	45	55	100

As medidas de associação e respectivos testes de hipóteses para dados organizados em tabelas de contingência estão disponíveis no SPSS através do menu `Analyze / Descriptive Statistics / Crosstabs`.

Primeiramente há que introduzir os dados da tabela de contingência e seleccionar o menu `Data / Weight cases` por forma a atribuir pesos correspondentes às frequências observadas para cada célula.


Crosstabs
x
ylinear
yquad
ylog
yexp
yind
ylinear2
ylinear1
Row(s):
Column(s):
Layer 1 of 1
Previous
Next
OK
Paste
Reset
Display clustered bar charts
Suppress tables
Exact...
Statistics...
Cells...
Format...
yind
ylinear2
ylinear1
var
var
Crosstabs: Statistics
Chi-square
Correlations
Continue
Cancel
Help
Nominal
Contingency coefficient
Phi and Cramér's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals:

1- O teste do χ^2

$$H_0: \text{ as variáveis são independentes} \quad vs \quad H_1: \text{ as variáveis são dependentes.}$$

Estatística de teste:

$$X^2 = \sum_{\text{todas as células}} \frac{(O_i - E_i)^2}{E_i},$$

onde E_i representa a frequência esperada e O_i a observada.

Quando o número de observações é elevado a distribuição da estatística X^2 é aproximadamente a do χ^2 e daí o nome do teste.

Rejeita-se a hipótese de independência entre as variáveis quando o valor da estatística de teste é superior a um certo valor crítico (reflectindo grandes desvios entre as frequências observadas e esperadas).

Resumindo:

O teste do χ^2 aplica-se sempre que quisermos averiguar a existência de dependência entre duas variáveis de tipo categórico.

Requisitos do teste: As frequências esperadas em cada classe não devem ser inferiores a 5 unidades sempre que o número total de observações é $n \leq 20$. Se $n > 20$ não deverá existir mais do que 20% das células com frequências esperadas inferiores a 5 nem deverá existir nenhuma célula com frequência esperada inferior a 1.

Inconvenientes do teste:

1. Uma vez que a distribuição da estatística de teste é apenas aproximada (assintótica), para amostras pequenas o valor do *p-value* poderá conter um erro apreciável. No caso de tabelas 2×2 e sempre que $n \leq 20$ deve-se recorrer ao **teste de Fisher** que fornece valores exactos para os *p-values* do teste.

2. Devido à natureza discreta da contagem das frequências o valor da estatística do χ^2 vem acrescida de um erro. No caso de tabelas 2×2 deve-se utilizar uma **correcção à continuidade** (fornecida pelo SPSS).

Inconvenientes da estatística do χ^2 enquanto medida de associação

A estatística X^2 utilizada no teste do χ^2 é uma medida de associação entre duas variáveis já que assume valores próximos de zero quando as variáveis são independentes e valores elevados (positivos) quando existe dependência. No entanto, ao contrário do que acontecia com os coeficientes de assimetria, esta medida não está limitada ao intervalo $[0, 1]$ e o seu valor máximo depende do número total de observações.

Coeficientes de associação para dados categóricos que se assemelham aos coeficientes de correlação:

1 - O coeficiente de Cramér

O coeficiente de Cramér é uma medida de associação entre duas variáveis medidas numa escala categórica. Portanto pode ser aplicado em situações onde a informação se encontra distribuída por categorias nominais não ordenáveis.

$$C = \sqrt{\frac{X^2}{n(l-1)}}$$

onde n representa o número total de observações, l representa o mínimo entre o número de linhas e colunas da tabela de contingência, e X^2 é o valor da estatística do teste de χ^2.

A partir do valor do coeficiente de Cramér também é possível efectuar um teste às hipóteses

$$H_0: \text{ as variáveis são independentes} \quad vs \quad H_1: \text{ as variáveis são dependentes.}$$

Vantagens do coeficiente de Cramér:

o seu valor está limitado ao intervalo $[0, 1]$.

quando as variáveis são totalmente independentes $C = 0$.

quanto maior a associação maior o valor do coeficiente.

o coeficiente pode ser determinado em situações onde mais nenhum coeficiente (dos já expostos) pode ser aplicado.

ao contrário da estatística X^2, o coeficiente pode ser aplicado para comparar tabelas de contingência de dimensão diferente ou baseadas em amostras de dimensão diferente.

Desvantagens do coeficiente:

quando $C = 1$ pode não haver associação perfeita entre as duas variáveis. A associação só é perfeita se o número de linhas for igual ao número de colunas.

o coeficiente de Cramér está sujeito aos mesmos pressupostos do teste do qui-quadrado se pretendermos testar o seu significado.

este coeficiente não deve ser comparado directamente com outros. Se os dados forem ordinais podemos calcular o coeficiente de Cramér mas não devemos comparar directamente o seu valor com o valor do coeficiente de Pearson. Embora o coeficiente aumente com o grau de associação as diferenças na magnitude não têm uma interpretação directa.

2 - O coeficiente Φ

O coeficiente Φ é muito semelhante ao coeficiente de Cramér e foi proposto inicialmente apenas para tabelas de contingência 2×2. Neste caso o teste de independência que se pode efectuar pode ser baseado no teste exacto de Fisher fornecendo valores mais exactos que os do coeficiente de Cramér. Para tabelas 2×2 com conteúdo representado pelas letras $\begin{array}{cc} \mathrm{A} & \mathrm{B} \\ \mathrm{C} & \mathrm{D} \end{array}$ o coeficiente é dado por

$$R_\phi = \frac{|AD - BC|}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Se retirarmos o módulo do numerador obtemos um coeficiente que pode assumir valores negativos detectando assim um sentido na associação entre as duas variáveis.

No que respeita a vantagens e desvantagens do coeficiente, elas são idênticas às do coeficiente de Cramér.