2.2.4. Medidas de correlação

Se duas variáveis têm relação probabilística linear entre si (observada no diagrama de dispersão), a força desta relação pode ser medida pela *covariância* entre estas variáveis, ou pelo seu *coeficiente de correlação linear*. Veremos abaixo como estas duas medidas são calculadas em dados de amostras.

2.2.4.1. Covariância

A medida da *covariância* entre duas variáveis é derivada da *variância*, que mede a dispersão de uma variável. Vimos (seção **2.2.2.4**) que a variância s^2 de uma variável X é calculada pelo somatório do quadrado dos desvios de cada valor de X em relação à média:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n}$$

A covariância entre duas variáveis X e Y, normalmente representada por *cov(X,Y),* é calculada numa amostra pelo produto do desvio de uma variável em relação à sua média, pelo desvio da outra, como na eq. (**1**).

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n} \qquad (1)$$

Não é difícil entender intuitivamente o raciocínio por trás desta fórmula. Veja por exemplo, a Fig. **1A**, que mostra o diagrama de dispersão *altura x peso* de uma amostra de ciclistas profissionais (idêntica a Fig. 2 da seção **2.1.7.1**). A média do peso *X* destes ciclistas é $\overline{X}$ = 71,7 kg; a média da altura é $\overline{Y}$ = 175,3 cm. O ciclista representado pelo ponto no canto superior direito tem altura X=188 cm e peso Y=80 kg. Ele está acima da média, tanto no peso quanto na altura; ambos os desvios são positivos, e seu produto será também positivo. O ciclista representado pelo ponto no outro extremo do gráfico, no canto inferior esquerdo, tem altura X=165 cm e peso Y=61 kg; ele está abaixo da média, tanto no peso quanto na altura. Os dois desvios serão portanto negativos, e seu produto será positivo. Para todos os pontos que estejam nos dois quadrantes destacados com fundo cinza no gráfico, o produto dos desvios será positivo; como eles são maioria, neste exemplo, o somatório deverá ser positivo. Para os pontos que estão nos dois quadrantes com fundo branco, porém, um desvio será positivo e o outro negativo; o produto será portanto negativo. Se estes pontos forem muito numerosos (como por exemplo na Fig. **3C**), o somatório será negativo, indicando uma covariância negativa.

A covariância mostra se a relação entre as duas variáveis é positiva ou negativa; porém, apresenta os mesmos dois problemas já encontrados na variância. Primeiro, por não ser uma medida relativa, não nos permite avaliar se a força da relação entre as duas variáveis é grande, ou não. Por exemplo, para os ciclistas, a covariância calculada foi:

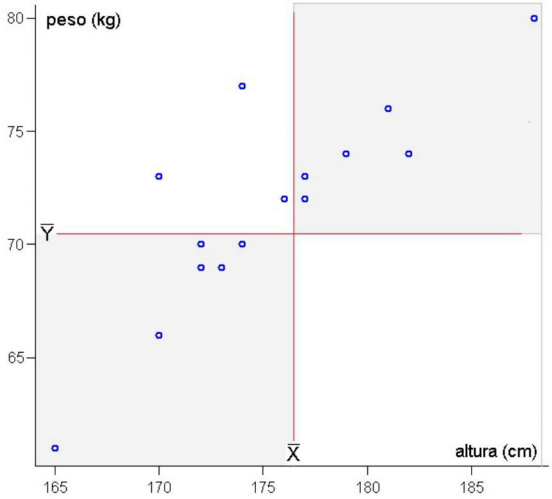
cov(altura,peso) = 21,9 cm.kg

Isto é muito ou pouco? Não podemos dizer, se não há um padrão de comparação. A covariância entre alturas e pesos nos dados de estudantes de Medicina (Fig. **1B**) é igual a

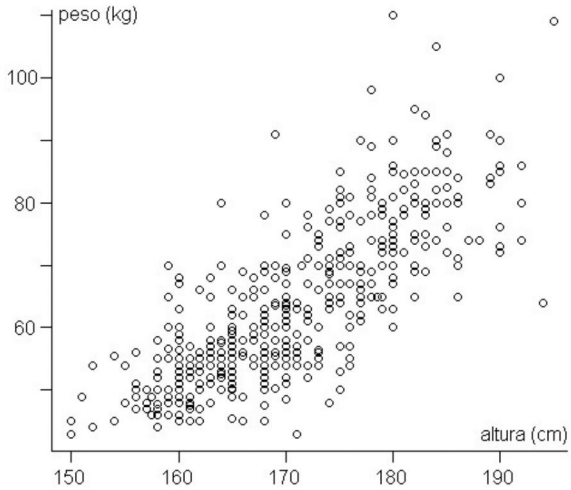
cov(altura,peso) = 82,6 cm.kg

No entanto, é claro no gráfico que a relação deve ser mais forte entre os ciclistas do que entre os estudantes. A razão disto é fácil de entender: os ciclistas têm todos um biótipo específico, característico daquele esporte; aqueles que não têm a relação desejada entre altura e peso provavelmente não conseguirão sucesso como profissionais. Os estudantes, por outro lado, não têm que atender a nenhuma exigência quanto à isto.

O segundo problema com a medida de covariância é que ela tem uma unidade, dada pelo produto das unidades das duas variáveis (no caso, kg x cm); não podemos por isso comparar o valor da covariância entre um par de variáveis com a de outro par, se as unidades forem diferentes. A covariância é por isso muito usada na teoria (na demonstração de teoremas, etc.), mas não muito na prática. Na seção seguinte veremos uma medida que evita estes dois problemas.



A. Ciclistas profissionais de estrada



B. Estudantes de Medicina da UFJF

Figura 1. Diagramas de dispersão *altura x peso*

Por fim, é importante ter sempre em mente que a covariância mede a relação probabilística *linear* entre duas variáveis; não serve se a relação for não-linear. Por exemplo, considere os três diagramas da Fig. **2**, feitos a partir de dados simulados. É bem evidente que as duas variáveis da Fig. **2A** têm uma relação probabilística não-linear muito forte, quase determinística; as da Fig. **2B** ainda também têm uma relação, porém bem mais fraca; as da Fig. **2C** não têm relação nenhuma. As covariâncias, no entanto, são praticamente nulas (0,00; -0,02 e 0,01, respectivamente), indicando que não há relação *linear* em nenhum dos três diagramas.

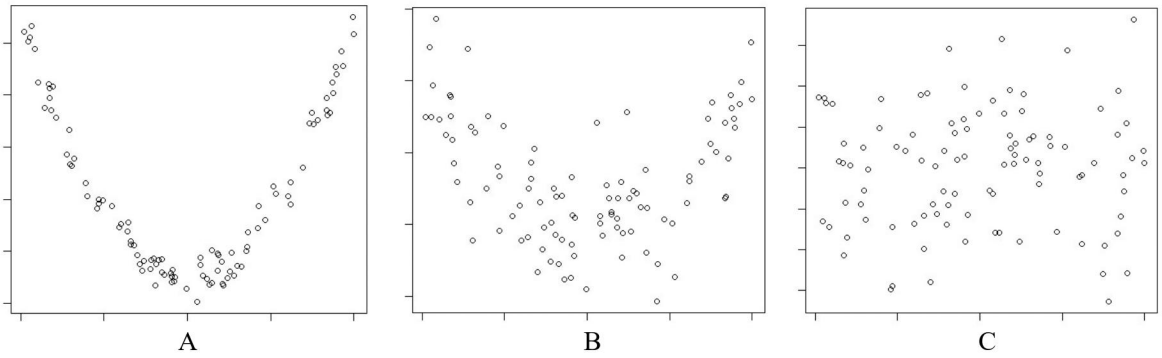


Figura 2. Três distribuições simuladas, com diferentes graus de relação não-linear

2.2.4.2. Coeficiente de correlação linear de Pearson

O *coeficiente de correlação linear de Pearson* é uma medida da força da relação probabilística linear entre duas variáveis. Francis Galton atribuiu a esta relação o nome de *correlação*, e propôs um índice para medi-la; este índice foi depois aperfeiçoado por Karl Pearson, donde o nome do coeficiente.

Em geral o valor deste coeficiente, calculado nos dados de uma amostra, é representado pela letra r ; o valor teórico para uma população é representado pela letra grega ρ (rô). Sua idéia básica é a mesma da covariância: usar o somatório dos produtos dos desvios de cada ponto em relação às médias das variáveis. No entanto, para estabelecer limites aos valores da medida, e para torná-la adimensional, a covariância é dividida pelo produtos dos desvios-padrões s_X e s_Y das duas variáveis, como na eq. (**2**).

$$r = \frac{\mathrm{cov}(X,Y)}{s_X s_Y} \qquad (2)$$

Demonstra-se que este coeficiente pode assumir valores no intervalo [-1,1]. Um valor de $r = \pm 1$ indica que as duas variáveis têm perfeita correlação linear, positiva ou negativa (o que é o mesmo que dizer que a relação entre elas é determinística); um valor de $r = 0$ indica que não há correlação alguma entre as variáveis.

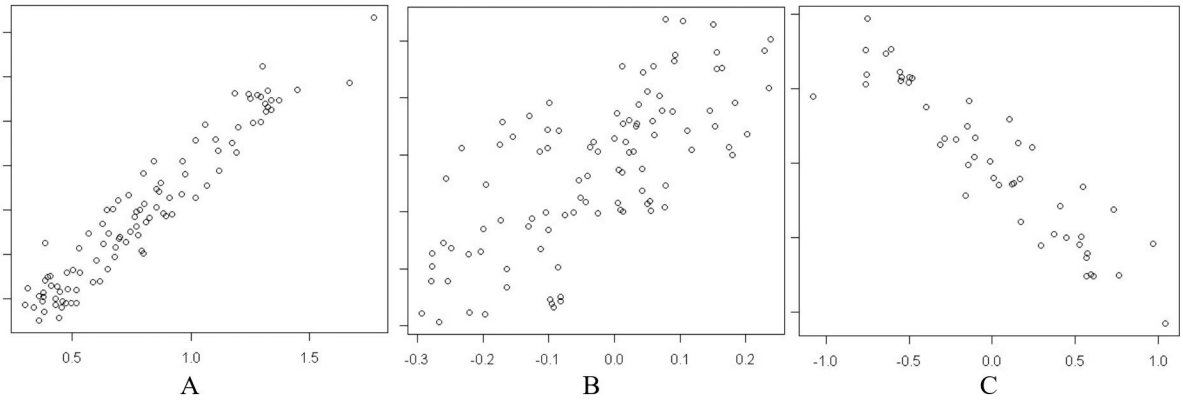


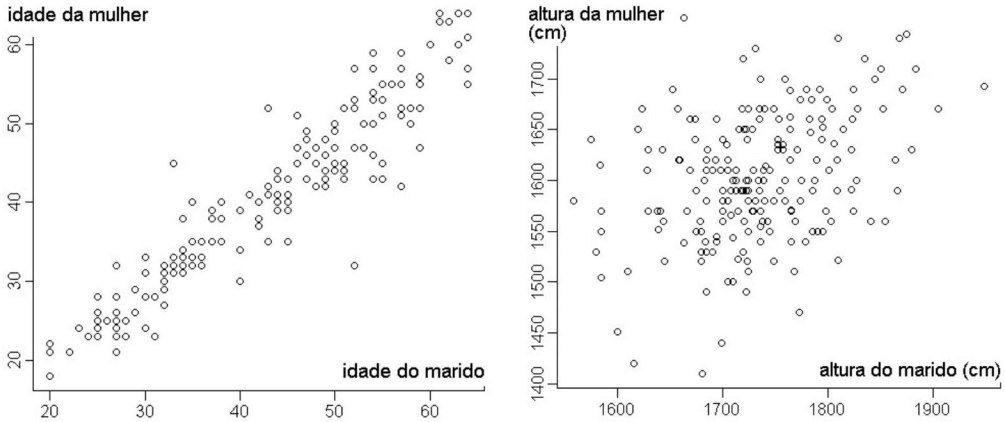
Figura 3. Três distribuições simuladas, com correlação linear

A Fig. **3** mostra três diagramas de dispersão com dados simulados (idênticos aos da Fig. 4 da seção **2.1.7**). Nos diagramas (A) e (C), as variáveis têm correlação muito forte, positiva e negativa, com coeficientes $r = 0{,}96$ e $r = -0{,}92$, respectivamente. No diagrama (B), a correlação positiva ainda é bastante evidente, mas menor, com $r = 0{,}71$ (na prática, correlações com $|r| > 0{,}7$ ainda são consideradas fortes, e são raramente encontradas).

Há três características importantes do coeficiente de Pearson que devem ser notadas: (i) o coeficiente é adimensional (não tem unidade); (ii) o coeficiente só mede a força de relações *lineares*; (iii) o coeficiente é muito sensível a valores discrepantes (*outliers*).

(i) O coeficiente não tem unidade

Este coeficiente é adimensional, e seu valor não depende das unidades das variáveis; podemos por isso usá-lo para comparar as correlações de pares de variáveis que têm unidades diferentes. Por exemplo, a Fig. **4A** mostra o diagrama de dispersão que relaciona as idades de maridos e mulheres, numa amostra de casais ingleses; a Fig. **4B**, o que relaciona as alturas nos mesmos casais. É bem evidente nos diagramas que existe uma relação linear positiva muito forte entre as idades, mas não tanto entre as alturas. Não podemos comparar a força destas relações usando a *covariância*, porque as unidades são diferentes; podemos contudo comparar usando o *coeficiente de correlação*. Isto confirma o que observamos nos diagramas: para as idades, o valor calculado do coeficiente é de $r = 0.94$, o que é uma correlação muito forte; para as alturas, apenas $r = 0.36$.



A. Idade do marido x idade da mulher B. Altura do marido x altura da mulher

Figura 4. Diagramas de dispersão das idades e alturas, em casais ingleses

(ii) O coeficiente mede a força da relação linear entre as variáveis

O coeficiente de Pearson não pode ser usado para avaliar relações não-lineares entre variáveis; não serve, por exemplo, para as variáveis nos diagramas da Fig. **2A** e **B**, que têm relações claramente não-lineares ($r = 0$, nos dois gráficos).

Na prática, quando em Estatística falamos em "correlação", geralmente estamos nos referindo a relações lineares. No R, por exemplo, o comando que calcula a correlação linear entre duas variáveis X e Y é `cor(X,Y)`; não é preciso especificar que a correlação seja linear. O estudo e a modelagem de relações não-lineares geralmente está além do escopo

da Estatística básica, e não existe uma medida geralmente aceita para medir a força destas relações.

(iii) O coeficiente é muito sensível a valores discrepantes

O coeficiente de Pearson é muito sensível, como também são a covariância e variância. Os gráficos da Fig. **5** mostram um exemplo extremo. Na amostra da Fig. **5A**, a correlação entre X e Y é obviamente nula. Na amostra da Fig. **5B**, introduzimos um ponto discrepante no canto superior direito do diagrama; como os desvios deste ponto em relação às média serão muito grandes, o somatório dos desvios é aumentado, e o coeficiente agora para a ser $r = 0{,}91$, o que indicaria uma correlação muito forte (que na verdade não existe, pois foi causada por apenas um ponto). Devido a esta sensibilidade, o coeficiente de Pearson deve ser usado com cuidado, especialmente se as amostras forem pequenas, porque um único valor discrepante pode vir a afetar o valor de r e sugerir uma correlação forte, que na verdade não existe.

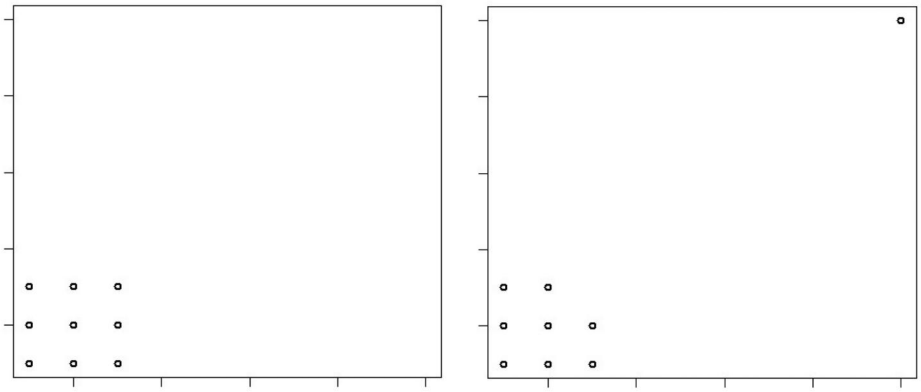


Figura 5. Efeito dos pontos discrepantes na correlação

2.2.4.3. Correlação × causalidade

Por fim, é importante lembrar de que uma *correlação* forte entre duas variáveis não implica *causalidade*. A correlação pode ser uma indicação de que a variação em uma das variáveis causa a variação na outra; porém, esta indicação, por si só, não é convincente. É preciso, além disso, verificar se existe alguma razão lógica que justifique a suposição de que uma destas variáveis afeta a outra.

Correlação fortes entre variáveis que aparentemente não têm nenhuma relação entre si podem aparecer por várias razões. A primeira é simplesmente a quantidade de informação disponível hoje em dia na *internet* e em outras fontes: há tantos dados, que é fácil encontrar amostras de duas variáveis quaisquer que têm por acaso uma forte correlação, principalmente se as amostras são pequenas. Exemplos disto podem ser vistos no site *http://tylervigen.com* : o número de doutorados em Engenharia está correlacionado com o consumo per capita de queijo mozarela em um país ($r = 0{,}96$); o número de doutorados em Sociologia nos EUA está correlacionado com o número de lançamentos mundiais de foguetes espaciais não-comerciais ($r = 0{,}79$); o número de pessoas que se afogam ao caírem

de um barco de pesca está correlacionado com a taxa de casamentos no estado de Kentucky (r = 0,95), etc. Correlações deste tipo são chamadas de "espúrias".

Outro razão para o surgimentos destas correlações é, simplesmente, a fraude. Se alguém quer provar alguma teoria de maneira fraudulenta, pode simplesmente inventar dados (o que acontece com mais freqüência do que se imagina); ou então, escolher dentro de um conjunto de dados aqueles que mais favorecem sua teoria, e descartar os outros. Por exemplo, suponha que tenhamos os dados mostrados na Fig. **6A**, que são amostras simuladas de duas populações independentes. Se quero usar estes dados para provar que existe uma correlação positiva entre as duas variáveis, posso simplesmente criar uma amostra selecionando os pontos marcados na Fig. **6B**, e esconder os outros; ou, se quero provar que a correlação é negativa, selecionar os da Fig. **6C**. Isto é chamado em inglês de *cherry picking* (catar cerejas), e é considerado uma forma grave de fraude científica.

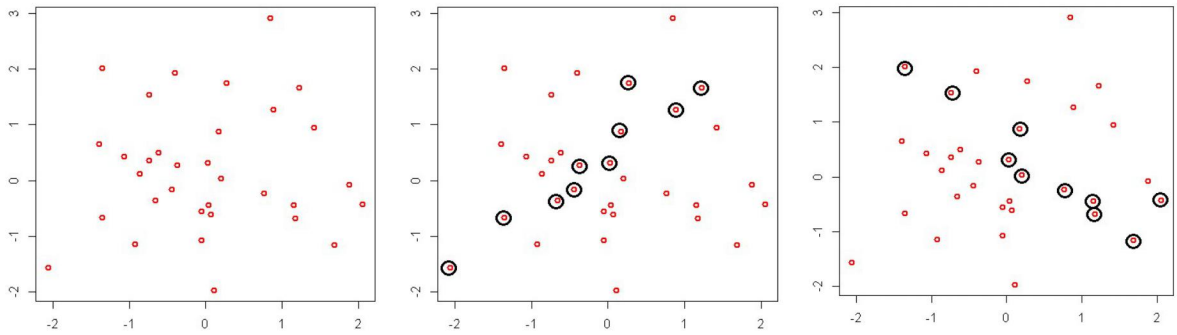


Figura 6. Selecionando pontos para criar correlações (*cherry picking*)

Uma terceira razão para a existência de correlação fortes entre duas variáveis que não têm na verdade nenhuma relação lógica entre si são as *variáveis de confundimento*: se duas variáveis X e Y estão associadas – quando X cresce, Y cresce também -, pode haver uma terceira variável W, oculta, que faz com que X e Y se movam na mesma direção. Por exemplo, existe uma correlação entre a porcentagem da população que tem computadores num país, e a porcentagem de pessoas que morrem de câncer. Isto não quer dizer que computadores causam câncer; no caso, a correlação é causada por uma terceira variável, o *grau de desenvolvimento* de um país. Quanto mais desenvolvido o país, maior o número de computadores e maior a expectativa de vida; a maior expectativa de vida, por sua vez, aumenta a porcentagem de pessoas com câncer (que é tipicamente uma doença de pessoas idosas).

Cada caso de correlação deve ser analisado por especialistas da área, que devem verificar se a suposta relação de causa e efeito entre as variáveis faz sentido - se podemos dizer que X realmente causa Y, ou se X e Y são ambos conseqüências de uma terceira variável W. As variáveis de confundimento são um problema, por exemplo, na pesquisa em Medicina, quando se deseja descobrir a causa da doença de Alzheimer. Há uma quantidade enorme de variáveis cuja relação com a incidência da doença deve ser investigada; algumas delas mostrarão correlação positiva, e será então preciso decidir se existe causalidade nesta relação.

Resumo

- Se duas variáveis têm uma relação probabilística linear entre si, dizemos que elas têm *correlação linear*;
- A força da relação probabilística entre duas variáveis X e Y pode ser medida pela *covariância* entre estas variáveis (representada por *cov(X,Y)*), ou pelo seu *coeficiente de correlação linear* (representado por *r*);

- A *covariância* é importante nas aplicações teóricas, mas não é muito usada nas aplicações práticas;
- O *coeficiente de correlação linear de Pearson* varia no intervalo [-1,1]. Os valores extremos deste intervalo, $r = 1$ e $r = -1$, indicam que existe uma relação determinística entre as variáveis;
- Este coeficiente é adimensional e pode ser usado para medir a correlação entre qualquer par de variáveis, independentemente de suas unidades ou de sua médias
- Uma correlação nula ($r = 0$) não significa que não haja relação entre as variáveis; pode haver uma relação não-linear entre elas;
- A existência de *correlação* entre duas variáveis não implica em *causalidade*! Se duas variáveis estão correlacionadas, isto nem sempre quer dizer que uma delas causa a outra.