Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
**IMD0601 - Bioestatística**

# Distribuição multivariada

Prof. Dr. Tetsu Sakamoto
Instituto Metrópole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

# Baixe a aula (e os arquivos)

- Para aqueles que não clonaram o repositório:

```
> git clone https://github.com/tetsufmbio/IMD0601.git
```

- Para aqueles que já tem o repositório local:

```
> cd /path/to/IMD0601

> git pull
```

# Revisão

**Probabilidade condicional**

**Independência entre os eventos**

**Distribuição multivariada**

- **Discreta e Contínua;**
- **Distribuição conjunta;**
- **Distribuição marginal;**
- **Distribuição condicional.**

# Distribuições multivariadas comuns

- **Multinomial**
- **Normal multivariada**
- **Distribuição de Dirichlet**

# Distribuição multinomial

O mais comum entre as distribuições multivariada discreta;

Extensão da distribuição binomial ➜ 3 ou mais resultados possíveis;

Considere o experimento:

**n** experimentos independentes;
Cada experimento ➜ **r** tipos diferentes de resultados;
Probabilidade de qualquer resultado ➜ constante $(p_1, p_2, \ldots, p_r)$;
Soma das probabilidades ➜ 1 $(p_1+p_2+ \ldots + p_r = 1)$

# Distribuição multinomial

Considere o experimento:

**n** experimentos independentes;
Cada experimento ➔ **r** tipos diferentes de resultados;
Probabilidade de qualquer resultado ➔ constante ($p_1, p_2, \ldots, p_r$);
Soma das probabilidades ➔ 1 ($p_1 + p_2 + \ldots + p_r = 1$)

Se contarmos o número de vezes que cada resultado ocorre, temos um conjunto de **r** variáveis aleatórias ($X_1, X_2, \ldots, X_r$), onde $\mathbf{X_j}$ é a contagem do j-ésimo resultado;

$$\mathbf{X_j = x_j}$$

$$x_1 + x_2 + \ldots + x_r = n$$

# Distribuição multinomial

Como estamos lidando com uma série de eventos independentes:

$p_1^{x1} p_2^{x2} \ldots p_r^{xr}$

Análise combinatória ➔ número de combinações possíveis considerando r grupos;

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

Coeficiente multinomial

Combinando as duas equações temos:

$$p(x_1, x_2, \ldots, x_r) = \frac{n!}{x_1! x_2! \ldots x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}$$

# Distribuição multinomial

Aplicação:

- Modelar a distribuição conjunta do número de genótipos observados.

Exemplo simples:

- Gene contendo 2 alelos (**A** e **a**);
- Genótipos possíveis ➔ **AA**, **Aa**, **aa**
- Número de indivíduos com cada genótipo ➔ $\mathbf{n_{AA}}$, $\mathbf{n_{Aa}}$, $\mathbf{n_{aa}}$
- Formalizando em um modelo probabilístico ➔ $X = n_{AA}$, $Y = n_{Aa}$, $Z = n_{aa}$
- Probabilidades ➔ $P_{AA}$, $P_{Aa}$, $P_{aa}$

$$p(X = n_{AA}, Y = n_{Aa}, Z = n_{aa}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(P_{AA})^{n_{AA}}(P_{Aa})^{n_{Aa}}(P_{aa})^{n_{aa}}$$

# Distribuição multinomial

Suponha que:

- 20 indivíduos
- Genótipos ➔ $n_{AA} = 4$, $n_{Aa} = 14$, $n_{aa} = 2$
- Estimativa dos parâmetros multinomiais:
    - $P_{AA} = 0.2$; $P_{Aa} = 0.7$; $P_{aa} = 0.1$

Como não temos muitos dados, podemos simular dados com estes parâmetros para gerar gráficos e verificar outros detalhes da distribuição.

# Distribuição normal multivariada

Amplamente utilizada em estatística multivariada;

Distribuição normal univariada:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

distribuição conjunta de X e Y ➡ produto das distribuições marginais de X e Y(caso independentes).

$$f(x,y) = f(x)f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(y-\mu)^2}{2\sigma^2}}$$

# Distribuição normal multivariada

$$f(x,y) = f(x)f(y) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu)^2}{2\sigma^2}}\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(y-\mu)^2}{2\sigma^2}}$$

Reduzindo a fórmula anterior considerando distribuições normais padrão.

$$f(x,y) = \frac{1}{\sqrt{2\pi}}e^{\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}e^{\frac{y^2}{2}} = \frac{1}{\sqrt{2\pi}}e^{\frac{x^2+y^2}{2}}$$

# Distribuição Dirichlet

**Versão multivariada da distribuição beta**;

Distribuição beta ➜ Usado para modelar dados na forma de **proporção**;

$$f(x) = \frac{1}{\beta(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1 \quad \beta(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Exemplo ➜ modelar dados de proporção de nucleotídeos (A, T, C, G) em uma sequência de DNA;

$X_1, X_2, \ldots , X_k$ ➜ conjunto de proporções;

$X_1 + X_2 + \ldots + X_k = 1$

$$f(X_1, X_2, \ldots, X_k) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod (X_i)^{\alpha_i - 1}$$

# Distribuição Dirichlet

Exemplo simples ➜ modelar a distribuição conjunta das proporções de purinas (A e G) e pirimidinas (C e T) em uma dada sequência:

k = 2;

X1 ➜ proporção de purinas (p1); X2 ➜ proporção de pirimidinas (p2);

alpha1 = 1; alpha2 = 2;

$$f(X_1, X_2) = \frac{\Gamma(\sum_{i=0}^{2} \alpha_i)}{\prod_{i=0}^{2} \Gamma(\alpha_i)} \prod_{i=0}^{2} (x_i)^{\alpha_i - 1}$$

$$f(X_1, X_2) = \frac{\Gamma(3)}{\Gamma(1)\Gamma(2)} (p_1)^{1-1} (p_2)^{2-1}$$

$$f(X_1, X_2) = \frac{\Gamma(3)}{\Gamma(1)\Gamma(2)} (p_1)^{1-1} (1 - p_1)^{2-1}$$

Equivalente a distribuição beta de α=α1=1, ß=α2=2