Correlação e Regressão Linear

A medida de correlação é o tipo de medida que se usa quando se quer saber se duas variáveis possuem algum tipo de relação, de maneira que quando uma varia a outra varia também. Baseado na medida de correlação entre duas variáveis, pode-se ter uma idéia sobre se o conhecimento de valores de uma das variáveis permite a previsão de valores da outra variável. Se uma variável tende a aumentar quando a outra aumenta, dizemos que a correlação é positiva. Por outro lado, se uma variável tende a diminuir quando a outra aumenta, dizemos que a correlação é negativa. Já uma correlação igual a zero indica que uma variação em uma das variáveis (aumento ou diminuição) não influencia a outra.

Pense nas seguintes afirmações:

1. Quanto mais velha a pessoa, de menos coisas ela se lembra;
2. Quanto mais se dá às crianças, mais elas querem;
3. As pessoas mais altas tendem a ter mais sucesso nas suas carreiras;
4. Quanto mais punição física as crianças recebem, mais agressivas elas vão ficar quando crescerem;
5. A estimulação cognitiva na infância aumenta a inteligência da pessoa;
6. Bons músicos são, em geral, bons em matemática;
7. Pessoas que são boas em matemática tendem a ser ruins em literatura;
8. Quanto mais se pratica um instrumento musical, menos erros são cometidos ao tocá-lo.

Estes são todos exemplos de casos de correlação entre duas variáveis. Cada afirmação propõe que duas variáveis estão correlacionadas, isto é, que elas co-variam no sentido de que:

- Quando uma variável aumenta a outra também aumenta (**correlação positiva**);
- Quando uma variável aumenta a outra diminui (**correlação negativa**).

Exercício: Quais dos casos acima são, em sua opinião, exemplos de correlação positiva e quais são exemplos de correlação negativa? Sugira outros exemplos de correlações positivas e negativas.

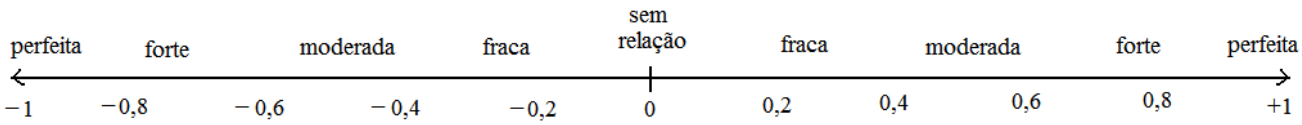
O primeiro passo para se verificar a validade de uma afirmação como as anteriores é *operacionalizar* as definições das variáveis envolvidas. Por exemplo, no caso da afirmação 7 o que se pode fazer para testá-la é olhar os resultados de provas de alunos de segundo grau nas duas matérias (matemática e literatura). No caso da afirmação 3, uma das variáveis pode ser medida diretamente (a altura), mas e a outra? Como *medir* o sucesso de alguém em uma carreira? Pelo salário, ou deve-se considerar alguma medida de "satisfação no emprego", e com que pesos? Isto é o que se quer dizer por operacionalização de uma variável.

Exercício: Proponha definições operacionais para as duas variáveis envolvidas em cada um dos exemplos anteriores e como elas devem ser medidas em um estudo de correlação.

Afirmações como "há uma correlação entre punição severa na infância e delinqüência na idade adulta", ou "punições severas na infância e delinqüência na idade adulta tendem a se correlacionar" são muito comuns em diversos meios (imprensa, universidade, governo, sistemas judiciário e penal, organizações não-governamentais, etc). Na verdade, nas duas afirmações estão faltando duas coisas importantes: (i) quão forte é a correlação; e (ii) quão significativa ela é. Força e significância são dois elementos importantes para se qualificar uma correlação, e elas não querem dizer a mesma coisa – como veremos.

A **força** de uma relação entre duas variáveis nos dá o grau com que uma variável tende a variar quando a outra varia. Ela é expressa em uma escala indo de −1 (correlação negativa perfeita) a +1 (correlação positiva perfeita). O nome que se dá à variável que mede a força de uma correlação (nessa escala de −1 a +1) é **coeficiente de correlação** (representado pela letra *r*).

As interpretações que se costumam dar aos significados dos valores do coeficiente de correlação dentro da sua faixa de valores possíveis são dadas abaixo:



Note que correlação negativa não quer dizer falta de correlação! O sinal do coeficiente de correlação tem como função apenas indicar se as duas variáveis se correlacionam de maneira diretamente proporcional ou inversamente proporcional, isto é, se quando uma aumenta a outra aumenta ou se quando uma aumenta a outra diminui. A força da correlação (positiva ou negativa) é dada pelo módulo do coeficiente de correlação: quanto maior o módulo, mais forte é a correlação. E correlação zero indica que não há qualquer relação entre as duas variáveis.

A técnica mais simples e provavelmente mais útil para se estudar a relação entre duas variáveis é o chamado **diagrama de dispersão**. O primeiro passo para a construção de um diagrama de dispersão é coletar pares de valores, um para a variável *X* e outro para a variável *Y*, onde cada par (*X*,*Y*) refere-se a um mesmo indivíduo (por exemplo, nota da prova de matemática e nota da prova de literatura de um aluno).

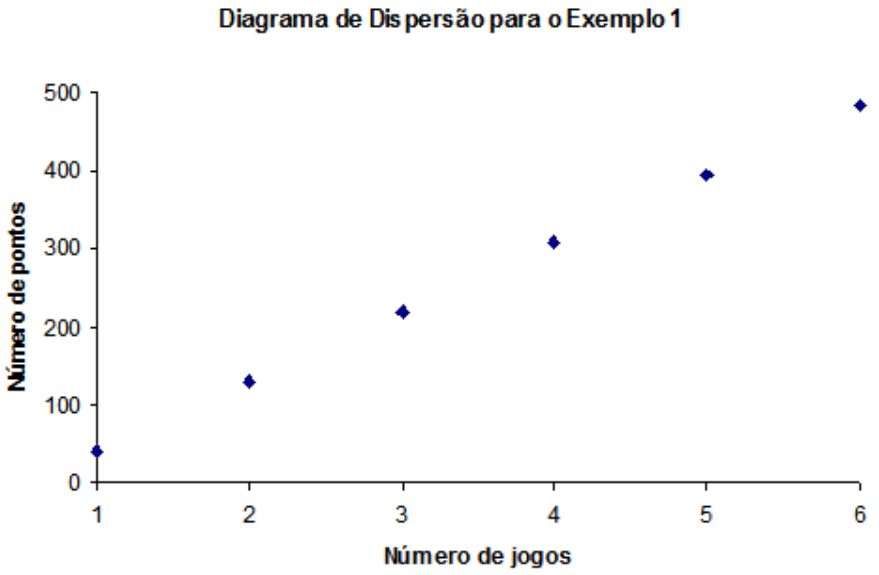
Supondo que foram coletados *n* pares de valores, (X_i,Y_i), $i = 1, ..., n$, para *n* indivíduos diferentes, o diagrama de dispersão é um gráfico cartesiano em que os valores da variável *X* são colocados no eixo horizontal (abscissa) e os valores da variável *Y* são colocados no eixo vertical (ordenada). Desta forma, cada um dos *n* pares de valores é representado graficamente como um único ponto. Olhando para o arranjo dos pontos no gráfico, pode-se discernir algum padrão que indique a possível forma funcional da relação entre os dados.

Exemplo 1: Suponha que uma criança esteja aprendendo a jogar um novo jogo de vídeo-game, por exemplo, um jogo em que a criança assuma o papel de uma personagem em um

mundo encantado que tenha como objetivo encontrar um certo tesouro. Durante a busca pelo tesouro, a personagem se movimenta por esse mundo encantado e vai enfrentando desafios de vários tipos. Cada vez que ela supera um desafio, ganha um certo número de pontos e novas habilidades que a ajudarão a achar o tesouro mais facilmente. Vamos supor que o aprendizado da criança em jogar esse novo jogo esteja sendo monitorado por um psicólogo. Pelas regras do acompanhamento, a cada dia a criança deve iniciar um jogo novo com a sua personagem sempre na mesma situação e com zero pontos.

Após seis jogos, o desempenho da criança resultou nos seguintes dados, apresentados em forma de tabela e na forma de um diagrama de dispersão (dados fictícios):

Número de jogos	Número de pontos
1	42
2	131
3	219
4	308
5	396
6	485

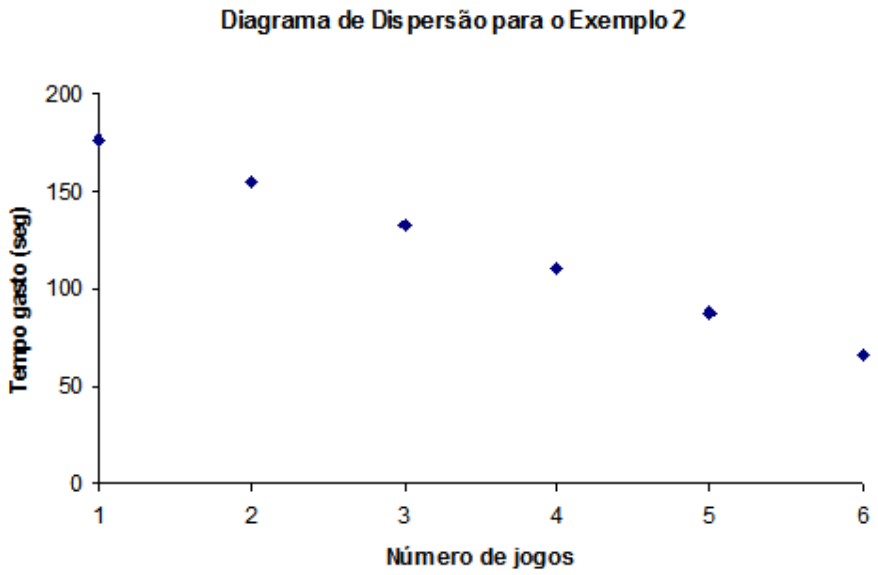


Observe que o diagrama de dispersão indica claramente que há uma relação positiva entre o número de pontos num jogo e o número de vezes que a criança o jogou: quanto mais vezes a criança repete o jogo, mais pontos ela faz. No caso, a correlação entre as duas variáveis é positiva e perfeita (coeficiente de correlação $r = +1$), mas veremos como calcular esse coeficiente depois.

Exemplo 2: Consideremos novamente o mesmo caso do exemplo anterior. A cada repetição do jogo, além de registrar o número de pontos que a criança faz, o psicólogo também registra o tempo gasto pela criança para completar o primeiro desafio do jogo.

Os resultados estão mostrados abaixo.

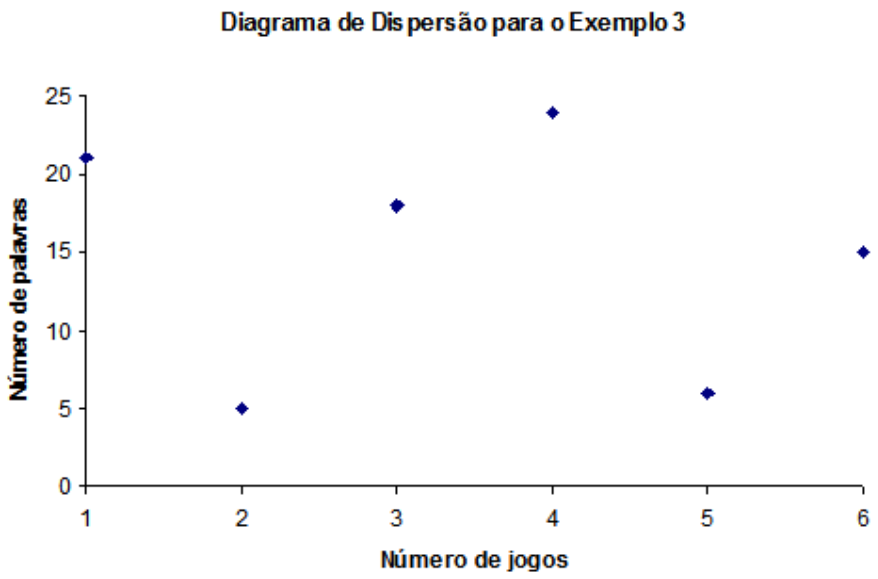
Número de jogos	Tempo gasto (seg.)
1	177
2	155
3	133
4	110
5	88
6	66



A correlação entre as duas variáveis é agora negativa e perfeita (coeficiente de correlação $r = -1$). Compare os dois diagramas de dispersão: quando a correlação é positiva, os pontos no diagrama de dispersão vão do quadrante inferior esquerdo ao quadrante superior direito; já quando a correlação é negativa, os pontos vão do quadrante superior esquerdo ao quadrante inferior direito.

Exemplo 3: Ainda considerando o mesmo caso dos dois exemplos anteriores, suponha que a cada repetição do jogo o psicólogo também anote quantas palavras a criança fala durante os primeiros 10 minutos de jogo. O resultado está dado abaixo.

Número de jogos	Número de palavras faladas
1	20
2	4
3	13
4	24
5	5
6	15

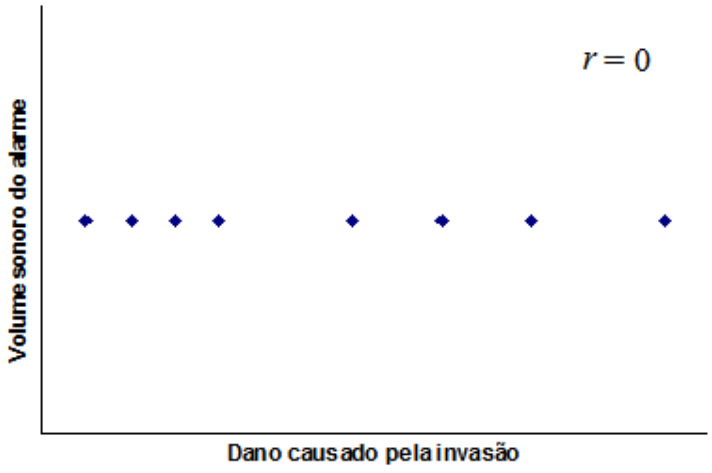


Neste último caso não há correlação entre as duas variáveis (o coeficiente de correlação vale $r = -0{,}02$).

Nos casos dos exemplos 1 e 2, em que as correlações são perfeitas (positiva ou negativa), é possível traçar uma reta no olho unindo todos os pontos. A equação dessa reta nos dá a relação quantitativa entre as duas variáveis (X e Y). Porém, quando a correlação não é perfeita (mesmo que seja forte) deve-se *calcular* essa reta matematicamente e não usar o *olhômetro*. A reta que dá a relação entre duas variáveis é chamada de **reta de regressão linear** e ela sempre pode ser calculada, mesmo que as varáveis não tenham qualquer correlação. Veremos como calculá-la mais tarde.

No exemplo 3, o valor do coeficiente de correlação é $r \cong 0$ porque as variações em Y não são afetadas pelas variações em X. Uma outra maneira de dizer isso é que o valor de Y não pode ser previsto a partir do conhecimento do valor de X. Para interpretar melhor o significado de $r = 0$, vejamos mais alguns casos em que isso ocorre.

Exemplo 4: Seja o diagrama de dispersão mostrado abaixo.

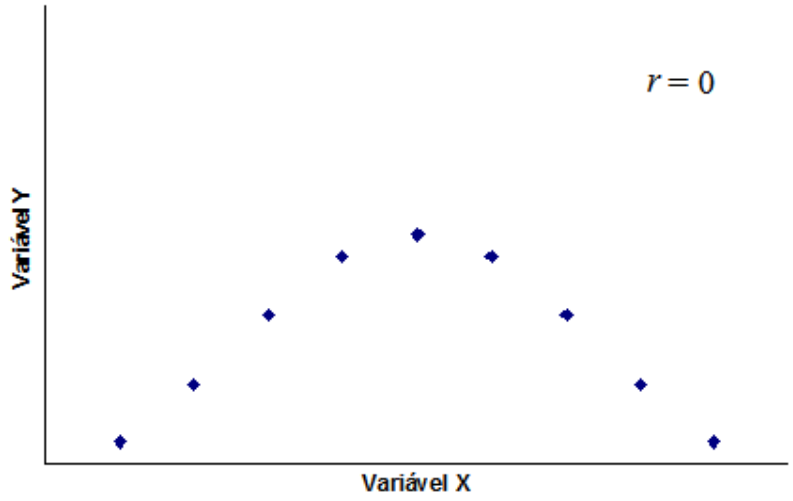


Este diagrama mostra, no eixo x, a quantidade de dano causado a uma família quando a sua casa é invadida por ladrões (em alguma escala predeterminada de dano) e, no eixo y, o volume do alarme sonoro que dispara quando a casa é invadida. Observe que, neste caso, r

= 0 porque o valor da variável *y* permanece constante independentemente do que aconteça com a variável *x*. O valor de *Y* pode ser previsto pelo diagrama (é sempre o mesmo!), mas o valor de *X* não. A única coisa que se pode prever a partir do conhecimento de *X* é que, se *X* tiver algum valor diferente de zero, haverá um valor de *Y*.

No começo desta aula foi escrito que "uma correlação igual a zero indica que uma variação em uma das variáveis (aumento ou diminuição) não influencia a outra". Isto só está correto para o caso de relações *lineares* entre variáveis. No caso de relações não-lineares, o coeficiente de correlação pode ter um valor próximo de zero e ainda assim elas estarem relacionadas. É por isso que a construção de um diagrama de dispersão é fundamental para o estudo da relação entre duas variáveis, pois ele permite que se visualize a relação entre elas. Vejamos um exemplo.

Exemplo 5: Seja o seguinte diagrama de dispersão.



Este diagrama tem uma forma curva, em forma de U invertido. Para este caso o cálculo do valor do coeficiente de correlação resulta em $r = 0$, mas mesmo assim vemos pelo gráfico que existe uma relação previsível entre *Y* e *X*. As variáveis *X* e *Y* não estão especificadas, mas pode-se pensar em algumas que possuam uma relação desse tipo. Por exemplo, temperaturas médias ao longo dos meses ano (começando a contar do inverno). Em psicologia, uma tal relação poderia descrever, por exemplo, o interesse de uma pessoa em realizar uma dada tarefa (como montar quebra-cabeças, por exemplo) em função do

número de vezes que ela repete a tarefa. No começo, o interesse cresce com número de repetições porque elas representam um desafio para a pessoa, mas depois que ela já atinge domínio sobre a tarefa o seu interesse decresce.

Exercício: pense em outras situações de interesse em psicologia que possam ser descritas por uma relação em forma de U invertido como a acima. Pense também em situações que possam ser descritas por uma relação em forma de U.

Relações entre duas variáveis como a do exemplo 5 são chamadas de **relações não-lineares** (simplesmente porque não se pode traçar uma linha reta que descreva a relação entre *X* e *Y*). Relações não-lineares são muito importantes por serem muito comuns – na natureza e nas relações humanas –, mas o seu estudo (com exceção de alguns casos simples) não será feito aqui.

Relações lineares também são importantes: (i) elas são aproximadamente válidas na natureza em algumas condições restritas; (ii) elas funcionam como bons modelos iniciais para um grande número de relações; e (iii) elas são simples, permitindo um tratamento matemático completo de forma analítica (isto é, não computacional).

O coeficente de correlação *r* é usado para medir a força de relações *lineares* entre duas variáveis Y e X. Quando $r = 0$, isto significa que não há relação linear entre as variáveis. Porém, *r* pode ser zero e ainda assim existir possivelmente alguma relação entre as duas variáveis, mas ela será necessariamente não-linear.

Vamos agora ver como calcular o coeficiente de correlação *r*. Antes de mais nada, é importante dizer que há mais de uma maneira de se definir o coeficiente de correlação matematicamente. Vamos apresentar aqui dois desses coeficientes: o **coeficiente de correlação de Pearson** e o **coeficiente de correlação de Spearman**.