Variáveis Aleatórias (continuação)   Prof. Waldecir Perrella

Distribuição Conjunta:

Para duas variáveis aleatórias *X* e *Y* define-se Função Distribuição Cumulativa CDF $F_{XY}(x,y)$ por:

$$P(X \leq x \quad e \quad Y \leq y) = F_{XY}(x,y) \tag{1}$$

e a Função Densidade de Probabilidade de Probabilidade PDF $f_{XY}(x,y)$ por:

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x\, \partial y} F_{XY}(x,y) \tag{2}$$

O evento de observar *X* no intervalo estar no intervalo (-∞, ∞) e *Y* estar no intervalo (-∞, ∞) é o evento certo e assim:

$$\int_{-\infty_1}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1 \tag{3}$$

A probabilidade de *X* estar no intervalo (x_1,x_2) e *Y* estar no intervalo (y_1,y_2) é dada por:

$$P(x_1 \leq X \leq x_2 \ e \ y_1 \leq Y \leq y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{XY}(x,y)dxdy \tag{4}$$

Quando se tem duas VAs *X* e *Y*, as PDFs individuais são chamadas de densidades marginais e são obtidas, respectivamente, por:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy \quad e \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx \tag{5}$$

O conceito de probabilidades condicionais pode ser aplicado para o caso de VA contínuas. Assim, define-se a PDF condicional $f_{X|Y}(x|y)$ como a densidade de probabilidade condicional de *x* dado o valor de *y*.

Os derivados para o caso discreto pode ser aplicado para PDF condicionais:

$$f_{X|Y}(x\,|\,y) f_Y(y) = f_{XY}(x,y) \tag{6}$$

$$f_{Y|X}(y\,|\,x) f_X(x) = f_{XY}(x,y) \tag{7}$$

$$f_{X|Y}(x\,|\,y) = \frac{f_{Y|X}(y\,|\,x) f_X(x)}{f_Y(y)} \tag{8}$$

A equação (8) é regra de Bayes para VAs contínuas. Quando se tem VAs mistas, isto é, discretas e contínuas, a regra de Bayes é dada por:

$$P_{X|Y}(x_i \mid y) f_Y(y) = f_{Y|X}(y \mid x_i) P_X(x_i) \tag{9}$$

onde X é uma VA discreta, Y é uma VA contínua, $P_X(x_i)$ é probabilidade de ocorrer o evento x_i e $P_{X|Y}(x_i|y)$ é a probabilidade de ocorrer o evento x_i dado que ocorreu o valor de y.

VAs contínuas X e Y são independentes se

$$f_{X|Y}(x \mid y) = f_X(x) \tag{10}$$

Aplicando a Regra de Bayes tem-se:

$$f_{XY}(x,y) = f_X(x) f_Y(y) \tag{11}$$

Valores Esperados e Correlação

Se $Z=g(X,Y)$ é uma função das VAs X e Y seu valor esperado é dado por:

$$E[g(X,Y)] = E(Z) = \int_{-\infty}^{\infty} z f_Z(z) dz = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dx\, dy \tag{12}$$

quando X e Y são contínuas e

$$E[g(X,Y)] = E(Z) = \sum_{i=-\infty}^{\infty} z_i P(z_i) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} g(x_i, y_j) P(x_i, y_j) \tag{13}$$

quando X e Y são discretas.

Define-se correlação das VAs X e Y por:

$$R_{XY} = E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f_{XY}(x,y) dx\, dy \tag{14}$$

quando X e Y são contínuas e

$$R_{XY} = E[XY] = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} x_i\, y_j\, P(x_i, y_j) \tag{15}$$

quando X e Y são discretas.

Define-se covariância das VAs X e Y por:

$$K_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{XY}(x,y) dx\, dy \tag{16}$$

quando X e Y são contínuas e

$$K_{XY} = E[(X-\mu_X)(Y-\mu_Y)] == \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}(x_i-\mu_X)(y_i-\mu_Y)P(x_i,y_j) \qquad 17$$

quando X e Y são discretas.

Expandindo a expressão de covariância tem-se:

$$K_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = R_{XY} - \mu_X\mu_Y \qquad 18$$

Se $K_{XY} = 0$ as VAs X e Y são ditas não correlacionadas e neste caso, usando a relação acima, $R_{XY} = E[XY] = \mu_X\mu_Y$. VAs estatisticamente independentes implica que $E[XY] = \mu_X\mu_Y$ e assim são sempre não correlacionadas, mas o contrário nem sempre é como é o caso de VAs gaussianas. Outro resultado importante é que a variância da soma de duas VAs e a soma das suas variâncias, quando elas são não correlacionas.

Define-se também, coeficiente de correlação r pela seguinte expressão:

$$r = \frac{K_{XY}}{\sigma_X\sigma_Y} \qquad 19$$

Pode-se mostrar as seguintes desigualdades:

$$-1 \le r \le 1 \qquad e \qquad [E(XY)]^2 \le E(X^2)E(Y^2) \qquad 20$$

Para entender-se o significado da definição de correlação e covariância, considere um experimento cuja saída é especificada por duas variáveis aleatórias X e Y. Considere ainda que não se tem nenhuma informação sobre o experimento. Nestas condições, será mostrado que é possível obter-se informações sobre a dependência ou independência entre as variáveis X e Y, observando-se as saídas de um grande número de realizações do experimento. Assim, será mostrado que um teste de correlação fornece alguma informação sobre a depend6encia entre a dependência entre variáveis.

Assim, considere o caso onde as variáveis X e Y são dependentes ou relacionadas tal que elas variam harmonicamente., isto é, se x aumenta, y aumenta e se x decresce, y decresce.

Como um exemplo considere a relação existente entre os números de publicações realizados por uma pessoa e suas respectivas notas, durante o período escolar. É razoável esperar que exista uma relação entre as duas quantidades. Para obter-se essa dependência é necessário estudar-se um grande número de cientistas e engenheiros para obter-se a nota e o número de publicações para cada pessoa. Isto pode ser considerado um experimento aleatório com as saídas x (Nota) e y (Número de publicações), onde essas saídas correspondem a uma realização deste experimento. Assim, pode-se obter um gráfico os pontos (x,y) para cada pessoa. Este gráfico é conhecido como diagrama de dispersão ou espalhamento (*scatter diagram*). Na Figura 1a tem-se um exemplo destes gráficos. Nesta figura observar-se que para um valor de x alto tem-se provavelmente um valor de y também alto. Observe o significado da palavra *provavelmente*. Não é sempre verdadeiro que para um y tem-se um alto x, mas isto será verdadeiro na maioria das vezes. Em outras palavras, existem poucos casos onde estudantes com notas baixas que produziram um número elevado de publicações e também poucos estudantes com altas notas e poucas publicações.

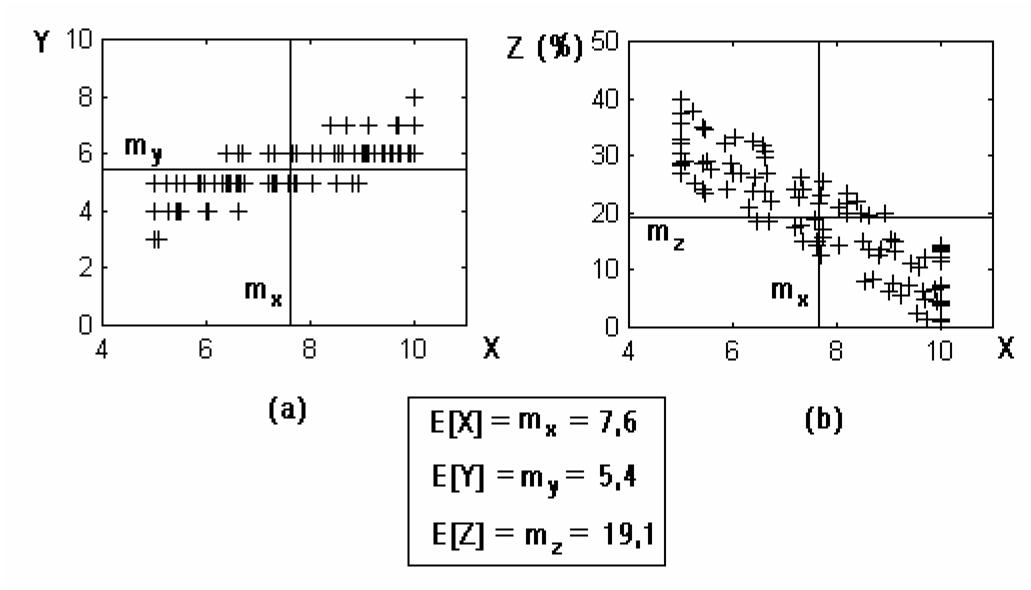

Figura 1 Diagrama de dispersão
(a).*X* = nota e *Y* = número de publicações
(b).*X* = nota e *Z* = porcentagem de faltas

Na Figura 1, também, são mostrados os valores médios das variáveis aleatórias envolvidas no experimento. Os valores médios foram estimados pela seguinte expressão:

$$\hat{E}[S] = \hat{m}_s = \frac{1}{N}\sum_{i=1}^{N} s_i \qquad (21)$$

onde *S* é a variável aleatória, o acento circunflexo significa estimativa e *N* é o número de pessoas observadas.

A variável aleatória $X - m_x$ representa a diferença entre a Nota real de cada pessoa e a Nota média do grupo observado e $Y - m_y$ representa o número real de publicações e a média de publicações do grupo observado. Em geral, uma pessoa, com uma nota acima da média, provavelmente, apresenta uma produção maior que o número médio de publicações. Obviamente, se $X - m_x$ é positivo então é mais provável que $Y - m_y$ seja positivo e se $X - m_x$ é negativo (abaixo da nota média), então é mais provável que $Y - m_y$ seja negativo (abaixo do número médio de publicações). Assim, a quantidade $(X - m_x) \bullet (Y - m_y)$ será positiva na maioria das realizações. Realizando-se este produto para cada pessoa e somando-se todos e então dividindo pelo número total de pessoas, obtêm-se uma estimativa do valor médio desse produto que é chamada de K_{XY}. A fórmula para a estimativa pode ser expressa por:

$$\hat{K}_{XY} = \hat{E}[(x-m_x)(y-m_y)] = \frac{1}{N}\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y) \qquad 22$$

Semelhantemente estima-se a correlação pela seguinte expressão:

$$\hat{R}_{XY} = \hat{E}[xy] = \frac{1}{N}\sum_{i=1}^{N} x_i y_i \qquad 23$$

Para o exemplo da Figura 1, obteve-se os seguintes valores, mostrados na Tabela 1

Tabela 1 Covariância, coeficiente de correlação e correlação para o exemplo da Figura 1

m_X	m_Y	m_Z	σ_X	σ_Y	σ_Z	K_{XY}	r_{XY}	R_{XY}	K_{XZ}	r_{XZ}	R_{XZ}
7,662	5,42	19,41	3,0	1,2	92	1,607	0,44	43,14	-15,17	-0,055	133,56

A seguir considera-se o caso onde as duas variáveis são relacionadas, mas variam em direção oposta. Como um exemplo deste tipo de correlação considere a relação entre as notas (variável *X*) e a percentagem de faltas (variável *Z*) dos estudantes. Neste caso, as variáveis mostram uma dependência no sentido negativo, isto é, quanto maior for a nota menor é porcentagem de faltas. O diagrama de dispersão é mostrado na Figura 1b. Assim, se $X - m_x$ é positivo então é mais provável que $Y - m_y$ seja negativo (abaixo da média de faltas) e se $X - m_x$ é negativo (abaixo da nota média), então é mais provável que $Y - m_y$ seja positivo (acima da média de faltas). Assim, a quantidade $(X - m_x)\bullet(Y - m_y)$ será negativa na maioria das realizações e a média desses produtos, K_{XY}, será negativa. Neste caso tem-se uma covariância negativa entre *X* e *Z*. Pode-se observar que a covariância negativa não significa que as variáveis não são relacionadas e sim dependentes, mas quando uma cresce e outra decresce e vice-versa. A Figura 1b ilustra esse caso e na Tabela 1 tem-se os valores estimados.

A seguir, considera-se o caso onde as variáveis *X* e W são tais que o valor de *X* não influencia no valor de *W*. Como um exemplo, considere a relação entre as Notas (Variável *X*) e o número que filhos (Variável W) de cada pessoa. É razoável esperar que as variações em *X* e *W* não apresente nenhum padrão de dependência. O diagrama de dispersão deve apresentar um padrão como mostrado na Figura 2.

Para o exemplo da Figura 2 obteve-se os valores mostrados na Tabela 2

Tabela 2 Covariância, coeficiente de correlação e correlação para o exemplo da Figura 2

m_X	m_W	σ_X	σ_W	K_{XW}	r_{XW}	R_{XW}
7,662	2,469	3,016	1,590	-0,1083	-0,0183	18,81

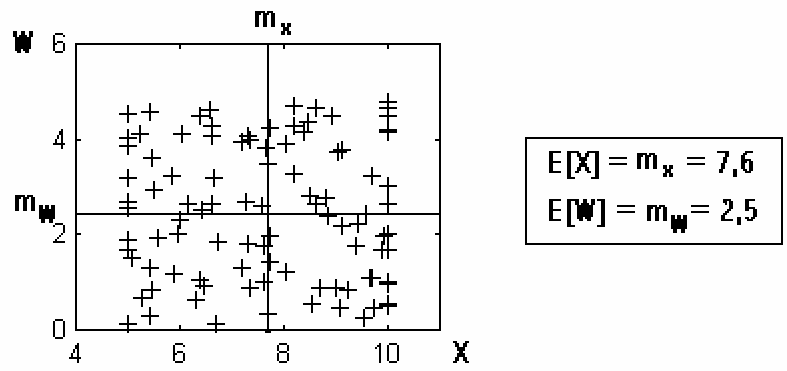

Figura 2. Diagrama de dispersão onde *X* = nota e *W* = número de filhos.

Neste caso, se $X - m_x$ é positivo então é igualmente provável que $Y - m_y$ seja negativo ou positivo. Assim, o produto $(X - m_x) \bullet (Y - m_y)$ tem a mesma chance de ser positivo e negativo média desses produtos, K_{XY}, será zero. Neste caso diz-se que as variáveis são não correlacionadas.