Medidas de Concordância para Variáveis Categorizadas

DISSERTAÇÃO DE MESTRADO

Fábia Filipa Nunes Camacho
MESTRADO EM MATEMÁTICA

UNIVERSIDADE da MADEIRA
A Nossa Universidade
www.uma.pt

setembro | 2013

Medidas de Concordância para Variáveis Categorizadas

DISSERTAÇÃO DE MESTRADO

Fábia Filipa Nunes Camacho
MESTRADO EM MATEMÁTICA

ORIENTAÇÃO
Rita Maria César e Sá Fernandes de Vasconcelos

Medidas de Concordância para Variáveis Categorizadas

DISSERTAÇÃO DE MESTRADO

Fábia Filipa Nunes Camacho

MESTRADO EM MATEMÁTICA

JÚRI:

Ana Maria Cortesão Pais Figueira da Silva Abreu
Sílvio Filipe Velosa
Rita Maria César e Sá Fernandes de Vasconcelos

ORIENTAÇÃO:
Rita Maria César e Sá Fernandes de Vasconcelos

Medidas de Concordância para Variáveis Categorizadas

Fábia Filipa Nunes Camacho

Setembro 2013

Agradecimentos

Com a conclusão deste trabalho e deste mestrado, concluo também uma fase da minha vida da qual muito me orgulho. Existiram altos e baixos e nem sempre foi fácil, mas são as dificuldades que nos tornam mais fortes e que nos fazem dar mais valor às nossas conquistas.

O meu primeiro e maior agradecimento devo sem qualquer dúvida aos meus pais que sempre fizeram tudo o que podiam, e se calhar por vezes até o que não podiam, para me proporcionarem a mim (e posteriormente ao meu irmão) uma boa vida, uma boa educação, uma infância feliz e cheia de amor. Sempre fizeram de tudo para que nunca, nem por um momento, nós pudéssemos duvidar de que o amor que sentem por nós é infinito e inesgotável. Sempre me deram força, ânimo e apoio para que eu alcançasse os meus sonhos e sempre fizeram questão de demonstrar o orgulho que têm em mim. Sem dúvida que sou o que sou hoje devido a eles, ao grande exemplo que me deram como pessoas, dos valores que desde cedo me incutiram tais como honestidade, empenho e força de vontade.

Também devo um grande obrigado à Prof. Dr.ª Rita Vasconcelos, minha orientadora nesta tese, cuja ajuda foi inestimável e a qual muito agradeço. Sem a sua preciosa ajuda, esta tese não teria o nível de qualidade alcançado. Considero-a um grande exemplo de profissional e uma pessoa excecional, pois nunca recusa ajudar quem a si recorre.

Ao meu maninho Tiago, que por ser ainda uma criança não entende a importância da conclusão de uma tese de mestrado e por isso sempre preferiu "boicotar" os meus momentos de estudo com brincadeiras, gargalhadas e carinhos. Mas o seu carinho e a sua admiração por mim sempre me deram muita força.

Às amigas que fiz nesta Universidade, Helena, Eva, Alexandra, Carla, Graça, Fernanda e Letícia. Convosco partilhei este percurso, juntas trabalhámos e estudámos mas também rimos, brincámos e nos divertimos como só nós sabemos. Fomos colegas de curso, colegas de estudo e de trabalho, mas também amigas e confidentes umas das outras, e por vezes até "psicólogas". Nunca deixamos de nos apoiar umas às outras e mesmo nos momentos menos bons, nunca deixei de sentir o vosso apoio.

Às amigas que já trazia, Ilda e Ana Claudia, companheiras de muitos momentos de diversão mas também de momentos de desânimo quando estes teimaram em surgir.

Partilhámos este percurso desde o momento em que soubemos ter entrado na Universidade e partilhámos também todas as emoções que esta experiência nos proporcionou. Mas sem nunca deixarmos que a distância quebrasse o elo de amizade que nos ligava.

Não posso também deixar de agradecer aos professores e colegas de licenciatura e de mestrado que contribuíram para a minha aprendizagem e que me transmitiram conhecimentos de um valor incalculável.

Agradeço também a todas as pessoas que passaram pela minha vida durante esta fase e que de alguma forma contribuíram para o meu sucesso ou que de alguma maneira me ajudaram, me ensinaram e me instruíram.

A todos os anteriormente mencionados, mais uma vez: MUITO OBRIGADO!

Resumo

Apesar desta tese se propor abordar exclusivamente variáveis categorizadas e medidas que descrevem a relação entre estas variáveis, considerámos extremamente relevante estender ao caso em que uma das variáveis em estudo é uma variável contínua. Tais medidas, quando usadas corretamente, fornecem uma descrição útil da estrutura implícita numa tabela de contingência.

Ao longo dos anos, muitos cientistas deduziram medidas de associação de acordo com o que se propunham avaliar. Perante um conjunto tão variado de medidas, o nosso objetivo foi sumarizá-las, classificá-las em grupos mais gerais e estabelecer em que situações é que a sua utilização é indicada.

Apesar de estarmos interessados em avaliar a concordância, realizámos que era essencial incluir nesta tese os testes de ajustamento e a análise de tabelas de contingência.

Palavras chave: testes de ajustamento, tabelas de contingência, associação de variáveis aleatórias, concordância, estimadores de máxima verosimilhança.

Abstract

Although our purpose in this thesis was to deal exclusively with categorical variables and summary numbers that describe relationships between cross-classified variables, we understand that it was of great relevance to include the case when one of the variables is a continuos variable. When properly used, they provide useful description of the structure displayed by a contigency table.

Over the years, many scientists have developed association measures for their needs. With so many different measures, our aim was to summarize them, sort them into general groups and establish when the calculation of each measure is indicated.

We were focused on evaluate the agreement, we realized that it was essential to include goodness-of-fit tests and contingency tables analysis.

Key-words: goodness-of-fit test, contigency tables, association between random variables, agreement, maximum likelihood estimators.

Índice

Capítulo 1

Testes de ajustamento para dados categorizados

1.1 Introdução

A razão da inclusão nesta tese de um capítulo sobre os testes de ajustamento, consiste no facto de um teste de ajustamento poder ser considerado um teste à concordância entre um modelo teórico (contemplado na hipótese nula) e um modelo empírico.

Podemos reduzir muitos dos testes de ajustamento ao teste da hipótese relativa aos parâmetros $\mathbf{\Pi} = (\pi_1, \ldots, \pi_k)$ de uma distribuição multinomial:

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{n!}{\pi_1! \ldots \pi_k!} \pi_1^{x_1} \ldots \pi_k^{x_k}$$

onde $\mathbf{\Pi}$ representa o vetor das probabilidades e os $x's$ são inteiros não negativos cuja soma é n. A variável aleatória multinomial $\mathbf{X}$ é muitas vezes baseada em $X_i, i = 1, \ldots, k$ sendo X_i o número de v.as. $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, independentes e identicamente distribuídas (i.i.d.) com $F(y; \Theta)$, pertencentes à classe C_i, onde $\{C_i; i = 1, \ldots, k\}$ é um conjunto de classes mutuamente exclusivas o qual esgota o conteúdo de probabilidade de F.

Num teste de ajustamento para dados categorizados queremos testar a hipótese nula

$$H_0 : \mathbf{\Pi} = \mathbf{\Pi}_0$$

onde $\mathbf{\Pi}_0 = (\pi_{0_i}, \ldots, \pi_{0_k})$ é o vetor de probabilidade pré-especificado.

Tendo em conta as expressões usualmente encontradas na literatura, utilizaremos frequentemente O_i para representar X_i, a frequência absoluta observada na classe $C_i, i = 1, \ldots, k$. Utilizaremos E_i para representar $n\pi_{0_i}$, a frequência esperada, sob H_0, na classe $C_i, i = 1, \ldots, k$.

Possivelmente a estatística mais utilizada é o X^2 de Pearson:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}. \tag{1.1}$$

Note-se que, sob H_0, cada O_i é uma v.a. Binomial $B(n, \pi_{0_i})$ com valor médio $\mu_i = n\pi_{0i}$, pelo que pode ser aproximada por uma v.a. de Poisson (μ_i). O Teorema do Limite Central contempla a aproximação da v.a. de Poisson pela v.a. Normal. Como o quadrado da v.a. Normal estandardizada é uma v.a. Qui-quadrado com um grau de liberdade, $\chi^2(1)$, podemos considerar estabelecer que, sob H_0 a estatística X^2 de Pearson tem uma distribuição aproximada da distribuição de uma v.a. $\chi^2(k-1)$.

1.2 Família de estatísticas do Qui-quadrado

As estatísticas mais frequentemente usadas em testes de ajustamento de frequências observadas (de Pearson, logaritmo da razão das verosimilhanças, logaritmo da razão das verosimilhanças transformado, etc.), são casos especiais da família de estatísticas

$$\left\{I^\lambda; \lambda \in \mathbb{R}\right\}$$

em que

$$2nI^\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} X_i \left\{ \left(\frac{X_i}{E_i}\right)^\lambda - 1 \right\}; \lambda \in \mathbb{R} \tag{1.2}$$

onde $\{X_i, i = 1, ..., k\}$ representam as frequências observadas e $\{E_i, i = 1, ..., k\}$ as frequências esperadas. $2nI^\lambda$ é definida por continuidade para $\lambda = 0, -1$

$$2nI^0 = \lim_{\lambda \to 0} 2nI^\lambda = 2\sum_{i=1}^{k} X_i \ln\left(\frac{X_i}{E_i}\right) \tag{1.3}$$

$$2nI^{-1} = \lim_{\lambda \to -1} 2nI^\lambda = 2\sum_{i=1}^{k} E_i \ln\left(\frac{E_i}{X_i}\right) \tag{1.4}$$

Da definição de $2nI^\lambda$ facilmente se vê que:

- quando $\lambda = 1$ obtém-se a estatística $X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ de Pearson;
- quando $\lambda = -\frac{1}{2}$, obtém-se a estatística de Freeman-Tukey $T^2 = 4\sum_{i=1}^{k} \left(\sqrt{O_i} - \sqrt{E_i}\right)^2$;
- quando $\lambda = -2$ obtém-se a estatística de Neyman modificada $NM^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{O_i}$;
- quando $\lambda = 0$ obtém-se a estatística $G^2 = 2\sum_{i=1}^{k} O_i \ln\left(\frac{O_i}{E_i}\right)$ do logaritmo da razão das verosimilhanças;
- quando $\lambda = -1$ obtém-se a estatística de Kullback do logaritmo da razão das verosimilhanças modificada $GH^2 = 2\sum_{i=1}^{k} E_i \ln\left(\frac{E_i}{O_i}\right)$.

Todas estas estatísticas tentam avaliar, de formas diferentes, o grau de afastamento entre as frequências observadas e as frequências esperadas. Note-se que usualmente nos referimos a estes afastamentos como discrepâncias, porque I^λ não é, gralmente, uma função distância. Todas estas estatísticas têm assintoticamente uma distribuição Qui-Quadrado com $k-1$ graus de liberdade, uma vez que:

$\triangleright$ se $F_E(.)$ é a função de distribuição exata de $2nI^\lambda$ para λ fixo e se $F_{\chi^2(k-1)}(.)$ é a função de distribuição da v.a. $\chi^2(k-1)$, então $F_E(t) = F_{\chi^2(k-1)}(t) + o(1)$ quando $n \to \infty$ e $\forall t$, sob H_0 e com k fixo.

Os cálculos que seguidamente apresentamos, mostram que este resultado é válido para qualquer valor do parâmetro λ.

Consideremos o teste:
$H_0 : \pi = \pi_0$, em que as probabilidades $\{\pi_{0_i}\}$ são completamente especificadas.
Para $\lambda \neq 0, -1$:

$$
\begin{aligned}
2nI^\lambda &= \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} X_i \left\{ \left(\frac{X_i}{E_i} \right)^\lambda - 1 \right\} = \\
&= \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} X_i \left\{ \left(\frac{X_i}{n\pi_{0_i}} \right)^\lambda - 1 \right\} = \\
&= \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} \left(\frac{X_i^{\lambda+1}}{(n\pi_{0_i})^\lambda} - X_i \right) = \\
&= \frac{2}{\lambda(\lambda+1)} \left[\sum_{i=1}^{k} \frac{X_i^{\lambda+1}}{(n\pi_{0_i})^\lambda} - n \right] = \\
&= \frac{2}{\lambda(\lambda+1)} \left[\sum_{i=1}^{k} n\pi_{0_i} \frac{X_i^{\lambda+1}}{(n\pi_{0_i})^{\lambda+1}} - n \right] = \\
&= \frac{2n}{\lambda(\lambda+1)} \left[\sum_{i=1}^{k} \pi_{0_i} \frac{X_i^{\lambda+1}}{(n\pi_{0_i})^{\lambda+1}} - 1 \right] = \\
&= \frac{2n}{\lambda(\lambda+1)} \left[\sum_{i=1}^{k} \pi_{0_i} \frac{X_i^{\lambda+1}}{(n\pi_{0_i})^{\lambda+1}} - \sum_{i=1}^{k} \pi_{0_i} \right] = \\
&= \frac{2n}{\lambda(\lambda+1)} \sum_{i=1}^{k} \pi_{0_i} \left(\frac{X_i^{\lambda+1}}{(n\pi_{0_i})^{\lambda+1}} - 1 \right) = \\
&= \frac{2n}{\lambda(\lambda+1)} \sum_{i=1}^{k} \pi_{0_i} \left[\left(1 + \frac{X_i - n\pi_{0_i}}{n\pi_{0_i}} \right)^{\lambda+1} - 1 \right]
\end{aligned}
\tag{1.5}
$$

Escrevendo

$$
V_i = \frac{X_i - n\pi_{0_i}}{n\pi_{0_i}} \tag{1.6}
$$

e expandindo em série de Taylor, vê-se que

$$
\begin{aligned}
2nI^{\lambda} &= \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{k}\pi_{0_i}\left[(\lambda+1)V_i+\frac{\lambda(\lambda+1)}{2}V_i^2+O\left(V_i^3\right)\right]= \qquad (1.7)\\
&= 2n\sum_{i=1}^{k}\left[\frac{(\lambda+1)}{\lambda(\lambda+1)}V_i\pi_{0_i}+\frac{\lambda(\lambda+1)}{2\lambda(\lambda+1)}V_i^2\pi_{0_i}+O\left(V_i^3\right)\right]=\\
&= 2n\sum_{i=1}^{k}\left[\frac{1}{\lambda}\frac{X_i-n\pi_{0_i}}{n\pi_{0_i}}\pi_{0_i}+\frac{V_i^2}{2}\pi_{0_i}+O\left(V_i^3\right)\right]=\\
&= 2n\left[\left(\sum_{i=1}^{k}\frac{1}{\lambda}\frac{X_i-n\pi_{0_i}}{n\pi_{0_i}}\pi_{0_i}\right)+\left(\sum_{i=1}^{k}\frac{V_i^2}{2}\pi_{0_i}+O\left(V_i^3\right)\right)\right]=\\
&= 2n\left[\frac{1}{\lambda}\sum_{i=1}^{k}\left(\frac{X_i}{n}-\pi_{0_i}\right)+\left(\sum_{i=1}^{k}\frac{V_i^2}{2}\pi_{0_i}+O\left(V_i^3\right)\right)\right]=\\
&= 2n\left[\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^{k}X_i-\sum_{i=1}^{k}\pi_{0_i}\right)+\left(\sum_{i=1}^{k}\frac{V_i^2}{2}\pi_{0_i}+O\left(V_i^3\right)\right)\right]=\\
&= 2n\left[\frac{1}{\lambda}\left(\frac{n}{n}-1\right)+\left(\sum_{i=1}^{k}\frac{V_i^2}{2}\pi_{0_i}+O\left(V_i^3\right)\right)\right]=\\
&= 2n\left[\sum_{i=1}^{k}\pi_{0_i}\frac{V_i^2}{2}+o\left(\frac{1}{n}\right)\right]
\end{aligned}
$$

sob o modelo H_0. Um resultado idêntico é válido para $\lambda = 0, -1$, também através duma expansão em série de Taylor [6].

Logo,

$$2nI^{\lambda} = 2nI^{1} + o(1)\,; \lambda \in \mathbb{R} \qquad (1.8)$$

Assim, cada uma daquelas estatísticas tem, assintoticamente, a mesma distribuição que a estatística X^2 de Pearson que sabemos ter assintoticamente a distribuição χ^2_{k-1} sob H_0.

Este resultado também se estende ao caso mais geral em que existem parâmetros não especificados em H_0 (Bishop, Fienberg and Holland, 1975, Cap. 14)[5].

Consideremos um vetor de parâmetros $\Theta = (\theta_1, \ldots, \theta_s) \in \mathbb{R}^s, s < k-1$, e uma função

$$\mathbf{f} : \mathbb{R}^s \to \Delta_k = \left\{ P = (P_1, \ldots, P_k) : P_1 \geq 0, i = 1, \ldots, k; \sum_{i=1}^{k} P_i = 1 \right\} \qquad (1.9)$$

tal que a cada vetor de parâmetros Θ corresponde um vetor de probabilidades $\Pi = (\Pi_1, \ldots, \Pi_k)$. Tem-se, então, que as duas hipóteses nulas seguintes são equivalentes:

$$H_0 : \Theta \in Q_0 \quad \text{e} \quad H_0 : \Pi \in \mathbf{f}\left(Q_0\right).$$

Sob certas condições de regularidade em $\mathbf{f}$ e em Q_0, e considerando um vetor $\widehat{\Theta}$ o melhor estimador assintoticamente Normal de $\Theta = (\theta_1, \ldots, \theta_s)$ e $\widehat{\Pi} = \mathbf{f}\left(\widehat{\Theta}\right)$, então, sob H_0, $2nI^{\lambda}$ converge em distribuição para uma v.a. $\chi^2\left(k - s - 1\right)$ quando $n \to \infty$.

Prova-se que as estatísticas da Família de estatísticas do Qui-Quadrado não são assintoticamente equivalentes sob $H_0 : \pi_{O_i} = \frac{1}{k}, i = 1, \ldots, k$.

Por exemplo, a estatística X^2 de Pearson $(\lambda = 1)$ tem $\mu_n \approx k$ e $\sigma_n^2 \approx 2k$, enquanto que a estatística G^2 do logaritmo da razão das verosimilhanças $(\lambda = 0)$ tem:

$$\mu_n \simeq 2k \left[\sum_{j=1}^{\infty} \ln\left(j\right).\frac{e^{-a}a^j}{(j-1)!} - \ln\left(a\right)\right]$$

e σ_n^2 também é diferente.

É esta sensibilidade a diferentes valores de λ que levou a explorar esta família de estatísticas do Qui-Quadrado, segundo vários critérios.

Qualquer decisão sobre qual das estatísticas deve ser usada num teste de ajustamento depende das respectivas performances de acordo com vários critérios.

Potência para amostras finitas:

O critério mais importante para comparar testes, nomeadamente a função potência, é para amostras finitas, muitas vezes, matematicamente intratável. No entanto, é acessível em computador para os testes baseados na Família de estatísticas do Qui-Quadrado. Na tabela seguinte foi calculada a potência exata dos testes $H_0 : \pi_{0_i} = \frac{1}{k}, i = 1, \ldots, k$.

$$H_1 : \pi_{1i} = \begin{cases} \lfloor 1 - \delta\left(k-1\right)\rfloor / k \ , \ i = 1, \ldots, k-1 \\ \left(1 + \delta/k\right) \ , \ i = k \end{cases} \tag{1.10}$$

em que $-1 \leq \delta \leq k - 1$ é fixo. Esta alternativa resulta de uma perturbação $\frac{\delta}{k}$ na k-ésima probabilidade enquanto que as restantes são ajustadas por forma à soma ser igual a 1.

Para $\delta > 0$ a potência exata aumenta com o λ. Para $\delta < 0$ a potência exata diminui com o λ. Quando a hipótese alternativa é deste tipo, é sempre possível aumentar a potência do teste escolhendo outros membros da Família de estatísticas do Qui-quadrado. Contudo, quando $|\lambda|$ aumenta, existe um limiar evidente a partir do qual a função potência pouco se altera conforme o λ varia.

λ	$\delta = 1.5$	$\delta = 0.5$	$\delta = -0.9$
-5.0	0.6316	0.1228	0.7434
-2.0	0.6500	0.1231	0.7434
-1.0	0.7960	0.1348	0.7342
-0.5	0.8009	0.1412	0.7263
-0.3	0.8525	0.1538	0.7108
0.0	0.8640	0.1567	0.7045
0.3	0.8640	0.1567	0.7045
0.5	0.8640	0.1567	0.7045
0.7	0.7640	0.1567	0.7045
1.0	0.8745	0.1629	0.5150
1.5	0.8855	0.1682	0.3844
2.0	0.8962	0.1725	0.3291
2.5	0.8982	0.1733	0.2780
5.0	0.9025	0.1743	0.2422

Potência exata para n=20, k=4; alternativas definidas por 1.10; nível de significância de 0.05.

Capítulo 2

Tabelas de contingência

Ao recolhermos observações de variáveis discretas, cujos possíveis valores representam categorias, estamos na área dos dados categorizados. O mesmo acontece se a variável em estudo for contínua e recorrermos a uma classificação, segundo um ou mais critérios, e contarmos o número de observações em cada uma das classes consideradas.

Se os nossos dados categorizados forem obtidos a partir do estudo simultâneo de duas ou mais variáveis estamos na área das tabelas de contingência.

O objetivo da recolha de dados categorizados consiste na análise da relação entre as variáveis categorizadas em estudo.

2.1 Tabelas de contingência bidimensionais

No caso de uma amostra classificada segundo duas variáveis qualitativas, os dados respeitantes às contagens são então apresentados numa tabela de contingência em que as linhas $L_1, L_2, ..., L_r$ dizem respeito às r categorias ou classes consideradas na primeira variável, seguidas por uma linha marginal de totais, e as colunas $C_1, C_2, ..., C_c$ dizem respeito às categorias ou classes da segunda variável sendo também seguidas de uma coluna marginal de totais.

2.1.1 Testes do Qui-quadrado

Existem dois testes do Qui-quadrado que podem ser aplicados a tabelas $r \times c$ e, num caso particular destas, a tabelas 2×2: o teste de independência e o teste de homogeneidade que embora testem hipóteses diferentes têm cálculos semelhantes. Estes testes são baseados nos seguintes pressupostos:

1. os dados são categorizados para $r \times c$ categorias mutuamente exclusivas;
2. os dados testados representam uma amostra aleatória de n observações e cada elemento da amostra só poderá aparecer numa única célula da tabela de contingência.

Independência

Um teste de independência visa testar a hipótese nula de que a variável cujas categorias estão apresentadas nas linhas de uma tabela e a variável cujas categorias estão apresentadas nas colunas de uma tabela de contingência não estão relacionadas, ou seja, testa a hipótese de independência entre ambas as variáveis aleatórias:

$$H_0 : p_{ij} = p_{i.} \times p_{.j}, \forall_{i,j}$$
$$H_1 : \exists_{i,j} : p_{ij} \neq p_{i.} \times p_{.j}$$

em que p_{ij} representa a probabilidade de um elemento pertencer simultaneamente à categoria i da variável representada nas linhas da tabela e à categoria j da variável representada nas colunas da tabela. $p_{i.}$ representa a probabilidade marginal de um elemento ser classificado na categoria i da primeira variável referida anteriormente independentemente da categoria da segunda variável à qual pertença. $p_{.j}$ é a probabilidade marginal de um elemento ser classificado na categoria j da segunda variável referida anteriormente independentemente da categoria da primeira variável à qual pertença.

Uma outra forma de dizer que as variáveis são independentes é dizer que não existe uma forma de predizer a que categoria um elemento poderá pertencer, quando se sabe a que categoria pertence para a outra variável.

A estatística de teste é dada através da fórmula:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.1}$$

que tem aproximadamente a distribuição de um χ^2com $(r-1)(c-1)$ graus de liberdade, quando a tabela de contingência é $r \times c$. O_{ij} representa o número de elementos da amostra que pertencem simultaneamente à categoria i da variável representada nas linhas da tabela e à categoria j da variável representada nas colunas da tabela e E_{ij} representa o número esperado de elementos classificados na célula (i, j), sob H_0. Se as variáveis em questão forem independentes, a frequência esperada para uma célula (i, j) será dada por:

$$E_{i,j} = n\widehat{p}_{i.} \times \widehat{p}_{.j} = n\frac{n_{i.}}{n} \times \frac{n_{.j}}{n} = \frac{n_{i.} \times n_{.j}}{n} \tag{2.2}$$

sendo $n_{i.}$ e $n_{.j}$ os totais marginais da linha i e da coluna j observados respetivamente; n representa o número total de elementos da amostra. Os estimadores de máxima verosimilhança de $p_{i.}$ e $p_{.j}$ são dados respetivamente por $\frac{n_{i.}}{n}$ e $\frac{n_{.j}}{n}$.

A aproximação da distribuição da estatística de teste X^2, quando a hipótese de independência é verdadeira, à distribuição Qui-quadrado, é válida para o caso de as frequências esperadas não serem demasiado pequenas. Este termo, um pouco vago, tem sido interpretado como as frequências esperadas serem iguais ou superiores a 5. Esta restrição tem sido considerada arbitrária por alguns autores e referir-nos-emos a ela na secção 2.1.3.

Se temos uma tabela de contingência 2×2 pode aplicar-se a correção de Yates para a continuidade (a distribuição Qui-quadrado, contínua, está a ser usada como uma

aproximação a uma distribuição de uma v.a. discreta, nomeadamente distribuição multinomial), com o objetivo de melhorar a aproximação da distribuição de X^2 à distribuição χ^2:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}. \tag{2.3}$$

Se a dimensão da amostra for grande esta correção tem pouco efeito no valor da estatística do Qui-quadrado.

A região de rejeição é dada por $X^2 \geq \chi^2_\alpha$ com $(r-1)(c-1)$ graus de liberdade.

Um dos pressupostos deste teste é o de que não existe uma determinação antes do início do estudo dos números totais marginais de dados a serem estudados, seja para a variável cujas categorias são representadas pelas linhas da tabela, seja para a variável cujas categorias são representadas pelas colunas da tabela.

Exemplo[26]: Um investigador quer determinar se existe relação entre o tipo de personalidade (introvertido, extrovertido) e a afiliação partidária. Duzentas pessoas foram recrutadas para este estudo e a cada uma delas foi feito um teste de personalidade. Depois, cada uma dessas pessoas indica qual a sua afiliação partidária (Democrata, Republicano). O número de indivíduos em cada categoria não foi determinado à partida, sendo que os dados são totalmente aleatórios. Os dados relativos a este estudo encontram-se na tabela seguinte:

		Afiliação Partidária		
		Democrata	**Republicano**	**Total**
Personalidade	**Introvertido**	30	70	100
	Extrovertido	60	40	100
	Total	90	110	200

H_0 : As variáveis "Personalidade"e "Afiliação Partidária"são independentes, isto é, $p_{ij} = p_{i.} \times p_{.j}, \forall_{i,j}$, $i, j = 1, 2$

H_1 : As variáveis "Personalidade"e "Afiliação Partidária"não são independentes, isto é, $\exists_{i,j} : p_{ij} \neq p_{i.} \times p_{.j}$, $i, j = 1, 2$

As frequências esperadas são:

$$\begin{aligned}
E_{11} &= \frac{n_{1.} \times n_{.1}}{n} = \frac{100 \times 90}{200} = 45 \\
E_{12} &= \frac{n_{1.} \times n_{.2}}{n} = \frac{100 \times 110}{200} = 55 \\
E_{21} &= \frac{n_{2.} \times n_{.1}}{n} = \frac{100 \times 90}{200} = 45 \\
E_{22} &= \frac{n_{2.} \times n_{.2}}{n} = \frac{100 \times 110}{200} = 55
\end{aligned}$$

Visto tratar-se de uma tabela 2×2, vamos utilizar a fórmula com a correção de Yates:

$$\begin{aligned} X^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}} = \\ &= \frac{(|30-45|-0.5)^2}{45} + \frac{(|70-55|-0.5)^2}{55} + \frac{(|60-45|-0.5)^2}{45} + \frac{(|40-55|-0.5)^2}{55} = \\ &= 4.67 + 3.82 + 4.67 + 3.82 = 16.98 \end{aligned}$$

A região de rejeição é dada por $X^2 \geq \chi^2_{0.05}(1) = 3.841$. Como $16.98 > 3.841$ rejeitamos a hipótese de independência entre as variáveis.

Se as variáveis aleatórias não são independentes torna-se útil recorrer a algum tipo de medida que quantifique o grau de associação entre elas. Algumas destas medidas serão abordadas no Capítulo 3.

Homogeneidade

O que distingue um teste de homogeneidade de um teste de independência é que no primeiro caso os totais marginais de linha ou de coluna da tabela de contingência são fixos à partida, ou seja, ao ser-nos dada uma tabela de contingência esta possui uma das margens fixas anteriormente à recolha dos dados. Temos, assim, probabilidades condicionais em cada célula da tabela e não conjuntas como nos testes de independência.

No caso de serem os totais de linha fixos à partida, a nossa hipótese H_0 será então:

$$H_0 : p_{j|1} = p_{j|2} = ... = p_{j|r} \ , \forall j \ , \ j = 1, \ldots, c$$

ou seja, testamos a existência de homogeneidade das r populações em estudo em relação a cada categoria da variável representada pelas colunas da tabela. A hipótese alternativa será:

$$H_1 : \exists j, \exists i, k : p_{j|i} \neq p_{j|k} \ , \ j = 1, \ldots, c \ , \ i, k = 1, \ldots, l \text{ e } i \neq k.$$

De resto, a forma como se calcula o teste é igual aos cálculos para o teste de independência.

Exemplo [26]: Duzentas pessoas escolhidas aleatoriamente são sujeitas a um teste de inteligência para o qual dispõem de uma hora. 100 destas pessoas efetuam este teste expostas a um ruído contínuo sendo-lhes dito que este deve-se a uma avaria num gerador. As outras 100 fazem o teste sem nenhum tipo de ruído de fundo. À saída do teste, um idoso com um braço engessado pede a cada uma das pessoas ajuda para levar uma embalagem pesada até ao seu carro. Os resultados foram os seguintes:

		Ajuda ao idoso		
		Sim	**Não**	**Total**
Expostos ao ruído	**Sim**	30	70	100
	Não	60	40	100
Total		90	110	200

Pretende-se avaliar se a exposição ao ruído afetou o comportamento dos indivíduos em relação à ajuda que lhes foi posteriormente pedida.

H_0 : A proporção de pessoas que ajudaram o idoso é igual tanto no grupo de pessoas expostas ao ruído como no grupo de pessoas não expostas ao ruído.

H_1 : A proporção de pessoas que ajudaram o idoso é diferente no grupo de pessoas expostas ao ruído e no grupo de pessoas não expostas ao ruído.

Ou seja,

$H_0 : p_{j|1} = p_{j|2}$, $j = 1, 2$ vs $H_1 : \exists j : p_{j|1} \neq p_{j|2}$, $j = 1, 2$.

As frequências esperadas são:

$$\begin{aligned} E_{11} &= \frac{n_{1.} \times n_{.1}}{n} = \frac{100 \times 90}{200} = 45 \\ E_{12} &= \frac{n_{1.} \times n_{.2}}{n} = \frac{100 \times 110}{200} = 55 \\ E_{21} &= \frac{n_{2.} \times n_{.1}}{n} = \frac{100 \times 90}{200} = 45 \\ E_{22} &= \frac{n_{2.} \times n_{.2}}{n} = \frac{100 \times 110}{200} = 55 \end{aligned}$$

Visto tratar-se de uma tabela 2×2, vamos utilizar a fórmula com a correção de Yates:

$$\begin{aligned} X^2 &= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}} = \\ &= \frac{(|30 - 45| - 0.5)^2}{45} + \frac{(|70 - 55| - 0.5)^2}{55} + \frac{(|60 - 45| - 0.5)^2}{45} + \frac{(|40 - 55| - 0.5)^2}{55} = \\ &= 4.67 + 3.82 + 4.67 + 3.82 = 16.98 \end{aligned}$$

A região de rejeição é dada por $X^2 \geq \chi^2_{0.05}(1) = 3.841$. Como $16.98 > 3.841$ rejeitamos a hipótese de que existe semelhança na proporção de pessoas que ajudaram o idoso. Olhando para a tabela é possível ver que o número de pessoas não expostas ao ruído que ajudaram o idoso é o dobro do número de pessoas que o fizeram tendo sido expostas ao ruído.

2.1.2 Teste Exato de Fisher

No caso de uma tabela 2×2 com células com frequências esperadas pequenas, uma alternativa ao teste de independência do Qui-quadrado é o Teste Exato de Fisher. Consideremos a seguinte tabela de contingência:

	Var. B		
Var. A	**Cat. 1**	**Cat. 2**	**Total**
Cat. 1	n_{11}	n_{12}	$n_{1.}$
Cat. 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

A probabilidade de obter qualquer tabela com as frequências $n_{11}, n_{12}, n_{21}, n_{22}$ quando todos os totais marginais são os da tabela acima é dada por:

$$\frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n_{11}!n_{12}!n_{21}!n_{22}!n!} \tag{2.4}$$

Esta é a fórmula utilizada pelo teste de Fisher para determinar a probabilidade de se obter a tabela resultante dos nossos dados e as probabilidades de todas as tabelas com os mesmos totais marginais e com valores ainda menores em células já consideradas como tendo um valor baixo na nossa tabela.

Suponhamos que era n_{11} o valor muito baixo da tabela. Então as outras tabelas cujas probabilidades iríamos calcular seriam:

$P_{n_{11}-1}$		
$n_{11}-1$	$n_{12}+1$	$n_{1.}$
$n_{21}+1$	$n_{22}-1$	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

,

$P_{n_{11}-2}$		
$n_{11}-2$	$n_{12}+2$	$n_{1.}$
$n_{21}+2$	$n_{22}-2$	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

$\cdots$

$\cdots$

$\cdots$

P_0		
0	$n_{12}+n_{11}$	$n_{1.}$
$n_{21}+n_{11}$	$n_{22}-n_{11}$	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

e $P = P_{n_{11}} + P_{n_{11}-1} + P_{n_{11}-2} + \cdots + P_0$.

Região de rejeição da hipótese de independência: $P \leq \alpha$ (nível de significância).

Exemplo: A tabela seguinte apresenta a relação entre os resultados obtidos por ecografia hepática e a deteção de certas aletrações químicas. A questão de interesse subjacente a esta recolha de dados é avaliar a capacidade das imagens por ecografia hepática fornecerem indicação da necessidade de realização daquelas análises químicas.

Imagens do	**Alterações químicas**		
Fígado	**Existentes (+)**	**Não existentes (-)**	**Total**
Anomalia (+)	7	4	11
Negativo (-)	3	5	8
Total	10	9	19

H_0 : Os resultados da ecografia hepática são independentes dos resultados das análises químicas em estudo.

H_1 : A probabilidade de o teste por ecografia ser positivo é maior quando existem alterações químicas.

A probabilidade de obter a tabela acima é dada por:

$$\frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n_{11}!n_{12}!n_{21}!n_{22}!n!} = \frac{11!8!10!9!}{7!4!3!5!19!} = 0.2000$$

A célula com a frequência mais baixa é $n_{21} = 3$. Calculemos agora a probabilidade de obter a tabela com $n_{21} - 1 = 2$ mas com os totais marginais iguais:

$$\frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{(n_{11}+1)!\,(n_{12}-1)!\,(n_{21}-1)!\,(n_{22}+1)!n!} = \frac{11!8!10!9!}{8!3!2!6!19!} = 0.0500$$

Calculemos a probabilidade de obter a tabela com $n_{21} - 2 = 1$ mantendo os totais marginais iguais:

$$\frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{(n_{11}+2)!\,(n_{12}-2)!\,(n_{21}-2)!\,(n_{22}+2)!n!} = \frac{11!8!10!9!}{9!2!1!7!10!} = 0.0048$$

Calculemos finalmente a probabilidade de obter a tabela com $n_{21} - 3 = 0$ mantendo os totais marginais iguais:

$$\frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{(n_{11}+3)!\,(n_{12}-3)!\,(n_{21}-3)!\,(n_{22}+3)!n!} = \frac{11!8!10!9!}{10!1!0!8!19!} = 0.0001$$

Somando todas estas probabilidades, vem que

$$\begin{aligned} P &= P_{n_{11}} + P_{n_{11}-1} + P_{n_{11}-2} + P_{n_{11}-3} = \\ &= 0.2000 + 0.0500 + 0.0048 + 0.0001 = 0.2549 \end{aligned}$$

Ao nível de significância de 0.05, temos: $0.2549 > 0.05$. Logo não rejeitamos a hipótese nula.

2.1.3 Frequências esperadas pequenas

Como vimos, a aproximação da estatística X^2 à distribuição χ^2 é feita tendo em conta o pressuposto de que os valores esperados não são muito pequenos. Este termo "muito pequenos"é vago e tem geralmente sido interpretado como significando que as frequências esperadas deveriam ser superiores a 5 para o teste do Qui-quadrado ser válido.

Cochran (1954) mostrou que esta regra é demasiado restritiva, e sugeriu que se relativamente poucas frequências esperadas forem inferiores a 5 (digamos 20%), podemos admitir uma célula com uma frequência esperada de 1. Até esta regra pode ser demasiado restritiva, uma vez que um trabalho mais recente de Lewontin e Felsenstein (1965) e Slakter (1966), mostra que muitas das frequências esperadas podem ser iguais a 1, sem afetar muito o teste. Lewontin e Felsenstein propuseram a seguinte regra para tabelas com 2 linhas: a tabela $2 \times c$ pode ser testada pelo critério convencional do χ^2 se todas as frequências esperadas são maiores ou iguais a 1. Estes autores afirmaram ainda que, até esta regra é extremamente conservativa e, na maioria dos casos, o critério do χ^2 pode ser usado com frequências esperadas superiores a 0.5 na menor célula.

Um processo que tem sido usado quase rotineiramente desde há muitos anos para ultrapassar o problema de frequências esperadas pequenas consiste em juntar categorias. Contudo, este pode ser criticado por diversas razões: em primeiro lugar, podemos perder uma considerável quantidade de informação ao juntarmos categorias, e isto pode prejudicar o interesse e a intenção do nosso estudo; em segundo lugar, a aleatoriedade da amostra pode ser afetada. Todo o raciocínio para o teste do χ^2 se baseia na aleatoriedade da amostra, e em que as categorias a que podem pertencer as observações são escolhidas à *priori*. Juntar categorias depois de possuirmos os dados pode afetar a natureza aleatória da amostra, com consequências que desconhecemos. Em último lugar, a maneira segundo a qual se juntam categorias pode ter um importante efeito nas inferências que fazemos. Concluindo, sempre que possível, devemos evitar juntar categorias.

2.1.4 Análise dos Resíduos

Um procedimento útil na identificação das células de uma tabela de contingência responsáveis pela significância do valor da estatística de teste num teste do Qui-quadrado consiste em analisar os desvios entre os valores observados e os esperados, medidos de uma forma apropriada. O modelo de independência (ou de homogeneidade) pode ajustar-se bem em algumas células da tabela de contingência e ajustar-se mal noutras células. Esta falta de ajustamento pode ajudar-nos a explicar associações entre variáveis. A forma mais óbvia de definir esses resíduos é subtraindo os valores esperados de cada célula ao valor observado em cada célula. No entanto este procedimento não forneceria resultados satisfatórios, uma vez que uma certa diferença tem maior relevância quando a frequência esperada é pequena do que quando é grande. Uma forma mais apropriada de calcular os resíduos e_{ij} é dada por:

$$e_{ij} = \frac{(n_{ij} - E_{ij})}{\sqrt{E_{ij}}} \tag{2.5}$$

onde E_{ij} é calculado através da equação 2.2. Estes termos são conhecidos como resíduos estandardizados e são tais que a estatística de teste do Qui-quadrado é dada por:

$$X^2 = \sum_{i=1}^{r} \sum_{i=1}^{c} e_{ij}^2. \tag{2.6}$$

Poderá ser intuitivo pensar que estes resíduos possam ser avaliados comparando-os com os percentis da distribuição Normal estandardizada, mas tal não é possível visto a variância de $\{e_{ij}\}$ ser sempre menor ou igual a um e até, em alguns casos, consideravelmente inferior a um. De facto, segundo Agrestti [2], estes resíduos são assintoticamente Normais, com valor médio zero, mas a variância dos $\{e_{ij}\}$ sob H_0, é igual a (número de graus de liberdade) / (número de células da tabela). Uma vez que o número de graus de liberdade é inferior ao número de células, a variância de e_{ij} pode ser menor do que um.

Uma análise mais precisa poderá ser alcançada usando os resíduos ajustados, d_{ij}:

$$d_{ij} = \frac{e_{ij}}{\sqrt{\left[\left(\frac{1-n_{i.}}{n}\right)\left(\frac{1-n_{.j}}{n}\right)\right]}}. \tag{2.7}$$

Quando as variáveis contidas na tabela de contingência são independentes, os resíduos ajustado são aproximadamente normalmente distribuídos com média zero e desvio-padrão um.

Exemplo[9]: Os dados da tabela seguinte representam o número de falhas nos pistões em cada componente (Norte, Central e Sul) de quatro compressores.

		Norte	**Central**	**Sul**	**Total**
	1	17	17	12	46
Compressor n^o	2	11	9	13	33
	3	11	8	19	38
	4	14	7	28	49
Total		53	41	72	166

Na tabela seguinte estão (a) os resíduos ajustados para as falhas dos pistões e (b) os resíduos estadardizados para as falhas dos pistões:

		Norte	**Central**	**Sul**
	(a)			
	1	0.86	2.27	-2.78
Compressor	2	0.19	0.38	-0.52
n^o	3	-0.45	-0.59	0.94
	4	-0.60	-2.01	2.32
	(b)			
	1	0.60	1.67	-1.78
Compressor	2	0.14	0.30	-0.35
n^o	3	-0.32	-0.45	0.62
	4	-0.41	-1.47	1.46

O valor da estatística X^2 para a tabela dada é 11.7 com seis graus de liberdade demonstrando uma evidência relativamente fraca de associação entre as variáveis ($valor - p = 0.069$). Quatro dos resíduos ajustados têm valor absoluto maior que 2,

sendo o maior correspondente à parte Sul do compressor 1. Neste exemplo, o uso de resíduos ajustados ao invés de resíduos estandardizados tem um efeito considerável na análise. O valor absoluto do maior resíduo estandardizado é 1.78 pelo que este resíduos não fornecem evidência suficiente da não existência de independência.

2.1.5 Combinação de Tabelas de Contingência 2×2

A combinação de informação recolhida por vários observadores pode levar à combinação de tabelas de contingência.

Em muitos estudos podemos, pois, ter de analisar dados de várias tabelas de contingência 2×2 todas referentes à mesma questão. Neste caso torna-se útil decidir de que forma poderemos combiná-las para testar a associação entre a variável cujas categorias se encontram dispostas nas linhas da tabela e a variável cujas categorias se encontram dispostas nas colunas da tabela.

Por exemplo, poderemos querer testar a associação entre o cancro do pulmão e o tabaco e os dados estarem em diferentes tabelas de contingência, cada uma delas referente a uma região do país. Como poderemos relacionar a informação contida nestas tabelas? Um método que poderá ser intuitivo é combinar todos os dados numa única tabela de contingência 2×2 e calcular a estatística do Qui-quadrado habitual. Mas este procedimento só será legítimo se as proporções correspondentes nas várias tabelas forem semelhantes. Este procedimento não deverá ser utilizado caso as proporções variem de tabela para tabela ou se houver suspeita de que haja variação, uma vez que neste caso os dados combinados não irão refletir de forma exata a informação contida nas tabelas originais. Retomando o exemplo do cancro do pulmão, poderemos estar na situação em que a ocorrência desta doença é mais frequente em algumas regiões do que noutras.

Outra técnica também muitas vezes utilizada é calcular o valor habitual da estatística do Qui-quadrado para cada tabela separadamente e depois somá-los. A estatística resultante deverá então ser comparada com o valor correspondente de χ^2 com g graus de liberdade onde g é o número de tabelas originais (esta definição baseia-se no facto de que a soma de g variáveis Qui-quadrado independentes, com 1 grau de liberdade cada, é ela própria distribuída como uma variável Qui-quadrado com g graus de liberdade). Este método não será também adequado uma vez que não tem em conta a direção das diferenças entre as proporções nas várias tabelas e consequentemente não é suficientemente potente para detetar a diferença que surge constantemente na mesma direção em todas ou quase todas as tabelas indivíduais.

Se existem outros fatores, para além das duas variáveis dicotómicas de maior interesse, que podem afetar a relação entre elas, então é importante ter estes fatores em consideração - são chamados fatores de confundimento na análise dos dados. Para corretamente termos acesso à relação entre as duas variáveis, será necessário considerar separadamente os dados para cada categoria do factor de confundimento. Numa análise à homogeneidade de dois grupos a hipótese nula especifica que para cada categoria da variável de confundimento a probabilidade de sucesso é a mesma nos dois grupos, mas possivelmente diferente de uma categoria da variável de confundimento para outra.

Três métodos mais adequados para combinar a informação contida em várias tabelas de contingência 2×2 serão apresentados nas secções seguintes.

Método da raíz do Qui-quadrado

Se os tamanhos das amostras não diferem muito entre si (digamos que não diferem mais do que numa razão de 2 para 1) e os valores das proporções se encontram entre aproximadamente 0.2 e 0.8 [9], então poderemos utilizar um método baseado nas somas das raízes das estatísticas do Qui-quadrado e que tem em consideração os sinais das diferenças das proporções. É fácil de ver que sob a hipótese de que as proporções são iguais, o valor de X para qualquer das tabelas 2×2 é aproximadamente normalmente distribuído com média zero e desvio-padrão igual a 1. Consequentemente, a soma destes valores de X para o conjunto total de g tabelas é aproximadamente normalmente distribuída com média zero e desvio-padrão $\sqrt{g}$. Então a estatística de teste para a hipótese de não existirem diferenças entre as proporções em todas as tabelas é:

$$Z = \sum_{i=1}^{g} \frac{X_i}{\sqrt{g}} \tag{2.8}$$

onde X_i é o valor da raíz quadrada da estatística X^2 usual para a $i - ésima$ tabela sempre acompanhado do respectivo sinal.

Exemplo[9]: Os dados contidos na tabela seguinte referem-se à incidência de tumores malignos (B) e benignos (A) nos hemisférios esquerdo e direito do córtex. Pretendemos testar a existência de associação entre o hemisfério e o tipo de tumor. Informação relativa a três sítios diferentes em cada hemisfério foi recolhida, mas um estudo inicial mostrou não existir razão para suspeitar de que qualquer relação entre o hemisfério e o tipo de tumor variaria de um sítio para outro, portanto foi indicada uma avaliação geral entre o tipo de tumor e o hemisfério.

Sítio	**Localização do tumor**	**A**	**B**	**Proporção de B**	X^2	X_i
1.	Hemisfério esquerdo	17	5	0.2273	1.7935	1.3392
	Hemisfério direito	6	5	0.4545		
		23	10			
2.	Hemisfério esquerdo	12	3	0.200	1.5010	1.2288
	Hemisfério direito	7	5	0.4167		
		19	8			
3.	Hemisfério esquerdo	11	3	0.2143	2.003	1.4155
	Hemisfério direito	11	9	0.4500		
		22	12			

Para cada um dos 3 sítios o número de pacientes é relativamente aproximado (33, 27 e 34 respetivamente) logo podemos aplicar o método da raíz do Qui-quadrado. O

valor de X^2 é inicialmente calculado para cada tabela separadamente (de notar que nenhum deles é significativo). É depois calculada a raíz quadrada destes valores e o sinal da diferença entre as proporções é atribuído a cada valor X_i. Para estes dados a diferença entre as proporções tem a mesma direção nas três tabelas, uma vez que a proporção de tumores malignos no hemisferio direito é sempre superior à proporção de tumores no hemisfério esquerdo. Logo o mesmo sinal é atribuído a cada X_i (se é positivo ou negativo é indiferente). Aplicamos então a estatística de teste:

$$\begin{aligned} Z &= \sum_{i=1}^{g} \frac{X_i}{\sqrt{g}} = \\ &= \frac{1.339 + 1.229 + 1.415}{\sqrt{3}} = \\ &= \frac{3.983}{\sqrt{3}} = 2.3 \end{aligned}$$

Este valor será comparado com os valores da tabela da distribuição normal estandardizada de onde vemos que é significativo ao de nível de significância de 0.05. Desta forma, considerando os três diferentes sítios dos hemisférios juntos, existe evidência de associação entre o tipo de tumor e o hemisfério do córtex.

Se para estes dados somássemos os valores da estatística Qui-quadrado de cada tabela, ou seja, $1.793+1.501+2.003 = 5.297$, obteríamos um resultado não significativo quando comparado com $\chi^2_{0.05}(3)$. Claramente neste caso, onde cada uma das diferenças tem a mesma direção, o método da raíz do Qui-quadrado é mais potente que aquele baseado na soma dos valores indivíduais do Qui-quadrado.

Método de Cochran

Se os tamanhos das amostras e as proporções não satisfazem as condições mencionadas pelo método anterior, então o método da raíz do Qui-quadrado tem tendência a perder potência. Métodos aplicados a tabelas correspondentes a amostras muito pequenas não conseguem detetar diferenças nas proporções de forma tão clara como quando são aplicados a tabelas provenientes de amostras maiores. Quando as diferenças no tamanho no total de observações para cada tabela são muito grandes, é necessário aplicar um método que pondere os dados das diferentes tabelas. Cochran (1954) sugeriu um método assim: a estatística de teste Y que é uma média ponderada das diferenças entre as proporções em cada tabela:

$$Y = \frac{\sum_{i=1}^{g} w_i d_i}{\sqrt{\sum_{i=1}^{g} w_i P_i Q_i}} \tag{2.9}$$

onde:

- g é o número total de tabelas 2×2;

- $P_i = (n_{i1}\widehat{p}_{i1} + n_{i2}\widehat{p}_{i2}) / (n_{i1} + n_{i2})$ sendo n_{i1} e n_{i2} os totais marginais dos dois grupos representados nas linhas (ou colunas) para a $i - ésima$ tabela, $\widehat{p}_{i1}$ e $\widehat{p}_{i2}$ as respectivas proporções, que são os estimadores de máxima verosimilhança das correspondentes proporções populacionais;
- $Q_i = (1 - P_i)$;
- $d_i = (\widehat{p}_{i1} - \widehat{p}_{i2})$;
- $w_i = n_{i1}n_{i2} / (n_{i1} + n_{i2})$.

Y é a média ponderada dos valores d_i na qual os pesos usados atribuem maior importância às diferenças baseadas em amostras grandes do que em amostras pequenas.

As hipóteses a formular são:

$$\begin{aligned} H_0 &: p_{i1} = p_{i2} \ , \forall i \ , \ i = 1, \ldots, g \\ H_1 &: \exists i : p_{i1} \neq p_{i2} \ , \ i = 1, \ldots, g \end{aligned}$$

Sob a hipótese de que as diferenças entre as proporções na população são zero para $i = 1, \ldots, g$, a estatística Y é normalmente distribuída com média zero e variância unitária.

Exemplo[9]: A tabela seguinte ilustra a incidência de tiques nervosos entre rapazes e raparigas com dificuldades em se integrarem com outras crianças, divididos por três classes etárias:

		Tiques			
Classe etária	**Sexo**	**Sim**	**Não**	**Total**	**Proporção com tiques**
5-9 anos	Rapazes	13	27	70	0.1857
	Raparigas	3	23	26	0.1154
	Total	16	80	96	0.1667
10-12 anos	Rapazes	26	56	82	0.3171
	Raparigas	11	29	40	0.2750
	Total	37	85	122	0.3033
13-15 anos	Rapazes	15	56	71	0.2113
	Rparigas	2	27	29	0.0690
	Total	17	83	100	0.1700

Os valores necessários ao cálculo da estatística Y são os seguintes:

- Classe etária 5-9 anos:

$$n_{11} = 70 \text{ e } n_{12} = 26,$$

$$\widehat{p}_{11} = 0.1857 \text{ e } \widehat{p}_{12} = 0.1154,$$

$$P_1 = 0.16671 \text{ e } Q_1 = 0.8333,$$

$$d_1 = 0.0703 \text{ e } w_1 = 18.96.$$

- Classe etária 10-12 anos:

$$n_{21} = 82 \text{ e } n_{22} = 40,$$

$$\widehat{p}_{21} = 0.3171 \text{ e } \widehat{p}_{22} = 0.2750,$$

$$P_2 = 0.3033 \text{ e } Q_2 = 0.6967,$$

$$d_2 = 0.0421 \text{ e } w_2 = 26.89.$$

- Classe etária 13-15 anos:

$$n_{31} = 71 \text{ e } n_{32} = 29,$$

$$\widehat{p}_{31} = 0.2113 \text{ e } \widehat{p}_{32} = 0.0690,$$

$$P_3 = 0.1700 \text{ e } Q_3 = 0.8300,$$

$$d_3 = 0.1423 \text{ e } w_2 = 20.59.$$

Calculando a estatística de teste:

$$\begin{aligned} Y &= \frac{\sum_{i=1}^{g} w_i d_i}{\sqrt{\sum_{i=1}^{g} w_i P_i Q_i}} = \\ &= \frac{18.96 \times 0.0703 + 26.89 \times 0.0421 + 20.59 \times 0.1423}{\sqrt{18.96 \times 0.1667 \times 0.8333 + 26.89 \times 0.3033 \times 0.6967 + 20.59 \times 0.1700 \times 0.8300}} = \\ &= \frac{1.333 + 1.132 + 2.93}{\sqrt{2.63 + 5.68 + 2.91}} = \frac{5.395}{\sqrt{11.22}} = 1.61 \end{aligned}$$

Consultando a tabela da distribuição normal vemos que a probabilidade correspondente é de 0.1074.

Se tivessemos juntando as três classes etárias e calculado um teste Qui-quadrado, teria sido obtido o valor 2.110, valor este correspondente a uma probabilidade de 0.2838 o que é mais do dobro do valor obtido pelo método de Cochran, o que demonstra a maior sensibilidade deste último.

Em casos onde a relação entre duas variáveis é claramente diferente de tabela para tabela, nenhum dos métodos anteriormente descritos será muito útil. Por exemplo, suponhamos que tinhamos apenas 2 tabelas em que as amostras possuem tamanhos idênticos; se as tabelas contiverem números de observações semelhantes e se as diferenças nas proporções de interesse para ambas as tabelas são grandes, mas com sinais contrários, então tanto a estatística do método da raíz do Qui-quadrado como o de Cochran terão valor aproximadamente zero e ambos os resultados não serão significativos. Torna-se assim necessário ter em conta que ambos os procedimentos são úteis essencialmente para detetar afastamentos da hipótese nula devido a diferenças constantes nas proporções de tabela para tabela. Deverá ser evitado a aplicação de qualquer um destes métodos a conjuntos de tabelas cujas diferenças variam em magnitude e em direção. Nestes casos não é mesmo aconselhável qualquer tipo de combinação de tabelas.

Teste de Mantel-Haenszel

O teste de Mantel-Haenszel (muitas vezes também chamado de Cochran-Mantel-Haenszel) é geralmente usado para repetições de testes de independência em tabelas $2 \times 2 \times k$. Temos três variáveis, duas delas são testadas para averiguar se existe independência entre elas e a terceira determina as repetições, ou seja, é um teste adequado para testar a hipótese nula de independência entre duas variáveis dicotómicas usando dados de uma população dividida em classes ou estratos.

Embora esta estatística não seja uma soma de estatísticas do Qui-quadrado calculadas para cada estrato individualmente, o processo é muito semelhante a isso. É mais potente do que simplesmente combinar Qui-quadrados indivíduais e menos susceptível ao problema de frequências esperadas pequenas nas tabelas indivíduais 2×2.

O cálculo da estatística de Mantel-Haenszel baseia-se no facto de que para cada tabela 2×2, as frequências observadas em qualquer uma das células, dados os totais marginais, determinam as frequências em cada uma das restantes células. Isto significa que pode ser criada uma estatística usando apenas os dados na célula $(1, 1)$ de cada uma das tabelas.

Existem algumas variações desta estatística, mas a mais usual é:

$$M^2 = \frac{\left(\left|\sum_{k=1}^{g} O_{11k} - \sum_{k=1}^{g} E_{11k}\right| - \frac{1}{2}\right)^2}{\sum_{k=1}^{g} \frac{n_{1.k} n_{2.k} n_{.1k} n_{.2k}}{n_{..k}^2 (n_{..k} - 1)}} \tag{2.10}$$

onde O_{11k} e E_{11k} são as frequências observadas e esperadas na célula n_{11} de cada uma das g tabelas 2×2 e no denominador estão os totais marginais e o total absoluto de

cada uma dessas tabelas. O denominador representa a variância do numerador. É feita uma correção para a continuidade, a correção de Yates, sendo que a isso corresponde a entrada $-\frac{1}{2}$ no numerador. Esta estatística tem uma distribuição aproximada da distribuição Qui-quadrado com 1 grau de liberdade.

O valor de M^2 aumenta na medida em que as diferenças entre os valores esperados e observados aumentam ou conforme o valor da variância (denominador) diminui.

Diferentes autores poderão apresentar a fórmula de Mantel-Haenszel de diferentes formas, mas são todas algébricamente equivalentes. Algumas dessas fórmulas poderão, por exemplo, não incluir a correção de Yates.

As hipóteses a formular para este teste são:

$$\begin{aligned} H_0 &: p_{k1} = p_{k2} \ , \forall k \ , k = 1, \ldots, g \\ H_1 &: \exists k : p_{k1} \neq p_{k2} \ , k = 1, \ldots, g \end{aligned}$$

Exemplo[17]: A Universidade da Califórnia em Berkeley investigou a discriminação de genéro nas admissões em 1973. Um exame superficial mostra que aproximadamente 45% dos candidatos masculinos foram admitidos naquele ano, enquanto que apenas 30% das candidatas foram admitidas. Aparentemente isto demonstra um caso de discriminação de genéro, mas as admissões nesta universidade são feitas por departamento pelo que será mais adequado olhar para os dados de admissões de cada um destes departamentos. A tabela seguinte mostra os dados referentes a 5 desses departamentos:

	Masculino		**Feminino**	
Departamento	**Aceite**	**Rejeitado**	**Aceite**	**Rejeitado**
A	353	207	17	8
B	120	205	202	391
C	138	279	131	244
D	53	138	94	299
E	22	351	24	317
Total	686	1180	508	1259

Olhando de uma forma geral, temos que 36.8% dos homens são admitidos enquanto que apenas 28.8% das mulheres são admitidas. No entanto se virmos por departamento, existem três nos quais a taxa de mulheres aceite é maior do que a taxa de homens.

Temos que as frequências observadas e esperadas para os dados anteriores são:

Departamento	O_{11k}	E_{11k}	**Variância**
A	353	354.19	5.572
B	120	114.00	47.861
C	138	141.63	44.340
D	53	48.08	24.251
E	22	24.03	10.753
Total	686	681.93	132.777

Aplicando a estatística de Mantel-Haenszel:

$$\begin{aligned} M^2 &= \frac{\left(\left|\sum_{k=1}^{5} O_{11k} - \sum_{k=1}^{5} E_{11k}\right| - \frac{1}{2}\right)^2}{\sum_{k=1}^{5} \frac{n_{1.k} n_{2.k} n_{.1k} n_{.2k}}{n_{..k}^2 (n_{..k}-1)}} = \\ &= \frac{\left(|686 - 681.93| - \frac{1}{2}\right)^2}{132.777} = 0.096 \end{aligned}$$

Como este valor é inferior a $3.841 = \chi^2_{0.05}(1)$ não rejeitamos a hipótese de independência entre as duas variáveis, pelo que, de uma maneira geral não existe evidência de discriminação de género nas admissões da Universidade de Berkeley.

2.1.6 Teste de McNemar

O teste de McNemar tem muitas aplicações em todas as ciências, em estudos em que o mesmo grupo é utilizado várias vezes ao longo do tempo para recolha de informação com o objetivo de estudar um processo evolutivo, visto que se foca nas células em que existem discordâncias e apenas estas são comparadas.

Pode ser utilizado em duas situações:

- para avaliar dados categorizados obtidos através de um estudo em que o mesmo grupo de indivíduos é sujeito a dois tipos de experiências;
- no caso de uma situação de antes e depois. Neste caso, o grupo de indivíduos em estudo são sujeitos a uma avaliação antes de serem submetidos a uma experiência e são também avaliados após essa mesma experiencia. Testa-se a existência de diferenças nas avaliações antes e após a experiência.

Dada uma tabela:

		Experiência 1	
		A presente	**A ausente**
Experiência	**A presente**	n_{11}	n_{12}
2	**A ausente**	n_{21}	n_{22}

H_0: A proporção de sucessos (A presente) é igual em ambas as experiências; no nosso caso virá $H_0 : p_{1.} = p_{.1}$.

H_1 : A proporção de sucessos (A presente) é diferente em ambas as experiências; no nosso caso virá $H_0 : p_{1.} \neq p_{.1}$.

Isto é equivalente a testar se a proporção de discordâncias num sentido é igual à proporção de discordâncias no outro sentido.

A estatística de teste é dada por:

$$X^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}} \tag{2.11}$$

e tem distribuição qui-quadrado com 1 grau de liberdade. Quando $X^2 \geq \chi^2_\alpha$ rejeitamos a hipótese nula.

Este teste baseia-se nos seguintes pressupostos:

1. os indivíduos em estudo foram selecionados aleatoriamente da população a que pertencem;

2. cada um dos $n_{11} + n_{12} + n_{21} + n_{22}$ pares de observações na tabela é independente dos restantes;

3. os dados referentes aos indivíduos em estudo deverão ser apresentados de forma dicotómica envolvendo duas categorias mutuamente exclusivas;

4. se n_{12} e n_{21} são pequenos, o teste de McNemar não deverá ser utilizado. Assim, se $\frac{n_{12}+n_{21}}{2} \leq 10$, deverá ser utilizado o teste a uma proporção em que $n = n_{12} + n_{21}$ e $X = n_{12}$ (ou n_{21}) em que X representa o número de sucessos na amostra, baseado na distribuição Binomial.

Atualmente existe uma extensão deste teste para tabelas $r \times r$ com $r > 2$. Neste caso, testamos se o canto superior direito da tabela é simétrico ao canto inferior esquerdo.

Temos a estatística de teste:

$$X^2 = \sum_{i=1}^{r} \sum_{j>i} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \tag{2.12}$$

que tem aproximadamente a distribuição de uma variável aleatória qui-quadrado com $\frac{r(r-1)}{2}$ graus de liberdade.

Exemplo 1 (caso de uma verdadeira experiência)[26]: Um psicólogo pretende testar um medicamento para combater a enurese comparando-o a um placebo. São administrados tanto o placebo quanto o medicamento a 100 crianças acompanhadas durante um período de 6 meses. Durante este estudo, cada criança é submetida a 6 tratamentos com placebo e 6 tratamentos com o medicamento, cada um com a duração de 1 semana, seguido sempre de uma semana de descanso para que não haja interferência entre tratamentos. É depois perguntado aos pais se houve resposta favorável após cada um dos tratamentos.

		Resposta favorável ao medicamento		
		Sim	**Não**	**Total**
Resposta favorável ao placebo	**Sim**	10	13	23
	Não	41	36	77
	Total	51	49	100

H_0 : A proporção de indivíduos na célula n_{12} é semelhante à da célula n_{21}.

$$\chi^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}} = \frac{(|13 - 41| - 1)^2}{13 + 41} = \frac{729}{54} = 13.5$$

Uma vez que $13.5 > \chi^2_{0.05}(1) = 3.841$, rejeitamos H_0. Portanto a proporção de indivíduos que respondeu favoravelmente ao medicamento é significativamente superior à proporção dos que responderam favoravelmente ao placebo.

Exemplo 2 (caso antes e depois)[1]: Suponhamos que se recorre a um painel de opinião para avaliar a alteração de base de apoio ao governo causada pela alterações legislativas que o OE introduz nas deduções do IRS. O resultado encontra-se na tabela seguinte:

		Depois	
		A favor	**Contra**
Antes	**A favor**	36	94
	Contra	108	12

H_0 : Não houve alteração no apoio dado ao governo.

$$\chi^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}} = \frac{(|12 - 36| - 1)^2}{(12 + 36)} = \frac{529}{48} = 11.02$$

Como $11.02 > 3.841 = \chi^2_{0.05}(1)$, rejeitamos H_0.

2.1.7 Teste de Cochran

Muitas vezes temos que o uso de um tratamento poderá ter dois possíveis resultados: o sucesso ou a falha. Se c tratamentos forem aplicados a um grupo de n elementos diferentes e independentes uns dos outros, os resultados podem ser traduzidos numa tabela de contingência $2 \times c$:

	Tratamentos				
	1	2	$\cdots$	c	**Total**
Sucesso	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
Falha	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	n

(uma linha para os sucessos e outra para os insucessos e uma coluna para cada tratamento) usualmente com uma margem fixa, e assim utilizar-se o teste do qui-quadrado para averiguar se existem diferenças entre os tratamentos. Contudo, muitas vezes é possível detetar diferenças subtis entre os tratamentos, ou seja, aumentar a potência do teste, se cada um dos n elementos for submetido aos c tratamentos.

	Tratamentos			
	T_1	T_2	$\cdots$	T_c
indiv. 1	x_{11}	x_{12}	$\cdots$	x_{1c}
$\vdots$	x_{21}	x_{22}	$\cdots$	x_{2c}
indiv. n	x_{n1}	x_{n2}	$\cdots$	x_{nc}

em que $x_{ij} = \begin{cases} 1, \text{ se o resultado = sucesso} \\ 0, \text{ se o resultado = falha} \end{cases}$.

O teste de Cochran é um método apropriado para este tipo de análise, pois destina-se a comparar c grupos ($c > 2$) emparelhados, onde os dados são dicotómicos, tendo por objetivo averiguar se as frequências ou proporções de sucesso dos c grupos diferem entre si significativamente.

Este teste tem por base os seguintes pressupostos:

1. os n elementos da amostra foram selecionados aleatoriamente da população a que pertencem;

2. os dados deverão estar dicotomizados.

Hipótese a testar:
H_0 : Os tratamentos têm efeitos iguais ou $H_0 : p_{i1} = p_{i2} = \ldots = p_{ic}$ em que p_{ik} = probabilidade de sucesso no elemento i com o tratamento k.
vs
H_1 : Existem diferenças entre os efeitos dos tratamentos ou $H_1 : \exists i : \exists j, k : p_{ij} \neq p_{ik}$
.

Estatística de teste:

$$Q = \frac{(c-1)\left[c\sum_{j=1}^{c} C_j^2 - \left(\sum_{j=1}^{c} C_j\right)^2\right]}{cn - \sum_{i=1}^{n} R_i^2} \tag{2.13}$$

onde c é o número de tratamentos, n é o número de elementos da amostra, C_i o número total de sucessos por coluna e R_i o número total de sucessos por linha, que, sob a hipótese de os tratamentos terem o mesmo efeito, têm aproximadamente uma ditribuição qui-quadrado com $(c-1)$ graus de liberdade.

Assim usando a estatística Q estamos a testar a homogeneidade dos tratamentos.

Região de rejeição de H_0: $Q \geq \chi^2_{\alpha-1}(k-1)$.

Exemplo[1]: Pretende-se estudar a influência da atitude de um entrevistador sobre as respostas de donas de casa a uma determinada pesquisa de opinião. Os três entrevistadores são treinados para efetuar as suas entrevistas de três maneiras diferentes, desde a mais cordial à mais brusca. Cada entrevistador questiona cada uma das 10 donas de casa. É lhes perguntado se concordam com as alterações feitas à circulação numa estrada da zona, sendo que 0 significa "não concordo"e 1 "concordo".

H_0 : A atitude do entrevistador não influenciou as respostas dadas pelas donas de casa.

H_1 : A atitude do entrevistador influenciou as respostas dadas pelas donas de casa.

Entrev. 1	**Entrev. 2**	**Entrev. 3**	$\mathbf{R}_i$	$\mathbf{R}_i^2$
0	0	0	0	0
1	1	0	2	4
0	1	0	1	1
0	0	0	0	0
1	0	0	1	1
1	1	0	2	4
1	1	0	2	4
0	1	0	1	1
1	0	0	1	1
0	0	0	0	0
$C_1 = 5$	$C_2 = 5$	$C_3 = 0$	$\sum R_i = 10$	$\sum R_i^2 = 16$

$$\begin{aligned} Q &= \frac{(c-1)\left[c\sum_{j=1}^{c} C_j^2 - \left(\sum_{j=1}^{c} C_j\right)^2\right]}{cn - \sum_{i=1}^{n} R_i^2} = \\ &= \frac{(3-1)\left[3\times(5^2+5^2+0^2) - (5+5+2)^2\right]}{(3\times 10) - 16} = \\ &= \frac{2\left[150-100\right]}{284} = 0.352 \end{aligned}$$

$\chi^2_{\alpha-1}(c-1) = \chi^2_{0.05}(2) = 5.991$

Como $0.352 < 5.991$ podemos concluir que a atitude do entrevistador não influenciou as respostas dadas pelas donas de casa.

Madansky (1963) no seu artigo "Tests of homogeneity for correlated samples"[21] generalizou o teste de Cochran ao caso em que cada tratamento pode ter um número arbitrário de resultados possíveis.

2.2 Tabelas de Contingência Tridimensionais

A extensão da análise de tabelas bidimensionais a tabelas tridimensionais não é simples. Teremos de considerar possíveis interações entre as diversas categorias de três fatores. Existir uma interação entre as três variáveis categorizadas significa que a relação entre duas delas não é a mesma em todas as categorias da outra variável. Mas podem existir interações apenas entre duas das três variáveis, como abordaremos nesta secção.

2.2.1 Teste de Independência

A hipótese de independência mútua das variáveis numa tabela de contingência tridimensional poderá ser formulada da seguinte maneira:

$$H_0 : p_{ijk} = p_{i..}p_{.j.}p_{..k}, \forall i, j, k, \text{ com } i = 1, \ldots, r, j = 1, \ldots, c \text{ e } k = 1, \ldots, l$$

onde p_{ijk} representa a probabilidade de uma observação estar na $ijk-ésima$ célula da tabela e $p_{i..}$, $p_{.j.}$ e $p_{..k}$ são as probabilidades marginais das primeira, segunda e terceira variáveis categorizadas em estudo, respetivamente. Esta é a hipótese equivalente à utilizada no caso de uma tabela de contingência bidimensional e também para testá-la é utilizado um processo análogo ao do caso bidimensional.

A hipótese alternativa será:

$$H_1 : \exists i, j, k : p_{ijk} \neq p_{i..}p_{.j.}p_{..k}.$$

Começamos por calcular as frequências esperadas que são depois comparadas com as frequências observadas utilizando a estatística Qui-quadrado. As frequências esperadas são calculadas através da fórmula:

$$E_{ijk} = n\widehat{p}_{i..}\widehat{p}_{.j.}\widehat{p}_{..k} \tag{2.14}$$

onde $\widehat{p}_{i..}$, $\widehat{p}_{.j.}$ e $\widehat{p}_{..k}$ são as probabilidades correspondentes estimadas. Os melhores estimadores são aqueles que derivam dos totais marginais de cada variável (são os estimadores de máxima verosimilhança), nomeadamente:

$$\widehat{p}_{i..} = \frac{n_{i..}}{n}, \widehat{p}_{.j.} = \frac{n_{.j.}}{n} \text{ e } \widehat{p}_{..k} = \frac{n_{..k}}{n}. \tag{2.15}$$

Substituindo estes valores em 2.14 fica:

$$E_{ijk} = n\frac{n_{i..}}{n}\frac{n_{.j.}}{n}\frac{n_{..k}}{n} = \frac{n_{i..}n_{.j.}n_{..k}}{n^2}. \tag{2.16}$$

Desta forma, a estatística de teste é calculada da forma usual:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}} \tag{2.17}$$

que tem aproimadamente uma distribuição Qui-quadrado com $rcl - r - c - l + 2$ graus de liberdade.

Exemplo[9]: A tabela seguinte ilustra um conjunto de dados referentes ao comportamento no que diz respeito ao suicídio, entre 3375 homens e mulheres, segundo 6 métodos (1 - envenenamento através de matéria sólida ou líquida, 2 - gás, 3 - enforcamento, sufocamento ou afogamento, 4 - Pistolas, facas ou explosivos, 5 - atirar-se de um penhasco/ponte, 6 - outros) e 3 classe etárias (A1 - 10 a 40 anos, A2 - 40 a 70 anos e A3 - mais de 70 anos):

	Método					
	1	**2**	**3**	**4**	**5**	**6**
Homens						
A1	398	121	455	155	55	124
A2	399	82	797	168	51	82
A3	93	6	316	33	26	14
Mulheres						
A1	259	15	95	14	40	38
A2	450	13	450	26	71	60
A3	154	5	185	7	38	10
Total	1753	242	2298	303	281	328

e os totais para as classes etárias são:

$$n_{A1} = 1769,\ n_{A2} = 2649 \text{ e } n_{A3} = 887.$$

Para calcular as frequências esperadas é só aplicar a equação 2.16, por exemplo para a célula n_{111} (Homem, método 1, classe etária 1 - 10 a 40 anos) temos:

$$\begin{aligned} E_{ijk} &= \frac{n_{i..}n_{.j.}n_{..k}}{N^2} = \frac{3375 \times 1769 \times 1753}{5305^2} = \\ &= \frac{10466067375}{28143025} = 371.89 \end{aligned}$$

As restantes frequências esperadas encontram-se representadas na tabela seguinte:

	Método					
	1	**2**	**3**	**4**	**5**	**6**
Homens						
A1	371.89	51.34	487.5	85.49	59.61	69.58
A2	556.89	76.88	730.02	128.02	89.27	104.20
A3	186.47	25.74	244.44	42.87	29.89	34.89
Mulheres						
A1	212.66	29.36	278.78	48.89	34.09	39.79
A2	318.46	43.96	417.46	73.21	51.05	59.59
A3	106.63	14.72	139.78	24.51	17.09	19.95

Aplicando a fórmula da estatística de teste do Qui-quadrado:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}} = 90.3$$

e os graus de liberdade são:

$$rcl - r - c - l + 2 = 3 \times 6 \times 2 - 3 - 6 - 2 + 2 = 27$$

pelo que $\chi^2_{0.05}(27) = 40.113$. Como $747.4 > 40.113$ rejeitamos a hipótese de independência entre as variáveis.

Se não rejeitamos a hipótese de independência anteriormente descrita, então uma análise mais pormenorizada da tabela não trará grandes vantagens. No entanto, se rejeitarmos a hipótese de independência não deveremos assumir que existe uma associação significativa entre todas as variáveis. Pode, por exemplo, ocorrer que existe uma associação entre duas das variáveis em estudo enquanto que a terceira variável é completamente independente das outras. Isto significa que a associação entre as duas variáveis é idêntica para todas as categorias da terceira variável. Neste caso, terá interesse testar a hipótese de independência parcial. Também pode ocorrer que duas das variáveis sejam independentes para cada um dos níveis da terceira variável mas cada uma delas poderá estar associada com a terceira, ou seja, as primeiras duas variáveis são independentes condicionalmente dado o nível da terceira. Tais condições poderão mais uma vez ser formuladas em termos de probabilidades, sendo as três possíveis hipóteses de independência parcial para uma tabela de contingência tridimensional:

1. $H_0^{(1)} : p_{ijk} = p_{i..}p_{.jk}$ - primeira variável independente das restantes;
2. $H_0^{(2)} : p_{ijk} = p_{.j.}p_{i.k}$ - segunda variável independente das restantes;
3. $H_0^{(3)} : p_{ijk} = p_{..k}p_{ij.}$ - terceira variável independente das restantes.

Pegando na hipótese $H_0^{(1)} : p_{ijk} = p_{i..}p_{.jk}$, $\forall i, j, k$ e estudando em mais detalhe, ela diz-nos que a probabilidade de uma observação ocorrer na célula ijk, ou seja, p_{ijk}, é dada pelo produto entre a probabilidade de que esta observação pertença à $i-ésima$ categoria da primeira variável, $p_{i..}$, e a probabilidade de que a observação pertença simultaneamente à categoria j da segunda variável e à categoria k da terceira variável. Se a hipótese é verdadeira isso implica que a primeira variável é independente das restantes, o que implica a veracidade das seguintes hipóteses compostas:

$$p_{ij.} = p_{i..}p_{.j.} \text{ e } p_{i.k} = p_{i..}p_{..k}, \tag{2.18}$$

uma vez que

$$p_{ij.} = \sum_k p_{ijk} \underset{p_{ijk}=p_{i..}\times p_{.jk}}{=} \sum_k p_{i..} \times p_{.jk} = p_{i..} \sum_k p_{.jk} = p_{i..} \times p_{.j.} \tag{2.19}$$

e

$$p_{i.k} = \sum_j p_{ijk} \underset{p_{ijk}=p_{i..}\times p_{.jk}}{=} \sum_j p_{i..} \times p_{.jk} = p_{i..} \sum_j p_{.jk} = p_{i..} \times p_{..k} \tag{2.20}$$

Para testar a hipótese $H_0^{(1)}$, o procedimento é exatamente o mesmo que explicado anteriormente, sendo que o cálculo das frequências esperadas é feito através da fórmula:

$$E_{ijk} = n\widehat{p}_{i..}\widehat{p}_{.jk} \tag{2.21}$$

sendo $\widehat{p}_{i..}$ e $\widehat{p}_{.jk}$ os estimadores de máxima verosimilhança das probabilidades $p_{i..}$ e $p_{.jk}$ obtidos da seguinte forma:

$$\widehat{p}_{i..} = \frac{n_{i..}}{n} \text{ e } \widehat{p}_{.jk} = \frac{n_{.jk}}{n}. \tag{2.22}$$

Desta forma temos:

$$E_{ijk} = n\frac{n_{i..}}{n}\frac{n_{.jk}}{n} = \frac{n_{i..}n_{.jk}}{n}. \tag{2.23}$$

A estatística de teste poderá ser calculada usado a fórmula 2.17 com $clr - cl - r + 1$ graus de liberdade.

2.2.2 Modelos Log-lineares

Apresentaremos, apenas, uma breve abordagem aos modelos log-lineares tridimensionais, sendo imediata a generalização a quatro ou mais variáveis.

Goodman desenvolveu o modelo log-linear como uma ferramenta analítica muito prática para abordar relações entre variáveis categorizadas. É uma técnica muito utilizada quando analisamos as interações entre três ou mais variáveis.

Apresentaremos o modelo para tabelas de contingência tridimensionais. Este modelo pode ser facilmente estendido a situações em que existam mais variáveis envolvidas.

Queremos descobrir que interações, no caso de existir alguma, representam o relacionamento entre as 3 variáveis.

Vamos identificar as três variáveis como A, B e C e suponhamos que têm I, J e K categorias respetivamente. Designamos por p_{ijk} a probabilidade teórica (desconhecida) de que uma observação escolhida ao acaso pertença à célula (i, j, k) da tabela, e seja $\upsilon_{ijk} = \ln(p_{ijk})$. Uma forma alternativa de definir p_{ijk} é como sendo a probabilidade de que um indivíduo, escolhido aleatoriamente da amostra em estudo, pertença simultaneamente à categoria i da variável A, à categoria j da variável B e à categoria k da variável C.

Um método simples para construir um modelo linear com os logaritmos das probabilidades das células, é por analogia com os modelos da Análise de Variância.

O modelo saturado fornece-nos uma expressão completa para as quantidades $\{\upsilon_{ijk}\}$ em termos de uma média geral, dos efeitos principais das variáveis A, B e C, das três interações entre variáveis duas a duas , AB, AC e BC e a interação entre as três variáveis ABC. Sobre ABC falaremos mais tarde. Usamos expoentes para simbolizar as variáveis envolvidas e indíces para simbolizar as categorias das variáveis, por exemplo, λ_{12}^{AC} refere-se à associação (interação) entre a categoria 1 da variável A e a categoria 2 da variável C. O modelo completo é dado por:

$$\upsilon_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}. \tag{2.24}$$

Para eliminar as redundâncias do modelo, de maneira a que o número de parâmetros não ultrapasse o número de células $(I \times J \times K)$, os $\lambda' s$ em 2.24 estão sujeitos às seguintes restrições:

$$\sum_i \lambda_i^A = \sum_j \lambda_j^B = \cdots = \sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = \cdots = \sum_k \lambda_{ijk}^{ABC} = 0. \tag{2.25}$$

Embora não pareça muito provável, a julgar pela complexidade do modelo, existe um algoritmo muito simples que leva à determinação dos valores de λ. Existem *softwares*, nomeadamente o SPSS, preparados para trabalhar com este tipo de modelos.

Se escrevermos

$$\upsilon_{...} = \sum_i \sum_j \sum_k \frac{\upsilon_{ijk}}{IJK} \tag{2.26}$$

de tal forma que $\upsilon_{...}$ é a média geral das log-probabilidades e

$$\upsilon_{i..} = \sum_j \sum_k \frac{\upsilon_{ijk}}{JK} \tag{2.27}$$

de modo que $\upsilon_{i..}$ é a média de todas as log-probabilidades envolvendo a categoria i da variável A, então, por substituição de 2.24 no lado direito das equações 2.26 e 2.27 obtemos:

$$\lambda_i^A = \upsilon_{i..} - \upsilon_{...}. \tag{2.28}$$

Note-se que os efeitos principais são funções das somas marginais dos logaritmos, mas não correspondem às somas marginais das probabilidades na escala original.

Assim, λ_i^A é uma medida da influência da categoria i da variável A, indicando quanto mais (ou menos) provável é a categoria i relativamente a todas as outras categorias da variável A.

Podemos obter fórmulas semelhantes para os outros $\lambda's$; por exemplo, se

$$\upsilon_{ij.} = \sum_k \frac{\upsilon_{ijk}}{K} \ , \ \upsilon_{i.k} = \sum_j \frac{\upsilon_{ijk}}{J} \ , \text{ etc.}$$

então:

$$\lambda_{ij}^{AB} = \upsilon_{ij.} - \upsilon_{i..} - \upsilon_{.j.} + \upsilon_{...} \tag{2.29}$$

$$\lambda_{ijk}^{ABC} = \upsilon_{ijk} - \upsilon_{ij.} - \upsilon_{i.k} - \upsilon_{.jk} + \upsilon_{i..} + \upsilon_{.j.} + \upsilon_{..k} - \upsilon_{...} \tag{2.30}$$

λ^{AB} representa a interação entre as variáveis A e B tomada em média relativamente a todas as tabelas das categorias de C. A relação 2.29 implica que λ_{ij}^{AB} indica até que ponto a ocorrência conjunta das categorias A_i e B_j é mais ou menos provável do que seria esperado se as variáveis A e B fossem independentes. De forma semelhante, λ^{ABC} que é por vezes chamada interação de 2^a ordem, mede a diferença na interação entre as variáveis A e B obtida para as tabelas nas categorias da variável C. λ_{ijk}^{ABC} indica até que ponto a interdependência das variáveis A e B é por si só dependente da categoria da variável C.

Assim, se alguma interação entre duas das variáveis é constante para as tabelas nas categorias da terceira variável, então a interação entre as três variáveis $\lambda^{ABC} = 0$. Em particular, se alguma interação entre duas das variáveis é zero, então a interação quanto muito é de 2^a ordem, isto é, a interação das três variáveis é nula. Isto leva-nos à definição de modelos hierárquicos. A família de modelos hierárquicos é definida como a família tal que se um qualquer termo em λ é igual a zero no modelo, todos os termos em λ com ele relacionados mas de ordem superior também devem ser iguais a zero no modelo. Assim, se $\lambda^{AB} = 0$, então $\lambda^{ABC} = 0$ (mas λ^{AC} e λ^{BC} podem ou não ser nulos). Também, se $\lambda^{AC} \neq 0$, então λ^A e λ^C têm de estar presentes no modelo.

Os dados para a situação geral $I \times J \times K$ consistem em frequências observadas $\{f_{ijk}\}$ nas várias células $\{(i, j, k)\}$ da classificação tridimensional. Escrevemos

$$y_{ijk} = \ln(f_{ijk}) \tag{2.31}$$

e obtemos as estimativas do parâmetro substituindo os $\upsilon's$ nas equações 2.29 e 2.30 pelos correspondentes $y's$. Por exemplo, o estimador de máxima verosimilhança de λ_{ij}^{AB} é:

$$\widehat{\lambda}_{ij}^{AB} = y_{ij.} - y_{i..} - y_{.j.} - y_{...} \tag{2.32}$$

onde

$$y_{ij.} = \sum_k \frac{y_{ijk}}{K} \ , \ y_{i..} = \sum_j \sum_k \frac{y_{ijk}}{JK} \ , \text{ etc.} \tag{2.33}$$

O objetivo de ajustar um modelo saturado é o de obter uma impressão da importancia relativa dos vários $\lambda's$. O modelo é aditivo nos $\lambda's$ de forma a que aqueles com valores próximos de zero tenham apenas uma ligeira importância. Equivalentemente, aqueles que diferem substancialmente de zero terão bastante importância. O que necessitamos então, é de um guia sobre o que constitui uma diferença substancial.

É fácil de perceber através de uma equação como a 2.29 que cada estimador $\widehat{\lambda}$ é uma combinação linear dos $\{y_{ijk}\}$.

As estatísticas referidas no capítulo 1, para os testes de ajustamento, nomeadamente a estatística X^2 de Pearson e a estatística G^2 do logaritmo da razão das verosimilhanças, são as tipicamente utilizadas no ajustamento dos modelos log-lineares.

O ajustamento de modelos não saturados pode levar ao recurso a processos iterativos para obtenção dos estimadores dos parâmetros do modelo. Mais uma vez, *softwares* tais como o SPSS, são uma preciosa ferramenta.

Note-se que nesta secção é feita, apenas, uma breve abordagem aos modelos log-lineares.

Capítulo 3

Medidas de Associação

3.1 Introdução

Neste capítulo serão abordados índices que descrevem a relação entre duas ou mais variáveis chamados **Medidas de Associação**. Quando devidamente usadas, estas medidas dão-nos uma descrição muito útil da estrutura das tabelas de dados em estudo.

Uma medida de associação deve fornecer uma resposta numérica simples à questão sobre a intensidade da relação entre duas variáveis. Mas, como veremos ao longo deste capítulo, medidas diferentes de associação focam-se em aspetos diferentes da relação entre as variáveis. É de notar que não existe uma razão para usar uma única medida de associação.

Grande parte dos avanços feitos sobre este tema devem-se a Goodman e Kruskal (1954[11], 1959[12], 1963[13] e 1972[14]) que afirmam que idealmente cada problema de investigação deveria possuir uma medida de associação própria, desenvolvida especificamente para as necessidades do problema em discussão. De facto, ao longo do tempo, foram sendo desenvolvidas um grande número de medidas de associação de forma que atualmente quando precisamos de utilizar alguma não existe necessidade de criar uma nova, sendo que o verdadeiro problema está em escolher sensatamente qual a melhor a utilizar em cada situação. Sendo assim, torna-se útil sumarizar as medidas de associação existentes, explicar como funcionam e em que situação devem ser aplicadas. É esse o objetivo deste capítulo.

3.1.1 Coeficiente de Correlação Linear de Pearson

Embora este trabalho tenha sido desenvolvido para a análise de dados categorizados, é fundamental apresentarmos aqui o coeficiente de correlação de Pearson para melhor entendermos os coeficientes que apresentaremos nas secções seguintes.

O coeficiente de correlação de Pearson é uma medida do grau de linearidade existente entre as variáveis aleatórias contínuas X e Y. É dado por:

$$\rho_{X,Y} = \frac{cov\,(X,Y)}{\sigma_X \sigma_Y}$$

em que $cov\,(X,Y) = E\left[(X - E\,(X))\,(Y - E\,(Y))\right]$, $\sigma_X^2 = E\left[X - E\,(X)\right]^2$ e $\sigma_Y^2 = E\left[Y - E\,(Y)\right]^2$. Varia entre -1 e 1 e é muitas vezes também chamado coeficiente de correlação do produto-momento. O estimador de $\rho_{X,Y}$ é dado por:

$$R_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]}\sqrt{\left[\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2\right]}}. \tag{3.1}$$

Embora o denominador da equação 3.1 seja sempre positivo, o numerador pode ser positivo, negativo ou igual a zero, permitindo assim que R também possa ser positivo, negativo ou zero, respetivamente.

O sinal de R indica a direção da associação entre ambas as variáveis. Um sinal positivo indica uma correlação positiva ou direta, ou seja, que um aumento no valor de uma das variáveis leva também a um aumento da outra variável; assim, na amostra que recolhemos, os valores mais elevados de uma variável terão tendência a aparecer associados aos valores mais elevados da outra variável. Um sinal negativo indica uma correlação negativa ou indireta, ou seja, que um aumento de valor de uma das variáveis seja acompanhado por um decréscimo no valor da outra variável; neste caso, na nossa amostra os valores elevados de uma das variáveis terão tendência a aparecer associados aos valores mais baixos da outra variável. Se $\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = 0$, então $R = 0$, e temos uma correlação nula o que indica que não existe uma associação linear entre as variáveis em questão (se as variáveis forem independentes então obviamente o coeficiente de correlação de Pearson é igual a zero uma vez que a covariância entre as variáveis é nula; no entanto se obtivermos um coeficiente de correlação muito próximo de zero a conclusão que devemos tirar é que a associação linear entre as variáveis é praticamente inexistente).

Quanto mais próximo de 1 estiver $|R|$, maior o grau de associação linear entre as variáveis. R não possui nenhuma unidade de medida visto que as unidades tanto de X como de Y aparecem tanto no numerador como no denominador e assim anulam-se aritmeticamente.

$\rho_{X,Y}$ mantém-se inalterado se adicionarmos constantes às variáveis ou se as variáveis forem multiplicadas por constantes com o mesmo sinal.

O quadrado do coeficiente de correlação de Pearson é o coeficiente de determinação, $R_{X,Y}^2$, que mede quanto da variabilidade total de Y é explicada por uma relação linear entre X e Y.

Exemplo[30]: Cálculo do coeficiente de correlação simples e do coeficiente de determinação para dados relativos ao comprimento de asa em cm (X) e comprimento da cauda em cm (Y) de 12 aves de uma determinada espécie.

X	Y
10.4	7.4
10.8	7.6
11.1	7.9
10.2	7.2
10.3	7.4
10.2	7.1
10.7	7.4
10.5	7.2
10.8	7.8
11.2	7.7
10.6	7.8
11.4	8.3

$$\begin{aligned}
\sum_{i=1}^{n} x_i &= 128.2 \text{ cm} \quad \sum_{i=1}^{n} y_i = 90.8 \text{ cm} \\
\sum_{i=1}^{n} x_i^2 &= 1371.32 \text{ cm}^2 \quad \sum_{i=1}^{n} y_i^2 = 688.40 \text{ cm}^2 \quad \sum_{i=1}^{n} x_i y_i = 971.37 \text{ cm}^2
\end{aligned}$$

$$\begin{aligned}
R_{X,Y} &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}} = \\
&= \frac{971.37 - \frac{128.2 \times 90.8}{12}}{\sqrt{\left[1371.32 - \frac{(128.2)^2}{12}\right]\left[688.40 - \frac{(90.8)^2}{12}\right]}} = \\
&= \frac{1.323}{\sqrt{1.717 \times 1.347}} = \frac{1.323}{1.521} = 0.87
\end{aligned}$$

O valor obtido para R é uma estimativa do coeficiente de correlação $\rho_{X,Y}$. Se queremos averiguar se existe de facto uma correlação entre X e Y na população, podemos testar $H_0 : \rho = 0$. Para testar esta hipótese, a estatística utilizada é:

$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \frown T_{n-2}. \tag{3.2}$$

Se queremos testar a hipótese $H_0 : \rho = \rho_0$ com $\rho_0 \neq 0$, a estatística de teste a utilizar é:

$$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) = \arctan R \tag{3.3}$$

que tende assintoticamente para uma variável aleatória Normal, cujo valor médio é:

$$E(Z) \cong \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) + \frac{\rho_0}{2(n-1)} \tag{3.4}$$

e cuja variância é:

$$var(Z) \cong \frac{1}{n-3}. \tag{3.5}$$

A aproximação à Normal é muito satisfatória para amostras de dimensão 20 a 25 no mínimo, mas é frequentemente utilizada para amostras mais pequenas. Desde que a dimensão das amostras não seja inferior a 20, podemos utilizar como valor médio de Z:

$$E(Z) \cong \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right). \tag{3.6}$$

Assim, a nossa estatística de teste será :

- para $n \geq 20$:

$$\frac{Z - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}} \frown N(0,1)$$

- para $n < 20$:

$$\frac{Z - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) + \frac{\rho_0}{2(n-1)}}{\frac{1}{\sqrt{n-3}}} \frown N(0,1)$$

Comparando dois coeficientes de correlação

Hipóteses sobre dois coeficientes de correlação podem ser testadas usando:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} \tag{3.7}$$

onde

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \tag{3.8}$$

Se $n_1 = n_2$, então a equação anterior torna-se

$$\sigma_{z_1-z_2} = \sqrt{\frac{2}{n-3}} \tag{3.9}$$

onde n é o tamanho de cada amostra. O procedimento mais utilizado para um teste bi-lateral à hipótese $H_0 : \rho_1 = \rho_2 \;\; vs \;\; H_1 = \rho_1 \neq \rho_2$ é utilizando a equação 3.7 como demonstrado no exemplo seguinte:

Exemplo[30]: Para uma amostra de 98 comprimentos de asas e caudas de uma espécie de ave, um coeficiente de correlação igual a 0.78 foi calculado. Uma segunda amostra de 95 medidas iguais de uma outra espécie, deu um coeficiente de correlação igual a 0.84. Vamos testar se existe igualdade dos coeficientes de correlação das duas populações.

$H_0 : \rho_1 = \rho_2$

$H_1 : \rho_1 \neq \rho_2$

$$\begin{array}{rcll} R_1 & = & 0.78 & R_2 = 0.84 \\ z_1 & = & 1.0454 & z_2 = 1.2212 \\ n_1 & = & 98 & n_2 = 95 \end{array}$$

$$Z = \frac{1.0454 - 1.2212}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} = \frac{-0.1758}{0.1463} = -1.202$$

$Z_{0.05(2)} = 1.960$

Não rejeitamos H_0.

Coeficiente de correlação comum Tal como no exemplo anterior, concluir que $\rho_1 = \rho_2$ poderá levar a dizer que ambas as amostras provêm da mesma população ou de populações com coeficientes de correlação idênticos. Neste caso, é possível combinar informação de ambas as amostras para calcular uma melhor estimativa para o único ρ. Podemos designar esta estimativa por coeficiente de correlação comum ou ponderado, obtido convertendo

$$z_w = \frac{(n_1 - 3)\, z_1 + (n_2 - 3)\, z_2}{(n_1 - 3) + (n_2 - 3)} \tag{3.10}$$

no seu respectivo valor R, que designaremos por R_w. Se ambas as amostras têm o mesmo tamanho, então a equação anterior simplifica para:

$$z_w = \frac{z_1 + z_2}{2}. \tag{3.11}$$

A conversão de z_w no coeficiente de correlação comum R_w pode fazer-se com recurso a uma calculadora ou numa folha de cálculo através da relação $R = \tanh z$.

Se temos que $\rho < 0.5$, uma melhor estimativa para o parâmetro obtém-se utilizando

$$z_w = \frac{(n_1 - 1)\, z_1^{'} + (n_2 - 1)\, z_2^{'}}{(n_1 - 1) + (n_2 - 1)} \tag{3.12}$$

onde

$$z_i' = z_i - \frac{3z_i + R_i}{4\,(n_i - 1)} \tag{3.13}$$

e z_i é dado pela equação:

$$z_i = 0.5 \ln\left(\frac{1 + R}{1 - R}\right). \tag{3.14}$$

O coeficiente de correlação comum para o exemplo dado anteriormente pode ser calculado por:

$$z_w = \frac{(n_1 - 3)z_1 + (n_2 - 3)\, z_2}{(n_1 - 3) + (n_2 - 3)} = \frac{95 \times 1.0454 + 92 \times 1.2212}{95 + 92} = 1.1319$$

Convertendendo em R_w dá

$$R_w = 0.81$$

Comparando mais do que dois coeficientes de correlação

Se temos k amostras e para cada uma delas foi calculado o respectivo R, é muitas vezes útil perceber se todas as amostras provêm de populações com ρ's idênticos ou não. Se não rejeitamos a hipótese $H_0 : \rho_1 = \rho_2 = ... = \rho_k$ então é possível combinar todas as amostras calculando apenas um R para estimar um único ρ. Testar esta hipótese implica transformar cada R num valor z. Depois podemos então calcular

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3)\, z_i^2 - \frac{\left[\sum_{i=1}^{k} (n_i - 3)\, z_i\right]^2}{\sum_{i=1}^{k} (n_i - 3)} \tag{3.15}$$

o qual pode ser considerado um valor da variável aleatória Qui-quadrado com $k - 1$ graus de liberdade.

Coeficiente de correlação comum Se H_0 não é rejeitada, então todos os k coeficientes de correlação das amostras são combinados para estimar um ρ comum à população. Um R comum pode ser obtido a partir da média ponderada de z,

$$z_w = \frac{\sum_{i=1}^{k} (n_i - 3)\, z_i}{\sum_{i=1}^{k} (n_i - 3)} \tag{3.16}$$

à qual corresponde o valor de R que designaremos por R_w. Se H_0 não for rejeitada, podemos testar $H_0 : \rho = 0 \;\; vs \;\; H_1 : \rho \neq 0$ pelo método atribuído a Neyman (1959) por Paul (1988):

$$Z = \frac{\sum_{i=1}^{k} n_i r_i}{\sqrt{N}} \tag{3.17}$$

onde $N = \sum_{i=1}^{k} n_i$, rejeitando H_0 quando $|Z| \geq Z_{\alpha(2)}$. Para um teste uni-lateral, $H_0 : \rho = 0 \;\; vs \;\; H_1 : \rho > 0$, rejeitamos H_0 se $Z \geq Z_{\alpha(1)}$ e $H_0 : \rho = 0 \;\; vs \;\; H_1 : \rho < 0$ é rejeitada se $Z \leq -Z_{\alpha(1)}$.

Exemplo[30]: Teste de hipóteses relativas a coeficientes de correlação de três variáveis aleatórias. Consideremos que temos três amostras de dimensões: $n_1 = 24$; $n_2 = 29$ e $n_3 = 32$ e representemos os coeficientes de correlação entre as três variáveis aleatórias por: $R_1 = 0.52$; $R_2 = 0.56$ e $R_3 = 0.87$.

Queremos testar:
$H_{0:}\rho_1 = \rho_2 = \rho_3$
$H_1 : \exists i, j : \rho_i \neq \rho_j$

i	R_i	z_i	z_i^2	n_i	$n_i - 3$	$(n_i - 3)\, z_i$	$(n_i - 3)\, z_i^2$
1	0.52	0.5763	0.3321	24	21	12.1023	6.9741
2	0.56	0.6328	0.4004	29	26	16.4528	10.4104
3	0.87	1.3331	1.7772	32	29	38.6599	51.5388
$\sum$					76	67.2150	68.9233

$$X^2 = \sum (n_i - 3)\, z_i^2 - \frac{\left[\sum (n_i - 3)\, z_i\right]^2}{\sum (n_i - 3)} = 68.9233 - \frac{(67.2150)^2}{76} = 9.478$$

$g.l. = k - 1 = 2$
$\chi^2_{0.05}\,(2) = 5.991$

Desta forma rejeitamos H_0.

Caso H_0 não tivesse sido rejeitada, teria sido útil calcular o coeficiente de correlação comum:

$$z_w = \frac{\sum (n_i - 3)\, z_i}{\sum (n_i - 3)} = \frac{67.2150}{76} = 0.884$$

Que corresponde a:

$$R_w = 0.71$$

3.2 Coeficiente de correlação ordinal de Spearman

O coeficiente de correlação de Spearman (ρ_s estimado por R_s), desenvolvido por Spearman (1904), é uma medida de associação não paramétrica entre duas variáveis ordinais. Pode ser usado como medida de correlação para alguns tipos de dados não numéricos. É um caso particular do coeficiente de correlação linear de Pearson, caso este fosse utilizado para duas variáveis ordinais, usando as ordens das observações (*ranks*). Temos n indivíduos e cada um deles contribui com informação para duas variáveis X e Y sendo que uma das seguintes situações é verdadeira:

- ambas as variáveis já estão organizadas numa escala ordinal;
- os dados originais provêm de uma variável na escala ordinal e de uma outra variável contínua que é depois transformada em categorias que podem ser representadas numa escala ordinal;
- os dados dizem respeito a duas variáveis contínuas que são depois transformadas em categorias que podem ser representadas numa escala ordinal e para as quais as inferências sobre o coeficiente de Pearson não podem ser calculadas visto não se confirmar algum dos seus pressupostos.

Em vez de trabalharmos com as observações tal e qual foram recolhidas, utilizaremos no cálculo do coeficiente de correlação de Spearman as ordens das observações.

O coeficiente de correlação ordinal de Spearman determina o grau de relação monotónica existente entre duas variáveis. Uma relação monotónica pode ser descrita como:

- crescente, onde um aumento numa das variáveis leva a um aumento na outra, ou seja, as ordens mais elevadas numa das variáveis estão associadas às ordens mais elevadas na outra (o que está associado a uma correlação positiva);
- decrescente, onde um aumento numa das variáveis leva a uma diminuição na outra, ou seja, as ordens mais elevadas de uma das variáveis estão associadas às ordens mais baixas da outra (associada a correlação negativa).

Após se ter atribuído um *rank* a cada observação de uma variável, a fórmula do coeficiente de correlação de Pearson poderá ser aplicada aos *ranks* para assim obtermos o coeficiente de correlação entre *ranks* de Spearman (R_s). Existe, no entanto, uma forma mais simples de cálculo:

$$R_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{3.18}$$

onde d_i é a diferença entre *ranks* ($d_i = rank\ X_i - rank\ Y_i$) nos pares de observações.

O valor de R_s, que pode ser considerado uma estimativa do coeficiente de correlação da população, ρ_s, pode variar entre -1 e 1, não tem unidades de medida e o grau de

associação é tanto mais forte quanto mais próximo de -1 ou de $+1$ R_s se encontra. O sinal de R_s indica a direção da relação monotónica: sinal positivo indica uma relação monotónica crescente e um sinal negativo indica uma relação monotónica decrescente. Quanto mais próximo de 0, menor é o grau de associação entre as variáveis.

Exemplo[30]: Relação entre as notas obtidas em testes de aptidão de Matemática (X) e Biologia (Y) de 10 estudantes.

Estudante (i)	(X_i)	***rank*** **de** X_i	(Y_i)	***rank*** **de** Y_i	$\mathbf{d}_i$	$\mathbf{d}_i^2$
1	57	4	83	7	-3	9
2	45	1	37	1	0	0
3	72	7	41	2	5	25
4	78	8	84	8	0	0
5	53	2	56	3	-1	1
6	63	5	85	9	-4	16
7	86	9	77	6	3	9
8	98	10	67	10	0	0
9	59	4	70	5	-1	1
10	71	6	59	4	2	4
						$\sum d_i^2 = 64$

$$\begin{aligned} R_s &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \\ r_S &= 1 - \frac{6 \times 64}{10^3 - 10} = 1 - 0.388 = 0.612 \end{aligned}$$

3.2.1 Correção para empates

Se temos empates nos *ranks* então atribuímos a cada uma das observações envolvidas no empate a média dos *ranks* correspondentes. Neste caso, R_s será mais corretamente calculado pela equação de R aplicada aos *ranks*, ou então por:

$$(R_s)_c = \frac{(n^3 - n)/6 - \sum d_i^2 - \sum t_X - \sum t_Y}{\sqrt{\left[(n^3 - n)/6 - 2\sum t_X\right]\left[(n^3 - n)/6 - 2\sum t_Y\right]}}$$

onde

$$\sum t_X = \frac{\sum (t_i^2 - t_i)}{12}$$

com t_i o número de valores iguais a x_i. O mesmo se aplica a Y:

$$\sum t_Y = \frac{\sum (t_i^2 - t_i)}{12}$$

Exemplo[30]:

X	**rank**	Y	**rank**	d_i	d_i^2
10.4	4	7.4	5	-1	1
10.8	8.5	7.6	7	1.5	2.25
11.1	10	7.9	11	-1	1
10.2	1.5	7.2	2.5	-1	1
10.3	3	7.4	5	-2	4
10.2	1.5	7.1	1	0.5	0.25
10.7	7	7.4	5	2	5
10.5	5	7.2	2.5	2.5	6.25
10.8	8.5	7.8	9.5	-1	1
11.2	11	7.7	8	3	9
10.6	6	7.8	9.5	-3.5	12.25
11.4	12	8.3	12	0	0
					$\sum d_i^2 = 42$

$H_0 : \rho_s = 0 \;\; vs \;\; H_1 : \rho_s \neq 0$

$$\begin{aligned} R_s &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \\ r_S &= 1 - \frac{6 \times 420}{1716} = 1 - 0.147 = 0.853 \end{aligned}$$

Aplicando a correção para empates:

Em X: $\sum t_X = \frac{\left(2^3-2\right)+\left(2^3-2\right)}{12} = 1$, correspondente às 2 observações com o valor 10.2 e 2 observações com o valor 10.8 (ordens 1.5 e 8.5).

Em Y: $\sum t_Y = \frac{\left(2^3-2\right)+\left(3^3-3\right)+\left(2^3-2\right)}{12} = 12$, correspondente às 2 observações com o valor 7.2, 3 com o valor 7.4 e 2 com o valor 7.8 (ordens 2.5, 5 e 9.5).

Logo

$$(r_s)_c = \frac{\left(12^3 - 12\right)/6 - 42 - 1 - 3}{\sqrt{\left[\left(12^3 - 12\right)/6 - 2 \times 1\right]\left[\left(12^3 - 12\right)/6 - 2 \times 1\right]}} = \frac{240}{284} = 0.845$$

Quando $\sum t_X$ e $\sum t_Y$ são iguais a 0 a equação de $(R_s)_c$ torna-se igual à de r_s. Na realidade só existem diferenças quando temos um grande número de empates.

O coeficiente de correlação de Spearman pode ser usado como uma estatística de teste para testar a independência entre duas variáveis aleatórias. No entanto, na hipótese alternativa deverá ser especificado qual o tipo de dependência que pode ser detetado, por exemplo, uma relação monotónica crescente.

3.3 Coeficiente de Correlação Ponto Bisserial

O coeficiente de correlação ponto bisserial representa um caso especial do coeficiente de correlação linear de Pearson. Este coeficiente é utilizado quando estamos na presença de uma variável contínua e a outra variável é categorizada dicotómica.

Um problema de alguma importância em aplicações estatísticas consiste em encontrar uma medida de associação entre uma variável aleatória discreta X, que toma valores 0 e 1, e uma variável aleatória contínua Y. O usual coeficiente de correlação produto-momento, $\rho_{X,Y}$, costuma ser usado para este fim. Recebeu então nome de coeficiente de correlação ponto bisserial devido à sua relação com o coeficiente de correlação bisserial proposto por Pearson para um problema semelhante.

A expressão deste coeficiente é dada por:

$$R_{pb} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}. \tag{3.19}$$

Uma alternativa a esta equação é:

$$R_{pb} = \left[\frac{\overline{Y}_1 - \overline{Y}_0}{\widetilde{S}_Y}\right]\sqrt{p_0 p_1}\sqrt{\frac{n}{n-1}} \tag{3.20}$$

onde:

- $\overline{Y}_1$ e $\overline{Y}_0$ são, respectivamente, os valores médios da variável Y para os indivíduos categorizados com o valor 1 e 0 na variável X;
- p_0 e p_1 são, respectivamente, as proporções de indivíduos categorizados como 0 e 1 na variável X;
- $\widetilde{S}_Y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n-1}}$, ou seja, o desvio-padrão da variável aleatória Y.

O valor do coeficiente de correlação ponto bisserial encontra-se no intervalo $[-1, 1]$. O sinal de R_{pb} não é relevante a não ser que a variável dicotomizada esteja ordenada, ou seja, utiliza-se o valor absoluto de R_{pb} pois o facto da variável dicotómica tomar valor 0 e 1 não significa que uma categoria seja superior à outra.

O quadrado do coeficiente de correlação ponto bisseral é o coeficiente de determinação, R^2_{pb}, que mede quanto da variabilidade total de Y é explicada pela variabilidade de X.

Exemplo[26]: É feito um estudo para determinar se a destreza manual (variável X) e a coordenação olho-mão (variável Y) estão correlacionadas. O estudo integra 10 indivíduos, 5 dextros e 5 canhotos, os quais foram sujeitos a um teste para avaliar o grau de coordenação olho-mão. Quanto maior o valor, melhor a coordenação. Os dextros estão codificados com o valor 1 e os canhotos com o valor 0.

Indivíduo	X	X^2	Y	Y^2	XY
1	1	1	11	121	11
2	1	1	1	1	1
3	1	1	0	0	0
4	1	1	2	4	2
5	1	1	0	0	0
6	0	0	11	121	0
7	0	0	11	121	0
8	0	0	5	25	0
9	0	0	8	64	0
10	0	0	4	16	0
$\sum$	5	5	53	473	14

$$R_{pb} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

$$\begin{aligned} r_{pb} &= \frac{14 - \frac{5\times 53}{10}}{\sqrt{\left[5 - \frac{5^2}{10}\right]\left[473 - \frac{53^2}{10}\right]}} = \frac{14 - 26.5}{\sqrt{2.5 \times 192.1}} = \\ &= \frac{-12.5}{21.91} = -0.57 \end{aligned}$$

$$r_{pb}^2 = (-0.57)^2 = 0.325$$

pelo que podemos afirmar que 32.5% da variabilidade de Y é explicada pela variável X.

3.4 Coeficiente de Correlação Bisserial

O coeficiente de correlação bisserial (R_b) é uma medida de associação entre duas variáveis que é utilizado quando ambas as variáveis são contínuas mas uma delas foi dicotomizada. O valor obtido para este coeficiente é uma estimativa do valor que seria obtido caso fosse calculado o coeficiente de correlação linear de Pearson sem que uma das variáveis fosse dicotomizada.

O coeficiente de correlação bisserial é baseado no pressuposto de que a distribuição subjacente a ambas as variáveis é contínua e Normal. Uma vez que a precisão de R_b depende fortemente da normalidade dos dados, ele não deve ser utilizado a não ser que haja uma forte evidência da normalidade da variável que é dicotomizada. Caso a distribuição da variável dicotómica se desvie consideravelmente da normalidade o valor obtido de R_b não será uma estimativa fiável da correlação. Uma consequência da

violação deste pressuposto é que o valor máximo de R_b excederá o valor 1. Lindeman et al. (1980) refere que os limites teóricos de R_b são $-\infty < R_b < +\infty$. Neste caso é aconselhável utilizar o coeficiente de correlação ponto bisserial, o qual apresentamos na secção anterior, ao invés do coeficiente de correlação bisserial.

Ao contrário do coeficiente de correlação ponto bisserial, o sinal de R_b deverá ser tomado em atenção uma vez que clarifica a natureza da relação entre as variáveis. Ao atribuir os valores 0 e 1 à variável dicotomizada, devemos ter em atenção o seguinte: o valor 0 deverá ser atribuído ao desempenho/valor/qualidade mais baixo enquanto que o valor 1 deverá ser atribuído ao desempenho/valor/qualidade mais alto.

O coeficiente de correlação bisserial pode ser obtido utilizando uma das seguintes equações:

$$R_b = \left[\frac{\overline{Y}_1 - \overline{Y}_0}{s_Y}\right]\left[\frac{p_0 p_1}{h}\right] \tag{3.21}$$

$$R_b = \left[\frac{\overline{Y}_1 - \overline{Y}_0}{\widetilde{s}_Y}\right]\left[\frac{p_0 p_1}{h}\right]\sqrt{\frac{n}{n-1}} \tag{3.22}$$

$$R_b = \frac{r_{pb}\sqrt{p_0 p_1}}{h} \tag{3.23}$$

onde:

- p_0 representa a proporção de casos dicotomizados com o valor 0;
- p_1 representa a proporção de casos dicotomizados com o valor 1;
- $\overline{Y}_1$ e $\overline{Y}_0$ são, respectivamente, os valores médios da variável Y para os indivíduos categorizados com o valor 1 e 0 na variável X;
- h é o valor da função densidade da variável aleatória Normal reduzida no ponto z_{p_1}, isto é, o valor da variável aleatória normal reduzida que deixa para a sua direita uma área igual a p_1.Tendo em consideração que a função densidade da variável aleatória normal reduzida é dada por $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$.
- $s_Y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n}}$;
- $\widetilde{s}_Y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n-1}}$.

A utilização da equação 3.23 só deve acontecer quando também foi calculado R_{pb} para os mesmos dados.

O erro padrão de R_b é dado por:

$$S_{R_b} = \frac{\frac{\sqrt{p_0 p_1}}{h} - R_b^2}{\sqrt{n}}. \tag{3.24}$$

Para o mesmo conjunto de dados, o valor absoluto de R_b será sempre maior que R_{pb} pois o valor de $\frac{\sqrt{p_0 p_1}}{h}$ é sempre superior a 1. Quanto mais próximos p_0 e p_1 estiverem, menor será a discrepância entre R_b e $R_{pb.}$

Exemplo[26]: É conduzido um estudo com vista a perceber se existe relação entre a coordenação olho-mão (Y) e o grau de inteligência (X). Cinco indivíduos com uma inteligência acima da média (1) e outros cinco com inteligência abaixo da média (0) foram sujeitos a um teste para avaliar a coordenação olho-mão, onde quanto maior o valor obtido, melhor a coordenação.

Indivíduo	X	Y	Y^2
1	1	11	121
2	1	1	1
3	1	0	0
4	1	2	4
5	1	0	0
6	0	11	121
7	0	11	121
8	0	5	25
9	0	8	64
10	0	4	16
$\sum$	5	53	473

$$\overline{y}_0 = \frac{11+11+5+8+4}{5} = 7.8$$

$$\overline{y}_1 = \frac{11+1+0+2+0}{5} = 2.8$$

$$\begin{aligned} S_Y &= \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n}} \\ s_Y &= \sqrt{\frac{473 - \frac{53^2}{10}}{10}} = \sqrt{\frac{192.1}{10}} = \\ &= 4.38 \end{aligned}$$

$$\begin{aligned} \widetilde{S}_Y &= \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n-1}} \\ \widetilde{s}_Y &= \sqrt{\frac{473 - \frac{53^2}{10}}{10-1}} = \sqrt{\frac{192.1}{9}} = \\ &= 4.62 \end{aligned}$$

Como $p_0 = 0.5$ e $p_1 = 0.5$, $z_{p_1} = 0$ e consequentemente $h = 0.3989$ que é o valor da função densidade da variável aleatória normal reduzida no ponto 0.

$$\begin{aligned} R_b &= \left[\frac{\overline{Y}_1 - \overline{Y}_0}{s_Y}\right]\left[\frac{p_0 p_1}{h}\right] \\ r_b &= \left[\frac{7.8 - 2.8}{4.38}\right]\left[\frac{0.5 \times 0.5}{0.3989}\right] = \\ &= 0.71544 \end{aligned}$$

ou

$$\begin{aligned} R_b &= \left[\frac{\overline{Y}_1 - \overline{Y}_0}{\widetilde{s}_Y}\right]\left[\frac{p_0 p_1}{h}\right]\sqrt{\frac{n}{n-1}} \\ r_b &= \left[\frac{7.8 - 2.8}{4.62}\right]\left[\frac{0.5 \times 0.5}{0.3989}\right]\sqrt{\frac{10}{10-1}} = \\ &= 0.71496 \end{aligned}$$

Note-se que a tabela de dados para as variáveis X e Y é exatamente a mesma neste exemplo e no exemplo da secção anterior. Tínhamos obtido um valor para o coeficiente de correlação ponto bisserial de -0.57 enquanto que o valor do coeficiente de correlação bisserial que obtivemos neste exemplo foi de 0.715. Não são comparáveis atendendo à diferença no intervalo de variação destes dois coeficientes e também por os pressupostos serem diferentes.

3.4.1 Coeficiente de Correlação Rank-Bisserial

O coeficiente de correlação rank-bisserial é muito semelhante ao coeficiente de correlação bisserial, só que em vez de termos uma das variáveis dicotómica e a outra variável contínua, temos uma variável dicotómica e a outra variável é ordinal. Este coeficiente varia entre -1 e 1.

A equação para o cálculo do coeficiente de correlação rank-bisserial é dada por:

$$R_{rb} = 2 \times \frac{\left(\overline{Y}_1 - \overline{Y}_0\right)}{n}$$

onde

- n é o número de pares de dados;
- $\overline{Y}_1$ e $\overline{Y}_0$ são, respectivamente, os valores médios da variável Y para os indivíduos categorizados com o valor 1 e 0 na variável X.

3.5 Coeficiente de Correlação Tetracórico

O coeficiente de correlação Tetracórico (R_{tet}) é utilizado quando estamos perante duas variáveis contínuas que foram transformadas em variáveis dicotómicas. O valor obtido para este coeficiente é uma estimativa do valor que se obteria para o coeficiente de correlação de Pearson caso se utilizassem as variáveis sem as dicotomizar.

O coeficiente de correlação Tetracórico é baseado no pressuposto de que as distribuições subjacentes a ambas as variáveis são contínuas e Normais. Vários autores sugerem que seja tomado em atenção que este coeficiente e o coeficiente de correlação bisserial são baseados em distribuições hipotéticas subjacentes às variáveis que não são diretamente observadas e que portanto deverão ser usados com cautela. Uma vez que a precisão de R_{tet} depende fortemente da normalidade das variáveis envolvidas, ele não deverá ser empregue caso não haja forte evidência de que é esse o caso.

Uma vez que a magnitude do erro padrão da estimativa de R_{tet} é maior que a magnitude do erro padrão da estimativa de R, o tamanho da amostra utilizada para calcular R_{tet} deverá ser relativamente grande, sendo que alguns autores afirmam que deverá ser duas vezes maior que o tamanho da amostra utilizada para calcular o coeficiente de correlação linear de Pearson. R_{tet} varia entre -1 e 1.

Duas expressões geralmente usadas para calcular o valor do coeficiente de correlação tetracórico são:

$$R_{tet} = \sin\left[90^o\left(\frac{n_{11}+n_{22}-n_{12}-n_{21}}{N}\right)\right] \tag{3.25}$$

e

$$R_{tet} = \cos\left(\frac{180^o}{1+\sqrt{\frac{n_{11}n_{22}}{n_{12}n_{21}}}}\right). \tag{3.26}$$

A equação 3.25 é mais adequada quando $p_0 = p_1 = 0.5$ para ambas as variáveis dicotómicas enquanto que a equação 3.26 é mais adequada para quando $p_0 \neq p_1$.

No caso de R_{tet} ter sinal negativo significa que um indivíduo classificado numa categoria de uma das variáveis estará mais provavelmente classificado na outra categoria da outra variável, um sinal positivo indica que existe a tendência para os indivíduos serem classificados na mesma categoria em ambas as variáveis.

Exemplo[26]: Foi pedido a 200 pessoas que respondessem Concordo (1) ou Discordo (0) às seguintes perguntas: "Acredito que o aborto deveria ser legalizado"(X) e "Acredito que assassinos devem ser condenados à morte"(Y). Os resultados obtidos encontram-se na tabela seguinte:

		X		
		0	**1**	**Total**
Y	**0**	30	70	100
	1	60	40	100
	Total	90	110	200

Como $p_{0_X} = (30+60)/200 = 0.45$ e $p_{1_X} = (70+40)/200 = 0.55$ são bastante aproximadas e $p_{0_Y} = (30+70)/200 = 0.5$ e $p_{1_Y} = (60+40)/200 = 0.5$ são iguais, ambas as equações terão valores aproximados:

$$\begin{aligned} R_{tet} &= \sin\left[90^o\left(\frac{n_{11}+n_{22}-n_{12}-n_{21}}{N}\right)\right] \\ r_{tet} &= \sin\left[90^o\left(\frac{30+40-70-60}{200}\right)\right] = \\ &= \sin\left[90^o\left(-0.3\right)\right] = \sin -27^o = -0.45 \end{aligned}$$

$$\begin{aligned} R_{tet} &= \cos\left(\frac{180^o}{1+\sqrt{\frac{n_{11}n_{22}}{n_{12}n_{21}}}}\right) \\ r_{tet} &= \cos\left(\frac{180^o}{1+\sqrt{\frac{30\times 40}{70\times 60}}}\right) = \cos\left(\frac{180^o}{1.53}\right) = \\ &= \cos 117.65^o = -0.46 \end{aligned}$$

O sinal negativo indica que os indivíduos que responderam "Concordo"a uma das perguntas, mais provavelmente terão respondido "Discordo"à outra pergunta, o que indica existir uma correlação negativa ou inversa.

O coeficiente de correlação Tetracórico foi generalizado ao caso em que as variáveis X e Y têm r e s categorias ordinais, respetivamente. Este coeficiente de correlação é chamado coeficiente de correlação Policórico. Olsson (1979) apresentou o estimador de máxima verosimilhança para este coeficiente.

3.6 Coeficiente de correlação V de Cramer

Este coeficiente não-paramétrico é indicado para medir a intensidade da associação entre duas variáveis nominais que são normalmente organizadas em tabelas de contingência:

		Variável 1			
		1	**2**	$\cdots$	**c**
	1	O_{11}	O_{12}	$\cdots$	O_{1C}
Variável	**2**	O_{21}	O_{22}	$\cdots$	O_{2C}
2	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	r	O_{L1}	O_{L2}	$\cdots$	O_{LC}

O coeficiente V de Cramer é um coeficiente calculado através da fórmula:

$$V = \sqrt{\frac{X^2}{n\left[\min\left(r,c\right)-1\right]}} \tag{3.27}$$

onde $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ é a estatística do Qui-quadrado e n corresponde ao número total de observações

Este coeficiente varia entre 0 e 1. Quando as variáveis são independentes $V = 0$. V é tanto maior quanto maior for a associação entre as variáveis. Ao contrário da estatística X^2, o coeficiente V pode ser aplicado para comparar tabelas de contingência de dimensão diferente ou baseadas em amostras de dimensão diferente.

O valor $V = 1$, não significa que exista correlação perfeita entre as variáveis. Isso só acontece quando o número de linhas e o número de colunas são iguais.

O coeficiente de Cramer está sujeito aos mesmos pressupostos do teste do qui-quadrado se pretendermos testar o seu significado. Este coeficiente não deve ser comparado diretamente com outros. Se os dados forem ordinais podemos calcular o coeficiente de Cramer mas não devemos comparar diretamente o seu valor com o valor do coeficiente de Pearson. Embora o valor do coeficiente aumente com o grau de associação, as diferenças na magnitude não têm uma interpretação direta.

A partir do coeficiente V de Cramer podemos efetuar um teste às hipóteses:
H_0 : As variáveis são independentes. $\quad vs \quad$ H_1 : As variáveis são dependentes.

No caso de varáveis dicotómicas o coeficiente V de Cramer pode ser substituído pelo coeficiente Phi (ϕ) (secção 3.7.2).

3.7 Coeficientes de Correlação para dados dicotómicos

Muito frequentemente surgem situações em que o objetivo é estudar a associação entre duas variáveis dicotómicas, como por exemplo em áreas da Biologia ou da Medicina. Temos os dados dispostos em tabelas de contingência bidimensionais, tabelas estas que nos dão informação sobre as frequências conjuntas das variáveis.

3.7.1 Coeficiente de Contingência (C)

O coeficiente de contingência (também conhecido por coeficiente de contigência de Pearson) é uma medida de associação que pode ser aplicada a tabelas de contingência $r \times c$ em que $r \geq 2$ e $c \geq 2$. O valor deste coeficiente é dado pela equação:

$$C = \sqrt{\frac{X^2}{X^2+n}} \tag{3.28}$$

onde X^2 é o valor da estatística do Qui-quadrado para a tabela de contingência dada e n o número total de observações.

Uma vez que n é sempre diferente de zero, C nunca terá valor 1, mesmo que exista uma associação perfeita entre as variáveis. Consequentemente $0 \leq C < 1$. Uma limitação deste coeficiente é que o seu limite superior é uma função do número de linhas e colunas da tabela $r \times c$ e é dado por:

$$C_{\max} = \sqrt{\frac{k-1}{k}} \tag{3.29}$$

onde $k = \min(r, c)$. Além disso, coeficientes de contigência que tenham sido calculados para diferentes tabelas, só poderão ser comparados se as tabelas tiverem a mesma dimensão. Uma forma de dar a volta a estes problemas consiste em utilizar um coeficiente ajustado:

$$C_{aj} = \frac{C}{C_{\max}}$$

e assim, quando existir associação perfeita entre as variáveis, teremos $C = C_{\max}$ e, consequentemente, $C_{aj} = 1$. Logo $0 \leq C_{aj} \leq 1$.

Exemplo[26]: Duzentas pessoas escolhidas aleatoriamente são sujeitas a um teste de inteligência para o qual dispõem de uma hora. 100 destas pessoas efetuam este teste expostas a um ruído contínuo sendo-lhes dito que este deve-se a uma avaria num gerador. As outras 100 fazem o teste sem nenhum tipo de ruído de fundo. À saída do teste um idoso com um braço engessado pede a cada uma das pessoas ajuda para levar uma embalagem pesada até ao seu carro. Os resultados foram os seguintes:

		Ajuda ao idoso		
		Sim	**Não**	**Total**
Expostos ao ruído	**Sim**	30	70	100
	Não	60	40	100
	Total	90	110	200

Para tabelas 2×2 o valor máximo de C será:

$$C_{\max} = \sqrt{\frac{k-1}{k}} = \sqrt{\frac{2-1}{2}} = 0.71$$

As frequências esperadas são:

$$\begin{aligned}
E_{11} &= \frac{n_{1.} \times n_{.1}}{n} = \frac{100 \times 90}{200} = 45 \\
E_{12} &= \frac{n_{1.} \times n_{.2}}{n} = \frac{100 \times 110}{200} = 55 \\
E_{21} &= \frac{n_{2.} \times n_{.1}}{n} = \frac{100 \times 90}{200} = 45 \\
E_{22} &= \frac{n_{2.} \times n_{.2}}{n} = \frac{100 \times 110}{200} = 55
\end{aligned}$$

Visto tratar-se de uma tabela 2×2, vamos utilizar a fórmula com a correção de Yates:

$$\begin{aligned} X^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}} = \\ &= \frac{(|30-45|-0.5)^2}{45} + \frac{(|70-55|-0.5)^2}{55} + \frac{(|60-45|-0.5)^2}{45} + \frac{(|40-55|-0.5)^2}{55} = \\ &= 4.67 + 3.82 + 4.67 + 3.82 = 16.98 \end{aligned}$$

Neste caso:

$$\begin{aligned} C &= \sqrt{\frac{X^2}{X^2+n}} = \sqrt{\frac{16.98}{16.98+200}} = \\ &= \sqrt{0.078} = 0.28 \end{aligned}$$

Pelo que é possível dizer que a associação entre as variáveis é fraca.

$$C_{aj} = \frac{C}{C_{\max}} = \frac{0.28}{0.71} = 0.39$$

o que reforça a conclusão tirada anteriormente.

3.7.2 Coeficiente ϕ

O coeficiente ϕ é um coeficiente muito semelhante ao V de Cramer e também ao coeficiente de contingência C. Em geral possuem valores muito próximos. Foi proposto inicialmente para tabelas de contingência 2×2, ou seja, para variáveis dicotómicas.

Neste caso podemos efetuar o teste de independência entre as duas variáveis a partir do coeficiente ϕ. O teste é baseado no teste exacto de Fisher, fornecendo valores mais exactos que os do coeficiente V de Cramer.

O coeficiente ϕ^2 está relacionado com a estatística do Qui-quadrado, tendo-se

$$\phi^2 = \frac{X^2}{N} \tag{3.30}$$

sendo N o número total de observações.

Consequentemente,

$$\phi = \sqrt{\frac{X^2}{N}}. \tag{3.31}$$

Calculando o valor de X^2 e extraindo a raíz quadrada com o sinal de $(n_{11}n_{22} - n_{12}n_{21})$, a fórmula seguinte é ainda equivalente:

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}} \tag{3.32}$$

no caso de utilizarmos frequências absolutas. Para o caso de usarmos proporções fica:

$$\phi = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{(p_{11}+p_{12})(p_{21}+p_{22})(p_{11}+p_{21})(p_{12}+p_{22})}}. \tag{3.33}$$

O valor de ϕ situa-se entre -1 e 1 quando as proporções marginais são iguais. Quando isso não acontece, os limites têm valores diferentes de -1 e de 1 e em geral não são simétricos em relação a zero. Nestes casos temos que:

$$\phi_{\max} = \sqrt{\frac{P'Q}{P^*Q'}} \tag{3.34}$$

e

$$\phi_{\min} = \sqrt{\frac{QQ'}{P^*P'}} \tag{3.35}$$

onde:

- P^* é a maior das somas marginais, ou seja, $P^* = \max\{P_{.1}, P_{.2}, P_{1.}, P_{2.}\}$;
- Q é a soma marginal para a mesma variável que P^* mas para a outra categoria;
- P' é a soma marginal para a mesma categoria com a mesma ordem que a de P^* mas para a outra variável;
- Q' é a soma marginal para a outra categoria e a outra variável.

De forma esquemática, podemos ter estes valores nas seguintes posições:

	0	1																		
0			P^*	ou	0			Q	ou		1	0	P'	ou		0	1	Q'
1			Q		1			P^*		0			Q'		1			P'
	P'	Q'				Q'	P'				P^*	Q				Q	P^*			

Quando se muda de categoria, muda-se de letra; quando se muda de variável, coloca-se um apóstrofo.

Para interpretar o coeficiente ϕ, é preciso compará-lo com os valores máximo e mínimo. Em muitas ocasiões o mais importante é a comparação de vários coeficientes ϕ o que se torna mais difícil quando eles não variam no mesmo intervalo. Uma solução simples será transformá-los em coeficientes R ou calcular a razão entre ϕ e $\phi_{\max}$, no caso de $\phi > 0$ e entre ϕ e $\phi_{\min}$ no caso de $\phi < 0$, para cada um deles.

Para determinar se um coeficiente ϕ é significativo basta compará-lo com $\phi_\alpha = \frac{z_{\alpha/2}}{\sqrt{N}}$, sendo α o nível de significância do teste e $z_{\alpha/2}$ o valor da variável aleatória Normal estandardizada que deixa para a direita uma área de $\alpha/2$. Valores de ϕ superiores a ϕ_α permitem concluir, ao nível de significância α, que ϕ é significativo.

Exemplo[8]: Queremos averiguar se 42 alunos, que responderam a duas determinadas perguntas de um teste (X e Y), ao ter sucesso numa delas também o terão na outra.

		Pergunta X		
		Fracasso	Êxito	Total
Pergunta Y	Fracasso	12	7	19
	Êxito	6	17	23
	Total	18	24	42

$$\begin{aligned}\phi &= \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}} = \\ &= \frac{(12\times 17)-(6\times 7)}{\sqrt{19\times 23\times 18\times 24}} = \frac{204-42}{\sqrt{188784}} = 0.373\end{aligned}$$

Como $\phi = 0.373 > 0$ calculamos:

$$\phi_{\max} = \sqrt{\frac{P'Q}{P^*Q'}} = \sqrt{\frac{23\times 18}{24\times 19}} = \sqrt{0.9079} = 0.953$$

A comparação entre os valores de ϕ e $\phi_{\max}$, permite-nos dizer que a associação entre as variáveis é fraca.

Para avaliar se ϕ é significativo, vamos calcular,

$$\phi_{0.01} = \frac{z_{0.005}}{\sqrt{N}} = \frac{2.576}{\sqrt{42}} = 0.3975 \text{ e } \phi_{0.05} = \frac{z_{0.025}}{\sqrt{N}} = \frac{1.960}{\sqrt{42}} = 0.3024$$

como $\phi > 0.3024 = \phi_{0.05}$, existe evidência de que ϕ é significativo ao nível de significância de 0.05.

3.7.3 Coeficiente Q de Yule

O coeficiente Q de Yule é uma medida de associação para variáveis dicotómicas, apresentadas em tabelas de contingência 2×2. É dado por:

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} \tag{3.36}$$

O intervalo de Q é $[-1, 1]$.

Se tivermos $Q = 1$ isto não significa necessariamente que a associação é perfeita. Se alguma das frequências observadas for igual a 0 o valor do Q de Yule poderá ser

-1 ou +1. Por isso, o significado de Q pode levar a conclusões incorretas quando o valor de uma das células da tabela for 0. Sendo assim, não é recomendável utilizar este coeficiente quando a frequência absoluta de alguma célula é muito pequena. Desde que o número total de observações seja razoavelmente grande, a distribuição de Q é normal com variância $\frac{1}{4}\left(1-Q^2\right)^2\left[\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right]$. Pode assim obter-se um intervalo de confiança aproximado para Q:

$$\widehat{Q} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{4}\left(1-\widehat{Q}^2\right)^2\left[\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right]}.$$

A significância do Q de Yule pode ser avaliada por:

$$z=\frac{Q}{\sqrt{\frac{1}{4}\left(1-Q^2\right)^2\left[\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right]}}. \tag{3.37}$$

3.8 *Odds ratio*

A *odds ratio* é uma medida de associação adequada para tabelas de contingência de qualquer dimensão, mas mais facilmente interpretadas no caso de tabelas 2×2. A *odds ratio* expressa o grau de associação entre duas variáveis num formato numérico diferente do habitual.

A *odds ratio* é uma das medidas mais utilizadas em investigação epidemiológica. Em certas condições, é uma aproximação de quão mais provável é um resultado estar presente (isto é, uma categoria de uma das variáveis) nos elementos de uma categoria da outra variável, do que na(s) outra(s) categoria(s) desta variável. Por exemplo, se Y representa a presença ou ausência de cancro do pulmão e se X representa se uma pessoa fuma ou não, então o valor 2 para uma estimativa da *odds ratio* indica que o cancro do pulmão é duas vezes mais frequente entre os fumadores do que entre os não fumadores, para a população em estudo.

O conceito de *odds* de um dado evento (X) ocorrer é dado por:

$$odds\,(X)=\frac{P\,(X \text{ ocorrer})}{P\,(X \text{ não ocorrer})}. \tag{3.38}$$

Temos as seguintes propriedades:

- $0 \leq odds \leq \infty$;
- a $odds > 1$ significa que a probabilidade do evento ocorrer é maior do que $\frac{1}{2}$. Quanto maior o valor da *odds*, maior a probabilidade do evento ocorrer;
- a $odds < 1$ significa que a probabilidade do evento ocorrer é menor do que $\frac{1}{2}$. Quanto menor o valor da *odds*, menor a probabilidade do evento ocorrer, sendo que o valor minímo que que a *odds* pode tomar é zero;

- a *odds* $= 1$ significa que a probabilidade do evento ocorrer é igual a $\frac{1}{2}$.

Consideremos uma tabela de contingência 2×2:

		Variável 2		
		B	$\overline{B}$	**Total**
Variável 1	**A**	n_{11}	n_{12}	$n_{1.}$
	$\overline{A}$	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	N

Uma outra forma de apresentar esta tabela será substituindo as frequências absolutas por proporções:

$$p_{ij} = \frac{n_{ij}}{N} \tag{3.39}$$

o que transforma a tabela em:

		Variável 2		
		B	$\overline{B}$	**Total**
Variável 1	**A**	p_{11}	p_{12}	$p_{1.}$
	$\overline{A}$	p_{21}	p_{22}	$p_{2.}$
	Total	$p_{.1}$	$p_{.2}$	1

Frequentemente, uma das duas características em estudo é influente na outra. Por exemplo, um estudo poderá considerar a idade da mãe como influente do peso do bebé à nascença. Uma forma de medir a probabilidade de obter o resultado em estudo quando o factor influente está presente é dado por:

$$\Omega_A = \frac{P\left(B|A\right)}{P\left(\overline{B}|A\right)}, \tag{3.40}$$

Ω_A indica a possibilidade (*odds*) de que B ocorra quando A está presente. Como $P\left(B|A\right)$ pode ser estimado pelo estimador de máxima verosimilhança

$$\widehat{P}\left(B|A\right) = \frac{\widehat{p}_{11}}{\widehat{p}_{1.}} \tag{3.41}$$

e $P\left(\overline{B}|A\right)$ por

$$\widehat{P}\left(\overline{B}|A\right) = \frac{\widehat{p}_{12}}{\widehat{p}_{1.}} \tag{3.42}$$

também é estimador de máxima verosimilhança, então Ω_A poderá ser estimado por

$$O_A = \frac{\widehat{p}_{11}/\widehat{p}_{1.}}{\widehat{p}_{12}/\widehat{p}_{1.}} = \frac{\widehat{p}_{11}}{\widehat{p}_{12}}. \tag{3.43}$$

Quando A não está presente, a probabilidade de ocorrência de B é definida por

$$\Omega_{\overline{A}} = \frac{P\left(B|\overline{A}\right)}{P\left(\overline{B}|\overline{A}\right)} \tag{3.44}$$

que pode ser estimado por

$$O_{\overline{A}} = \frac{\widehat{p}_{21}/\widehat{p}_{2.}}{\widehat{p}_{22}/\widehat{p}_{2.}} = \frac{\widehat{p}_{21}}{\widehat{p}_{22}}. \tag{3.45}$$

As duas *odds*, Ω_A e $\Omega_{\overline{A}}$ poderão ser combinadas de forma a nos darem uma medida de associação (Yule, 1900):

$$Q = \frac{\Omega_A - \Omega_{\overline{A}}}{\Omega_A + \Omega_{\overline{A}}}. \tag{3.46}$$

Uma outra, também da autoria de Yule (1912), é dada por:

$$Y = \frac{\sqrt{\Omega_A} - \sqrt{\Omega_{\overline{A}}}}{\sqrt{\Omega_A} + \sqrt{\Omega_{\overline{A}}}}. \tag{3.47}$$

A medida de associação baseada em Ω_A e $\Omega_{\overline{A}}$ com maior uso é simplesmente a razão entre as duas:

$$\omega = \frac{\Omega_A}{\Omega_{\overline{A}}} \tag{3.48}$$

que é chamada *odds ratio*, ou razão das possibilidades. Pode ser estimada pela *odds ratio* amostral:

$$\widehat{w} = \frac{O_A}{O_{\overline{A}}} = \frac{\widehat{p}_{11}/\widehat{p}_{12}}{\widehat{p}_{21}/\widehat{p}_{22}} = \frac{\widehat{p}_{11}\widehat{p}_{22}}{\widehat{p}_{12}\widehat{p}_{21}}. \tag{3.49}$$

Se $P\left(B|A\right) = P\left(B|\overline{A}\right)$, o que indica independência ou falta de associação entre ambas as características, então Ω_A e $\Omega_{\overline{A}}$ são também iguais, pelo que $\omega = 1$. Se $P\left(B|A\right) > P\left(B|\overline{A}\right)$, então $\Omega_A > \Omega_{\overline{A}}$ e $\omega > 1$. Se $P\left(B|A\right) < P\left(B|\overline{A}\right)$ então $\Omega_A < \Omega_{\overline{A}}$ e $\omega < 1$. Quando $1 < \omega < \infty$, os indivíduos na primeira linha têm uma maior probabilidade de sucesso do que aqueles na segunda linha. Quando $0 < \omega < 1$ acontece o contrário.

Valores de ω distantes de 1.0 numa dada direção demonstram associação forte. Dois valores representam a mesma associação mas em direcções opostas, quando uma é o inverso da outra. Por exemplo, quando $\omega = 0.25$ a probabilidade de sucesso na linha 1 é 0.25 vezes a probabilidade de sucesso na linha 2, ou de forma equivalente, a probabilidade de sucesso na linha 2 é $\frac{1}{0.25} = 4.0$ vezes a probabilidade de sucesso na linha 1. Quando a ordem das linhas ou das colunas é revertida, o novo valor de ω é o inverso do original.

O erro padrão para a *odds ratio* estimada é dado por:

$$s.e.(\widehat{\omega}) = \frac{\widehat{\omega}}{\sqrt{N}}\sqrt{\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}}. \tag{3.50}$$

Uma fórmula equivalente para o erro padrão usa as frequências absolutas:

$$s.e.(\widehat{\omega}) = \frac{\widehat{\omega}}{\sqrt{N}}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{3.51}$$

Podemos também utilizar esta medida mas com as frequências absolutas, sendo o estimador da máxima verosimilhança

$$\widehat{\omega} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Exemplo[10]: Foram selecionados 200 nascimentos, de uma população de 1000, para estudar a possível associação entre o peso do recém-nascido e a idade da mãe. Consideremos $A =$ *mãe com idade* $\leq$ *20 anos* e $\overline{A} =$ *mãe com idade* $>$ *20 anos.* Temos também que $B =$ *peso do recém-nascido* $\leq$ *2500 gramas* e $\overline{B} =$ *peso do recém-nascido* $>$ *2500 gramas.* A tabela seguinte ilustra os dados recolhidos:

	Peso à nascença		
Idade da Mãe	**B**	$\overline{B}$	**Total**
A	10	40	50
$\overline{A}$	15	135	150
Total	25	175	200

Em termos de proporções, esta tabela torna-se:

	Peso à nascença		
Idade da Mãe	**B**	$\overline{B}$	**Total**
A	0.050	0.200	0.25
$\overline{A}$	0.075	0.675	0.75
Total	0.125	0.875	1

ou seja, a proporção de nascimentos nos quais a mãe tinha 20 anos ou menos e o recém-nascido pesava 2500 gramas ou menos é estimada por $p(A \text{ e } B) = p_{11} = 0.050$. A proporção de nascimentos em que a mãe tinha 20 anos ou menos é estimada por $p(A) = p_{1\cdot} = 0.25$ e a proporção de nascimentos em que o recém-nascido pesava 2500 gramas ou menos é $p(B) = p_{\cdot 1} = 0.125$. Portanto a *odds* estimada de uma mãe com 20 anos ou menos dar à luz um bebé com 2500 gramas ou menos é:

$$O_A = \frac{\widehat{p}_{11}}{\widehat{p}_{12}} = \frac{0.050}{0.200} = 0.25$$

ou seja, a cada 4 quatro nascimentos de bebés com mais de 2500 gramas, filhos de mães com 20 anos ou menos, haverá um nascimento de um bebé com peso inferior a 2500 gramas.

A *odds ratio* estimada é:

$$\widehat{\omega} = \frac{O_A}{O_{\overline{A}}} = \frac{\widehat{p}_{11}\widehat{p}_{22}}{\widehat{p}_{12}\widehat{p}_{21}} = \frac{0.050 \times 0.675}{0.200 \times 0.075} = 2.25,$$

indicando que a probabilidade de um bebé nascer com peso baixo em mães mais jovens é 2.25 vezes superior à mesma probabilidade mas para mães com idade superior a 20 anos.

O erro padrão é dado por:

$$\begin{aligned} s.e.(\widehat{\omega}) &= \frac{\widehat{\omega}}{\sqrt{N}}\sqrt{\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}} = \\ &= \frac{2.25}{\sqrt{200}}\sqrt{\frac{1}{0.050} + \frac{1}{0.200} + \frac{1}{0.075} + \frac{1}{0.675}} = \\ &= 1.00 \end{aligned}$$

3.8.1 Testar hipóteses acerca das *odds ratio*

Centremos a nossa atenção em tabelas 2×2 com frequências marginais $n_{1.}$, $n_{2.}$, $n_{.1}$ e $n_{.2}$ correspondentes às frequências observadas. Suponhamos também que o valor da *odds ratio* subjacente é igual a ω. As frequências esperadas E_{ij} associadas a ω são tais que:

- os E_{ij}'s são consistentes com os dados originais no sentido em que eles mantêm os totais marginais:

	factor B		
factor A	**Presente**	**Ausente**	**Total**
Presente	E_{11}	E_{12}	$n_{1.}$
Ausente	E_{21}	E_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

- os E_{ij}'s são consistentes com o valor de ω no sentido em que

$$\frac{E_{11}E_{22}}{E_{12}E_2} = \omega. \tag{3.52}$$

A hipótese de que o valor subjacente da *odds ratio* é igual a ω poderá ser testada comparando o valor da estatística de teste:

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(|n_{ij} - E_{ij}| - \frac{1}{2}\right)}{E_{ij}} \tag{3.53}$$

onde $E_{ij} = \frac{n_{1.}n_{.j}}{N}$, com o valor da variável aleatória Qui-quadrado com 1 grau de liberdade.

Quando ω é o valor hipotético da *odds ratio*, e quando $\omega \neq 1$, as frequências esperadas poderão ser calculadas como se define de seguida. Seja:

$$X = \omega\left(n_{1.} + n_{.1}\right) + \left(n_{2.} + n_{.1}\right) \tag{3.54}$$

e

$$Y = \sqrt{X^2 - 4n_{1.}n_{.1}\omega\left(\omega - 1\right)}. \tag{3.55}$$

Então:

$$\begin{aligned} E_{11} &= \frac{X - Y}{2\left(\omega - 1\right)}, \\ E_{12} &= n_{1.} - E_{11}, \\ E_{21} &= n_{.1} - E_{11}, \\ E_{22} &= n_{.2} - n_{1.} + E_{11.} \end{aligned} \tag{3.56}$$

O resultado seguinte foi provado por Stevens (1951) e Conrfield (1956). Quando as frequências marginais são fixadas e quando ω é o valor da *odds ratio*, n_{ij} (para qualquer uma das quatro células da tabela) é aproximadamente normalmente distribuído com média E_{ij} e erro padrão $\frac{1}{\sqrt{W}}$, onde

$$W = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{E_{ij}}$$

e os E_{ij}'s são definidos como em 3.56.

Através da regressão logística, é possível obter estimativas da *odds ratio* através das estimativas dos coeficientes do modelo. Estas estimativas, os correspondentes testes de significância e intervalos de confiança podem ser obtidos diretamente através do *software* SPSS.

3.8.2 Risco Relativo

Risco relativo é um quociente de probabilidades: incidência de um resultado nos elementos pertencente a uma categoria sobre a incidência desse resultado nos restantes elementos.

O risco relativo permite que se compare as probabilidades relativas de um resultado estar presente, isto é, o risco relativo é a probabilidade de contrair uma doença dado que o indivíduo faz parte de um determinado grupo (geralmente o grupo que se considera

ter o maior risco) dividido pela probabilidade de contrair a doença sendo membro do outro grupo (geralmente o grupo com menor risco). No caso de estarmos presente uma tabela do tipo:

	Contrai doença	**Não contrai doença**	**Total**
Grupo 1	n_{11}	n_{12}	$n_{1.}$
Grupo 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

as probabilidade serão então dadas por:

$$p\left(\text{Contrair a doença/Grupo 2}\right) = \frac{n_{21}}{n_{21}+n_{22}}$$

$$p\left(\text{Contrair a doença/Grupo 1}\right) = \frac{n_{11}}{n_{11}+n_{12}}.$$

A equação para o risco relativo é então:

$$RR = \frac{n_{21}/\left(n_{21}+n_{22}\right)}{n_{11}/\left(n_{11}+n_{12}\right)} \tag{3.57}$$

ou também:

$$RR = \frac{n_{11}n_{21}+n_{12}n_{21}}{n_{11}n_{21}+n_{11}n_{22}}.$$

Exemplo[10]: Na tabela seguinte estão dados relativos a 200 indivíduos que acabaram de estar em contacto com um determinado animal, divididos em dois grupos: um em que os membros desse grupo lavam as mãos imediatamente após o contacto com o animal, e outro onde os indivíduos não lavam as mãos após o contacto com o animal.

	Contrai a doença		
	Sim	**Não**	**Total**
Lava as mãos	30	70	100
Não lava as mãos	60	40	100
Total	90	110	200

$$\begin{aligned}\widehat{P}\left(\text{Contrair a doença/Não lava as mãos}\right) &= \frac{n_{21}}{n_{21}+n_{22}} = \\ &= \frac{60}{60+40} = \frac{60}{100} = \\ &= 0.6\end{aligned}$$

$$\begin{aligned}\widehat{P}\left(\text{Contrair a doença/Lava as mãos}\right) &= \frac{n_{11}}{n_{11}+n_{12}} = \\ &= \frac{30}{30+70} = \frac{30}{100} = \\ &= 0.3\end{aligned}$$

$$\begin{aligned}RR &= \frac{n_{21}/\left(n_{21}+n_{22}\right)}{n_{11}/\left(n_{11}+n_{12}\right)} \\ \widehat{RR} &= \frac{0.6}{0.3} = 2\end{aligned}$$

o que significa que alguém que não lava as mãos logo após o contacto com o animal tem uma probabilidade duas vezes maior de contrair a doença do que alguém que lava as mãos após ter estado em contacto com o animal.

Se revertermos a equação:

$$\begin{aligned}RR &= \frac{n_{11}/\left(n_{11}+n_{12}\right)}{n_{21}/\left(n_{21}+n_{22}\right)} \\ \widehat{RR} &= \frac{0.3}{0.6} = 0.5\end{aligned}$$

o que indica que alguém que lava as mãos logo após ter estado em contacto com o animal tem metade da probabilidade de contrair a doença de alguém que não lava as mãos.

Como referimos no início da secção sobre a *odds ratio*, o seu valor, muitas vezes, aproxima-se do valor do risco relativo. Esta aproximação é válida, quando a probabilidade do resultado estar presente é pequena quer num grupo de elementos da amostra quer nos restantes grupos.

O estimador do risco relativo terá uma distribuição aproximadamente Normal para amostras suficientemente grandes. Infelizmente, na prática, esta exigência sobre a dimensão da amostra excede as dimensões das amostras da maioria dos estudos. Assim, as referências são usualmente baseadas na distribuição da estatística $\ln\left(\widehat{RR}\right)$, cuja distribuição se aproxima da distribuição Normal para amostras de dimensão muito inferior. Mais uma vez, o recurso aos *softwares* específicos, permitem a obtenção rápida dos resultados.

3.9 Lambda de Goodman e Kruskal

É uma medida de associação para duas variáveis categorizadas, com uma interpretação muito clara em termos de probabilidades.

Se escolhermos um indivíduo ao acaso de uma população e não tivermos nenhum tipo de informação sobre a que linha ou coluna este indivíduo pertence, a melhor previsão sobre a que célula ele pertence, será a correspondente à linha e coluna com maior densidade populacional marginal. Denotam-se a linha e coluna com maior densidade populacional marginal por $p_{m.}$e $p_{.m}$ respetivamente. Claramente a probabilidade de errar em cada caso é de $(1 - p_{m.})$ e $(1 - p_{.m})$ respetivamente.

Se soubermos a que coluna pertence, a melhor previsão para a linha será a que corresponde à maior densidade na coluna dada. Dada a coluna j denotamos a densidade máxima por p_{mj}. A probabilidade de cometer um erro nesta coluna é $p_{.j} - p_{mj}$ e sobre todas as colunas é de $1 - \sum_{j=1}^{c} p_{mj}$.

A diferença entre as duas probabilidades de erro é dada por:

$$(1 - p_{m.}) - \left(1 - \sum_{j=1}^{c} p_{mj}\right) = \sum_{j=1}^{c} p_{mj} - p_{m.}.$$

Esta expressão é então dividida pela probabilidade de errar na previsão da linha e assim obtém-se o Lambda de Goodman e Kruskal:

$$\lambda_{m.} = \frac{\sum_{j=1}^{c} p_{mj} - p_{m.}}{1 - p_{m.}}. \tag{3.58}$$

A medida de associação $\lambda_{m.}$ representa a redução proporcional do erro para a previsão das linhas quando a coluna é conhecida. De forma semelhante, calcula-se este valor para a previsão da coluna quando a linha é conhecida:

$$\lambda_{.m} = \frac{\sum_{i=1}^{r} p_{im} - p_{.m}}{1 - p_{.m}}. \tag{3.59}$$

As duas medidas de associação acima descritas são assimétricasno sentido em que uma das variáveis é predicta usando a outra. Uma medida de associação simétrica que combina a lógica usada anteriormente é dada por

$$\lambda = \frac{\left[\frac{1}{2}\sum_{j=1}^{c} p_{mj} + \frac{1}{2}\sum_{i=1}^{r} p_{im}\right] - \frac{1}{2}\left[p_{.m} - p_{m.}\right]}{1 - \frac{1}{2}\left[p_{.m} + p_{m.}\right]}. \tag{3.60}$$

O intervalo de variação destas medidas de associação é $[0, 1]$.

Exemplo[18]: As observações foram obtidas de uma grande população de cidadãos em vários municípios. Os valores na tabela são as densidades populacionais. A cada cidadão foi pedido que respondessem a uma questão sobre a seriedade da situação da criminalidade na vizinhança. Foi-lhes pedida também informação sobre a sua idade. Os totais dão-nos as densidades marginais para a opinião e idade.

Idade	Opinião sobre a situação da criminalidade				
	Nada séria	**Ligeir. séria**	**Moder. séria**	**Muito séria**	**Totais**
Menos de 30	0.015	0.076	0.121	0.055	0.267
30 a 39	0.017	0.117	0.111	0.037	0.282
40 a 49	0.012	0.074	0.104	0.032	0.222
50 a 59	0.007	0.034	0.072	0.020	0.133
Mais de 60	0.001	0.027	0.038	0.030	0.096
Totais	0.052	0.328	0.446	0.174	1.000

As medidas de associação de lambda são dadas por:

$$\begin{aligned}\lambda_{.m} &= \frac{\sum_{j=1}^{c} p_{im} - p_{.m}}{1 - p_{m.}} = \\ &= \frac{(0.121 + 0.117 + 0.104 + 0.072 + 0.038) - 0.446}{1 - 0.446} = \\ &= 0.01\end{aligned}$$

$$\begin{aligned}\lambda_{m.} &= \frac{\sum_{j=1}^{c} p_{mj} - p_{m.}}{1 - p_{m.}} = \\ &= \frac{(0.017 + 0.117 + 0.121 + 0.055) - 0.282}{1 - 0.282} = \\ &= 0.039\end{aligned}$$

$$\begin{aligned}\lambda &= \frac{\left[\frac{1}{2}\sum_{j=1}^{c} p_{mj} + \frac{1}{2}\sum_{i=1}^{r} p_{im}\right] - \frac{1}{2}\left[p_{.m} - p_{m.}\right]}{1 - \frac{1}{2}\left[p_{.m} + p_{m.}\right]} = \\ &= \frac{\frac{1}{2}(0.452) + \frac{1}{2}(0.310) + \frac{1}{2}(0.282) + \frac{1}{2}(0.446)}{1 - \frac{1}{2}(0.282) + \frac{1}{2}(0.446)} = 0.027\end{aligned}$$

Todas estas três medidas de previsão de associação sugerem que a associação é extremamente fraca.

3.9.1 Inferência para Lambda

Uma tabela de contingência pode ser usada para fazer inferências sobre as medidas de associação populacionais definidas nesta secção. Um estimador para as medidas de associação de Goodman-Kruskal pode ser usado com este propósito, substituindo as probabilidades $p_{i.}$ e $p_{.j}$ pelas frequências amostrais das células, n_{ij}, poderemos obter

os estimadores para as medidas de Goodman-Kruskal. Para medir a previsibilidade da coluna dada a linha, o estimador é dado por

$$\widehat{\lambda}_{.m} = \frac{\sum_{i=1}^{r} n_{im} - n_{.m}}{n - n_{.m}}, \tag{3.61}$$

onde n_{im} denota o maior valor de n_{ij} na linha i e $n_{.m}$ é o maior valor dos totais marginais $n_{.j}$.

Em amostras grandes a estatística $z = \frac{\widehat{\lambda}_{.m} - \lambda_{.m}}{\widehat{\sigma}_{\lambda_{.m}}}$ tem uma distribuição Normal estandardizada. O estimador de $\widehat{\sigma}_{\lambda_{.m}}$ é dado por

$$\widehat{\sigma}^2_{\lambda_{.m}} = \left(n - \sum_{i=1}^{r} n_{im} \right) \frac{\sum n_{im} + n_{.m} - 2n^*_{.m}}{(n - n_{.m})^3}, \tag{3.62}$$

onde $n^*_{.m}$ é a soma dos n_{im}. Esta estatística não pode ser aplicada se $\widehat{\lambda}_{.m} = 0$ ou se $\widehat{\lambda}_{.m} = 1$. As hipótese $H_0 : \lambda_{.m} = 0$ e $H_0 : \lambda_{.m} = 1$ são rejeitadas a não ser que $\widehat{\lambda}_{.m} = 0$ ou $\widehat{\lambda}_{.m} = 1$, respetivamente.

De uma forma semelhante, a medida de previsibilidade para a linha dada a coluna $\widehat{\lambda}_{m.}$ é dada por

$$\widehat{\lambda}_{m.} = \frac{\sum_{j=1}^{c} n_{mj} - n_{m.}}{n - n_{m.}}. \tag{3.63}$$

onde n_{mj} é o valor mais alto de n_{ij} na coluna j e $n_{m.}$ é o maior valor dos totais marginais $n_{i.}$.

Exemplo[18]: Antes da obrigatoriedade no uso do cinto de segurança nos automóveis, na província de Alberta um estudo foi feito para determinar a utilidade dos cintos de segurança na prevenção de ferimentos. Uma amostra de 86 769 relatórios de acidentes automóveis foram estudados. Em cada relatório, a gravidade dos ferimentos do condutor foi classificada em uma de quatro categorias e também foi anotado o uso ou não de cinto de segurança.

	Gravidade dos ferimentos do condutor				
Cinto de segurança	**Nenhum**	**Mínimo**	**Ligeiro**	**Grave/Fatal**	**Totais**
Sim	12813	647	359	42	13861
Não	65963	4000	2642	303	72908
Totais	78776	4647	3001	345	86769

$$\begin{aligned} \widehat{\lambda}_{.m} &= \frac{\sum_{j=1}^{c} n_{im} - n_{.m}}{n - n_{.m}} = \\ &= \frac{(12813 + 65963) - 78776}{86769 - 78776} = 0 \end{aligned}$$

$$\begin{aligned}\widehat{\lambda}_{m.} &= \frac{\sum_{i=1}^{r} n_{mj} - n_{m.}}{n - n_{m.}} = \\ &= \frac{(65963 + 4000 + 2642 + 303) - 72908}{86769 - 72908} = 0\end{aligned}$$

Ambas as medidas de associação têm resultado 0. A categoria sem cinto de segurança domina a categoria com cinto de segurança em todos os níveis de gravidade dos ferimentos; a categoria "Nenhum ferimento" domina todas os outros níveis de gravidade tanto para o uso de cinto de segurança como para o não uso de cinto de segurança.

Capítulo 4

Medidas de Concordância

4.1 Introdução

A distinção entre concordância e associação para dados nominais consiste no facto de, para que duas observações concordem, elas deverão pertencer a categorias idênticas, enquanto que para que duas observações estejam perfeitamente associadas é apenas necessário que possamos prever a categoria a que pertence uma das observações a partir do conhecimento da categoria a que pertence a outra observação. Assim, uma tabela de contingência pode exibir um grau de associação elevado ao mesmo tempo que possui um grau de concordância alto ou baixo.

A medida de concordância mais simples é a que se baseia na proporção da população cuja categorização de ambas as variáveis é idêntica, isto é, $\sum_i p_{ii}$. Os valores que esta proporção simples pode tomar são afetados pelos totais marginais.

Muitas outras medidas de concordância foram desenvolvidas ao longo dos tempos, sendo que algumas das mais utilizadas são discutidas nas secções seguintes.

4.2 Gamma (γ) de Goodman e Kruskal

Goodman e Kruskal (1954) propuseram uma medida de concordância aplicada a tabelas de contingência $r \times c$ quando as categorias de ambas as variáveis são ordinais. Consideramos que (x_i, y_i) e (x_j, y_j) são concordantes se quando $x_i < x_j$ se tem $y_i < y_j$ e quando $x_i > x_j$ temos $y_i > y_j$. São discordantes quando $x_i > x_j$ e $y_i < y_j$ ou $x_i < x_j$ e $y_i > y_j$. Seja C o número de pares concordantes e D o número de pares discordantes.

Dos $(C + D)$ pares de observações que são concordantes ou discordantes, a proporção $\frac{C}{(C+D)}$ é concordante e a porporção $\frac{D}{(C+D)}$ é discordante. A diferença entre estas proporções é um estimador da medida de concordância *gamma*:

$$\widehat{\gamma} = \frac{(C - D)}{(C + D)} \tag{4.1}$$

A expressão da medida de concordância gamma definida para a população, é:

$$\gamma = \frac{p_c - p_d}{p_c + p_d} \tag{4.2}$$

onde

$$p_c = 2\sum\sum_{i<k}\sum\sum_{j<l} p_{ij}p_{kl} \quad \text{e} \quad p_d = 2\sum\sum_{i<k}\sum\sum_{j>l} p_{ij}p_{kl} \tag{4.3}$$

são as probabilidades de concordância e discordância para um par de observações selecionadas ao acaso. O número 2 está presente em ambas as fórmulas pois a primeira observação pode ser o par (i, j) e a segunda o par (k, l) ou o contrário.

O valor de $\widehat{\gamma}$ (e γ) é simétrico, ou seja, é o mesmo quer consideremos a v.a. X, a v.a. Y ou ambas como variável resposta. $\widehat{\gamma}$ varia entre -1 e 1 sendo que valores absolutos mais elevados denotam uma associação mais forte. Temos $\widehat{\gamma} = 1$ quando $D = 0$ e $\widehat{\gamma} = -1$ quando $C = 0$.

O valor $|\gamma| = 1$ implica que a relação é monótona mas não estritamente monótona. Por outras palavras, se $\gamma = 1$ para um par de observações (X_a, Y_a) e (X_b, Y_b) onde $X_a < X_b$, temos que $Y_a \leq Y_b$ mas não necessariamente que $Y_a < Y_b$. A independência estatística de X e Y implica que $\gamma = 0$ mas o contrário não é verdade.

Para tabelas de contingência 2×2, a expressão de $\widehat{\gamma}$ simplifica-se:

$$\widehat{\gamma} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = Q. \tag{4.4}$$

Esta medida foi introduzida por G. Udny Yule em 1900 que lhe atribuiu o símbolo Q em honra de Adolphe Quetelet, um astrónomo-sociólogo-estatístico belga. Também chamado por Q de Yule (abordado na secção 3.7.3), relaciona-se com a *odds ratio* $\widehat{\theta} = n_{11}n_{22}/\left(n_{12}n_{21}\right)$ pela fórmula:

$$\widehat{\gamma} = \frac{\widehat{\theta} - 1}{\widehat{\theta} + 1}. \tag{4.5}$$

Para tabelas 2×2, gamma é uma função de $\widehat{\theta}$ monótona que transforma um intervalo $[0, \infty]$ em $[-1, 1]$.

Exemplo[?]: Consideremos a tabela seguinte que relaciona a *Felicidade* com o *Rendimento Familiar.*

	Felicidade			
Rendimento Familiar	**Muito Feliz**	**Feliz**	**Pouco Feliz**	**Total**
Acima da média	272	294	49	615
Na média	454	835	131	1420
Abaixo da média	185	527	208	920
Total	911	1656	388	**2955**

Seja "Muito Feliz" a categoria mais alta da variável *Felicidade* e "Acima da média"a categoria mais alta da variável *Rendimento Familiar*.

Consideremos duas pessoas, uma englobada na célula (*Acima da média*, *Muito Feliz*) e uma outra na célula (*Na média*, *Feliz*). Este par de pessoas é concordante uma vez que a primeira pessoa está classificada acima da segunda tanto para a variável *Rendimento Familiar* como na variável *Felicidade*. Então cada uma das 272 pessoas consideradas na célula (*Acima da média*, *Muito Feliz*) formam pares concordantes quando combinadas com cada uma das 835 pessoas classificadas com (*Média*, *Feliz*), logo temos $272 \times 835 = 227120$ pares concordantes provenientes destas duas células. Na realidade as 272 pessoas classificadas na célula (*Acima da média*, *Muito Feliz*) formam pares concordantes com todos os $(835 + 131 + 527 + 208) = 1701$ classificados em categorias inferiores em ambas as variáveis.

$$\begin{aligned} C &= 272\,(835 + 131 + 527 + 208) + 294\,(131 + 208) \\ +454\,(527 + 208) + 835 \times 208 &= 1069708 \end{aligned}$$

$$\begin{aligned} D &= 49\,(454 + 835 + 185 + 527) + 294\,(454 + 185) \\ +131\,(185 + 527) + 835 \times 185 &= 533662 \end{aligned}$$

O valor de $\widehat{\gamma}$ para este exemplo é dado por:

$$\widehat{\gamma} = \frac{C - D}{C + D} = \frac{1069708 - 533662}{1069708 + 533662} = \frac{536046}{1603370} \simeq 0.334$$

Temos que $\frac{2}{3}$ dos pares são concordantes e $\frac{1}{3}$ é discordante e a diferença entre as proporções é 0.334, o que indica haver uma ligeira tendência para que a felicidade aumente conforme aumenta o rendimento familiar. A proporção de pares concordantes é superior à de pares discordantes em 0.334.

4.3 Tau de Kendall

O τ de Kendall poderá ser usado em alternativa ao coeficiente de correlação de Spearman. É uma medida de correlação de *ranks*, ou seja, que se foca na semelhança entre os valores dos *ranks* atribuídos às observações de X e de Y.

Para qualquer amostra com n observações existem $\binom{n}{2}$ possíveis pares de comparações de pontos (x_i, y_i) e (x_j, y_j). Consideramos que (x_i, y_i) e (x_j, y_j) são concordantes se quando $x_i < x_j$ se tem $y_i < y_j$ e quando $x_i > x_j$ temos $y_i > y_j$. São discordantes quando $x_i > x_j$ e $y_i < y_j$ ou $x_i < x_j$ e $y_i > y_j$. Seja C o número de pares concordantes e D o número de pares discordantes. O tau de Kendall é dado por:

$$\tau = \frac{(C - D)}{\binom{n}{2}} \tag{4.6}$$

Claramente τ (também conhecido como tau-a, nesta forma sem correção para empates) varia entre -1 e 1. Se $\tau = 1$, temos que a concordância é perfeita. Se $\tau = -1$, então a discordância é perfeita. Se X e Y são independentes, então τ será aproximadamente zero. O tau de Kendall pode ser visto como um estimador da diferença entre a probabilidade de concordância e a probabilidade de discordância.

Um forma um pouco mais simples de entender como se fazem estes cálculos, poderá ser a que Siegel[27] mostra com o exemplo seguinte:

Exemplo[27]: Suponhamos que se pediu a um júri composto por dois juízes que avaliassem quatro textos, escritos por pessoas diferentes, e os classificassem tendo em conta a qualidade do estilo expositivo. Os *ranks* obtidos são os seguintes:

	Texto a	**Texto** b	**Texto** c	**Texto** d
Juíz X	3	4	2	1
Juíz Y	3	1	4	2

Se reordenarmos estes valores de forma a que os *ranks* dos Juíz X estejam pela ordem natural $(1, 2, 3, \ldots)$ obtemos:

	Texto d	**Texto** c	**Texto** a	**Texto** b
Juíz X	1	2	3	4
Juíz Y	2	4	3	1

Podemos agora analisar o grau de correspondência entre os *ranks* atribuídos por cada um dos juízes. Como já temos os *ranks* do Juiz X por ordem natural, vemos agora quantos pares de *ranks* do Juiz Y estão também numa ordem correta (natural).

Começamos pelo valor mais à esquerda, 2 e comparamos com cada um dos outros valores. O primeiro par a ser comparado será 2 e 4. Está na ordem correta, pois o 2 precede o 4. Então atribuímos a este par o valor $+1$. O par seguinte, 2 e 3, também está em ordem natural, logo também atribuímos $+1$. Já o par 2 e 1, não está em ordem natural pois o 1 precede o 2 e não o contrário, logo atribuímos -1 a este par. Fazemos o mesmo para todos os outros pares e no fim fazemos a soma de todos os valores:

$$S = (+1) + (+1) + (-1) + (-1) + (-1) + (-1) = -2$$

Na equação de τ substituímos o numerador por este valor e no denominador atribuímos a n o número de casos que se comparou, ou seja, 4, e calculamos:

$$\tau = \frac{S}{\binom{n}{2}} = \frac{-2}{\binom{4}{2}} = \frac{-2}{6} = -0.33$$

Então $\tau = -0.33$ é a medida de concordância entre os *ranks* atribuídos pelo Juíz X e os *ranks* atribuídos pelo Juíz Y.

Se na comparação de (x_i, y_i) e (x_j, y_j), $x_i = x_j$ ou $y_i = y_j$ ou ambos, a comparação é chamada de empate e não conta como concordante ou discordante. Se existe um grande número de empates, o denominador de τ deverá ser corrigido substituindo $\binom{n}{2}$ por:

$$\left(\left[\binom{n}{2} - T_X\right]\left[\binom{n}{2} - T_Y\right]\right)^{\frac{1}{2}}$$

onde $T_X = \frac{1}{2}\sum t_X\,(t_X - 1)$, sendo t o número de observações empatadas em cada grupo de empates na variável X e $T_Y = \frac{1}{2}\sum t_Y\,(t_Y - 1)$ o mesmo, mas para a variável Y.

O τ de Kendall pode ser usado como uma estatística de teste num teste de hipóteses que pretenda verificar se duas variáveis poderão ser consideradas estatisticamente dependentes. Sob a hipótese nula da independência de X e Y, a distribuição de τ tem um valor esperado igual a zero.

Em amostras grandes (segundo alguns autores, é suficiente que n seja maior do que 10) a distribuição da estatística $3\tau\sqrt{n\,(n-1)}/\sqrt{2\,(2n+5)}$ é aproximadamente normal padrão, se X e Y são independentes (τ é aproximadamente normal com média zero e variância dada por $\frac{2(2n+5)}{9n(n-1)}$). A distribuição desta estatística aproxima-se da normalidade mais rapidamente que a distribuição de $R_s\sqrt{n-1}$ para a correlação de Spearman (R_s).

4.3.1 Comparação entre o coeficiente de correlação de Spearman R_s e τ

Estes coeficientes não são diretamente comparáveis pois ambos têm uma escala própria subjacente. Isto significa que se medirmos o grau de correlação entre as variáveis A e B usando R_s e depois compararmos A e C usando τ, não poderemos dizer que A está mais próxima de B do que de C ou vice-versa, visto estarmos a usar dois coeficientes de correlação incomparáveis.

Contudo, ambos os coeficientes usam a mesma quantidade de informação sobre os dados e assim ambos têm o mesmo poder para detetar a existência de associação na população. Ou seja, a distribuição de R_s e τ é tal que ambos rejeitarão a hipótese nula (de independência entre determinadas variáveis de uma população) para o mesmo nível de significância.

4.4 Tau-b de Kendall

Em 1945 Kendall propôs uma medida baseada na diferença $C - D$ que usa pares de observações empatadas em uma ou ambas as observações. Considere de novo o exemplo da secção 4.2, em que X é o rendimento familiar e Y corresponde ao grau e felicidade. O número de pares T_X empatados na variável linha X é dado por:

$$T_X = \sum_{i=1}^{r} \frac{n_{i.}\left(n_{i.}-1\right)}{2} = \frac{615 \times 614 + 1420 \times 1419 + 920 \times 919}{2} = 1619035 \qquad (4.7)$$

e pares T_Y empatados na variável coluna Y são dados por:

$$T_Y = \sum_{j=1}^{c} \frac{n_{.j}\left(n_{.j}-1\right)}{2} = \frac{911 \times 910 + 1656 \times 1655 + 388 \times 387}{2} = 1859923. \qquad (4.8)$$

Os pares empatados em X e Y são pares de observações provenientes da mesma célula. O número total destes pares é dado por:

$$T_{XY} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}\left(n_{ij}-1\right)}{2} = \frac{272 \times 271 + 294 \times 293 + \ldots + 208 \times 207}{2} = 717793. \qquad (4.9)$$

Para n observações, o número total de pares decompõe-se em:

$$\frac{n\left(n-1\right)}{2} = C + D + T_X + T_Y - T_{XY}. \qquad (4.10)$$

Nesta fórmula T_{XY} é subtraído pois os pares empatados em X e Y são também contados em T_X e T_Y. Note-se que $T_{XY} \leq T_X + T_Y$.

A fórmula do *Tau-b* para uma amostra é dada por:

$$\widehat{\tau}_b = \frac{C - D}{\sqrt{\left[\frac{n(n-1)}{2} - T_X\right]\left[\frac{n(n-1)}{2} - T_Y\right]}} \qquad (4.11)$$

enquanto que para a população é dado por:

$$\tau_b = \frac{p_c - p_d}{\sqrt{\left(1 - \sum_i p_{i.}^2\right)\left(1 - \sum_j p_{.j}^2\right)}}. \qquad (4.12)$$

Tal como o Gamma falado anteriormente, o Tau-b é simétrico. Uma vez que $C + D$ nunca é maior que $\left[n\left(n-1\right)/2 - T_X\right]$ ou $\left[n\left(n-1\right)/2 - T_Y\right]$, também nunca poderá exceder a sua média geométrica, que é o denominador de $\widehat{\tau}_b$. Assim, $|\widehat{\tau}_b| \leq |\widehat{\gamma}|$. Para tabelas 2×2, $\widehat{\tau}_b$ é idêntico à correlação entre X e Y.

Na realidade $\widehat{\tau}_b$ é um tipo de correlação até para tabelas $r \times c$, usando *sign scores* para os pares de observações. Para cada par de observações (x_a, y_a) e (x_b, y_b) seja

$$x_{ab} = sign\left(x_a - x_b\right) \text{ e } y_{ab} = sign\left(y_a - y_b\right) \qquad (4.13)$$

onde $sign\left(u\right) = 1$ se $u > 0$, $sign\left(u\right) = -1$ se $u < 0$ e $sign\left(u\right) = 0$ se $u = 0$. Os *sign scores* $\{x_{ab}\}$ indicam se x_a é maior ou menor que x_b e o mesmo acontece para $\{y_{ab}\}$.

Temos as seguintes propriedades:

- $x_{ab} = -x_{ba}$ e $y_{ab} = -y_{ba}$;
- $x_{ab}y_{ab} = 1$ se o par em questão for concordante e será igual a -1 se o par em questão for discordante;
- $x_{ab}^2 = 1$ para pares não empatados em X e $x_{ab}^2 = 0$ para pares empatados em X. De forma semelhante, $y_{ab}^2 = 1$ para pares não empatados em Y e $y_{ab}^2 = 0$ para pares empatados em Y.

Para $n\,(n-1)$ pares ordenados de observações (a, b) com $a \neq b$,

- $\sum\sum_{a\neq b} x_{ab}y_{ab} = 2\,(C - D)$;
- $\sum\sum_{a\neq b} x_{ab} = \sum\sum_{a\neq b} y_{ab} = 0$;
- $\sum\sum_{a\neq b} x_{ab}^2 = 2\left[\frac{n(n-1)}{2} - T_X\right]$;
- $\sum\sum_{a\neq b} y_{ab}^2 = 2\left[\frac{n(n-1)}{2} - T_Y\right]$.

Cada par é usado duas vezes nestes somatórios de forma a que $\sum\sum_{a\neq b} x_{ab} = \sum\sum_{a\neq b} y_{ab} = 0$ devido à relação $x_{ab} = -x_{ab}$ e $y_{ab} = -y_{ab}$. A correlação amostral entre $\{x_{ab}\}$ e $\{y_{ab}\}$ é portanto:

$$\frac{\sum\sum_{a\neq b} x_{ab}y_{ab}}{\sqrt{\left(\sum\sum_{a\neq b} x_{ab}^2\right)\left(\sum\sum_{a\neq b} y_{ab}^2\right)}} = \frac{C - D}{\sqrt{\left[\frac{n(n-1)}{2} - T_X\right]\left[\frac{n(n-1)}{2} - T_Y\right]}} = \widehat{\tau}_b. \qquad (4.14)$$

Para o exemplo da secção 4.2 temos:

$$\begin{aligned}
\widehat{\tau}_b &= \frac{C - D}{\sqrt{\left[\frac{n(n-1)}{2} - T_X\right]\left[\frac{n(n-1)}{2} - T_Y\right]}} = \\
&= \frac{1069708 - 533662}{\sqrt{\left(\frac{2955\times 2954}{2} - 1619035\right)\left(\frac{2955\times 2954}{2} - 1859923\right)}} = \\
&= \frac{536046}{\sqrt{(4364535 - 1619035)\,(4364535 - 1859923)}} = 0.204
\end{aligned}$$

Ou seja, esta medida mostra uma tendência relativamente fraca para que pessoas com maiores níveis de rendimento sejam mais felizes.

4.5 *d* de Somers

Somers propôs uma medida similar a gamma e a tau-b, mas que trata Y como uma variável resposta e X como uma variável explicativa. Para esta medida, pares não empatados em X servem de base. O d de Somers para uma amostra é dado por:

$$d = \frac{C - D}{\frac{n(n-1)}{2} - T_X} \tag{4.15}$$

Dos pares não empatados em X, d é a diferença entre as proporções de pares concordantes e discordantes.

Uma vez que o denominador de d é pelo menos tão grande quanto o denominador de $\widehat{\gamma}$, $|d| \leq |\widehat{\gamma}|$. Para que $|d| = 1$ é necessário que exista uma monotonia mais estrita que para $|\widehat{\gamma}| = 1$, sendo que C ou D tem de ser igual a 0 e nenhum dos pares não empatados em X pode estar empatado em Y. O d de Somers para a população é dado por:

$$\Delta = \frac{p_c - p_d}{1 - \sum_i p_{i.}^2}. \tag{4.16}$$

Para o exemplo da Secção 4.2, temos:

$$\begin{aligned} d &= \frac{C - D}{\frac{n(n-1)}{2} - T_X} = \frac{1069708 - 533662}{\frac{2955 \times 2954}{2} - 1619035} = \\ &= \frac{536046}{4364535 - 1619035} = 0.195 \end{aligned}$$

4.6 Kappa de Cohen

Para escalas nominais, a medida de concordância mais usada é o *Kappa* (Cohen,1960). Compara a concordância observada com a esperada caso as avaliações fossem independentes. O coeficiente *Kappa* trata as classificações como nominais e portanto trata as discordâncias para categorias ordenadas que estão próximas, da mesma forma que as discordâncias para categorias que estão distantes.

Uma situação comum na qual são usadas tabelas de contingência $r \times r$ é quando dois observadores classificam separadamente uma amostra de sujeitos usando a mesma escala categórica. Por exemplo, dois psiquiatras poderão classificar pacientes de acordo com determinados teste de diagnóstico. Estes dados são muitas vezes utilizados para averiguar se a escala categórica é ou não de confiança, geralmente avaliando o grau de concordância entre ambos os observadores. Um indicador de concordância intuitivo e simples será a proporção de pacientes classificados na mesma categoria por ambos os observadores, P_o:

$$P_o = \sum_{i=1}^{r} \frac{n_{ii}}{N}. \tag{4.17}$$

No entanto esta não é uma medida de concordância adequada uma vez que não tem em conta a concordância entre os observadores que pode ser atribuída ao acaso, pois é possível que os observadores tenham atribuído alguns sujeitos a determinadas categorias aleatoriamente tendo em conta as suas taxas marginais. O indicador de concordância para este tipo de processo, isto é, para a concordância devida apenas ao acaso, é dado por:

$$P_c = \frac{1}{N}\left(\sum_{i=1}^{r} \frac{n_{i.} n_{.i}}{N}\right). \tag{4.18}$$

Quando existe concordância total entre os dois observadores, $P_o = 1$. A diferença $P_o - P_c$ representa o excesso sobre a concordância atribuído apenas ao acaso. O excesso máximo possível é dado por $1 - P_c$. A razão entre estas duas diferenças é denominada *Kappa*:

$$k = \frac{P_o - P_c}{1 - P_c} \tag{4.19}$$

que é uma medida de concordância com as seguintes propriedades:

- Se existe concordância total, $k = 1$.
- Se a concordância observada é maior do que a concordância devida ao acaso, $k > 0$.
- Se a concordância observada é igual à concordância devida ao acaso, $k = 0$.
- Se a concordância observada é menor do que a concordância devida ao acaso, $k < 0$, sendo que o seu valor minímo depende das distribuições marginais.

A variância de k pode ser estimada através de:

$$\begin{aligned} var\,(k) \;=\; & \frac{1}{n\,(1-P_c)^4}\{\sum_{i=1}^{r} p_{ii}\,[(1-P_c) - (p_{.i} + p_{i.})\,(1-P_o)] + \\ & + (1-P_o)^2 \sum_{i=1, i\neq j}^{r}\sum_{j=1}^{r} p_{ij}\,(p_{.i} + p_{j.})^2 - (P_o P_c - 2P_c + P_o)^2\} \end{aligned} \tag{4.20}$$

onde p_{ij} representa a proporção de observações na ij-ésima célula e $p_{i.}$ e $p_{.i}$ são as proporções marginais totais.

Exemplo[9]: Dois psiquiatras avaliaram 118 pacientes segundo 5 diferentes categorias de diagnóstico:

		Psiquiatra 1				
		D1	**D2**	**D3**	**D4**	**D5**
Psiquiatra 2	**D1**	22	2	2	0	0
	D2	5	7	14	0	0
	D3	0	2	36	0	0
	D4	0	1	14	7	0
	D5	0	0	3	0	3

$$\begin{aligned} P_o &= \sum_{i=1}^{r} \frac{n_{ii}}{N} = \frac{22}{118} + \frac{7}{118} + \frac{36}{118} + \frac{7}{118} + \frac{3}{118} = \\ &= 0.636 \end{aligned}$$

$$\begin{aligned} P_c &= \frac{1}{N}\left(\sum_{i=1}^{r} \frac{n_{i.}n_{.i}}{N}\right) = \\ &= \frac{1}{118}\left[\frac{26\times 27}{118} + \frac{26\times 12}{118} + \frac{38\times 71}{118} + \frac{22\times 7}{118} + \frac{6\times 3}{118}\right] = \\ &= 0.008\left[5.95 + 2.64 + 22.86 + 1.31 + 0.15\right] = 0.278 \end{aligned}$$

$$k = \frac{P_o - P_c}{1 - P_c} = \frac{0.636 - 0.278}{1 - 0.278} = 0.496$$

Pelo valor de k ($k = 0.496$), podemos dizer que existe concordância moderada.

Para calcular a variância, precisamos transformar a tabela de frequências absolutas numa tabela de proporções:

		Psiquiatra 1					
		D1	**D2**	**D3**	**D4**	**D5**	**Totais**
Psiquiatra 2	**D1**	0.186	0.017	0.017	0.000	0.000	0.220
	D2	0.042	0.059	0.119	0.000	0.000	0.220
	D3	0.000	0.017	0.305	0.000	0.000	0.322
	D4	0.000	0.008	0.119	0.059	0.000	0.186
	D5	0.000	0.000	0.025	0.000	0.025	0.051
	Totais	0.229	0.102	0.585	0.059	0.025	1

$$\begin{aligned} var\,(k) &= \frac{1}{n\,(1-P_c)^4}\{\sum_{i=1}^{r} p_{ii}\left[(1-P_c) - (p_{.i} + p_{i.})\,(1-P_o)\right] + \\ &\quad + (1-P_o)^2 \sum_{i=1, i\neq j}^{r}\sum_{j=1}^{r} p_{ij}\,(p_{.i} + p_{j.})^2 - (P_oP_c - 2P_c + P_o)^2\} \\ &= 0.0036 \end{aligned}$$

4.6.1 Kappa Ponderado

Quando as categorias estão ordenadas, o grau de discordância depende da diferença entre as classificações, isto é, entre os resultados obtidos por cada uma das variáveis. A medida *Kappa ponderado* (Spitzer et al. 1976) usa ponderações $\{w_{ij}\}$ que satisfazem a condição $0 \leq w_{ij} \leq 1$ com todos $w_{ii} = 1$ e todos $w_{ij} = w_{ji}$ para descrever a proximidade da concordância. Escolhas frequentes para as ponderações são:

$$\left\{ w_{ij} = \frac{1 - |i-j|}{r-1} \right\} \text{ e } \left\{ w_{ij} = \frac{1 - (i-j)^2}{(r-1)^2} \right\}.$$

Para ambas estas ponderações, a concordância é mais fraca e a discordância mais forte para células mais distantes da diagonal principal.

A proporção da concordância ponderada observada é:

$$P_{o(w)} = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} p_{ij} \tag{4.21}$$

onde as proporções são calculadas através de $\frac{n_{ij}}{N}$. A proporção da concordância ponderada esperada devido ao acaso será:

$$P_{c(w)} = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} p_{i.} p_{.j} \tag{4.22}$$

O *Kappa Ponderado* é dado por:

$$k_w = \frac{P_{o(w)} - P_{c(w)}}{1 - P_{c(w)}}. \tag{4.23}$$

O denominador iguala o numerador quando $P_{o(w)} = 1$, correspondendo a uma concordância perfeita. $k_w = 0$ quando o modelo de independência é sustentado e $k_w = 1$ quando existe concordância perfeita. Quanto mais forte for a concordância ponderada, mais alto o valor de k_w para as distribuições marginais dadas. De notar que quando $w_{ij} = 0$ para todos $i \neq j$, então o *Kappa ponderado* é igual ao *Kappa* geral.

A interpretação da magnitude do *Kappa ponderado* é semelhante à do *Kappa* geral. Com $\left\{ w_{ij} = \frac{1-(i-j)^2}{(r-1)^2} \right\}$, k_w é uma medida de correlação intraclasses para uma Análise de Variância dupla tratando os indivíduos avaliados e os avaliadores como amostras aleatórias de indivíduos e avaliadores. É uma medida da proporção da variabilidade explicada.

Uma desvantagem das medidas *Kappa* e *Kappa Ponderado* é que os seus valores dependem fortemente das distribuições marginais. O mesmo processo de avaliação de diagnóstico pode produzir diferentes valores de k e k_w. Valores de $\widehat{k_w}$ para diferentes tabelas devem ser comparados apenas se forem usadas as mesmas ponderações e margens semelhantes. Graham e Jackson (1993) denotaram que $\widehat{k_w}$ descreve melhor associação do que concordância.

4.7 Coeficiente de Concordância de Kendall

O conceito de correlação entre duas variáveis pode ser expandido para considerar concordância entre mais do que duas variáveis. Esta concordância é facilmente medida não-parametricamente pelo Coeficiente de Concordância de Kendall. Um uso comum deste coeficiente de concordância é para expressar a intensidade da concordância entre os vários *rank*ings.

Várias fórmulas equivalentes para o coeficiente de Kendall podem ser encontradas em vários textos. Uma de fácil uso é:

$$W = \frac{\sum R_i^2 - \frac{(\sum R_i)^2}{n}}{\frac{M^2(n^3-n)}{12}} \tag{4.24}$$

onde M é o número de variáveis correlacionadas e n o número de observações por variável. R_i é igual à soma dos *ranks* na linha i. Outra fórmula equivalente é:

$$W = \frac{12\sum R_i^2 - 3M^2 n\,(n+1)^2}{M^2\,(n^3-n)}. \tag{4.25}$$

O valor de W pode variar entre 0 (quando não existe concordância) e 1 (quando existe perfeito acordo entre o *ranking* de todos os grupos).

Poderemos questionar se o valor calculado para W é significativo, isto é, se representa uma concordância significativamente diferente de zero na respetiva população. Uma forma simples de o fazer envolve a relação entre o coeficiente de concordância W e o Qui-quadrado de Friedman, X_r^2:

$$X_r^2 = M\,(n-1)\,W. \tag{4.26}$$

Assim podemos converter um W no seu equivalente X_r^2 e depois ver na tabela de valores críticos (Anexo E) o seu valor. Se n ou M são maiores que os valores encontrados nesta tabela, então poderemos assumir que X_r^2 será aproximadamente uma v.a. χ^2 com $n-1$ graus de liberdade.

Exemplo[30]: A cada uma de 3 crianças foi pedido que avaliasse e ordenasse por ordem de preferência seis sabores de gelado. Queremos saber se as três avaliações foram semelhantes.

H_0 : Não existe concordância entre as classificações dadas pelas três crianças.
H_1 : Existe concordância entre as classificações dadas pelas três crianças.

	Sabores					
Criança	**1**	**2**	**3**	**4**	**5**	**6**
1	5	1	3	2	4	6
2	6	2	3	1	5	4
3	6	3	2	1	4	5
Soma (R_i)	17	6	8	4	13	15

$M = 3$ e $n = 6$

$$W = \frac{\sum R_i^2 - \frac{(\sum R_i)^2}{n}}{\frac{M^2(n^3-n)}{12}} = \frac{(17^2+6^2+8^2+4^2+13^2+15^2) - \frac{63^2}{6}}{\frac{3^2(6^3-6)}{12}} = \\ = \frac{137.5}{157.5} = 0.873$$

$$X_r^2 = M\,(n-1)\,W = 3\,(6-1)\,0.873 = 13.095$$

Pela tabela $(\chi_r^2)_{0.05,3,6} = 7.000$. Como $X_r^2 > (\chi_r^2)_{0.05,3,6}$ rejeitamos H_0 logo existe concordância.

Exemplo[30]: Coeficiente de Concordância de Kendall para três diferentes partes do corpo de 12 aves.

Ave	**Asa(cm)**		**Cauda(cm)**		**Bico(mm)**		**Soma dos**
i	**Valor**	***rank***	**Valor**	***rank***	**Valor**	***rank***	***ranks*** (R_i)
1	10.4	4	7.4	5	17	5.5	14.5
2	10.8	8.5	7.6	7	17	5.5	21
3	11.1	10	7.9	11	20	9.5	30.5
4	10.2	1.5	7.2	2.5	14.5	2	6
5	10.3	3	7.4	5	15.5	3	11
6	10.2	1.5	7.1	1	13	1	3.5
7	10.7	7	7.4	5	19.5	8	20
8	10.5	5	7.2	2.5	16	4	11.5
9	10.8	8.5	7.8	9.5	21	11	29
10	11.2	11	7.7	8	20	9.5	28.5
11	10.6	6	7.8	9.5	18	7	22.5
12	11.4	12	8.3	12	22	12	36

$M = 3$ e $n = 12$

H_0 : Não existe concordância entre as três variáveis.
H_1 : Existe concordância entre as três variáveis.

Sem correção para empates

$$W = \frac{\sum R_i^2 - \frac{\left(\sum r_i^2\right)^2}{n}}{\frac{M^2(n^3-n)}{12}} = \frac{1175.5}{1287} = 0.913$$

$$X_r^2 = M\,(n-1)\,W = 3\,(12-1)\,0.913 = 30.129$$

$g.l. = n - 1 = 11$
Como $\chi^2_{0.05}\,(11) = 19.675$ então rejeitamos H_0.

4.7.1 Coeficiente de concordância de Kendall com *ranks* empatados

Se temos empates nos *ranks* dentro de qualquer dos M grupos, então a média desses *ranks* é atribuída a cada uma das observações envolvidas no empate. W será calculado com uma correção para empates:

$$W_c = \frac{\sum R_i^2 - \frac{(\sum R_i)^2}{n}}{\frac{M^2(n^3-n) - M\sum t}{12}} \tag{4.27}$$

onde

$$\sum t = \sum_{i=1}^{m} \left(t_i^3 - t_i\right) \tag{4.28}$$

com t_i o número de empates no i-ésimo grupo de empates e m é o número de grupos com *ranks* empatados.

W_c não diferirá muito de W a não ser que tenhamos um grande número de empates.

Uma equação equivalente para o cálculo de W_c é:

$$W_c = \frac{12\sum R_i^2 - 3Mn\,(n+1)^2}{M^2\,(n^3-n) - M\sum t}. \tag{4.29}$$

Continuação do exemplo anterior usando a correção para empates:

Grupo "Asa": 2 empates com 10.2 cm ($t_1 = 2$); 2 empates com 10.8 cm ($t_2 = 2$)

Grupo "Cauda": 2 empates com 7.2 cm ($t_3 = 2$); 3 empates com 7.4 cm ($t_4 = 3$); 2 empates com 7.8 cm ($t_5 = 2$)

Grupo "Bico": 2 empates com 17 mm ($t_6 = 2$); 2 empates com 20mm ($t_7 = 2$).

$$\begin{aligned}
\sum t &= \sum_{i=1}^{7} \left(t_i^3 - t_i\right) = \\
&= \left(2^3-2\right) + \left(2^3-2\right) + \left(2^3-2\right) + \left(3^3-3\right) + \left(2^3-2\right) + \left(2^3-2\right) + \left(2^3-2\right) = \\
&= 60
\end{aligned}$$

$$W_c = \frac{\sum R_i^2 - \frac{\left(\sum R_i^2\right)^2}{n}}{\frac{M^2(n^3-n)-M\sum t}{12}} = \frac{1175.5}{\frac{1544-3\times 60}{12}} = 0.924$$

Testar a significância de W_c:
$(X_r^2)_c = M\,(n-1)\,W_c = 3\,(12-1)\,0.924 = 30.492$

Como $\chi^2_{0.05}\,(11) = 19.675$ neste caso também se rejeita H_0.

4.7.2 Relação entre W e R_s

É interessante ver que W está relacionado com o valor médio de todos os possíveis coeficientes de correlação de *ranks* de Spearman que podem ser obtidos através de todos os pares de variáveis. Estes coeficientes de correlação poderão ser listados em forma de matriz:

$$\begin{array}{ccccc} (R_s)_{11} & (R_s)_{12} & (R_s)_{13} & \cdots & (R_s)_{1M} \\ (R_s)_{21} & (R_s)_{22} & (R_s)_{23} & \cdots & (R_s)_{2M} \\ (R_s)_{31} & (R_s)_{32} & (R_s)_{33} & \cdots & (R_s)_{3M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (R_s)_{M1} & (R_s)_{M2} & (R_s)_{M3} & \cdots & (R_s)_{MM} \end{array}$$

Cada elemento da diagonal principal, $(R_s)_{ii}$, tem valor igual a 1.0 e cada valor abaixo da diagonal principal é o mesmo que se encontra acima da diagonal principal em posição idêntica, uma vez que $(R_s)_{ik} = (R_s)_{ki}$. Existem $M!/[2(M-2)!]$ possibilidades diferentes de R_s's para M variáveis.

No exemplo das aves, tínhamos três R_s's: $(R_s)_{12}$ para os comprimentos das asas e da cauda; $(R_s)_{13}$ para os comprimentos das asas e do bico; $(R_s)_{23}$ para os comprimentos da cauda e do bico. A matriz dos coeficientes de correlação de *ranks* de Spearman, usando correção para empates, seria:

$$\begin{array}{ccc} 1.000 & 0.852 & 0.917 \\ 0.852 & 1.000 & 0.890 \\ 0.917 & 0.890 & 1.000 \end{array}$$

Para o exemplo das crianças seria:

$$\begin{array}{ccc} 1.000 & 0.771 & 0.771 \\ 0.771 & 1.000 & 0.890 \\ 0.771 & 0.890 & 1.000 \end{array}$$

Sendo a média de R_s denotada por $\overline{R}_s$, a relação com W é:

$$W = \frac{(M-1)\,\overline{R}_s + 1}{M} \tag{4.30}$$

portanto,

$$\overline{R}_s = \frac{MW_c - 1}{M - 1} \tag{4.31}$$

Se existem empates, então as duas equações acima relacionam W_c e $\left(\overline{R}_s\right)_c$ da mesma maneira que W e $\overline{R}_s$ estão relacionados. Enquanto que W pode variar entre 0 e 1, $\overline{R}_s$ varia de $(-1)/((M-1))$ a 1. Para o exemplo das aves, $(\overline{R}_s)_c = (0.852 + 0.917 + 0.890)/3 = 0.886$. Para o exemplo das crianças, $\left(\overline{R}_s\right)_c = 0.809$, e é possível ver que as duas equações anteriores se mantêm.

Se $M = 2$ (ou seja, existem apenas duas variáveis sendo correlacionadas), então tanto R_s como W podem ser calculados:

$$W = \frac{\overline{R}_s + 1}{2} \tag{4.32}$$

e

$$r_s = 2W_c - 1. \tag{4.33}$$

Quando $M = 2$, o uso de R_s é preferível pois varia entre -1 e 1 e existem tabelas de valores críticos mais completas disponíveis.

Se se conclui que há concordância significativa para cada um de dois grupos de dados, podemos querer averiguar se a concordância dentro de cada grupo é a mesma para ambos os grupos. Por exemplo, os dados no exemplo das crianças são valores para preferências no sabor dos gelados avaliados por três crianças. Poderíamos ter um conjunto de dados semelhantes para a preferência demonstrada por vários adultos em relação aos mesmos sabores. Se existisse concordância significativa entre as crianças assim como entre os adultos, poderíamos questionar se o consenso entre crianças é o mesmo que entre adultos. Um teste com este propósito foi apresentado por Schucany (1975) e Frawley (1973), com elaboração de Li e Schucany (1975). No entanto, o teste de hipóteses nem sempre é conclusivo no que diz respeito à concordância entre os dois grupos e foi criticado por Hollander e Sethuraman (1978) que propuseram um procedimento diferente. Serlin e Marascuilo (1983) reexaminaram ambas as hipóteses assim como as comparações múltiplas.

4.8 Correlação Ponderada entre *ranks*

A correlação entre *ranks* abordada no coeficiente de correlação de Spearman dá semelhante ênfase a cada par de observações, mas existem situações em que é de interesse saber se existe correlação entre as observações mais altas (ou mais baixas) das variáveis aleatórias em estudo.

Para estes casos é melhor ter um procedimento que dê mais peso à concordância entre amostras onde as observações tenham *ranks* mais altos ou mais baixos. Quade e Salama (1992) referem-se a este método como "Correlação Ponderada entre *ranks*".

O coeficiente de correlação de Pearson terá de ser calculado usando os números de Savage:

$$R_T = \frac{\sum_{i=1}^{n} (S_i)_1 (S_i)_2 - n}{(n - S_1)} \tag{4.34}$$

onde os números de Savage (que também podem ser encontrados na tabela do Anexo D) são dados por

$$S_i = \sum_{j=i}^{n} \frac{1}{j}. \tag{4.35}$$

Testar a significância de R_T significa testar $H_0 : \rho_T = 0 \; vs \; H_1 : \rho_T > 0$.

A tabela no Anexo F dá-nos os valores críticos para R_T. Para amostras maiores que as referenciadas na tabela, é válida uma aproximação à v.a. normal:

$$Z = \frac{R_T}{\sqrt{n-1}} \tag{4.36}$$

$R_T = 1.0$ quando existe concordância perfeita entre os *ranks* de dois grupos de dados. Se os *ranks* são completamente opostos nas duas amostras então $R_T = -1.0$ apenas se $n = 2$; este valor aproxima-se de -0.645 à medida que n aumenta.

Quando existem empates nos *ranks*, devemos usar a média dos números de Savage das posições de empate e calcular R_T usando a fórmula 3.1 do coeficiente de correlação de Pearson (página 36).

Quando fazemos esta análise de correlação testamos as concordâncias para os *ranks* do topo da lista.

Exemplo[30]: Um estudo determinou a importância relativa de oito fatores ecológicos no sucesso de uma determinada espécie de pássaro num habitat específico. Um outro estudo fez o mesmo para outra espécie no mesmo habitat. Pretende-se determinar se os mesmos fatores ecológicos têm importância semelhante para ambas as espécies, ou seja, queremos ver se existe uma correlação positiva entre os fatores mais importantes para uma espécie e os mais importantes para a outra. Trata-se então de um teste unilateral.

H_0 : Os mesmos fatores ecológicos não são os mais importantes para ambas as espécies de pássaros.

H_1 : Os mesmos fatores ecológicos são os mais importantes para ambas as espécies de pássaros.

			Número de Savage (S_i)		
factor (i)	**Espécie 1**	**Espécie 2**	**Espécie 1**	**Espécie 2**	$(S_i)_1 (S_i)_2$
A	1	1	2.718	2.718	7.388
B	2	2	1.718	1.718	2.952
C	3	3	1.218	1.218	1.484
D	4	7	0.885	0.268	0.237
E	5	8	0.635	0.125	0.079
F	6	6	0.435	0.435	0.189
G	7	5	0.268	0.635	0.170
H	8	4	0.125	0.885	0.111
		Somatório	8.002	8.002	12.610

$n = 8$

$$\begin{aligned} R_T &= \frac{\sum_{i=1}^{n} (S_i)_1 (S_i)_2 - n}{(n - S_1)} \\ r_T &= \frac{12.610 - 8}{8 - 2.718} = 0.873 \end{aligned}$$

Temos pela tabela do anexo F, um valor crítico igual a 0.692 ($\alpha = 0.05$), o que resulta na rejeição de H_0. Assim, concluímos existir concordância significativa entre os fatores ecológicos mais importantes para as duas espécies de pássaro.

4.9 Concordância Top-Down

Quando existem mais do que dois grupos de *ranks* podemos querer avaliar apenas os *ranks* mais altos (ou os mais baixos). No primeiro exemplo da secção 4.2, podemos desejar saber se as crianças concordam nos sabores preferidos, sendo menos importante se concordam nos menos apreciados. Tal como na situação de correlação que abordámos na secção anterior, podemos aplicar os scores de Savage, S_i (Anexo D), e o coeficiente de concordância é dado por:

$$C_T = \frac{1}{M^2 (n - S_1)} \left(\sum_{i=1}^{n} R_i^2 - M^2 n \right). \tag{4.37}$$

A estatística de teste para avaliar a significância deste coeficiente é

$$X_T^2 = M (n - 1) C_T \tag{4.38}$$

que segue aproximadamente uma distribuição Qui-quadrado com $n - 1$ $g.l.$ (Iman e Conover, 1987). Aqui n e M são os mesmos que nos cálculos anteriores: cada um dos

M grupos tem n *ranks*. Também R_i é a soma dos scores de Savage, para todos os M grupos, no *rank* i e S_1 é o primeiro score de Savage.

Exemplo[30]: Queremos avaliar se existe concordância significativa entre as crianças relativamente ao melhor sabor de gelado (dados do primeiro exemplo da secção 4.7). Na tabela estão os scores de Savage em vez dos *ranks*, atribuindo-se estes scores por ordem crescente, ou seja, S_1 é atribuído ao *rank* 1 (o sabor preferido), S_2 ao *rank* 2 e assim sucessivamente.

	Sabores					
Criança	**1**	**2**	**3**	**4**	**5**	**6**
1	0.367	2.450	0.950	1.450	0.617	0.167
2	0.167	1.450	0.950	1.240	0.367	0.617
3	0.167	0.950	1.450	1.240	0.617	0.367
Soma (R_i)	0.701	4.850	3.350	3.930	1.601	1.151

H_0 : Não existe concordância nos sabores preferidos pelas três crianças.
H_1 : Existe concordância nos sabores preferidos pelas três crianças.

$$\begin{aligned} C_T &= \frac{1}{M^2\left(n-S_1\right)}\left(\sum_{i=1}^{n} R_i^2 - M^2 n\right) = \\ &= \frac{1}{3^2\left(6-2.450\right)}\left[\left(0.701^2+4.850^2+3.350^2+3.930^2+1.601^2+1.151^2\right)-3^2\times 6\right] = \\ &= 0.03130\times 0.5693 = 0.01782 \end{aligned}$$

$$\begin{aligned} X_T^2 &= M\left(n-1\right)C_T \\ X_T^2 &= 3\left(6-1\right)0.01782 = 0.267 \end{aligned}$$

$g.l. = n-1 = 5$
$\chi^2_{0.05}\left(5\right) = 11.070$
Assim, não rejeitamos H_0.

Poderíamos fazer o mesmo teste para os sabores menos apreciados. Neste caso atribuir-se-iam os scores de Savage por ordem inversa, ou seja, S_1 atribuído ao *rank* 6, S_2 ao *rank* 5 e assim sucessivamente. Neste caso teríamos $C_T = 0,8222$ e $X_T^2 = 12,333$, o que resultaria na rejeição de H_0, isto é, de não existir concordância nos sabores menos preferidos.

Capítulo 5

Alguns testes não paramétricos para a distribuição de variáveis categorizadas ordinais

5.1 Introdução

Muitas vezes, ao pretendermos comparar duas ou mais distribuições, deparamo-nos com dados provenientes da observação de variáveis aleatórias categorizadas ordinais. Neste capítulo serão abordados os testes mais frequentemente usados e que podem ser realizados com recurso a *softwares* estatísticos.

5.2 Teste de Mann-Whitney

Este é um teste não paramétrico que usa os *ranks* de duas amostras aleatórias independentes para testar a hipótese de essas amostras serem provenientes de populações com a mesma distribuição relativamente à variável em estudo, ou seja, que as duas populações são idênticas no que concerne esta variável.

Seja $X_1, \ldots, X_n$ a amostra aleatória da primeira população e $Y_1, \ldots, Y_m$ a amostra aleatória da segunda população. Juntam-se as observações de ambas as amostras e ordenam-se por ordem crescente as $n + m$ observações, atribuindo a cada uma delas um *rank* ou ordem (de 1 a $n + m$). Se existirem observações iguais atribui-se a cada uma delas a média dos *ranks* envolvidos no empate. Seja $R(X_i)$ o *rank* atribuído à observação X_i da primeira população e $R(Y_j)$ o *rank* atribuído à observação Y_j da segunda população. Seja $N = n + m$.

A hipótese a ser testada é $H_0 : F(x) = G(x), \forall x \quad vs \quad H_1 : \exists x, F(x) \neq G(x)$, sendo $F(x)$ a função de distribuição da variável aleatória X em estudo para a primeira população e $G(x)$ a função de distribuição de X para a segunda população. Se admitirmos que a haver alguma diferença entre as funções de distribuição, essa diferença

diz respeito ao valor médio das populações, então testamos: $H_0 : E(X) = E(Y)$ vs $H_1 : E(X) \neq E(Y)$.

A estatística de teste é dada por:

$$T = \sum_{i=1}^{n} R(X_i). \tag{5.1}$$

Este teste possui uma tabela própria que contém valores para $n, m \leq 20$. Nos casos em que $n > 20$ ou $m > 20$, a região de rejeição é $T < w_{\frac{\alpha}{2}}$ ou $T > w_{1-\frac{\alpha}{2}}$ em que o quantil $w_{1-\frac{\alpha}{2}}$ é calculado da seguinte forma:

$$w_{1-\frac{\alpha}{2}} = n(N+1) - w_{\frac{\alpha}{2}}. \tag{5.2}$$

e $w_{\frac{\alpha}{2}}$ pode ser aproximado por:

$$w_{\frac{\alpha}{2}} = \frac{n(N+1)}{2} + z_{\frac{\alpha}{2}}\sqrt{\frac{nm(N+1)}{12}} \tag{5.3}$$

onde $z_{\frac{\alpha}{2}}$ é o valor da variável aleatória Normal reduzida que deixa para a esquerda uma área de $\frac{\alpha}{2}$.

Se temos muitos pares, é indicado utilizarmos a estatística de teste :

$$T_1 = \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}} \tag{5.4}$$

que segue a distribuição Normal estandardizada.

O teste de Mann-Whitney para duas amostras independentes foi posteriormente generalizado ao problema de analisar k amostras, $k > 2$, por Kruskal e Wallis.

5.3 Teste de Kruskal-Wallis

Este teste é apropriado para testar se, num conjunto de k ($k \geq 2$) amostras aleatórias extraídas de k possivelmente diferentes populações, as k populações são idênticas. No caso de $k = 2$, este teste é idêntico ao teste de Mann-Whitney.

Para este teste temos os seguintes pressupostos:

1. cada amostra foi selecionada de forma aleatória da população que representa;
2. as k amostras são independentes umas das outras.

Seja:

Amostra 1	**Amostra 2**	$\cdots$	**Amostra** k
$X_{1,1}$	$X_{2,1}$	$\cdots$	$X_{k,1}$
$X_{1,2}$	$X_{2,2}$	$\cdots$	$X_{k,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1,n_1}	X_{2,n_2}	$\cdots$	X_{k,n_k}

Juntam-se as amostras numa só e ordena-se por ordem crescente atribuindo um *rank* a cada uma (de 1 até N). No caso de observações empatadas atribui-se a média dos *ranks* envolvidos.

Seja $R(X_{ij})$ o *rank* atribuído à observação X_{ij}. Sendo R_i a soma dos *ranks* atribuídos às observações da amostra i, calcula-se este valor para todas as amostras.

Hipótese a ser testada:
H_0 : Todas as funções de distribuição da v.a. X para as k populações são idênticas.
vs
H_1 : Pelo menos uma das populações tem tendência a originar observações maiores do que as outras.

Ou, admitindo que todas as populações têm a mesma distribuição,
$H_0 : E(X_1) = E(X_2) = \cdots = E(X_k)$. *vs* $H_1 : \exists i, j : E(X_i) \neq E(X_j)$.

Estatística de teste:

$$T = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1). \tag{5.5}$$

Região de rejeição: $T > w_{1-\alpha}$, onde $w_{1-\alpha}$ pode ser obtido da seguinte forma:

- Se $k \leq 5$ e as dimensões das amostras são pequenas ($n < 8$ [30]), então podemos consultar uma tabela própria para este teste;
- Quando $k > 5$ e/ou temos amostras grandes, podemos considerar uma aproximação à variável aleatória Qui-quadrado com $k - 1$ *g.l.*

Exemplo[26]: Um psicólogo conduziu um estudo para avaliar se o ruído atrapalha ou não o processo de aprendizagem. Quinze indivíduos são distribuídos aleatoriamente por 3 grupos e a cada um deles são dados 20 minutos para memorizar 10 sílabas sem sentido sendo avaliados no dia seguinte. Os indivíduos do grupo 1 estudam as sílabas numa sala em silêncio. Os 5 indivíduos do grupo 2 estudam numa sala com ruído moderado. Os indivíduos do grupo 3 fazem o estudo numa sala muito ruídosa. O número de sílabas corretas em cada grupo é:

Indivíduo	**Grupo 1**	**Grupo 2**	**Grupo 3**
1	8	7	4
2	10	8	8
3	9	5	7
4	10	8	5
5	9	5	7

Existem indícios de que o barulho afetou a capacidade de aprendizagem dos indivíduos?

É necessário juntar as amostras e atribuir a cada observação o *rank* correspondente:

Indivíduo	**Grupo 1**	R_1	**Grupo 2**	R_2	**Grupo 3**	R_3
1	8	9.5	7	6	4	1
2	10	14.5	8	9.5	8	9.5
3	9	12.5	5	3	7	6
4	10	14.5	8	9.5	5	3
5	9	12.5	5	3	7	6
$\sum R_i$		63.5		31		25.5

$$\begin{aligned} T &= \frac{12}{N\left(N+1\right)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3\left(N+1\right) = \\ &= \frac{12}{15\left(15+1\right)}\left[\frac{63.5^2}{5}+\frac{31^2}{5}+\frac{25.5^2}{5}\right] - 3\left(15+1\right) = \\ &= 0.05\left[806.45+192.2+130.05\right] - 48 = \\ &= 8.435 \end{aligned}$$

Como as amostras têm dimensão 5, procuramos o valor correspondente na tabela. Como $H_{0.05,5,5,5} = 5.780$, rejeitamos a hipótese de que existem diferenças na aprendizagem pelos indivíduos dos três diferentes grupos.

5.3.1 Correção para empates

Algumas fontes sugerem que no caso de existirem demasiados empates na distribuição dos *ranks*, o valor da estatística de Kruskal-Wallis deverá sofrer um ligeiro ajuste. A correção para empates resultará num ligeiro aumento no valor de T mas tornando o teste mais potente.

A equação para esta correção é dada por:

$$C = 1 - \frac{\sum_{i=1}^{s}\left(t_i^3 - t_i\right)}{N^3 - N} \tag{5.6}$$

onde s representa o número de conjuntos de empates e t_i o número de observações empatadas no $i-ésimo$ conjunto de empates.

A estatística de teste passa a ser então:

$$T_C = \frac{T}{C}. \tag{5.7}$$

Para o caso do exemplo anterior teríamos:

$$\begin{aligned}\sum_{i=1}^{s}\left(t_i^3 - t_i\right) &= \left[3^3-3\right]+\left[3^3-3\right]+\left[4^3-4\right]+\left[2^3-2\right]+\left[2^3-2\right] = \\ &= 24+24+60+6+6 = 120\end{aligned}$$

$$C = 1 - \frac{\sum_{i=1}^{s}\left(t_i^3 - t_i\right)}{N^3 - N} = 1 - \frac{120}{15^3 - 15} = 0.964$$

$$T_C = \frac{T}{C} = \frac{8.435}{0.964} = 8.75$$

o que não altera as conclusões tiradas anteriormente.

5.3.2 Comparações Múltiplas

Se rejeitamos a hipótese nula podemos utilizar os processo das comparações múltiplas para determinar quais os pares de populações que diferem entre si.

Consideram-se diferentes as populações i e j que satisfazem a seguinte inequação:

$$\left|\frac{R_i}{n_i} - \frac{R_j}{n_j}\right| > t_{1-\frac{\alpha}{2}}\left(S^2\frac{N-1-T}{N-k}\right)^{\frac{1}{3}}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)^{\frac{1}{2}} \tag{5.8}$$

em que:

- R_i e R_j são as somas das ordens das duas amostras;
- $t_{1-\frac{\alpha}{2}}$ é o quantil $1-\frac{\alpha}{2}$ da distribuição t com $N-k$ $g.l.$;
- $S^2 = \frac{1}{N-1}\left(\sum R\left(X_{ij}\right)^2 - N\frac{(N+1)^2}{4}\right)$.

Utiliza-se o mesmo α que para o teste de Kruskal-Wallis.

5.4 Teste de Wilcoxon

O teste de Wilcoxon é um teste não-paramétrico usado quando uma variável é observada na mesma população em dois momentos diferentes, obtendo-se assim duas amostras emparelhadas. O objetivo é usar os *ranks* associados à amostra das diferenças dos pares de observações para testar a hipótese da variável em estudo ter a mesma distribuição nos dois momentos de observação.

Testam-se as hipóteses:
H_0 : As amostras provêm de populações com a mesma distribuição, para a v.a. em estudo.
vs
H_1 : As amostras provêm de populações com distribuições diferentes, para a v.a. em estudo.

Para cada par de valores calcula-se a diferença d subtraindo o segundo valor ao primeiro, mantendo os sinais e eliminando todos aqueles cuja diferença seja zero:

Amostra 1	**Amostra 2**	d
x_1	y_1	$d_1 = x_1 - y_1$
x_2	y_2	$d_2 = x_1 - y_2$
$\vdots$	$\vdots$	$\vdots$
x_N	y_N	$d_N = x_N - y_N$

Ordenam-se as diferenças por ordem crescente usando o seu valor absoluto e atribui-se uma ordem a cada uma, ou seja, um *rank*. No caso das diferenças negativas, acrescenta-se a cada *rank* o sinal negativo. Se existirem empates, atribui-se o valor médio dos *ranks* envolvidos no empate. Determina-se a soma dos valores absolutos dos *ranks* negativos ($\sum R_-$) e a soma dos *ranks* positivos ($\sum R_+$).

Estatística de Teste:

- Se $N \leq 25$, ou seja, amostras pequenas: seja T a menor das somas determinados anteriormente, ou seja, $T =$ *Soma dos ranks positivos ou negativos*, conforme o que for menor em valor absoluto. Região de Rejeição: $T <$ *valor obtido para o* α *escolhido* (tabela própria do teste - Anexo C).

- Se $N > 25$:

$$Z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} \backsim N(0,1). \tag{5.9}$$

Região de Rejeição: $Z < z_{\frac{\alpha}{2}}$ ou $Z > z_{\frac{\alpha}{2}}$.

Se a igualdade seguinte não for respeitada, então algum erro foi cometido durante os cálculos:

$$\sum R_+ + \sum R_- = \frac{N\,(N+1)}{2}. \tag{5.10}$$

Quando $\sum R_+$ e $\sum R_-$ são equivalentes, então ambos os valores irão igualar $[n\,(n+1)]\,/4$. Este valor é conhecido como o valor esperado da estatística T de Wilcoxon. Se o valor de $\sum R_+$ é significativamente superior ao de $\sum R_-$, então existe indicação de que a população da Amostra 1 tenha *ranks* mais altos do que a população da Amostra 2. Por outro lado, quando $\sum R_-$ tem um valor significativamente superior ao de $\sum R_+$ existe uma maior probabilidade de que a população da Amostra 2 possua *ranks* mais altos do que a população da Amostra 1.

Exemplo[1]: Considere os dados da tabela seguinte onde estão indicados os desempenhos de 15 crianças ao construir torres de blocos em duas experiências.

Criança	**1^a Exp.**	**2^a Exp.**	d	**ranks** d	**rank atribuído**
A	30	30	$30-30=0$		
B	19	6	$19-6=13$	6	6
C	19	14	$19-14=5$	4.5	4.5
D	23	8	$23-8=15$	8.5	8.5
E	29	14	$29-14=15$	8.5	8.5
F	178	52	$175-52=26$	14	14
G	42	14	$42-14=28$	10	10
H	20	22	$20-22=-2$	1	-1
I	12	17	$12-17=-5$	4.5	-4.5
J	39	8	$39-8=1$	11	11
K	14	11	$14-11=3$	2.5	2.5
L	81	30	$81-30=51$	13	13
M	17	14	$17-14=3$	2.5	2.5
N	31	17	$31-17=14$	7	7
O	52	15	$52-15=37$	12	12

H_0 : Não existem diferenças entre os desempenhos na primeira e na segunda experiência.

H_1 : Existem diferenças entre os desempenhos na primeira e na segunda experiência.

$$\begin{aligned} \sum R_- &= 5.5 \\ \sum R_+ &= 99 \end{aligned}$$

$$T = 5.5$$

Como $N = 14 \leq 25$, $T = 5.5$ será a estatística de teste com $\alpha = 0.05$, logo pela tabela do Anexo C o valor crítico do teste é 21. Portanto, como $5.5 < 21$, rejeitamos H_0, pelo que se conclui existirem diferenças nos desempenhos das crianças na primeira e na segunda experiência.

5.4.1 correção para a continuidade

Poderá ser aplicada uma correção para a continuidade à estatística de Wilcoxon. Essa correção implica uma ligeira diminuição no valor absoluto da estatística ao ser subtraido 0.5 ao valor absoluto do numerador da equação de Z. Neste caso temos:

$$Z = \frac{\left|T - \frac{N(N+1)}{4}\right| - 0.5}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} \tag{5.11}$$

5.4.2 Correção para empates

Alguns autores recomendam que se aplique uma correção para empates na fórmula de Z. Esta correção resultará num ligeiro aumento do valor absoluto de Z, mas a não ser que estejamos perante um grande número de empates, a diferença para o valor de Z não corrigido será mínima.

Teríamos então:

$$Z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24} - \frac{\sum t_i^3 - t_i}{48}}}. \tag{5.12}$$

5.5 Teste de Friedman

Este teste é utilizado em dados ordinais para testar as diferenças na distribuição da variável aleatória em estudo nas k ($k > 2$) populações, com base em k amostras emparelhadas. Os dados deste teste são *ranks*, dispostos numa tabela de dupla entrada com k colunas e N linhas. Aos valores de cada linha atribuem-se *ranks* de 1 a k e depois determinam-se as somas dos *ranks* em cada coluna (R_j).

Como em outros testes falados anteriormente, este teste poderá ser utilizado numa situação em que a mesma variável é observada em k momentos diferentes.

Os N indivíduos das amostras foram selecionados aleatoriamente.

Hipóteses a testar:

H_0 : As k amostras provêm de populações com a mesma distribuição, para a v.a. em estudo.

vs

H_1 : As k amostras provêm de populações com distribuições diferentes, para a v.a. em estudo.

Estatística de teste:

$$X_r^2 = \frac{12}{Nk\,(k+1)} \sum_{j=1}^{k} (R_j)^2 - 3N\,(k+1) \tag{5.13}$$

Região de rejeição:

- Existe uma tabela específica para este teste para valores de $k = 3$ e N de 2 a 9 e para $k = 4$ e N de 2 a 4 (Anexo E). Se o valor obtido neste tabela não supera α, rejeita-se H_0.
- Para valores de N e k maiores do que aqueles que constam da tabela, podemos usar a distribuição Qui-quadrado com $k - 1$ *g.l.*. Neste caso, se o valor da estatística de teste é maior do que o quantil da distribuição qui-quadrado correspondente ao nível de significância escolhido rejeita-se H_0.

Exemplo: Um bloco consiste de um grupo de quatro animais (porquinhos-da-Índia) que podemos assegurar que estarão em condições ambientais idênticas. Em cada bloco, a cada animal é atribuída aleatoriamente uma de quatro dietas experimentais de tal forma que cada animal em cada bloco receberá uma dieta diferente. Os resultados observados das dietas são categorizados de acordo com vários parâmetros (aumento de peso, incidência de doenças, etc.), sendo 1 a pior categoria e 9 a melhor categoria. A tabela seguinte, contém estes resultados:

	Dietas			
	I	**II**	**III**	**IV**
Bloco A	9	4	1	7
Bloco B	6	5	2	8
Bloco C	9	1	2	6

Atribui-se *ranks* por linha e em seguida calculam-se os totais por coluna:

	Dietas			
	I	**II**	**III**	**IV**
Bloco A	4	2	1	3
Bloco B	3	2	1	4
Bloco C	4	1	2	3
R_j	11	5	4	10

$$
\begin{aligned}
X_r^2 &= \frac{12}{Nk\,(k+1)} \sum_{j=1}^{k} (R_j)^2 - 3N\,(k+1) = \\
&= \frac{12}{3 \times 4\,(4+1)} \left[11^2 + 5^2 + 4^2 + 10^2\right] - 3 \times 3\,(4+1) = \\
&= \frac{12}{60} \times 262 - 45 = 7.4
\end{aligned}
$$

Temos um valor crítoco igual a 6.500 obtido pela tabela deste teste. Como $6.500 < 7.4$ que é o valor obtido através da estatística de teste, não temos evidência suficiente para rejeitar a hipótese de que as 4 dietas produzam resultados idênticos.

5.5.1 Correção para empates

No caso de um número excessivo de empates, poderá ser mais adequado aplicar uma correção para empates à equação de X_r^2. Esta correção levará a um ligeiro aumento do valor de X_r^2 mas que resultará num teste mais potente.

O factor de correção é dado por:

$$
C = 1 - \frac{\sum_{i=1}^{s} (t_i^3 - t_i)}{n\,(k^3 - k)} \tag{5.14}
$$

onde s é o número de conjuntos de empates e t_i é o número de *ranks* empatados no $i - ésimo$ conjunto de empates.

Temos então a fórmula para a estatística de Friedman com correção para empates a ser dada por:

$$
X_{r_C}^2 = \frac{X_r^2}{C}. \tag{5.15}
$$

Capítulo 6

Conclusão

Os temas focados nesta tese estão longe de ser esgotados pelos resultados aqui apresentados. Dedicámo-nos a fazer uma recolha de métodos que elucidasse a procura de medidas para avaliar a concordância em dados categorizados. Com o intuito de escrever sobre a análise de concordância para variáveis categorizadas, constatámos que tal não era possível sem estendermos o nosso estudo a outros métodos de análise de dados categorizados. É assim que surgem os testes de ajustamento e a análise de tabelas de contingência onde frequentemente é referida a concordância como a igualdade de proporções. A relação entre medidas de associação e testes de independência é, por vezes, confusa. Abordámos estes dois assuntos separadamente. Usamos os testes de independência para determinar se existe uma relação entre as duas variáveis categorizadas, mas usamos uma medida de associação para nos ajudar a compreender o tipo e intensidade desta relação.

Os testes não paramétricos para comparar distribuições de v.as., também desempenham um papel importante na análise de dados categorizados e são, por este motivo, apresentados neste trabalho. No entanto, um estudo mais aprofundado, revelou-nos que existem evoluções recentes nas aplicações de alguns testes que os tornam ligeiramente mais potentes. Apesar de as referirmos neste trabalho, poderão ser estudadas mais pormenorizadamente na literatura específica. Note-se que os *softwares* da especialidade incluem versões muito atualizadas destes testes.

Muitos assuntos tratados nesta tese estão incluídos na análise estatística de uma base de dados recolhida para um projeto na área da Medicina (no qual participàmos), cujo objetivo é avaliar a colonoscopia virtual como técnica de eleição no rastreio dos cancros colo-rectais, superando os resultados até agora dominados pela colonoscopia ótica. Teremos assim, uma excecional forma de utilizar os dados desta base de dados, nomeadamente na realização de testes que avaliem a concordância entre os dois métodos e na comparação dos valores obtidos para as medidas de associação e de concordância.

Será a continuação do trabalho realizado até este momento e que cremos que resultará em conclusões cientificamente interessantes.

Bibliografia

[1] Abreu, A. M., (2010/2011) - *Apontamentos das aulas de "Complementos de Estatística"*. Universidade da Madeira.

[2] Agresti, A., (2010). *Analysis of ordinal categorical data*, John Wiley & Sons Inc.

[3] Agresti, A., (2002). *Categorical data Analysis*, John Wiley & Sons Inc.

[4] Armitage, P., Berry, G., (1994). *Statistical Methods in Medical Research*, Blackwell Science.

[5] Bishop, Y. M. M., Fienberg, S. E., Holland, P.W., (2007). *Discrete Multivariate Analysis Theory and Practice*. Springer.

[6] Cressie, N., Read, T. R. C. (1984), *Multinomial Goodness-of-fit tests*. Em Journal of the Royal statistical society, series B, vol. 46, n^{o}3.

[7] Cureton, E. E., (1956). Rank-bisserial Correlation. Em *Psychometrika*, vol. 21, n^{o}3.

[8] D'Hainaut, L., (1992). *Conceitos e métodos da estatística*. Fundação Calouste Gulbenkian.

[9] Everitt, B. S., (1994). *The analysis of contigency tables*, Chapman & Hall.

[10] Fleiss, J. L., (1981). *Statistical methods for rates and proportions*, John Wiley & Sons Inc.

[11] Goodman, L.A., Kruskal, W. H. (1954). Measures of association for cross classifications. Em *Journal of the American Statistical Association*, Vol. 49, n^o 268.

[12] Goodman, L.A., Kruskal, W. H. (1959). Measures of association for cross classifications. II: Further discussion and references. Em *Journal of the American Statistical Association*, Vol. 54, n^o 285.

[13] Goodman, L.A., Kruskal, W. H. (1963). Measures of association for cross classifications. III: Aproximate Sampling Theory. Em *Journal of the American Statistical Association*, Vol. 58.

[14] Goodman, L.A., Kruskal, W. H. (1972). Measures of association for cross classifications. IV: Simplification of Asymptoctic Variances. Em *Journal of the American Statistical Association*, Vol. 67, nº 338.

[15] Hanbook of Biological Statistics. *http://udel.edu/˜mcdonald/statcmh.html*. Consultado em 13/08/2013.

[16] Hosmer, D. W., Lemeshow, S., (2000). *Applied Logistic Regression*, John Wiley & Sons Inc.

[17] Howell, D. C., (2010). *Statistical Methods for Psychology*, Wadsworth Cengage Learning.

[18] Jobson, J. D., (1991). *Applied Multivariate Data Analysis, Volume I: Regression and experimental Design*, Springer.

[19] Jobson, J. D., (1992). *Applied Multivariate Data Analysis, Volume II: Categorical and multivariate methods*, Springer.

[20] Lira, S. A., (2004). *Análise de Correlação: abordagem teórica e de construção dos coeficientes com aplicações*. Dissertação de Pós-graduação. Universidade Federal do Paraná. Curitiba. Brasil.

[21] Madansky, A. (1963). Tests of homogeneity for correlated samples. Em *Journal of the American Statistical Association*, Vol. 58.

[22] Maroco, J., (2010). *Análise Estatística com utilização do SPSS*, Edições Sílabo.

[23] Matthews, D.E., Farewell, V.T., (1996). *Using and Understanding Medical Statistics,* Karger.

[24] Pestana, D. D., Velosa, S. F., (2008). *Introdução à probabilidade e à estatística Volume I*, Fundação Calouste Gulbenkian - Serviço de Educação e Bolsas.

[25] Planet Math. *http://planetmath.org/cramersv*. Consultado em 15/08/2013.

[26] Sheskin, D. J., (2000). *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC

[27] Siegel, S., (1990). *Nonparametric statistics for the behavioral sciences*, Other McGraw-Hill.

[28] Upton, G. J. G., (1978). *The Analysis of Cross-tabulated Data*, John Wiley & Sons.

[29] Vasconcelos, R., (2011/2012) - *Apontamentos das aulas de "Análise de dados multivariados"*. Universidade da Madeira.

[30] Zar, J. H., (1996). *Biostatistical Analysis*. Prentice Hall International Editions.

Anexos

A - Tabela de Valores Críticos da Distribuição do Qui-Quadrado

Tabela que fornece valores x tais que $P(\chi^2_\alpha > x) = p$.

	Probabilidade							
g.l.	**0.50**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.828
2	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.816
3	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.467
5	4.351	6.626	9.236	11.070	12.833	15.086	16.750	20.515
6	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.458
7	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.322
8	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.528
14	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.123
15	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.697
16	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
17	16.338	20.489	24.769	27.587	30.191	33.409	35.713	40.790
18	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.820
20	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.315
21	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.797
22	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
23	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
24	23.337	28.241	33.196	36.415	39.364	42.980	45.559	51.179
25	24.337	29.339	34.382	37.652	40.646	44.34	46.928	52.620
30	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.703
35	34.336	40.223	46.059	49.802	53.203	57.342	60.275	66.619
40	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.402

B - Tabela da distribuição Normal Estandardizada

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

C - Tabela de valores críticos para o teste de Wilcoxon

		α			
	(uni-lateral)	**0.005**	**0.01**	**0.025**	**0.05**
	(bilateral)	**0.01**	**0.02**	**0.05**	**0.10**
	5				1
	6			1	2
	7		0	2	4
	8	0	2	4	6
	9	2	3	6	8
	10	3	5	8	11
	11	5	7	11	14
	12	7	10	14	17
	13	10	13	17	21
	14	13	16	21	26
	15	16	20	25	30
	16	19	24	30	36
	17	23	28	35	41
n	**18**	28	33	40	47
	19	32	38	46	54
	20	37	43	52	60
	21	43	49	59	68
	22	49	56	66	75
	23	53	62	73	83
	24	61	69	81	92
	25	68	77	90	101
	26	76	85	98	110
	27	84	93	107	120
	28	92	102	117	130
	29	100	111	127	141
	30	109	120	137	152

D - Números de Savage, S_i

	$i =$									
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
3	1.833	0.833	0.333							
4	2.083	1.083	0.583	0.250						
5	2.283	1.283	0.783	0.450	0.200					
6	2.450	1.450	0.950	0.617	0.367	0.167				
7	2.593	1.593	1.093	0.756	0.510	0.310	0.143			
8	2.718	1.718	1.218	0.885	0.635	0.435	0.268	0.125		
9	2.829	1.829	1.329	0.996	0.746	0.546	0.379	0.236	0.111	
10	2.929	1.929	1.429	1.096	0.846	0.646	0.479	0.336	0.211	0.100
11	3.020	2.020	1.520	1.187	0.937	0.373	0.570	0.427	0.302	0.191
12	3.103	2.103	1.603	1.270	1.020	0.820	0.653	0.510	0.385	0.274
13	3.180	2.180	1.680	1.347	1.097	0.897	0.730	0.587	0.462	0.351
14	3.252	2.251	1.752	1.418	1.168	0.968	0.802	0.659	0.534	0.423
15	3.318	2.318	1.818	1.485	1.235	1.035	0.868	0.725	0.600	0.489
16	3.381	2.381	1.881	1.547	1.297	1.097	0.931	0.788	0.663	0.552
17	3.440	2.440	1.940	1.606	1.356	1.156	0.990	0.847	0.722	0.611
18	3.495	2.495	1.995	1.662	1.412	1.212	1.045	0.902	0.777	0.666
19	3.548	2.548	2.048	1.714	1.464	1.264	1.098	0.955	0.830	0.719
20	3.598	2.598	2.098	1.764	1.514	1.314	1.148	1.005	0.880	0.769

	$i =$									
n	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
11	0.091									
12	0.174	0.083								
13	0.251	0.160	0.077							
14	0.323	0.232	0.148	0.071						
15	0.389	0.298	0.215	0.138	0.067					
16	0.452	0.361	0.278	0.201	0.129	0.062				
17	0.510	0.420	0.336	0.259	0.188	0.121	0.059			
18	0.566	0.475	0.392	0.315	0.244	0.177	0.114	0.056		
19	0.619	0.528	0.445	0.368	0.296	0.230	0.167	0.108	0.053	
20	0.669	0.578	0.495	0.418	0.346	0.280	0.217	0.158	0.103	0.050

E - Tabela da distribuição X_r^2 de Friedman

k	N	α		
(n)	(M)	**0.10**	**0.05**	**0.01**
3	**3**	6.000	6.000	
3	**4**	6.000	6.500	8.000
3	**5**	5.200	6.400	8.400
3	**6**	5.330	7.000	9.000
3	**7**	5.429	7.143	8.857
3	**8**	5.250	6.250	9.000
3	**9**	5.556	6.222	9.556
3	**10**	5.000	6.200	9.600
3	**11**	4.909	6.545	9.455
3	**12**	5.167	6.167	9.500
3	**13**	4.769	6.000	9.385
3	**14**	5.143	6.143	9.000
3	**15**	4.933	6.400	8.933
4	**2**	6.000	6.000	
4	**3**	6.600	7.400	9.000
4	**4**	6.300	7.800	9.600
4	**5**	6.360	7.800	9.960
4	**6**	6.400	7.600	10.200
4	**7**	6.429	7.800	10.371
4	**8**	6.300	7.650	10.350
4	**9**	6.467	7.800	10.867
4	**10**	6.360	7.800	10.800
4	**11**	6.382	7.909	11.073
4	**12**	6.400	7.900	11.100
4	**13**	6.415	7.985	11.123
4	14	6.343	7.886	11.143
4	**15**	6.440	8.040	11.240

		α		
n	M	**0.10**	**0.05**	**0.01**
5	**2**	7.200	7.600	8.000
5	**3**	7.467	8.533	10.133
5	**4**	7.600	8.800	11.200
5	**5**	7.680	8.960	11.680
5	**6**	7.733	9.067	11.867
5	**7**	7.771	9.143	12.114
5	**8**	7.800	9.300	12.300
5	**9**	7.733	9.244	12.444
5	**10**	7.760	9.280	12.480
6	**2**	8.286	9.143	9.714
6	**3**	8.714	9.857	11.762
6	**4**	9.000	10.286	12.714
6	**5**	9.000	10.486	13.229
6	**6**	9.048	10.571	13.619
6	**7**	9.122	10.674	13.857
6	**8**	9.143	10.714	14.000
6	**9**	9.127	10.778	14.143
6	**10**	9.143	10.800	14.299

F - Tabela de Valores Críticos de R_T

	α					
n	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
3	1.000	1.000	1.000	1.000	1.000	1.000
4	0.870	0.942	1.000	1.000	1.000	1.000
5	0.905	0.959	0.977	1.000	1.000	1.000
6	0.676	0.810	0.887	0.943	0.969	1.000
7	0.622	0.738	0.836	0.906	0.934	0.977
8	0.575	0.692	0.779	0.865	0.904	0.960
9	0.530	0.654	0.742	0.826	0.871	0.936
10	0.492	0.620	0.707	0.793	0.840	0.913
11	0.461	0.589	0.677	0.762	0.812	0.890
12	0.435	0.560	0.650	0.735	0.786	0.868
13	0.412	0.535	0.625	0.711	0.762	0.847
14	0.393	0.513	0.602	0.689	0.740	0.827
15	0.389	0.486	0.565	0.680	0.688	0.826
16	0.376	0.470	0.546	0.656	0.665	0.798
17	0.364	0.454	0.528	0.635	0.644	0.773
18	0.353	0.440	0.512	0.615	0.625	0.750
19	0.343	0.428	0.497	0.598	0.607	0.728
20	0.334	0.416	0.483	0.581	0.591	0.709
21	0.325	0.405	0.470	0.566	0.576	0.691
22	0.317	0.395	0.459	0.552	0.562	0.674
23	0.310	0.386	0.448	0.538	0.549	0.659
24	0.303	0.377	0.438	0.526	0.537	0.644
25	0.297	0.368	0.428	0.515	0.526	0.631
26	0.291	0.361	0.419	0.504	0.515	0.618
27	0.285	0.354	0.411	0.494	0.505	0.606
28	0.280	0.347	0.403	0.484	0.496	0.595
29	0.275	0.340	0.395	0.475	0.487	0.584
30	0.270	0.334	0.388	0.466	0.478	0.574