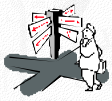

Testes Não Paramétricos

CAC (Coeficientes de Associação e Correlação)

Prof. Lorí Viali, Dr.
http://www.mat.ufrgs.br/viali/
viali@mat.ufrgs.br

Medidas de Associação, Correlação e Testes de Significância

Em muitas situações é necessário saber se dois conjuntos de dados estão relacionados e com que intensidade ocorre esta relação. Medidas destinadas a determinar o grau de relacionamento entre duas ou mais variáveis são denominadas medidas de associação (variáveis qualitativas) ou correlação (variáveis quantitativas).

Estas medidas são expressas através de um número, que geralmente varia no intervalo de -1 a 1 e são denominados de coeficientes de associação ou de correlação.

O Coeficiente de Contingência C

Conceito

O coeficiente de contingência C é uma medida associação entre dois conjuntos de atributos. É útil quando se dispõem apenas de dados apresentados em escala nominal em um ou nos dois conjuntos de atributos.

Para determinar esta medida não é necessário dispor as variáveis em uma determinada maneira. Não importa quem seja linha e quem seja coluna, o valor obtido será o mesmo.

Para calcular o coeficiente de contingência C os dados devem ser apresentados em uma tabela de contingência como a ilustrada a seguir. Os dados podem ser divididos em qualquer número de categorias, isto é, a tabela pode ser do tipo kxr, onde k = número de colunas e r = número de linhas.

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

	A_1	B_2	...	B_k	Total
B_1	x_{11}	x_{12}	...	x_{1k}	$s_{1.}$
B_2	x_{21}	x_{22}	...	x_{2k}	$s_{2.}$
	...	...	...	...	...
B_r	x_{r1}	x_{r2}	...	x_{rk}	$s_{r.}$
Total	$s_{.1}$	$s_{.2}$	...	$s_{.k}$	$s_{..}$

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

O coeficiente de contingência pode, então, ser obtido através da seguinte expressão:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Onde

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

é o qui-quadrado calculado conforme já visto.

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Exemplo

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Considere-se os valores os valores da tabela como sendo o resultado das variáveis: "Grau de instrução" (coluna) e "Procedência" (linha). Determinar o grau de associação entre as duas variáveis.

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

	Prim. Grau	Seg. Grau	Superior	Total
Capital	4	5	6	15
Interior	11	4	3	18
Outra	2	3	2	7
Total	17	12	11	40

O qui-quadrado será:

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 5{,}0989$$

O coeficiente de contingência será:

$$C = \sqrt{\frac{\chi 2}{n+\chi 2}} = \sqrt{\frac{5{,}0989}{40+5{,}0989}} = 0{,}34$$

O teste de significância para o coeficiente de contingência

Uma vez observado uma relação entre dois conjuntos de atributos em amostras, quer-se determinar se é plausível concluir pela associação desses mesmos atributos na população de onde foram retiradas as amostras.

Ao se testar a significância de uma medida de associação, está-se na realidade testando a hipótese de nulidade de que não existe associação na população, isto é, que o valor observado poderia ter ocorrido aleatoriamente entre as amostras mesmo que as populações não apresentam qualquer relação.

Para testar a hipótese de nulidade, determina-se a distribuição amostral da estatística, neste caso, a medida de associação, sob H_0. Utiliza-se, então, uma prova estatística adequada para determinar, a um nível de significância pré-fixado, se o valor observado pela estatística considerada pode ter provavelmente ocorrido sob H_0.

Embora, muitas estatísticas de associação possam ser determinadas por este método o coeficiente de contingência C, constitui um caso especial. Uma das razões por que não se pode utilizar a distribuição amostral de C para testar um determinado valor observado, reside na considerável complexidade matemática de tal procedimento.

Outra razão é que no desenvolvimento do cálculo de C, já se calcula de forma intermediária uma estatística que constituí uma indicação simples e adequada da significância de C.

Tal estatística é o χ^2. Pode-se determinar se um valor de C difere significativamente de um valor causal simplesmente determinando se um valor de χ^2 é significativo.

Para qualquer tabela de contingência kxr pode-se determinar a significância do grau de associação pela estatística C, determinando a probabilidade de ocorrência, sob H_0, de valores tão grandes quanto o valor observado de χ^2, com $gl = (k - 1)(r - 1)$.

Se essa probabilidade não supera α, pode-se rejeitar a hipótese de nulidade, àquele nível. Se o qui-quadrado baseado nos valores amostrais é significativo, pode-se concluir que, na população, a associação entre os dois conjuntos é diferente de zero.

Exemplo

No exemplo anterior foi determinado que o coeficiente de associação entre as variáveis: escolaridade e procedência é $C = 0,34$. Para chegar a este valor foi utilizado o valor $\chi^2 = 5,0989$. É este valor que vai ser usado para testar a significância de C.

Nesse caso o grau de liberdade será $gl = (3 - 1)(3 - 1) = 4$.

A significância do resultado encontrado, isto é, 5,0989 é 27,73%.

Assim não é possível afirmar que existe associação na população.

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Limitações do Coeficiente de Contingência

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

A grande aplicabilidade e a determinação relativamente fácil de C podem dar a entender que se trata de uma medida ideal de associação. Este não é o caso, no entanto, em razões das limitações desta estatística.

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Em geral, pode-se dizer que um coeficiente de associação (correlação) deve apresentar pelo menos as seguintes características:

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

- Onde houver completa falta de associação o coeficiente deve dar zero.
- Quando as variáveis são completamente dependentes entre si, isto é, estão perfeitamente relacionadas o coeficiente deve ser igual a 1.

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

O coeficiente C tem a primeira destas características, mas não a segunda. Ele é zero quando não existe associação, mas não atinge o valor um, quando a relação é perfeita, sendo esta a primeira limitação do coeficiente de contingência C.

O limite superior de C é uma função do número de categorias. Quando $k = r$, o limite superior de C, isto é, o valor que deveria ocorrer se as variáveis tivessem uma relação perfeita é:

$$\sqrt{\frac{k-1}{k}}$$

Por exemplo, o limite superior de C para uma tabela 2x2 é igual a 0,71. Para uma tabela 3x3, o máximo que C pode atingir é um valor de 0,82.

O fato de o valor máximo de C, depender de **k** e **r** é uma segunda limitação, pois dois coeficientes de contingência só serão comparáveis se provierem de tabelas com o mesmo número de linhas e colunas.

Uma terceira limitação de C é que os dados devem se prestar para o cálculo do χ^2 antes que C possa ser convenientemente utilizado, isto é, o cálculo de C sofre das mesmas limitações do cálculo do qui-quadrado.

Uma última limitação de C é que ele não é diretamente comparável com nenhuma outra medida de associação (correlação), como por exemplo, o coeficiente de Pearson, o de Spearman ou o de Kendall.

A despeito destas limitações o coeficiente de contingência é uma medida útil pela sua larga aplicabilidade, pois não exige suposições sobre a forma da população de escores, não exige continuidade da variável em estudo e requer apenas mensuração nominal.

Isto faz do coeficiente de contingência uma medida que pode ser aplicada em situações em que nenhuma outra pode ser aplicada.

Exercício

Resolva o exercício um do Laboratório Sete.

O Coeficiente V de Crámer

Considerações

Apesar de sua popularidade o coeficiente de contingência tem a desvantagem de que o número de linhas e colunas influencia o resultado. A alternativa é utilizar o coeficiente V (de Cramer), definido por:

$$V = \sqrt{\frac{\chi^2}{n.(k-1)}}$$

Onde:

n = tamanho da amostra

k = min {linhas, colunas}

Exercício

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Resolva o exercício dois do Laboratório sete.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

O Coeficiente de Correlação por postos de Spearman

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Considerações

Dentre todas as estatísticas com base em postos, o coeficiente de correlação de Spearman foi a que surgiu primeiro e é talvez a mais conhecida hoje. A sua principal vantagem é não exigir normalidade dos dados.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Esta estatística, por vezes designada "rho" (ρ), é representada, aqui por r_S. É uma medida de associação que exige que as duas variáveis tenham mensuração pelo menos ordinal para que os postos possam ser determinados.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Determinação

Suponha que existam n pares ordenados por postos representando duas variáveis. Por exemplo, um grupo de estudantes ordenado de acordo com suas notas no vestibular de uma universidade e também de acordo com sua classificação ao fim do primeiro ano.

Representando os escores do vestibular por: X_1, X_2, ..., X_n e os escores da classificação ao final do primeiro ano por: Y_1, Y_2, ..., Y_n, pode-se utilizar uma medida de correlação por postos para determinar o relacionamento entre as duas variáveis.

A correlação entre a classificação no vestibular e a classificação ao fim do primeiro ano seria perfeita se e somente se $X_i + Y_i = C$ = Constante, para todo "i". Portanto, parece lógico usar as diversas diferenças: $d_i = X_i - Y_i$ como indicativo da diferença entre os dois conjuntos de postos.

Suponha que o aluno A tenha obtido o primeiro lugar no vestibular, mas ao fim do primeiro ano esteja em sexto lugar. Neste caso, $d = 1 - 6 = -5$. Um aluno B, por outro lado, ficou em nono lugar no vestibular e agora, ao final do primeiro ano, é o segundo colocado. O valor de d para ele é então: $d = 9 - 2 = 7$.

O valor das diversas diferenças "d" fornece uma ideia do relacionamento entre as duas variáveis. Se a relação entre os dois conjuntos de postos fosse perfeita, todos os valores de "d" seriam zero. Quanto maiores os diversos valores de "d", menor será a associação entre as duas variáveis.

A utilização direta das diferenças (d) para o cálculo do coeficiente de correlação acarreta dificuldades. Por exemplo, os valores negativos e positivos se cancelam se forem somados. Por isso é utilizado o valor de d ao quadrado, d^2, para eliminar esta dificuldade.

A expressão para o cálculo do coeficiente de correlação de Spearman é baseada no cálculo do coeficiente de Pearson (estatística paramétrica) r, onde:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Onde: $x = X - \overline{X}$

$y = Y - \overline{Y}$

Mas quando X e Y são postos, $r = r_S$ *e a soma de n inteiros: 1, 2, ..., n é dada por:*

$$\sum X = \sum Y = \frac{n(n+1)}{2}$$

E a soma dos quadrados dos postos, isto é, $1^2 + 2^2 + \ldots + n^2$ *é dada por:*

$$\sum X^2 = \sum Y^2 = \frac{n(n+1)(2n+1)}{6}$$

Como: $x = X - \overline{X}$ *, então:*

$$\sum x^2 = \sum (X - \overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n} = \sum X^2 - n\overline{X}$$

Mas:

$$\sum X = \sum Y = \frac{n(n+1)}{2}$$

e:

$$\sum X^2 = \sum Y^2 = \frac{n(n+1)(2n+1)}{6}$$

Assim:

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

$$= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n^3 - n}{12} = \sum y^2$$

Mas: $d = x - y$.

Então $d^2 = (x - y)^2 = x^2 + y^2 - 2xy$

Assim:

$$\Sigma d^2 = \Sigma(x - y)^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy$$

Pela expressão do cálculo do coeficiente de correlação de Pearson, tem-se:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = r_S$$

Então: $\sum xy = r_S\sqrt{\sum x^2 \sum y^2}$

e $\sum d^2 = \sum x^2 + \sum y^2 - 2r_S\sqrt{\sum x^2 \sum y^2}$

Portanto:

$$r_S = \frac{\sum x^2 + \sum y^2 - \sum d^2}{2\sqrt{\sum x^2 \sum y^2}}$$

Substituindo Σx^2 e Σy^2 na expressão e simplificando, tem-se:

$$r_S = 1 - \frac{6\sum d^2}{n^3 - n}$$

E x e m p l o

Determinar o coeficiente de correlação de Spearman para as variáveis: X e Y do exercício três do laboratório sete.

	X	Y
1	5	6
2	9	16
3	17	18
4	1	1
5	2	3
6	21	21
7	3	7
8	29	20
9	7	15
10	100	22

	X	Y	P_X	P_Y	d_i
1	5	6	4	3	1
2	19	16	6	6	0
3	17	18	7	7	0
4	1	1	1	1	0
5	2	3	2	2	0
6	21	21	8	9	-1
7	3	7	3	4	-1
8	29	20	9	8	1
9	7	15	5	5	0
10	100	22	10	10	0
Total	---	---	---	---	0

O valor do coeficiente de correlação será então:

$$r_S = 1 - \frac{6.4}{10^3 - 10} = 0{,}9760$$

Empates

Ocasionalmente podem ocorrer empates entre os escores de dois valores na mesma variável. Quando isto ocorre, a cada um deles é atribuído a média dos postos que seriam atribuídos caso o empate não ocorresse, isto é, adota-se o procedimento usual.

Quando a proporção de empates é grande torna-se necessário a utilização de um fator de correção.

O efeito de postos empatados na variável X ou Y, reduz a soma dos quadrados. Portanto, quando houver empates é necessário corrigir a soma dos quadrados.

Neste caso: $T = \frac{t^3 - t}{12}$

Onde t = número de observações empatadas em determinado posto.

A soma dos quadrados corrigida será então:

$$\sum x^2 = \frac{n^3 - n}{12} - \sum T_X \quad e \quad \sum y^2 = \frac{n^3 - n}{12} - \sum T_Y$$

$$\sum x^2 = \frac{n^3 - n}{12} - \sum T_X \quad e \quad \sum y^2 = \frac{n^3 - n}{12} - \sum T_Y$$

$\sum T$, onde a soma de T indica o somatório sobre os vários valores de T para todos os grupos de observações empatadas.

Assim se o número de empates for considerável o cálculo do coeficiente de correlação de Spearman deve ser realizado por:

$$r_S = \frac{\sum x^2 + \sum y^2 - \sum d^2}{2\sqrt{\sum x^2 \sum y^2}}$$

Onde:

$$\sum x^2 = \frac{n^3 - n}{12} - \sum T_X \quad e \quad \sum y^2 = \frac{n^3 - n}{12} - \sum T_Y$$

Teste para o Coeficiente de Correlação de Spearman

Se as amostras utilizadas no cálculo do coeficiente de correlação de Spearman foram selecionadas aleatoriamente, então pode-se utilizar os seus valores para testar se as variáveis correspondentes estão associadas na população, isto se r_S pode ser considerado diferente de zero.

Pequenas Amostras

Suponha verdadeira a hipótese de nulidade, isto é, suponha-se que $\rho_S = 0$. Se as amostras são aleatórias, então para uma dada ordem dos escores de X, todas as ordens possíveis dos escores Y tem a mesma probabilidade.

Para n valores existem n! ordenações possíveis dos escores X que podem ocorrer com qualquer ordenação dos escores Y. Como essas ordenações são igualmente prováveis, a probabilidade de ocorrência de determinada ordenação dos escores X conjuntamente com dada ordenação dos escores Y é 1 / n!.

A cada uma das possíveis ordenações de Y está associado um valor de r_S. A probabilidade de ocorrência, sob H_0, de cada valor de r_S é então proporcional ao número de permutações que originam aquele valor.

Aplicando a fórmula do cálculo do r_S pode-se perceber que:

Se n = 2, então r_S só pode assumir os valores -1 e +1. Cada um destes valores tem probabilidade 1/2.

Se n = 3, então os possíveis valores de r_S são -1, -1/2, +1/2 e +1. Cada um destes valores tem probabilidade de ocorrência, sob H_0, respectivamente de: 1/6, 1/3, 1/3 e 1/6.

A tabela P (Siegel, pg. 315) fornece os valores críticos unilaterais de r_S, obtidos por este método. Para n variando de 4 a 30, a tabela fornece o valor de r_S com a probabilidade associada, sob H_0, para p = 0,05, e p = 0,01.

Exemplo

Suponha que 12 pares das variáveis X e Y forneceram um coeficiente de correlação r_S = 0,82. Verifique se é possível afirmar que esse valor é significativamente maior do que zero a uma probabilidade de 1%.

Pela tabela P vê-se que esse valor é significativo ao nível $p < 0,01$ (teste unilateral). Pode-se então rejeitar a hipótese concluindo que, na população estudada, as duas variáveis estão positivamente associadas.

Grandes Amostras

Quando n é 10 ou mais, a significância de um valor obtido de r_S, sob a hipótese de nulidade, pode ser comprovado através de (Kendall, 1948):

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

O Coeficiente de Correlação por Postos de Kendall

Conceito

O coeficiente de correlação por postos de Kendall, τ (tau) é uma medida de associação para variáveis ordinais. Neste caso, τ dará uma medida do grau de associação entre os dois conjuntos de postos.

A distribuição amostral de τ, sob H_0 é conhecida e pode, portanto ser testada. Uma vantagem de τ sobre o coeficiente r_S é que τ pode ser generalizado para um coeficiente de correlação parcial que será visto posteriormente.

Suponha-se que se peça a dois juízes X e Y, para atribuir postos a quatro objetos. Por exemplo, poderíamos solicitar que classificassem quatro ensaios por ordem de qualidade de estilo.

Represente-se os quatro ensaios por a, b, c e d. Os postos obtidos foram:

Ensaio	a	b	c	d
Juiz X	3	4	2	1
Juiz Y	3	1	4	2

Reordenando os ensaios, de forma que os postos atribuídos pelo juiz X apareçam na ordem natural (1, 2, ..., n), tem-se:

Ensaio	a	b	c	d
Juiz X	1	2	3	4
Juiz Y	2	4	3	1

Temos agora condições de determinar o grau de correspondência entre os julgamentos de X e de Y. Os postos atribuídos pelo juiz X já estando na ordem natural, passa-se a determinar quantos pares de postos atribuídos pelo juiz Y se acham em sua ordem correta (natural) em relação ao outro.

Considera-se primeiro todos os pares de postos em que figura o posto 2 do juiz Y - o posto mais à esquerda em seu conjunto. O primeiro par, 2 e 4, está na ordem correta, isto é, 2 precede 4. Como a ordem é "natural", atribui-se o escore +1 a este par.

Os postos 2 e 3 constituem o segundo par, que também está na ordem correta (o 2 vem antes do 3), recebendo, assim, também o escore +1. O terceiro par consiste dos postos 2 e 1.

Esses escores não estão na ordem "natural", pois 2 não vem antes do 1. Atribui-se então ao par o escore -1. O total dos escores de todos os pares de postos que incluem o posto 2 é: +1 + 1 - 1 = 1.

Considera-se, em seguida, todos os pares possíveis de postos que incluem o posto 4 (segundo posto do juiz Y a contar da esquerda) e um outro posto que o segue. Um par é o 4 e 3 cujos elementos não estão em ordem, recebendo, por isso, o escore -1. O total destes escores é: -1 -1 = -2.

Considerando agora o posto 3 e os seguintes, obtém-se um único par: 3 e 1, cujos elementos não estão em ordem natural; o par recebe o escore -1. O total de todos os escores assim atribuídos é: 1 - 2 -1 = -2.

Qual é o total máximo possível que se pode obter para os escores atribuídos a todos os pares de postos do juiz Y?

Obter-se-ia o total máximo se os postos dos juízes X e Y tivessem apresentado perfeita concordância, porque então, colocados os postos de X em sua ordem natural, cada par de postos do juiz Y se apresentaria também na ordem natural, recebendo, assim, o escore +1.

O total máximo possível, no caso de uma concordância perfeita entre X e Y, seria 6.

O grau de relacionamento entre os dois conjuntos de postos é dado pela razão do total efetivo de escores + 1 e -1, para o total máximo possível.

O coeficiente de correlação por postos de Kendall é a razão:

τ = *(total efetivo) / (total máximo possível)* = -2 / 6 = - 0,33.

Isto é, τ = *-0,33 é uma medida da concordância entre os postos atribuídos aos ensaios pelos juízes X e Y.*

Pode-se considerar τ *como função do número mínimo de inversões ou permutas entre elementos vizinhos, necessário para transformar um posto em outro. Este coeficiente é uma espécie de coeficiente de desordenamento.*

Método

Viu-se que:

τ = *(total efetivo) / (total máximo possível)*

Em geral, o escore máximo possível será:

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

Anotando por S a soma dos escores +1 e -1 para todos os pares, tem-se:

$$\tau = \frac{S}{\frac{n(n-1)}{2}} = \frac{2S}{n(n-1)}$$

Onde n = número de pares envolvidos.

O cálculo de S pode ser abreviado da seguinte forma:

Após colocados em sua ordem natural os postos do juiz X, os postos correspondentes do Juiz Y se apresentam na seguinte ordem:

Juiz Y: 2 4 3 1

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Pode-se determinar o valor de S partindo do primeiro número à esquerda e contando o número de postos à sua direita que são superiores. Deste número subtrai-se o número de postos à direita que são inferiores. Procedendo desta forma para todos os postos e somando os resultados se obtém S.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Assim, para os valores acima, os postos à direta de 2 e superiores a 2 são 3 e 4, e o 1 é inferior. O posto 2 contribuí, então, com 2 - 1 = 1 para o valor de S.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Para o posto 4 existe 0 valores superiores e dois inferiores, então sua contribuição é: 0 - 2 = -2. Para o posto 3, existe à direita apenas um inferior, então sua contribuição para S é 0 - 1 = -1 .

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

O total destas contribuições é então de:

1 - 2 - 1 = -2 = S.

Conhecido S pode-se aplicar a expressão para o cálculo do coeficiente τ para os postos atribuídos pelos dois juízes:

$$\tau = \frac{S}{\frac{n(n-1)}{2}} = \frac{2S}{n(n-1)} = \frac{2(-2)}{4(4-1)} = \frac{-4}{12} = -0{,}33$$

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Exemplo

Abaixo as variáveis autoritarismo e aspirações de status social para 12 estudantes. Calcular o valor de τ para os dados.

1	2	3	4	5	6	7	8	9	10	11	12
3	4	2	1	8	11	10	6	7	12	5	9
3	6	5	1	10	9	8	3	4	12	7	11

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Para calcular τ vamos reordenar os estudantes de modo que o primeiro conjunto de postos se apresente na ordem natural:

1	2	3	4	5	6	7	8	9	10	11	12
1	2	3	4	5	6	7	8	9	10	11	12
1	5	2	6	7	3	4	10	11	8	9	12

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

1	2	3	4	5	6	7	8	9	10	11	12
1	2	3	4	5	6	7	8	9	10	11	12
1	5	2	6	7	3	4	10	11	8	9	12

Dispostos em sua ordem natural os postos de X, determinamos o valor de S para os postos de Y:

$$S = (11 - 0) + (7 - 3) + (9 - 0) + (6 - 2) + (5 - 2) + (6 - 0) + (5 - 0)(2 - 2) + (1 - 2) + (2 - 0) + (1 - 0) = \mathbf{44}$$

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

O posto relativo a autoritarismo mais à esquerda é 1. Este posto tem 11 postos superiores a sua direita e nenhum que lhe seja inferior. Sua contribuição para S é, pois, (11 - 0). O posto 5 contribui com (7 - 3) para S , pois a sua direita existem 7 superiores e a sua esquerda estão 3 postos que lhe são inferiores. E assim por diante.

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Sabido que S = 44 e n = 12, aplica-se então a expressão do cálculo do coeficiente.

$$\tau = \frac{S}{\frac{n(n-1)}{2}} = \frac{2S}{n(n-1)} = \frac{2.44}{12(12-1)} = \frac{88}{132} = 0{,}67$$

Esse valor representa o grau de relacionamento entre o autoritarismo e as aspirações de status social dos 12 estudantes.

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Empates

Quando há empate entre duas ou mais observações de X ou de Y, atribui-se as observações empatadas a média dos postos que lhes caberiam se não houvesse empate.

O efeito dos empates consiste em modificar o denominador da fórmula de τ.

Assim, a expressão para o cálculo do coeficiente quando ocorrem empates é:

$$\tau = \frac{S}{\sqrt{\frac{1}{2}n(n-1) - T_x}\sqrt{\frac{1}{2}n(n-1) - T_y}}$$

Onde $T_X = ½\sum t(t - 1)$ e $T_Y = ½\sum t(t - 1)$, onde t é o número de observações empatadas em cada grupo de empates nas variáveis X e Y.

Teste para o Coeficiente Tau de Kendall

Se uma amostra aleatória for extraída de uma população em que X e Y não estão relacionados e se atribuem aos elementos da amostra postos relativos à X e Y, então, para uma dada ordem de postos de X todas as ordens possíveis de postos de Y são igualmente verossímeis.

Isto é, para uma dada ordem dos postos de X, qualquer ordem de Y tem a mesma probabilidade de ocorrência que qualquer outra ordem. Suponhamos os valores de X dispostos na ordem natural 1, 2, 3, ..., n.

Para tal ordenação dos postos de X, todas as $n!$ ordens possíveis dos postos de Y são igualmente prováveis sob H_0. Portanto, qualquer ordenação em particular dos postos de Y tem probabilidade $1/n!$ de ocorrência, sob H_0.

Probabilidades de τ, sob H_o para $n = 4$.

Valor de τ	Frequência de ocorrência sob H_0	Probabilidade de ocorrência sob H_o
-1,00	1	1/14
-0,67	3	3/14
-0,33	5	5/24
0,00	6	6/24
0,33	3	5/24
0,67	5	3/24
1,00	1	1/24

A cada uma das n! disposições possíveis de postos de Y acha-se associado um valor de τ. Esses valores possíveis do índice variarão de +1 a - 1 e podem ser dispostos em uma distribuição de frequências.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Por exemplo, para n = 4, há 4! = 24 ordenações possíveis dos postos de Y e a cada uma delas está associado um valor de τ. A tabela anterior fornece a frequência de ocorrência sob H_0. A medida que n cresce é cada vez mais trabalhoso construir as distribuições.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

A medida que n cresce a distribuição de τ tende para uma normal de nédia $\mu_\tau = 0$ e desvio padrão dado por:

$$\sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$$

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

O Coeficiente de Correlação Parcial de Postos de Kendal

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Quando se observa uma correlação entre duas variáveis existe sempre a possibilidade de que tal correlação seja devida à associação de cada uma das duas variáveis com uma terceira variável.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Por exemplo, em um grupo de pessoas de diversas idades, pode-se verificar uma alta correlação entre a amplitude do vocabulário e a altura.

Tal correlação, entretanto, pode não refletir um relacionamento verdadeiro ou direto entre as duas variáveis mas resultado do fato de que tanto a amplitude do vocabulário quanto a altura estão relacionados com uma terceira variável a idade.

Problemas deste tipo podem ser abordados através da determinação de um coeficiente de correlação parcial. Na correlação parcial os efeitos de uma terceira variável Z sobre as variáveis X e Y são controlados mantendo-a constante.

Ao planejar o experimento, pode-se adotar dois caminhos. Introduzir controles experimentais com o propósito de eliminar a influência da terceira variável ou utilizar métodos estatísticos para eliminar tal influência.

Por exemplo, para se estudar a relação entre a capacidade de memorização e a capacidade para resolver certos tipos de problemas será necessário controlar o efeito das diferenças de inteligência.

Uma alternativa é escolher pessoas com o mesmo nível de inteligência. Se isto não for possível, pode-se aplicar então o controle estatístico.

Com a correlação parcial o efeito da inteligência sobre a relação entre memorização e capacidade de resolução de problemas poderá ser determinada de forma direta ou não-contaminada.

Suponha que os postos de 4 pessoas em relação a 3 variáveis X, Y e Z foram obtidos. Deseja-se determinar a correlação entre X e Y quando Z é controlada.

Pessoas	a	b	c	d
Posto de Z	1	2	3	4
Posto de X	3	1	2	4
Posto de Y	2	1	3	4

Para cada uma das variáveis sabe-se que há $\binom{4}{2}$ pares de postos possíveis. Colocados os postos de Z em sua ordem natural, observa-se cada par de postos possível em X, Y e Z. Atribuí-se o sinal + aos pares em que o posto mais baixo precede o posto mais alto e um sinal – caso contrário.

Suponha que os postos de 4 pessoas em relação a 3 variáveis X, Y e Z foram obtidos. Deseja-se determinar a correlação entre X e Y quando Z é controlada.

Par	(a, b)	(a, c)	(a, d)	(b, c)	(b, d)	(c, d)
Z	+	+	+	+	+	+
X	-	-	+	+	+	+
Y	-	+	+	+	+	+

As informações obtidas são resumidas em uma tabela 2x2.

Sinal do Par	+	-
+	(+, +)	(+, -)
-	(-, +)	(-, -)

No primeiro par (a, b) tanto X quanto Y discordam do sinal de Z então a frequência vai para a célula D (-, -). No segundo par (c, d) Y concorda com Z mas X não. A frequência é registrada na célula C (-, +).

Os pares restantes apresentam todos o mesmo sinal e portanto a frequência vai para a célula A (+, +). Em resumo, tem-se:

Sinal do Par	+	-	Total
+	4	0	4
-	1	1	2
Total	5	1	6

O coeficiente de correlação por postos de Kendall entre duas variáveis (X, Y) considerando constante uma terceira variável (Z) é dado então por:

$$\tau_{XY.Z} = \frac{AD - BC}{\sqrt{(A+D)(C+D)(A+C)(B+D)}}$$

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Para os dados sendo analisados, tem-se:

$$\tau_{XY.Z} = \frac{AD - BC}{\sqrt{(A+D)(C+D)(A+C)(B+D)}} = \frac{4.1 - 0.1}{\sqrt{(4+0)(1+1)(4+1)(0+1)}} = \frac{4}{\sqrt{4.2.5.1}} = \frac{4}{\sqrt{40}} = \sqrt{0,4} = 0,6325$$

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

A correlação entre X e Y, com o efeito de Z constante é então: $\tau_{XY.Z}$ = 0,63. Se fosse calculado a correlação entre X e Y sem considerar Z o resultado seria: τ = 4/6 = 0,67.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

A expressão para o cálculo do coeficiente de correlação parcial por postos de Kendall é algumas vezes denominada de "Coeficiente Phi" e pode-se mostrar que:

$$\tau_{XY.Z} = \sqrt{\frac{\chi^2}{n}}$$

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

A maneira de calcular o CCPPK não é prática quando n é grande. Nesse caso, pode-se utilizar a seguinte expressão alternativa devida a Kendall:

$$\tau_{XY.Z} = \frac{\tau_{XY} - \tau_{XZ}\tau_{YZ}}{\sqrt{(1-\tau_{XZ}^2)(1-\tau_{YZ}^2)}}$$

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

O teste de associação linear ou Qui-quadrado de Mantel-Haenszel

X_i Y_j	1	2	...	k	Total
1	f_{11}	f_{12}	...	f_{1k}	r_1
2	f_{21}	f_{22}	...	f_{2k}	r_2
...	...	...	...	...	...
l	f_{l1}	f_{l2}	...	f_{lr}	r_l
Total	c_1	c_2	...	c_k	W

O qui-quadrado de Mantel-Haenszel também denominado de teste qui-quadrado de associação linear por linear é uma medida de significância para variáveis ordinais. Ele é utilizado para testar a significância do relacionamento linear entre duas variáveis ordinais, porque é mais poderoso do que o qui-quadrado de Pearson.

O qui-quadrado de Mantel-Haenzel não é adequado para variáveis nominais. Se ele for significativo então é possível dizer que o aumento de uma variável está associado com o aumento (ou decréscimo, para relacionamentos negativos) da outra variável.

Como outras estatísticas que utilizam o qui-quadrado ele não deve ser utilizado com valores baixos de frequências.

O teste de associação linear de Mantel-Haenszel é dado por:

$$X^2_{MH} = (W - 1)r^2$$

onde r é o coeficiente de correlação de Pearson definido conforme o apresentado a seguir. O grau de liberdade da estatística é 1.

O algoritmo para o cálculo do coeficiente de correlação de Pearson para uma tabela de contingência é dado por:

$$r = \frac{cov(X,Y)}{\sqrt{S_X S_Y}}$$

Onde:

$$Cov(X,Y) = \sum x_i y_j f_{ij} - \left(\sum_{i=1}^{\ell} x_i r_i\right)\left(\sum_{j=1}^{k} y_j j\right) / W$$

e:

$$S_X = \sum_{i=1}^{\ell} x_i^2 r_i - \left(\sum_{i=1}^{\ell} x_i r_i\right)^2 / W$$

$$S_X = \sum_{j=1}^{k} y_j^2 c_j - \left(\sum_{j=1}^{k} y_j c_j\right)^2 / W$$

E x e m p l o

	1	2	3	Total
1	4	5	6	15
2	11	4	3	18
3	2	3	2	7
Total	17	12	11	40

Onde:

$$Cov(X,Y) = \sum x_i y_j f_{ij} - \left(\sum_{i=1}^{\ell} x_i r_i\right)\left(\sum_{j=1}^{k} y_j j\right) / W = -3,20$$

e:

$$S_X = \sum_{i=1}^{\ell} x_i^2 r_i - \left(\sum_{i=1}^{\ell} x_i r_i\right)^2 / W = 20,40$$

$$S_X = \sum_{j=1}^{k} y_j^2 c_j - \left(\sum_{j=1}^{k} y_j c_j\right)^2 / W = 27,10$$

$W = 40.$

Portanto:

$$\chi^2_{MH} = (W - 1)r^2 = (40 - 1).(-0,14)^2$$

$$= 0,7224.$$

http://www.statsguides.bham.ac.uk/

Guias do SPSS 9, 10 e Minitab 12.

http://www.uc.edu/sashtml/proc/zompmeth.htm

Fórmulas de vários tipos de coeficientes

http://www.nyu.edu/its/socsci/Docs/correlate.html

Correção de empates para o cc de Spearman

http://www.uc.edu/sashtml/stat/chap28/sect20.htm

SHESKIN, David J. **Handbook of Parametric and Nonparametric Statistical Procedures**. 4th ed. Boca Raton (FL): Chapman & Hall/CRC, 2007.