MEDIDAS DE ASSOCIAÇÃO

OTTO SCHMIDT

DISSERTAÇÃO APRESENTADA

AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO GRAU DE MESTRE

EM

ESTATÍSTICA

ORIENTADOR:

Prof. Dr. ADOLPHO WALTER PIMAZONI CANTON

- SÃO PAULO, DEZEMBRO DE 1978 -

Dedico este trabalho

à minha mulher, Maria Lúcia

aos meus pais, Vital e Maria Eleonora

aos meus filhos, Maria Cristina, Otto e Ricardo

AGRADECIMENTOS

Se não fosse o apoio moral e material recebido de tantas pessoas este trabalho não chegaria ao seu término. Fica aqui registrada a minha eterna gratidão a todos que direta ou indiretamente contribuiram para a sua execução.

Desejo destacar os que tiveram participação direta.

Do Instituto de Matemática e Estatística da U.S.P.

- *Prof. Dr. Adolpho Walter Pimazoni Hasenclever, não só pelo fato de ter sido o meu orientador mas também pela acolhida pronta, amiga e estimuladora que me proporcionou;*
- *Prof. Dr. Clóvis de Araujo Peres, pelo estímulo constante, pela sua amizade e pelo auxílio que prestou discutindo e dando sugestões sobre este trabalho;*
- *Prof. Dr. Pedro Alberto Morettin, pela sua compreensão e espírito de colaboração ;*
- *os professores, funcionários e colegas pela atenção e incentivo.*

Do Museu de Zoologia da U.S.P.

- *Dr. Paulo Emílio Vanzolini, pelo apoio e pela orientação nas aplicações da estatística;*
- *os zoólogos Dr. José Luiz Moreira Leme e José Lima de Figueiredo pela cooperação e interesse demonstrados;*
- *os colegas, zoólogos, estagiários e funcionários, em particular os da Seção de Herpetologia , pelo incentivo e colaboração.*

Finalmente, quero agradecer ao Dr. Messias Carlos Galvão Bueno, do Instituto Básico de Biologia Médica e Agrícola da UNESP, pela sua participação e ao Sr. João Baptista Esteves de Oliveira pela inestimável colaboração ao datilografar esta dissertação.

ÍNDICE

PREFÁCIO

Este trabalho trata do problema de medir a associação, ou a correlação, entre duas variáveis.

O termo correlação já está consagrado pelo uso e será empregado como sinônimo de associação, do mesmo modo que na literatura consultada sobre o assunto, nos casos em que os níveis das variáveis consideradas são representados por números e esses números são diretamente envolvidos no cálculo da associação.

Quando se mede associação, o objetivo não é verificar se existe relação formal entre as variáveis, o que se quer é saber como elas variam conjuntamente. Não existe distinção entre as variáveis, tal como variável dependente e independente. Não é demais alertarmos aqui que, significância da correlação não implica necessariamente uma relação causal. Pode ser que, após constatar a existência de associação entre duas variáveis, um pesquisador, apoiado em outras considerações, inerente ao seu campo de atuação, chegue a relações causais entre as mesmas; mas na maioria das vezes são conclusões que fogem do domínio da estatística.

De modo geral, o problema se apresenta da seguinte maneira: existe uma população de objetos, ou indivíduos, e cada um deles é classificado segundo dois ou mais critérios

(variáveis), mensuráveis ou não (o importante é que um indivíduo possa ser classificado, sem ambigüidades, como pertencendo a um dos níveis do critério considerado), e queremos saber se, dados dois critérios, certos níveis de um tendem a ocorrer junto com certos níveis do outro mais freqüentemente do que poderia ser atribuído simplesmente ao acaso.

Na maioria das vezes um pesquisador estará interessado, não somente em detetar, mas também em quantificar essa associação, mesmo porque ela poderá servir como um critério de diferenciação entre populações de objetos ou indivíduos.

Às medidas de associação daremos a denominação de índices ou coeficientes.

De acordo com as diferentes escalas de medição temos diferentes maneiras de medir associação. No presente trabalho consideramos as variáveis envolvidas, medidas pelo menos em escala ordinal com poucas exceções; numa a exigência é mais fraca pois uma das variáveis poderá ser nominal e em outras duas a restrição é mais forte, as duas variáveis deverão ser medidas em escala pelo menos intervalar.

No capítulo 1 depois de destacarmos alguns casos da literatura científica, onde existiu a preocupação de estudar associação, tomaremos contato com a terminologia utilizada no restante do trabalho e relacionaremos as principais propriedades requeridas de uma medida de associação. Além

disso mostraremos que é possível medir associação entre duas variáveis utilizando uma terceira variável como controle (medidas de associação parcial). Com o intuíto de dar uma primeira visão, geral, das medidas a serem estudas nos próximos capítulos faremos um breve resumo das mesmas fornecendo também informações históricas.

No capítulo 2 serão estudados os principais coeficientes de associação total destacando-se as suas interpretações, os estimadores amostrais, as distribuições amostrais exatas ou aproximadas (sempre que possível) desses estimadores e, quando for o caso, algumas considerações sobre o tamanho das amostras suficiente para o uso das aproximações.

No capítulo 3, estudaremos os principais coeficientes de associação parcial nos mesmos moldes do capítulo anterior.

No capítulo 4, serão apresentadas de maneira muito breve outras medidas de associação que são encontradas na literatura.

CAPÍTULO 1

INTRODUÇÃO

Freqüentemente, em aplicações estatísticas aos vários ramos das ciências, defrontamo-nos com a necessidade de medir o grau de *associação* entre duas variáveis. Um zoólogo poderá estar interessado em verificar se o número de anéis corporais de um lagarto está associado, por exemplo, com a amplitude de temperatura média anual (Vanzolini, 1968). Um antropólogo poderá desejar medir a associação entre peso, área da superfície e estatura, de indivíduos do sexo masculino de uma determinada população, com temperaturas médias das regiões em que nasceram (Newmann e Munro, 1955). Um meteorologista, por seu turno, estará interessado em associar suas previsões com as ocorrências de um determinado fenômeno atmosférico (Finley, 1884). Em todos esses casos os pesquisadores estarão utilizando *coeficientes ou índices de associação*.

Em termos gerais podemos dizer que duas variáveis X e Y são *positivamente associadas* se existe uma tendência para que valores altos de X ocorram juntos com valores altos de Y e valores baixos de X ocorram juntos com valores baixos de Y. De modo análogo diremos que X e Y são *negativamente associa-*

das se existe uma tendência para que valores altos de X ocorram juntos com valores baixos de Y e que valores baixos de X ocorram juntos com valores altos de Y. Diremos simplesmente que X e Y são associadas quando ocorer um dos dois casos.

Os índices quantitativos de associação (correlação) C(X,Y) são, de modo geral, padronizados e deles são requeridas certas propriedades de simetria e invariância:

(i) $-1 \leq C(X,Y) \leq 1$ ou $0 \leq C(X,Y) \leq 1$, onde no primeiro conjunto de limites, os valores 1 e -1 são atingidos em casos de associação extrema ou *perfeita*, respectivamente, *positiva* e *negativa*. O segundo conjunto de limites aplica-se àqueles índices que não distinguem associação positiva da negativa.

(ii) $C(X,Y) = C(Y,X)$ e $C(X,Y) = -C(-X,Y) = -C(X,-Y) = C(-X,-Y)$

(iii) Se X e Y são transformadas em novas variáveis $X' = f(X)$ e $Y' = g(Y)$ então $C(X',Y') = C(X,Y)$ para determinadas funções f e g. Em particular teremos: *invariância linear* se $f(X) = a_X + b_X X$ e $g(Y) = a_Y + b_Y Y$ com b_X e b_Y positivos; *invariância monotônica*, que é requerida se o índice deve ser adequado a dados ordinais, quando f e g são funções monotônicas crescentes.

Um outro aspecto a ser considerado, que aparece também em aplicações estatísticas, é o da necessidade de medirmos a *associação*, entre as variáveis X e Y, *controlada* por uma terceira variável Z. Essa medida é comumente denominada *asso*

ciação parcial (correlação parcial) e será denotada por C(X,Y|Z) onde C indica associação e | indicada controlada por. A título de exemplo podemos considerar o caso em que alguns pacientes são tratados de uma certa doença e outros não, e queremos medir a associação entre recuperação e tratamento controlada por uma terceira variável, digamos sexo dos pacientes. Um outro exemplo, agora real, é aquele em que Mosimann (1956) estudou correlações parciais entre diferentes escudos do plastrão de tartarugas, de uma determinada espécie, quando o comprimento total do plastrão era mantido constante.

Este trabalho tem por objetivo apresentar as medidas mais comuns de associação total e parcial, bem como, o seu emprego.

Nos capítulos seguintes estudaremos as principais medidas de associação total mas faremos agora, a título de apresentação, um breve resumo das mesmas:

Coeficiente de correlação de Pearson (ρ) introduzido por Pearson (1896) e definido como

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var } X}\ \sqrt{\text{var } Y}},$$

é conhecido também como "coeficiente de correlação momento-produto" e exige que as variáveis sejam medidas em escala, pelo menos, intervalar.

Coeficiente de correlação de postos de Spearman (ρ_s), introduzido por Spearman (1904). É uma medida de associação que exige que as variáveis X e Y sejam medidas pelo menos em escala ordinal. Se $R(X_i)$ e $R(Y_i)$ são os postos do indivíduo i, $i = 1,2,\ldots,n$, de acordo com as variáveis X e Y, respectivamente, a definição de ρ_s será:

$$\rho_s(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}[R(X_i)-\overline{R(X)}][R(Y_i)-\overline{R(Y)}]}{\left\{\frac{1}{n}\sum_{i=1}^{n}[R(X_i)-\overline{R(X)}]^2\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}[R(Y_i)-\overline{R(Y)}]^2\right\}^{1/2}}$$

onde

$$\overline{R(X)} = \frac{1}{n}\sum_{i=1}^{n}R(X_i) \quad \text{e} \quad \overline{R(Y)} = \frac{1}{n}\sum_{i=1}^{n}R(Y_i).$$

Coeficientes de associação ordinal (τ_a, γ e τ_b), que, da mesma forma que ρ_s, exigem as variáveis X e Y medidas pelo menos em escala ordinal. Se (X_1,X_2) e (Y_1,Y_2) são duas observações sobre (X,Y) nós diremos que elas são: *concordantes* se $X_1 > X_2$ e $Y_1 > Y_2$ ou $X_1 < X_2$ e $Y_1 < Y_2$, *discordantes* se $X_1 > X_2$ e $Y_1 < Y_2$ ou $X_1 < X_2$ e $Y_1 > Y_2$, *empatadas* se $X_1 = X_2$ e/ou $Y_1 = Y_2$. Considerando p_C, p_D e p_T como as probabilidades de que um par de observações, escolhido aleatoriamente, seja, respectivamente, concordante, discordante e empatado, as definições dos coeficientes τ_a, γ e τ_b são dadas por:

a) $\tau_a(X,Y) = p_C - p_D$ que é conhecido como *índice de associação ordinal "tau-a" de Kendall* ou ainda *índice de associação ordinal incondicional*. Esse coeficiente foi, segundo Kendall (1955), con

siderado por Greiner (1909) e Esscher (1924) como um método para estimar correlações em uma população normal mas foi redescoberto por Kendall (1938) que considerou-o puramente como um coeficiente de postos; Kruskal (1958) acredita que o precursor desse coeficiente foi Fechner em 1897.

b) $\gamma(X,Y) = \frac{p_C - p_D}{p_C + p_D}$, $p_T \neq 1$ denominado *Índice de associação "gama" de Goodman e Kruskal* mas conhecido também como *Índice condicional de associação ordinal*. Foi proposto por Goodman e Kruskal (1954).

c) $\tau_b(X,Y) = \frac{p_C - p_D}{\sqrt{1-p_X}\ \sqrt{1-p_Y}}$, $p_X \neq 1$ e $p_Y \neq 1$, onde p_X (p_Y) é a probabilidade de que duas observações aleatórias tenham o mesmo valor de X(Y). É o *coeficiente de associação "tau-b" de Kendall* muito conhecido pelo nome de "coeficiente de correlação de postos de Kendall" ou ainda "coeficiente de correlação sinal-diferença". Foi introduzido por Kendall (1945).

É interessante notar que se $p_T = 0$ teremos $\tau_a = \tau_b = \gamma$.

Coeficiente de associação δ-generalizado ($\bar{\delta}$). Recentemente, Agresti (1978) apresentou algumas medidas para descrever o grau de diferença entre dois grupos e entre vários grupos sobre uma variável resposta categórica ordinal. Dentre essas medidas destacaremos δ-generalizado por ser adequado para medir associação entre uma variável nominal X e outra variável ordinal Y. Sejam Y_I e Y_J observações sobre a variável resposta para membros dos grupos I e J selecionados, ao acaso e

independentemente, de acordo com a distribuição marginal da variável nominal. O coeficiente $\bar{\delta}$ é definido como o valor esperado de $|P(Y_I>Y_J) - P(Y_I<Y_J)|$ com relação a (I,J), isto é,

$$\bar{\delta} = \underset{(I,J)}{E} \left| P[Y_I>Y_J \mid (I,J)] - P[Y_I<Y_J \mid (I,J)] \right|$$

Índices do tipo redução proporcional em risco (RPR).

Neste item não apresentaremos nenhum novo índice de associação total; apenas mostraremos que alguns dos índices estudados podem ser interpretados segundo este conceito.

Os índices do tipo RPR são medidas que não encaram a correlação como algo que descreve a população. Nestes casos ela é vista, um tanto operacionalmente, como medindo o valor de se conhecer alguma coisa sobre uma das variáveis quando necessitamos saber algo sobre a outra.

Vamos supor que desejamos predizer a componente Y de uma observação aleatória do par (X,Y) e que sofreremos uma perda $L(Y_1,Y_2)$, não negativa, se nossa predição for Y_1 quando o verdadeiro valor da variável é Y_2. Consideremos duas situações diferentes: (1) não temos informação alguma antes de predizer Y, (2) conhecemos o valor de X antes da predição. Se R_1 e R_2 são as esperanças da perda, ou *riscos*, nas duas situações, definimos

$$RPR(X,Y) = 1 - \frac{R_2}{R_1}$$

que é o índice adequado para medir o valor de conhecer X antes de predizer Y e é denominado "índice de redução proporcional em risco" da situação 2 quando comparada com a situação 1.

Este índice não exige que X e Y sejam nem mesmo ordinais e da maneira como foi definido não é simétrico. É possível redefini-lo de modo a satisfazer essa propriedade conforme veremos mais tarde. Dessa maneira geral quem primeiro formalizou esse conceito de associação foi Goodman e Kruskal (1954).

A seguir apresentaremos também, de modo resumido, os mais conhecidos coeficientes de correlação e associação parcial que serão tratados nos capítulos seguintes.

Dependendo da forma de *controlar pela variável* Z diferentes medidas de associação parcial são definidas.

Coeficiente de correlação parcial "momento-produto" de Pearson $\rho(X,Y|Z)$. Neste caso o conceito de controle considerado é "*ajustando pela* Z". Vamos supor que f e g são funções adequadas para predizer X e Y à pratir de Z, respectivamente, de modo que os *resíduos* $X' = X-f(Z)$ e $Y' = Y-g(Z)$ são concentrados em torno de zero. É então definido o coeficiente de correlação parcial como $C(X,Y|Z) = C(X',Y')$, isto é, a correlação total entre os resíduos. Em particular se o critério de concentração é variância, então f e g são as funções de regres

são, isto é, as médias condicionais de X e Y dada Z, neste caso se o coeficiente de correlação momento-produto é usado, nós obtemos o coeficiente de *correlação parcial momento-produto* $\rho(X,Y|Z)$. Se além disso, as funções de regressão são lineares em Z e as variâncias condicionais não dependem de Z, então o mesmo resultado pode também ser obtido diretamente à partir da familiar fórmula de correlação parcial

$$\rho(X,Y|Z) = \frac{\rho(X,Y)-\rho(X,Z)\rho(Y,Z)}{\sqrt{1-\rho^2(X,Z)}\ \sqrt{1-\rho^2(Y,Z)}}$$

que muitas vezes é utilizada como definição de correlação parcial.

Coeficiente de associação parcial de Davis [$\delta(X,Y|Z)$]. Proposto por Davis (1967) é baseado no importante conceito de controle "*mantendo Z constante*" e no coeficiente de associação ordinal γ de Goodman e Kruskal. Davis considerou o caso em que X, Y e Z são variáveis categóricas de modo que a população pode ser representada por uma tabela de contingência de 3-entradas. Vamos supor que p_i é a probabilidade de que uma observação casual possua o i-ésimo valor de Z e que p_{C_i} (p_{D_i}, p_{T_i}) seja a probabilidade de que um par casual de observações seja empatado em Z, em seu i-ésimo valor, e também concordante (discordante, empatado) com relação a X e Y de modo que $p_{C_i}+p_{D_i}+p_{T_i} = p_i^2$. Nessas condições Davis definiu o seu índice de correlação parcial como

$$\gamma(X,Y|Z) = \frac{\sum_i p_{C_i} - \sum_i p_{D_i}}{\sum_i p_{C_i} + \sum_i p_{D_i}}$$

Coeficiente de correlação parcial de Kendall [$\phi(X,Y|Z)$]. Foi proposto por Kendall (1942) e na sua definição exige-se que X, Y e Z sejam medidas pelo menos em escala ordinal. Vamos supor, para simplificar, que, para o momento, empates são impossíveis. Classificaremos os pares de observações casuais, por exemplo (X_1,Y_1,Z_1) e (X_2,Y_2,Z_2), quanto ao fato de mostrarem X e Y concordantes ou discordantes com Z e arranjaremos as probabilidades desses eventos em uma tabela 2×2:

		X e Z	
		CONCORDANTES	DISCORDANTES
Y e Z	CONCORDANTES	p_0	p_X
	DISCORDANTES	p_Y	p_Z

em seguida definiremos o "coeficiente de correlação parcial de Kendall" por

$$\phi(X,Y|Z) = \frac{p_0 p_Z - p_X p_Y}{\sqrt{(p_0+p_X)(p_Y+p_Z)(p_0+p_Y)(p_X+p_Z)}}$$

Coeficiente de associação parcial baseado em pareamento [$\theta(X,Y|Z)$]. Foi introduzido por Quade (1971). Este coeficiente exige que X e Y sejam pelo menos ordinais porém Z é uma variável sem

qualquer restrição quanto à escala de medição podendo ser nominal e/ou multivariada. Baseia-se em associação em termos de concordância e discordância de pares de observações e no conceito de controle *"mantendo Z constante"* ou, mais precisamente, em termos da noção de *pareamento*.

De modo geral é estabelecida uma regra pela qual possa ser decidido se duas observações são *pareadas* sem ambiguidades.

Chamaremos de PAREAMENTO o evento em que um par casual de observações é *pareado* e de C (D) o evento em que o par é concordante (discordante) com relação a X e Y. Assumiremos que P{PAREAMENTO} > 0. Assim definimos o *índice de associação pareada"*

$$\theta(X,Y|Z) = P\{C|\text{PAREAMENTO}\} - P\{D|\text{PAREAMENTO}\}$$

Coeficientes de correlação parcial como extensão do conceito de redução proporcional em risco. Do mesmo modo que para correlação total iremos mostrar que alguns dos índices de associação parcial podem ser interpretados segundo este conceito.

De modo geral, será feita uma afirmação sobre Y, sujeita a perdas especificadas em caso de erro, em cada uma das situações: (1) temos informação de Z, mas não de X, antes da predição; (2) temos informação sobre Z e X antes da predição. Desse modo a redução proporcional em perda esperada (risco) para a situação 2 quando comparada com a situação 1 pode ser tomada como uma medida de correlação parcial.

CAPÍTULO 2

MEDIDAS DE ASSOCIAÇÃO TOTAL

No capítulo 1 vimos as definições e denominações mais comuns das medidas mais utilizadas de associação total. Houve oportunidade, também, de dar algumas informações de interesse puramente histórico. No presente capítulo estudaremos essas medidas fazendo considerações à respeito das possíveis interpretações que elas possam ter, apresentando os estimadores amostrais correspondentes e, sempre que possível, a distribuição assintótica dos mesmos.

2.1 - COEFICIENTE DE CORRELAÇÃO MOMENTO-PRODUTO DE PEARSON (ρ)

Definimos no capítulo 1, como sendo

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var } X}\ \sqrt{\text{var } Y}}$$

Como consequência direta da desigualdade de Cauchy-Schwarz conclui-se que $0 \leq \rho^2 \leq 1$ e, portanto, que

$$-1 \leq \rho \leq 1$$

Vamos supor que as variáveis aleatórias X e Y têm distribuição normal bi-variada. Nessas condições $\rho(X,Y)$ e o

modelo de regressão linear de uma variável sobre a outra, relacionam-se de modo muito interessante.

Sejam μ_X e μ_Y as médias populacionais e σ_X^2 e σ_Y^2 as variâncias populacionais de X e Y. Vamos denotar a covariância de X e Y por σ_{XY}. Assim podemos descrever

$$\rho(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Para um dado valor x de X existe uma subpopulação de valores de Y correspondendo a X = x. Sua distribuição, a distribuição condicional de Y dado X = x, é normal uni-variada com média

$$\mu_{Y \cdot X} = \mu_Y + \frac{\sigma_{XY}}{\sigma^2} (x - \mu_X)$$

denominada *esperança condicional de Y dado X=x* (ou então regressão de Y sobre X). A variância dessa distribuição, denominada *variância condicional de Y dado X=x* é

$$\sigma^2 = \sigma_Y^2 (1 - \rho^2)$$

e portanto podemos escrever

$$\sigma^2 = \frac{\sigma_Y^2 - \sigma^2}{\sigma_Y^2}$$

o que permite interpretarmos o quadrado de $\rho(X,Y)$ como sendo a proporção da variância de Y que é explicada pela regressão de Y sobre X.

Se definirmos $e = Y - \mu_{X \cdot Y}$ (isto é, e mede o desvio de Y à partir de sua média em $X = x$) a distribuição condicional de e dado $X = x$ é normal com média 0 e variância σ^2. Podemos então escrever

$$Y = \mu_{Y \cdot X} + e = \mu_Y - \frac{\sigma_{XY}}{\sigma_X^2} \mu_X + \frac{\sigma_{XY}}{\sigma_X^2} x + e = \beta_0 + \beta_1 x + e$$

onde

$$\beta_0 = \mu_Y - \beta_1 \mu_X, \quad \beta_1 = \frac{\sigma_{XY}}{\sigma_X^2} \quad e \quad e \sim N(0, \sigma^2)$$

Notemos que este é justamente o modelo de regressão linear simples de Y sobre X.

Então podemos dizer que $\rho^2(X,Y)$ é a proporção da variância de Y "explicada" pela regressão linear de Y sobre X. Quando $\rho = 0$ nós concluímos que $\sigma^2 = \sigma_Y^2$, o que significa que nada da variância de Y é explicada pela regressão de Y sobre X. Quando $\rho = \pm 1$, então $\sigma^2 = 0$, o que implica que toda a variância de Y é explicada pela regressão de Y sobre X, isto é, a relação entre X e Y é perfeitamente linear.

Consideremos agora o estimador amostral (r) de ρ.

Se $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ é uma amostra casual, de tamanho n, da população o "coeficiente de correlação amostral momento-produto", $r(X,Y)$, pode ser definido como

$$r(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n\right]\left[\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2/n\right]}}$$

e do mesmo modo que para ρ temos

$$-1 \leq r \leq 1$$

A variância de r para grandes amostras é uma expressão que envolve todos os momentos de segunda e quarta ordem da população amostrada. No caso normal, isto é, (X,Y) tem distribuição normal bi-variada, a expressão simplifica-se para

$$\text{var } r = (1-\rho^2)^2/n$$

embora seja de pouco valor prático, pois a distribuição de r tende muito lentamente para a normal e seria necessário utilizar $n \geq 500$.

O problema foi resolvido por Fisher (1921) que propos a seguinte transformação para r:

$$v = \frac{1}{2} \log_e\left(\frac{1+r}{1-r}\right)$$

que, para amostras normais, é muito mais próxima da distribuição normal com média aproximada

$$\mu_v = E(v) \cong \frac{1}{2} \log_e\left(\frac{1+\rho}{1-\rho}\right)$$

e variância aproximada

$$\sigma_v^2 = \text{var } r \cong \frac{1}{n-3}.$$

independente de ρ.

Vamos supor que queremos testar H_0: $\rho = \rho_0$ onde ρ_0 é uma constante $\neq \pm 1$.

Fazemos a transformação de r para v que, sob a hipótese nula, é aproximadamente normal com média

$$\mu_v = \frac{1}{2} \log_e \left(\frac{1+\rho_0}{1-\rho_0} \right)$$

e variância $\sigma_v^2 = \frac{1}{n-3}$.

Computamos $z = \frac{v-\mu_v}{\sigma_v}$ que, sob H_0 é aproximadamente N(0,1) quando n é grande. O nível crítico observado (P-value) depende da hipótese alternativa e H_0 é rejeitada se $P < \alpha$.

Um intervalo de confiança $100(1-\alpha)\%$ para μ_v é (v_1, v_2) onde $v_1 = v - \sigma_v Z_{1-(\alpha/2)}$ e $v_2 = v + \sigma_v Z_{1-(\alpha/2)}$.

Usando a transformação inversa

$$r = \frac{e^{2v}-1}{e^{2v}+1},$$

obtemos os limites de confiança para ρ.

O intervalo de confiança pode também ser usado para testar H_0: $\rho = \rho_0$ contra H_1: $\rho \neq \rho_0$ simplesmente rejeitan-

do H_0 ao nível α se o intervalo exclui ρ_0.

Existem tabelas para as transformações de v para r e de r para v. Do mesmo modo existem cartas ("charts") que permitem obter os intervalos de confiança graficamente (Pearson e Hartley, 1966).

Quando os testes de significância ou limites de confiança são críticos e uma melhor aproximação é desejada, Hotelling (1953) recomenda uma outra transformação (v*) que é obtida por

$$v^* = v - \frac{3v+r}{4n}$$

a sua variância $\sigma^2_{v^*} = \frac{1}{n-1}$. Sokal e Rohlf (1969) fazem recomendações acerca das aproximações no que diz respeito ao tamanho (n) da amostra: para $n \geq 50$ a aproximação de Fisher é adequada (sendo tolerável para $n \geq 25$). A aproximação sugerida por Hotteling é razoavelmente sastifatória para $n \geq 10$.

A variância de r para grandes amostras é encontrada em Kendall e Stuart (1977).

Kendall e Stuart (1973) recomendam que, em trabalhos práticos, só devemos usar ρ, como medida de associação, em casos de variação normal ou quase normal. Nesse caso podemos verificar facilmente que $\rho = \frac{\beta_1 \sigma_X}{\sigma_Y}$ e teremos $\rho = 0$ se e somente se $\beta_1 = 0$. Podemos testar H_0: $\rho = 0$ usando, então, o teste "t" comumente utilizado para testar H_0: $\beta_1 = 0$. Esse tes-

te pode ser posto na forma

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

onde t_0 tem distribuição "t" de Student com n-2 graus de liberdade.

2.2 - COEFICIENTE DE CORRELAÇÃO DE POSTOS DE SPEARMAN (ρ_s)

Nós vimos que o "coeficiente de correlação momento-produto" de Pearson é adequado para variáveis métricas X e Y, no entanto existem situações em que desejamos medir associação entre variáveis que não podem ser medidas de maneira objetiva, isto é, não podem ser expressas por números de unidades de um tipo objetivo. Tomemos o exemplo de Yule e Kendall (1940) onde se deseja medir associação entre "habilidade" matemática e musical em uma classe de estudantes. Uma maneira muito comum de medir habilidade seria atribuir notas aos estudantes mas uma objeção importante é que diferentes examinadores poderiam atribuir diferentes notas a um mesmo aluno. Uma correlação entre notas obtidas em matemática e música dependeria, de certo modo, do examinador e não refletiria acuradamente a relação entre as duas qualidades.

Uma maneira de contornar a situação seria arranjar os estudantes "em ordem" de habilidade sem tentar medi-la numericamente. Embora houvesse, ainda, divergência de opinião

entre os examinadores ela não seria tão freqüente como no caso anterior. Atribuiríamos então a cada estudante um número que indicasse a sua posição, entre os demais, para a habilidade em questão. Os alunos estariam assim *"ordenados"* e o número correspondente a cada aluno seria o seu *"posto"*.

No capítulo 1 nós definimos o "coeficiente de correlação de postos de Spearman" (ρ_s) como

$$\rho_s = \frac{\frac{1}{n}\sum_{i=1}^{n}[R(X_i)-\overline{R(X)}][R(Y_i)-\overline{R(Y)}]}{\left\{\frac{1}{n}\sum_{i=1}^{n}[R(X_i)-\overline{R(X)}]^2\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}[R(Y_i)-\overline{R(Y)}]^2\right\}^{1/2}}$$

onde X_i e Y_i representam as medidas, em escala pelo menos ordinal, do indivíduo i (i=1,2,...,n) com relação às variáveis X e Y; $R(X_i)$ e $R(Y_i)$ representam os postos do indivíduo i para X e Y, respectivamente; $\overline{R(X)}$ e $\overline{R(Y)}$ são as médias

$$\frac{1}{n}\sum_{i=1}^{n}R(X_i) \text{ e } \frac{1}{n}\sum_{i=1}^{n}R(Y_i).$$

Notemos que ρ_s é calculado pela mesma expressão que ρ, isto é, é o coeficiente de correlação de Pearson calculado para os postos $R(X_i)$ e $R(Y_i)$. Daí o seu nome de "coeficiente de correlação de postos".

Uma forma conveniente para o cálculo de ρ_s é obtida quando fazemos a suposição de que *não existem empates* entre os X_i's e/ou Y_i's.

Nesse caso:

$$\overline{R(X)} = \frac{1}{n}\sum_{i=1}^{n} R(X_i) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \frac{n+1}{2}$$

e analogamente $\overline{R(Y)} = \frac{n+1}{2}$. Temos também que

$$\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2 = \sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2 = \sum_{i=1}^{n} [i^2 - i(n+1) + \frac{(n+1)^2}{4}] =$$

$$= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4} =$$

$$= \frac{n(n^2-1)}{12}$$

e analogamente

$$\sum_{i=1}^{n} [R(Y_i) - \overline{R(Y)}]^2 = \frac{n(n^2-1)}{2}$$

Portanto temos para ρ_s

$$\rho_s = \frac{\sum_{i=1}^{n} [R(X_i) - \frac{n+1}{2}][R(Y_i) - \frac{n+1}{2}]}{n(n^2-1)/12}$$

ou ainda

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} [R(X_i) - R(Y_i)]^2}{n(n^2-1)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3-n} \therefore \rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3-n},$$

onde $d_i = [R(X_i) - R(Y_i)]$, que por ser uma forma adequada para o cálculo de ρ_s, tem sido muitas vezes utilizada como de

finição deste coeficiente.

No caso de dois ou mais indivíduos receberem o mesmo "posto" na mesma variável atribuímos a cada indivíduo a média dos postos que lhes seriam atribuídos caso não houvesse empate.

A proporção de empates sendo pequena, o efeito sobre o valor do coeficiente será desprezível mas, se for grande, um fator de correção deverá ser considerado no cálculo do mesmo.

Foi visto, à pouco, que $\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2 = \frac{n^3-n}{12}$ quando não há empates na variável X. Havendo empates o efeito será uma redução no valor de $\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2$. Torna-se, então, necessário corrigir a soma de quadrados e o fator de correção é

$$T_X = \frac{t^3-t}{12}$$

onde t é o número de observações empatadas em um posto. Para empates em vários postos consideraremos $\sum T_X$, onde o somatório leva em conta os vários valores de T_X para todos os grupos de observações empatadas e $\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2$ será agora, com a devida correção, igual a

$$\frac{n^3-n}{12} - \sum T_X$$

e, analogamente, para $\sum_{i=1}^{n} [R(Y_i) - \overline{R(Y)}]^2$ teremos

$$\frac{n^3-n}{12} - \sum T_Y$$

Se existir uma grande quantidade de empates devemos usar a fórmula de definição de ρ_s, mas existe outra que facilita os cálculos quando usamos essas correções. Essa fórmula é

$$\rho_s = \frac{\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2 + \sum_{i=1}^{n} [R(Y_i) - \overline{R(Y)}]^2 - \sum_{i=1}^{n} d_i^2}{2\sqrt{\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2 \sum_{i=1}^{n} [R(Y_i) - \overline{R(Y)}]^2}}$$

onde

$$\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2 = \frac{n^3-n}{12} - \sum T_X$$

e

$$\sum_{i=1}^{n} [R(Y_i) - \overline{R(Y)}]^2 = \frac{n^3-n}{12} - \sum T_Y$$

Vamos considerar agora o problema da estimação de ρ_s por meio de uma amostra casual, de tamanho n, da população. Denominemos r_s o estimador de ρ_s.

O estimador óbvio em caso de não haver (ou haver poucos) empates será

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

ou então, para grande quantidade de empates:

$$\rho_s = \frac{\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2 + \sum_{i=1}^{n} [R(Y_i) - \overline{R(Y)}]^2 - \sum_{i=1}^{n} d_i^2}{2\sqrt{\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2 \sum_{i=1}^{n} [R(Y_i) - \overline{R(Y)}]^2}}$$

onde os elementos que aparecem nas duas formas têm o mesmo significado que possuiam anteriormente só que agora referem-se aos indivíduos da amostra. Em outras palavras $R(X_i)$ é o posto em relação a x_i, valor amostral observado da variável aleatória X_i; analogamente para $R(Y_i)$.

Sob a hipótese nula de independência pode-se chegar à distribuição amostral exata de r_s para cada valor de n, porém torna-se muito difícil, do ponto de vista prático, obter essas distribuições para valores muito grandes de n. O procedimento é o seguinte:

Extraindo uma amostra casual de postos, de X e de Y, da população em estudo teremos, para uma dada ordenação de Y, igual probabilidade para qualquer ordenação de X, sob H_0. O mesmo acontecerá se invertermos os papéis de X e Y. Para n indivíduos serão n! ordenações possíveis, em relação à variável X, que poderão ocorrer com qualquer ordenação em relação a Y. Como essas n! ordenações são equiprováveis teremos o valor $\frac{1}{n!}$ para a probabilidade de ocorrência de determinada ordenação dos valores de X em conjunto com dada ordenação dos valores de Y. A cada ordenação de Y associa-se um valor de r_s. A probabilidade de ocorrência, sob a hipótese

nula, de qualquer valor particular de r_s é, portanto, proporcional ao número de permutações que dão origem àquele valor. Ao calcularmos r_s para n = 2 teremos apenas dois valores possíveis, +1 e -1, cada um com probabilidade 1/2 de ocorrência sob H_0. Para n = 3 os valores possíveis são -1, $-\frac{1}{2}$, 1/2 e 1 com probabilidades, respectivamente, 1/6, 1/3, 1/3, e 1/6 sob H_0.

Siegel (1975) apresenta uma tabela que dá os valores críticos de r_s com probabilidades associadas p = 0,05 e p = 0,01, sob H_0: $\rho_s = 0$, para n de 4 a 30. (para testes unilaterais).

Agora, para grandes amostras ($n \geq 10$) o teste de significância de um valor de r_s, obtido sob hipótese de independência, é realizado por meio de

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

onde t tem distribuição t de Student com n-2 graus de liberdade.

2.3 - COEFICIENTES DE ASSOCIAÇÃO ORDINAL (τ_a, γ e τ_b)

Na apresentação sumária destes coeficientes, que fizemos no capítulo anterior, definimos os pares de observações aleatórias (sobre duas variáveis X e Y medidas em escala pelo menos ordinal) como sendo *concordantes*, *discordantes* ou

empatados de acordo com as várias possibilidades de combinar as ordenações, de X e Y nos mesmos. Se X e Y são variáveis métricas podemos colocar essas definições em outros termos: serão *concordantes* (*discordantes*) se estiverem numa linha de inclinação positiva (negativa) e *empatados* se estiverem numa linha horizontal, ou vertical, ou se forem coincidentes. Ainda de um outro modo podemos dizer que são *concordantes*, *discordantes* ou *empatados* se o produto $(X_1-X_2)(Y_1-Y_2)$ for, respectivamente, positivo, negativo ou nulo.

As probabilidades de que um par aleatório seja concordante, discordante, empatado, empatado em X, empatado em Y foram respectivamente denotadas por p_C, p_D, p_T, p_X e p_Y, donde concluimos que

$$p_C + p_D + p_T = 1 \quad \text{e} \quad p_T = p_X + p_Y - p_{XY}$$

onde p_{XY} denota a probabilidade do par aleatório ser empatado em X e em Y.

Vamos estudar mais detalhadamente os coeficientes τ_a, γ e τ_b apresentando as possíveis interpretações, a relação existente entre seus valores, os estimadores amostrais e considerações sobre testes de significância dos mesmos baseados nas distribuições assintóticas.

O *coeficiente de associação ordinal* τ_a *de Kendall*, foi definido como

$$\tau_a(X,Y) = p_C - p_D$$

e pode ser interpretado diretamente como sendo a diferença entre a probabilidade de um par aleatório ser concordante e a probabilidade do mesmo ser discordante. Teremos associação perfeita positiva, $\tau_a = 1$, se pares aleatórios forem concordantes com probabilidade um, isto é, $p_C = 1$, e associação perfeita negativa $\tau_a = -1$, se os pares forem discordantes com probabilidade um, isto é, $p_D = 1$. Esses seriam os casos em que a população está sobre alguma curva crescente (1º caso) ou decrescente (2º caso) monotonicamente.

É interessante notarmos que se $p_T \neq 0$, τ_a não atingirá os limites 1 e -1 e temos

$$|\tau_a| \leq 1 - p_T$$

O *coeficiente de associação ordinal* γ *de Goodman e Kruskal* cuja definição foi dada por

$$\gamma(X,Y) = \frac{p_C - p_D}{p_C + p_D}, \qquad p_T \neq 1$$

pode ser interpretado como a diferença entre as probabilidades condicionais, $p_C/(p_C+p_D)$ e $p_D/(p_C+p_D)$, de que duas observações aleatórias sejam concordantes e discordantes dado que não são empatadas. Essa é a razão pela qual é também conhecido como "coeficiente de associação ordinal condicional".

Notemos que $-1 \leq \gamma \leq 1$ e que γ atingirá os limites 1 ou -1 sempre que $p_D = 0$ ou $p_C = 0$ independentemente da existência de empates.

O *coeficiente de associação ordinal* τ_b *de Kendall* que é dado por

$$\tau_b(X,Y) = \frac{p_C - p_D}{\sqrt{1-p_X}\ \sqrt{1-p_Y}}, \qquad p_X \neq 1 \text{ e } p_Y \neq 1$$

além de não ter interpretação fácil é bem mais difícil de manusear e calcular do que τ_a e γ.

É fácil mostrar que $0 \leq |\tau_a| \leq |\tau_b| \leq |\gamma| \leq 1$; para isso basta lembrar que os numeradores, nas definições desses coeficientes, são todos iguais e mostrar que os denominadores apresentam a relação

$$0 \leq |p_C + p_D| \leq |\sqrt{(1-p_X)(1-p_Y)}| \leq 1$$

o que é um fato pois é óbvio que

$$0 \leq p_X \leq p_T \leq 1 \quad \text{e} \quad 0 \leq p_Y \leq p_T \leq 1;$$

isso acarreta

$$0 \leq 1-p_T \leq 1-p_X \leq 1 \quad \text{e} \quad 0 \leq 1-p_T \leq 1-p_Y \leq 1$$

logo

$$(1-p_T)^2 \leq (1-p_X)(1-p_Y) \leq 1 \quad \therefore \quad |1-p_T| \leq |\sqrt{(1-p_X)(1-p_Y)}| \leq 1$$

e como

$$0 \leq p_C + p_D = 1 - p_T \text{ temos } 0 \leq |p_C+p_D| \leq |\sqrt{(1-p_X)(1-pY)}| \leq 1$$

e então

$$0 \leq \frac{|p_C - p_D|}{|p_C + p_D|} \leq \frac{|p_C - p_D|}{|\sqrt{(1-p_X)(1-p_Y)}|} \leq |p_C - p_D| \leq 1$$

$$0 \leq |\tau_a| \leq |\tau_b| \leq |\gamma| \leq 1$$

Essa relação nos leva à conclusão de que

$$0 \leq |\tau_b| \leq 1$$

e portanto $-1 \leq \tau_b \leq 1$ e teremos

$$\tau_b = 1 \text{ quando } p_C = 1$$

$$\tau_b = -1 \text{ quando } p_D = 1$$

Vamos supor, agora, que $(X_1,Y_1), (X_2,Y_2), \ldots, (X_n,Y_n)$ é uma amostra casual da variável aleatória bidimensional (X,Y). Dessas n observações é possível formar $N = \frac{n(n-1)}{2}$ pares de observações.

Sejam, então, para esses N pares de observações: N_C N_D, N_T, N_X e N_Y, respectivamente, o número dos que são concordantes, discordantes, empatados, empatados em X, empatados em Y.

Então $N_C + N_D + N_T = N$.

Os estimadores óbvios de τ_a, γ e τ_b serão denominados por t_a, G e t_b e definidos como

$$t_a(X,Y) = \frac{N_C - N_D}{N} = \frac{2(N_C - N_D)}{n(n-1)}$$

$$G(X,Y) = \frac{N_C - N_D}{N_C + N_D}, \qquad N_T \neq N$$

$$t_b(X,Y) = \frac{N_C - N_D}{\sqrt{(N-N_X)(N-N_Y)}}, \qquad N_X \neq N \text{ e } N_Y \neq N$$

onde as estimativas de p_C, p_D, p_T, p_X e p_Y são, respectivamente, $\frac{N_C}{N}$, $\frac{N_D}{N}$, $\frac{N_T}{N}$, $\frac{N_X}{N}$ e $\frac{N_Y}{N}$.

Da mesma forma que para os coeficientes populacionais pode-se mostrar que

$$0 \leq |t_a| \leq |t_b| \leq |G| \leq 1$$

Vamos ver um exemplo simples de cálculo dos três coeficientes (Quade, 1966).

Consideremos as alturas e pesos de uma amostra de 4 homens como indicado na tabela

HOMEM	1º	2º	3º	4º
ALTURA (pol)	69	72	71	69
PESO (lb)	180	196	167	148

O primeiro homem é *concordante* com o segundo que é maior e mais pesado, é *discordante* com o terceiro que é maior mas de menor peso, e é *empatado* com o quarto que tem a mesma altura. Os três pares de homens que restam, (2º,3º), (2º,4º) e (3º,4º) são todos *concordantes*.

Temos então

$$N_C = 4$$

$$N_D = 1$$

$$N_T = 1$$

$$N = N_C + N_D + N_T = 6$$

$$N_X = 1$$

$$N_Y = 0$$

dai

$$t_a = \frac{N_C - N_D}{N} = \frac{4 - 1}{6} = \frac{3}{6} = 0,5$$

$$G = \frac{N_C - N_D}{N_C + N_D} = \frac{4 - 1}{4 + 1} = \frac{3}{5} = 0,6$$

$$t_b = \frac{N_C - N_D}{\sqrt{N-N_X}\ \sqrt{N-N_Y}} = \frac{4 - 1}{\sqrt{(6-1)}\ \sqrt{(6-0)}} = \frac{3}{\sqrt{30}} = 0,5477.$$

Em seguida veremos os desvios-padrão assintóticos para t_a e G. Esses resultados encontram-se, por exemplo, em Quade (1971) onde são obtidos como casos especiais do "coeficente de associação pareada". A abordagem de Quade baseia se em estatísticas-U e a de Goodman e Kruskal (1963 e 1972) no "método delta".

Com base nesses resultados temos que

$$Z_{\tau_a} = \frac{t_a - \tau_a}{S_a} \quad \text{e} \quad Z_\gamma = \frac{G-\gamma}{S_G}$$

têm distribuições assintóticas normais com média zero e variância 1, onde

$$S_a = \frac{2\sqrt{n\Sigma(C_i-D_i)^2 - [\Sigma(C_i-D_i)]^2}}{n(n-1)\sqrt{n}}$$

e

$$S_G = \frac{4\sqrt{n\Sigma C_i^2(\Sigma D_i)^2 - 2\Sigma C_i \Sigma D_i \Sigma C_i D_i + (\Sigma C_i)^2 \Sigma D_i^2}}{(\Sigma C_i + \Sigma D_i)^2}$$

C_i (D_i) é o número de observações concordantes (discordantes) com a observação (X_i, Y_i), $i=1,2,\ldots,n$.

Vamos supor que Z^* seja o valor crítico da $N(0,1)$ tal que $P\{|Z| \leq Z^*\} = 1 - \alpha$. Então um intervalo de confiança central para τ_a, com coeficiente de confiança aproximado $(1-\alpha)$ é

$$t_a \pm Z^* S_a$$

Agora suponhamos que desejamos testar a hipótese de que o índice populacional τ_a é igual a um valor especificado τ_a'. Um teste bi-lateral de tamanho aproximado α rejeita a hipótese H_0: $\tau_a = \tau_a'$ se $|t_a - \tau_a'| \geq Z^* S_a$. O nível de significância (observado) "a posteriori" do valor t_a obtido na amostra para um teste bi-lateral é dado por:

$$P = P\left\{|Z| \geq \frac{|t_a - \tau_a'|}{S_a}\right\}$$

O que foi exemplificado para t_a vale para G e procedimentos unilaterais podem ser obtidos da maneira óbvia. Quade (1971) afirma que esses resultados aproximados deverão ser, provavelmente, suficientemente acurados para a maioria das situações práticas se o tamanho amostral n é ao me

nos 20, no caso de t_a e ao menos 50, no caso de G.

A hipótese de associação ordinal nula é a mesma independentemente do índice pelo qual é expressa, isto é

$$H_a: \tau_a = 0,\ H_d: \gamma = 0 \text{ e } H_b: \tau_b = 0$$

são equivalentes e valem se e somente se $p_C = p_D$ (assumindo que a probabilidade de um empate não é 1). Desse modo para testar a hipótese de associação ordinal nula nós podemos calcular simplesmente

$$\frac{t_a}{s_a} = \frac{\Sigma(C_i - D_i)\sqrt{n}}{2\sqrt{n\Sigma(C_i - D_i)^2 - [\Sigma(C_i - D_i)]^2}}$$

e rejeitá-la se este valor exceder Z*, o valor crítico ao nível α, da distribuição normal padrão.

Desse modo, nós testamos a hipótese de associação ordinal nula e a rejeição da mesma implica certamente que X e Y não são independentes. No entanto aceitar essa hipótese não implica em aceitar que X e Y sejam independentes pois sabemos que associação nula não implica necessariamente independência.

Podemos testar a hipótese de independência pois, do mesmo modo que para r_s, se não houver empates é possível obter, sob tal hipótese nula, a distribuição exata de t_a (e como já tivemos oportunidade de mostrar, de G e t_b também, pois os três são idênticos quando não há empates). Na realidade essa distribuição exata, para cada n, é mais facilmente ob-

tida do que para r_s pois existem fórmulas de recorrência. Valores críticos para testar independência, obtidos dessas distribuições amostrais exatas, são apresentados em Quade (1966) para n de 4 a 12.

Para grandes amostras ele recomenda (Quade, op.cit) em geral que os procedimentos que vimos anteriormente sejam seguidos e a hipótese de associação-ordinal nula seja testada em vez da hipótese de independência. No entanto apresenta a variância de t_a sob hipótese de independência de X e Y.

$$\sigma^2_{t_a} = \sigma^2 = \frac{2}{9n(n-1)} [2(n-2)(1-p_{XX})(1-p_{YY}) + 9(1-p_X)(1-p_Y)]$$

onde p_X e p_Y já foram definidos e p_{XX} (p_{YY}) é a probabilidade de que três observações aleatórias tenham o mesmo valor de X (Y).

Assumindo X e Y independentes e a amostra representada em uma tabela de 2-entradas um estimador $\hat{\sigma}^2$ não viciado de σ^2 pode ser obtido ao substituirmos as várias probabilidades pelas estimativas

$$\hat{p}_X = \frac{\sum A_i^2 - n}{n(n-1)} = \frac{N_X}{N}$$

$$\hat{p}_Y = \frac{\sum B_j^2 - n}{n(n-1)} = \frac{N_Y}{N}$$

$$\hat{p}_{XX} = \frac{\sum A_i^3 - 3\sum A_i^2 + 2n}{n(n-1)(n-2)}$$

$$\hat{p}_{YY} = \frac{\sum B_j^3 - 3\sum B_j^2 + 2n}{n(n-1)(n-2)}$$

onde os A's e os B's são os totais para linhas e colunas na tabela de contingência que representa a amostra.

Trataremos então $(t_a/\hat{\sigma})$ como N(0,1) para testar independência contra a alternativa $p_C \neq p_D$.

Atingiremos claramente um limite superior para σ^2 se fizermos p_X, p_Y, p_{XX} e p_{YY} iguais a zero e então

$$\sigma_U^2 = \frac{2(2n+5)}{9n(n-1)} = \frac{2n+5}{9N}$$

que será o valor exato de $\hat{\sigma}^2$ se não há empates na amostra. Então se tratarmos (t_a/σ_U) como N(0,1) teremos um teste conservativo que produzirá um nível de significância "a posteriori" (p_U) maior do que o verdadeiro nível p. Por outro lado, como $p_{XX} \leq p_X$ e $p_{YY} \leq p_Y$, é facilmente verificado que um limite inferior para σ^2 é

$$\sigma_L^2 = \frac{2(2n+5)(1-p_X)(1-p_Y)}{9n(n-1)}$$

e um estimador não viciado para σ_L^2 será

$$\hat{\sigma}_L^2 = \frac{2(2n+5)(N-N_X)(N-N_Y)}{9n(n-1)N^2} = \frac{(2n+5)(N-N_X)(N-N_Y)}{9N^3}$$

se a hipótese é verdadeira. Então se tratarmos $(t_a/\hat{\sigma}_L)$ como uma N(0,1), teremos um teste anti-conservativo que produzi-

rá um nível de significância "a posteriori" P_L menor do que o verdadeiro P. Isto é, $P_L \leq P \leq P_U$, e em particular se para qualquer α pré-especificado tivermos $P_U < \alpha$ segue-se que $P < \alpha$ também, ou se $P_L > \alpha$ segue-se que $P > \alpha$; portanto em um ou outro caso não haverá necessidade de efetuar os cálculos mais complicados que produziriam um P mais exato.

Quando os testes de independência vistos são aplicados no caso em que X e Y são normalmente distribuídos eles têm eficiência relativa assintótica de 91% em relação ao teste mais poderoso que no caso é aquele baseado no coeficiente de correlação de Pearson. Esta eficiência relativa assintótica vale igualmente para o teste baseado em r_s de Spearman (Quade, 1966).

2.4 - COEFICIENTE DE ASSOCIAÇÃO δ-GENERALIZADO ($\bar{\delta}$)

Lembremos, antes de tudo, que o uso desta medida exige que Y seja uma variável categórica ordinal; X poderá ser uma variável nominal que, freqüentemente, em trabalhos práticos, representará um conjunto de grupos, tais como, localidades, tipos religiosos, raças, etc...

Consideremos então r grupos para X e c níveis para a variável resposta ordinal Y.

Para a população objeto de estudo usaremos a seguinte notação:

ρ_{ij} é a probabilidade de um membro selecionado ao acaso

ser classificado no grupo i e nível de resposta j;

$\rho_i = \sum_{j=1}^{c} \rho_{ij}$;

$\tilde{\rho}_{ij} = \rho_{ij}/\rho_i$ é a probabilidade condicional de um membro selecionado ao acaso ser classificado no nível de resposta j dado que pertence ao i-ésimo grupo;

em todos os casos acima $1 \leq i \leq r$ e $1 \leq j \leq c$.

Sejam I e J variáveis aleatórias representando os índices de dois grupos distintos selecionados de acordo com a distribuição marginal $\{\rho_i\}$ da variável nominal e sejam também Y_I, Y_J observações sobre Y para membros desses grupos selecionados ao acaso e independentemente.

Feitas essas suposições $\bar{\delta}$ foi definido como

$$\bar{\delta} = \underset{(I,J)}{E} \left| P[Y_I > Y_J \mid (I,J)] - P[Y_I < Y_J \mid (I,J)] \right|$$

e escrevendo de outra forma temos

$$\bar{\delta} = \underset{(I,J)}{E} \left| \delta_{IJ} \right| = \sum_{i<j} \lambda_{ij} \left| \delta_{ij} \right|$$

onde

$$\lambda_{ij} = \rho_i \rho_j / \sum_{a<b} \rho_a \rho_b$$

e

$$\delta_{ij} = P(Y_i > Y_j) - P(Y_j < Y_i) = \sum_{k>\ell} \tilde{\rho}_{ik} \tilde{\rho}_{j\ell} - \sum_{k<\ell} \tilde{\rho}_{ik} \tilde{\rho}_{j\ell}$$

Convém notarmos que δ_{ij} é justamente o índice δ definido por Agresti, no mesmo trabalho, para medir o grau de diferença entre dois grupos i e j sobre a resposta categórica ordinal; dai a denominação δ-generalizado para $\bar{\delta}$.

Da última expressão para $\bar{\delta}$ podemos interpretá-lo como uma média ponderada dos valores absolutos de δ entre os possíveis pares de grupos. Quanto aos pesos λ_{ij} atribuídos a esses pares é possível interpretá-los como a probabilidade de dois grupos selecionados ao acaso (sem reposição, de acordo com $\{\rho_i\}$) produzirem o par (i,j).

Quando $r = 2$ temos $\bar{\delta} = |\delta|$.

Não é difícil concluir que $0 \leq \bar{\delta} \leq 1$ e que

$\bar{\delta} = 0$ se e somente se $\delta_{ij} = 0$, $\forall i,j$, $i < j$

$\bar{\delta} = 1$ se e somente se $|\delta_{ij}| = 1$, $\forall i,j$, $i < j$.

Vamos mostrar agora que $\bar{\delta}$ pode ser interpretado como uma diferença entre duas probabilidades:

Dizemos que o grupo i é *mais alto* do que o grupo j se $\delta_{ij} > 0$ e definimos um par de respostas (y_i, y_j) provenientes dos grupos i e j como possuindo *"ordem semelhante"* à daqueles grupos se o membro do grupo mais alto tem o maior posto, isto é, se $y_i - y_j$ tem o mesmo sinal que δ_{ij}; em outras palavras, o par tem "ordem semelhante" se o membro do par com maior posto médio (quando postos são computados usando somente aqueles grupos) tem posto real mais alto. Analogamen-

te definimos um par possuindo "*ordem não-semelhante*" à dos grupos se o membro do grupo mais alto tem o posto inferior. Sejam S e D os eventos em que um par escolhido ao acaso tenha "ordem semelhante" e "ordem não-semelhante" respectivamente, e seja U o evento em que os membros do par são classificados em grupos diferentes (isto é, eles são não-empatados na variável de classificação dos grupos). Vamos definir agora

$$G_i^+ = \{j: \delta_{ij}>0\},\ G_i^- = \{j: \delta_{ij}<0\}$$

$$R_{i\ell}^{(s)} = \sum_{j\in G_i^+} \sum_{k<\ell} \rho_{jk} + \sum_{j\in G_i^-} \sum_{k>\ell} \rho_{jk}$$

$$R_{i\ell}^{(d)} = \sum_{j\in G_i^+} \sum_{k>\ell} \rho_{jk} + \sum_{j\in G_i^-} \sum_{k<\ell} \rho_{jk}$$

Notemos que $R_{i\ell}^{(s)}$ ($R_{i\ell}^{(d)}$) é a probabilidade de um par de membros, um do grupo i apresentando resposta ℓ e o outro aleatório ter "ordem semelhante" ("ordem não-semelhante"). Segue-se que

$$P(S) = \sum_{i,\ell} \rho_{i\ell} R_{i\ell}^{(s)},\ P(D) = \sum_{i,\ell} \rho_{i\ell} R_{i\ell}^{(d)},\ P(U) = 2\sum_{i<j} \rho_i \rho_j \text{ e}$$

$$P(S|U) - P(D|U) = \sum_{i,\ell} \rho_{i\ell}(R_{i\ell}^{(s)} - R_{i\ell}^{(d)})/2\sum_{a<b} \rho_{ab} =$$

$$= \left\{\sum_i \sum_{j\in G_i^+} \left[\sum_{k<\ell} \rho_{i\ell}\rho_{jk} - \sum_{k>\ell} \rho_{i\ell}\rho_{jk}\right] + \sum_i \sum_{j\in G_i^-} \left[\sum_{k>\ell} \rho_{i\ell}\rho_{jk} - \sum_{k<\ell} \rho_{i\ell}\rho_{jk}\right]\right\} / 2\sum_{a<b} \rho_a\rho_b =$$

$$= \left\{\sum_{i<j} \left|\sum_{k<\ell} \rho_{i\ell}\rho_{jk} - \sum_{k>\ell} \rho_{i\ell}\rho_{jk}\right|\right\} / \sum_{a<b} \rho_a\rho_b = \sum_{i<j} \rho_i\rho_j |\delta_{ij}| / \sum_{a<b} \rho_a\rho_b = \bar{\delta},$$

isto é, $P(S|U) - P(D|U) = \bar{\delta}$ ou, em palavras, $\bar{\delta}$ é a diferença entre a probabilidade condicional de selecionar um par de membros com "ordem semelhante" e a probabilidade condicional de selecionar um par com "ordem não-semelhante", dado que os membros estão em grupos diferentes.

Vamos supor que exista uma indexação dos grupos tal que $\delta_{ij} \geq 0$ sempre que $i > j$; diremos então que os grupos são *"consistentemente ordenados"*. Em adição, suponhamos que a variável cujos níveis são os r grupos é também ordinal e tem os níveis ordenados naturalmente de acordo com uma indexação que produz uma ordenação consistente. Há então uma relação monotônica entre a variável resposta e os grupos. Nesse contexto, um par tendo "ordem semelhante" é simplesmente um par "concordante" e $\bar{\delta}$ torna-se uma medida de associação ordinal para uma tabela $r \times c$ que recebe o nome de "d de Somers", isto é, d é a diferença entre as probabilidades de um par aleatório de membros ser "concordante" e "discordante" dado que os membros são classificados em grupos diferentes.

Em seguida examinemos o que existe de teoria amostral para o coeficiente $\bar{\delta}$. Para isso vamos supor que temos amostras, de tamanho n_i, para cada um dos r grupos e sejam, então, n_{ij}, $1 \leq i \leq r$ e $1 \leq j \leq c$, as freqüências para cada uma das celas. Temos então, que

$$n = \sum_{i,j} n_{ij} = \sum_i n_i, \quad P_{ij} = \frac{n_{ij}}{n}, \quad P_i = \frac{n_i}{n}, \quad \tilde{P}_{ij} = \frac{n_{ij}}{n}$$

onde P_{ij}, P_i e $\tilde{P}_{ij}$ são análogos amostrais de ρ_{ij}, ρ_i e $\tilde{\rho}_{ij}$.

Sejam também, $\hat{\lambda}_{ij}$ e d_{ij} os análogos amostrais de λ_{ij} e δ_{ij}. Nessas condições definimos $\bar{d}$, o estimador amostral de $\bar{\delta}$, do seguinte modo

$$\bar{d} = \sum_{i<j} \hat{\lambda}_{ij} |d_{ij}|$$

onde

$$\hat{\lambda}_{ij} = P_i P_j \Big/ \sum_{a<b} P_a P_b = n_i n_j \Big/ \sum_{a<b} n_a n_b$$

e

$$d_{ij} = \sum_{k>\ell} \tilde{P}_{ik} \tilde{P}_{j\ell} - \sum_{k<\ell} \tilde{P}_{ik} \tilde{P}_{j\ell} = \frac{\sum_{k>\ell} n_{ik} n_{j\ell} - \sum_{k<\ell} n_{ik} n_{j\ell}}{n_i n_j}$$

Podemos então obter $\bar{d}$ em função das n_i e n_{ij}'s

$$\bar{d} = \frac{\sum_{i<j} \left| \sum_{k>\ell} n_{ik} n_{j\ell} - \sum_{k<\ell} n_{ik} n_{j\ell} \right|}{\sum_{a<b} n_a n_b}$$

Utilizando o mesmo "método delta" usado por Goodman e Kruskal na obtenção da variância assintótica do coeficiente γ, Agresti conclui que $\sqrt{n}(\bar{d}-\bar{\delta})/\hat{\sigma}_{\bar{d}}$ tem assintoticamente, distribuição N(0,1) onde

$$\hat{\sigma}_{\bar{d}} = \left\{ \sum_{i,\ell} P_{i\ell} [\bar{d}(1-P_i) - (\hat{R}_{i\ell}^{(s)} - \hat{R}_{i\ell}^{(d)})]^2 \right\} \Big/ \left(\sum_{a<b} P_a P_b \right)^2$$

sendo $\hat{R}_{i\ell}^{(s)}$ e $\hat{R}_{i\ell}^{(d)}$ os análogos amostrais de $R_{i\ell}^{(s)}$ e $R_{i\ell}^{(d)}$ definidos anteriormente, onde os ρ_{ij}'s foram substituídos pelos

P_{ij}'s. ($\hat{G}_i^+$ e $\hat{G}_i^-$ são os análogos amostrais para G_i^+ e G_i^-).

Apenas para exemplificar consideremos os dados da tabela abaixo baseados em um estudo efetuado em 1976 durante uma "primary" presidencial no Estado de Wisconsin, E.U.A. (Hedlund, 1978).

Cada pessoa, em uma amostra casual de 1083 eleitores da área metropolitana de Milwaukee, foi classificada, entre outros aspectos, quanto à afiliação partidária (Republicano, Democrata, Independente) e ideologia política (Conservador, moderado, liberal). Um dos objetivos do estudo era medir associação entre afiliação partidária e ideologia política.

AFILIAÇÃO PARTIDÁRIA	IDEOLOGIA POLÍTICA			TAMANHO DA AMOSTRA
	CONSERVADOR	MODERADO	LIBERAL	
Democrata (D)	0,251	0,391	0,358	399
Independente (I)	0,300	0,447	0,253	470
Republicano (R)	0,593	0,366	0,070	214

Foram obtidos os valores: $d_{D,I} = 0,110$; $d_{D,R} = 0,435$ e $d_{I,R} = 0,347$ utilizando $d_{ij} = \sum_{k>\ell} \tilde{P}_{ik}\tilde{P}_{i\ell} - \sum_{k<\ell} \tilde{P}_{ik}\tilde{P}_{j\ell}$ já que temos os $\tilde{P}_{ij}$'s. Nessa amostra existe uma ordenação consistente R < I < D para os grupos da variável nominal afiliação partidária.

Calculando $\bar{d} = \sum_{i<j} P_i P_j |d_{ij}| / \sum_{i<j} P_i P_j$ com $P_i = \frac{n_i}{n}$ foi obtido $\bar{d} = 0,248$.

Para o cálculo de $\hat{\sigma}_{\bar{d}}$ vamos identificar D, I e R com

os valores 1, 2 e 3 respectivamente

$$G_i^+ = \{j: j>i\},\ \hat{G}_i^- = \{j: j<i\}$$

e os valores de $\hat{R}_{i\ell}^{(s)}$ e $\hat{R}_{i\ell}^{(d)}$ são dados na tabela abaixo onde os valores de $\hat{R}_{i\ell}^{(d)}$ encontram-se entre parênteses.

AFILIAÇÃO PARTIDÁRIA	IDEOLOGIA POLÍTICA		
	CONSERVADOR	MODERADO	LIBERAL
DEMOCRATA (1)	0 (0,384)	0,247 (0,124)	0,507 (0)
INDEPENDENTE (2)	0,276 (0,080)	0,249 (0,106)	0,183 (0,236)
REPUBLICANO (3)	0,580 (0)	0,242 (0,222)	0 (0,560)

Por exemplo:

$\hat{R}_{22}^{(s)} = 0,132 + 0,117 = 0,249$ e $\hat{R}_{22}^{(d)} = 0,092 + 0,014 = 0,106$

$\sum_{a<b} P_a P_b = 0,3185$

temos então, usando a fórmula já vista para $\hat{\sigma}_{\bar{d}}$,

$$\hat{\sigma}_{\bar{d}}^2 = \frac{1}{(0,3185)^2}\{0,092[(0,248)(0,802)-(0-0,384)]^2 + \ldots +$$

$$+ 0,014[(0,248)(0,566) - (0-0,560)]^2\} = 0,790$$

logo $\sigma_{\bar{d}} = 0,889$, e como a distribuição assintótica de

$$\sqrt{n}(\bar{d}-\bar{\delta})/\hat{\sigma}_{\bar{d}} \text{ é } N(0,1)$$

um intervalo de confiança aproximado com coeficiente de confiança de 95% para $\bar{\delta}$ é $\bar{d} \pm 1,96\hat{\sigma}_{\bar{d}}/\sqrt{n}$ ou

$$0,248 \pm 1,96(0,889)/\sqrt{1083} \text{ ou } 0,248 \pm 0,053$$

2.5 - ÍNDICES DO TIPO REDUÇÃO PROPORCIONAL EM RISCO (RPR)

A fórma básica desses índices foi dada no capítulo 1 e é

$$RPR(X,Y) = 1 - \frac{R_2}{R_1}$$

onde R_1 e R_2 são as perdas esperadas, ou riscos, em duas situações diferentes para a predição de Y de um par aleatório (X,Y) retirado da população. Como R_2 é obviamente menor do que R_1 (porque na situação 2 contaremos com a informação sobre X antes da predição ao passo que na situação 1 isso não acontece) concluímos que $0 \leq RPR \leq 1$. O valor 1 será alcançado quando o conhecimento de X reduz o risco a zero, isto é $R_2 = 0$. Teremos $RPR(X,Y) = 0$ quando, conhecer X, não tem valor nenhum para a predição de Y, isto é, quando $R_2 = R_1$. O sentido da associação é ignorado para estes índices e conforme já foi dito X e Y não necessitam nem mesmo ser ordinais.

Do modo como foi definido, RPR não é simétrico mas existe a possibilidade de defini-lo para que satisfaça essa propriedade; basta supor que, com igual probabilidade, seremos convidados a predizer Y ou X e que na situação 2 seremos informados do valor da outra variável antes de fazer a predição da que está sendo solicitada. Redefinindo R_1, R_2 e, consequentemente, RPR(X,Y) da maneira obviamente adequada teremos um índice satisfazendo a requerida propriedade de si-

metria. Uma generalização que poderia ser feita seria exigir predições mais gerais a respeito das variáveis.

Existe a possibilidade de interpretar os índices de associação já definidos segundo o conceito de RPR. Veremos a título de exemplo, as interpretações para ρ e para τ_a.

Suponhamos que devemos predizer Y com perda igual ao "erro quadrático". Na situação 1 o risco será mínimo se utilizarmos como predição a média de Y e será igual à variância $\sigma^2(Y)$; na situação 2 o risco mínimo é alcançado pelo uso da média condicional de Y dada X, sendo portanto igual à variância condicional de Y dada X, isto é, $\sigma^2(Y|X)$. Teremos então

$$RPR(X,Y) = 1 - \frac{\sigma^2(Y|X)}{\sigma^2(Y)}$$

e acontece que isso é justamente o quadrado de $\rho(X,Y)$ de Pearson desde que a média condicional seja uma função linear de X.

Vamos supor agora que duas observações (X_1,Y_1) e (X_2,Y_2) serão tomadas ao acaso e que devemos predizer se $Y_1 < Y_2$ ou $Y_1 > Y_2$. Se a predição for correta a perda é zero, se incorreta a perda é 1, a não ser que $Y_1 = Y_2$ (embora não seja permitida essa previsão) e nesse caso a perda é $\frac{1}{2}$. Sem a informação prévia sobre a ordenação de X_1 e X_2 nosso risco é $R_1 = \frac{1}{2}$ independentemente da estratégia empregada na pre

dição (podemos, por exemplo, lançar uma moeda). Se formos informados da ordenação de X_1 e X_2 antes de fazer a predição poderemos adotar o seguinte procedimento para obter risco mínimo: se $X_1 = X_2$ lançar uma moeda, de outro modo predizer a ordem de Y_1 e Y_2 de maneira a tornar o par $[(X_1,Y_1),(X_2,Y_2)]$ "concordante" (discordante) se $\tau_a(X,Y)$ é positivo (negativo). O risco será

$$R_2 = \min(p_C,p_D) + \frac{1}{2}\, p_T$$

e RPR(X,Y) será

$$PRP(X,Y) = 1 - \frac{\min(p_C,p_D) + \frac{1}{2}\, p_T}{\frac{1}{2}} = 1 - 2\min(p_C,p_D) - p_T =$$

$$= (p_C+p_D) - 2\,\min(p_C,p_D) = |\tau_a|.$$

Interpretações semelhantes para $|\gamma|$ e $|\tau_b|$ poderão ser encontradas, respectivamente, em Costner (1965) e Wilson (1969).

CAPÍTULO 3

MEDIDAS DE ASSOCIAÇÃO PARCIAL

Conforme já tivemos oportunidade de mencionar as diferentes medidas de associação parcial podem ser obtidas à partir das diferentes medidas de associação total variando a maneira de controlar pela terceira variável. Neste capítulo estudaremos apenas os coeficientes mais comumente encontrados na literatura. Seguiremos, no estudo das principais medidas de associação parcial, o mesmo esquema utilizado no capítulo 1 para as medidas de associação total.

3.1 - COEFICIENTE DE CORRELAÇÃO PARCIAL MOMENTO-PRODUTO DE PEARSON $\rho(X,Y|Z)$

O coeficiente de correlação parcial momento-produto de Pearson foi definido na capítulo 1 e dissemos que a conhecida fórmula de correlação parcial

$$\rho(X,Y|Z) = \frac{\rho(X,Y) - \rho(X,Z)\rho(Y,Z)}{\sqrt{1-\rho^2(X,Z)}\ \sqrt{1-\rho^2(Y,Z)}}$$

podia ser utilizada como definição do mesmo. Agindo desse modo, o coeficiente poderá ser calculado por essa fórmula mes-

mo que a distribuição conjunta de X, Y e Z não seja normal multivariada (e essa é uma suposição necessária para, partindo-se da definição anterior, chegar a essa fórmula). No entanto, já foi dito no capítulo 2 que o uso do coeficiente de correlação total de Pearson só deve ser usado como medida de associação no caso das variáveis apresentarem distribuição conjunta normal bivariada. O mesmo acontece com o coeficiente de correlação parcial pois somente nesses casos é possível utilizar os resultados assintóticos já obtidos para correlação total. De acordo com Fisher (1924) a distribuição do coeficiente de correlação parcial baseado em n observações é a mesma do coeficiente de correlação total baseado em (n-1) observações.

O coeficiente de correlação parcial amostral definido para estimar $\rho(X,Y|Z)$ é

$$r(X,Y|Z) = \frac{r(X,Y) - r(X,Z)r(Y,Z)}{\sqrt{1-r^2(X,Z)}\ \sqrt{1-r^2(Y,Z)}} .$$

Temos então, para o caso de distribuição normal multivariada para (X,Y,Z), usando os resultados acima, que:

Para testarmos a hipótese H_0: $\rho(X,Y|Z) = 0$ utilizamos

$$t = \frac{r(X,Y|Z)\sqrt{n-3}}{\sqrt{1-r^2(X,Y|Z)}}$$

onde t tem, sob H_0, distribuição "t" de Student com (n - 3) graus de liberdade.

A transformação de Fisher vista para o coeficiente de correlação total (r) pode também ser utilizada para testar H_0: $\rho(X,Y|Z) = \rho_0(X,Y|Z)$. Então temos

$$v = \frac{1}{2} \log_e \frac{1+r(X,Y|Z)}{1-r(X,Y|Z)}$$

que, sob H_0, tem distribuição normal aproximada com média

$$\mu_v = \frac{1}{2} \log_e \frac{1+\rho_0(X,Y|Z)}{1-\rho_0(X,Y|Z)}$$

e variância

$$\sigma_v^2 = \frac{1}{n-4}$$

Usamos então a estatística

$$Z = \frac{v-\mu_v}{\sigma_v}$$

que é aproximadamente N(0,1), quando n é grande, para testar a hipótese em questão.

Como um exemplo do cálculo dos coeficientes de correlação parcial (e total) momento-produto vamos utilizar os dados da tabela 3.1 (Adaptado de Steel and Torrie, 1960), onde temos as porcentagens de nitrogênio (X_1) e cloro (X_2) e o logaritmo do tempo de queima (X_3), em segundos, de 30 a-

TABELA 3.1 - Porcentagem de Nitrogênio (X_1) e Cloro (X_2), e logaritmo de tempo de queima das folhas (X_3), em segundos, em amostras de tabaco (Adaptado de Steel and Torrie, 1960).

AMOSTRA Nº	X_1	X_2	X_3
1	3,05	1,45	0,34
2	4,22	1,35	0,11
3	3,34	0,26	0,38
4	3,77	0,23	0,68
5	3,52	1,10	0,18
6	3,54	0,76	0,00
7	3,74	1,59	0,08
8	3,78	0,39	0,11
9	2,92	0,39	1,53
10	3,10	0,64	0,77
11	2,86	0,82	1,17
12	2,78	0,64	1,01
13	2,22	0,85	0,89
14	2,67	0,90	1,40
15	3,12	0,92	1,05
16	3,03	0,97	1,15
17	2,45	0,18	1,49
18	4,12	0,62	0,51
19	4,61	0,51	0,18
20	3,94	0,45	0,34
21	4,12	1,79	0,36
22	2,93	0,25	0,89
23	2,66	0,31	0,91
24	3,17	0,20	0,92
25	2,79	0,24	1,35
26	2,61	0,20	1,33
27	3,74	2,27	0,23
28	3,13	1,48	0,26
29	3,49	0,25	0,73
30	2,94	2,22	0,23
$\sum X_i$	98,36	24,23	20,58
$\bar{X}_i$	3,28	0,81	0,69
$\sum X_i^2$	332,3352	30,1907	20,8074

mostras de tabaco. Queremos calcular $r(X_1,X_2)$ $r(X_1,X_3)$, $r(X_2,X_3)$, $r(X_1,X_3|X_2)$ e $r(X_2,X_3|X_1)$. Para isso além dos valores de $\sum X_i$, $\bar{X}_i$ e $\sum X_i^2$ que constam da tabela foram calculados $\sum X_1X_2 = 81,5834$; $\sum X_1X_3 = 61,6502$; $\sum X_2X_3 = 12,4103$; $\sum(X_1-\bar{X}_1)^2 = 9,8455$; $\sum(X_2-\bar{X}_2)^2 = 10,6209$; $\sum(X_3-\bar{X}_3)^2 = 6,6895$; $\sum(X_1-\bar{X}_1)(X_2-\bar{X}_2) = 2,1413$; $\sum(X_1-\bar{X}_1)(X_3-\bar{X}_3) = -5,8248$; $\sum(X_2-\bar{X}_2)(X_3-\bar{X}_3) = -4,2115$ e então

$$r(X_1,X_2) = 0,2094 \quad r(X_1,X_3) = -0,7177 \quad r(X_2,X_3) = -0,4998$$
$$r(X_1,X_3|X_2) = -0,7238 \quad r(X_2,X_3|X_1) = -0,5133.$$

Se quizessemos testar a hipótese $H_0: \rho(X_2,X_3|X_1) = 0$ calcularíamos

$$t = \frac{r(X_2,X_3|X_1)\sqrt{n-3}}{\sqrt{1-r^2(X_2,X_3|X_1)}}$$

onde t é "t" de Student com (n-3) graus de liberdade.

Esse cálculo foi feito e o resultado obtido, t=-3,11, mostra que a correlação parcial entre porcentagem de cloro e o logaritmo de tempo de queima, em segundos, controlado pela porcentagem de nitrogênio é significantemente diferente de zero ao nível de 1%. (Nos testes bilaterais o sinal de t é ignorado).

3.2 - COEFICIENTE DE ASSOCIAÇÃO PARCIAL DE DAVIS $\gamma(X,Y|Z)$

Na sua definição consideramos X, Y e Z variáveis categóricas, pelo menos ordinais, de modo que a população pos

sa ser representada em uma tabela de 3-entradas. Definimos: p_i como a probabilidade de uma observação casual possuir o i-ésimo valor de Z; p_{C_i} (p_{D_i}, p_{T_i}) como a probabilidade de um par casual de observações ser empatado em Z, em seu i-ésimo valor, e concordante (discordante, empatado) com relação a X e Y. Dessas definições concluímos: $p_{C_i} + p_{D_i} + p_{T_i} = p_i^2$.

O coeficiente de associação parcial de Davis foi definido, então, no capítulo 1 como sendo:

$$\gamma(X,Y|Z) = \frac{\sum_i p_{C_i} - \sum_i p_{D_i}}{\sum_i p_{C_i} + \sum_i p_{D_i}}$$

mas $\sum_i p_{C_i}$ $\left(\sum_i p_{D_i}\right)$ é a probabilidade total de que um par aleatório seja empatado em Z e concordante (discordante) com respeito a X e Y; podemos interpretar $\gamma(X,Y|Z)$ como a diferença entre a probabilidade de um par aleatório, empatado em Z mas não em X e/ou Y, ser concordante com respeito a X e Y e a probabilidade de ser discordante.

É possível escrever o coeficiente de Davis em função dos coeficientes "gama" de Goodman e Kruskal calculados em cada nível i de Z. Para isso podemos escrever "gama" em cada um desses níveis

$$\gamma_i = \frac{p_{C_i} - p_{D_i}}{p_{C_i} + p_{D_i}}$$

Desse modo o coeficiente de Davis ficará

$$\gamma(X,Y|Z) = \frac{\sum_{i=1}^{n} (p_{C_i} + p_{D_i})\gamma_i}{\sum_{i=1}^{n} (p_{C_i} + p_{D_i})}$$

e uma nova interpretação pode ser dada para $\gamma(X,Y|Z)$: é a média ponderada das associações condicionais na qual o peso dado à i-ésima associção é porporcional à probabilidade de um par aleatório de observações ser empatado em Z em seu i-ésimo valor mas não empatado em X ou Y.

O estimador amostral de $\gamma(X,Y|Z)$, bem como, considerações sobre sua distribuição assintótica amostral serão consideradas mais adiante em conjunção com o "coeficiente de associação parcial baseado em pareamento".

3.3 - COEFICIENTE DE CORRELAÇÃO PARCIAL DE KENDALL $\phi(X,Y|Z)$

Suponhamos X, Y e Z medidas pelo menos em escala ordinal e que empates são impossíveis.

Vamos definir então p_0, p_X, p_Y e p_Z do seguinte modo:

p_0 é a probabilidade de um par casual de observações ser *não-discordante*, o que significa serem concordantes em relação a X e Z e a Y e Z, o que implica serem concordantes em relação a X e Y;

p_X é a probabilidade de um par casual de observações ser X-*discordante*, significando que são discordantes em re

lação a X e Z e concordantes em relação a Y e Z o que implica serem discordantes em relação a X e Y;

p_Y é a probabilidade de um par casual de observações ser Y-*discordante*, significando que são concordantes em relação a X e Z e discordantes em relação a Y e Z o que implica serem discordantes com relação a X e Y;

p_Z é a probabilidade do par ser Z-*discordante* significando serem discordantes com relação a X e Z e a Y e Z, implicando, então, serem concordantes em relação a X e Y.

Uma conclusão imediata é que

$$p_C = p_0 + p_Z \quad \text{e} \quad p_D = p_X + p_Y$$

onde p_C e p_D tem o mesmo significado dado nas definições anteriores.

Kendall, ao definir o seu coeficiente $\phi(X,Y|Z)$, que como vimos é dado por

$$\phi(X,Y|Z) = \frac{p_0 p_Z - p_X p_Y}{\sqrt{(p_0+p_X)(p_Y+p_Z)(p_0+p_Y)(p_X+p_Z)}},$$

utilizou o seguinte argumento: se os pares de observações "não-discordantes" e "Z-discordantes" predominam sobre os "X-discordantes" e "Y-discordantes" a correlação parcial é positiva e em caso contrário, é negativa; se eles apresentarem mesma proporção a correlação parcial é nula.

Devemos notar que se $\phi(X,Y|Z) = 0$ então $p_0p_Z - p_Xp_Y = 0$ e portanto $p_0/p_X = p_Y/p_Z$ o que significa que entre pares concordantes e discordantes com relação a X e Z existe mesma proporção quer sejam eles concordantes ou discordantes em relação a Y e Z.

Se $\phi(X,Y|Z) = 1$ nós temos

$$(p_0p_Z - p_Xp_Y)^2 = (p_0+p_X)(p_Y+p_Z)(p_0+p_Y)(p_X+p_Z)$$

e dai:

$$4p_0p_Xp_Yp_Z + p_0^2(p_Xp_Y+p_Xp_Z+p_Yp_Z) + p_X^2(p_0p_Y+p_0p_Z+p_Yp_Z) +$$

$$+ p_Y^2(p_0p_X+p_0p_Z+p_Xp_Z) + p_Z^2(p_0p_Y+p_0p_X+p_Xp_Y) = 0$$

Como p_0, p_X, p_Y e p_Z não podem ser negativos a expressão à esquerda anular-se-á somente se dois deles forem iguais a zero. Para auxiliar o raciciínio vamos representar essas probabilidades em uma tabela 2×2 como no capítulo 1.

		X e Z	
		CONCORDANTES	DISCORDANTES
Y e Z	CONCORDANTES	p_0	p_X
	DISCORDANTES	p_Y	p_Z

Se uma coluna ou uma linha da tabela for constituída de zeros temos o caso em que X ou Y estão em perfeita concordân-

cia ou discordância com Z, mas neste caso $\phi(X,Y|Z)$ não está definido. Temos então que considerar somente as diagonais constituídas de zeros. Se $p_0 = 0$ e $p_Z = 0$, X e Y discordam completamente em suas concordâncias com Z e então $\phi(X,Y|Z) = -1$, se $p_X = 0$ e $p_Y = 0$, X e Y concordam completamente e $\phi(X,Y|Z) = 1$.

O símbolo ϕ para representar o coeficiente de Kendall, foi usado por Quade (1971) devido ao fato desse coeficiente ser formalmente o mesmo "coeficiente ϕ" definido para medir dependência em tabelas de contingência 2×2. Em seu livro, Kendall (1955) utilizou o símbolo $\tau_{QR.P}$ para medir a correlação entre Q e R controlada por P.

Nesse mesmo trabalho Kendall chega à conclusão que é possível calcular o seu coeficiente de correlação parcial utilizando a fórmula de correlação parcial de Pearson onde os coeficientes de correlação total de Pearson são substituídos pelos coeficientes τ de associação ordinal total.

Para correlação entre X e Y controlada por Z teríamos:

$$\phi(X,Y|Z) = \frac{\tau(X,Y) - \tau(X,Z)\tau(Y,Z)}{\sqrt{[1-\tau^2(X,Z)][(1-\tau^2(Y,Z)]}}$$

Embora tenhamos feito a suposição inicial de que empates eram impossíveis a fórmula acima é válida nos casos de empates desde que a medida τ_b de associação ordinal total de

Kendall seja utilizada (Hawkes, 1971). No entanto Somers (1966) prefere descarta-se dos pares empatados e calcular $\phi(X,Y|Z)$ unicamente à partir dos não-empatados.

Ainda não é conhecida a distribuição amostral do coeficiente de associação parcial $\phi(X,Y|Z)$ de Kendall, não havendo, portanto, possibilidade de serem testadas hipóteses sobre valores observados do mesmo em amostras casuais de uma população.

3.4 - COEFICIENTE DE ASSOCIAÇÃO PARCIAL BASEADO EM PAREAMENTO $\theta(X,Y|Z)$

Para este coeficiente X e Y devem ser pelo menos ordinais porém Z pode ser nominal.

Estabelecida uma regra de modo que um par casual de observações sobre (X,Y,Z) possa ser classificado como "pareado" sem ambiguidades, definimos PAREAMENTO como sendo o evento em que o par casual é "pareado". Assumindo que

$$P\{\text{PAREAMENTO}\} > 0$$

definimos o coeficiente de associação parcial pareada como

$$\phi(X,Y|Z) = P\{C|\text{PAREAMENTO}\} - P\{D|\text{PAREAMENTO}\}$$

onde C (D) é o evento em que um par aleatório de observações (X_1,Y_1,Z_1) e (X_2,Y_2,Z_2), é concordante (discordante) com relação a X e Y.

A interpretação de $\theta(X,Y|Z)$ é obtida diretamente da definição como sendo a diferença entre as probabilidades con

dicionais de concordância e discordância de um par aleatório de observações dado que o mesmo é pareado.

Concluímos que $-1 \leq \theta(X,Y|Z) \leq 1$ e que

$\theta(X,Y|Z) = 1$ se todos os pares pareados são concordantes, isto é, $P\{C|PAREAMENTO\} = 1$

$\theta(X,Y|Z) = -1$ se todos os pares pareados são discordantes, isto é, $P\{D|PAREAMENTO\} = 1$

$\theta(X,Y|Z) = 0$ se os pares pareados têm igual probabilidade de serem concordantes ou discordantes, isto é, $P\{C|PAREAMENTO\} = P\{D|PAREAMENTO\}$

Suponhamos, agora, que o coeficiente θ deva ser estimado de uma amostra casual de n observações (X_i, Y_i, Z_i), $i=1,2,\ldots,n$.

Entre os $N = n(n-1)/2$ pares de observações possíveis seja N_M o número dos que são pareados e dentre estes consideremos N_{CM} (N_{DM}) como o número de concordantes (discordantes) com relação a X e Y.

O estimador natural de θ é

$$T(X,Y|Z) = \frac{N_{CM} - N_{DM}}{N_M}$$

que é a diferença entre as proporções dos pares pareados, na amostra, que são concordantes e discordantes. (Se acontecer que uma amostra não contenha pares pareados poderemos definir T arbitrariamente como sendo igual a zero).

Temos que $-1 \leq T(X,Y|Z) \leq 1$ e que

$T(X,Y|Z) = 1$ se todos os pares pareados observados forem concordantes;

$T(X,Y|Z) = -1$ se esses pares forem todos discordantes, e

$T(X,Y|Z) = 0$ se a proporção dos pares pareados concordantes for igual à dos pareados discordantes.

Consideremos agora a distribuição amostral de T (Quade, 1971).

Para cada $i=1,2,\ldots,n$ seja M_i o número de observações (X_j,Y_j,Z_j), $j \neq i$, que são pareadas com (X_i,Y_i,Z_i) e seja W_i o número destas que são concordantes com (X_i,Y_i,Z_i) menos o número das que são discordantes.

Desse modo $\sum M_i = 2N_M$ (porque cada par pareado aparece duas vezes) e $\sum W_i = 2(N_{CM}-N_{DM})$. Portanto

$$T(X,Y|Z) = \frac{\sum W_i}{\sum M_i}$$

Esse método de cálculo leva a uma fórmula conveniente para o erro padrão assintótico de T

$$S_T = \frac{2}{(\Sigma M_i)^2} \sqrt{\Sigma W_i^2 (\Sigma M_i)^2 - 2\Sigma W_i \Sigma M_i \Sigma W_i M_i + (\Sigma W_i)^2 \Sigma M_i^2}$$

A distribuição de T é assintoticamente normal, isto é, para n grande a quantidade $\frac{T-\theta}{S_T}$ é aproximadamente N(0,1)

Assim, pelo menos para grandes amostras é possível fazer inferências baseadas no "coeficiente de associação pareada".

Se Z_α é o valor crítico para a variável $Z \sim N(0,1)$ de modo que $P\{|Z| \geq Z_\alpha\} = \alpha$ nós teremos, por exemplo, um intervalo de confiança (bi-lateral) com coeficiente de confiança $100(1-\alpha)\%$ dado por

$$T - S_T Z_\alpha \leq \theta \leq T + S_T Z_\alpha$$

e a hipótese H_0: $\theta = \theta_0$ pode ser rejeitada em favor da hipótese alternativa H_1: $\theta \neq \theta_0$ se e somente se o valor θ_0 cai fora desse intervalo de confiança.

Testes e intervalos de confiança uni-laterais podem ser construídos também, da maneira óbvia.

Como caso especial, a hipótese H_0: $\theta = \theta$ pode ser rejeitada se e somente se $\left|\frac{T}{S_T}\right| \geq Z_\alpha$. Contudo para essa hipótese nula poderá ser preferível um teste envolvendo somente os W's: rejeitar H_0 se e somente se

$$\frac{\sqrt{n}\bar{W}}{2\sqrt{\Sigma(W_i - \bar{W})^2}} \geq Z_\alpha \quad \text{onde} \quad \bar{W} = \frac{\Sigma W_i}{n} \; .$$

O "coeficiente de associação parcial pareada" pode ser interpretado como uma versão generalizada de correlação parcial no sentido de correlação condicional média. Vamos supor, para simplificar, que Z é uma variável aleatória puramente discreta. Seja E(z), para cada valor de $Z = z$, o evento em que duas observações aleatórias (X_1, Y_1, Z_1) e (X_2, Y_2, Z_2) tem $z_1 = z_2 = z$, isto é, são empatadas com respeito a Z em z.

Nessas condições teremos

$$\tau(X,Y|Z) = P\{C|E(z)\} - P\{D|E(z)\}$$

Vamos, agora, construir uma associação condicional média ponderando as associações condicionais em z proporcionalmente à probabilidade de observar um par empatado em z.

Isto é,

$$\tau(X,Y|Z) = \frac{\sum_z P\{E(z)\}\tau(X,Y|Z=z)}{\sum_z P\{E(z)\}}$$

mas como

$$P\{C|E(z)\} = \frac{P\{C \text{ e } E(z)\}}{P\{E(z)\}} \quad \text{e} \quad P\{D|E(z)\} = \frac{P\{D \text{ e } E(z)\}}{P\{E(z)\}}$$

podemos escrever

$$P\{E(z)\tau(X,Y|Z=z) = P\{C \text{ e } E(z)\} - P\{D \text{ e } E(z)\}$$

Vamos denominar de EMPATE o evento em que um par de observações casual é empatado em Z, isto é, EMPATE é a união de todos os eventos E(z), então:

$$\sum_z P\{E(z)\} = P\{\text{EMPATE}\}$$

e

$$\sum_z P\{E(z)\}\tau(X,Y|Z=z) = P\{C \text{ e EMPATE}\} - P\{D \text{ e EMPATE}\}$$

e portanto

$$\tau(X,Y|Z) = P\{C|EMPATE\} - P\{D|EMPATE\}$$

mas então, se definirmos, para $\theta(X,Y|Z)$, PAREAMENTO = EMPATE teremos

$$\theta(X,Y|Z) = \tau(X,Y|Z)$$

isto é, o "índice de associação parcial pareada" é uma verdadeira associação parcial, no sentido de associação parcial média, se duas observações são definidas como PAREADAS quando seus valores de Z são iguais.

Se a função de probabilidade da variável discreta Z é $h(z)$ de modo que $P\{E(z)\} = h^2(z)$ o coeficiente de associação parcial pode ser escrito como

$$\tau(X,Y|Z) = \frac{\sum h^2(z)\tau(X,Y|Z=z)}{\sum h^2(z)}$$

Se, no entanto, Z é uma variável contínua, com função de densidade $h(z)$, podemos escrever a expressão análoga

$$\tau(X,Y|Z) = \frac{\int h^2(z)\tau(X,Y|Z)dz}{\int h^2(z)dz}$$

onde $\tau(X,Y|Z=z)$ é a associação dentro da distribuição condicional de X e Y dada $Z = z$. Mas agora uma amostra casual não terá pares empatados nos quais basear uma estimativa amostral de τ. Em vez de exigir que todos os pares usados no in

dice amostral de associação parcial sejam exatamente empatados, nós relaxamos a exigência para permitir que pares sejam considerados pareados embora só praticamente empatados. Chamaremos de TOLERÂNCIA a discrepância máxima permitida entre duas observações para que sejam consideradas pareadas, por exemplo se PAREAMENTO = EMPATE então a tolerância é zero.

O índice amostral T de associação pareada estima estritamente $\theta(X,Y|Z)$ mas em qualquer situação real, se a tolerância for pequena, θ será essencialmente equivalente a $\tau(X,Y|Z)$. Contudo os dois índices populacionais não serão ordinàriamente muito idênticos em valor. Um exemplo particular é o caso em que X e Y são condicionalmente independentes dada $Z = z$ para todo z. Isto é suficiente, embora não seja necessário, para implicar que cada associação condicional $\tau(X,Y|Z=z) = 0$ e portanto que a associação parcial $\tau(X,Y|Z) = 0$ também, mas isso não implica que $\theta(X,Y|Z) = 0$.

Agora vejamos como os coeficientes de associação total τ_a de Kendall, γ de Goodman e Kruskal e o coeficiente de associação parcial de Davis são casos particulares de $\theta(X,Y|Z)$.

De fato, $\theta(X,Y|Z)$ é na realidade uma família de coeficientes que se diferenciam pelas suas definições de PAREAMENTO, sendo que a única restrição é $P\{\text{PAREAMENTO}\} > 0$.

Vamos supor que duas observações são *sempre pareadas*. Nesse caso $P\{\text{PAREAMENTO}\} = 1$ e

$$\theta(X,Y|Z) = P\{C\} - P\{D\} = \tau_a(X,Y)$$

Como $M_i = n-1$, $i=1,2,3,\ldots,n$, podemos mostrar que neste caso, $T = t_a$.

Além disso se C_i (D_i) é o número de observações concordantes (discordantes) com a observação (X_i,Y_i,Z_i) o erro padrão assintótico de t_a será:

$$S_a = \frac{2}{n(n-1)} \sqrt{\Sigma W_i^2 - (\Sigma W_i)^2/n}$$

onde $W_i = C_i - D_i$, para $i=1,2,3,\ldots,n$.

Agora vamos supor que duas observações são pareadas se e somente se são *não-empatadas em X ou Y*. Nesse caso PAREAMENTO é a união dos eventos C e D e portanto

$$\theta(X,Y|Z) = \frac{p\{C\} - p\{D\}}{P\{C\} + P\{D\}} = \gamma(X,Y)$$

Aqui $W_i = C_i - D_i$ e $M_i = C_i + D_i$, $i=1,2,3,\ldots,n$ e portanto $T = G$ e o erro padrão assintótico de G é dado por

$$S_G = \frac{4}{(\Sigma C_i + \Sigma D_i)^2} \sqrt{\Sigma C_i^2 (\Sigma D_i)^2 - 2\Sigma C_i \Sigma D_i \Sigma C_i D_i + (\Sigma C_i)^2 \Sigma D_i^2}$$

Finalmente vamos supor que duas observações são pareadas se e somente se *são empatadas em Z mas não em X ou Y*

Com essa definição:

$$\theta(X,Y|Z) = \frac{P\{C \text{ e EMPATE}\} - P\{D \text{ e EMPATE}\}}{P\{C \text{ e EMPATE}\} + P\{D \text{ e EMPATE}\}} = \gamma(X,Y|Z)$$

que é o "coeficiente de associação parcial de Davis".

Se C_i (D_i) é agora redefinido como o número de observações concordantes (discordantes) com a observação (X_i,Y_i,Z_i) com relação a X e Y e também empatadas com ela em Z, então $W_i = C_i - D_i$ e $M_i = C_i + D_i$ da mesma forma que para o coeficiente de associação total $\gamma(X,Y)$, e o erro padrão assintótico do coeficiente de Davis tem a mesma fórmula de S_G.

Para ilustrar o método de computação de T vamos considerar um exemplo (Quade, 1971) baseado na Tabela 3.2.

Seja X o resultado de exame, uma variavel ordinal registrada como A, B, C, D ou F; e seja Y a variável métrica altura, registrada em polegadas. A variável controle Z é bivariada e a primeira componente é a variável nominal sexo (Z_1) e a segunda componente é quociente de inteligência QI, (Z_2).

O índice amostral de associação pareada (T) entre resultado de exame e altura, controlado por sexo e QI, isto é, entre X e Y dadas Z_1 e Z_2, é obtido usando os valores M_i e W_i, definidos juntamente com o índice T, que aparecem na última coluna da tabela. Neste cálculo duas crianças são consideradas como pareadas se elas são do mesmo sexo e diferem em QI por não mais do que 10 unidades. A primeira criança,

TABELA 3.2 - Sexo, QI, altura e resultado do exame final para uma classe de crianças de quarta-série (dados ficticios) (Adaptado de Quade 1971, Tabela 6.1, pg. 31).

	RESULTADO DE EXAME	ALTURA (pol.)	SEXO	QI	PAREADOS EM SEXO e QI	
i	X	Y	Z_1	Z_2	M	W
1	F	50	M	85	2	2
2	D	58	M	92	5	-1
3	D	54	M	93	6	5
4	A	56	M	96	5	-1
5	C	55	M	100	6	2
6	C	58	M	102	6	-1
7	B	57	M	103	5	1
8	C	53	M	109	5	1
9	F	54	M	115	4	-2
10	B	57	M	118	5	2
11	A	49	M	120	4	-4
12	D	52	M	123	4	0
13	B	60	M	128	3	0
14	C	51	F	83	1	-1
15	B	50	F	86	1	-1
16	C	52	F	98	3	-2
17	D	57	F	99	3	-1
18	F	53	F	105	5	2
19	C	53	F	106	5	0
20	A	54	F	111	4	2
21	C	55	F	114	4	0
22	C	51	F	121	3	1
23	C	52	F	131	3	2
24	A	55	F	135	2	2
25	B	54	F	140	2	2

por exemplo, é pareada com exatamente duas outras, isto é, a segunda e a terceira (por conveniência do cálculo manual os dados foram arranjados segundo as variáveis controle), portanto $M_i = 2$, e ela é concordante com ambas – em particular

ela é a menor das três e também recebeu o grau mais baixo — portanto $W_i = 2$. Os valores de M_i e W_i para as outras 24 crianças podem ser conferidos de modo semelhante. Podemos então computar $\sum M_i = 96$ o que indica que existem 48 pares pareados de crianças e $\sum W_i = 10$, indicando que há 5 pares concordantes a mais do que discordantes; portanto o índice

$$T = \sum W_i / \sum M_i = 10/96 = 0{,}104.$$

Tendo calculado $\sum M_i^2 = 422$, $\sum M_i W_i = 50$ e $\sum W_i^2 = 90$ nós encontramos $S = 0{,}191$. Portanto o índice é menor do que seu erro padrão e certamente não significantemente diferente de zero no sentido estatístico. Se essa amostra pudesse ser considerada como grande, nós poderíamos tomar $T/S = 0{,}545$ como $N(0,1)$ e testar H_0: $\theta = 0$, e teríamos também o intervalo de confiança com coeficiente de confiança 95%, por exemplo, $T \pm 1{,}96\ S$ ou $(-0{,}270,\ 0{,}469)$ para o índice populacional θ. No entanto é conveniente ressaltar que temos somente 25 observações e 48 pares pareados que não são independentes uns dos outros.

Vamos agora ver um outro exemplo que usa os dados de Hajda, citados por Davis (1967), que foram obtidos de uma inspeção amostral de mulheres de Baltimore. X é uma dicotomia assumindo os valores "alta" e "baixa" se a entrevistada tinha idade, respectivamente, acima ou abaixo de 45 anos; Y é uma outra dicotomia, assumindo os valores "alta" e "baixa" se ela tinha ou não lido um livro recentemente, Z distin

TABELA 3.3

IDADE X	LEITURA DE LIVRO Y	EDUCAÇÃO Z	FREQÜÊNCIA F	C	D	T	W	M_1	M_2
ALTA	ALTA	UNIVERSITÁRIA	104	46	0	302	46	348	46
	BAIXA		36	0	163	185	-163	348	163
BAIXA	ALTA		163	0	36	312	-36	348	36
	BAIXA		46	104	0	244	104	348	104
ALTA	ALTA	COLEGIAL	159	327	0	627	327	954	327
	BAIXA		179	0	290	664	-290	954	290
BAIXA	ALTA		290	0	179	775	-179	954	179
	BAIXA		327	159	0	795	159	954	159
ALTA	ALTA	MENOS QUE COLEGIAL	54	133	0	412	133	545	133
	BAIXA		335	0	24	521	-24	545	24
BAIXA	ALTA		24	0	315	210	-335	545	335
	BAIXA		133	54	0	491	54	545	54

gue três categorias de realização educacional "universitária", "colegial" e "menos do que colegial". Duas definições de pareamento serão consideradas: a primeira produzindo um coeficiente de correlação parcial direto, declara duas observações, *pareadas* se são empatadas em Z; enquanto que a segunda, produzindo o coeficiente de associação parcial de Davis, declara-as *pareadas* somente se elas são empatadas em Z mas não em X ou Y. A Tabela 3.3 mostra os cálculos com algum detalhe. Lá estão listados os 12 possíveis valores de

(X,Y,Z) e a freqüência F com que cada um ocorre na amostra. Então são mostradas quantas observações são empatadas em Z e concordantes (discordantes, empatadas) com respeito a X e Y com cada uma das observações num dado valor, denominadas C (D,T). Nós temos que W = C - D, para a primeira definição de pareamento, $M_1 = C + D + T$, e para a segunda, $M_2 = C + D$.

Em um ou outro caso

$$T = \frac{\Sigma FW}{\Sigma FM}$$

e

$$S = \frac{2}{(\Sigma FM)^2} \sqrt{\Sigma FM^2 (\Sigma FW)^2 - 2\Sigma FM\Sigma FW\Sigma FMW + (\Sigma FM)^2 \Sigma FW^2}$$

No segundo caso, isto é, $M_2 = C + D$, a fórmula equivalente para S em termos de C's e D's é, na forma de dados grupados

$$S = \frac{4}{(\Sigma FC + \Sigma FD)^2} \sqrt{\Sigma FC^2 (\Sigma FD)^2 - 2\Sigma FC\Sigma FD\Sigma FCD + (\Sigma FC)^2 \Sigma FD^2}$$

Para o exemplo, $\Sigma FW = -3718$ e $\Sigma FW^2 = 5572914$. Para a primeira definição de pareamento

$\sum FM = 1330092$, $\sum FM^2 = 1073601726$ e $\sum FMW = -1531320$

produzindo para o coeficiente de associação parcial amostral T = -0,0028 com S = 0,0112. Para a segunda definição, $\sum FM = 259554$, $\sum FM^2 = \sum FW^2$ (esta igualdade pode valer sempre que X e Y são dicotômicas, mas não em geral) e $\sum FMW = -1070650$, pro-

duzindo o coeficiente de associação parcial de Davis T = = -0,0143 com S = 0,0581. É interessante notar que se calculássemos |T/S| para testar a hipótese, H_0: $\theta = 0$, teríamos aproximadamente o mesmo valor e consequentemente um mesmo nível de significância.

Para o mesmo exemplo foram calculados:

$t_a(X,Y) = -0,0596$

$\gamma(X,Y) = -0,2412$

$t_b(X,Y) = -0,1206$ $\qquad$ $t_b(X,Z) = -0,2442$

$t_b(Y,Z) = 0,4139$ $\qquad$ e $\qquad$ $t_b(X,Y|Z) = -0,0221$

pelo método sugerido por Hawkes para $\phi(X,Y|Z)$, isto é, usando a mesma fórmula do coeficiente de correlação parcial momento-produto, considerando os empates, utilizando "tau-b" de Kendall para as correlações totais. Agora calculando pelo método de Somers, isto é, descartando-nos dos empates obtemos:

		X e Z	
		CONCORDANTES	DISCORDANTES
Y e Z	CONCORDANTES	68987	180932
	DISCORDANTES	15600	27456

$$\phi(X,Y|Z) = -0,0674.$$

3.5 - EXTENSÃO DO CONCEITO DE REDUÇÃO PROPORCIONAL EM RISCO À ASSOCIAÇÃO PARCIAL

O *coeficiente de correlação parcial momento-produto de Pearson* pode ser obtido da mesma forma que o coeficiente de correlação total especificando a perda como o erro quadrático em predizer Y.

O *coeficiente de correlação parcial de Davis* pode ser obtido diretamente da interpretação, em termos de RPR, do coeficente $\gamma(X,Y)$ de Goodman e Kruskal se a afirmação à respeito de Y que devemos fazer é a predição da ordenação de Y em duas observações aleatórias empatadas em Z, onde na situação 1: seremos informados apenas do valor comum de Z e na situação 2: seremos, além disso, informados da ordenação de X.

CAPÍTULO 4

OUTRAS MEDIDAS DE ASSOCIAÇÃO

Neste capítulo iremos discorrer de maneira muito breve sobre outras medidas de associação encontradas na literatura mas que foram omitidas dos capítulos anteriores, entre outras razões, pelo fato de terem sido propostas para medir associação em situações muito especiais; isso é o que acontece com aquelas definidas à partir de tabelas 2×2, embora a omissão, para algumas, seja apenas aparente pois elas são casos especiais das que estudamos. Uma outra causa das omissões é a dificuldade maior para interpretar e calcular que algumas medidas apresentam quando comparadas com aquelas já estudadas por nós.

4.1 - MEDIDAS BASEADAS NA ESTATÍSTICA "QUI-QUADRADO" (χ^2)

Para definir estas medidas necessitamos fazer algumas suposições:

Vamos supor que X e Y sejam variáveis categóricas nominais e cuja distribuição conjunta pode ser representada em uma tabela de duas entradas como segue:

X \ Y	1	2	...	q	TOTAIS
1	n_{11}	n_{12}	...	n_{1q}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2q}	$n_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
p	n_{p1}	n_{p2}	...	n_{pq}	$n_{p\cdot}$
TOTAIS	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot q}$	$n_{\cdot\cdot}$

onde n_{ij} é o número de pares de observações de (X,Y) que são classificadas na classe i da variável X e na classe j da variável Y, i=1,2,...,p; j=1,2,...,q.

As medidas que estudamos nos capítulos 1, 2 e 3 dependiam da ordem das linhas e colunas mas as que iremos ver agora são invariantes a qualquer permutação de linhas e colunas.

Se uma população finita tem $n\cdot\cdot$ membros e é apresentada em uma tabela $p \times q$ como a que foi apresentada definimos a estatística "qui-quadrado" como

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot}}$$

e se fizermos $f_{ij} = \frac{n_{ij}}{n_{\cdot\cdot}}$ teremos

$$\chi^2 = n_{\cdot\cdot} \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}}$$

ou

$$\chi^2 = n_{..}\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{f_{ij}^2}{f_{i\cdot}f_{\cdot j}} - 1\right] = n_{..}\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right]$$

Notemos que χ^2 ~~existe~~ cresce com $n_{..}$ e por essa razão não é uma medida adequada para associação.

Numa tentativa de remediar a situação foi definida

$$\phi^2 = \chi^2/n_{..}$$

que é conhecida como "*contingência média quadrática*", mas ela depende do tamanho da tabela. Vejamos como isso acontece: suponhamos que $p = q$ (isto é, temos uma tabela quadrada) e que além disso existe uma perfeita associação entre X e Y, isto é, $n_{i\cdot} = n_{\cdot i} = n_{ii}$ para todo i. Calculando χ^2 pelas fórmulas acima chegamos ao resultado $\chi^2 = n_{..}(p-1)$ e portanto que $\phi^2 = (p-1)$ e este é o valor máximo para ϕ^2 em uma tabela quadrada. Vamos supor agora que $p \neq q$. Nesse caso teremos uma associação perfeita quando todas as entradas estiverem concentradas na diagonal maior. O número de celas nessa diagonal é $\min(p,q)$ e desse modo teremos $\chi^2 = n_{..}\min[(p-1),(q-1)]$ e portanto $\phi^2 = \min[(p-1),(q-1)]$.

Foram feitas algumas tentativas para normalizar ϕ^2 de modo que sua variação fosse no intervalo convencional [0,1].

Tschuprow sugeriu como fator normalizante de ϕ^2 a média geométrica entre (p-1) e (q-1):

$$T = \left\{ \frac{\chi^2/n_{..}}{[(p-1)(q-1)]^{1/2}} \right\}^{1/2}$$

Maung (1941) sugeriu o máximo de ϕ^2, isto é,

$$\min[(p-1),(q-1)]$$

como fator normalizante, de modo que

$$C = \left\{ \frac{\chi^2/n_{..}}{\min[(p-1)(q-1)]} \right\}^{1/2}$$

Uma outra medida baseada em ϕ^2, conhecida como "*coeficiente de contingência*" foi atribuída a Pearson (1904) por Yule e Kendall (1940):

$$P = \left(\frac{\phi^2}{1+\phi^2} \right)^{1/2} = \left(\frac{\chi^2}{n_{..}+\chi^2} \right)^{1/2}$$

Assumindo distribuição normal bivariada para (X,Y), com "coeficiente de correlação momento-produto" igual a ρ, representada numa tabela de contingência é possível mostrar que $P^2 \longrightarrow \rho^2$ em probabilidade à medida que o número de categorias na tabela aumenta. Se X e Y são independentes não existe associação, consequentemente se $P = 0$ nós temos $\chi^2 = 0$ o que implica (ver na definição de χ^2) cada

$$(n_{ij} - n_{i.}n_{.j}/n_{..}) = 0 \therefore n_{ij} = n_{i.}n_{.j}/n_{..}$$

e X e Y são independentes. Temos claramente, $0 \leq P \leq 1$ mas co

mo P depende de ϕ^2, ele de modo geral não atinge o limite 1 pois dependerá, com ϕ^2, do tamanho da tabela.

Dos dois coeficientes anteriores, C e T, somente C poderá, sempre, atingir o limite +1. T só atingirá no caso da tabela ser quadrada.

4.2 - MEDIDAS BASEADAS EM PREDIÇÃO ÓTIMA

Vamos continuar supondo que X e Y sejam variáveis categóricas nominais e que a distribuição conjunta das mesmas pode ser representada em uma tabela $p \times q$ como a do item anterior.

Suponhamos que nosso objetivo é predizer a que classe de variável Y pertence uma observação. Se a classe de X, para a observação, é desconhecida, o melhor que podemos fazer é escolher a classe de Y com maior total marginal, isto é, o valor de m satisfazendo

$$f_{\cdot m} = \max\{f_{\cdot 1}, \ldots, f_{\cdot q}\}$$

(estamos supondo a tabela construída em termos das f_{ij}). A nossa probabilidade de erro é, neste caso, $P_1 = 1 - f_{\cdot m}$.

Agora, vamos supor que a classe de X, para a observação, é conhecida como sendo a. Então é claro que na nossa predição de Y devemos considerar apenas a linha a da tabela e nossa melhor predição é a classe de Y que corresponde à ce

la de maior freqüência na linha a, isto é, o valor de m_a satisfazendo

$$f_{am_a} = \max\{f_{a1}, \ldots, f_{aq}\}$$

Notemos que m_a varia de linha para linha, pois do contrário a informação sobre a classe de X seria desnecessária. Considerando que existem p linhas cada uma ocorrendo com freqüência $f_{i.}$ o nosso erro, neste caso, será

$$P_2 = 1 - \sum_{i=1}^{p} f_{im_i}, \text{ onde } \sum_{i=1}^{p} f_{im_i}$$

é a soma das freqüências máximas para cada linha da tabela.

Como uma medida do poder preditivo de X para Y foi sugerida por Goodman e Kruskal (1954) a medida

$$\lambda_y = \frac{P_1 - P_2}{P_1} = \frac{\sum_{i=1}^{p} f_{im_i} - f_{\cdot m}}{1 - f_{\cdot m}}$$

que pode ser interpretada como redução proporcional em erro devida ao conhecimento da classe de X.

Teríamos uma definição análoga para a predição de X

$$\lambda_x = \frac{\sum_{j=1}^{q} f_{m_j j} - f_{m\cdot}}{1 - f_{m\cdot}}$$

Se as predições de X à partir de Y ou de Y à partir de X forem igualmente importantes poderá ser obtida uma medida simétrica considerando a predição de X metade das vezes e a predição de Y a outra metade. A probabilidade de erro quando uma classe preditora é desconhecida será então

$$P_1 = 1 - \frac{1}{2}(f_{\cdot m} + f_{m\cdot})$$

e quando a classe preditora é conhecida a probabilidade de erro será

$$P_2 = 1 - \frac{1}{2}\left(\sum_{i=1}^{p} f_{im_i} + \sum_{j=1}^{q} f_{m_j j} \right)$$

Então definimos

$$\lambda = \frac{P_1 - P_2}{P_1} = \frac{\sum_{i=1}^{p} f_{im_i} + \sum_{j=1}^{q} f_{m_j j} - f_{\cdot m} - f_{m\cdot}}{2 - f_{\cdot m} - f_{m\cdot}}$$

ou em termos dos n_{ij}

$$\lambda = \frac{\sum_{i=1}^{p} n_{im_i} + \sum_{j=1}^{q} n_{m_j j} - n_{\cdot m} - n_{m\cdot}}{2n_{\cdot\cdot} - n_{\cdot m} - n_{m\cdot}}$$

Então λ é redução proporcional em erro devida ao uso das classes preditoras quando as direções de predição são igualmente importantes.

λ é indeterminada se toda a população está numa única cela, de outro modo $0 \leq \lambda \leq 1$. $\lambda = 1$ se e só se a população inteira está em celas isoladas, isto é, que são as únicas celas não nulas da sua linha e sua coluna.

4.3 - CORRELAÇÃO CANÔNICA

Uma abordagem diferente para medir associação entre variáveis nominais é atribuir, a cada classe das variáveis, escores e então calcular o coeficiente de correlação momento-produto de Pearson com esses escores. Surge a questão: como escolher os escores? Uma resposta razoável seria escolher os escores que tornem máxima a correlação entre X e Y. Esse então é o problema da correlação canônica (Anderberg, 1973).

4.4 - MEDIDAS ENTRE VARIÁVEIS BINÁRIAS

Suponhamos que as variáveis X e Y sejam expressas como dicotomias, isto é, presente-ausente, sim-não, alto-baixo, etc. Para esses casos podemos simplificar a tabela utilizada nos itens anteriores

X \ Y	1	0	TOTAIS
1	a	b	a+b
0	c	d	c+d
TOTAIS	a+c	b+d	n

É comumente convencionado empregar 1 e 0 para denominar as classes de X e de Y. Essas denominações são muitas vezes empregadas como escores.

A primeira medida que vamos definir é o *coeficiente* ϕ de associação.

Utilizando o sistema de escores 1,0, como vimos acima, e calculando o "coeficiente de correlação momento-produto" de Pearson para a população representada na tabela 2×2 teremos

$$\sum_{i=1}^{n} x_i = a+b \qquad \sum_{i=1}^{n} y_i = a+c$$

$$\sum_{i=1}^{n} x_i y_i = a \qquad \sum_{i=1}^{n} x_i^2 = a+b \qquad \sum_{i=1}^{n} y_i^2 = a+c$$

onde x_i, y_i são os escores atribuídos ao indivíduo i da população de acordo com as classes de X e Y que ele pertence. Temos então para ρ

$$\rho = \frac{a - (a+b)(a+c)/n}{\sqrt{[a+b-(a+b)^2/n][a+c-(a+c)^2/n]}} = \frac{ad - bc}{\{(a+b)(c+d)(a+c)(b+d)\}^{1/2}}$$

e essa é a definição do coeficiente ϕ.

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Calculando a estatística χ^2 para essa mesma tabela e lembrando que $p = q = 2$ chegaremos ao interessante resultado

$$\rho^2 = \phi^2 = \chi^2/n = T^2 = c^2$$

Como ρ é uma medida invariante por transformações lineares, os dois escores 0 e 1 são arbitrários porque podem ser transformados em qualquer outro par de escores. Isso nos leva a concluir que qualquer par de escores é ótimo e o problema da correlação canônica reduz-se também à equação de ϕ.

Yule (1912) propôs duas medidas de associação em tabelas 2×2

$$Y = \frac{(ad)^{1/2} - (bc)^{1/2}}{(ad)^{1/2} + (bc)^{1/2}}$$

e

$$Q = \frac{ad-bc}{ad+bc}$$

Ambas estão relacionadas por

$$Q = \frac{2Y}{(1+Y^2)}$$

Tanto Y como Q independem dos totais marginais da tabela.

Goodman e Kruskal (1954) mostraram que a medida Q proposta por Yule é a versão 2×2 da sua medida de associação ordinal γ.

Edwards (1963) mostrou que uma medida de associação em uma tabela 2×2 deveria ser uma função do quociente-

cruzado $R = bc/ad$ e concluiu que

$$Y = \frac{1-R^{1/2}}{1+R^{1/2}} \quad e \quad Q = \frac{1-R}{1+R}$$

desde que $ad \neq 0$.

Além dessas medidas encontramos uma grande quantidade de coeficientes de associação, principalmente em trabalhos de taxonomia numérica, zoogeografia e ecologia, que são definidos à partir da contagem, e representação em tabelas 2×2, do número total de coincidências e não-coincidências, em relação às classes de X e Y, dos indivíduos da população.

4.5 - ASSOCIAÇÃO QUADRANTE

Esta é talvez a medida mais simples de associação entre duas variáveis aleatórias e relaciona-se diretamente à soma das probabilidades de um ponto, representando um par de valores das variáveis, estar em quadrantes opostos pela origem de um sistema coordenado cartesiano natural.

Chamando o par de variáveis de (X,Y) e sendo (x_0,y_0) um valor fixado de (X,Y) temos que as medidas quadrantes de associação são baseadas em

$$P\{(X>x_0 \text{ e } Y>y_0) \text{ ou } (X<x_0 \text{ e } Y<y_0)\}$$

ou de modo mais conveniente

$$P\{(X-x_0)(Y-y_0)>0\}$$

e é a probabilidade de que os desvios de X e Y em relação a x_0 e y_0 tenham os mesmos sinais, isto é, que (X,Y) esteja no 1º e 3º quadrantes considerando como origem o ponto (x_0,y_0).

Se fizermos (x_0,y_0) = (Med X, Med Y) onde onde Med indica a *mediana* definimos

$$\sigma_s = P\{(X-MedX)(Y-MedY)>0\}$$

que é a probabilidade de que os desvios de X e Y à partir de suas medianas tenham o mesmo sinal. Obviamente $0 \leq \sigma_s \leq 1$ e

$\sigma_s = 1$ se e somente se (X-MedX) e (Y-MedY) são positivos ou negativos ao mesmo tempo com probabilidade um.

$\sigma_s = 0$ se e somente se (X-MedX) e (Y-MedY) têm sinais diferentes com probabilidade 1.

Se X e Y são independentes $\sigma_s = \frac{1}{2}$ (mas a recíproca não é necessariamente verdadeira).

Podemos considerar analogamente

$$\sigma_d = P\{(X-MedX)(Y-MedY)<0\}$$

ou a probabilidade de que (X,Y) esteja no 2º e 4º quadrantes considerando a origem em (MedX,MedY).

Claramente $\sigma_s + \sigma_d = 1$.

É definida então a medida de associação

$$Q = \sigma_s - \sigma_d = 2\sigma_s - 1$$

que é a diferença entre as probabilidade de que desvios de

X e Y em relação às suas medianas tenham sinais iguais e diferentes.

Teremos

$$Q = 1 \text{ se e somente se } \sigma_s = 1$$

$$Q = -1 \text{ se e somente se } \sigma_d = 1$$

$$Q = 0 \text{ se e somente se } \sigma_s = \sigma_d = 1/2$$

e $-1 \leq Q \leq 1$.

- ¤ -

BIBLIOGRAFIA

Agresti,A. (1978) - "Describing differences on an ordered categorical response", *Technical Report Nº 137*, University of Florida,September, 1978.

Anderberg,M.R. (1973) - *Cluster analysis for applications*,Academic Press, New York and London.

Costner,H.L. (1965) - "Criteria for measures of association", *American Sociol. Rev., 30 (3)*, June, 1965, 341-353.

Davis,J.A. (1967) - "A partial coefficient for Goodman and Kruskal's gamma", *J.A.S.A. 62*, 189-193.

Edwards,A.W.F. (1963) - "The measure of association in a 2 × 2 table",*J. Roy. Statist. Soc. Ser.A 126, PT1*, 109-114.

Esscher,F. (1924) - "On a method of determining correlation from the ranks of variates", *Skand. Akt. 7*, 201.

Finley,J.P. (1884) - "Tornado Predictions", *The American Meteorological Journal, 1*, 85-88.

Fisher,R.A. (1921) -"On the probable error of a coefficient of correlation deduced from a small sample", *Metron, 1, Nº4*, 1.

————————(1924) -"The distribution of the partial correlation coefficient", *Metron, 3*, 329.

Goodman,L.A. & W.H.Kruskal (1954) - "Measures of association for cross classifications", *J.A.S.A., 49*, 723-764.

———————— (1963) - Measures of association for cross classifications, III: Aproximate sampling theory", *J.A.S.A.,58*, 310-364.

———————— (1972) -"Measures of association for cross classifications,IV: Simplification of assintotic variances", *J.A.S.A.,67*, 415-421.

Greiner,R. (1909) - "Über das Fehlersystem der Kollektivmasslehre", *Zeitschrift für Mathematik und Physik, 57, 121,* 225 e 337.

Hawker,R.K. (1971) - "The multivariate analysis of ordinal measures", *American Journal of Sociology, 76,* 908-926.

Hedlund,R.D. (1978) - "Cross-over Voting in a 1976 Open Presidential Primary", *Public Opinion Quarterly,41,* 498-514.

Hotelling,H. (1953) - "New light on the correlation coefficient and its transforms", *J. Roy. Statist. Soc. Ser.B, 15,* 193.

Kendall, M.G. (1938) - "A new measure of rank correlation", *Biometrika, 30,* 81.

———————— (1942) - "Partial rank correlation, *Biometrika, 32,* 277.

———————— (1945) -"The treatment of ties in raking problems", *Biometrika, 33,* 239.

———————— (1955) - *Rank correlation methods*, Second Edition,Griffin, London.

Kendall,M.G & A.Stuart (1973) - *The Advanced Theory of Statistics*, Volume 2. Third Edition, Griffin, London.

———————————— (1977) - *The Advanced Theory of Statistics*, Volume 1. Fourth Edition, Griffin, London.

Kruskal,W.H. (1958) - "Ordinal measures of association", *J.A.S.A., 53,* 814-861.

Maung,K. (1941) - "Measure of association in a contingency table with special reference to the pigmentation of hair and eye color os scottish school children", *Ann. Eugenics, 11,* 189-223.

Mosimann,J.E. (1956) - "Variation and relative growth in the plastral scutes of the turtle", *Kinosternon integrun* Leconte. *Misc.Publ. Mus. Zool. Univ. Mich. nº 97,* 43 pp.

Newman, R.W. & E.H.Munro (1955) - "The relation of climate and body si-

ze in U.S. males", *Am. J. Phys. Anthrop., N.S., 13*, 1-17.

Pearson,E.S. & H.O.Hartley (1966) - *Biometrika Tables for Statisticians*, Volume 1, 3rd. Ed., Cambridge Univ. Press, Cambridge.

Pearson,K. (1896) - "Regression, Heredity and Panmixia", *Phil.Trans.Roy. Soc. Series A, vol. 187*, 253.

———— (1904) - "On the theory of contingency and its relation to association and normal correlation", *Drapers' Company Research Memoirs, Biometric,Series I*; Dulau & Co., London.

Quade,D. (1966) - *Order association*, Notas mimeografadas da Universidade da Carolina do Norte, E.U.A.

———— (1971) "Nonparametric Partial Correlation", *Report Nº SW 13/71 of the Mathematical Center 2e Boerhaavesstraat 49*, Amsterdam, the Netherlands.

Siegel,S. (1975) - *Estatística não paramétrica para as ciências do comportamento*, trad. Alfredo Alves de Faria, McGraw-Hill do Brasil,São Paulo.

Sokal,R.R. & F.J.Rohlf (1969) - *Biometry.The Principles and Practice of Statistics in Biological Research*, W.H.Freeman and Company,San Francisco.

Somers,R.H. (1966) - "An approach to the multivariate analysis of ordinal data", *American Sociol. Rev.,33*, 971-977.

Spearman,C. (1904) - "The proof and measurement of association between two things", *Am. J. Psych., 15*, 88.

Steel,R.G.D. & J.H.Torrie (1960) - *Principles and Procedures of Statistics*, McGraw-Hill Book Co., Inc., N. York, Toronto, London.

Vanzolini, P.E. (1968) - "Environmental temperature and number of body annuli in *Amphisbaena alba*: notes on a cline (Sauria, Amphisbaenidae)" *Papéis Avulsos Zool., S.Paulo 21(23)* 231-241

Wilson,T.P. (1969) - "A proportional-reduction-in-error interpretation for Kendall's tau-b", *Social Forces, 47*, 340-342.

Yule, G.U. (1912) - "On the methods of measuring association between two atributes", *Roy. Stat. Soc.*, 75, 579-652.

Yule,G.U & M.G.Kendall (1940) - *An introduction to the theory of statistics*", Twelfth Edition, Revised, Griffin, London.

(JB)