3.3.5. Modelo de distribuição binomial

O modelo mais importante de VAD será para nós o modelo *binomial*, que surge quando um experimento de Bernoulli é repetido um número fixo de vezes, e contamos o número de sucessos obtidos.

3.3.5.1. Introdução

Para mostrar a origem destes modelo, retornaremos ao problema do lançamento de três dados.

Exemplo 1: lançamento de três dados

Um dado é lançado três vezes (ou três dados são lançados simultaneamente, o que dá no mesmo) e contamos o número X de vezes em que aparece a face 6. Para calcular as probabilidades por meio de técnicas de enumeração, usamos um diagrama de árvore (seção **3.1.6**) e a partir dele calculamos a distribuição de probabilidades na Tab. **1**.

Tabela 1. Distribuição de probabilidades no lançamento de três dados (X: número de dados que mostram a face 6)

X	p(x)
0	0,5787
1	0,3472
2	0,0694
3	0,0046
Σ	1,0000

A partir da tabela, podemos calcular o valor esperado e a variância da variável X:

$E(X) = 0 \times 0{,}5787 + 1 \times 0{,}3472 + 2 \times 0{,}0694 + 3 \times 0{,}0046 = 0{,}5$

$V(X) = (0-0{,}5)^2 \times 0{,}5787 + (1-0{,}5)^2 \times 0{,}3472 + (2-0{,}5)^2 \times 0{,}0694 + (3-0{,}5)^2 \times 0{,}0046$
$= 0{,}4167$

Tornaremos o problema um pouco mais complicado, aumentando o número de dados.

Exemplo 2: lançamento de quatro dados

Lançamos quatro dados simultaneamente e contamos o número X de vezes em que aparece a face 6. Usando a notação do modelo de Bernoulli, definimos:

sucesso: face 6
fracasso: qualquer outra face
p=P(*sucesso*)=1/6
q=P(*fracasso*)=1−p=5/6

A variável de interesse será X: número de dados que mostram a face 6.

Calcular a probabilidade de todos os dados mostrarem a face 6 (isto é, X=4) é relativamente fácil:

$$P(X=4) = p\times p\times p\times p = (1/6)^4 \approx 0{,}0008$$

A probabilidade de não ocorrer nenhum sucesso também é fácil de calcular:

$$P(X=0) = q\times q\times q\times q = (5/6)^4 \approx 0{,}4823$$

Para calcular a probabilidade P(X=1) de obtermos exatamente 1 sucesso (nem mais, nem menos) nas quatro tentativas, temos que listar as várias seqüências de lançamentos que podem resultar em *um* sucesso e *três* fracassos:

SFFF, FSFF, FFSF, FFFS

As probabilidades destes ramos são:

P(SFFF) = **p**qqq
P(FSFF) = q**p**qq
P(FFSF) = qq**p**q
P(FFFS) = qqq**p**

Note que todos estes ramos têm a mesma probabilidade, $p^1q^3 = (1/6)(5/6)^3 \approx 0{,}0965$. Como são quatro ramos,

$$P(X=1) = 4\times(1/6)(5/6)^3 \approx 0{,}3858$$

Para calcularmos a probabilidade P(X=2) de dois sucessos, enumeramos os ramos que resultam em dois sucessos:

SSFF, SFSF, SFFS, FSSF, FSFS, FFSS

Todos estes ramos têm dois sucessos e dois fracassos, apenas em posições diferentes (o que não altera as probabilidades). A probabilidade de cada ramo será portanto:

$$p^2q^2 = (1/6)^2(5/6)^2 = 25/1296 \approx 0{,}0193$$

Como são seis ramos de igual probabilidade,

$$P(X=2) = 6\times(1/6)^2(5/6)^2 \approx 6\times 0{,}0193 \approx 0{,}1157$$

A probabilidade P(X=3) pode ser calculada agora a partir das outras já calculadas:

$$P(X=3) = 1 - [P(X=0)+P(X=1)+P(X=2)+P(X=4)] \approx 0{,}0154$$

A distribuição de probabilidades resultante está mostrada na Tabela **2**.

Tabela 2. Probabilidades no lançamento de quatro dados (X: nº. de dados que mostram a face "6")

X	p(x)
0	0,4823
1	0,3858
2	0,1157
3	0,0154
4	0,0008
Σ	1,0000

Neste exemplo, podemos ver que cada seqüência de sucessos e fracassos (cada ramo da árvore) tem como probabilidade

$$p^{(\text{número de sucessos})}q^{(\text{número de fracassos})}$$

Se fizermos

n: número de tentativas
x: número de sucessos,
$n - x$: número de fracassos

podemos escrever que cada seqüência terá probabilidade igual a $p^x q^{(n-x)}$. O problema agora é calcular quantas destas seqüências existem, para cada valor de x. Para isto, usamos análise combinatória: em cada ramo há n posições (no exemplos acima, n dados); se destas n posições quero escolher x para os sucessos e $n - x$ para os fracassos, quantas possibilidades tenho para esta escolha? O número de possibilidades será o número de combinações das n posições x a x, dado por:

$$C_n^x = \frac{n!}{x!(n-x)!}$$

onde $n!$ (fatorial de n) é dado por:

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$$

Por exemplo, se são n=4 dados, o número de maneiras em que podem ser obtidos x=2 sucessos será:

$$C_n^x = \frac{n!}{x!(n-x)!} = \frac{4!}{2!(4-2)!} = 6$$

Para cada número x de sucessos existem C_n^x ramos, e a probabilidade de cada ramo é dada por $p^x q^{(n-x)}$; o modelo para cálculo das probabilidades será portanto obtido pela multiplicação da probabilidade de cada ramo pelo número de ramos:

$$P(X = x) = C_n^x . p^x . q^{(n-x)}$$

3.3.5.2. Função de probabilidades e parâmetros

O modelo binomial tem portanto sua função de probabilidades dada pela expressão:

$$P(X = x) = C_n^x p^x q^{(n-x)}$$

cujos parâmetros são:

n número de repetições
p probabilidade de sucesso em cada repetição

Note que o valor *q* que aparece na fórmula não é considerado um dos parâmetros, porque ele é dado simplesmente por $q=1-p$. Alguns livros, para manter a notação mais coerente, preferem escrever o modelo como:

$$P(X=x)=C_n^x p^x (1-p)^{(n-x)}$$

Este modelo é representado abreviadamente por *B(n,p)*; a letra B de "binomial" e os dois parâmetros *n* e *p*. (O modelo é chamado de "binomial", porque sua fórmula é a do *binômio de Newton*, usado na Álgebra para calcular potências da soma de dois termos). Para dizer que uma variável X segue o modelo binomial *B(n,p)*, usamos a notação:

$X \sim B(n,p)$

Pode ser demonstrado (mas isto não será feito aqui) que o valor esperado e a variância de uma variável que segue este modelo podem ser calculados em função dos parâmetros, da forma:

$E(X) = np \qquad V(X) = npq$

O modelo binomial pode ser usado em problemas que têm em comum estas características:

(i) Compõem-se de um número fixo *n* de repetições de um *experimento de Bernoulli*. Cada repetição é chamada de uma *tentativa*;

(ii) Cada tentativa pode resultar em dois resultados. O resultado que nos interessa contar é geralmente chamado de *sucesso*, o outro de *fracasso* (no problema acima, obtemos um *sucesso* quando o dado mostra a face 6, e *fracasso* quando mostra qualquer outra face);

(iii) As tentativas são *probabilisticamente independentes*. Isto quer dizer que o resultado de uma tentativa não afeta as probabilidades da tentativa seguinte, e que as probabilidades de *sucesso* e de *fracasso* serão constantes em todas as tentativas. Representamos a probabilidade de sucesso em cada tentativa por *p* e a probabilidade de fracasso por *q*;

(iv) A variável de interesse é *X*, o número de *sucessos* obtidos dentro das *n* tentativas.

Exemplo 1 (cont.) – Lançamento de três dados

Voltando ao Exemplo 1, do lançamento de três dados, e usando as fórmulas acima:

$n=3$
X: número de dados que mostram a face "6"
Sucesso: face "6" → $p=P(S)=1/6$
Fracasso: qualquer outra face → $q=P(F)=1-p=5/6$
$E(X) = np = 0{,}5 \qquad V(X) = npq = 3 \times 1/6 \times 5/6 \approx 0{,}4167$

Exemplo 2 (cont.) – Lançamento de quatro dados

Voltando ao Exemplo 2 acima, do lançamento de quatro dados:

n=4, p=1/6

E(X) = np = 4 × 1/6 ≈ 0,67

V(X) = npq = 4 × 1/6 × 5/6 ≈ 0,56

3.3.5.3. Simetria ou assimetria da distribuição binomial

A simetria ou assimetria da distribuição binomial é controlada pelo parâmetro p, a probabilidade de sucesso em cada tentativa. Se p=0,5, como num problema de *cara ou coroa*, a distribuição é simétrica. Se são feitas n=10 tentativas, o valor de X com maior probabilidade será X=5, e as probabilidades decrescem simetricamente à medida que o valor de X se afasta de 5:

P(X=4) = P(X=6)

P(X=3) = P(X=7), etc.

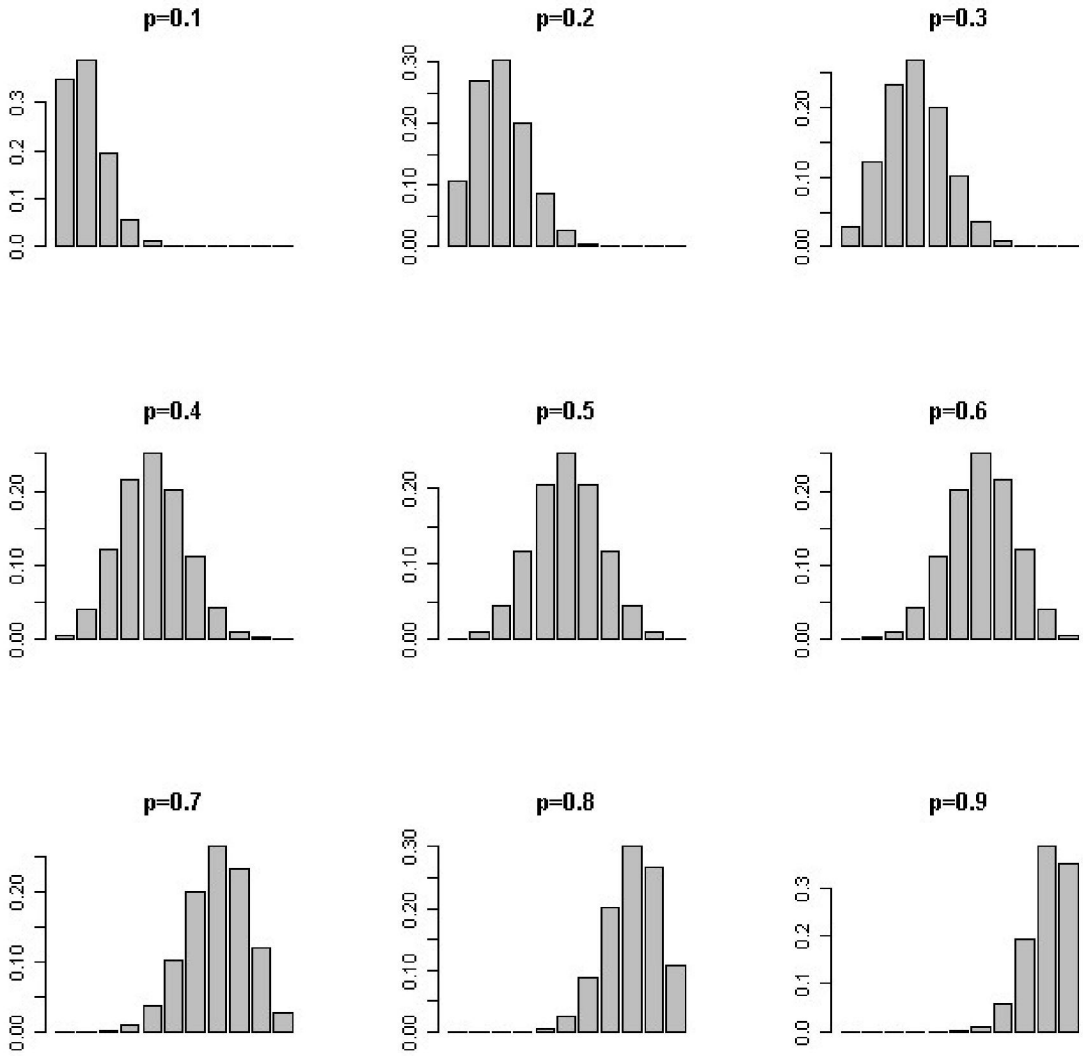

Figura 1. Distribuições binomiais para n=10, diferentes valores de p

Esta simetria é devida a uma propriedade das combinações, mencionada na seção **3.1.4.4**:

$$C_n^r = C_n^{n-r}$$

Portanto,

$$C_{10}^4 = C_{10}^6$$

$$C_{10}^3 = C_{10}^7 \text{ , etc.}$$

A Fig. **1** mostra distribuições binomiais para n=10, e nove valores diferentes de *p*. Se $p < 0{,}5$, a distribuição será assimétrica positiva, como nos gráficos da primeira linha da figura; se $p > 0{,}5$, será assimétrica negativa, como nos gráficos da terceira linha. Note que a distribuição com p=0,1 é a imagem refletida num espelho da distribuição com p=0,9; a distribuição com p=0,2 é a imagem refletida da distribuição com p=0,8, etc.

Exemplo 3 – Número de meninas em famílias de 12 crianças

A Tabela **3**, bastante conhecida na literatura de Estatística, mostra os resultados de um levantamento feito entre 6115 famílias que tinham 12 crianças, na Saxônia (um dos estados da Alemanha). O número *X* de meninas em cada família foi anotado, e a distribuição de frequências desta variável está nas colunas 2 e 3 da tabela (freqüências observadas e relativas). As colunas 4 e 5 mostram a distribuição de probabilidades de X dada por um modelo binomial B(n=12, p=0,5), e o número esperado de famílias em cada linha, calculado a partir destas probabilidades.

Tabela 3. Número de meninas em famílias de 12 crianças na Saxônia (Geissler, 1889)

	valores observados		valores calculados, supondo p = 0,5	
***X*: número de meninas**	**freqüência absoluta**	**freqüência relativa**	**probabilidade**	**freqüência esperada**
0	7	0.0011	0.0002	1
1	45	0.0074	0.0029	18
2	181	0.0296	0.0161	99
3	478	0.0782	0.0537	328
4	829	0.1356	0.1208	739
5	1112	0.1818	0.1934	1183
6	1343	0.2196	0.2256	1379
7	1033	0.1689	0.1934	1183
8	670	0.1096	0.1208	739
9	286	0.0468	0.0537	328
10	104	0.0170	0.0161	99
11	24	0.0039	0.0029	18
12	3	0.0005	0.0002	1
	6115	1	1	6115

Na tabela, podemos ver que as frequências de *X* previstas pelo modelo binomial são bastante próximas das observadas na amostra, mas há algumas discrepâncias. O gráfico da Fig. **2**, mostra que a distribuição da variável não é totalmente simétrica (como deveria ser se p=0,5); existem mais famílias com 5 meninas do que com 7, mais famílias com 4 meninas do que com 8, etc. Há *mais* meninas do lado esquerdo do gráfico do que seria esperado se as probabilidades de nascimento fossem iguais, e *menos* meninas do lado direito; isto

indica que a média desta distribuição será menor do que a média teórica dada pelo modelo. De fato, a média de *X* observada foi de 5,77, menor do que o valor esperado do modelo, que é

$$E(X) = np = 12 \times 0{,}5 = 6$$

Nasciam, portanto, menos meninas do que meninos. Esta diferença na verdade já era esperada, pois é um fato, bem conhecido dos demógrafos, que nascem mais meninos do que meninas, em quase todos os países do mundo. A razão entre o número de nascimentos de meninos em relação ao de meninas é geralmente medida pela *razão de sexo* (*sex ratio*), que é o número de meninos nascidos para cada 100 meninas, num intervalo de tempo; a média mundial é de cerca de 105/100. (Veremos este assunto novamente na Seção **4.4**)

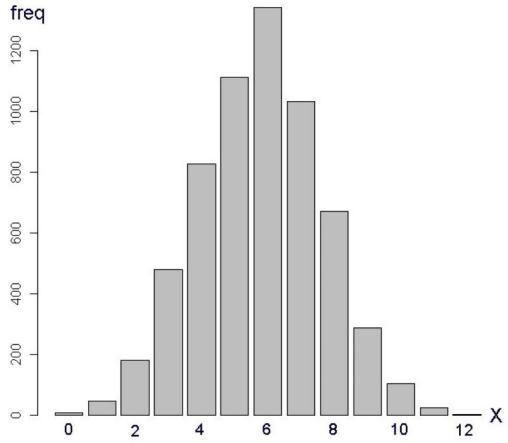

Figura 2. Distribuição do número de meninas em famílias com 12 crianças (Saxônia, 1889)

3.3.5.4. Aproximações da distribuição binomial, para *n* grande

A distribuição binomial é freqüentemente usada em problemas onde as amostras são muito grande. Por exemplo, numa pesquisa para estimar a intenção de votos de eleitores, antes de uma eleição onde haja apenas dois candidatos; ou numa pesquisa para avaliar a porcentagem de crianças de uma região que tomaram uma certa vacina. Em ambos exemplos, as amostras serão muito grandes, da ordem de milhares de elementos. Isto traz problemas para os cálculos, porque a função de probabilidades

$$P(X = x) = C_n^x . p^x . q^{(n-x)}$$

utiliza combinações, e estas se baseiam em fatoriais:

$$C_n^x = \frac{n!}{x!(n-x)!}$$

Os fatoriais porém não podem ser calculados para números muito grandes. O fatorial de 69, por exemplo, é o maior que pode ser obtido na maioria das calculadoras científicas.

$$69! = 1.71 \times 10^{98}$$

(Para dar uma idéia da dimensão deste número: os físicos estimam que o número de átomos existentes no universo é da ordem de *apenas* 10^{78}...). Existem maneiras de calcular

aproximadamente fatoriais de números maiores, mas mesmo estes cálculos têm limites. No R, por exemplo, o maior fatorial que pode ser calculado por aproximação usando o pacote estatístico básico “stats” é :

$$170! = 7.25 \times 10^{306}$$

Veremos mais adiante dois modelos, o de Poisson (Seção **3.3.6**) e o normal ou gaussiano (Seção **3.4.4**), que podem ser usados em algumas circunstâncias como aproximação da binomial para amostras grandes.