Estatística Aplicada em Ciências e Engenharia

A. Rita Gaio
Departamento de Matemática - FCUP
argaio@fc.up.pt

October 8, 2013

Índice Geral

Chapter 1

Covariância e correlação

1.1. Distribuições conjuntas

X, Y variáveis aleatórias, espaço de probabilidade $(\Omega = \mathbb{R}^2, \mathcal{A}, P)$

- **Função de densidade conjunta de X e Y, $f(x,y)$**

  X, Y discretas: $f(x,y) \geq 0$, $\sum_x \sum_y f(x,y) = 1$ $P(X = x, Y = y) = f(x,y)$

  X, Y contínuas: $f(x,y) \geq 0$, $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)\, dxdy = 1$ $P((X,Y) \in B) = \int\int_B f(x,y)\, dxdy$

- **Funções de densidade marginal, $f_X(x)$ e $f_Y(y)$**

  X, Y discretas: $f_X(x) = \sum_y f(x,y) = P(X = x)$, $f_Y(y) = \sum_x f(x,y) = P(Y = y)$

  X, Y contínuas: $f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\, dy$, $f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\, dx$

	y_1	y_2	$\dots$	y_n	$f_X(x_i)$
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$	$\dots$	$f(x_1,y_n)$	$f_X(x_1) = \sum_j f(x_1,y_j)$
x_2	$f(x_2,y_1)$	$f(x_2,y_2)$	$\dots$	$f(x_2,y_n)$	$f_X(x_2) = \sum_j f(x_2,y_j)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	$f(x_m,y_1)$	$f(x_m,y_2)$	$\dots$	$f(x_m,y_n)$	$f_X(x_m) = \sum_j f(x_m,y_j)$
$f_Y(y_i)$	$f_Y(y_1) = \sum_i f(x_i,y_1)$	$f_Y(y_2) = \sum_i f(x_i,y_2)$	$\dots$	$f_Y(y_n) = \sum_i f(x_i,y_n)$	1

- **Função de distribuição conjunta de X e Y, $F(x,y)$**

  X, Y discretas: $F(x,y) = P(X \leq x, Y \leq y) = \sum_{u \leq x} \sum_{v \leq y} f(u,v)$

  X, Y contínuas: $F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\, dudv$

1.2. Medidas de tendência central, dispersão e correlação

As definições anteriores generalizam-se de forma óbvia a $n \geq 2$ variáveis aleatórias X_1, X_2, ..., X_n.
Espaço de probabilidade $(\Omega = \mathbb{R}^n, \mathcal{A}, P)$
$X_1, \ldots, X_n$ variáveis aleatórias com densidade de probabilidade conjunta $f(x) = f(x_1, \ldots, x_n)$

- **Função de densidade marginal de X_j, $f_{X_j}(x)$:**

X_1, X_2, ..., X_n discretas: $$f_{X_j}(x_j) = \sum_{y_1,\ldots,y_{j-1},y_{j+1},\ldots,y_n} f(y_1, \ldots, y_{j-1}, x_j, y_{j+1}, \ldots, y_n)$$

X_1, X_2, ..., X_n contínuas: $$f_{X_j}(x_j) = \int f(y_1, \ldots, y_{j-1}, x_j, y_{j+1}, \ldots, y_n)\, dy_1 \ldots dy_{j-1} dy_{j+1} \ldots dy_n$$

	Distribuição Discreta	Distribuição Contínua
Média de X_j	$\mu_{X_j} = \sum_{x_j} x_j f_{X_j}(x_j)$	$\mu_{X_j} = \int_{-\infty}^{+\infty} x_j f_{X_j}(x_j) dx_j$
Variância de X_j	$\sigma^2_{X_j} = \sum_k (x_j - \mu_{X_j})^2 f_{X_j}(x_j)$	$\sigma^2_{X_j} = \int_{-\infty}^{+\infty} (x_j - \mu_{X_j})^2 f_{X_j}(x_j) dx_j$
Desvio Padrão de X_j	$\sigma_{X_j} = \sqrt{\text{variância de } X_j}$	$\sigma_{X_j} = \sqrt{\text{variância de } X_j}$
Covariância entre X_j e X_k	$\text{Cov}(X_j, X_k) = E[(X_j - E(X_j))(X_k - E(X_k))] = \sum_x (x_j - \mu_{X_j})(x_k - \mu_{X_k}) f(x)$	$\text{Cov}(X_j, X_k) = E[(X_j - E(X_j))(X_k - E(X_k))] = \int_{-\infty}^{+\infty} (x_j - \mu_{X_j})(x_k - \mu_{X_k}) f(x) dx$
Correlação entre X_j e X_k	$\text{Cor}(X_j, X_k) = \rho_{jk} = \frac{\text{Cov}(X_j, X_k)}{\sigma_{X_j} \sigma_{X_k}}$	$\text{Cor}(X_j, X_k) = \rho_{jk} = \frac{\text{Cov}(X_j, X_k)}{\sigma_{X_j} \sigma_{X_k}}$

Considere-se o caso particular em que existem apenas duas variáveis aleatórias, X e Y, com função densidade de probabilidade conjunta dada por $f(x,y)$. Se X e Y forem discretas (resp. contínuas), as funções de densidade marginal são

$$f_X(x) = \sum_y f(x,y) \quad \text{e} \quad f_Y(y) = \sum_x f(x,y), \qquad \left(\text{resp. } f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy \quad \text{e} \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx\right).$$

Tem-se ainda a seguinte tabela:

	Distribuição Discreta	Distribuição Contínua
Médias	$\mu_X = \sum_x x f_X(x)$ $\mu_Y = \sum_y y f_Y(y)$	$\mu_X = \int_{-\infty}^{+\infty} x f_X(x) dx$ $\mu_Y = \int_{-\infty}^{+\infty} y f_Y(y) dy$
Variâncias	$\sigma_X^2 = \sum_x (x-\mu_X)^2 f_X(x)$ $\sigma_Y^2 = \sum_y (y-\mu_Y)^2 f_Y(y)$	$\sigma_X^2 = \int_{-\infty}^{+\infty} (x-\mu_X)^2 f_X(x) dx$ $\sigma_Y^2 = \int_{-\infty}^{+\infty} (y-\mu_Y)^2 f_Y(y) dy$
Desvios Padrão	$\sigma_X = \sqrt{\text{variância de } X}$ $\sigma_Y = \sqrt{\text{variância de } Y}$	$\sigma_X = \sqrt{\text{variância de } X}$ $\sigma_Y = \sqrt{\text{variância de } Y}$
Covariância entre X e Y	$\text{Cov}(X,Y)$ $= E[(X-E(X))(Y-E(Y))]$ $= \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y)$	$\text{Cov}(X,Y)$ $= E[(X-E(X))(Y-E(Y))]$ $= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x-\mu_X)(y-\mu_Y) f(x,y) dx dy$
Correlação entre X e Y	$\text{Cor}(X,Y) = \rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$	$\text{Cor}(X,Y) = \rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$

1.3. Variáveis aleatórias independentes

X, Y variáveis aleatórias, espaço de probabilidade $(\Omega = \mathbb{R}^2, \mathcal{A}, P)$

- $X, Y \in \mathbb{R}$ dizem-se **independentes** se para quaisquer conjuntos $A, B \subset \mathbb{R}$ se tem

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

- Se f_X, f_Y e f_{XY} são as funções (densidade) de probabilidade marginal de X, marginal de Y e conjunta, respectivamente, então X e Y são independentes se e só se

$$f_{XY}(x, y) = f_X(x) f_Y(y), \quad \text{para qualquer } (x, y).$$

  Por outras palavras, a densidade conjunta do par (X, Y) é dada pelo produto das densidades marginais de X e de Y.

Observação: A definição anterior pode ser aplicada a variáveis contínuas ou discretas.

Teorema 1.4.1 *Sejam X e Y duas variáveis aleatórias e seja $c \in \mathbb{R}$ uma constante. Então*

(a) $E(cX) = cE(X)$

(b) $E(X+Y) = E(X) + E(Y)$

(c) se X e Y são independentes então

$$E(XY) = E(X)E(Y).$$

Prova:

(a) Suponhamos que X é uma variável aleatória discreta e seja f a sua função densidade de probabilidade. Então

$$E(cX) = \sum_x cxf(x) = c\sum_x xf(x) = cE(X).$$

No caso de X ser uma variável aleatória contínua, basta substituir os somatórios na dedução anterior por integrais (de $-\infty$ a $+\infty$).

(b) Suponhamos que X e Y são variáveis aleatórias discretas e seja $f(x,y)$ a sua função densidade de probabilidade conjunta. Então

$$\begin{aligned} &E(X+Y) \\ &= \sum_x \sum_y (x+y) f(x,y) \\ &= \sum_x x \left(\sum_y f(x,y) \right) + \sum_y y \left(\sum_x f(x,y) \right) \end{aligned}$$

i.e.,

$$\begin{aligned} E(X+Y) &= \sum_x x f_X(x) + \sum_y y f_Y(y) \\ &= E(X) + E(Y). \end{aligned}$$

No caso de X e Y serem variáveis aleatórias contínuas, basta substituir os somatórios na dedução anterior por integrais (de $-\infty$ a $+\infty$).

(c) Suponhamos que X e Y são variáveis aleatórias discretas e seja $f(x,y)$ a sua função densidade de probabilidade conjunta. Como X e Y são independentes

$$f(x,y) = f_X(x) f_Y(y).$$

Vem então

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy f(x,y) \\ &= \sum_x \sum_y xy f_X(x) f_Y(y) \\ &= \sum_x x f_X(x) \left(\sum_y y f_Y(y) \right) \\ &= E(X)E(Y). \end{aligned}$$

No caso de X e Y serem variáveis aleatórias contínuas, basta substituir os somatórios na dedução anterior por integrais (de $-\infty$ a $+\infty$).

Teorema 1.4.2 *Sejam X e Y duas v.a. Então*

(a) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ *(simetria)*

(b) $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$

(c) X, Y *independentes* $\Longrightarrow \mathrm{Cov}(X,Y) = 0$.

NOTA: A afirmação recíproca de (c) não é verdadeira: existem variáveis aleatórias com covariância nula que não são independentes.

Prova: A demonstração de (a) é trivial.

(b) Designemos a média de X (resp. Y) indiferenciadamente por $E(X)$ ou μ_X (resp. $E(Y)$ ou μ_Y). Pela definição de covariância entre X e Y e pelas propriedades da média tem-se

$$\begin{aligned}
&\mathrm{Cov}(X,Y)\\
&= E((X-\mu_X)(Y-\mu_Y))\\
&= E(XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y)\\
&= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y\\
&= E(XY) - E(Y)E(X)\\
&\quad - E(X)E(Y) + E(X)E(Y)\\
&= E(XY) - E(X)E(Y).
\end{aligned}$$

(c) Basta aplicar a alínea (b) anterior e o resultado (c) do teorema (1.4.1), $E(XY) = E(X)E(Y)$.

Corolário 1.4.3 *(a)* $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$

(b) $\mathrm{Var}(X) = E(X^2) - (E(X))^2$

Resulta ainda do teorema anterior que a covariância é uma função bilinear.

Corolário 1.4.4

(a) $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1,Y) + \mathrm{Cov}(X_2,Y)$

(b) $\mathrm{Cov}(X, Y_1 + Y_2) = \mathrm{Cov}(X,Y_1) + \mathrm{Cov}(X,Y_2)$

(c) $\mathrm{Cov}(cX,Y) = \mathrm{Cov}(X,cY) = c\,\mathrm{Cov}(X,Y)$.

Prova: Provamos apenas a alínea (a) sendo as demonstrações das outras alíneas absolutamente idênticas. Tem-se então

$$\begin{aligned}
&\mathrm{Cov}(X_1 + X_2, Y)\\
&= E((X_1+X_2)Y) - E(X_1+X_2)E(Y)\\
&= E(X_1Y) + E(X_2Y)\\
&\quad - E(X_1)E(Y) - E(X_2)E(Y)\\
&= \mathrm{Cov}(X_1,Y) + \mathrm{Cov}(X_2,Y).
\end{aligned}$$

Teorema 1.4.5 *Sejam X e Y duas variáveis aleatórias e c uma constante real. Então*

(a) $\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$

(b) $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X,Y)$

(c) se X e Y são independentes,

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

(d) $\mathrm{Cov}(X,c) = 0$

Prova:

(a)
$$\begin{aligned}\mathrm{Var}(cX) &= E[(cX - E(cX))^2] \\ &= E[(cX - cE(X))^2] \\ &= E[c^2(X - E(X))^2] \\ &= c^2 E[(X - E(X))^2] \\ &= c^2\,\mathrm{Var}(X).\end{aligned}$$

(b) Usando o facto de, para uma variável aleatória Z, se ter $\mathrm{Var}(Z) = E(Z^2) - E(Z)^2$, algumas propriedades da média e (a) do teorema (1.4.2) obtem-se

$$\begin{aligned}&\mathrm{Var}(X+Y) \\ &= E((X+Y)^2) - (E(X+Y))^2 \\ &= E(X^2 + 2XY + Y^2) - (E(X) + E(Y))^2 \\ &= E(X^2) + 2E(XY) + E(Y^2) \\ &\quad - (E(X))^2 - 2E(X)E(Y) - (E(Y))^2 \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(E(XY) - E(X)E(Y)) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).\end{aligned}$$

Também se podia ter deduzido o resultado a partir da igualdade $\mathrm{Var}(X_1 + X_2) = \mathrm{Cov}(X_1 + X_2, X_1 + X_2)$ e da bilinearidade da covariância.

A demonstração para $\mathrm{Var}(X - Y)$ é absolutamente análoga.

(c) Resulta da alínea (b) anterior e de (b) do teorema (1.4.2).

(d) Trivial.

A bilinearidade da covariância implica também os seguintes resultados (onde letras maíusculas designam, como de costume, variáveis aleatórias)

Corolário 1.4.6

(a)
$$\begin{aligned}&\mathrm{Cov}(X+Y, Z+W) \\ &= \mathrm{Cov}(X,Z) + \mathrm{Cov}(X,W) \\ &\quad + \mathrm{Cov}(Y,Z) + \mathrm{Cov}(Y,W)\end{aligned}$$

(b)
$$\begin{aligned}&\mathrm{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1^n} b_j Y_j\right) \\ &= \sum_{i,j} a_i b_j \mathrm{Cov}(X_i, X_j)\end{aligned}$$

(c)
$$\begin{aligned}&\mathrm{Var}(X_1 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) \\ &= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j)\end{aligned}$$

(d)
$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j);$$

em notação matricial,
$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = a^t \Sigma a$$

onde $a = (a_1, \ldots, a_n)^t$ *e* Σ *é a matriz* $n \times n$ *das covariâncias* $\mathrm{Cov}(X_i, X_j)$.

Exemplo 1: Calcule a covariância $\text{Cov}(X_i, X_j)$, para todos os pares (i, j), correspondente a uma distribuição multinomial com K categorias, $(X_1, \ldots, X_K) \sim Mult(n, (p_1, \ldots, p_K))$.

Começamos por observar que

$$X_i \sim B(n, p_i), \quad i = 1, ..., K.$$

Suponhamos primeiro que $i = j$. Tem-se

$$\text{Cov}(X_i, X_j) = \text{Cov}(X_i, X_i) = \text{Var}(X_i) = np_i(1 - p_i)$$

dado que a distribuição de X_i é $B(n, p_i)$.
Suponhamos agora que $i \neq j$. Da fórmula

$$\text{Var}(X_i + X_j) = \text{Var}(X_i) + \text{Var}(X_j) + 2\text{Cov}(X_i, X_j)$$

resulta que

$$\text{Cov}(X_i, X_j) = \frac{1}{2}\left(\text{Var}(X_i + X_j) - \text{Var}(X_i) - \text{Var}(X_j)\right).$$

Observando que $X_i + X_j \sim B(n, p_i + p_j)$, uma vez que $X_i + X_j$ conta os sucessos i e os sucessos j em n acontecimentos, teremos

$$\text{Cov}(X_i, X_j) = n(p_i+p_j)(1-(p_i+p_j)) - np_i(1-p_i) - np_j(1-p_j)$$

e portanto, para $i \neq j$,

$$\text{Cov}(X_i, X_j) = -np_ip_j.$$

Observar que a covariância é negativa. De facto, se houver uma grande quantidade de sucessos i em n acontecimentos, teremos necessariamente uma baixa quantidade de sucessos j nesses n acontecimentos, e vice-versa.

Exemplo 2: Seja X (resp. Y) o nível de creatinina em indivíduos de raça branca (resp. africana) com doença renal. Se $Z = \frac{X+Y}{2}$ (nível de creatinina médio de um indivíduo branco e de um indivíduo africano), $E(X) = 1.3$, $E(Y) = 1.5$ e $\text{Var}(X) = \text{Var}(Y) = 0.25$, então:

(a) $E(Z) = \frac{1}{2}(E(X) + E(Y)) = \frac{1}{2}(1.3 + 1.5) = 1.4$

(b) $\text{Var}(Z) = \frac{1}{4}(\text{Var}(X + Y)) = \frac{1}{4}(0.25 + 0.25)$
$= 0.125$, uma vez que X e Y são independentes.

Exemplo 3: Seja X a v.a. que representa o nível de creatinina em em indivíduos de raça branca com doença renal medido em Janeiro 2007, e Y a v.a. que representa o nível de creatinina dos mesmos indivíduos medido 1 ano depois.
A v.a. $Z = X - Y$ representa a variação ocorrida nos valores de creatinina no período de 1 ano.
Supondo $E(X) = E(Y) = 1.5$, $\text{Var}(X) = \text{Var}(Y) = 0.25$ e $\rho_{XY} = 0.5$, tem-se

(a) $E(Z) = E(X) - E(Y) = 0$

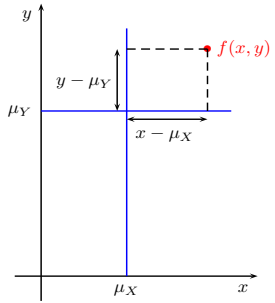
(b) $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$
$= 0.25 + 0.25 - 2(\sqrt{0.25}\sqrt{0.25}(0.5)) = .25$

1.4.1. Interpretação geométrica da covariância

Caso discreto: Para interpretar os valores positivos ou negativos da covariância entre duas v.a. X e Y discretas,

$$\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y)),$$

pode-se proceder da seguinte forma. Considere-se o referencial $(\mathcal{O}=(0,0),x,y)$ e nele o ponto (μ_X,μ_Y) como origem de novos eixos coordenados.



Em relação ao novo sistema de eixos tem-se

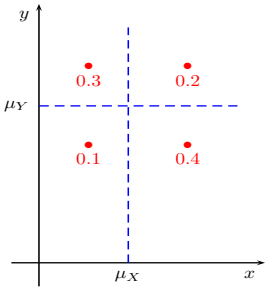
$(x-\mu_X)(y-\mu_Y) > 0$ nos $1^{\underline{o}}$ e $3^{\underline{o}}$ quadrantes
$(x-\mu_X)(y-\mu_Y) < 0$ nos $2^{\underline{o}}$ e $4^{\underline{o}}$ quadrantes.

Assim, podemos pensar que X e Y *variam da mesma forma* se existe uma probabilidade elevada de encontrar valores elevados de X (valores acima da média) associados a valores elevados de Y, ou valores abaixo da média de X associados a valores abaixo da média de Y (*i.e.*, os valores de (X,Y) com maior probabilidade de ocorrência estão nos primeiro e terceiro quadrantes). As diferenças $x-\mu_X$ e $y-\mu_Y$ para os pontos com maior probabilidade de ocorrência são positivas e portanto a covariância é positiva.
Pelo contrário, X e Y *variam de forma oposta* se existe uma grande probabilidade de valores baixos da média de X estarem associados a valores acima da média de Y, e vice-versa (portanto os valores de (X,Y) com maior probabilidade de ocorrência estão nos segundo e quarto quadrantes). As diferenças $(x-\mu_X)(y-\mu_Y)$ para os pontos com maior probabilidade de ocorrência são agora negativas e portanto a covariância é negativa.

Exemplo: Considere-se a v.a. bidimensional (X,Y) com função densidade de probabilidade dada por

	Y	
X	3	5
1	0.1	0.3
3	0.4	0.2

Os pontos com maior probabilidade de ocorrência estão nos segundo e quarto quadrantes portanto a covariância é negativa.



Caso contínuo: No caso contínuo procura-se identificar a região do plano cujos pontos possuem maior densidade de probabilidade, sendo então possível discernir se as v.a. X e Y mostram tendência mais ou menos intensa, do ponto de vista probabilístico, para variar no mesmo sentido ou em sentido oposto.
Se a variação for no mesmo sentido (valores elevados de X ocorrendo essencialmente associados a valores elevados de Y ou valores baixos de X ocorrendo essencialmente associados a valores baixos de Y) então a covariância é positiva; Se a variação for em sentido oposto (valores elevados de X ocorrendo essencialmente associados a valores baixos de Y ou valores baixos de X ocorrendo essencialmente associados a valores elevados de Y) então a covariância é negativa.

1.4.2. Covariância nula

Pelo teorema (1.4.2) sabemos que

$$X, Y \text{ independentes} \Longrightarrow \text{Cov}(X, Y) = 0.$$

Isto significa que

- se $\text{Cov}(X, Y) \neq 0$ então X e Y não são independentes
- se $\text{Cov}(X, Y) = 0$ então X e Y podem ou não ser independentes, *i.e.*, nada se pode concluir quando à sua independência.

Atente-se nos seguintes exemplos:

Exemplo 1. Considerem-se duas variáveis aleatórias discretas X e Y, com função densidade de probabilidade (f.d.p.) conjunta dada por

	Y	
X	3	5
1	0.3	0.3
3	0.2	0.2

Tem-se

$$f_X(1) = 0.6, \quad f_X(3) = 0.4$$
$$f_Y(3) = 0.5, \quad f_Y(5) = 0.5$$

donde

$$\begin{aligned} E(X) &= 1\, f_X(1) + 3\, f_X(3) = 1.8, \\ E(Y) &= 3\, f_Y(3) + 5\, f_Y(5) = 4, \end{aligned}$$

e

$$E(XY) \sum_{x,y} xy f(x, y) = 7.2$$

portanto

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0.$$

Numa situação deste tipo nada se pode concluir acerca da independência das duas variáveis. É preciso usar a definição. Ora

$$\begin{aligned} 0.3 = f(1,3) &= f_X(1) f_Y(3) = 0.6 \times 0.5 \\ 0.3 = f(1,5) &= f_X(1) f_Y(5) = 0.6 \times 0.5 \\ 0.2 = f(3,3) &= f_X(3) f_Y(3) = 0.4 \times 0.5 \\ 0.2 = f(3,5) &= f_X(3) f_Y(5) = 0.4 \times 0.5 \end{aligned}$$

logo X e Y são v.a. independentes.
Este é um exemplo em que $\text{Cov}(X, Y) = 0$ e X e Y são independentes.

Exemplo 2. A v.a. bidimensional (X, Y) com função densidade de probabilidade (f.d.p.)

$$f(x, y) = \begin{cases} \frac{5x^2 e^{-yx^2}}{(1+|x|)^{11}} & y > 0, \ -\infty < x < +\infty \\ 0 & \text{caso contrário} \end{cases}$$

tem covariância igual a zero (cálculos não apresentados aqui) mas não é possível obter a decomposição

$$f(x, y) = f_X(x) f_Y(y)$$

pelo que X e Y não são independentes.
Este é um exemplo em que $\text{Cov}(X, Y) = 0$ e X e Y não são independentes.

Exemplo 3 Considere-se $Z \sim N(0,1)$, $X = Z$ e $Y = Z^2$. Tem-se que

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E(XY) - E(X)E(Y) \\ &= E(Z^3) - E(Z)E(Z^2) \\ &= 0 \end{aligned}$$

porque todos os momentos ($E(Z^k)$) de ordem ímpar da normal reduzida $N(0,1)$ são nulos (pela simetria da distribuição). Contudo, naturalmente que X e Y são (muito!) dependentes.

Proposição 1.4.7 *Sejam X e Y são duas variáveis aleatórias independentes. Para quaisquer funções f e g tem-se*

$$E\left(f(X)g(Y)\right) = E\left(f(X)\right) E\left(g(Y)\right).$$

1.4.3. Covariância amostral

Considere-se uma amostra aleatória

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

de tamanho n de um par de variáveis aleatórias contínuas (X, Y).
A **covariância amostral** entre X e Y é dada por

$$\hat{\text{Cov}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

ou alternativamente, pode ser obtida de

$$\hat{\text{Cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \frac{1}{n^2} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i.$$

As fórmulas anteriores são o correspondente amostral de

$$\text{Cov}(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right)$$

e

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

respectivamente.

1.5. Coeficiente de correlação de Pearson

Sejam X e Y duas variáveis aleatórias (quantitativas). Duas situações extremas que podem acontecer são

- X e Y são independentes; neste caso

$$\text{Cov}(X,Y) = 0.$$

- $X = Y$; neste caso

$$\begin{aligned}\text{Cov}(X,Y) &= \text{Cov}(X,X) = \text{Var}(X)\\ &= \sigma_X\sigma_X = \sigma_X\sigma_Y.\end{aligned}$$

Uma vez que as unidades de $\text{Cov}(X,Y)$ são

$$(\text{unidades de } X) \times (\text{unidades de } Y)$$

torna-se difícil interpretar o grau de associação entre duas variáveis a partir da magnitude da covariância.

Uma quantidade adimensional [a] é o **coeficiente de correlação**, ρ_{XY}, definido por

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Também se usa a notação

$$\rho_{XY} = \rho = \text{Cor}(X,Y) = Corr(X,Y).$$

[a] portanto independente das unidades de X e de Y

A correlação é uma versão estandardizada da covariância e tem-se que

$$\text{Cor}(X,Y) = \text{Cor}\left(\frac{X-\mu_X)}{\sigma_x}, \frac{Y-\mu_Y}{\sigma_Y}\right).$$

De facto, pela bilinearidade da função de covariância,

$$\begin{aligned}&\text{Cor}\left(\frac{X-\mu_X}{\sigma_x}, \frac{Y-\mu_Y}{\sigma_Y}\right)\\ &= \text{Cov}\left(\frac{X-\mu_X}{\sigma_x}, \frac{Y-\mu_Y}{\sigma_Y}\right)\\ &= \frac{1}{\sigma_X\sigma_Y}\text{Cov}(X-\mu_X, Y-\mu_Y)\\ &= \frac{1}{\sigma_X\sigma_Y}\text{Cov}(X,Y)\\ &= \text{Cor}(X,Y).\end{aligned}$$

Exercício: Provar que a correlação é invariante por transformações lineares das variáveis aleatórias, i.e.,

$$\text{Cor}(X,Y) = \text{Cor}(aX+b, cY+d),$$

para quaisquer constantes reais a, b, c e d.

Teorema 1.5.1

$$-1 \leq \text{Cor}(X,Y) \leq 1.$$

Dem. Sem perda de generalidade, suponha que X e Y estão estandardizadas, para média 0 e desvio-padrão 1. Temos então

$$\begin{aligned} 0 \leq \mathrm{Var}(X+Y) &= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y) \\ &= 1+1+2\mathrm{Cov}(X,Y) \\ 0 \leq \mathrm{Var}(X-Y) &= \mathrm{Var}(X)+\mathrm{Var}(Y)-2\mathrm{Cov}(X,Y) \\ &= 1+1-2\mathrm{Cov}(X,Y) \end{aligned}$$

donde resulta o pretendido.

Ressalva-se aqui que uma correlação alta entre duas variáveis aleatórias não tem necessariamente a ver com uma relação de causalidade. Na verdade:

- uma correlação alta pode acontecer porque X causa Y
- uma correlação alta pode acontecer porque Y causa X
- uma correlação alta pode acontecer porque um terceiro factor, directa ou indirectamente, causa X e Y
- ocorreu um acontecimento improvável.

Teorema 1.5.2 *Para variáveis aleatórias X e Y com variância finita tem-se*

$$|E(XY)|^2 \leq E(X^2)E(Y^2).$$

Em particular,

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$$

Este resultado também é conhecido por **desigualdade de Cauchy-Schwarz em Teoria de Probabilidades**.

Dem.: A prova usa o facto de a esperança matemática do produto XY de duas variáveis aleatórias definir um produto interno no espaço das variáveis aleatórias, e a desigualdade de Cauchy-Schwarz usual da Álgebra Linear: para quaisquer dois vectores u e v num espaço de Hilbert

$$|\langle u, v, |\rangle \leq \|u\|\|v\|.$$

Em alternativa, sejam a e b constantes reais. Tem-se

$$E\left(aX+bY)^2\right) = a^2E(X^2)+2abE(XY)+b^2E(Y^2) \geq 0.$$

Se $E(X^2) > 0$, fazendo

$$a = -\frac{E(XY)}{E(X^2)} \quad e \quad b = 1$$

vem

$$-\frac{E(XY)^2}{E(X^2)} + E(Y^2) \geq 0$$

donde

$$(E(XY))^2 \leq E(X^2)E(Y^2).$$

Se $E(X^2) = 0$, escolhemos $\beta = 1$ e obtemos

Dem.: Provamos apenas (d), por ser mais difícil, sendo que as restantes provas devem ser encaradas como exercícios.
Sem perda de generalidade, suponhamos que X e Y são variáveis centradas (com média zero) - confirme que não há perda de generalidade nesta suposição.
A parte de que a relação linear entre X e Y implica um valor de 1 para o coeficiente de correlação linear resulta da bilinearidade da covariância e das propriedades do desvio padrão.
Suponhamos agora que $|\rho_{X,Y}| = 1$. Temos $\mathrm{Cov}(X,Y) = E(XY)$, $E(X^2) = \mathrm{Var}(X)$ e $E(Y^2) = \mathrm{Var}(Y)$. Pelo corolário da desigualdade de Cauchy-Schwarz obtem-se

$$|\mathrm{Cov}(X,Y)| = \sigma_X \sigma_Y$$

o que é equivalente a ter

$$P(aX + bY = 0) = 1, \quad \text{para quaisquer } a, b \in \mathbb{R}.$$

Re-escrevendo, e supondo $b \neq 0$,

$$P\left(Y = -\frac{a}{b}X\right) = 1$$

o que implica uma relação linear entre X e Y com probabilidade 1.

O coeficiente de correlação mede o grau de **associação linear** entre as variáveis;

- sinal de ρ_{XY}: indica se a correlação é positiva ou negativa (resp., se X e Y variam da mesma forma ou de forma oposta)
- valor absoluto de ρ_{XY}: mede a intensidade da associação linear.

Existem três situações importantes a distinguir:

- $\rho_{XY} = 0$: X e Y dizem-se não correlacionados e não existe associação linear entre X e Y. Contudo, se:
  - X e Y são linearmente relacionadas, ou
  - (X, Y) segue uma distrib. normal bivariada

  então $\rho_{XY} = 0$ implica independência.
- $\rho_{XY} \approx 1$

  existe uma associação linear positiva muito forte; essencialmente, X e Y da mesma forma
- $\rho_{XY} \approx -1$

  existe uma associação linear negativa muito forte; essencialmente, X e Y de forma oposta

Observação: quando a associação entre as variáveis não é linear ou pelo menos uma das variáveis é ordinal, o coeficiente de correlação não detecta possíveis associações. Nesse caso deve-se usar, por ex., o **coeficiente de correlação de Spearman**, que é uma estatística não paramétrica (que veremos mais tarde).

Exercício: Suponha que o vector aleatório $Z = (X, Y) \in \mathbb{R}^2$ segue uma distribuição normal bivariada, $Z \sim N(\mu, \Sigma)$, com $\mu \in \mathbb{R}^2$ e $\Sigma \in \mathbb{R}^2 \times \mathbb{R}^2$, digamos

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Recorde que se tem $E(X) = \mu_X$, $E(Y) = \mu_Y$, $\mathrm{Var}(X) = \sigma_X^2$, $\mathrm{Var}(Y) = \sigma_Y^2$ e $\mathrm{Cor}(X, Y) = \rho$.
A função densidade de probabilidade de $Z = (X, Y)$ é dada por

$$f(z) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(z-\mu)^t\Sigma^{-1}(z-\mu)\right).$$

Mostre que $\rho = 0$ implica X e Y independentes.

1.6. Coeficiente de correlação amostral de Pearson

Considere-se uma realização de uma amostra aleatória

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

de tamanho n de um par de variáveis aleatórias contínuas (X, Y).
O coeficiente de correlação linear **amostral** de Pearson associado à amostra é o estimador

$$\begin{aligned} & r_{xy}{}^{a} \\ &= \frac{SS_{xy}}{\sqrt{SS_{xx}}\sqrt{SS_{yy}}} \\ &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}} \\ &= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}\sqrt{\sum_{i=1}^n y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n y_i\right)^2}} \end{aligned}$$

do coeficiente de correlação de Pearson (populacional) entre x e y, onde

$$\begin{aligned} SS_{xx} &= \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2 \\ SS_{yy} &= \sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n y_i\right)^2 \end{aligned}$$

[a] também designado por $\widehat{\rho}$

$$\begin{aligned} SS_{xy} &= \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) \\ &= \sum_{i=1}^n x_i y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) \end{aligned}$$

são a variação amostral de x, a variação amostral de y, e a variação conjunta amostral entre x e y, respectivamente.[a]
Usando notação de álgebra linear, podemos escrever

$$r_{xy} = \frac{(x - \bar{x}I)^t (y - \bar{y}I)}{\|x - \bar{x}I\|_2 \|y - \bar{y}I\|}$$

onde

$$\|v\|_2 = \sqrt{\sum_{i=1}^n v_1^2 + \cdots + v_n^2}$$

para qualquer vector $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$.
Evidentemente:

(a) para vectores de observações x e y centrados tem-se

$$r_{xy} = \frac{x^t y}{\|x\|_2 \|y\|_2}$$

(b) para vectores de observações x e y estandardizados, tem-se

$$r_{xy} = x^t y.$$

[a] são variações, não variâncias; na verdade, são os numeradores das variâncias e covariância respectivas.

Proposição 1.6.1 *Seja* $(x_1, y_1), \ldots, (x_n, y_n)$ *uma realização de uma amostra aleatória do vector aleatório* (X, Y)*, e* a *e* b *constantes reais. Faça-se* $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *e* $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$*. Têm-se as seguintes propriedades:*

(a) $r_{x,x} = 1$

(b) $r_{x,y} = r_{y,x}$

(c) $r_{ax+b,y} = \text{sign}(a) r_{x,y}$

(d) se x *e* y *são vectores centrados, então*

$$r_{x,y} = \cos(\theta)$$

onde θ *é o ângulo definido pelos vectores* $x, y \in \mathbb{R}^n$*.*

(e) $-1 \leq r_{x,y} \leq 1$

(f) $|r_{x,y}| = 1$ *se e só se* $y = ax + b$ *para algumas constantes reais* a *e* b *com* $a \neq 0$*.*

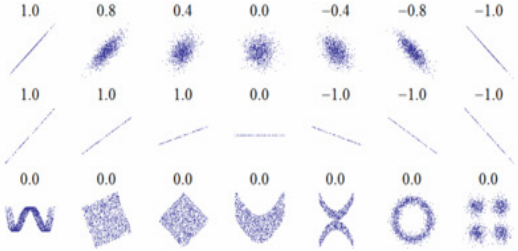
(g) $r_{x,y} = 0$ *se e só se* x *e* y *são vectores ortogonais.*

Dem.: A demonstração de (e) usa a desigualdade de Cauchy-Schwarz. Escrevendo $x^t y = \langle x, y \rangle$ e a desigualdade, ie,

$$|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2$$

vem o resultado.
A alínea (f) decorre de (d).

Alguns gráficos de dispersão e correspondentes coeficientes de correlação:



(Imagem retirada da Wikipedia)

Alguns autores propõem valores de corte para o coeficiente de correlação de forma a obter uma classificação qualitativa da correlação entre as v.a. em causa. Uma dessas regras empíricas é:

Correlação	$\lvert\rho_{XY}\rvert$
fraca	$[0.1, 0.3[$
moderada	$[0.3, 0.6[$
forte	$[0.6, 1[$

Contudo a interpretação do valor de ρ_{XY} deve depender essencialmente do contexto em que o problema é colocado.

1.7. Teste de correlação de Pearson

- Dados:
  - amostra aleatória $(x_1, y_1), \ldots, (x_n, y_n)$ de tamanho n de um par de variáveis aleatórias contínuas (X, Y)
  - a distribuição do par (X, Y) é aproximadamente uma **distribuição normal**. [a]

- Cálculos auxiliares: O coeficiente de correlação linear **amostral** de Pearson é o estimador

$$r_{xy}{}^{b} = \frac{SS_{xy}}{SS_{xx}SS_{yy}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

do coeficiente de correlação de Pearson entre X e Y, onde

$$\begin{aligned} SS_{xx}{}^{2} &= \sum_{i=1}^{n}(x_i - \overline{x})^2, \\ SS_{yy}{}^{2} &= \sum_{i=1}^{n}(y_i - \overline{y})^2, \\ SS_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \end{aligned}$$

são, respectivamente, a variação amostral de x, a variação amostral de Y e a variação conjunta de x e y amostral.

Nota: X, Y independentes $\Longrightarrow \mathrm{Cor}(X, Y) = 0$

[a] O teste continua válido se **pelo menos** uma das variáveis seguir uma distribuição normal

[b] também designado por $\widehat{\rho}$

- Pretende-se testar a hipótese nula

  H_0: $\mathrm{Cor}(X, Y) = 0$

  contra a alternativa

  $H_1 : \mathrm{Cor}(X, Y) \neq 0$

  (em particular, X e Y não são independentes)

  - Estatística de teste:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad \sim t(n-2)$$

  - Decisão: Rejeitar H_0 com nível de significância α se $|\, t \,| > t_{1-\alpha/2}(n-2)$

- Pretende-se testar a hipótese nula

  $H_0 : \mathrm{Cor}(X, Y) = \rho_0 \neq 0$

  contra a alternativa

  $H_1 : \mathrm{Cor}(X, Y) \neq \rho_0$

  - Estatística Z de Fisher:

$$z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \overset{a}{\sim} N(\mu_Z, \sigma_Z^2)$$

    onde

$$\mu_Z = \frac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right), \quad \sigma_Z = \frac{1}{\sqrt{n-3}}$$

  - Decisão: Rejeitar H_0 com nível de significância α se $|\frac{z-\mu_Z}{\sigma_Z}| > N_{1-\alpha/2}$

Se $n < 25$ a aproximação da distribuição da estatística de Fisher pela normal não é recomendada. Hotelling em 1953 sugeriu um outro procedimento que dá bons resultados para $n > 10$.

Sob a mesma hipótese nula

$H_0 : \text{Cor}(X, Y) = \rho_0 \neq 0$

contra a alternativa

$H_1 : \text{Cor}(X, Y) \neq \rho_0$

- Estatística de Hotelling: seja

$$Z_r^* = Z_r - \frac{3Z_r + r}{4n};$$

para $n > 10$ tem-se a distribuição assimptótica

$$Z^* = \frac{Z_r^* - Z_\rho^*}{1/\sqrt{n-1}} \overset{a}{\sim} N(0, 1)$$

O teste prossegue agora da forma usual.

Instruções em **R**:

```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         conf.level = 0.95, ...)
```

onde

- x e y representam as duas variáveis em causa
- *alternative* diz respeito à formulação da hipótese alternativa
- *method* indica o coeficiente de correlação a ser usado no teste

1.7.1. Exemplo 1

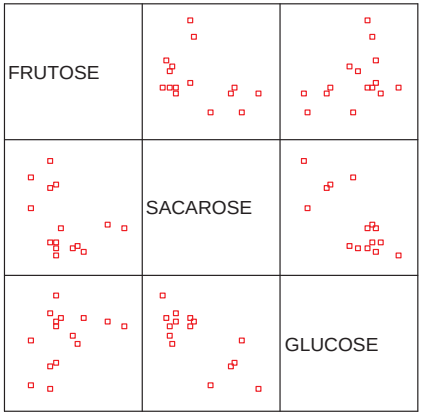
A tabela seguinte representa as concentrações (g l^{-1}) de frutose, sacarose e glucose para análises realizadas a 15 amostras de sumo de maçã de uma determinada marca.

Frutose	Sacarose	Glucose
40	20	6
49	27	11
47	26	10
47	34	5
40	29	16
49	6	26
47	10	22
51	14	21
49	10	20
49	8	19
55	8	17
59	7	21
68	15	20
74	14	19
57	9	15

Problema: Com um nível de significância de 0.05 para cada teste, testar correlação de Pearson nula para as variáveis acima mencionadas, duas a duas.

Correlação de Pearson (r)

	Frutose	Sacarose	Glucose
Frutose	1.000	-.380 p=.162	.373 p=.171
Sacarose		1.000	-.775 p=.001
Glucose			1.000

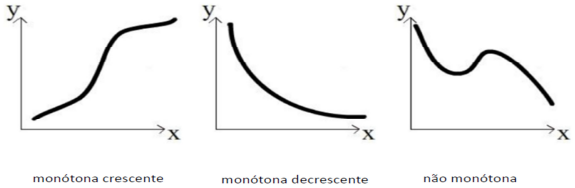


Resposta: Admitindo normalidade para os possíveis pares de variáveis podemos concluir o seguinte:

- Com um nível de significância de 0.05, a sacarose e a glucose têm correlação não nula (não sendo por isso variáveis independentes).
- Com um nível de significância de 0.05, não se pode rejeitar a hipótese de a frutose e a sacarose terem correlação nula.
- Com um nível de significância de 0.05, não se pode rejeitar a hipótese de a frutose e a glucose terem correlação nula.

1.8. Coeficiente de correlação de Spearman

Spearman (1904) desenvolveu um coeficiente de correlação ordinal que usa os ranks (postos, ordens) em vez dos valores exactos. Ao contrário do coeficiente de correlação de Pearson, que detecta tendências lineares, o coeficiente de correlação de Spearman identifica tendências monótonas, portanto mais gerais do que as primeiras.



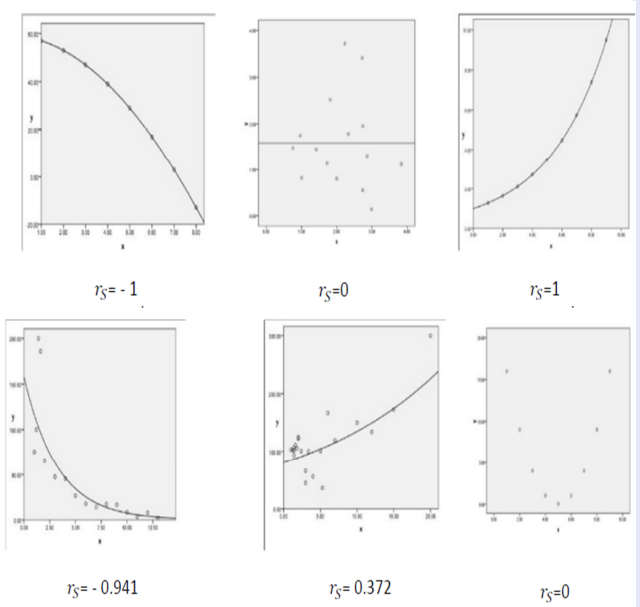
Resulta imediatamente do facto de o coeficiente de correlação de Spearman poder ser visto como um coeficiente de correlação de Pearson que

Proposição 1.8.1

$$-1 \leq r_S \leq 1$$

Quanto maior for o valor em módulo do coeficiente de correlação de Spearman, maior será a relação monótona entre as variáveis em causa. Um sinal positivo para r_S indica uma monotonia crescente enquanto que um sinal negativo sugere uma monotonia decrescente.

Apresentam-se agora os seguinte exemplos:



Acrescenta-se ainda que o coeficiente de correlação de Spearman é menos sensível à presença de valores extremos (*outliers*) do que o coeficiente de correlação de Pearson. A figura seguinte ilustra este ponto.

1.9. Teste de correlação de Spearman

Spearman (1904) desenvolveu um coeficiente de correlação ordinal que usa os ranks (postos, ordens) em vez dos valores exactos.

- Dados: amostra aleatória $(x_1, y_1), \ldots, (x_n, y_n)$ de tamanho n de um par de variáveis aleatórias ordinais ou contínuas (X, Y)
- Cálculos auxiliares:
  - Seja ξ_i a ordem dentro da amostra $x_1, \ldots, x_n$ da coordenada x_i do par (x_i, y_i).
  - Seja η_i a ordem dentro da amostra $y_1, \ldots, y_n$ da coordenada y_i do par (x_i, y_i).
  - Ficamos então com pares

$$(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n).$$

- O coeficiente de correlação de Spearman coincide com o coeficiente de correlação de Pearson dos ranks ξ e η, sendo portanto dado por

$$r_S = \frac{\sum_{i=1}^n \xi_i \eta_i - \frac{1}{n}\left(\sum_{i=1}^n \xi_i\right)\left(\sum_{i=1}^n \eta_i\right)}{\sqrt{\sum_{i=1}^n \xi_i^2 - \frac{1}{n}\left(\sum_{i=1}^n \xi_i\right)^2}\sqrt{\sum_{i=1}^n \eta_i^2 - \frac{1}{n}\left(\sum_{i=1}^n \eta_i\right)^2}}.$$

Quando não há empates ou quando os há em baixo número (usando os ranks médios para os empates), a fórmula anterior pode ser escrita na forma

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

onde $d_i = \eta_i - \xi_i$. Tem-se

$$-1 \leq r_S \leq 1.$$

Se as v.a. X e Y forem independentes então

$$r_S \approx 0.$$

- Pretende-se testar a hipótese nula

  H_0: As ordens de uma variável não se correlacionam com as ordens da outra variável: $\rho_S = 0$

  contra uma das alternativas:

  H_1 : As variáveis aleatórias X e Y não são independentes.

  H_1' : Existe uma tendência de valores elevados de X estarem associados a valores elevados de Y.

  H_1'' : Existe uma tendência de valores baixos de X estarem associados a valores baixos de Y.

- Estatística de teste:

$$r_S \sim \mathcal{S}(n)$$

sendo que a distribuição $S(n)$ encontra-se tabelada e não depende da distribuição das variáveis X e Y originais.

Para n grande e sob H_0 tem-se

$$\sqrt{n-1}\, r_S \overset{a}{\sim} N(0,1)$$

- Decisão:

  Rejeitar H_0 e aceitar H_1 com nível de significância α se $|\widehat{\rho}_S| > \mathcal{S}_{1-\alpha/2}(n)$

  Rejeitar H_0' e aceitar H_1' com nível de significância α se $\widehat{\rho}_S > \mathcal{S}_{1-\alpha}(n)$

  Rejeitar H_0'' e aceitar H_1'' com nível de significância α se $-\widehat{\rho}_S > \mathcal{S}_{1-\alpha}(n)$

Observação: Este teste não assume hipóteses sobre a distribuição do par (X, Y).

Instruções em **R**:

```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         conf.level = 0.95, ...)
```

onde

- x e y representam as duas variáveis em causa
- *alternative* diz respeito à formulação da hipótese alternativa
- *method* indica o coeficiente de correlação a ser usado no teste

1.9.1. Exemplo 2

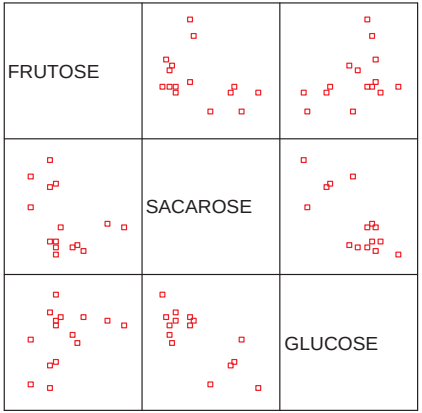
A tabela seguinte representa as concentrações (g l^{-1}) de frutose (F), sacarose (S) e glucose (G) para análises realizadas a 15 amostras de sumo de maça de uma determinada marca.

			Ordem			Diferenças		
F	S	G	F	S	G	d_{FS}	d_{FG}	d_{SG}
40	20	6	1.5	11	2	-1.5	-0.5	9
49	27	11	7.5	13	4	-5.5	3.5	9
47	26	10	4	12	3	-8	1	9
47	34	5	4	15	1	-11	3	14
40	29	16	1.5	14	6	-12.5	-4.5	8
49	6	26	7.5	1	14	6.5	-6.5	-13
47	10	22	4	6.5	13	-2.5	-9	-6.5
51	14	21	10	8.5	12.5	1.5	-2.5	-4
49	10	20	7.5	6.5	10.5	1	-3	-4
49	8	19	7.5	3.5	8.5	4	-1	-5
55	8	17	11	3.5	7	7.5	4	-3.5
59	7	21	13	2	12.5	11	0.5	-10.5
68	15	20	14	10	10.5	4	3.5	-0.5
74	14	19	15	8.5	8.5	6.5	6.5	0
57	9	15	12	5	5	7	7	0

Problema: Com um nível de significância de 0.05 para cada teste, testar correlação de Spearman nula para as variáveis acima mencionadas, duas a duas.

Correlação de Spearman ($\widehat{\rho}_S$)

	Frutose	Sacarose	Glucose
Frutose	1.000	-.477	.392
		p=.072	p=.148
Sacarose		1.000	-.679
			p=.005
Glucose			1.000



Resposta: Pode-se concluir o seguinte:

- Com um nível de significância de 0.05, a sacarose e a glucose têm correlação não nula (não sendo por isso variáveis independentes).
- Com um nível de significância de 0.05, não se pode rejeitar a hipótese de a frutose e a sacarose terem correlação nula.
- Com um nível de significância de 0.05, não se pode rejeitar a hipótese de a frutose e a glucose terem correlação nula.