O Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é baseado na idéia de variância, dada na aula 6. Como visto naquela aula, quando temos uma amostra composta por *n* dados, a variância da amostra é dada por,

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1},$$

onde $\bar{x}$ é a média dos *n* valores. A variância é uma medida quadrática da dispersão dos *n* dados em torno da sua média. Um valor grande de s^2 indica que os dados estão localizados a grandes distâncias da média, enquanto que um valor pequeno indica que eles estão localizados a pequenas distâncias.

Observando a fórmula para a variância, vemos que ela pode ser escrita como,

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)}{n-1},$$

ou seja, ela é dada pela somatória do *produto* de dois termos iguais, o desvio de cada dado x_i em relação à media.

A variância é uma grandeza estatística usada quando se trabalha com apenas uma variável *X*. Quando se trabalha com duas variáveis, *X* e *Y*, um tipo de pergunta que se costuma fazer é: quando os valores x_i da variável *X* variam em relação à sua média $\bar{x}$, ficando acima ou abaixo dela, como se comportam os valores y_i da variável *Y* em relação à sua média $\bar{y}$?

Dentre as várias possibilidades para esta pergunta, há três casos importantes:

1. Quando um valor x_i da variável *X* varia em relação à sua média $\bar{x}$ ficando acima dela, o valor correspondente y_i da variável *Y* também varia em relação à sua média $\bar{y}$ ficando acima dela; e quando um valor x_i da variável *X* varia em relação à sua média $\bar{x}$ ficando abaixo dela, o valor correspondente y_i da variável *Y* também varia em relação à sua média $\bar{y}$ ficando abaixo dela.

2. Quando um valor x_i da variável X varia em relação à sua média $\bar{x}$ ficando acima dela, o valor correspondente y_i da variável Y também varia em relação à sua média $\bar{y}$, mas ficando abaixo dela; e quando um valor x_i da variável X varia em relação à sua média $\bar{x}$ ficando abaixo dela, o valor correspondente y_i da variável Y também varia em relação à sua média $\bar{y}$, mas ficando acima dela.
3. Os valores y_i da variável Y variam em relação à sua média $\bar{y}$ de forma completamente independente da variação dos valores x_i da variável X em relação à sua média $\bar{x}$.

Nos dois primeiros casos, dizemos que as variáveis X e Y **co-variam**, pois a variação de uma em relação à sua média está associada à variação da outra em relação à sua média. No primeiro caso, dizemos que as variáveis X e Y têm **covariância positiva** e no segundo caso dizemos que elas têm **covariância negativa**. No terceiro caso, dizemos que as duas variáveis têm **covariância nula**.

A covariância entre duas variáveis, X e Y, é *quantificada* por uma fórmula similar à da variância:

$$\mathrm{COV}(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

Note que ela é idêntica à fórmula para o cálculo da variância, só que agora ele considera os desvios das duas variáveis, X e Y, em relação às suas respectivas médias.

Para entender o significado da covariância, considere um exemplo em que a variável X varia entre −1 e +1 (com média 0) e a variável Y varia entre −2 e +2 (também com média 0).

Suponhamos que tenham sido feitas 5 pares de medidas para as duas variáveis e que os resultados sejam os dados pela tabela abaixo:

i	X	Y
1	+1	+2
2	–1	–2
3	+0,5	+1
4	–0,5	–1
5	0	0

Em tal caso, o valor da covariância de X e Y é:

$$s_{XY} = \frac{1}{4}\Big((1-0)(2-0) + (-1-0)(-2-0) + (0{,}5-0)(1-0) + (-0{,}5-0)(-1-0) + (0-0)(0-0)\Big) = \frac{5}{4} = 1{,}25.$$

Este é um valor *positivo*. Note que ele é positivo porque *todos* os termos da soma acima são positivos (ou nulos). Isto ocorre porque sempre que um valor de X está abaixo da média o valor correspondente de Y também está abaixo da média. Portanto, os desvios dos dois em relação às suas respectivas médias são negativos e o produto deles dá um termo positivo. Da mesma forma, quando um valor de X está acima da média o valor correspondente de Y também está acima da média e o produto dos dois desvios em relação à média é positivo.

Suponhamos agora um caso em que os 5 pares de medidas fossem os seguintes:

i	X	Y
1	+1	–2
2	–1	+2
3	+0,5	–1
4	–0,5	+1
5	0	0

Neste caso, o valor da covariância de X e Y é:

$$s_{XY} = \frac{1}{4}\Big((1-0)(-2-0) + (-1-0)(2-0) + (0{,}5-0)(-1-0) + (-0{,}5-0)(1-0) + (0-0)(0-0)\Big) = -\frac{5}{4} = -1{,}25.$$

O valor da covariância é negativo (e tem o mesmo módulo que o anterior). Isto ocorre porque, agora, quando um valor de X está acima da média o valor correspondente de Y está abaixo da média e vice-versa. Portanto, os desvios que aparecem em cada termo terão sempre os sinais trocados e seus produtos serão negativos.

Exercício: Gere outros casos com 5 pares de valores para as variáveis X e Y do exemplo acima e calcule os valores das covariâncias para cada um deles. Faça um auto-teste para ver se você consegue prever que valor a covariância terá em cada um dos casos. Tente encontrar as situações em que essa covariância tem o máximo valor positivo e o mínimo valor negativo. Os módulos da covariância nesses dois casos são iguais ou diferentes? Ao fazer esses experimentos, você acabará por adquirir um entendimento (um *insight*) sobre o significado da covariância.

Se você de fato fez o exercício acima, você notou que a covariância de X e Y varia entre dois valores iguais em módulo, um positivo e outro negativo. Além disso, se fossemos dar uma unidade ao valor da covariância, ela seria igual ao produto da unidade da variável X pela unidade da variável Y (por exemplo, nos casos dos exemplos de 1 a 3 da aula 8 as unidades da covariância seriam, respectivamente, pontosXjogos, segundosXjogos e palavrasXjogos).

É interessante definir um índice para medir o grau de variação conjunta entre as variáveis X e Y que tenha duas características: (a) sua faixa de variação esteja limitada ao intervalo entre -1 e $+1$; e (b) seja adimensional, ou seja, não tenha unidades. Esse índice pode ser obtido a partir da covariância de X e Y dividindo-a pelo produto dos desvios padrões das variáveis X e Y:

$$r \equiv \frac{s_{XY}}{s_X s_Y} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_Y}.$$

Este índice é chamado de **coeficiente de correlação de Pearson**.

Exercício: Volte aos exemplos estudados acima e calcule, para cada um deles, o valor de *r* segundo a fórmula dada acima. Verifique que os valores de *r* estão sempre no intervalo entre –1 e +1.

Dica: Para fazer o exercício acima de uma maneira mais rápida, implemente no Excel uma planilha como a mostrada abaixo. A primeira coluna contém os índices das variáveis *X* e *Y*. A segunda coluna contém os valores da variável *X* e, logo abaixo, a média e o desvio padrão desses valores. A terceira coluna contém a mesma coisa que a coluna 2, só que para a variável *Y*. A quarta coluna contém os valores dos desvios $(x - \bar{x})$. A quinta coluna contém os valores dos desvios $(y - \bar{y})$. Finalmente, a sexta coluna contém os valores dos produtos dos desvios e logo abaixo, o valor de *r* calculado pela fórmula acima.

	A	B	C	D	E	F
1	i	x	y	(x-xmed)	(y-ymed)	(x-xmed)(y-ymed)
2		1	1	2	1	2	2
3		2	-1	-2	-1	-2	2
4		3	0,5	1	0,5	1	0,5
5		4	-0,5	-1	-0,5	-1	0,5
6		5	0	0	0	0	0
7	media		0	0	r		1
8	desvpad		0,790569	1,581139			

Os comandos usados para gerar a planilha acima estão mostrados abaixo:

	A	B	C	D	E	F
1	i	x	y	(x-xmed)	(y-ymed)	(x-xmed)(y-ymed)
2		1	1	2	=B2-B7	=C2-C7	=D2*E2
3		=A2+1	-1	-2	=B3-B7	=C3-C7	=D3*E3
4		=A3+1	0,5	1	=B4-B7	=C4-C7	=D4*E4
5		=A4+1	-0,5	-1	=B5-B7	=C5-C7	=D5*E5
6		=A5+1	0	0	=B6-B7	=C6-C7	=D6*E6
7	media	=MÉDIA(B2:B6)	=MÉDIA(C2:C6)		r	=SOMA(F2:F6)/(4*(B8*C8))	
8	desvpad	=DESVPAD(B2:B6)	=DESVPAD(C2:C6)				

Olhando para a fórmula do coeficiente de correlação de Pearson, vemos que ela não é difícil de ser calculada, mas é tediosa – especialmente para um *n* grande – e erros podem ser cometidos caso se queira calculá-la à mão. Desta forma, recomendo que vocês usem sempre um programa como o Excel para o cálculo de *r*.

Caso não se tenha acesso a um computador, mas a uma calculadora científica que calcule o desvio-padrão pode-se usar a fórmula anterior. Caso só se tenha acesso a uma calculadora comum, então não há jeito e deve-se fazer o cálculo à mão. Nesse último caso, recomendo que se use a fórmula dada a seguir, obtida por uma expansão dos termos quadráticos que aparecem na fórmula para *r*.

-----Dedução de uma fórmula alternativa para *r* (você pode pular este pedaço se quiser)----

O desvio padrão de uma variável *X* é,

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

de maneira que o produto dos desvios padrões de *X* e *Y* vale,

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

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

Substituindo esta expressão na fórmula para *r*, temos:

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

Expandindo os termos no numerador e no denominador:

$$r = \frac{\sum_{i=1}^{n}(x_i y_i) - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y}}{\sqrt{\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2\right].\left[\sum_{i=1}^{n}y_i^2 - 2\bar{y}\sum_{i=1}^{n}y_i + n\bar{y}^2\right]}} = \frac{\sum_{i=1}^{n}(x_i y_i) - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y}}{\sqrt{\left[\sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right].\left[\sum_{i=1}^{n}y_i^2 - \frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}\right]}} \Rightarrow$$

$$\Rightarrow r = \frac{n\sum_{i=1}^{n}(x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right].\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}.$$

----------Fim da dedução da fórmula alternativa para *r*----------

Talvez você não se convença de que a fórmula acima permite um cálculo feito à mão mais rápido de *r* que a fórmula anterior. A única maneira de se convencer disso é fazendo uma experiência à mão. Escolha um dos exemplos dados anteriormente e calcule *r* segundo as duas fórmulas, usando papel e lápis. Meça os tempos gastos para o cálculo de *r* pelas duas fórmulas e verifique por você mesmo qual é menor (avalie também, de forma subjetiva, qual das duas maneiras é a menos tediosa; verifique se as duas medidas se correlacionam!).

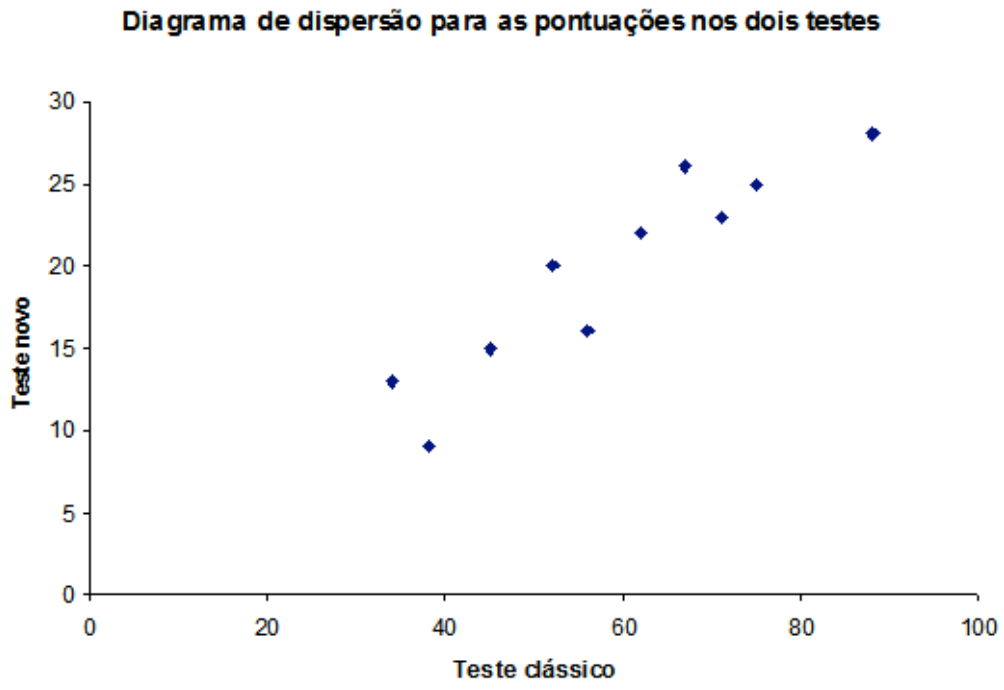
Vamos agora, para reforçar, fazer mais um exercício do cálculo do coeficiente de correlação de Pearson, só que usando o Excel.

Exemplo 6: Suponha que exista um teste psicológico clássico e de validade comprovada para a avaliação do desempenho de crianças em ler textos em voz alta. Vamos supor que um pedagogo tenha proposto um novo teste para avaliar esse desempenho. Uma maneira de verificar se o novo teste oferece uma avaliação tão boa quanto o teste clássico é ver como eles se correlacionam.

Considere que os dois testes tenham sido aplicados a uma mesma amostra composta por 10 crianças. As pontuações obtidas por elas estão dadas na tabela abaixo (vamos supor que o teste clássico atribua escores entre 0 e 100 e que o teste novo atribua escores entre 0 e 30).

	Pontuação no teste	
Criança	**Teste Clássico**	**Teste Novo**
1	67	26
2	71	23
3	45	15
4	56	16
5	62	22
6	38	9
7	52	20
8	75	25
9	88	28
10	34	13

Usando o Excel, implemente uma planilha como a do exemplo anterior e calcule r (arredonde-o até a segunda casa decimal). Você deve obter o valor $r = 0{,}93$. Isto indica que existe uma correlação muito forte entre os dois testes. A verificação disso pode ser feita traçando-se o diagrama de dispersão para as duas variáveis:



Observe que os pontos não caem todos exatamente sobre uma linha reta como nos exemplos anteriores. Isto decorre do fato de que a correlação não é perfeita (r não é exatamente igual a 1).

O Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman, denotado por r_S ou ρ, é usado quando os valores das variáveis X e Y são ordenados por ranking (ou postos). Ele mede a correlação entre as posições dos pares (x, y) nos *rankings* das variáveis X e Y.

Assim como o coeficiente de correlação de Pearson, o coeficiente de correlação de Spearman r_S varia entre -1 e $+1$ e é interpretado da mesma maneira (-1 indicando correlação negativa perfeita e $+1$ indicando correlação positiva perfeita).

Para ordenar os valores de uma variável por ranking, dá-se ao valor mais alto a posição 1, ao segundo maior valor a posição 2, ao terceiro maior valor a posição 3, etc. Caso haja valores iguais, dá-se a cada um deles o valor da média das posições que seriam ocupadas por eles no ranking, calculada somando-se essas posições e dividindo-se pelo número de valores iguais.

Para mostrar como calcular o coeficiente de correlação de Spearman, vamos dar um exemplo. Suponha que foram coletadas as notas das provas finais de português e de matemática de uma amostra de dez crianças de uma escola. Os dados estão apresentados na tabela abaixo.

Criança	**Nota de português (x)**	**Nota de matemática (y)**	**Posição no ranking de português**	**Posição no ranking de matemática**	**Diferença entre as posições (d)**	d^2
1	6,5	7,2	4	4,5	−0,5	0,25
2	3,0	6,0	9	6	3	9
3	8,3	7,2	1	4,5	−3,5	12,25
4	8,0	5,0	2	7	−5	25
5	5,5	8,5	7	1	6	36
6	6,4	4,8	5	8	−3	9
7	7,0	4,5	3	9	−6	36
8	2,5	4,0	10	10	0	0
9	5,1	7,6	8	3	5	25
10	5,8	7,9	6	2	4	16

As primeiras três colunas dão, respectivamente, o índice da criança (de 1 a 10), a sua nota em português e a sua nota em matemática. As outras duas colunas dão, respectivamente, as posições de cada criança nos rankings das notas de português e de matemática. Note que as crianças de números 1 e 3 tiraram a mesma nota em matemática. Elas deveriam ocupar as posições de números 4 e 5, de maneira que as duas receberam a posição média 4,5 (isto é, (4+5)/2).

As últimas duas colunas são usadas para o cálculo de r_S. A primeira dá as diferenças entre as posições das crianças nos rankings de português e de matemática e a segunda dá os quadrados dessas diferenças. A fórmula para o cálculo do coeficiente de correlação de Spearman é a seguinte:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d^2}{n(n^2 - 1)}.$$

Aplicando a fórmula acima aos dados da tabela, obtêm-se o valor $r_S = -0{,}02$. Este valor – muito próximo de zero – indica que não há correlação entre as posições rankeadas dos alunos em português e em matemática.

De onde vem esta fórmula? Se você fizer o calculo da correlação entre as posições nos rankings das notas de português e de matemática usando a fórmula do coeficiente de correlação de Pearson, você verá que o valor obtido para r será praticamente o mesmo que o obtido acima para r_S. Na realidade, o coeficiente de correlação de Spearman é o coeficiente de correlação de Pearson, só que aplicado às posições das variáveis (X, Y) nos seus respectivos rankings. A fórmula de Spearman é apenas uma maneira mais fácil de calcular o coeficiente de correlação de Pearson neste caso.

Nota: mesmo quando se calcula a correlação entre as posições nos rankings dos valores de X e Y usando-se a fórmula do coeficiente de correlação de Pearson, ainda assim chama-se o resultado do cálculo de coeficiente de correlação de Spearman (indicado por r_S).

Em resumo, o coeficiente de correlação de Spearman (r_S ou ρ) indica o grau de correlação entre duas variáveis, X e Y, usando suas posições nos seus respectivos rankings, ao invés dos seus valores reais. O valor de r_S é o mesmo que seria obtido caso se calculasse o coeficiente de correlação de Pearson para as posições nos rankings, só que a fórmula de Spearman é mais fácil de ser usada nesse caso.

Significância do Coeficiente de Correlação

Na aula 8, falamos que há duas coisas importantes que se deve saber sobre um coeficiente de correlação. Uma é a sua força, que é dada pelo módulo de r (ou r_s). A outra é a sua significância. Não temos condições de mostrar para vocês neste curso como fazer um teste matemático para avaliar a significância do coeficiente de correlação. O que vamos fazer, portanto, é apresentar o *conceito* de significância do coeficiente de correlação. Espera-se que com isso vocês apreciem a importância de se saber a significância de um coeficiente de correlação (independentemente de quão forte ou fraco ele seja).

Para apresentar o conceito de significância, consideremos o seguinte exemplo. Suponha que a Comissão de Graduação (CG) da Universidade tenha pedido a duas turmas diferentes do curso de pedagogia, por exemplo, a turma do segundo ano e a turma do terceiro ano, para avaliar três professores do curso. Vamos supor que cada turma tenha feito duas disciplinas diferentes com cada professor, de maneira que os alunos tenham um bom conhecimento deles e de suas capacidades didáticas. A avaliação consiste de questionários distribuídos entre os alunos, cobrindo temas como dedicação do professor, capacidade de explicação, conhecimento da matéria, educação etc.

Vamos supor que os questionários foram preenchidos por todos os alunos (infelizmente, algo raro na realidade) e entregues à CG. Os membros da CG fizeram, então, os cálculos das pontuações dos professores segundo a avaliação de cada turma (que envolveram médias das notas dadas pelos alunos de cada turma).

Os resultados estão apresentados na tabela abaixo, tanto em termos da pontuação de cada professor como em termos da sua posição no ranking (vamos supor que as notas de cada professor variam numa escala de 1 a 5 e que dois professores não possam ter notas iguais).

	Turma do 2º ano		**Turma do 3º ano**	
Professor	**Nota**	**Posição no ranking**	**Nota**	**Posição no ranking**
A	3,5	2	4	2
B	4	1	4,5	1
C	3,2	3	3	3

Vamos considerar apenas as colunas que dão as classificações dos professores em termos da sua posição no ranking. Observe que as posições dos três professores nos rankings das duas turmas são as mesmas. Isto implica que a correlação das avaliações dos três professores feitas pelas duas turmas é perfeita. De fato, o cálculo do coeficiente de correlação de Spearman nos dá $r_S = 1$ (mostre isso como exercício).

A questão que se põe é: o coeficiente de correlação é +1, indicando uma correlação perfeita, mas será que isso não ocorreu apenas por coincidência? Dito de outra forma, quão significante é o valor $r_S = 1$?

Em estatística, aborda-se a questão da significância de um resultado usando-se o conceito de hipótese nula. A **hipótese nula** (H_0) simplesmente assume que um dado resultado estatístico foi obtido apenas por acaso, devido a flutuações probabilísticas dos eventos sendo medidos, e não devido a um efeito real que cause o resultado. Sempre que se trabalha com uma hipótese para explicar um dado fenômeno, temos que considerar a possibilidade de pelo menos uma hipótese concorrente a ela. No caso da estatística, a hipótese concorrente é chamada de **hipótese alternativa** (H_A).

No nosso exemplo, o resultado empírico é que a correlação entre as avaliações dos três professores feitas pelas duas turmas é perfeita (r_S = 1). Como hipótese nula, vamos considerar que esse resultado é pura coincidência e, portanto, que nada de mais profundo possa ser retirado dele. Já a hipótese alternativa considera o contrário, que o resultado é devido a uma real similaridade das opiniões dos alunos das duas turmas sobre os três professores, ou seja, que a correlação é **significante**.

Uma vez que as duas hipóteses tenham sido enunciadas e os seus significados estejam bem claros na mente do pesquisador, a estratégia usada pela Estatística consiste em atacar a hipótese nula. Isso é feito com a seguinte pergunta: se a hipótese nula estiver correta e o resultado obtido for devido apenas ao acaso, qual a probabilidade de que ele ocorra?

No nosso exemplo, se o resultado obtido for apenas obra do acaso, para calcular a probabilidade dele temos que considerar todos os resultados possíveis.

De quantas maneiras os três professores podem ser ordenados pela turma do 2º ano? Como temos 3 professores, o número de maneiras é 3! = 3.2.1 = 6 (veja abaixo).

Professor	Possibilidades de rankeamento pela turma do 2º ano					
A	1	1	2	2	3	3
B	2	3	1	3	1	2
C	3	2	3	1	2	1

Igualmente, os 3 professores podem ser ordenados pela turma do 3º ano de 6 maneiras diferentes (as mesmas mostradas acima).

Desta forma, como cada turma pode ordenar os 3 professores de 6 maneiras diferentes, o número de possíveis maneiras diferentes em que os 3 professores podem ser rankeados pelas duas turmas em conjunto é 6x6 = 36 (veja abaixo).

Maneiras	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Prof.	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º
A	1 1	1 1	1 2	1 2	1 3	1 3	1 1	1 1	1 2	1 2	1 3	1 3	2 1	2 1	2 2	2 2	2 3	2 3
B	2 2	2 3	2 1	2 3	2 1	2 2	3 2	3 3	3 1	3 3	3 1	3 2	1 2	1 3	1 1	1 3	1 1	1 2
C	3 3	3 2	3 3	3 1	3 2	3 1	2 3	2 2	2 3	2 1	2 2	2 1	3 3	3 2	3 3	3 1	3 2	3 1
r_S	1	0,5	0,5	-0,5	-0,5	-1	0,5	1	-0,5	0,5	-1	-0,5	0,5	-0,5	1	-1	0,5	-0,5

Maneiras	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Prof.	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º	2º 3º
A	2 1	2 1	2 2	2 2	2 3	2 3	3 1	3 1	3 2	3 2	3 3	3 3	3 1	3 1	3 2	3 2	3 3	3 3
B	3 2	3 3	3 1	3 3	3 1	3 2	1 2	1 3	1 1	1 3	1 1	1 2	2 2	2 3	2 1	2 3	2 1	2 2
C	1 3	1 2	1 3	1 1	1 2	1 1	2 3	2 2	2 3	2 1	2 2	2 1	1 3	1 2	1 3	1 1	1 2	1 1
r_S	-0,5	0,5	-1	1	-0,5	0,5	-0,5	-1	0,5	-0,5	1	0,5	-1	-0,5	-0,5	0,5	0,5	1

A última linha dessa tabela dá os valores do coeficiente de correlação de Spearman r_S para cada um dos 36 casos possíveis. Vemos que em seis deles a correlação entre as avaliações das duas turmas é positiva e perfeita ($r_S = 1$).

Dadas todas as possibilidades acima, a probabilidade de se obter uma situação como a do exemplo, em que $r_S = 1$ é:

$$p = \frac{\text{número de maneiras em que } r_S = 1 \text{ pode ocorrer}}{\text{número de combinações possíveis}} = \frac{6}{36} = 0{,}167.$$

Isto quer dizer que caso o resultado obtido no experimento seja fruto de mero acaso, a probabilidade de ele ocorrer é de 16,7%.

Se você achar que este valor de probabilidade é suficientemente baixo (onde o critério de definição de "suficientemente baixo" tem que ser previamente definido), você pode rejeitar a hipótese nula. A idéia é a de que, se a probabilidade de o resultado ser obtido por mero acaso for muito baixa, deve-se considerar que a hipótese do acaso não é suficientemente forte para explicar o ocorrido e, portanto, que a hipótese alternativa tem mais chances de oferecer uma explicação melhor.

Normalmente, o limiar do valor de probabilidade abaixo do qual a hipótese nula é rejeitada é 5% (p = 0,05). Se a probabilidade do evento caso a hipótese nula esteja certa for menor que 5%, rejeita-se a hipótese nula; caso a probabilidade seja maior que 5%, não se pode rejeitar a hipótese nula.

A probabilidade de se obter a classificação dos dois professores do nosso exemplo por mero acaso é de 16,7%, um valor bem maior que 5%. Sendo assim, mesmo com o valor de r_S tendo sido máximo, não se pode concluir que ele é significante. A probabilidade de que ele tenha sido gerado apenas pelo acaso é muito grande.

O que aconteceria se a CG pedisse para as duas turmas avaliar quatro professores e os dois rankings feitos por elas fossem perfeitamente iguais, de maneira que novamente r_S = 1? Neste caso, cada turma poderia ordenar os quatro professores de 4! = 4.3.2.1 = 24 maneiras diferentes e o número de pareamentos possíveis dos dois ordenamentos seria 24x24 = 576. Desses, há 24 deles em que os rankeamentos são exatamente iguais. Desta forma, a probabilidade de se obter um pareamento idêntico por puro acaso é, neste caso,

$$p = \frac{24}{576} = 0{,}042.$$

Este valor é agora menor que o limiar de rejeição da hipótese nula, p = 0,05. Neste segundo caso, o acaso não é suficiente para explicar o resultado obtido (a um nível de significância de 0,05) e, portanto, a hipótese nula deve ser rejeitada. Conclui-se que a correlação é agora significativa.

Esses dois exemplos são muito simples e tiveram por objetivo apenas ilustrar o conceito de significância. Espero que eles tenham mostrado que um valor alto do coeficiente de correlação, tanto faz se for r ou r_S, não é suficiente para se acreditar que existe de fato uma relação entre as duas variáveis.

Quando o tamanho *n* da amostra é pequeno, é relativamente fácil (isto é, muito provável) obter um pareamento perfeito entre as duas variáveis apenas pelo acaso. Em um caso assim, tanto as regras da pesquisa científica como as regras do bom senso nos indicam que não devemos considerar um valor alto de *r* como significante.

Por outro lado, quando *n* for grande, mesmo valores baixos do coeficiente de correlação (por exemplo, $r = 0{,}2$ ou menor) são difíceis de ser obtidos por mero acaso e, portanto, são considerados como significantes. É por isso que é muito comum hoje em dia lermos reportagens em jornais dizendo que algum estudo mostrou que "existe uma pequena, mas significante" tendência de algo (em geral, associada a uma pesquisa sobre a correlação entre algum alimento e alguma condição da saúde humana).