1 Distribuição Multinomial

Material didático para a disciplina Probabilidade I, ministrada em 2019.2, elaborado pelos Professores Maurício Mota e Gualberto Agamez.

A primeira distribuição discreta multivariada é a distribuição polinomial ou multinomial. Ela é aplicada para resolver problemas como o enunciado a seguir.

Exemplo 1: Nas experiências que Mendel realizou com ervilhas, ele observou 315 redondas e amarelas, 108 redondas e verdes, 101 enrugadas e amarelas e 32 enrugadas e verdes. De acordo com sua teoria da hereditariedade, os números deveriam estar na proporção $9 : 3 : 3 : 1$. Há alguma evidência para se duvidar de sua teoria, nos níveis de significância $0,01$? e $0,05$?

Esta distribuição generaliza a distribuição binomial que divide uma população em duas classes disjuntas e exaustivas. Agora, a população é dividida em k classes disjuntas e exaustivas com $P(C_i) = p_i$ para $i = 1, 2, \ldots, k$ e $p_1 + p_2 + \ldots + p_k = 1$. Uma amostra aleatória com reposição de tamanho n é retirada e seja

X_i= número de elementos da categoria i na amostra, para $i = 1, 2, \ldots, k$.

A distribuição conjunta do vetor $(X_1, X_2, \ldots, X_k)$ é multinomial com parâmetros $n, p_1, p_2, \ldots, p_k$. Note que sempre vale

$$X_1 + X_2 + \ldots + X_k = n,$$

evidenciando a dependência entre elas.

Voltando ao exemplo dado temos que a nossa população de cruzamentos é dividida em 4 classes a saber:

Classe 1: As ervilhas redondas e amarelas.
Classe 2: As ervilhas redondas e verdes.
Classe 3: As ervilhas enrugadas e amarelas.
Classe 4: As ervilhas enrugadas e verdes.

Do enunciado temos que $x_1 = 315$, $x_2 = 108$, $x_3 = 101$ e $x_4 = 32$ e

$$n = x_1 + x_2 + x_3 + x_4 = 315 + 108 + 101 + 32 = 556.$$

Estas são as frequências observadas. Geralmente colocamos em forma de vetor, portanto

$$O = (O_1, O_2, O_3, O_4) = (315, 108, 101, 32).$$

Para achar $p_i,\ i = 1, 2, 3, 4$ com $p_1 + p_2 + p_3 + p_4 = 1$ temos a informação

$$9 : 3 : 3 : 1.$$

Isto quer dizer que de 16 $(9 + 3 + 3 + 1 = 16)$ cruzamentos 9 pertencem à classe 1, 3 à classe 2, 3 à classe 3 e 1 à classe 4. Logo

$$p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad \text{e} \quad p_4 = \frac{1}{16}.$$

Qual a distribuição de $X_i, \ i = 1, 2, 3, 4$? De acordo com o enunciado temos que:

$$X_i \sim \text{Binomial}(n, p_i),$$

e

$$E_i = \mathbb{E}(X_i) = n \times pi = 556 \times p_i,$$

o que acarreta

$$E_1 = 556 \times \frac{9}{16} = 312,75,$$

assim, em média, são esperadas 312,75 ervilhas redondas e amarelas.

$$E_2 = 556 \times \frac{3}{16} = 104,25,$$

assim, em média, são esperadas 104,25 ervilhas enrugadas e amarelas.

$$E_3 = 556 \times \frac{3}{16} = 104,25,$$

assim, em média, são esperadas 104,25 ervilhas redondas e verdes. Finalmente,

$$E_4 = 556 \times \frac{1}{16} = 34,75,$$

assim, em média, são esperadas 34,75 ervilhas enrugadas e verdes.

Estas são as frequências esperadas se nosso modelo é verdadeiro. Elas colocadas em forma vetorial temos:

$$E = (E_1, E_2, E_3, E_4) = (312,75; 104,25; 104,25; 34,75).$$

Queremos testar a hipótese:

H_0 : O modelo proposto se ajusta aos dados.
H_1 : O modelo proposto não se ajusta aos dados.

Vamos definir a estatística do teste:

$$Q = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i}.$$

Se H_0 é verdade temos que:

$$Q \sim \chi^2(gl = c - 1 - m),$$

em que c é o número de classes e m é o número de parâmetros linearmente independentes que você utilizou para calcular as frequências esperadas.

No nosso exemplo temos que:

$$c = 4, \quad m = 0 \quad \text{e} \quad gl = 3.$$

Seja $Q_{tab} = q_{(1-\alpha;\, gl)}$ o quantil de ordem $1 - \alpha$ de Q. Logo

$$q_{(0{,}99;\, 3)} = 11,3 \quad \text{e} \quad q_{(0{,}95;\, 3)} = 7,81.$$

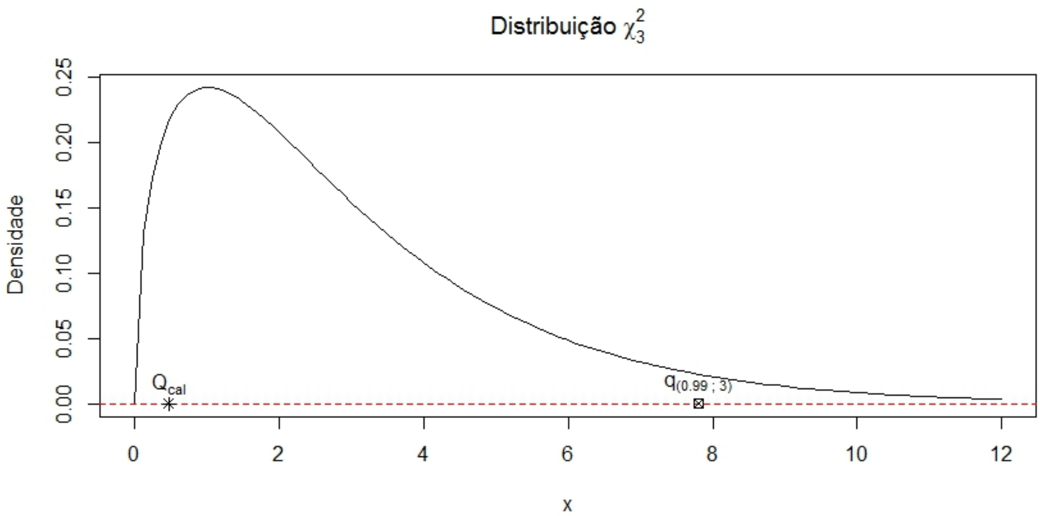
A regra de decisão será:

$$\text{Rejeitar } H_0 \text{ se } Q_{cal} \geq Q_{tab} = q_{(1-\alpha;\, gl)}.$$

Vamos calcular

$$\begin{aligned} Q_{cal} &= \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} \\ &= \frac{(315 - 312,75)^2}{312,5} + \frac{(101 - 104,25)^2}{104,25} + \frac{(108 - 104,25)^2}{104,25} + \frac{(32 - 34,75)^2}{34,75} \\ &= 0,470. \end{aligned}$$

Para $\alpha = 0,01$ temos que

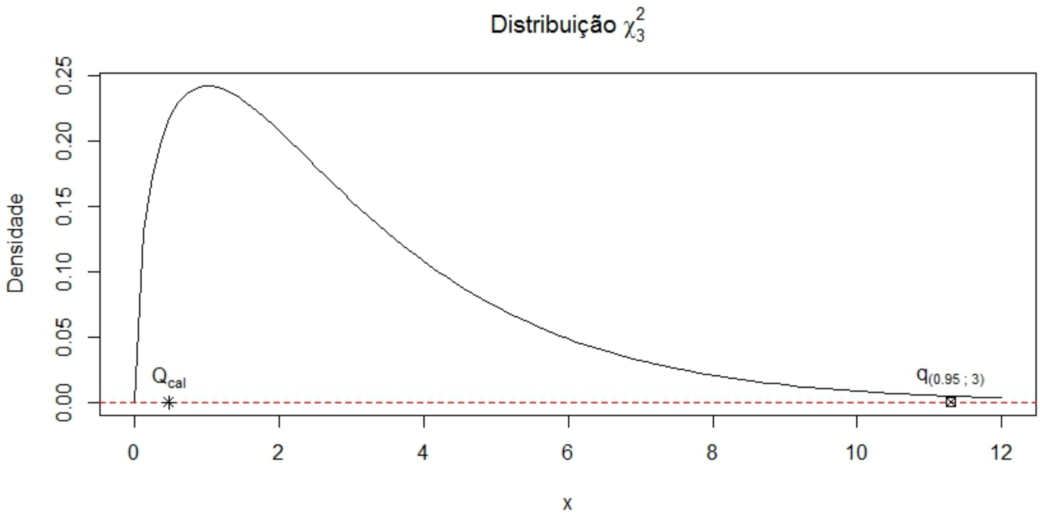


Portanto

$$Q_{cal} = 0,47 < Q_{tab} = 7,81.$$

Assim, concluímos que a teoria e a prática estão em consonância.

Para $\alpha = 0,05$ temos que



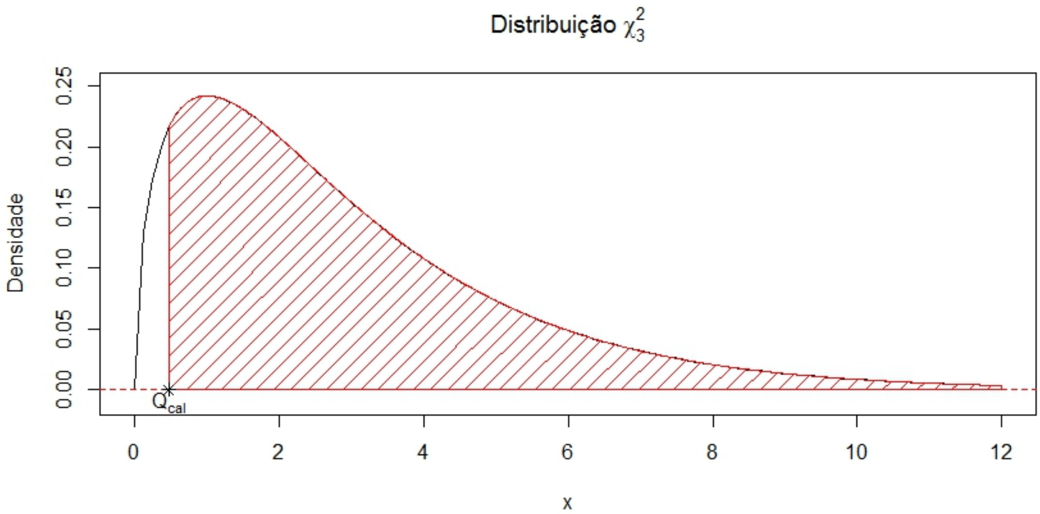
Portanto

$$Q_{cal} = 0,47 < Q_{tab} = 11,3.$$

Assim, concluímos que a teoria e a prática estão em consonância.

Hoje em dia calculamos o menor nível de significância que é chamado de nível descritivo ou p-valor do teste. Neste caso é dado por

$$nd = \hat{\alpha} = P(Q \geq Q_{cal}) = 0,9254.$$



A regra de decisão agora fica:
Se o nível de significância α for maior que o nível descritivo $\hat{\alpha}$ rejeitar a hipótese nula.

Vamos fazer pelo pacote R.

```
# Frequências observadas

O <- c(315,108,101,32)
c <- length(O); c
# [1] 4

# Tamanho da amostra

n <- sum(O); n
# [1] 556

# Probabilidades dadas pelo modelo teórico.

p <- c(9,3,3,1)/16; p
[1] 0.5625 0.1875 0.1875 0.0625

require(MASS)
fractions(p)
# [1] 9/16 3/16 3/16 1/16

# Cálculo das Frequências Esperadas.

m <- 0
E <- n*p; E
# [1] 312.75 104.25 104.25  34.75

Aux <- (O-E)^2/E; Aux
# [1] 0.01618705 0.13489209 0.10131894 0.21762590

Q_cal <- sum(Aux); Q_cal; round(Q_cal,2)
# [1] 0.470024
# [1] 0.47

alfa <- 0.01
gl=c-m-1; gl
# [1] 3

Q_tab <- qchisq(1-alfa,3); Q_tab; round(Q_tab,2)
# [1] 11.34487
# [1] 11.34

Q_cal >  Q_tab
# [1] FALSE
# Não rejeitar H_0.
```

```
alfa <- 0.05
Q_tab <- qchisq(1-alfa,3); Q_tab; round(Q_tab,2)
# [1] 7.814728
# [1] 7.81

Q_cal >  Q_tab
# [1] FALSE
# Não rejeitar H_0.

# Cálculo do nível descritivo.
nd <- 1-pchisq(Q_cal,gl); nd; round(nd,4)
# [1] 0.9254259
# [1] 0.9254


# Fazer o teste diretamente.

#?chisq.test

a <- chisq.test(O, p=p)
names(a)
#  [1] "statistic" "parameter" "p.value"   "method"    "data.name"
#  [6] "observed"  "expected"  "residuals" "stdres"
a

#  Chi-squared test for given probabilities

#  data:  O
#  X-squared = 0.47002, df = 3, p-value = 0.9254

a$expected
# [1] 312.75 104.25 104.25  34.75
```

Na literatura existe um outro teste que pode ser aplicado nessa situação. É o teste da razão de verossimilhança generalizada cuja estatística do teste é dada por:

$$D = 2\sum_{i=1}^{c})_i \ \log(O_i/E_i).$$

Se H_0 é verdade temos que:

$$D \sim \chi^2(gl = c - 1 - m),$$

e m que c é o número de classes e m é o número de parâmetros linearmente independentes que você utilizou para calcular as frequências esperadas. Vamos fazer pelo R:

```
> O= c(315,108,101,32)
> E=c( 312.75, 104.25, 104.25,  34.75)
```

```
>
> d_i= O*log(O/E);d_i
[1]  2.258074  3.816652 -3.198806 -2.638197
>
> D_cal=2*sum(d_i);D_cal
[1] 0.4754452
>
> c=4;m=0
> gl=c-1-m;gl
[1] 3
> # Cálculo do nível descritivo.
> nd <- 1-pchisq(D_cal,gl); nd; round(nd,4)
[1] 0.9242519
[1] 0.9243
```

Note que a conclusão é a mesma. Para um bom ajuste em ambos os testes é preciso que as frequências esperadas sejam pelo menos 5.

Vamos estudar o caso $k = 3$ da multinomial conhecido como distribuição trinomial.

1.1 Introdução.

Para facilitar a vida vamos fazer $X_1 = X$, $X_2 = Y$ e $X_3 = Z$. Considere uma população de tamanho N onde os elementos são divididos em 3 categorias C_1, C_2 e C_3 mutuamente exclusivos e exaustivos ($C_i \cap C_j = \emptyset$, $C_1 \cup C_2 \cup C_3 = \Omega$). Além disso a probabilidade de um elemento pertença à categoria i, $i = 1, 2, 3$ é

$$P(C_i) = p_i, \quad p_1 + p_2 + p_3 = 1.$$

Uma amostra aleatória de tamanho n com reposição é retirada . Sejam

X= número de elementos da categoria 1 na amostra.

Y= número de elementos da categoria 2 na amostra.

$Z = n - X - Y$= número de elementos da categoria 3 na amostra.

A distribuição conjunta de (X, Y, Z) é dada por:

$$P(X = x, Y = y, Z = z) = \frac{n!}{x!y!z!}\, p_1^x\, p_2^y\, p_3^z, \qquad x + y + z = n.$$

Para entender a fórmula dada inicialmente, calcule a probabilidade de aparecer na ordem x elementos da classe C_1, y elementos da classe C_2 e z elementos da classe C_3. Isto é,

$$P(C_1C_1 \ldots C_1\; C_2C_2 \ldots C_2\; C_3C_3 \ldots C_3) = p_1^x\, p_2^y\, p_3^z,$$

devido a independência das n repetições.

Qualquer permutação com repetição de $C_1C_1 \ldots C_1\ C_2C_2 \ldots C_2\ C_3C_3 \ldots C_3$ fornecerá x elementos da classe C_1, y elementos da classe C_2 e z elementos da classe C_3. Assim haverá

$$\frac{n!}{x!\ y!\ z!}$$

permutações distintas. Pelo Princípio Fundamental de contagem temos:

$$P(X=x, Y=y, Z=z) = \frac{n!}{x!y!z!}\ p_1^x\ p_2^y\ p_3^z, \qquad x+y+z=n.$$

Na próxima seção vamos apresentar a função de probabilidade conjunta de (X,Y).

1.2 Origem

Quando estudamos a distribuição binomial contamos o número de sucessos X e o número de fracassos $Y = n - X$ em n repetições independentes de um experimento de uma população dividida em uma partição com duas classes onde $n(C_1) = a$, $n(C_2) = b$ e $N = a + b$ é o tamanho da população. Sejam

$$p_1 = \frac{a}{a+b} = \frac{a}{N} \quad \text{e} \quad p_2 = \frac{b}{a+b} = \frac{b}{N}.$$

Assim $p_1 + p_2 = 1$ e $p_i,\ i = 1,2$ é a probabilidade de um elemento da população pertencer à classe $i,\ i = 1,2$.

Na realidade estamos interessados na variável aleatória X=número de elementos da classe 1 (sucesso) na amostra. Sabemos que

$$P(X=x) = \binom{n}{x}\ p_1^x\ p_2^{n-x}\ I_A(x),\ A = \{0,1,\ldots,n\}.$$

Para a demonstração que é um legítima função de probabilidade usará o famoso Binômio de Newton que será recordado agora.

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i\ b^{n-i}.$$

Note que dividindo por $(a+b)^n$ temos

$$1 = \sum_{i=0}^{n} \binom{n}{i} \frac{a^i\ b^{n-i}}{(a+b)^n}.$$

Mas

$$(a+b)^n = (a+b)^i\ \ (a+b)^{n-i}.$$

Logo,

$$1 = \sum_{i=0}^{n} \binom{n}{i} \frac{a^i\ b^{n-i}}{(a+b)^i\ \ (a+b)^{n-i}},$$

e que pode ser rearranjado como

$$
\begin{aligned}
1 &= \sum_{i=0}^{n} \binom{n}{i} \frac{a^i}{(a+b)^i} \; \frac{b^{n-i}}{(a+b)^{n-i}} \\
&= \sum_{i=0}^{n} \binom{n}{i} \left[\frac{a}{a+b}\right]^i \; \left[\frac{b}{a+b}\right]^{n-i} \\
&= \sum_{i=0}^{n} \binom{n}{i} p_1^i \, p_2^{n-i} \quad \text{em que} \quad p_2 = 1 - p_1.
\end{aligned}
$$

O binômio de Newton pode ser estendido para o trinômio

$$(a + b + c)^n.$$

Para $n = 2$ temos:

$$(a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc.$$

Qualquer elemento genérico T_{ij} é da forma

$$\frac{2!}{i!\, j!\, (2 - i - j)!}\, a^i\, b^j\, c^{2-i-j}, \; 0 \leq i + j \leq 2,$$

Vamos estudar com detalhes:

$$
\begin{aligned}
T_{00} &= \frac{2!}{0!\, 0!\, (2-0-0)!}\, a^0\, b^0\, c^{2-0-0} = c^2 \\
T_{01} &= \frac{2!}{0!\, 1!\, (2-0-1)!}\, a^0\, b^1\, c^{2-0-1} = 2bc \\
T_{02} &= \frac{2!}{0!\, 2!\, (2-0-2)!}\, a^0\, b^2\, c^{2-0-2} = b^2 \\
T_{10} &= \frac{2!}{1!\, 0!\, (2-1-0)!}\, a^1\, b^0\, c^{2-1-0} = 2ac \\
T_{11} &= \frac{2!}{1!\, 1!\, (2-1-1)!}\, a^1\, b^1\, c^{2-1-1} = 2ab \\
T_{20} &= \frac{2!}{2!\, 0!\, (2-2-0)!}\, a^2\, b^0\, c^{2-2-0} = a^2
\end{aligned}
$$

Para $n = 3$ temos

$$(a + b + c)^3 = a^3 + b^3 + c^3 + 3a^2b + 3a^2c + 3ab^2 + 3ac^2 + 3bc^2 + 6abc.$$

O termo genérico

$$\frac{n!}{i!\, j!\, k!}, \quad i + j + k = n,$$

é conhecido na literatura como coeficiente trinomial e será denotado por:

$$\binom{n}{i, j, k} = \frac{n!}{i!\ j!\ k!}, \quad i + j + k = n.$$

Ele aparece como o número de maneiras de dividir uma população com n elementos distintos em 3 subpopulações de tamanhos i, j e k, $i + j + k = n$. Isto pode ser feito de:

$$\binom{n}{i} \binom{n-i}{j} \binom{n-i-j}{k} = \frac{n!}{i!\ j!\ k!}.$$

Assim,

$$(a + b + c)^n = \sum_{\substack{0 \le i,\ j,\ k \le n \\ i+j+k=n}} \frac{n!}{i!\ j!\ k!}, \quad a^i\ b^j\ c^k.$$

Sejam

$$p_1 = \frac{a}{a + b + c}, \quad p_2 = \frac{b}{a + b + c} \quad \text{e} \quad p_3 = \frac{c}{a + b + c},$$

as proporções de elementos na população com a característica $i = 1, 2, 3$ na população. É claro que

$$p_1 + p_2 + p_3 = 1.$$

Assim vamos usar o teorema multinomial

$$1 = (p_1 + p_2 + p_3)^n = \sum_{\substack{0 \le i,\ j,\ k \le n \\ i+j+k=n}} \frac{n!}{i!\ j!\ k!}, \quad p_1^i\ p_2^j\ p_3^k.$$

Suponha agora que a nossa população é particionada em k classes (partição) e seja $p_i, \quad i = 1, 2, \ldots, k$ a probabilidade de um indivíduo da população pertencer a i-ésima classe. Como $p_1 + p_2 + \ldots + p_k = 1$ vamos usar a expansão multinomial

$$1 = (p_1 + p_2 + \ldots + p_k)^n = \sum_{\substack{0 \le x_1,\ x_2,\ \ldots, x_k \le n \\ x_1 + x_2 + \ldots + x_k = n}} \frac{n!}{x_1!\ x_2!\ \ldots,\ x_k!}, \quad p_1^{x_1}\ p_2^{x_2}\ \ldots,\ p_k^{x_k}.$$

Este coeficiente

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1! x_2! \ldots x_k!}$$

é conhecido como coeficiente multinomial.

Este modelo de probabilidade é conhecido na literatura como distribuição polinomial ou multinomial que será apresentado e discutido neste material.

1.3 Função Densidade de Probabilidade Conjunta

Dados $X = x$ e $Y = y$, o único valor possível para Z é $n - x - y$. Assim lembrando que $p_3 = 1 - p_1 - p_2$ temos que a função de probabilidade conjunta de (X, Y) é dada por:

$$f(x,y) = \frac{n!}{x!y!(n-x-y)!}\, p_1^x\, p_2^y\, (1-p_1-p_2)^{n-x-y}\; I_{\{0,1,\ldots,n\}}\,(x)\;\; I_{\{0,1,\ldots,n-x\}}\,(y).$$

Assim dizemos que (X, Y) tem uma distribuição bivariada de parâmetros n, p_1 e p_2.

Notação: $(X, Y) \sim Trinomial(n, p_1, p_2)$.

2 Estudo das Marginais

2.1 Marginal de X

Fato 1:

$$X \sim B(n, p_1)$$

.

Prova.

$$\begin{aligned}
f(x) = P(X = x) &= \frac{n!\, p_1^x}{x!\,(n-x)!} \sum_{y=0}^{n-x} \frac{(n-x)!}{y!(n-x-y)!}\, p_2^y\, (1-p_1-p_2)^{n-x-y} \\
&= \binom{n}{x}\; p_1^x \sum_{y=0}^{n-x} \binom{n-x}{y}\; p_2^y\, (1-p_1-p_2)^{n-x-y} \\
&= \binom{n}{x}\; p_1^x\, (1-p_1-p_2+p_2)^{n-x} \\
&= \binom{n}{x}\; p_1^x\, (1-p_1)^{n-x},
\end{aligned}$$

que é a função de probabilidade da Binomial de parâmetros n e p_1.

2.2 Marginal de Y

Fato 2:

$$Y \sim B(n, p_2)$$

.

Prova.

$$
\begin{aligned}
f(y) = P(Y=y) &= \frac{n!\, p_2^y}{y!\,(n-y)!} \sum_{x=0}^{n-y} \frac{(n-y)!}{x!(n-x-y)!}\, p_1^y\, (1-p_1-p_2)^{n-x-y} \\
&= \binom{n}{y}\, p_2^y \sum_{x=0}^{n-y} \binom{n-y}{x}\, p_1^x\, (1-p_1-p_2)^{n-x-y} \\
&= \binom{n}{y}\, p_2^y\, (1-p_2-p_1+p_1)^{n-y} \\
&= \binom{n}{y}\, p_2^y\, (1-p_2)^{n-y},
\end{aligned}
$$

que é a função de probabilidade da Binomial de parâmetros n e p_2.

3 Estudo das Condicionais.

3.1 Condicional de Y dado X=x.

Fato 3:

$$
Y|X=x \sim B\left(n-x, p=\frac{p_2}{1-p_1}\right)
$$

.

Prova.

Seja $f(y|x) = P(Y=y|X=x)$ a distribuição condicional de $Y|X=x$. Assim,

$$
\begin{aligned}
P(Y=y|X=x) &= \frac{P(X=x,Y=y)}{P(X=x)} \\
&= \frac{\dfrac{n!}{x!\,y!\,(n-x-y)!}\, p_1^y\;\, p_2^y\, (1-p_1-p_2)^{n-x-y}}{\dfrac{n!}{x!\,(n-x)!}\;\, p_1^x\, (1-p_1)^{n-x}} \\
&= \frac{(n-x)!}{y!\,(n-x-y)!}\; \frac{p_2^y\, (1-p_1-p_2)^{n-x-y}}{(1-p_1)^y\;(1-p_1)^{n-x-y}} \\
&= \binom{n-x}{y} \left(\frac{p_2}{1-p_1}\right)^y \left(1-\frac{p_2}{1-p_1}\right)^{n-x-y},
\end{aligned}
$$

que é a função de probabilidade da Binomial de parâmetros $n-x$ e $p = \dfrac{p_2}{1-p_1}$

3.2 Condicional de X dado Y=y.

Fato 4:

$$
X|Y=y \sim B\left(n-y, p=\frac{p_1}{1-p_2}\right)
$$

.

Prova.

Seja $f(x|y) = P(X = x|Y = y)$ a distribuição condicional de $X|Y = y$. Assim,

$$
\begin{aligned}
P(Y = y|X = x) &= \frac{P(X = x, Y = y)}{P(Y = y)} \\
&= \frac{\dfrac{n!}{x!\, y!\, (n-x-y)!}\ p_1^y\ p_2^y\ (1-p_1-p_2)^{n-x-y}}{\dfrac{n!}{y!\, (n-y)!}\ p_2^y\ (1-p_2)^{n-y}} \\
&= \frac{(n-y)!}{x!\, (n-x-y)!}\ \frac{p_1^x\ (1-p_1-p_2)^{n-y-x}}{(1-p_2)^x\ (1-p_2)^{n-y-x}} \\
&= \binom{n-y}{x} \left(\frac{p_1}{1-p_2}\right)^x \left(1 - \frac{p_1}{1-p_2}\right)^{n-y-x},
\end{aligned}
$$

que é a função de probabilidade da Binomial de parâmetros $n - x$ e $p = \dfrac{p_1}{1-p_2}$

4 Momentos da Trinomial.

Sabemos que:

a. $\mathbb{E}(X) = np_1$, $E(X^2) = np_1\,[1 + (n-1)p_1]$ e $\mathrm{Var}(X) = np_1(1-p_1)$.

b. $\mathbb{E}(Y) = np_2$, $\mathbb{E}(Y^2) = np_2\,[1 + (n-1)p_2]$ e $\mathrm{Var}(X) = np_2(1-p_2)$.

Vamos mostrar usando propriedades de esperança e variância condicionais.

Fato 5: A esperança condicional de $Y|X$ é dada por:

$$
\mathbb{E}(Y|X) = (n - X)\,\frac{p_2}{1-p_1}. \tag{1}
$$

Fato 6: A variância condicional de $Y|X$ é dada por:

$$
\begin{aligned}
\mathrm{Var}(Y|X) &= (n-X)\,\frac{p_2}{1-p_1}\,\frac{1-p_1-p_2}{1-p_1} \\
&= (n-X)\,\frac{p_2\,(1-p_1-p_2)}{(1-p_1)^2}.
\end{aligned} \tag{2}
$$

Obtenha a esperança de Y usando a condicional:

$$
\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}\left[\mathbb{E}(Y|X)\right] \\
&= \mathbb{E}\left[(n-X)\ \frac{p_2}{1-p_1}\right] \\
&= (n-\mathbb{E}(X))\ \frac{p_2}{1-p_1} \\
&= (n-np_1)\ \frac{p_2}{1-p_1} \\
&= n(1-p_1)\ \frac{p_2}{1-p_1} \\
&= np_2.
\end{aligned}
$$

Obtenha a variância de Y usando a condicional:

$$
\begin{aligned}
\mathrm{Var}(Y) &= E\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}\left[\mathbb{E}(Y|X)\right] \\
&= \mathbb{E}\left[(n-X)\ \frac{p_2\,(1-p_1-p_2)}{(1-p_1)^2}\right] + \mathrm{Var}\left[(n-X)\ \frac{p_2}{1-p_1}\right] \\
&= (n-\mathbb{E}(X))\ \frac{p_2}{1-p_1} + \left[\frac{p_2}{1-p_1}\right]^2\ \mathrm{Var}(n-X) \\
&= (n-np_1)\ \frac{p_2\,(1-p_1-p_2)}{(1-p_1)^2} + \left[\frac{p_2}{1-p_1}\right]^2\ \mathrm{Var}(X) \\
&= n(1-p_1)\ \frac{p_2\,(1-p_1-p_2)}{(1-p_1)^2} + \left[\frac{p_2}{1-p_1}\right]^2\ np_1(1-p_1) \\
&= np_2\left[\frac{1-p_1-p_2}{1-p_1} + \frac{p_2p_1}{1-p_1}\right] \\
. \quad &= np_2\left[\frac{1-p_1-p_2+p_1p_2}{1-p_1}\right] \\
&= np_2\left[\frac{(1-p_1)(1-p_2)}{1-p_1}\right] \\
&= np_2(1-p_2).
\end{aligned}
$$

Fato 7: Mostre que $\mathbb{E}(XY) = n(n-1)p_1\,p_2$ usando condicional.

Prova.

$$\begin{aligned}
\mathbb{E}(XY) &= \mathbb{E}\left[\mathbb{E}(XY|X)\right] \\
&= \mathbb{E}\left[X\mathbb{E}(Y|X)\right] \\
&= \mathbb{E}\left[X(n-X)\ \frac{p_2}{1-p_1}\right] \\
&= \mathbb{E}\left[X(n-X)\ \right]\ \frac{p_2}{1-p_1} \\
&= \left[n\mathbb{E}(X)-\mathbb{E}(X^2)\ \right]\ \frac{p_2}{1-p_1} \\
&= \left[n^2p_1-np_1(1-p_1)-n^2p_1^2\ \right]\ \frac{p_2}{1-p_1} \\
&= \left[n^2p_1(1-p_1)-np_1(1-p_1)\right]\ \frac{p_2}{1-p_1} \\
&= \left[n^2p_1-np_1\right]p_2 \\
&= n(n-1)p_1p_2.
\end{aligned}$$

Fato 8: Mostre que $\text{Cov}(X,Y)=-np_1\,p_2$.

Prova.

$$\begin{aligned}
\text{Cov}(X,Y) &= \mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y) \\
&= n(n-1)p_1p_2-n^2p_1p_2 \\
&= -np_1\,p_2(-n+1+n) \\
&= -np_1\,p_2.
\end{aligned}$$

Poderíamos provar usando condicional.

$$\begin{aligned}
\text{Cov}(X,Y) &= \text{Cov}(X,E[Y|X]) \\
&= \text{Cov}\left(X,(n-X)\ \frac{p_2}{1-p_1}\right) \\
&= \frac{p_2}{1-p_1}\ \text{Cov}[(X,(n-X)] \\
&= \frac{p_2}{1-p_1}\ [\text{Cov}(X,n)-\text{Cov}(X,X)] \\
&= \frac{p_2}{1-p_1}\ [0-\text{Var}(X)] \\
&= \frac{p_2}{1-p_1}\ [-np_1(1-p_1)] \\
&= -np_1p_2.
\end{aligned}$$

Fato 9: O coeficiente de correlação, ρ, entre X e Y é dado por:

$$\rho=-\sqrt{\frac{p_1\,p_2}{(1-p_1)(1-p_2)}}.$$

Prova.

$$\begin{aligned}
\rho &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \\
&= \frac{-np_1p_2}{\sqrt{np_1(1-p_1)np_2(1-p_2)}} \\
&= -\frac{p_1p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} \\
&= -\sqrt{\frac{p_1p_2}{(1-p_1)(1-p_2)}}.
\end{aligned}$$

Note que o coeficiente de correlação independente do tamanho n da amostra e do tamanho N da população.

5 Função Geradora de Probabilidades da Trinomial

A função geradora de probabilidades momentos bivariada de $(X,Y) \sim$ Trinomial (n,p_1,p_2), é dada por:

$$G_{X,Y}(t_1,t_2) = (p_1t_1 + p_2\, t_2 + 1 - p_1 - p_2)^n$$

em que t_1 e t_2 são reais.

Prova.

$$\begin{aligned}
G_{(X,Y)}\,(t_1,t_2) &= \mathbb{E}\left[t_1^X\, t_2^Y\right] \\
&= \sum_{x=0}^{n}\sum_{y=0}^{n-x} t_1^x\, t_2^y \frac{n!}{x!\; y!\; (n-x-y)!}\, p_1^x\;\; p_2^y\; (1-p_1-p_2)^{n-x-y} \\
&= \sum_{x=0}^{n}\sum_{y=0}^{n-x} \frac{n!}{x!\; y!\; (n-x-y)!}\, (p_1\; t_1)^x\;\; (p_2\; t_2)^y\;\, ; (1-p_1-p_2)^{n-x-y} \\
&= (p_1\; t_1 + p_2\; t_2 + 1 - p_1 - p_2)^n\,,
\end{aligned}$$

usando o teorema multinomial.

6 Função Geradora de Momentos da Trinomial

A função geradora de momentos bivariada de $(X,Y) \sim$ Trinomial (n,p_1,p_2), é dada por:

$$M_{X,Y}(t_1,t_2) = \left(p_1e^{t_1} + p_2\; e^{t_2} + 1 - p_1 - p_2\right)^n$$

em que t_1 e t_2 são reais.

Prova.

$$\begin{aligned}
M_{X,Y}(t_1,t_2) &= \mathbb{E}\left[e^{t_1X+t_2Y}\right] \\
&= \mathbb{E}\left[e^{t_1X}e^{t_2Y}\right] \\
&= \sum_{x=0}^{n}\sum_{y=0}^{n-x} e^{t_1}\, e^{t_2}\frac{n!}{x!\; y!\; (n-x-y)!}\, p_1^x\;\; p_2^y\; (1-p_1-p_2)^{n-x-y} \\
&= \sum_{x=0}^{n}\sum_{y=0}^{n-x} \frac{n!}{x!\; y!\; (n-x-y)!}\, (p_1\, e^{t_1})^x\;\; (p_2\, e^{t_2})^y\; (1-p_1-p_2)^{n-x-y} \\
&= \left(p_1e^{t_1}+p_2\, e^{t_2}+1-p_1-p_2\right)^n,
\end{aligned}$$

usando o teorema multinomial.

Vamos obter a função geradora de momentos de X.

$$\begin{aligned}
M_X(t) &= M_{(X,Y)}\,(t,0) \\
&= \left(p_1e^t+p_2\, e^0+1-p_1-p_2\right)^n \\
&= \left(p_1e^t+p_2+1-p_1-p_2\right)^n \\
&= \left(p_1e^t+1-p_1\right)^n,
\end{aligned}$$

que é a função geradora de momentos da binomial de parâmetros n e p_1.

Usando a função geradora de momentos vamos calcular novamente $\mathbb{E}(XY)$. Sabemos que

$$\mathbb{E}(XY)=\frac{\partial^2 M(t_1,t_2)}{\partial t_1\, \partial t_2}\Big|_{(0,0)}.$$

$$\begin{aligned}
\frac{\partial M}{\partial t_1} &= n\left(p_1\, e^{t_1}+p_2\, e^{t_2}+1-p_1-p_2\right)^{n-1}\; p_1\, e^{t_1} \\
&= np_1e^{t_1}\left(p_1e^{t}+p_2\, e^{t_2}+1-p_1-p_2\right)^{n-1},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 M}{\partial t_1\, \partial t_2} &= np_1\, e^{t_1}\left[p_1e^{t_1}+p_2e^{t_2}+(1-p_1-p_2)^{n-2}\right](n-1)p_2\, e^{t_2} \\
&= n(n-1)p_1p_2\;\; \left[p_1e^{t_1}+p_2e^{t_2}+(1-p_1-p_2)^{n-2}\right]
\end{aligned}$$

e

$$\begin{aligned}
\frac{\partial^2 M(t_1,t_2)}{\partial t_1\, \partial t_2}\Big|_{(0,0)} &= n(n-1)p_1p_2\;\; \left[p_1e^0+p_2e^0+(1-p_1-p_2)^{n-2}\right] \\
&= n(n-1)p_1p_2\;\; \left[p_1+p_2+(1-p_1-p_2)^{n-2}\right] \\
&= n(n-1)p_1p_2=\mathbb{E}(XY).
\end{aligned}$$

6.1 Transformações Importantes

Vamos estudar algumas transformações envolvendo a trinomial (X, Y).

a. $Z = n - X - Y \sim B(n, p_3 = 1 - p_1 - p_2)$.

Prova: A função geradora de momentos de Z é dada por:

$$\begin{aligned} M_Z(t) &= \mathbb{E}(e^{tZ}) \\ &= \mathbb{E}(e^{t(n-X-Y)}) \\ &= e^{tn}\ E(e^{-tX-tY}) \\ &= e^{tn}\ M(-t,-t) \\ &= e^{tn}\ \left(p_1 e^{-t} + p_2\ e^{-t} + 1 - p_1 - p_2\right)^n \\ &= \left(p_1 + p_2 + (1 - p_1 - p_2)e^t\right)^n \\ &= \left(p_3\ e^t + 1 - p_3\right), \end{aligned}$$

que é a função geradora de momentos da Binomial de parâmetros n e $p_3 = 1 - p_1 - p_2$.

b. $S = X + Y \sim Bin(n, p = p_1 + p_2)$.

Prova: A função geradora de momentos de S é dada por:

$$\begin{aligned} M_S(t) &= \mathbb{E}(e^{tS}) \\ &= \mathbb{E}(e^{t(X+Y)}) \\ &= \mathbb{E}(e^{tX+tY}) \\ &= M(t,t) \\ &= \left(p_1 e^t + p_2\ e^t + 1 - p_1 - p_2\right)^n \\ &= \left[(p_1 + p_2)e^t + (1 - p_1 - p_2)\right]^n, \end{aligned}$$

que é a função geradora de momentos da Binomial de parâmetros n e $p = p_1 + p_2$.

c. Obtenha a distribuição conjunta de $S = X + Y$ e $V = X$. Em seguida calcule a marginal de S.

Solução: É claro que $s \geq v$. Assim,

$$
\begin{aligned}
f(s,v) &= P(S=s,V=v) \\
&= P(X+Y=s,X=v) \\
&= P(X=v,Y=s-v) \\
&= \frac{n!}{v!(s-v)!)(n-s)!}\, p_1^v\, p_2^{s-v}\, (1-p_1-p_2)^{n-s}\, I_{\{0,1,\ldots,n\}}\,(v)\ \ I_{\{0,1,\ldots,n-v\}}\,(s-v) \\
&= \frac{n!}{s!(n-s)!}\ \frac{s!}{v!(s-v)!}(1-p_1-p_2)^{n-s}\quad p_1^v\, p_2^{s-v}\, I_{\{0,1,\ldots,n\}}\,(v)\ \ I_{\{0,1,\ldots,s\}}\,(v) \\
&= \binom{n}{s}\binom{s}{v}\, p_1^v\, p_2^{s-v}\, I_{\{0,1,\ldots,n\}}\,(s)\ \ I_{\{0,1,\ldots,s\}}\,(v)
\end{aligned}
$$

A marginal de S é dada por:

$$
\begin{aligned}
P(S=s) &= \binom{n}{s}\,(1-p_1-p_2)^{n-s}\,\sum_{v=0}^{s}\binom{s}{v}\, p_1^v\, p_2^{s-v} \\
&= \binom{n}{s}\,(p_1+p_2)^s\,(1-p_1-p_2)^{n-s}\quad I_{\{0,1,\ldots,n\}}\,(s)
\end{aligned}
$$

d. Obtenha a distribuição conjunta de (X,Z).

Solução: A função de probabilidade conjunta de (X,Z) é dada por

$$
\begin{aligned}
P(X=x,Z=z) &= P(X=x,n-X-Y=z) \\
&= P(X=x,Y=n-x-z) \\
&= \frac{n!}{x!z!(n-x-z)!}\, p_1^x\, p_2^{n-x-z}\,(1-p_1-p_2)^z\, I_{\{0,1,\ldots,n\}}\,(x)\ \ I_{\{0,1,\ldots,n-x\}}\,(z) \\
&= \frac{n!}{x!z!(n-x-z)!}\, p_1^x\, p_2^{n-x-z} p_3^z\quad I_{\{0,1,\ldots,n-x\}}\,(z) \\
&= \frac{n!}{x!z!(n-x-z)!}\, p_1^x\, p_3^z\,(1-p_1-p_3)^{n-x-z}\quad I_{\{0,1,\ldots,n-x\}}\,(z)
\end{aligned}
$$

com $p_3 = 1-p_1-p_2$. Além disso (X,Z) tem uma distribuição trinomial de parâmetros n, p_1 e p_3.

e. Obtenha a distribuição conjunta de (Y,Z).

Solução: A função geradora de momentos de (Y,Z) é dada por

$$
\begin{aligned}
M_{(Y,Z)}(t_1,t_2) &= \mathbb{E}(e^{t_1 Y + t_2 Z}) \\
&= \mathbb{E}(e^{t_1 Y + t_2 (n-X-Y)}) \\
&= e^{n\,t_2}\mathbb{E}(e^{-t_2 X + (t_1 - t_2) Y}) \\
&= e^{n\,t_2}\ M_{(X,Y)}(-t_2, t_1 - t_2) \\
&= e^{n\,t_2}\ \left(p_1\ e^{-t_2} + p_2\ e^{t_1 - t_2} + p_3\right)^n \\
&= \left(p_1 + p_2\ e_1^t + p_3\ e^{t_2}\right)^n \\
&= \left(1 - p_2 - p_3 + p_2\ e^{t_1} + p_3\ e^{t_2}\right)^n,
\end{aligned}
$$

que é a função geradora de momentos da Trinomial de parâmetros n , p_2 e p_3.

6.2 Comentários

Comentário 1. Se (X, Y) tiver distribuição trinomial de parâmetros n, p_1 e p_2 a variável aleatória $S = X + Y$ é a soma de duas variáveis aleatórias binomiais dependentes em que:

$$X \sim B(n, p_1), \quad Y \sim B(n, p_2), \quad S = X + Y \sim B(n, p_1 + p_2),\ 0 < p_1 + p_2 < 1.$$

Comentário 2. Se $X \sim B(n_1, p)$, $Y \sim B(n_2, p)$, independentes então

$$S = X + Y \sim B(n_1 + n_2, p).$$

Comentário 3. No caso da trinomial $S = X + Y$ representa o número de elementos das categorias 1 e 2 na amostra. A probabilidade de sucesso é então

$$p_{12} = P(C_1 \cup C_2) = p_1 + p_2.$$

Agora a nossa população está dividida em duas classes: C_{12} e C_3 e recaímos no caso da Binomial pois as repetições são independentes logo

$$S = X + Y \sim B(n, p_1 + p_2).$$

Comentário 4. Podemos obter a $\text{Cov}(X, Y)$ usando o fato que

$$S = X + Y \sim B(n, p_1 + p_2).$$

Prova. Como

$$\text{Var}(S) = \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y),$$

temos:

$$
\begin{aligned}
\mathrm{Cov}(X,Y) &= \frac{1}{2}\,[\mathrm{Var}(S)-\mathrm{Var}(X)-\mathrm{Var}(Y)]\\
&= \frac{1}{2}\,[n(p_1+p_2)(1-p_1-p_2)-np_1(1-p_1)-np_2(1-p_2)]\\
&= \frac{n}{2}\,\left[p_1(1-p_1)-p_1p_2+p_2(1-p_1)-p_2^2-p_1(1-p_1)-p_2(1-p_2)\right]\\
&= \frac{n}{2}\,\left[-p_1p_2+p_2-p_2p_1-p_2^2-p_2+p_2^2\right]\\
&= \frac{n}{2}\,[-2p_1p_2]\\
&= -np_1p_2.
\end{aligned}
$$

Vamos aplicar a teoria apresentada através de um exemplo.

6.3 Exemplo

Uma amostra contém 20% de bolas amarelas, 30% de bolas brancas e as restantes são vermelhas. Uma amostra aleatória de 5 bolas com reposição é retirada. Sejam

X= número de bolas amarelas na amostra.

Y= número de bolas brancas na amostra.

Z= número de bolas vermelhas na amostra.

A classe C_1 é formada pelas bolas amarelas, a C_2 pelas brancas e C_3 pelas vermelhas. Assim,

$$P(C_1)=p_1=0,2 \quad P(C_2)=p_2=0,3 \quad \text{e} \quad P(C_3)=p_3=0,5.$$

Responda o que se pede:

a. Qual a distribuição conjunta de (X,Y), de (X,Z) e de (Y,Z)?

Solução: (X,Y) tem distribuição trinomial de parâmetros $n=5$,e $p_1=0,2$ e $p_2=0,3$. A f.p.c. de (X,Y) é dada por:

$$P(X=x,Y=y)=\frac{5!}{x!y!(5-x-y)!}\,(0,2)^x\,(0,3)^y\,(0,5)^{5-x-z}\;\;I_{\{0,1,\ldots,5\}}\,(x)\;\;I_{\{0,1,\ldots,5-x\}}\,(y).$$

(X,Z) tem distribuição trinomial de parâmetros $n=5$,e $p_1=0,2$ e $p_3=0,5$.

A f.p.c. de (X,Z) é dada por:

$$P(X=x,Z=z)=\frac{5!}{x!z!(5-x-z)!}\,(0,2)^x\,(0,5)^z\,(0,3)^{5-x-z}\;\;I_{\{0,1,\ldots,5\}}\,(x)\;\;I_{\{0,1,\ldots,5-x\}}\,(z).$$

Finalmente (Y,Z) tem distribuição trinomial de parâmetros $n=5$, $p_2=0,3$ e $p_3=0,5$.

A f.p.c. de (Y, Z) é dada por:

$$P(Y = y, Z = z) = \frac{5!}{y!z!(5-y-z)!} (0,3)^x \, (0,5)^z \, (0,2)^{5-y-z} \;\; I_{\{0,1,\ldots,5\}}\,(y) \;\; I_{\{0,1,\ldots,5-y\}}\,(z).$$

b. Sabendo que saiu 2 bolas brancas, qual a distribuição do número de bolas amarelas?

Solução: A distribuição de $X|Y = 2$ é binomial de parâmetros $n - y = 5 - 2 = 3$ e $p = \dfrac{p_1}{1 - p_2} = \dfrac{2}{7}$.

c. Calcule os coeficientes de correlação das variáveis X,Y e Z.

Solução: O coeficiente de correlação entre X e Y é dado por

$$\rho_{XY} = -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}} = -\sqrt{\frac{0,2 \times 0,3}{0,8 \times 0,7}} = -\sqrt{\frac{3}{28}}.$$

O coeficiente de correlação entre X e Z é dado por:

$$\rho_{XZ} = -\sqrt{\frac{p_1 p_3}{(1-p_1)(1-p_3)}} = -\sqrt{\frac{0,2 \times 0,5}{0,8 \times 0,5}} = -\sqrt{\frac{1}{4}} = -\frac{1}{2}.$$

O coeficiente de correlação entre Y e Z é dado por:

$$\rho_{YZ} = -\sqrt{\frac{p_2 p_3}{(1-p_2)(1-p_3)}} = -\sqrt{\frac{0,3 \times 0,5}{0,7 \times 0,5}} = -\sqrt{\frac{3}{7}}.$$

d. Qual a distribuição de $S_{XY} = X + Y$, $S_{XZ} = X + Z$ e $S_{YZ} = Y + Z$?

Solução: Assim,

$$S_{XY} = X + Y \sim B(n, p_1 + p_2) = B(5; 0,5),$$

$$S_{XZ} = X + Z \sim B(n, p_1 + p_3) = B(5; 0,7),$$

$$S_{YZ} = Y + Z \sim B(n, p_2 + p_3) = B(5; 0,8).$$

e. Qual a f.g.m. de (X, Y)?

Solução:

$$M(t_1, t_2) = \left(0,2e^{t_1} + 0,3e^{t_2} + 0,5\right)^5 .$$

f. Qual a probabilidade de que haja mais bolas amarelas?

Solução: Seja E o evento que diz que há mais bolas amarelas na amostra. Este evento é equivalente ao evento $X \geq 3$, lembrando que $X \;\; B(5; 0,2)$ é o número de bolas amarelas. Assim

$$P(E) = P(X \geq 3) = \sum_{x=3}^{5} \binom{5}{x} (0,2)^x \, (0,8)^{5-x} = 0,051 + 0,006 + 0^+ = 0,057.$$

de acordo com a tabela I do Bussab& Morettin, pgs 494,495.

Vamos obter o resultado usando o R.

```
PE <- 1- pbinom(2,5,0.2); PE; round(PE,3)
# [1] 0.05792
# [1] 0.058

x <- 0:5; px <- dbinom(x,5,0.2)
tab <- cbind(x,px); round(tab,5)

#      x      px
# [1,] 0 0.32768
# [2,] 1 0.40960
# [3,] 2 0.20480
# [4,] 3 0.05120
# [5,] 4 0.00640
# [6,] 5 0.00032

Px <- cumsum(px)
tab <- cbind(x,px,Px);round(tab,5)

#      x      px      Px
# [1,] 0 0.32768 0.32768
# [2,] 1 0.40960 0.73728
# [3,] 2 0.20480 0.94208
# [4,] 3 0.05120 0.99328
# [5,] 4 0.00640 0.99968
# [6,] 5 0.00032 1.00000
```

```
px[4]+px[5]+px[6]
# [1] 0.05792
```

Vamos resolver utilizando a função *multinom* do R.

```
n <- 5
p <- c(2,3,5)/10; p
# [1] 0.2 0.3 0.5

x1 <- c(3,0,2)
px1 <- dmultinom(x1,n,p); px1
# [1] 0.02

x2 <- c(3,1,1)
px2<- dmultinom(x2,n,p); px2
# [1] 0.024

x3 <- c(3,2,0)
px3<- dmultinom(x3,n,p); px3
# [1] 0.0072

x4 <- c(4,0,1)
px4 <- dmultinom(x4,n,p); px4
# [1] 0.004

x5 <- c(4,1,0)
px5 <- dmultinom(x5,n,p); px5
# [1] 0.0024

x6 <- c(5,0,0)
px6 <- dmultinom(x6,n,p); px6
# [1] 0.00032

teste <- px1 + px2 + px3 + px4 + px5 + px6; teste
# [1] 0.05792
```

Vamos pensar numa justificativa melhor. Temos que resolver as desigualdades:

$$X > Y \quad \text{e} \quad X > 5 - X - Y.$$

ou

$$X > Y \quad \text{e} \quad 2X > 5 - Y.$$

ou

$$X > Y \quad \text{e} \quad X > \frac{5 - Y}{2}.$$

O que acarreta

$$X > \max\left(Y, \frac{5-Y}{2}\right).$$

g. Gere uma amostra aleatória de (X, Y, Z) trinomial de parâmetros $n = 5$ e $(p_1, p_2, p_3) = c(0,2; 0,3; 0,5)$.

```
# Vamos gerar uma amostra aleatória de tamanho 10 de
# (X,Y,Z)-Multinomial(5;0,2;0,3;0,5),

set.seed(32)
rmultinom(10, size = 5, prob = c(0.2,0.3,0.5))

# [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
# [1,]    1    2    0    2    1    1    1    0    2     1
# [2,]    2    2    4    2    1    1    2    3    1     3
# [3,]    2    1    1    1    3    3    2    2    2     1

# Cada coluna é um elemento da amostra:
# v_1=c(1,2,2), v-2=c(2,2,1),....v10=c(1,3,1)
```

h. Gere uma amostra de tamanho 100000 da nossa distribuição e calcule a frequência relativa do evento E. Comente o resultado!!!!

7 Generalização

A função de probabilidade de $(X_1, X_2, \ldots, X_k) \sim multinomial(n, p_1, p_2, \ldots, p_k)$ é dada por:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!x_2!\ldots x_k!}\; p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k},$$

Com $x_1 + x_2 + \ldots + x_k = n$ e $p_1 + p_2 + \ldots p_k = 1$.

Sua função geradora de probabilidades é dada por:

$$G_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = (p_1\; t_1 + p_2\; t_2 + \ldots + p_n)^n$$

Sua função geradora de momentos é dada por:

$$M_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = \left(p_1\; e^{t_1} + p_2\; e^{t_2} + \ldots + p_k\right)^n$$

$$M_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = (p_1\; t_1 + p_2\; t_2 + \ldots + p_k)^n$$

Propriedades da Multinomial.

Fato 7.1

$$X_i \sim \text{Binomial}(n, p_i),\ i = 1, 2, \ldots, k.$$

Prova. A função geradora de probabilidades de X_i é dada por:

$$\begin{aligned} G_{X_i}(t) &= \mathbb{E}\left[t^{X_i}\right] \\ &= G_{X_1,\ldots,X_n}(1, 1, \ldots, t, \ldots, 1) \\ &= (p_1 + p_2 + \ldots + p_i\ t + p_n)^n \\ &= (p_i\ t + p_1 + p_2 + \ldots + p_n)^n \\ &= (p_i\ t + 1 - p_i)^n\,, \end{aligned}$$

que é a função geradora de probabilidades da binomial de parâmetros n e p_i.

Fato 7.2

$$S_{ij} = X_i + X_j \sim \text{Binomial}(n, p_i + pj),\ i, j = 1, 2, \ldots, k, i \neq j.$$

Prova. A função geradora de probabilidades de S_{ij} é dada por:

$$\begin{aligned} G_{S_{ij}}(t) &= \mathbb{E}\left[t^{S_{ij}}\right] \\ &= \mathbb{E}\left[t^{X_i}\ t^{X_j}\right] \\ &= G_{X_1,\ldots,X_n}(1, 1, \ldots, t, \ldots, t, \ldots, 1) \\ &= (p_1 + p_2 + \ldots + p_i\ t + p_j t p_n)^n \\ &= ((p_i + p_j)\ t + p_1 + p_2 + \ldots + p_n)^n \\ &= ((p_i + p_j)\ t + 1 - p_i - p_j)^n\,, \end{aligned}$$

que é a função geradora de probabilidades da binomial de parâmetros n e $p_i + p_j$.

Fato 7.3 A covariância entre X_i e X_j é dada por:

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

Fato 7.4 A correlação entre X_i e X_j é dada por:

$$\text{Corr}(X_i, X_j) = -\sqrt{\frac{p_i\, p_j}{q_i\, q_j}},$$

$q_i = 1 - p_i$ e $q_j = 1 - p_j$.

Fato 7.5 A distribuição condicional de X_i dado que $X_j = x$ é binomial de parâmetros $m = n - x$ e $p = \dfrac{p_i}{1 - p_j}$.

$$X_i | X_j = x \sim \text{Binomial}\ \ (m = n - x\ ,\ p = \frac{p_i}{1 - p_j}).$$

8 Exercícios

1. A lei de Hardy-Weiberg diz que, sob certas condições, as frequências relativas dos genótipos AA,Aa e aa ocorrerem em uma população são, respectivamente, $p_1 = \theta^2$, $p_2 = \theta(1-\theta)$, e $p_3 = (1-\theta)^2$, com $\theta < 1$, n elementos dessa população infinita são selecionados aleatoriamente e seus genótipos anotados: Sejam

   $X=$ número de elementos do tipo AA entre os n selecionados;

   $Y=$ número de elementos do tipo Aa entre os n selecionados;

   $Z=$ número de elementos do tipo aa entre os n selecionados.

   a. Qual a distribuição de probabilidade do vetor aleatório (X, Y)? Qual a distribuição de probabilidade de X?
   b. Qual a covariância entre X e Y?
   c. Seja $T = \dfrac{2X+Y}{2n}$. Qual a média e a variância de T? O que representa T?
   d. Determine como função de θ a probabilidade de que não haja indivíduos do tipo AA na amostra.
   e. Determine como função de θ a probabilidade de que dentre 12 indivíduos selecionados haja dois indivíduos do tipo AA, quatro do tipo Aa e seis do tipo aa.
   f. Para que valor de θ a probabilidade do item e é máxima?

2. Em um processo para a obtenção de folhas de flandres(folha de ferro laminada revestida de estanho), a probabilidade de uma chapa ser boa é 0,90, de ter somente uma falha é 0,06 e de ter mais de uma falha é 0,04. Qual é a probabilidade que numa amostra aleatória de seis chapas tenhamos 3 chapas perfeitas, 2 com apenas uma falha e 1 com mais de uma falha?

3. Em um canteiro de obras existem 3 alojamentos onde dormem os trabalhadores. Se os alojamentos são escolhidos aleatoriamente pelos empregados, qual é a probabilidade de que não tenhamos alojamento vazio se no momento existem 5 empregados?

4. A vida útil da camada de pavimentação de um trecho de 1km de uma estrada pode ser descrito através de uma lei de probabilidade chamada lognormal. Através da mesma tem-se os seguintes dados: $P(A) = P(T < 1) = 0,01$ e $P(B) = P(T < 3) = 0,50$, em que A e B são os eventos que 1 km de estrada necessite de reparo em 1 e 3 anos respectivamente. Para um trecho de estrada de 4 km e supondo que a vida de qualquer quilômetro de estrada é independente dos outros, determine a probabilidade de:

   a. não ocorrerem reparos no primeiro ano;
   b. ocorrerem reparos em dos 4 km no primeiro ano;
   c. ocorrerem reparos nos 3 primeiros anos de uso.

5. Em uma cidade no horário noturno 30% dos aparelhos de TV estão ligados em noticiários, 25% assistem a uma certa comédia e o restante em outros programas. Qual é a probabilidade de uma dada pesquisa com 7 telespectadores escolhidos

aleatoriamente mostre que exatamente 3 assistam aos noticiários e pelo menos 2 a comédia?

6. Uma rodovia está dividida em 8 trechos de igual comprimento, cada qual sob a jurisdição de uma guarnição de uma polícia rodoviária e todos igualmente perigosos. Sabendo-se que nessa rodovia há, em média, 6 desastres por dia, calcular a probabilidade de que, em determinado dia haja 4 trechos sem nenhum acidente, 3 trechos com um desastre cada e um trecho com mais de um desastre.

7. Uma peça é considerada boa (tipo 1) se sua dimensão L está entre l_1 e l_2; recuperável (tipo 2) se $L > l_2$ e perdida (tipo 3) se $L < l_1$. Em uma caixa temos 3 peças do tipo 1, 4 do tipo 2 e 3 do tipo 3. Um aluno escolhe uma peça e a medida dessa dimensão é verificada com o auxílio de um aparelho de precisão, após o que devolve a peça dentro da caixa. Seis alunos fazem essa operação. Qual é a probabilidade de entre os seis obtermos o seguinte resultado, 3 peças do tipo 1, 2 peças do tipo 2 e 1 do tipo 3?

8. Um antropólogo está interessado no tipo de sangue dos habitantes de uma certa ilha. Uma amostra de 770 habitantes forneceu os seguintes resultados 180, 360, 132 e 98. De acordo com suas especulações sobre a história racial da ilha os possuidores dos 4 tipos de sangue (A, B, AB e O) deveriam estar na proporção 0,16; 0,48; 0,20 e 0,16, respectivamente. Qual a sua conclusão para $\alpha = 0,05$?

9. Em uma das experiências de Mendel, o resultado foi 355 ervilhas amarelas e 123 verdes. Isto concorda com a teoria de Mendel, de acordo com a qual no presente caso, a relação de ervilhas amarelas/verdes deve ser 3:1? adotar $\alpha = 5\%$.

10. Um número é selecionado ao acaso de um intervalo $\{x; 0 < x < 2\}$. Seja $A_i = \{x; \frac{i-1}{2} < x \leq \frac{i}{2}\}$ em que $i = 1, 2, 3, 4.$. Uma certa hipótese atribui probabilidades p_{i0} a esses conjuntos de acordo com

$$p_{i0} = \int_{A_i} \frac{2-x}{2} dx, \quad i = 1, 2, 3, 4.$$

Se as frequências observadas dos conjuntos $A_i's$ são, respectivamente, 30, 30, 10 e 10, poderíamos aceitar H_0 ao nível de 5%.

11. O número de suicídios na cidade de São Paulo por semana é uma v.a. com distribuição de Poisson com $\lambda = 0,7$ suicídios/semana. Isto é verdade ao nível de 5% de significância?

Numero de suicidios	0	1	2	3 ou mais
Frequencia observada	32	17	8	3

12. Calcule a probabilidade de que, lançando-se seis dados, obtenha-se três pontos ímpares, dois pontos seis e um ponto quatro.

13. Conhecem-se as seguintes probabilidades:

- 0,6 de uma declaração de imposto de renda ser preenchida corretamente;
- 0,2 de uma declaração conter somente erros que favorecem ao contribuinte;

- 0,1 de uma declaração conter somente erros que favorecem ao fisco;
- 0,1 de conter ambos os tipos de erros.

Calcular a probabilidade de que, em 10 declarações selecionadas aleatoriamente, seis estejam corretas, duas contenham erros a favor do contribuinte e uma contenha erros a favor do fisco.