UNIVERSIDADE DE SÃO PAULO
Faculdade de Zootecnia e Engenharia de Alimentos

ZAB1111 – Estatística Básica

Prof. César Gonçalves de Lima   cegdlima@usp.br

Aula 11 – PROBABILIDADE (5)
VARIÁVEIS ALEATÓRIAS BIDIMENSONAIS

9. VARIÁVEIS ALEATÓRIAS BIDIMENSIONAIS

Nas pesquisas científicas é comum avaliar mais de uma característica (variável aleatória) nos indivíduos em estudo.

Quando avaliadas nos mesmos indivíduos essas variáveis tendem a apresentar correlação (positiva ou negativa), indicando dependência nos seus resultados.

Existem técnicas estatísticas específicas usadas no estudo simultâneo de duas ou mais variáveis quantitativas discretas ou contínuas.

Objetivo: Estudar um par de variáveis aleatórias discretas.

Exemplo 9.1. Em um levantamento de preços de dois dos ingredientes mais caros de uma ração experimental, obteve-se o preço de um quilograma do ingrediente A (variável X) e o preço de um quilograma do ingrediente B (variável Y). A distribuição conjunta de probabilidades resultante foi:

$x \setminus y$	4,00	5,00	6,00	$P(X = x)$
9,00	0,01	0,04	0,30	0,35
10,00	0,02	0,20	0,03	0,25
11,00	0,30	0,07	0,03	0,40
$P(Y = y)$	0,33	0,31	0,36	1

A partir desta distribuição conjunta podemos obter as distribuições marginais de probabilidades dos preços dos ingredientes A e B:

x	9	10	11
$P(X = x)$	0,35	0,25	0,40

$E(X)$= 10,05 reais
$var(X)$= 0,7475 reais^2
$dp(X) = 0,86$ reais

y	4	5	6
$P(Y = y)$	0,33	0,31	0,36

$E(Y) = 5,03$ reais
$var(Y) = 0,6891$reais^2
$dp(Y) = 0,83$ reais

Vamos iniciar os estudos relacionando os preços dos dois ingredientes, verificando se são independentes (ou não!).

9.1. DISTRIBUIÇÃO CONDICIONAL

Definição 9.1 A probabilidade condicional da v.a. X assumir o valor x_i, dado que a v.a. Y assume o valor k, é definida como:

$$P(X = x_i | Y = k) = \frac{P(X = x_i; Y = k)}{P(Y = k)}$$

para todos os valores da variável X.

A <u>esperança condicional</u> e a <u>variância condicional</u> da variável X dado que $Y = k$ são definidas como:

$$E(X|Y = k) = \sum_i x_i \, P(X = x_i | Y = k)$$

$$var(X|Y = k) = \sum_i [x_i - E(X|Y = k)]^2 \, P(X = x_i | Y = k)$$

Com os dados do Exemplo 9.1, obtenha a distribuição de probabilidades, a média e a variância condicional do preço de um kg do ingrediente A (variável X), sabendo-se que o preço de um kg do ingrediente B (variável Y) foi de R$ 5,00.

x_i	9	10	11
$P(X{=}x_i \mid Y{=}5)$	0,1290	0,6452	0,2258

em que $P(X{=}x_i|Y = 5) = \dfrac{P(X{=}x_i|Y{=}5)}{0,31}$ para $x_i = 9, 10$ e 11.

Com base nestas informações podemos obter:

$E(\mathrm{X} \mid \mathrm{Y} = 1) = 10{,}10$ reais

$var(\mathrm{X}|\mathrm{Y} = 5) = 102{,}1903 - 10{,}10^2 = 0{,}34$ reais^2

Ou seja, nos locais onde o ingrediente B custa R$5,00, espera-se encontrar um kg do ingrediente A custando, em média, R$10,10.

Uma medida de variabilidade deste preço é o seu desvio padrão, que é igual a $\sqrt{0{,}34} =$ R$0,59 (cinquenta e nove centavos).

Definição 9.2. As variáveis aleatórias X e Y que assumem os valores $x_1, x_2, \ldots$ e $y_1, y_2, \ldots$, respectivamente, são chamadas **independentes** se e somente se para todo par de valores (x_i, y_j) de X e Y, tem-se que:

$$P(X = x_i; Y = y_j) = P(X = x_i)P(Y = y_j)$$

Nota: Se a igualdade falhar para um par (x_i, y_j) já podemos concluir que as variáveis X e Y são dependentes.

Para saber se os preços dos dois ingredientes mais caros da ração são independentes, escolhemos o par de valores $(X = 9; Y = 4)$ e calculamos:

$$P(X = 9, Y = 4) = 0{,}01$$

$$P(X = 9) * P(Y = 4) = (0{,}35)*(0{,}33) = 0{,}1155$$

$$P(X = 9, Y = 4) \neq P(X = 9) * P(Y = 4)$$

$\therefore$ Como a igualdade já falhou para este par de valores, podemos concluir que as variáveis X e Y não são independentes!

$\therefore$ Existe uma relação de dependência linear entre os preços dos ingredientes A e B!

Problema: Como quantificar esta dependência?

9.2. COVARIÂNCIA E CORRELAÇÃO ENTRE VARIÁVEIS ALEATÓRIAS

Objetivo: Quantificar a possível relação de dependência entre duas variáveis aleatórias.

Definição 9.2. A **covariância** é uma medida da relação linear entre as variáveis X e Y e é definida por:

$$cov(X;\ Y) = E\{[X - E(X)][Y - E(Y)]\} = E(XY) - E(X)E(Y)$$

onde $E(XY) = \sum_i \sum_j x_i y_j\ P\left(\mathrm{X} = x_i, \mathrm{Y} = y_j\right)$.

- $cov(X;Y)$ pode assumir qualquer valor em $\mathbb{R}$ (números reais), ou seja, $-\infty < cov(X;\ Y) < \infty$.
- Não dá para saber pelo valor de $cov(X;Y)$ se o grau de dependência entre X e Y é grande (forte) ou pequeno (fraco).

- $cov(X;Y) > 0$ indica que as variáveis X e Y são diretamente proporcionais.
- $cov(X;Y) < 0$ indica que as variáveis X e Y são inversamente proporcionais.
- $cov(X;Y) = 0$ indica que X e Y não são correlacionadas.

Observe que:

- Se X e Y são independentes $\Rightarrow E(XY) = E(X)E(Y) \Rightarrow cov(X;Y) = 0$, ou seja, duas v.a.'s independentes têm covariância nula.
- Mas se $cov(X;Y) = 0$ não podemos garantir que X e Y sejam independentes. Neste caso só podemos dizer que X e Y não são correlacionados.

No Exemplo 9.2 temos que:

$$E(XY) = 49{,}99 \Rightarrow cov(X,Y) = 49{,}99 - (10{,}05)(5{,}03) = -0{,}5615$$

Como $cov(\text{X; Y}) < 0$ nós concluímos que os preços dos dois ingredientes são <u>inversamente proporcionais</u>, quando o preço de um dos ingredientes é alto, o preço do outro ingrediente tende a ser mais baixo.

Para quantificar o grau de relacionamento entre duas variáveis é mais comum utilizarmos o **coeficiente de correlação linear**.

Definição 9.4. O coeficiente de correlação linear é uma medida do grau de dependência linear entre duas v.a.'s X e Y, sendo definido como:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)\,var(Y)}}, \quad \text{com } -1 \leq \rho(X,Y) \leq 1$$

- A dependência linear entre as variáveis X e Y é perfeita quando $\rho(X,Y) = +1$ ou quando $\rho(X,Y) = -1$
- Quanto mais próximo de $+1$ (ou de -1) estiver o valor de $\rho(X,Y)$, maior e positivo (ou negativo) é o grau de dependência entre as duas variáveis.
- Quando $\rho(X,Y) = 0$ concluímos que não existe qualquer relação de dependência entre as variáveis X e Y. Dizemos que X e Y não são correlacionadas.

x	y	$P(x;y)$
9	4	0,01
9	5	0,04
9	6	0,30
10	4	0,02
10	5	0,20
10	6	0,03
11	4	0,30
11	5	0,07
11	6	0,03

No exemplo dos preços os ingredientes, temos:

$$\rho(X,Y) = \frac{-0{,}5615}{\sqrt{(0{,}7475)(0{,}6891)}} = -0{,}7824$$

Existe uma forte correlação linear negativa entre os preços dos dois ingredientes mais caros usados na ração de frangos, indicando que em locais em que um dos ingredientes é mais caro, o outro tende a ser mais barato.

Exemplo 9.2. Em um estudo sobre a rotatividade de mão de obra especializada na lavoura foram avaliadas as variáveis X: "número de empregos que o trabalhador teve nos cinco últimos anos" e Y: "salário atual, em número de salários mínimos".

Pergunta: O salário atual de um trabalhador na lavoura depende do número de empregos do trabalhador nos últimos cinco anos?

$y \backslash x$	1	2	3	4	$P(Y = y)$
3	0	0	0,10	0,10	0,20
5	0,05	0,05	0,10	0,10	0,30
7	0,05	0,20	0,05	0	0,30
10	0,10	0,05	0,05	0	0,20
$P(X = x)$	0,20	0,30	0,30	0,20	1,00

Utilizando as fórmulas já conhecidas obtemos:

- $E(X) = 2{,}5$ empregos $\qquad E(Y) = 6{,}2$ salários mínimos
- $var(X) = 1{,}05 \qquad var(Y) = 5{,}56$
- $E(XY) = 14{,}05 \quad \Rightarrow cov(X;Y) = 14{,}05 - (2{,}5)(6{,}2) = -1{,}45$

$$\Rightarrow \rho(X,Y) = \frac{-1{,}45}{\sqrt{(1{,}05)(5{,}56)}} = -0{,}6001$$

Existe uma <u>correlação negativa</u> e <u>relativamente alta</u> entre o salário atual e o número de empregos dos trabalhadores na lavoura nos últimos cinco anos.

Existe uma tendência de menores salários para trabalhadores com maior número de empregos nos últimos cinco anos, ou uma tendência de valorizar com maiores salários os trabalhadores com menor número de empregos nos últimos cinco anos.

9.3. FUNÇÕES DE VARIÁVEIS ALEATÓRIAS

Vamos estudar a distribuição de probabilidades das seguintes funções de variáveis aleatórias discretas:

$S = X + Y$ (soma)

$D = X - Y$ (diferença)

$V = XY$ (produto)

Com este estudo será possível verificar algumas propriedades interessantes sobre a média dessas novas variáveis, como:

- $E(S) = E(X + Y) = E(X) + E(Y)$
- $E(D) = E(X - Y) = E(X) - E(Y)$
- Quando X e Y são independentes: $E(V) = E(XY) = E(X)E(Y)$

Outras propriedades interessantes que poderão ser comprovadas, referem-se à variância de funções de variáveis:

- $var(S) = var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$.
- $var(D) = var(X - Y) = var(X) + var(Y) - 2cov(X, Y)$.
- Quando X e Y <u>são independentes</u>: $var(X + Y) = var(X - Y) = var(X) + var(Y)$

Vamos confirmar essas propriedades (da média e da variância) com as variáveis X e Y do Exemplo 9.1, em que X: "preço de um quilograma do ingrediente A" e Y: "preço de um quilograma do ingrediente B".

A distribuição conjunta de probabilidades de X e Y é:

$x \setminus y$	4,00	5,00	6,00	$P(X=x)$
9,00	0,01	0,04	0,30	0,35
10,00	0,02	0,20	0,03	0,25
11,00	0,30	0,07	0,03	0,40
$P(Y=y)$	0,33	0,31	0,36	1

Para obter as distribuições de probabilidades das novas variáveis, devemos organizar essas informações de uma forma diferente.

1) Obtenção das distribuições de probabilidades das novas variáveis:

x	y	$P(x;y)$	$S = X+Y$	$D = X-Y$	$V = XY$
9	4	0,01	13	5	36
9	5	0,04	14	4	45
9	6	0,30	15	3	54
10	4	0,02	14	6	40
10	5	0,20	15	5	50
10	6	0,03	16	4	60
11	4	0,30	15	7	44
11	5	0,07	16	6	55
11	6	0,03	17	5	66

As distribuições de probabilidades das novas variáveis aleatórias são:

Soma:

$s = x + y$	13	14	15	16	17
$P(S = s)$	0,01	0,06	0,80	0,10	0,03

$E(\mathrm{S}) = 15{,}08$

$var(\mathrm{S}) = 0{,}3136$

Diferença:

$d = x - y$	3	4	5	6	7
$P(D = d)$	0,30	0,07	0,24	0,09	0,30

$E(\mathrm{D}) = 5{,}02$

$var(\mathrm{D}) = 2{,}5596$

Produto:

$v = xy$	36	40	44	45	50	54	55	60	66
$P(V = v)$	0,01	0,02	0,30	0,04	0,20	0,30	0,07	0,03	0,03

$E(\mathrm{V}) = 49{,}99$ $\qquad$ $var(\mathrm{V}) = 32{,}9899$

2) Verificando as propriedades:

Da média:

$$E(\mathrm{S}) = E(X) + E(Y) = 10{,}05 + 5{,}03 = 15{,}08$$

$$E(\mathrm{D}) = E(X) - E(Y) = 10{,}05 - 5{,}03 = 5{,}02$$

Da variância:

$$var(\mathrm{S}) = var(\mathrm{X+Y}) = 0{,}7475 + 0{,}6891 + 2(-0{,}5615) = 0{,}3136$$

$$var(\mathrm{D}) = var(\mathrm{X-Y}) = 0{,}7475 + 0{,}6891 - 2(-0{,}5615) = 2{,}5596$$

Perceba que estes resultados coincidem com aqueles obtidos a partir das distribuições de probabilidades.

Mais exemplos de funções de variáveis aleatórias:

$Z_1 = aX + b$ $\quad E(Z_1) = aX + b$ $\quad var(Z_1) = a^2 var(X)$

$Z_2 = aX + bY$ $\quad E(Z_2) = aE(X) + bE(Y)$

$$var(Z_2) = a^2 var(X) + b^2 var(Y) + 2ab\,cov(X, Y)$$

Lembrete:

$$var(X) = E(X^2) - [E(X)]^2$$