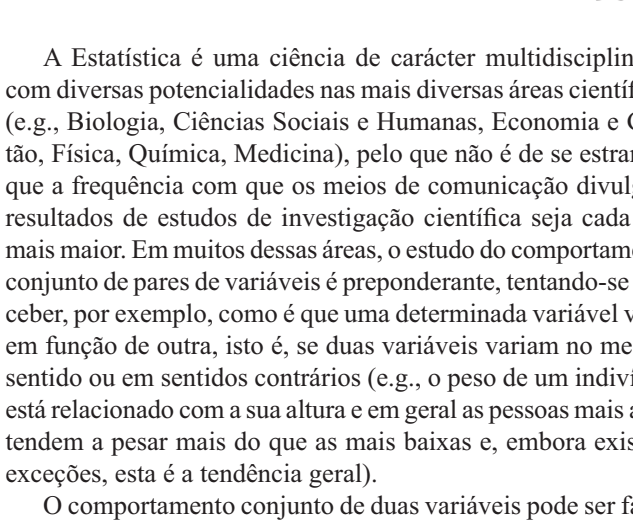
Por: Áurea Sousa
Professora Auxiliar do Departamento de Matemática e Estatística da Faculdade de Ciências e Tecnologia da Universidade dos Açores
aurea.st.sousa@uac.pt

Coeficiente de Correlação de Pearson e Coeficiente de correlação de Spearman. O que medem e em que situações devem ser utilizados?

A Estatística é uma ciência de carácter multidisciplinar e com diversas potencialidades nas mais diversas áreas científicas (e.g., Biologia, Ciências Sociais e Humanas, Economia e Gestão, Física, Química, Medicina), pelo que não é de se estranhar que a frequência com que os meios de comunicação divulgam resultados de estudos de investigação científica seja cada vez mais maior. Em muitos dessas áreas, o estudo do comportamento conjunto de pares de variáveis é preponderante, tentando-se perceber, por exemplo, como é que uma determinada variável varia em função de outra, isto é, se duas variáveis variam no mesmo sentido ou em sentidos contrários (e.g., o peso de um indivíduo está relacionado com a sua altura e em geral as pessoas mais altas tendem a pesar mais do que as mais baixas e, embora existam exceções, esta é a tendência geral).

O comportamento conjunto de duas variáveis pode ser facilmente observado através de um diagramade dispersão ou gráfico de dispersão, em que são utilizadas coordenadas cartesianas para exibir os valores de um conjunto de dados como uma coleção de pontos, em que os valores de uma variável são lidos no eixo horizontal (eixos das abcissas) e os da outra no eixo vertical (eixo das ordenadas). Note-se que, neste tipo de representação, cada um dos indivíduos corresponde a um ponto definido pelos valores de ambas as variáveis para esse indivíduo. Voltando ao exemplo da relação entre o peso e a altura, é apresentado abaixo um diagrama de dispersão que, embora seja referente a um conjunto de dados muito pequeno, permite ilustrar e clarificar este tipo de gráfico e a relação entre estas duas variáveis



Olhando para um gráfico desta natureza, é possível verificar o que sucede a uma variável, X, quando outra variável, Y, varia. Este tipo de gráfico permite visualizar tendências importantes sugeridas pelos dados, podendo-se imaginar uma linha que passa entre os pontos, vislumbrando-se algumas vezes uma relação linear entre as duas variáveis, outras vezes uma relação não linear e em outros casos a ausência de relação /associação. Existe correlação linear quando é possível ajustar à "nuvem" de pontos uma linha reta.

A primeira representação gráfica bidimensional é atribuída à Francis Galton (1822 - 1911), sendo de referir que a criação do gráfico de dispersão tal como é conhecido atualmente foi atribuída a John F. W. Herschel (1792-1871). A associação entre duas variáveis quantitativas é preferencialmente expressa por um coeficiente de correlação, por exemplo, através do coeficiente de correlação de Pearson, também designado por "coeficiente de correlação produto-momento" ou simplesmente por "ρ de Pearson", e do coeficiente de correlação de Spearman, também conhecido como coeficiente de correlação ordinal de Spearman, os quais variam entre -1 e +1 e não dependem das unidades de medida das variáveis, o que facilita a sua interpretação. No seguimento deste texto, veremos em que situações estes coeficientes devem ser aplicados.

O sentido e a intensidade da relação/associação linear existente entre duas variáveis quantitativas podem ser avaliados através do coeficiente de correlação linear de Pearson, o qual é adequado à avaliação de relações lineares.Embora o desenvolvimento deste coeficiente seja comumente atribuído a Karl Pearson, a sua origem remonta ao trabalho conjunto de Karl Pearson e de Francis Galton. O valor zero indica a inexistência de uma relação linear entre as duas variáveis, sendo de salientar que quanto mais próximo de 1 for o valor absoluto deste coeficiente mais forte é a relação linear entre as duas variáveis.O sinal indica o sentido da relação entre as duas variáveis (um sinal positivo indica que as duas variáveis variam no mesmo sentido, enquanto um sinal negativo indica que as variáveis variam em sentido inverso) e o valor deste coeficiente indica a magnitude/intensidade da relação linear entre as variáveis.Por exemplo, a venda de gelados geralmente é positivamente correlacionada com o aumento da temperatura (quanto maior a temperatura, maior é o número de gelados vendidos/consumidos); a venda de carros pode estar correlacionada negativamente com o aumento da taxa de desemprego (quanto maior a taxa de desemprego, menor o número de carros vendidos). No entanto, convém ter em atenção que a aplicação deste coeficiente a dados não lineares poderá não captar corretamente a intensidade da relação entre as variáveis. A inclinação dos pontos da esquerda para a direita, num diagrama de dispersão, sugere a existência de uma correlação positiva entre as variáveis. No limite, isto é, se a correlação for "perfeita" - como é o caso da correlação de uma variável consigo própria - o coeficiente de correlação será igual a 1. Por outro lado, a inclinação da direita para a esquerda indica a existência de uma correlação negativa. No limite, isto é, se a correlação for "perfeita" o coeficiente de correlação será igual a -1.

Há ainda a salientar o facto de que correlação não significa causalidade, isto é a observação da existência de uma relação/associação entre varáveis não implica necessariamente uma relação do tipo causa-efeito entre estas. Por exemplo, imaginemos um gráfico que relacione os danos causados pelos incêndios com o número de bombeiros disponíveis para os combater. Em geral, quantos mais danos mais bombeiros, mas isto não significa que os bombeiros são responsáveis pelos danos. Neste caso, há uma terceira variável que estabelece a relação de causa-efeito, nomeadamente a magnitude do incêndio.

Em 1994 o jornal "The New York Times" publicou um artigo que referia que os países com maior consumo de vinho tinham uma menor taxa de mortalidade associada a doenças cardíacas. No entanto, isso não significa necessariamente que exista uma relação de causa-efeito entre estas duas variáveis, embora essa possibilidade não possa ser descartada, sendo necessário mais algum trabalho de investigação, tendo em atenção outras variáveis (por vezes, uma multiplicidade de variáveis). Note-se que, geralmente, os países produtores de vinho são também os maiores consumidores desta bebida, pelo que essa taxa pode ser afetada por outras variáveis, tais como os hábitos/estilos de vida, a dieta mediterrânea ou o clima desses países.

É necessário ter, ainda, em atenção que a inexistência de uma correlação linear entre duas variáveis não significa que não se verifique outro tipo de correlação, por exemplo, exponencial, pelo que a representação gráfica pode ser uma mais-valia nesse contexto.

O coeficiente de Spearman avalia a intensidade e o sentido da relação monótona entre duas variáveis que estejam no mínimo numa escala ordinal, tem em consideração as ordens atribuídas às observações, em vez dos valores originais, e pode ser aplicado tanto no caso de dados lineares como no caso de dados não lineares. É de salientar, ainda, que os valores deste coeficiente podem ser calculados usando a fórmula do coeficiente de correlação de Pearson após a substituição dos valores originais das variáveis pelas respetivas ordens atribuídas. Assim, este coeficiente não é sensível a assimetrias na distribuição, nem à presença de outliers (valores atípicos), não exigindo que os dados provenham de duas populações com distribuições normais. Aplica-se também no caso de variáveis intervalares/rácio como alternativa ao coeficiente de correlação de Pearson, quando não é verificado o pressuposto de normalidade. No caso em que se observam alguns pontos muito afastados dos restantes, ou em que exista uma relação monótona crescente ou decrescente entre as variáveis em formato de curva, é recomendável a utilização do coeficiente de correlação de Spearman em vez do coeficiente de correlação de Spearman.

É importante consciencializar o leitor para o facto de que, enquanto o coeficiente de correlação de Pearson avalia relações lineares entre as variáveis, o coeficiente de correlação de Spearman avalia relações monótonas, quer estas sejam lineares ou não. É, ainda, de referir que quando a relação entre as variáveis não é monótona, nenhum destes dois coeficientes consegue captar corretamente a intensidade da relação entre as duas variáveis. Se estes dois coeficientes tomarem valores semelhantes, a relação entre as variáveis é provavelmente linear. Se o valor do coeficiente de correlação de Spearmanfor superior ao de Pearson, estamos provavelmente perante uma relação não linear monótona. Se o valor do coeficiente de correlação de Pearsonfor superior ao do coeficiente de correlação de Spearman, pode tratar-se de um artefacto devido à presença de outliers. Finalmente, se os valores de ambos os coeficientes forem negligenciáveis, a relação entre as duas variáveis poderá ser não linear e não monótona ou poderá não existir relação/correlação entre estas (ausência de correlação).

O facto de existir uma correlação entre duas variáveis na amostra (subconjunto da população) não significa necessariamente que haja correlação na população, pelo que é pertinente a realização do teste de significância associado a um coeficiente de correlação para se testar a hipótese nula de não haver correlação entre as duas variáveis na população, ou seja, a hipótese nula do coeficiente de correlação ser zero na população. Sempre que esta hipótese nula ou inicial é rejeitada conclui-se que há uma correlação, estatisticamente significativa, entre as duas variáveis na população.

Se a observação dos pontos referentes a um diagrama de dispersão sugerir uma relação linear entre duas variáveis, pode ser de interesse representar este padrão através de uma reta (de regressão), a qual pode ser ajustada usando o método dos mínimos quadrados. Nesse caso, a relação entre duas variáveis pode ser descrita pela chamada equação de regressão, em que a variável resposta (dependente), geralmente designada por Y, mede o resultado de um estudo e a variável explicativa (independente), geralmente designada por X, procura explicar os resultados observados. Daí a relação entre correlação e regressão, outro dos temas fascinantes na análise de dados nas diversas áreas científicas.