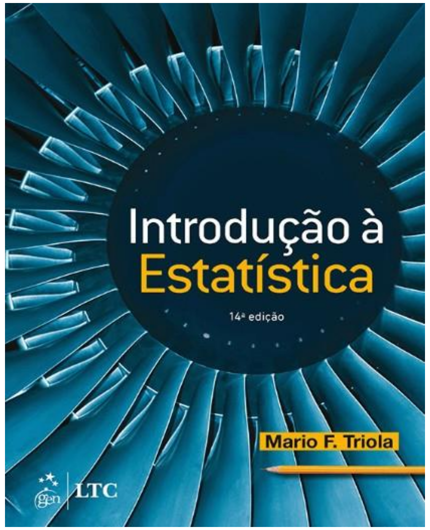
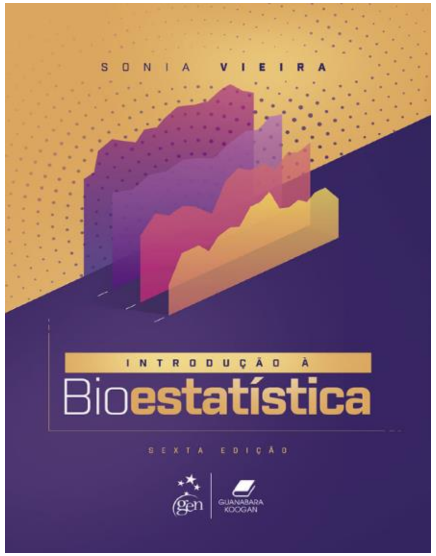
Aula 5 - Probabilidade, Variáveis Aleatórias e Principais Distribuições

Ao final desta aula você deverá entender

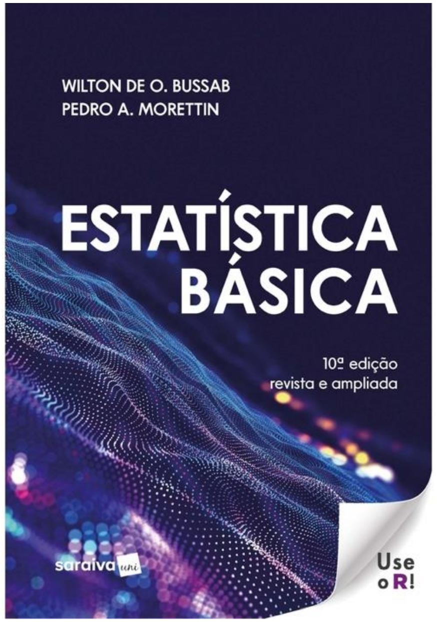
- Fundamentos da Probabilidade
- Variáveis Aleatórias e Distribuições de Probabilidade
- Distribuições de Probabilidade Discretas:
  - Binomial
  - Poisson
- Distribuições de Probabilidade Contínuas:
  - A Distribuição Normal
- Qual distribuição usar



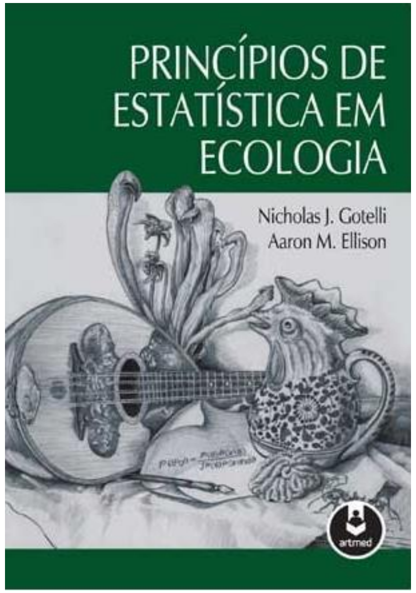
Cap 5, 6



Cap 8, 14, 15



Cap 5, 6, 7

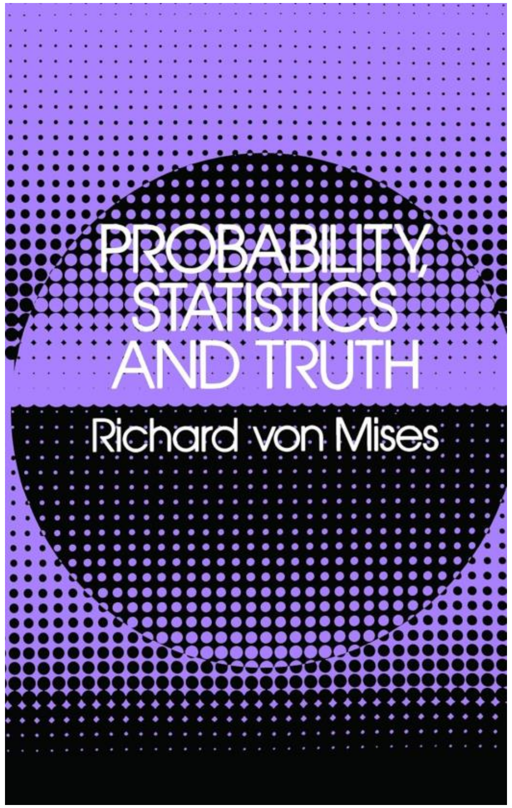


Cap 1, 2

Fundamentos da Probabilidade

O que é Probabilidade?

- Medida da incerteza de um evento
- Em biologia, muitos fenômenos são influenciados pelo acaso
- **Exemplos:**
  - A herança de um alelo específico de um gene.
  - O sexo de um filhote.
  - A captura de um animal em uma armadilha.



Porque precisamos de probabilidade em Biologia?

- **Exemplos Reais**
  - Genética: Probabilidade de herdar uma característica (ex: cor dos olhos)
  - Ecologia: Distribuição de espécies em um habitat
  - Epidemiologia: Risco de contrair uma doença
- **Ferramenta Essencial**
  - Base para testes estatísticos, modelagem e tomada de decisões

Conceitos Fundamentais

- **Experimento Aleatório:** Um processo que pode ter diferentes resultados, mesmo quando repetido da mesma maneira.
  - *Exemplo:* Cruzamento de duas plantas de ervilha heterozigotas (Aa x Aa).
- **Espaço Amostral (S):** O conjunto de todos os resultados possíveis de um experimento aleatório.
  - *Exemplo:* Os possíveis genótipos dos descendentes: {AA, Aa, aA, aa}.
- **Evento (E):** Um subconjunto do espaço amostral; um ou mais resultados de interesse.
  - *Exemplo:* O descendente ter o genótipo "aa".

Exemplo Genético Clássico

Ao cruzar duas plantas de ervilha heterozigotas (Aa x Aa), o genótipo do descendente é um evento aleatório. O espaço amostral é {AA, Aa, aA, aa}.

P(aa) = 25%

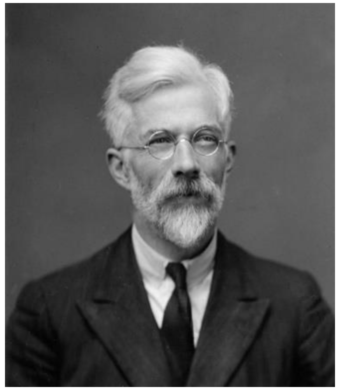
A probabilidade de um descendente ser homozigoto recessivo (aa).

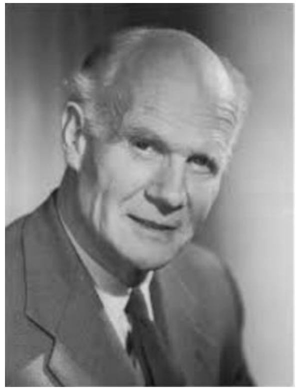
Definições de Probabilidade

- **Axiomática:** Define 3 axiomas que qualquer medida de probabilidade deve obedecer para ser matematicamente consistente.
  - Axioma 1: Não-negatividade
  - Axioma 2: Certeza (ou Normalização)
  - Axioma 3: Aditividade
- **Clássica:** Todos os resultados são igualmente prováveis.
  - $P(E) = \frac{(\text{Número de resultados favoráveis})}{(\text{Número total de resultados})}$
  - *Exemplo:* A probabilidade de um descendente de um cruzamento Aa x Aa ser "aa" é 1/4.

Definições de Probabilidade

- **Frequentista:** Baseada na frequência relativa de um evento *em um* **grande número de repetições**
  - $P(E) = \lim_{n \to \infty} \left(\frac{n(E)}{n} \right)$
  - *Exemplo:* Se em 1.000 nascimentos em um hospital, 510 são meninos, a probabilidade de nascer um menino é estimada em 0,51.
- **Bayesiana:** probabilidade é uma medida do **grau de crença** ou confiança na veracidade de uma proposição, **com base nas informações** e evidências **disponíveis**
  - Não é uma propriedade física e objetiva do mundo (como a frequência de caras em uma moeda), mas sim uma representação do nosso estado de conhecimento (ou incerteza) sobre algo
  - A principal força da abordagem Bayesiana é sua capacidade de atualizar crenças. À medida que novas evidências surgem, podemos usar um mecanismo formal para revisar nossa probabilidade inicial.





Teorema de Bayes

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Onde:

- **P(H|E) - Probabilidade a Posteriori:** É o que queremos descobrir. A nossa **crença atualizada** na hipótese H, *depois* de observar a evidência E.
- **P(H) - Probabilidade a Priori (Prior):** É a nossa **crença inicial** na hipótese H, *antes* de considerarmos a nova evidência.
- **P(E|H) - Verossimilhança (Likelihood):** A probabilidade de observar a evidência E, *assumindo que a hipótese H é verdadeira*.
- **P(E) - Evidência Marginal:** A probabilidade total de observar a evidência E. Atua como uma constante de normalização.
- O fluxo de pensamento é: **Crença Inicial (Prior) + Nova Evidência (Likelihood) → Crença Atualizada (Posterior)**

Comparação entre Frequentista e Bayesiana

Característica	Abordagem Frequentista	Abordagem Bayesiana
Definição Principal	Frequência relativa de um evento em um grande número de ensaios repetidos.	Grau de crença ou confiança na veracidade de uma proposição.
Natureza	Objetiva (a probabilidade é uma propriedade do mundo físico).	Subjetiva (a probabilidade é um estado de conhecimento de um observador).
O que descreve?	A taxa com que eventos ocorrem em experimentos.	A incerteza sobre qualquer proposição, seja ela repetível ou não.
Parâmetros do Modelo	Parâmetros (como a média μ) são constantes fixas, mas desconhecidas.	Parâmetros são variáveis aleatórias sobre as quais temos crenças (distribuições).
Resultado Principal	Intervalos de confiança e valores-p.	Distribuições de probabilidade a posteriori para os parâmetros.

Regras da Probabilidade

- **Regra da Adição**
  - **Eventos Mutuamente Exclusivos**
    - P(A ou B) = P(A) + P(B)
    - *Exemplo:* A probabilidade de um descendente ser AA **ou** aa.
  - **Eventos Não Exclusivos**
    - P(A ou B) = P(A) + P(B) - P(A ∩ B)
    - Exemplo: Animal mamífero **ou** carnívoro
- **Regra da Multiplicação**
  - **Eventos Independentes**
    - P(A e B) = P(A) * P(B)
    - *Exemplo:* A probabilidade de o primeiro filho de um casal ser menino **e** o segundo também
  - **Eventos dependentes**
    - P(A e B) = P(A) * P(A|B)
    - Exemplo: Probabilidade de sortear duas bolas vermelhas sem reposição.
- **Probabilidade Condicional:** A probabilidade de um evento A ocorrer, **dado que** o evento B já ocorreu.
  - P(A|B) = P(A e B) / P(B)
  - *Exemplo:* A probabilidade de um indivíduo ter uma doença, dado que o teste deu positivo.

Atividade - "Probabilidade na Genética” (5 min)

Variáveis Aleatórias e Distribuições de Probabilidade

Variáveis Aleatórias

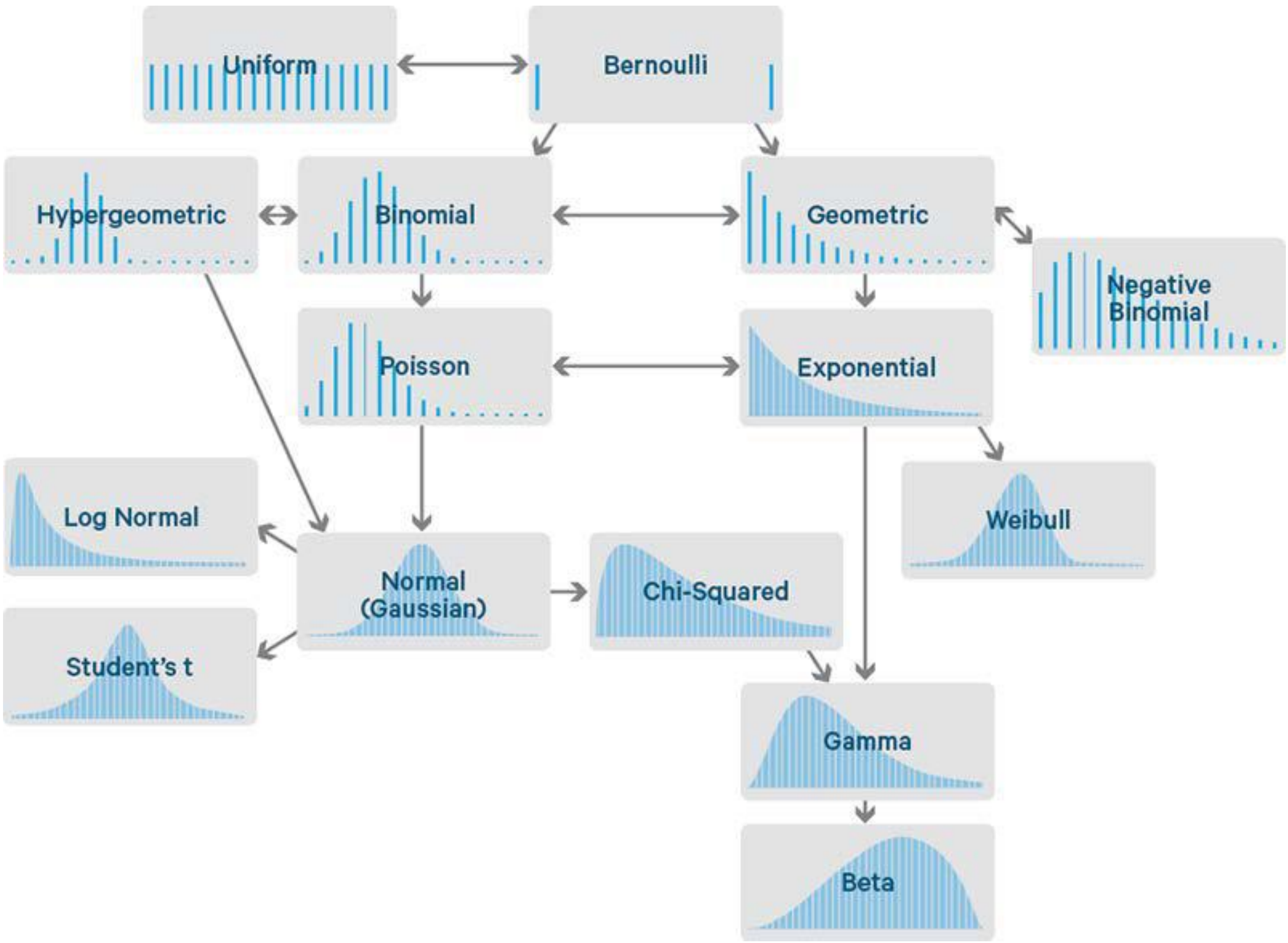
- Uma **variável aleatória (v.a.)** é uma variável cujo valor é uma realização (resultado) de um experimento aleatório (e.g., sorteio)
- **Tipos de Variáveis Aleatórias:**
  - **Discreta:** Assume um número ***finito*** ou contável de valores.
    - *Exemplos:*
      - Número de filhotes em uma ninhada.
      - Número de colônias de bactérias em uma placa de Petri.
  - **Contínua:** Pode assumir qualquer valor ***em um intervalo***.
    - *Exemplos:*
      - Altura de uma planta.
      - Peso de um animal.
      - Concentração de uma substância no sangue.

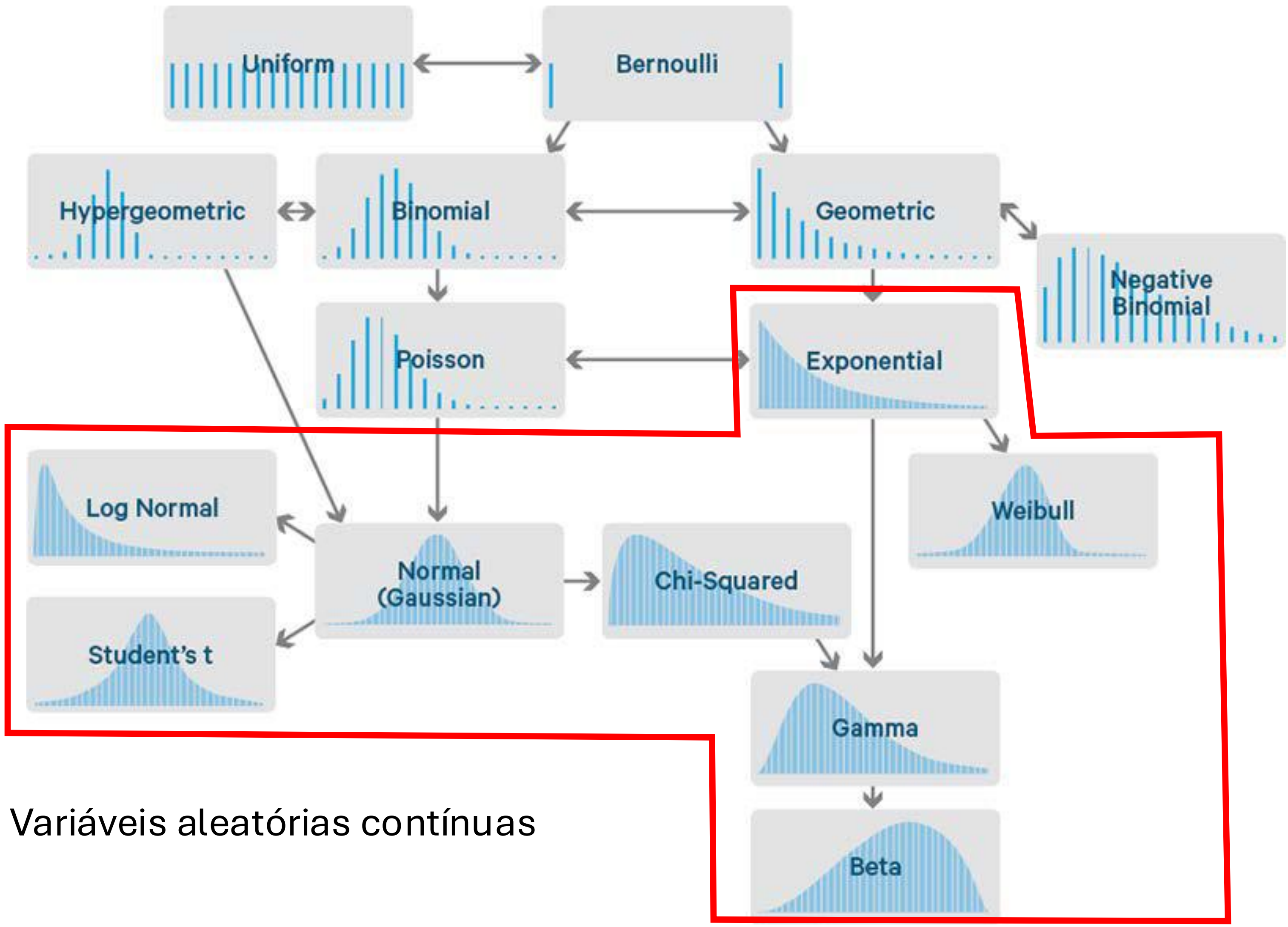
Variáveis Aleatórias

- **Funções:**
  - **Discretas:** Função de Probabilidade $P(X = x)$
  - **Contínuas:** Função Densidade de Probabilidade $f(x)$
- **Medidas:**
  - **Esperança (Média):**
    - $E(X) = \sum x\, P(X = x)$ (discreta)
    - $E(X) = \int x\, f(x)\, dx$ (contínua)
  - **Variância:** $Var(X) = E[X^2] - (E[X])^2$

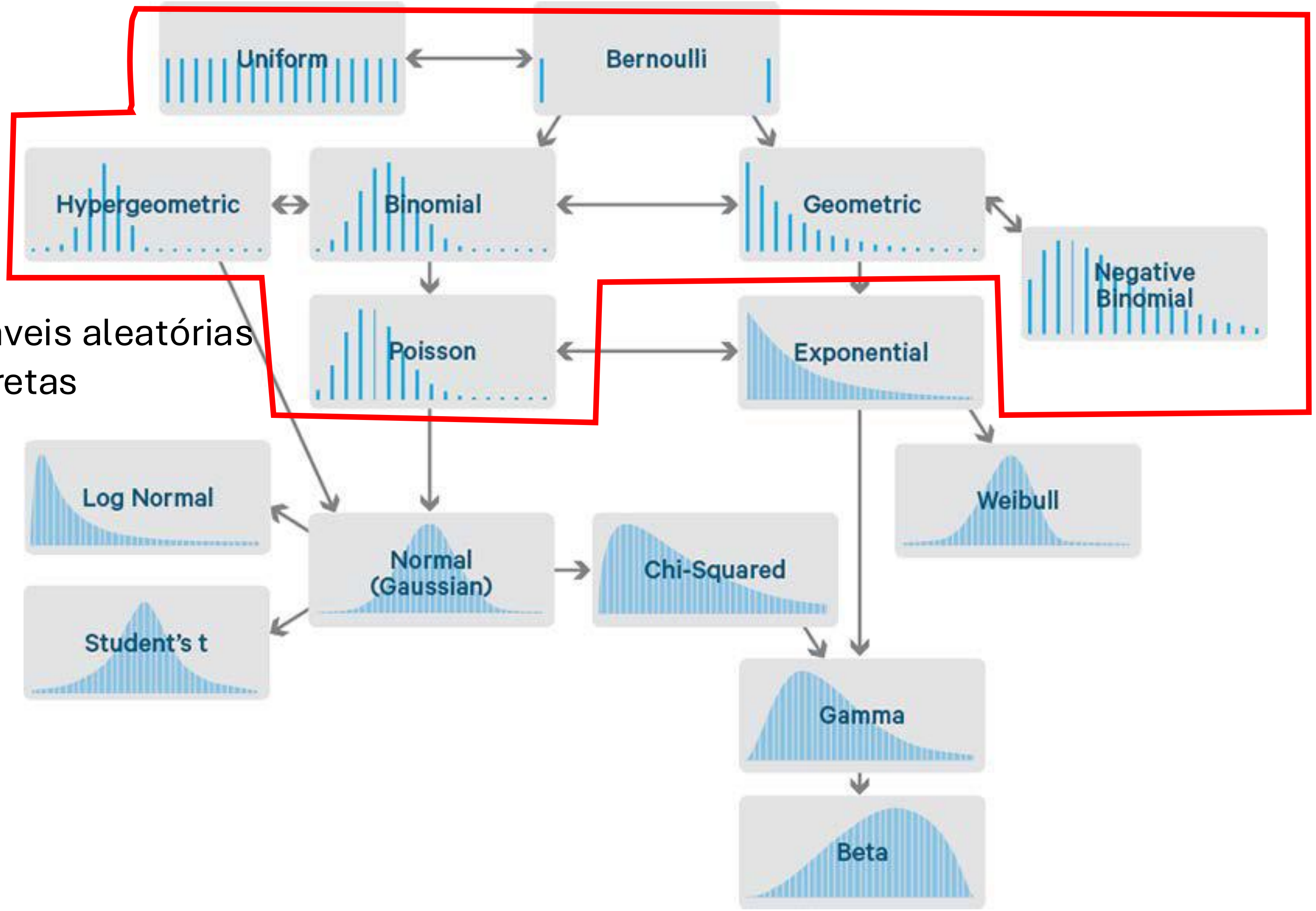
Distribuição de Probabilidade

- Uma **distribuição de probabilidade** descreve como as probabilidades se distribuem sobre os valores de uma variável aleatória.
- Para v.a. discretas, é uma lista de todos os valores possíveis e suas probabilidades.
- Para v.a. contínuas, é descrita por uma curva de densidade de probabilidade.

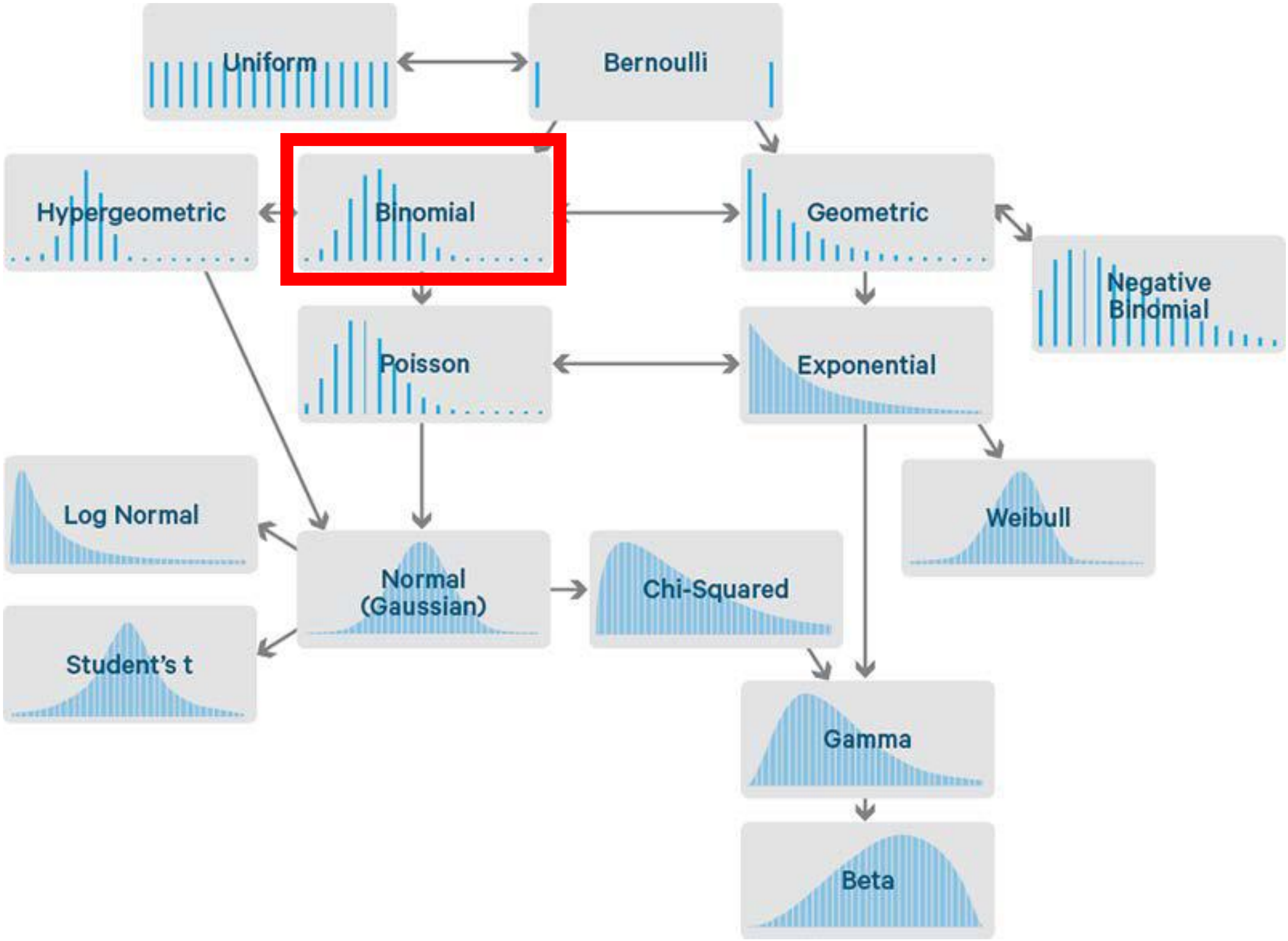

Uniform
Bernoulli
Hypergeometric
Binomial
Geometric
Negative Binomial
Poisson
Exponential
Log Normal
Weibull
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta



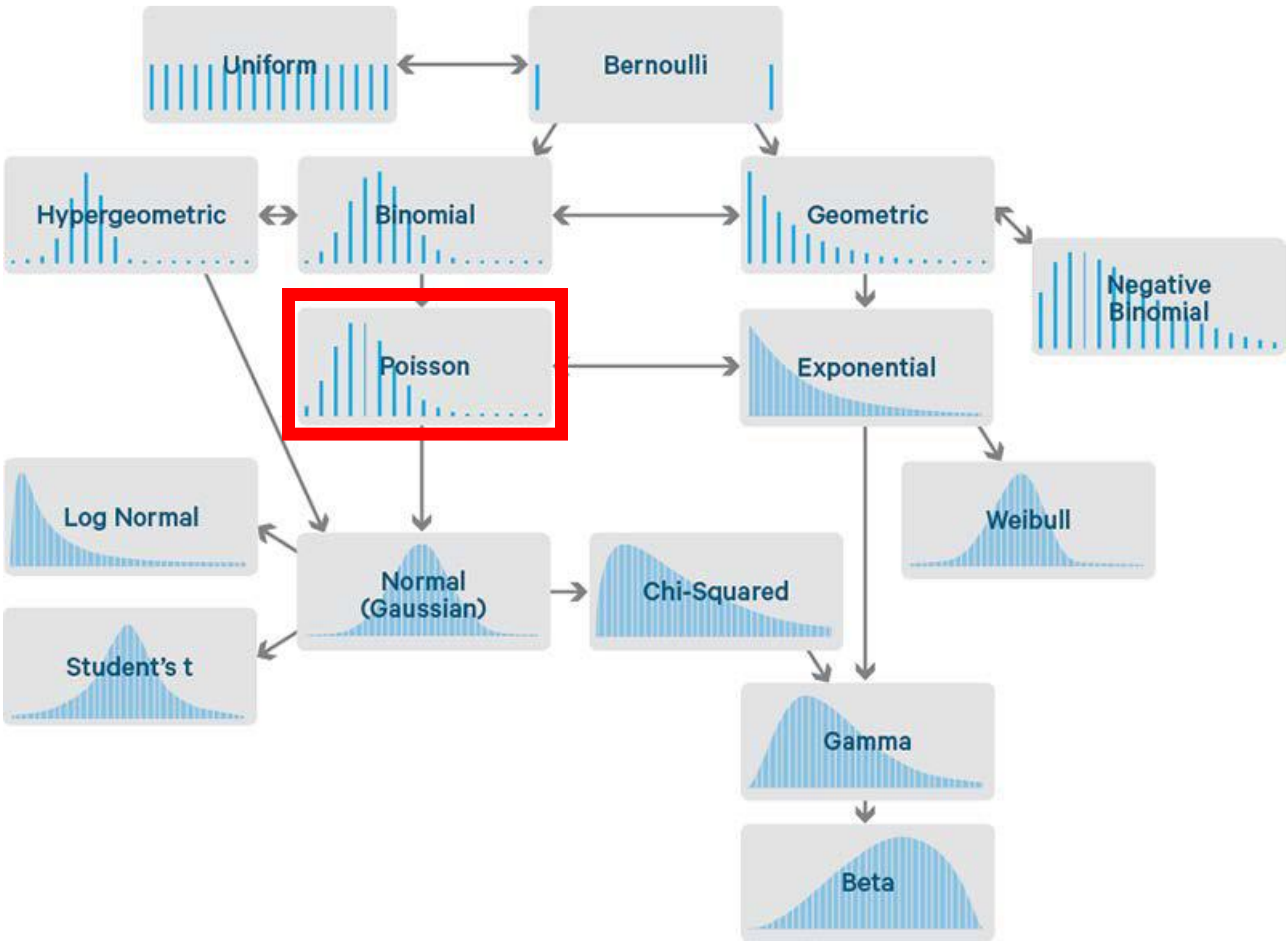
Variáveis aleatórias contínuas

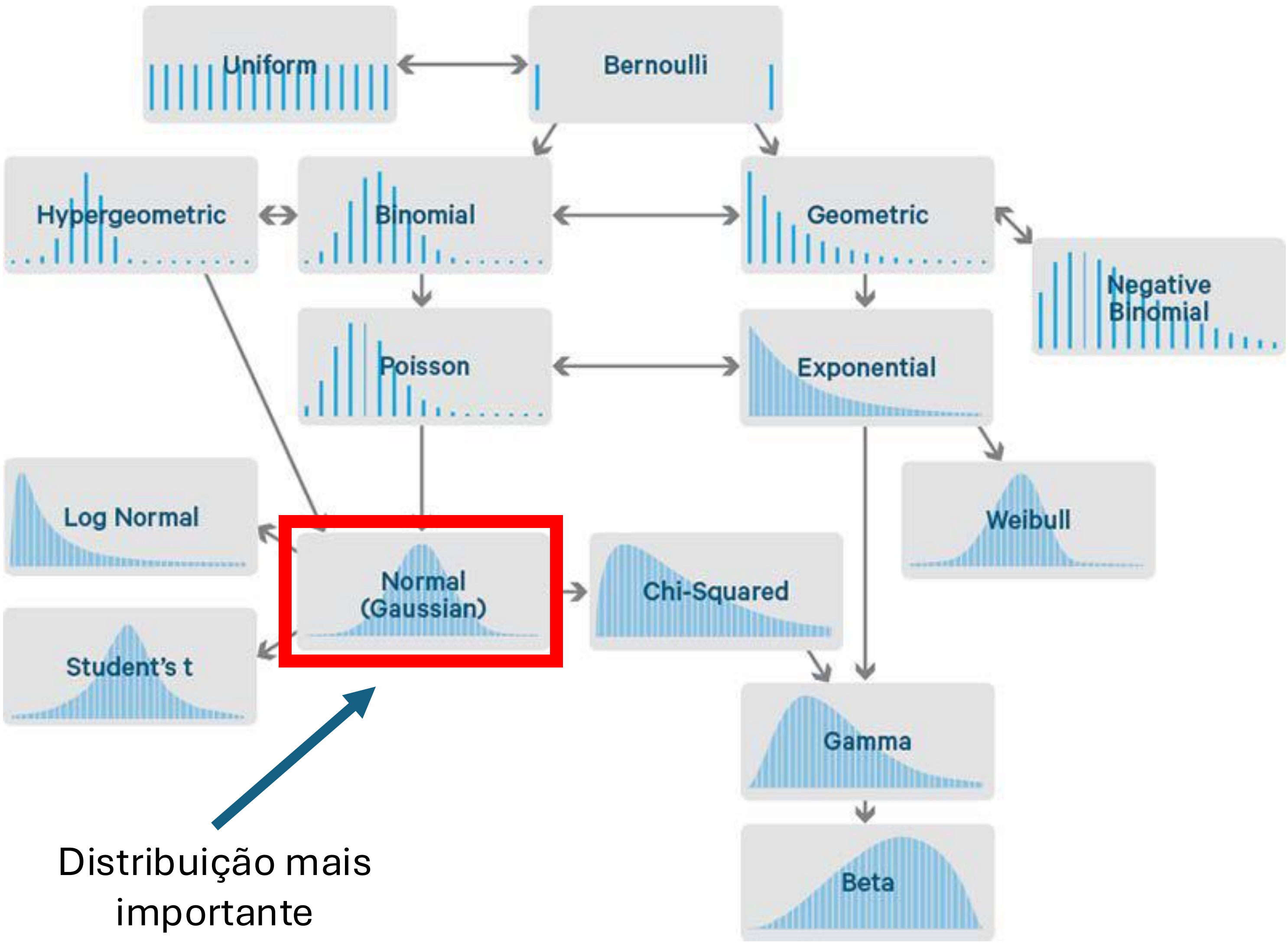


Variáveis aleatórias discretas



Uniform
Bernoulli
Hypergeometric
Binomial
Geometric
Negative Binomial
Poisson
Exponential
Log Normal
Weibull
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta


Uniform
Bernoulli
Hypergeometric
Binomial
Geometric
Negative Binomial
Poisson
Exponential
Log Normal
Weibull
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta

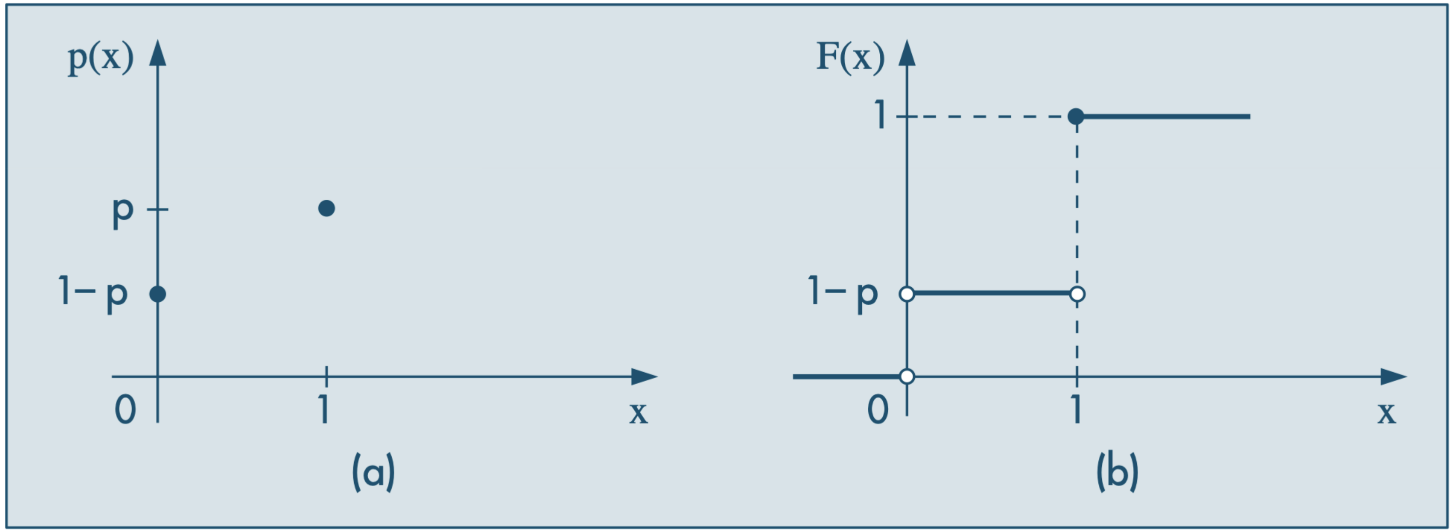


Distribuição mais importante
Para descrever var aleatórias contínuas

Distribuição Bernoulli

- **Definição:** Experimento com dois resultados (sucesso/fracasso)
- **Parâmetro:** *p* (probabilidade de sucesso)

Figura 6.10: Distribuição de Bernoulli (a) f.p. (b) f.d.a.



- **Exemplo Biológico:** Sobrevivência de um organismo após tratamento

Distribuição Binomial

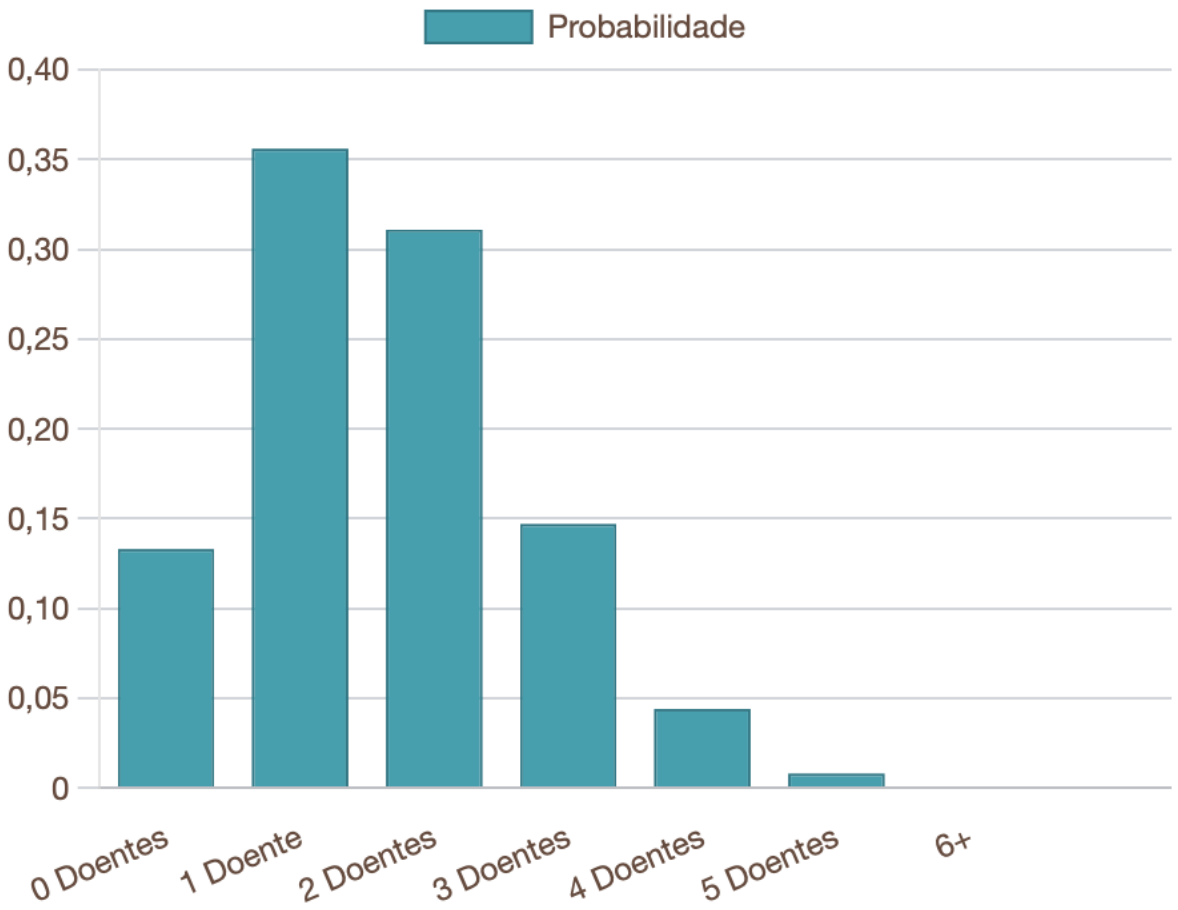
- Modelo para v.a. discretas.
- **Condições para usar a Binomial:**
  - Número fixo de tentativas (*n*).
  - Cada tentativa tem apenas dois resultados possíveis (sucesso ou fracasso).
  - A probabilidade de sucesso (*p*) é a mesma em todas as tentativas.
  - As tentativas são independentes.
- **Fórmula:**

$$P(X = x) \ = \left(\frac{n}{x}\right) p^{x}(1-p)^{n-x}$$

- **Exemplo em Biologia:**
  - Herança mendeliana (proporção de fenótipos)
  - Um casal tem 5 filhos. Qual a probabilidade de que exatamente 3 sejam meninos (considerando *p* =0.5)?
  - Sucesso/fracasso em experimentos de germinação

Distribuição Binomial

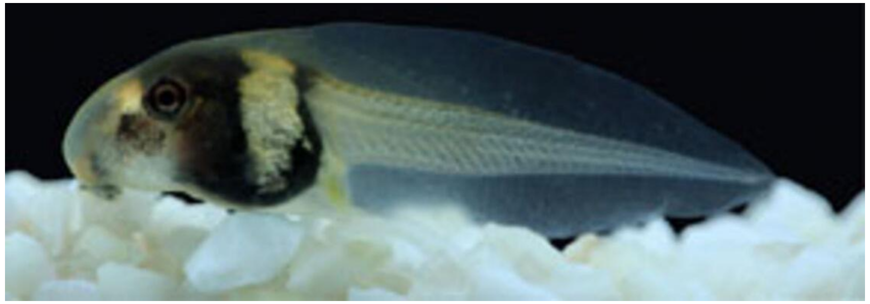
Modela o número de 'sucessos' em **'n'** tentativas independentes.



Contexto: Um gene recessivo causa uma doença com 25% de chance em descendentes de pais portadores. O gráfico mostra a probabilidade do número de filhotes doentes em uma ninhada de 8.

Distribuição Binomial – Exemplo 1

Há uma probabilidade de 0,3 de um girino, ao forragear em um corpo d’água, ser predado por uma larva de odonata. Determine as probabilidades de que, dentre seis girinos que estão forrageando no corpo d’água, 0, 1, 2, 3, 5 ou 6 sejam predados.



k ?

n ?

p ?

Distribuição Binomial – Exemplo 1

k = 0, 1, 2, 3, 4, 5 e 6

n = 6

p = 0,3

$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$p(0) = \binom{6}{0}(0{,}30)^0(0{,}70)^6 \approx 0{,}118$$

$$p(1) = \binom{6}{1}(0{,}30)^1(0{,}70)^5 \approx 0{,}303$$

$$p(2) = \binom{6}{2}(0{,}30)^2(0{,}70)^4 \approx 0{,}324$$

$$p(3) = \binom{6}{2}(0{,}30)^3(0{,}70)^3 \approx 0{,}185$$

$$p(4) = \binom{6}{4}(0{,}30)^4(0{,}70)^2 \approx 0{,}060$$

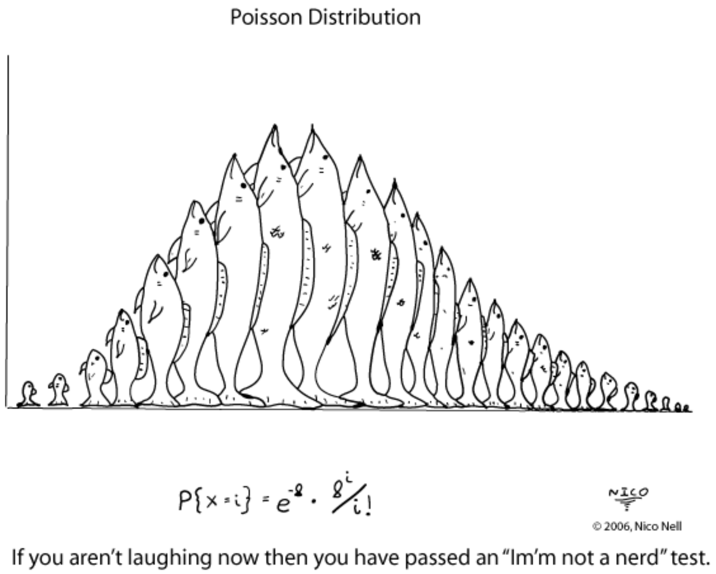
$$p(5) = \binom{6}{5}(0{,}30)^5(0{,}70)^1 \approx 0{,}010$$

$$p(6) = \binom{6}{6}(0{,}30)^6(0{,}70)^0 \approx 0{,}001$$

Distribuição de Poisson

- Modelo para v.a. discretas.
- Usada para modelar o número de ocorrências de um evento em um intervalo fixo de tempo ou espaço.
- **Condições:**
  - Os eventos ocorrem a uma taxa média constante (λ).
  - Os eventos são independentes.
- **Fórmula:**

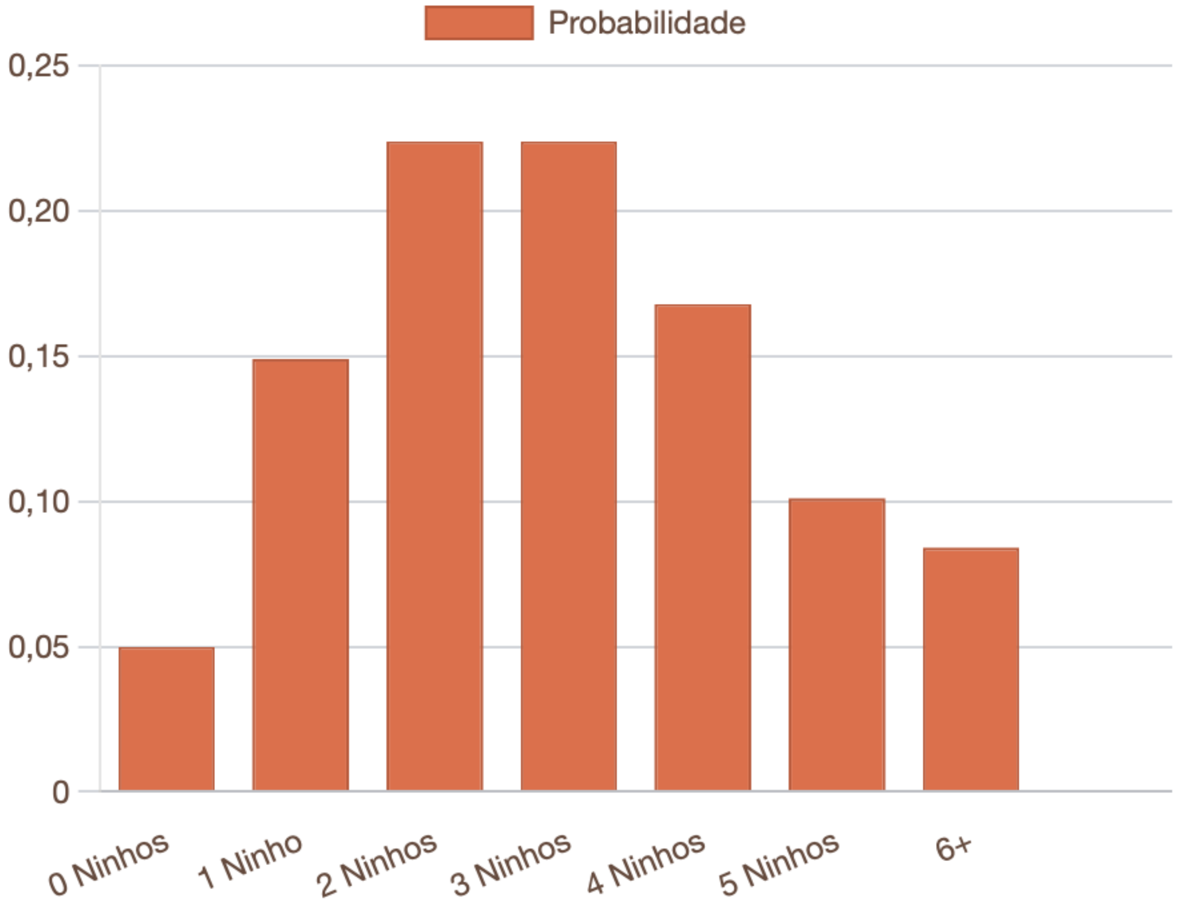
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$



- **Exemplos em Biologia:**
  - Número de mutações em um gene por geração.
  - Número de ninhos de uma espécie de ave por hectare de floresta.
  - Contagem de células em quadrantes microscópicos
  - Ocorrência de espécies em áreas delimitadas

Distribuição de Poisson

Modela o número de eventos raros ocorrendo em um intervalo de tempo ou espaço.



Contexto: Em média, um biólogo encontra 3 ninhos de uma ave rara por km² em uma reserva. O gráfico mostra a probabilidade de encontrar um certo número de ninhos em um km² específico.

Distribuição Poisson – Exemplo 1

- Suponha que observamos o número de visitas de borboletas em uma flor durante um período de 15 minutos.

- A probabilidade de uma borboleta chegar é a mesma para quaisquer dois períodos de tempo de igual cumprimento.

- A visita ou não de uma borboleta em qualquer período de tempo é independente da visita ou não de uma outra borboleta em qualquer outro período de tempo.

Distribuição Poisson – Exemplo 1

Suponha que o número médio de borboletas que vistiam a flor no período de 15 minutos é 10, então

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{10^x e^{-10}}{x!}$$

X: número de borboletas que chegam em qualquer período de 15 minutos

A probabilidade de 5 chegadas em 15 minutos

$$P(X = 5) == \frac{10^5 e^{-10}}{5!} = 0{,}0378$$

Distribuição Binomial e Poisson

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

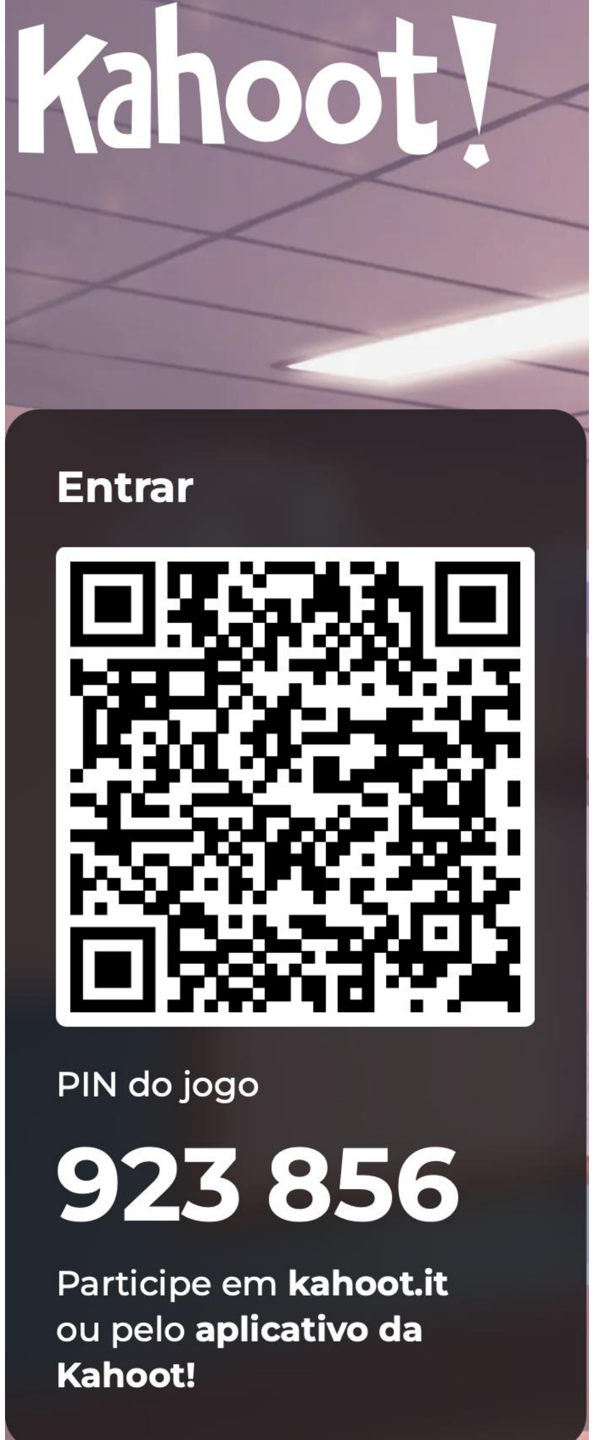
Se **N** for grande e ***p*** for pequeno, o evento é raro. Na prática:

$N \geq 50$ ou 100 e $Np < 5$ ou 10, então o evento é raro.

Neste caso, a distribuição binomial é muito aproximada da de Poisson, com

$$\boldsymbol{\lambda = Np}$$

Atividade 2 - "Qual distribuição usar?” (5 minutos)

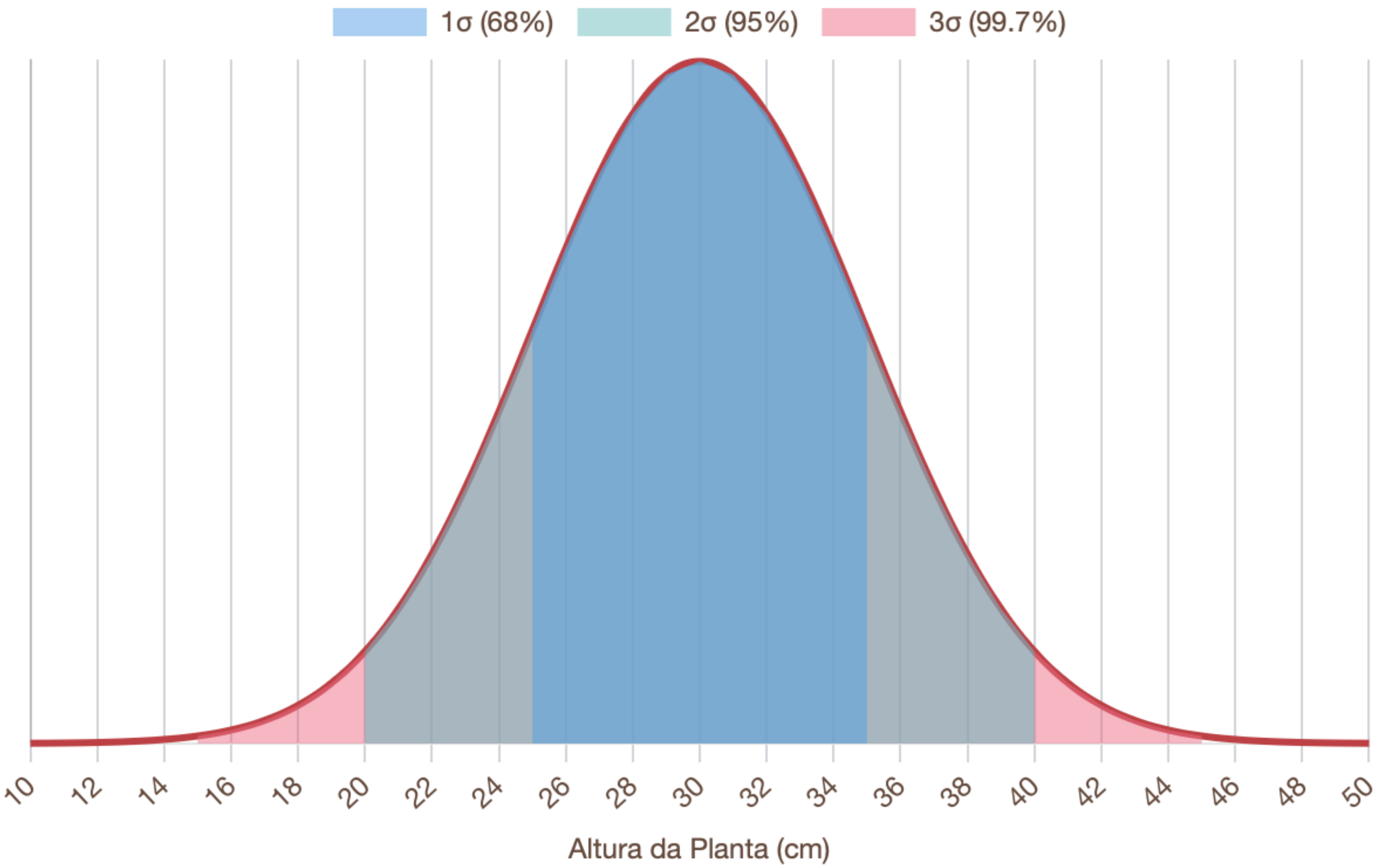


A Distribuição Normal

- A mais importante das distribuições contínuas.
- A curva Normal (ou Gaussiana) é em forma de sino, simétrica em torno da média.
- A importância da distribuição normal como um modelo de fenômenos quantitativos é devido em parte ao Teorema do Limite Central.
- **Parâmetros:**
  - **Média (μ):** centro da distribuição.
  - **Desvio Padrão** (σ): dispersão dos dados.
- **Exemplos em Biologia:**
  - Altura e peso de indivíduos em uma população.
  - Pressão sanguínea.
  - Comprimento das asas de uma espécie de borboleta.

A Curva da Vida: A Distribuição Normal

Muitas variáveis contínuas na biologia, como altura, peso e pressão sanguínea, seguem a famosa "curva em sino". Ela é simétrica em torno da média (μ) e sua dispersão é medida pelo desvio padrão (σ).



A Regra Empírica

Para dados normalmente distribuídos:

~68%

dos dados estão a 1σ da média.

~95%

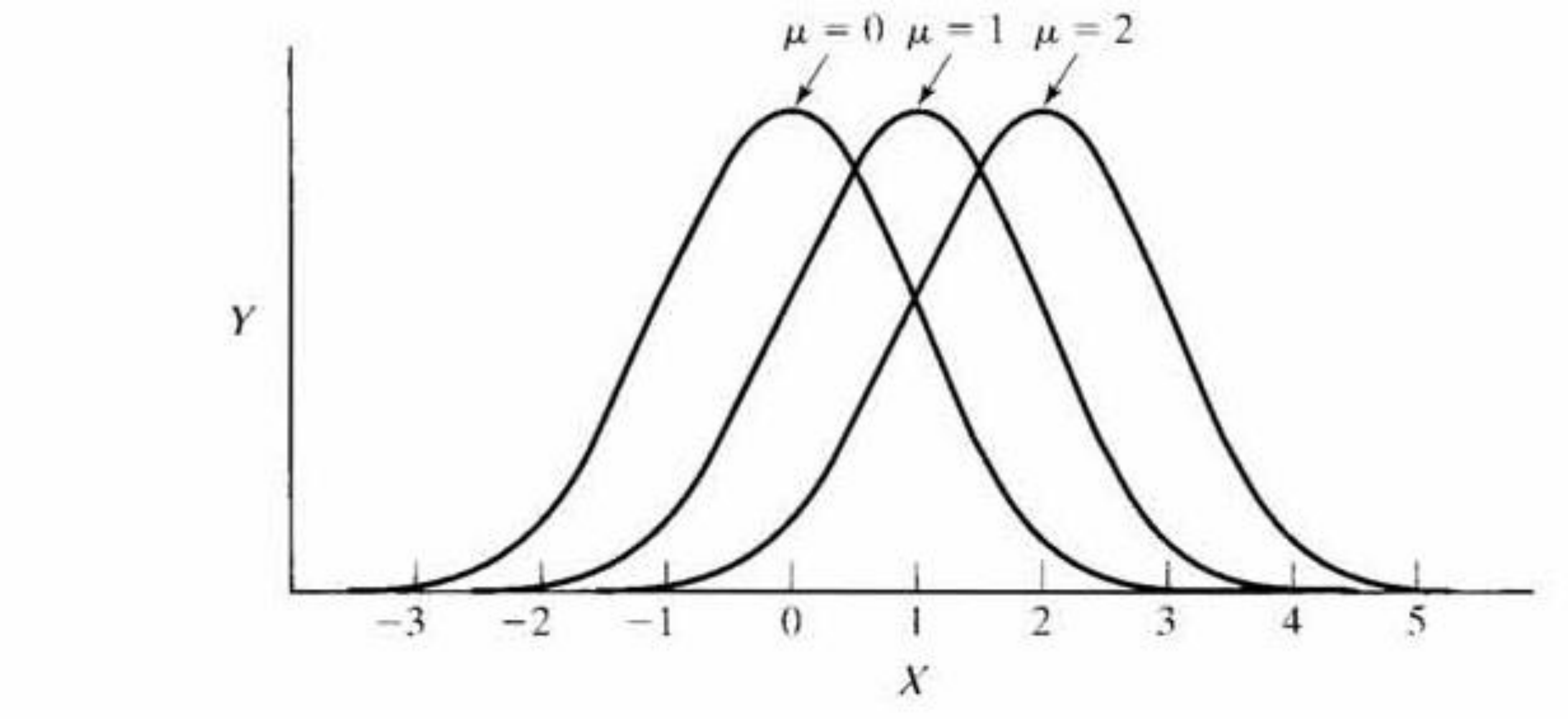
dos dados estão a 2σ da média.

~99.7%

dos dados estão a 3σ da média.

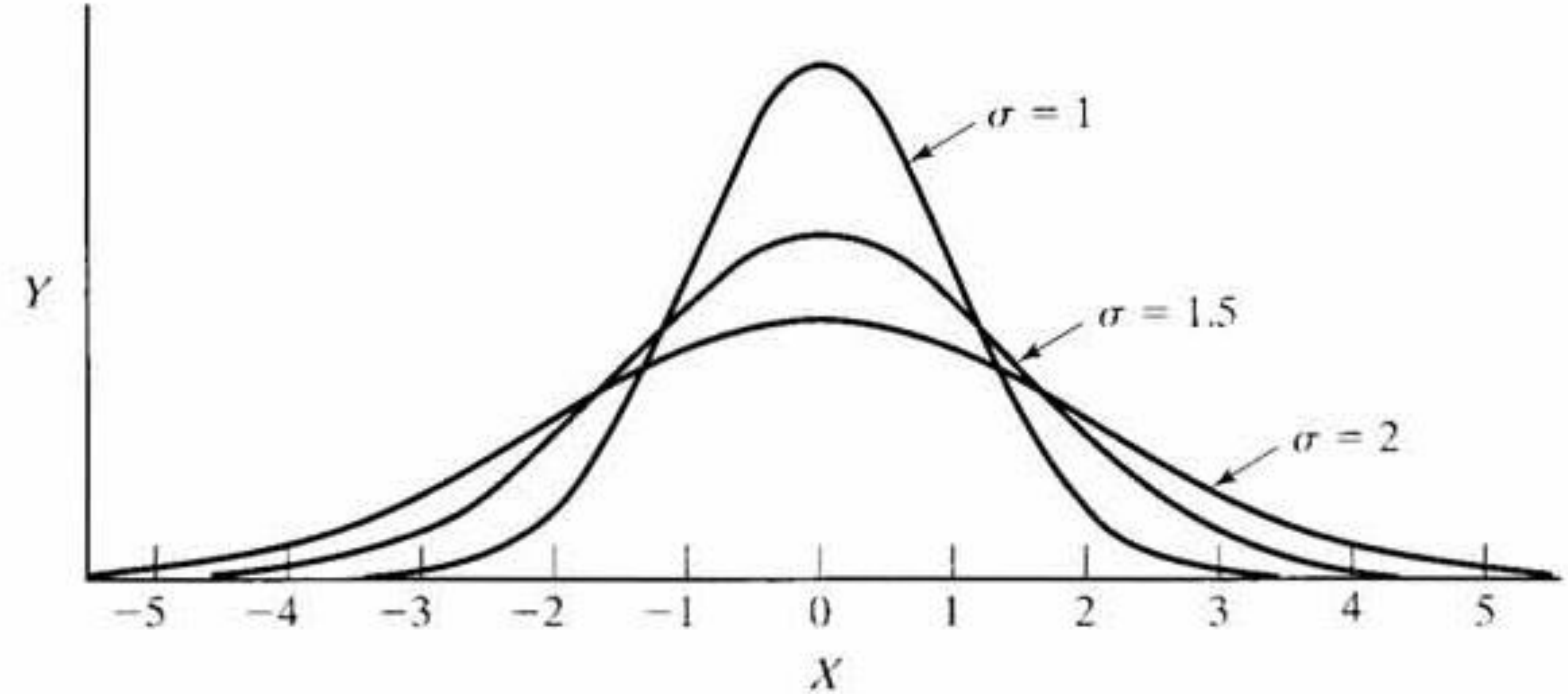
Contexto: Altura de uma população de plantas, com média (μ) de 30 cm e desvio padrão (σ) de 5 cm. A maioria das plantas terá altura entre 25 e 35 cm.

Variando a média (tendência central)



Zar 2010

Variando a variância (medida de dispersão)



Zar 2010

Distribuição *t* de Student

- Usada quando **σ é desconhecido** e **amostras são pequenas**
- Importante em testes de hipóteses

$$E(t) = 0, \quad \mathrm{Var}(t) = \frac{\nu}{\nu - 2}, \nu > 2,$$

Figura 7.28: A distribuição *t* de Student e a distribuição normal padrão.

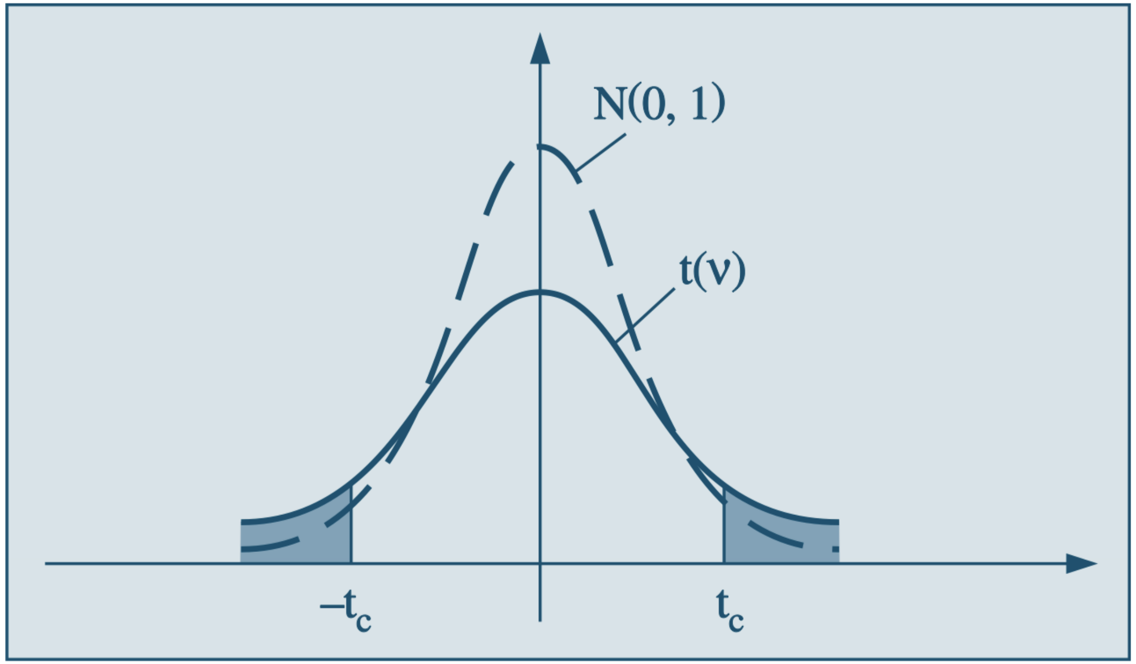


Tabela 7.2: Modelos para variáveis contínuas.

Modelo	$f(x)$	Parâmetros	$E(X)$, $\mathrm{Var}(X)$
Uniforme	$1/(\beta-\alpha),\ \alpha<x<\beta$	$\alpha,\ \beta$	$(\alpha+\beta)/2,\ (\beta-\alpha)^2/12$
Exponencial	$1/\beta\ e^{-t/\beta}, t>0$	β	$\beta,\ \beta^2$
Normal	$\dfrac{1}{\sigma\sqrt{2\pi}}\ \exp\left\{\left(\dfrac{x-\mu}{\sigma}\right)^2\right\}, -\infty<x<\infty$	$\mu,\ \sigma$	$\mu,\ \sigma^2$
Gama	$\beta^{-\alpha}/\Gamma(\alpha)\ \ x^{\alpha-1}\ e^{-x/\beta}, x>0$	$\beta>0,\ \alpha>0$	$\alpha\beta,\ \alpha\beta^2$
Qui-quadrado	$\dfrac{2^{-\nu/2}}{\Gamma(\nu/2)}\ y^{\nu/2-1}e^{-y/2},\ y>0$	ν	$\nu,\ 2\nu$
t-Student	$\dfrac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}}\left(1+\dfrac{t^2}{\nu}\right)^{-(\nu+1)/2}, -\infty<t<\infty$	ν	$0,\ \nu/(\nu-2)$
F-Snedecor	$\dfrac{\Gamma\left(\dfrac{(\nu_1+\nu_2)}{2}\right)}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)}\left(\dfrac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}\dfrac{w^{\frac{\nu_1-2}{2}}}{\left(1+\dfrac{\nu_1 w}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}}},\ w>0.$	$\nu_1,\ \nu_2$	$\dfrac{\nu_2}{\nu_2-2},\ \dfrac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$

Usos e Aplicações da Distribuição Normal

A Importância da Distribuição Normal Padrão

- A **Distribuição Normal Padrão** tem média 0 e desvio padrão 1.
- Podemos converter qualquer variável normal para a normal padrão usando o **Escore-Z**:

$$Z = \frac{(X - \mu)}{\sigma}$$

- O Escore-Z nos diz quantos desvios padrão um valor está da média.
- **Aplicação:** Podemos calcular a probabilidade de um valor ocorrer em qualquer distribuição normal.
  - *Exemplo:* Se a altura média dos homens em uma população é 175 cm com um desvio padrão de 7 cm, qual a probabilidade de um homem ter mais de 182 cm?

Cálculo passo-a-passo do Escore-Z

- Formula: $Z = \frac{(X-\mu)}{\sigma}$
- **X** = O valor que queremos investigar (182 cm)
- **μ** = A média da população (175 cm)
- **σ** = O desvio padrão da população (7 cm)
- **Passo 1: Calcular o Escore-Z**
  1. Z = (182 - 175) / 7
  2. Z = 7 / 7
  3. **Z = 1**
- Isso significa que a altura de 182 cm está exatamente **1 desvio padrão acima da média**.

Cálculo passo-a-passo do Escore-Z

- Agora, usamos as propriedades da curva normal padrão:
- A curva é simétrica, e 50% dos valores estão acima da média e 50% estão abaixo.

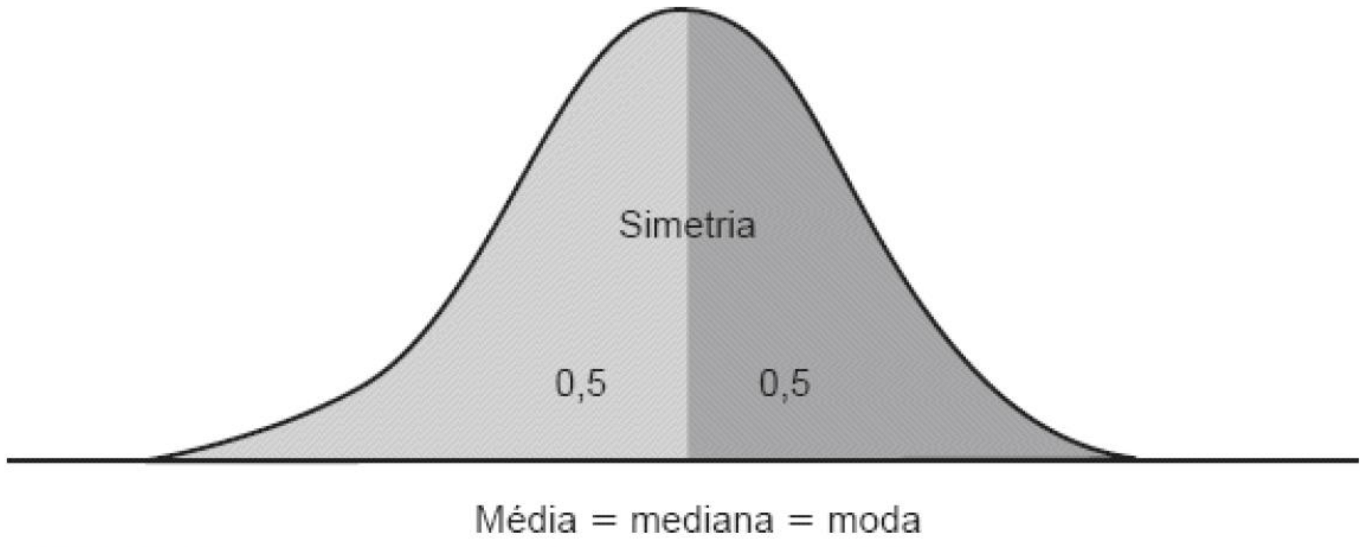
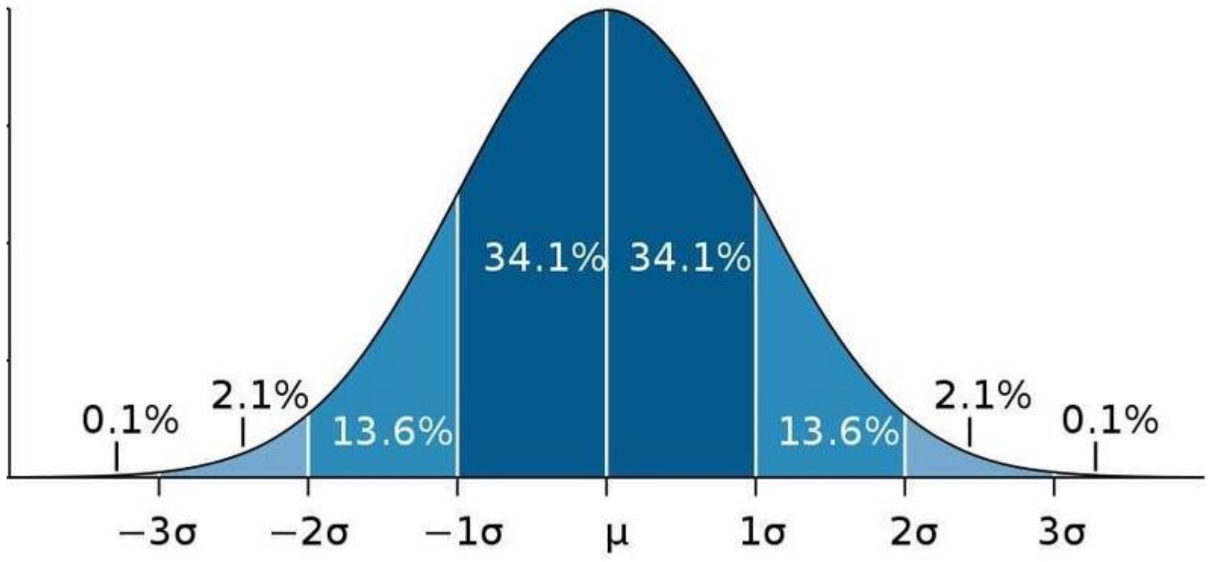


Figura 8.4 Distribuição normal: simetria em torno da média

Sonia Vieira 2022

Cálculo passo-a-passo do Escore-Z

- Pela "Regra Empírica (68-95-99.7)", sabemos que aproximadamente **68%** dos valores estão dentro de 1 desvio padrão da média (entre Z = -1 e Z = 1).
- Como a curva é simétrica, metade dessa área (34%) está entre a média (Z=0) e Z=1.



https://openbooks.library.umass.edu/p132-lab-manual/chapter/the-normal-distribution-and-standard-deviation/

Cálculo passo-a-passo do Escore-Z

- Se 50% de todos os homens têm altura acima da média e 34% estão entre a média e 182 cm (Z=1), a probabilidade de um homem ter *mais* de 182 cm é a área restante na cauda superior da curva.

- Probabilidade = 50% - 34% = 16%

- **Resposta:**

- A probabilidade de um homem selecionado ao acaso nessa população ter mais de 182 cm de altura é de aproximadamente **16%**.

Z-score

- Conceito:

Distância que um valor observado está da média em unidades de desvio padrão

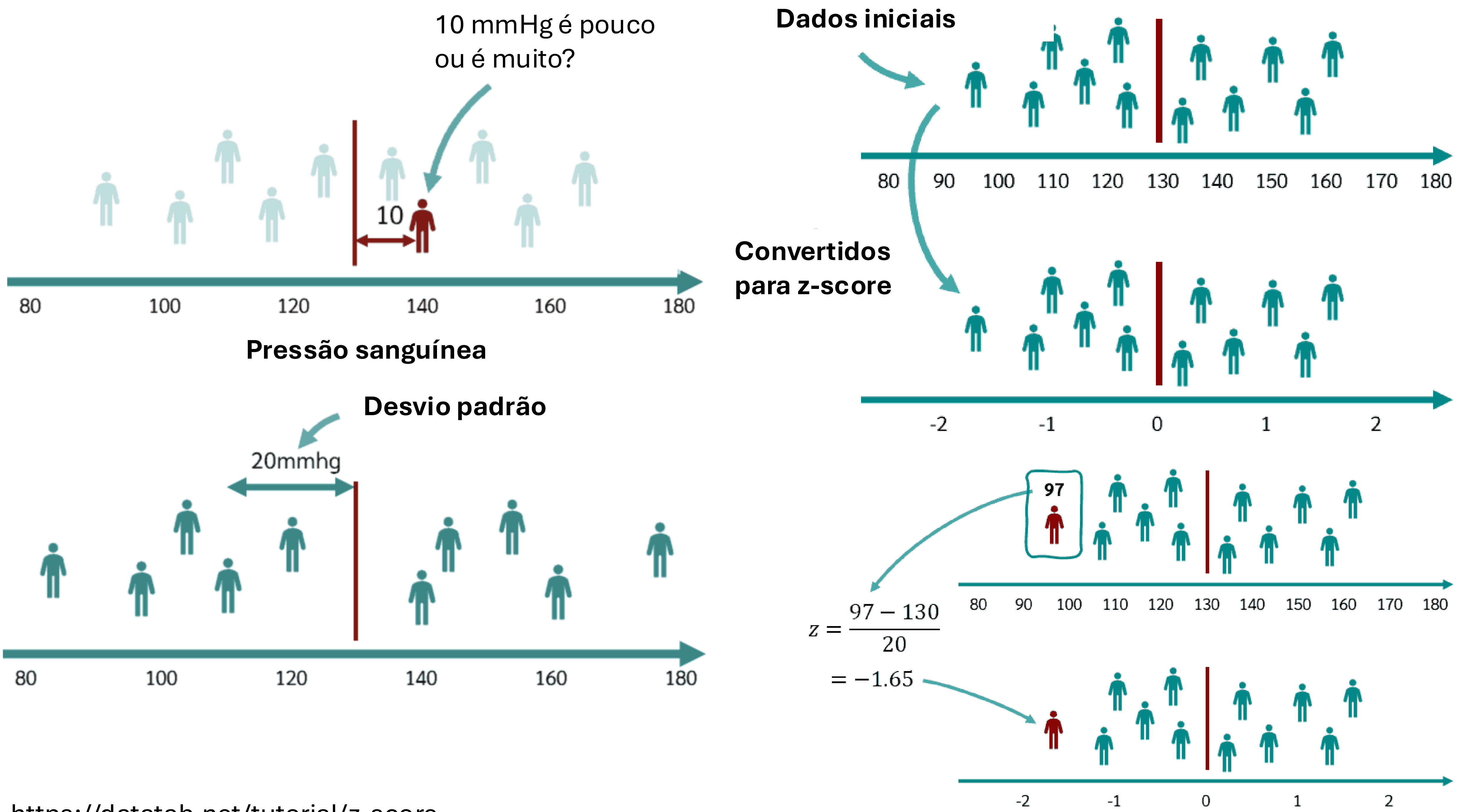
Valor observado

Valor médio

$$z = \frac{x - \mu}{\sigma}$$

z-Score

Desvio padrão



https://datatab.net/tutorial/z-score

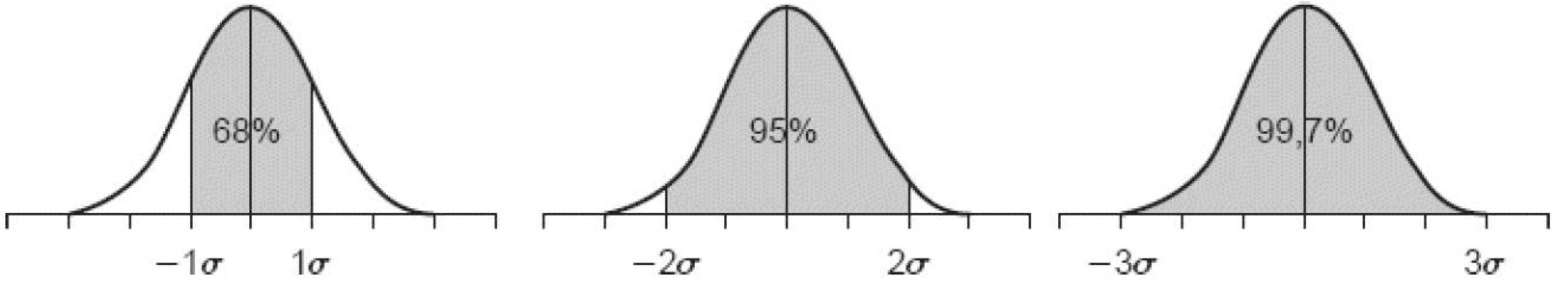


Figura 8.5 Probabilidades na distribuição normal

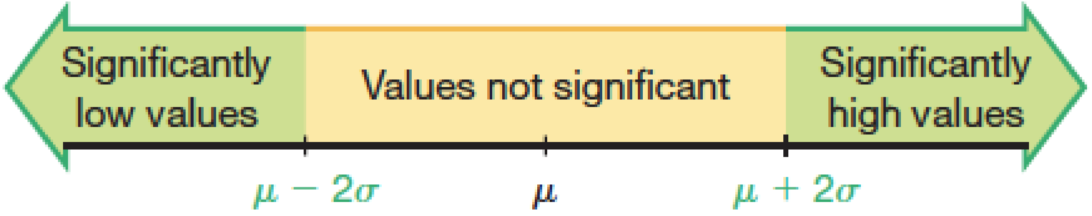


FIGURE 3-3 Range Rule of Thumb for Identifying Significant Values

Triola, M. 2024

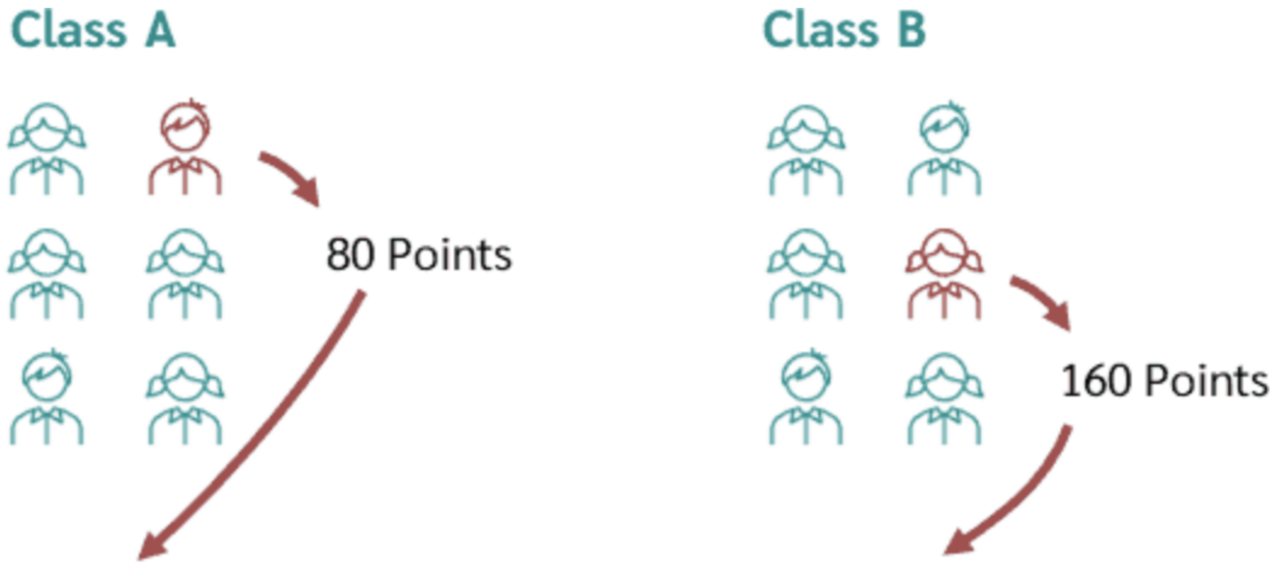
Por que isso é importante?

Porque para distribuições simétricas, sabemos que 95% dos dados estarão a uma distância de 2 desvios padrões da média! 😉

Por que isso é importante?

Permite comparar amostras vindas de populações diferentes

We now want to compare the performance of Max from class A, who scored 80 points, with the performance of Emma from class B, who scored 160 points.



$$z = \frac{80 - 70}{10} = 1 \qquad z = \frac{160 - 140}{20} = 1$$

https://datatab.net/tutorial/z-score

Dinâmica - "O Normal e o Raro" (5 minutos)

- Formem 4 grupos de 4-5 alunos, cada.

- Utilizem o escore-z para responder se as seguintes situações representam eventos raros ou comuns
  - Raros: dentro do intervalo de 2 desvios padrões
  - Comuns: acima (positivo) ou abaixo (negativo) de 2 desvios padrões

- Cada grupo discute e apresenta sua resposta.

Dinâmica - "O Normal e o Raro" (5 minutos)

- ***Grupo 1:*** A concentração de glicose no sangue de uma população de peixes segue uma distribuição normal com média 60 mg/dL e desvio padrão 5 mg/dL. Um peixe com 75 mg/dL seria considerado 'normal' ou 'raro'? Justifiquem usando o Escore-Z.
- ***Grupo 2:*** O comprimento dos ovos de uma espécie de beija-flor segue uma distribuição normal com média de 15 mm e desvio padrão de 0,5 mm. Um biólogo encontra um ovo com 16,5 mm de comprimento. Este ovo tem um tamanho comum ou incomum para a espécie?

Dinâmica - "O Normal e o Raro" (5 minutos)

- ***Grupo 3:*** A pressão arterial sistólica em um grupo de capivaras adultas em repouso é normalmente distribuída, com uma média de 130 mmHg e desvio padrão de 10 mmHg. Uma capivara examinada apresentou uma pressão de 145 mmHg. Este valor é considerado normal ou raro?
- ***Grupo 4:*** O peso dos frutos de uma variedade de tomate segue uma distribuição normal com média de 150 g e desvio padrão de 12 g. Em uma colheita, um tomate pesou apenas 120 g. Este peso é normal ou atípico para esta variedade?

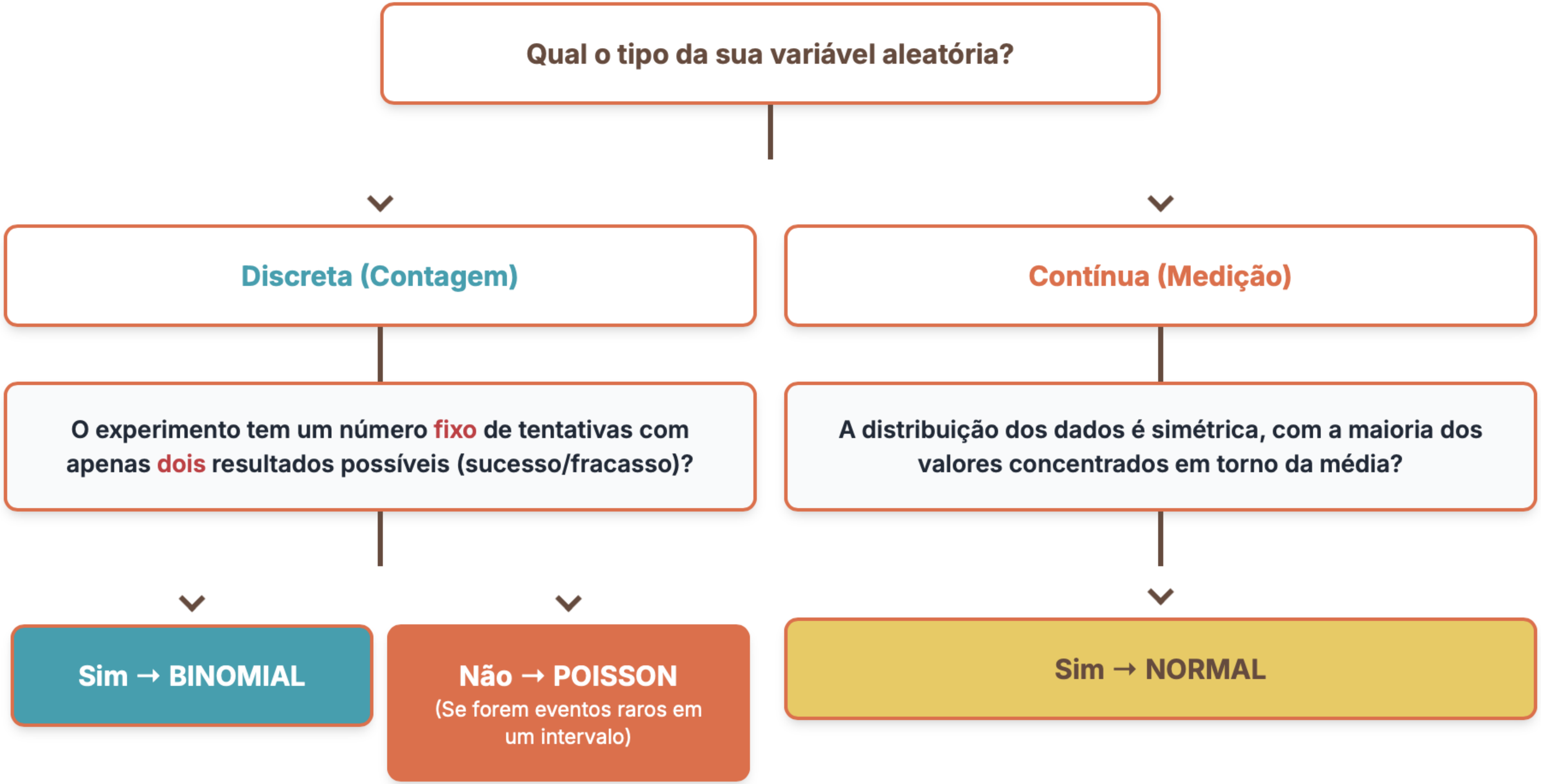
Como escolher qual distribuição usar?

Comparação entre distribuições

Distribuição	Tipo de Variável	Parâmetros Chave	Média (μ)	Variância ($\sigma2$)	Caso de Uso Biológico Clássico
Binomial	Discreta	n (ensaios), p (prob. de sucesso)	np	np(1−p)	Número de indivíduos infectados em uma amostra de n.
Poisson	Discreta	λ (taxa média)	λ	λ	Número de mutações raras em uma fita de DNA.
Normal	Contínua	μ (média), σ (desv. padrão)	μ	$\sigma2$	Distribuição de uma característica quantitativa como altura ou peso.

Qual Modelo Usar? Um Guia de Decisão

A escolha do modelo probabilístico correto é crucial. Siga este fluxo para decidir qual distribuição se ajusta melhor ao seu problema biológico.



Recaptulando

- Probabilidade: a linguagem da incerteza.
- Variáveis Aleatórias: discretas e contínuas.
- Distribuições de Probabilidade: modelos para o comportamento de variáveis aleatórias.
  - Binomial: contagem de sucessos em 'n' tentativas.
  - Poisson: contagem de eventos em um intervalo.
  - Normal: modelo para dados contínuos e simétricos.