Estatística para Ciências de Dados

Aula 3: Distribuições de Probabilidade

Mariana Cúri
ICMC/USP
mcuri@icmc.usp.br

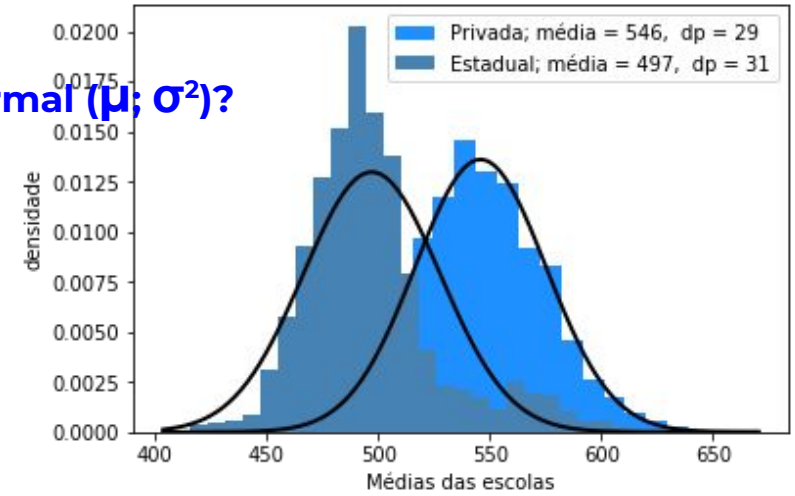
Conteúdo

1. Modelos discretos
    a. Bernoulli
    b. Binomial
    c. Geométrica
    d. Binomial negativa
    e. Hipergeométrica
    f. Poisson
2. Modelos contínuos
    a. Exponencial
    b. Normal
3. Resultados assintóticos

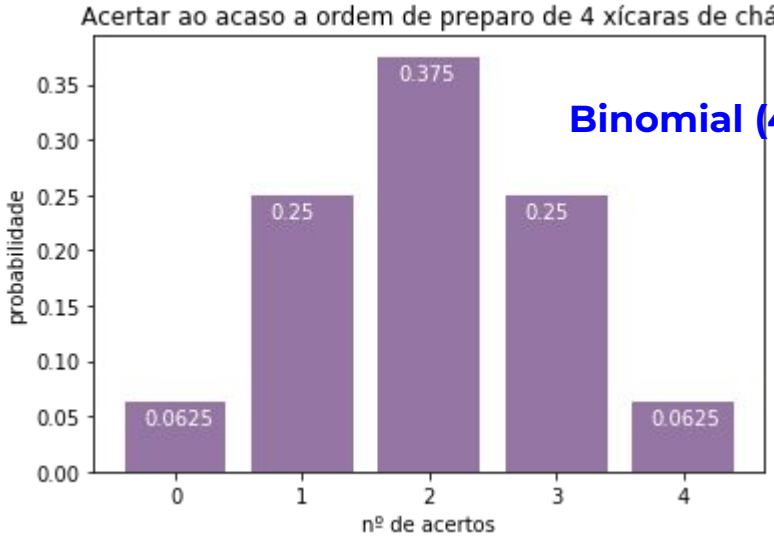
Motivação

Na prática, supõe-se que a variável de interesse, X, segue determinada distribuição de probabilidades na população, ou seja, define-se o modelo probabilístico.

Normal (μ; σ^2)?



Binomial (4; ½)?



Motivação

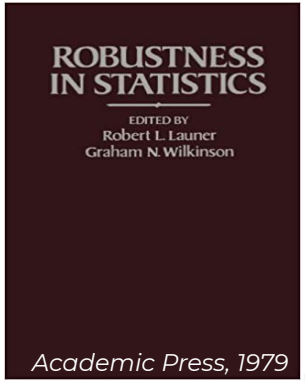
ROBUSTNESS IN THE STRATEGY OF SCIENTIFIC MODEL BUILDING[†]

G. E. P. Box

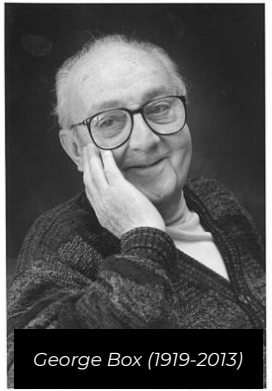
Robustness may be defined as the property of a procedure which renders the answers it gives insensitive to departures,

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do



Academic Press, 1979



George Box (1919-2013)

Modelos discretos

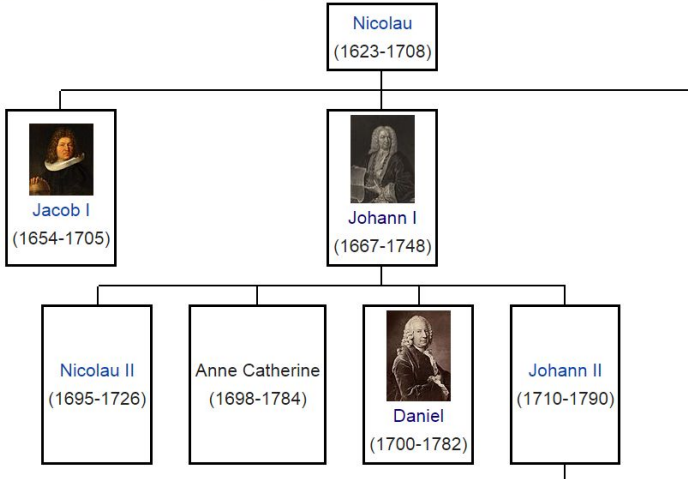
Modelo	$f(x)$	Suporte	parâmetros	$E(X)$	V(X)
Bernoulli	$p^x(1-p)^{1-x}$	$x = 0, 1$	p	p	$1-p$
Binomial	$C_x^n p^x (1-p)^{n-x}$	$x = 0, 1, \cdots, n$	n, p	np	$np(1-p)$
Geométrica	$p(1-p)^{x-1}$	$x = 1, 2, \cdots$	p	$1/p$	$(1-p)/p^2$
Binomial Negativa	$C_{k-1}^{x-1} p^k (1-p)^{x-k}$	$x = k, k+1, \cdots$	k, p	k/p	$k(1-p)/p^2$
Hipergeométrica	$\frac{C_x^k C_{n-x}^{N-k}}{C_n^N}$	$max\{0, n-(N-k)\} \leq x \leq min\{n, k\}$	N, n, k	nk/N	$\frac{N-n}{N-1} n \frac{k}{N}(1-\frac{k}{N})$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	$x = 0, 1, \cdots$	λ	λ	λ

Modelos discretos: Bernoulli

Um experimento que resulta em sucesso ou fracasso

Exs:

- peça com ou sem defeito de uma linha de produção
- resultado + ou - de um exame para COVID-19
- tirar 6 ou outro valor no lançamento de um dado
- transmissão de dados com ou sem erro
- acertar ou errar um lance livre no basquete



Fonte: https://pt.wikipedia.org/wiki/Fam%C3%ADlia_Bernoulli

X ~ Bernoulli (p): $f(x) = p^x (1-p)^{1-x}, x = 0, 1$

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

Modelos discretos: Binomial

$X_1 = x_1 = 0$ ou 1 — X_1 ~ Bernoulli (p)

$X_2 = x_2 = 0$ ou 1 — X_2 ~ Bernoulli (p)

$X_3 = x_3 = 0$ ou 1 — X_3 ~ Bernoulli (p)

$X_4 = x_4 = 0$ ou 1

...

$X_n = x_n = 0$ ou 1 — X_n ~ Bernoulli (p)

independentes

$$Y = \sum_{i=1}^{n} X_i \sim \text{Binomial(n,p)}$$

$$f(y) = C_y^n p^y (1-p)^{n-y},$$
$$y = 0, 1, \cdots, n$$

Y: nº de sucessos em n repetições independentes do experimento Bernoulli (p)

Modelos discretos

Problema: contratar um jogador para uma posição num time de basquete; bom potencial de arremesso

Tradução: alta probabilidade de acertar um arremesso *(prob p de sucesso)*

X_{Lo} ~ Bernoulli (p_{Lo})

Y_{Lo} ~ Binomial (n=5,p_{Lo})

Y_{MJ} ~ Binomial (n=5,p_{MJ})

n=10, 20?

Y_{KB} ~ Binomial (n=5,p_{KB})

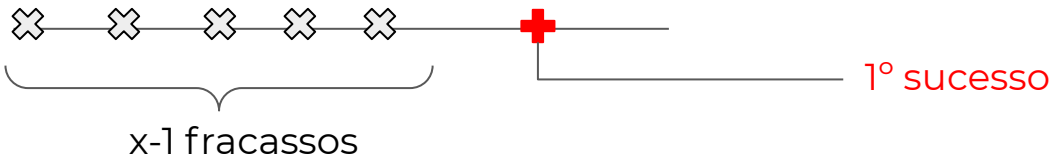
n=100?

Modelos discretos: Geométrica

X: nº de repetições de Bernoulli's (p, indep.) até a ocorrência do 1º sucesso

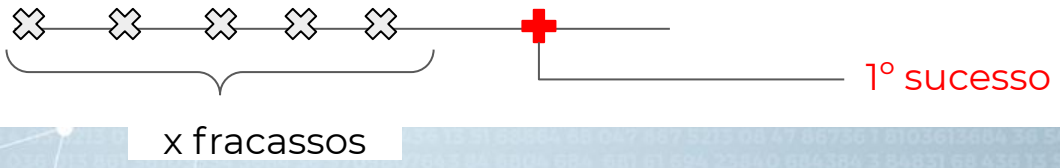
X ~ Geométrica(p)

$$f(x) = p(1-p)^{x-1}, x = 1, 2, \cdots$$



<u>Outra definição</u>: nº de repetições que antecedem o 1º sucesso

$$f(x) = p(1-p)^{x}, x = 0, 1, \cdots$$

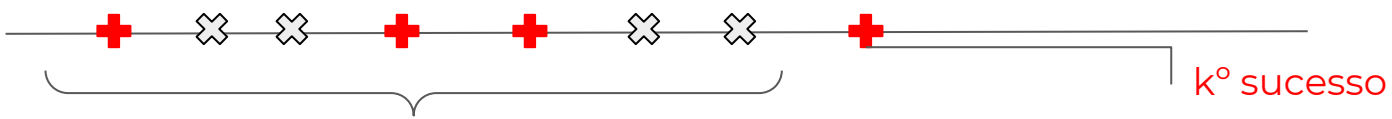


Modelos discretos: Binomial Negativa

X: n° de repetições de Bernoulli's (p, indep.) até a ocorrência do k° sucesso, k≥1

X ~ Binomial Negativa(k,p)

$$f(x) = C_{k-1}^{x-1} p^k (1-p)^{x-k}, x = k, k+1, \cdots$$



x-1 experimentos,

sendo k-1 sucessos

Modelos discretos: exemplo 1

Suponha que seu filho adora jogar basquete e que erra 3 arremessos a cada 10. Como ele sempre pede para ficar mais um pouco jogando antes de ir embora, você pensa responder sempre da mesma forma, para ser consistente. Entre as duas opções seguintes:

1) Mais 5 lances livres e vamos embora
2) Apenas lances livres e vamos embora quando você errar

Qual é a estratégia que permite que ele jogue mais, em média?

Modelos discretos: exemplo 1

1) 5 vezes. ←

2) X : nº de arremessos até o 1º IR fora.

$X \sim$ Geométrica $\left(p = \frac{3}{10} = 0{,}3\right)$

$$E(X) = \frac{1}{0{,}3} = 3{,}3 //$$

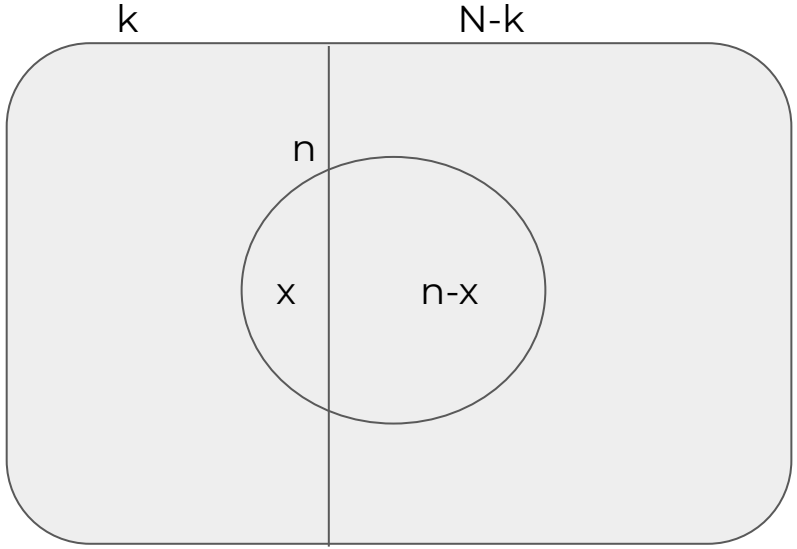
Modelos discretos: Hipergeométrica

X nº de sucessos em uma amostra de tamanho n (sem reposição) de uma **população finita**, de tamanho N, que contém k sucessos (k,n≤N).

X ~ Hipergeométrica(N,n,k)

$$f(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},$$

$$max\{0, n-(N-k)\} \leq x \leq min\{n,k\}$$



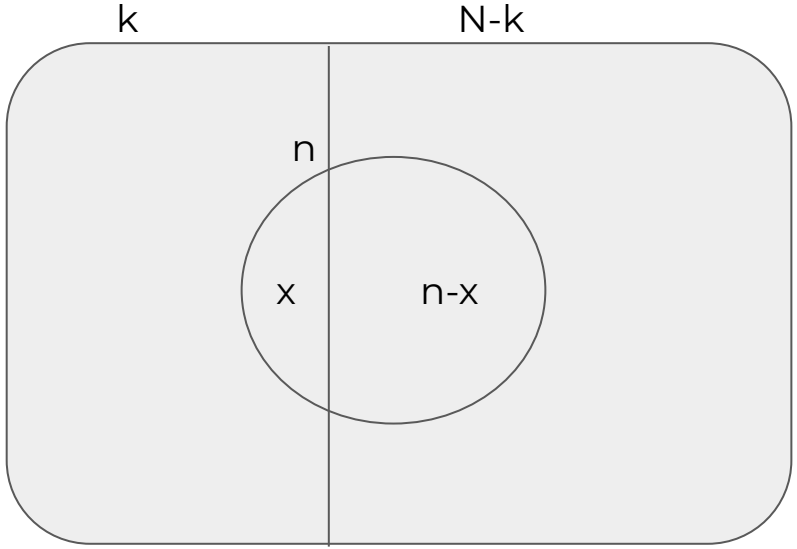
Modelos discretos: Hipergeométrica

X nº de sucessos em uma amostra de tamanho n (sem reposição) de uma **população finita**, de tamanho N, que contém k sucessos (k,n≤N).

X ~ Hipergeométrica(N,n,k)

$$f(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},$$

$$max\{0, n-(N-k)\} \leq x \leq min\{n,k\}$$



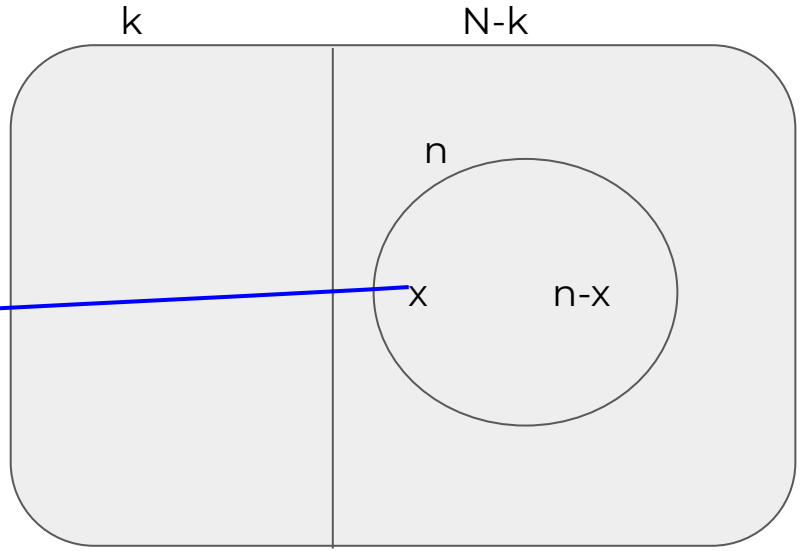
Modelos discretos: Hipergeométrica

X nº de sucessos em uma amostra de tamanho n (sem reposição) de uma **população finita**, de tamanho N, que contém k sucessos (k,n≤N).

X ~ Hipergeométrica(N,n,k)

$$f(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},$$

$$max\{0, n-(N-k)\} \leq x \leq min\{n, k\}$$



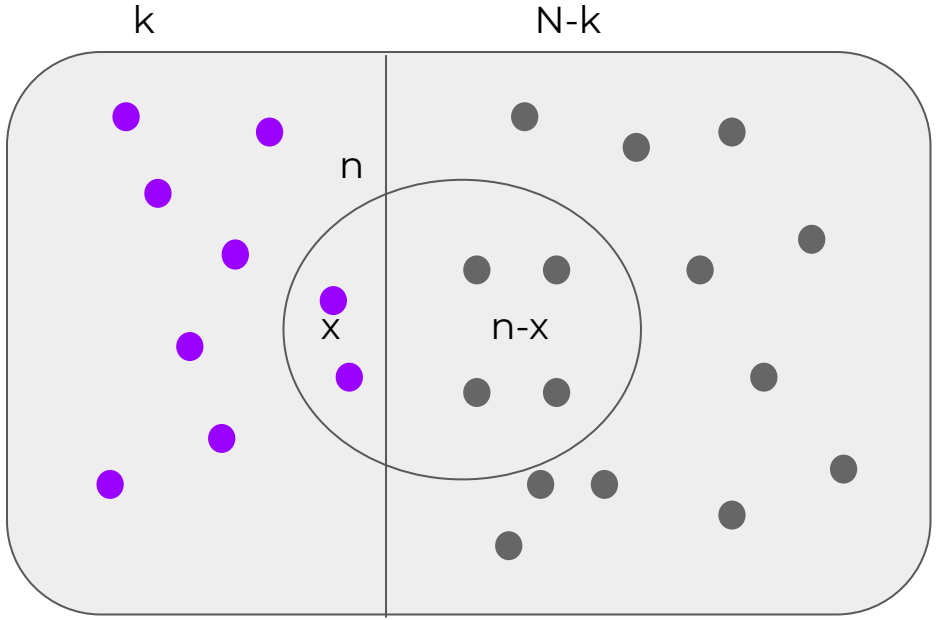
Modelos discretos: Hipergeimétrica

Modelo	$f(x)$	Suporte	parâmetros	$E(X)$	V(X)
Bernoulli	$p^x(1-p)^{1-x}$	$x = 0, 1$	p	p	$1-p$
Binomial	$C_x^n p^x (1-p)^{n-x}$	$x = 0, 1, \cdots, n$	n, p	np	$np(1-p)$
Geométrica	$p(1-p)^{x-1}$	$x = 1, 2, \cdots$	p	$1/p$	$(1-p)/p^2$
Binomial Negativa	$C_{k-1}^{x-1} p^k (1-p)^{x-k}$	$x = k, k+1, \cdots$	k, p	k/p	$k(1-p)/p^2$
Hipergeométrica	$\frac{C_x^k C_{n-x}^{N-k}}{C_n^N}$	$max\{0, n-(N-k)\} \leq x \leq min\{n, k\}$	N, n, k	nk/N	$\frac{N-n}{N-1} n \frac{k}{N}(1-\frac{k}{N})$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	$x = 0, 1, \cdots$	λ	λ	λ

fator de correção para população finita

Modelos discretos: exemplo 2

Processo de captura e recaptura para estimar tamanho populacional



Captura
k animais marcados

Recaptura
n animais
x são marcados
Qual o N?

https://www.tamar.org.br/noticia1.php?cod=830

Modelos discretos: Poisson

Exemplo: indústria de peças automobilísticas

- fabrica *n* peças por dia
- probabilidade *p* da fabricação gerar uma peça defeituosa
- X: número de peças fabricadas com defeito no dia
- Se *p* é constante e as peças são com ou sem defeito de forma independente, então:

X ~ Binomial (*n*,*p*) e E(X) = np = λ

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \frac{\lambda}{n}^x (1-\frac{\lambda}{n})^{n-x}$$

Modelos discretos: Poisson

Se $n \uparrow$ e $p \downarrow$ tal que E(X) = np = λ se mantém constante, então:

$$lim_{n\to\infty} f(x) = lim_{n\to\infty} \binom{n}{x} \frac{\lambda}{n}^{x} (1 - \frac{\lambda}{n})^{n-x} = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0, 1, \cdots$$

$\therefore$ X ~ Poisson(λ)

Processo de Poisson

Seja X o número de fissuras em um fio de cobre de 1m de comprimento, com um número médio de fissuras igual a λ:

- particionando o comprimento do fio em ($n, \uparrow$) subintervalos bem pequenos, t.q.
- a probabilidade de um subintervalo ter mais de uma fissura é 0 (desprezível),
- os subintervalos têm mesma probabilidade, $p=\lambda/n$, ($\downarrow$) de apresentar uma fissura, proporcional ao comprimento do subintervalo, e
- os subintervalos apresentam ou não uma fissura de forma independente,

então X ~ Poisson(λ)

Modelos discretos: exemplo 3

A avaliação final de um curso à distância consta de uma prova com 10 questões de múltipla escolha, cada uma com 5 alternativas de resposta. Aprovação no curso requer pelo menos 6 questões corretas.

a) Se um aluno responde a todas as questões baseado em palpite (“chute”), qual a probabilidade de ser aprovado?

b) O curso, a cada ano de oferecimento, tem 200 alunos matriculados. Qual é o número médio de alunos sem nenhum conhecimento que são aprovados no curso? Use a aproximação pela Poisson.

c) Qual é a probabilidade de que esse curso tenha no máximo 2 alunos sem nenhum conhecimento aprovados em dois anos de seu oferecimento?

Modelos discretos: exemplo 3

(a)

X : nº de questões corretas na prova

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - \sum_{x=0}^{5} \binom{10}{x} 0{,}2^{x} \cdot (1-0{,}2)^{10-x} = 1 - 0{,}994 = 0{,}006$$

$P(X \leq 5)$ ↓

$X \sim \text{Binomial}(n=10, 0.2)$

(b) $E(X) = \underbrace{np}_{\lambda} = 200 \cdot 0{,}006 = 1{,}2$

(c) Y : nº alunos s/ conhecimento aprovados no curso

$Y \sim \text{Poisson}\left(\lambda = np = \frac{1{,}2}{\text{10 alunos}}\right)$

$\lambda^{*} = 1{,}2 \cdot 2 = 2{,}4$ / 20 alunos

$$P(Y \leq 2) = \sum_{y=0}^{2} \frac{e^{-2{,}4} \cdot 2{,}4^{y}}{y!} = 0{,}5697.$$

Modelos contínuos: Exponencial

X: distância entre dois eventos sucessivos em um processo de poisson com média λ (por unidade de medida)

X ~ Exponencial(λ)

Justificativa: (do exemplo anterior)

N seja o número de fissuras em x metros de um fio de cobre, sendo λ a média de fissuras por metro.

N ~ Poisson(λx)

$$P(X > x) = P(N = 0) = \frac{e^{-\lambda x}\lambda x^0}{0!} = e^{-\lambda x} \Longrightarrow F(X) = 1 - P(X > x) = 1 - e^{-\lambda x}, x \geq 0$$

Modelos contínuos: Exponencial

X ~ Exponencial(λ)

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

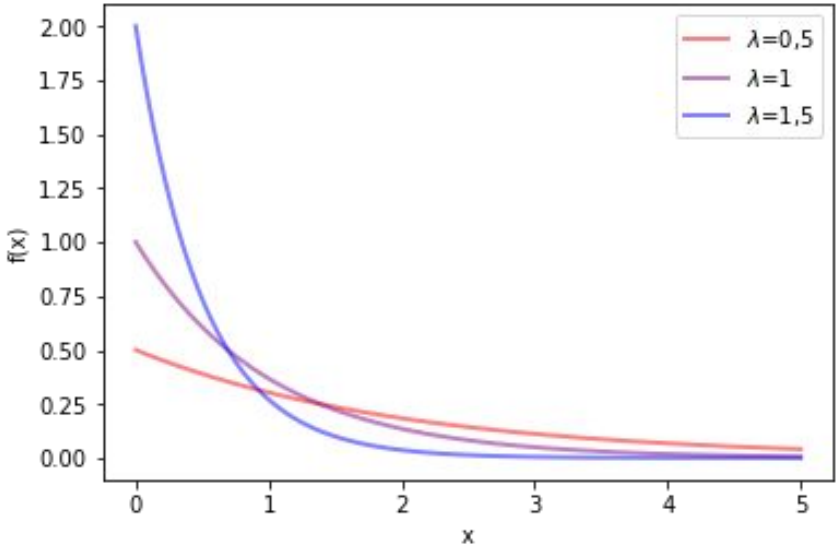
$$E(X) = 1/\lambda$$

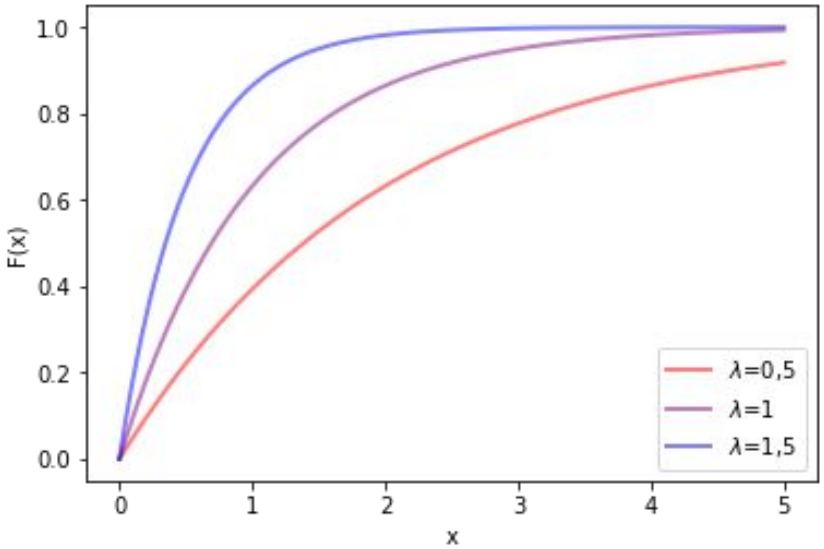
$$V(X) = 1/\lambda^2$$

Notação alternativa:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, x \geq 0$$

$$E(X) = \beta$$





Modelos contínuos: Exponencial

Propriedade:

Falta de memória

Se X ~ Exponencial(λ)

$P(X < t_1+t_2 \mid X > t_1) = P(X < t_2)$

ou

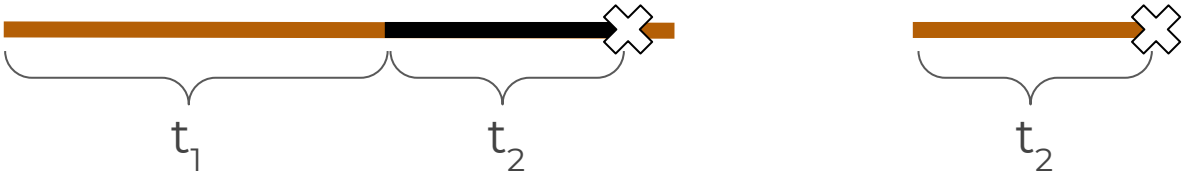
$P(X > t_1+t_2 \mid X > t_1) = P(X > t_2)$

Demonstração:

$$P(X < t_1 + t_2 \mid X > t_1) = \frac{P(t_1<X<t_1+t_2)}{P(X>t_1)}$$

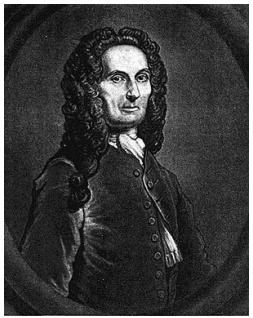
$$= \frac{F(t_1+t_2)-F(t_1)}{1-F(t_1)} = \frac{(1-e^{-\lambda(t_1+t_2)})-(1-e^{-\lambda t_1})}{e^{-\lambda t_1}}$$

$$= \frac{e^{-\lambda t_1}(1-e^{-\lambda t_2})}{e^{-\lambda t_1}} = 1 - e^{-\lambda t_2} = F(t_2)$$



Modelos contínuos: Normal

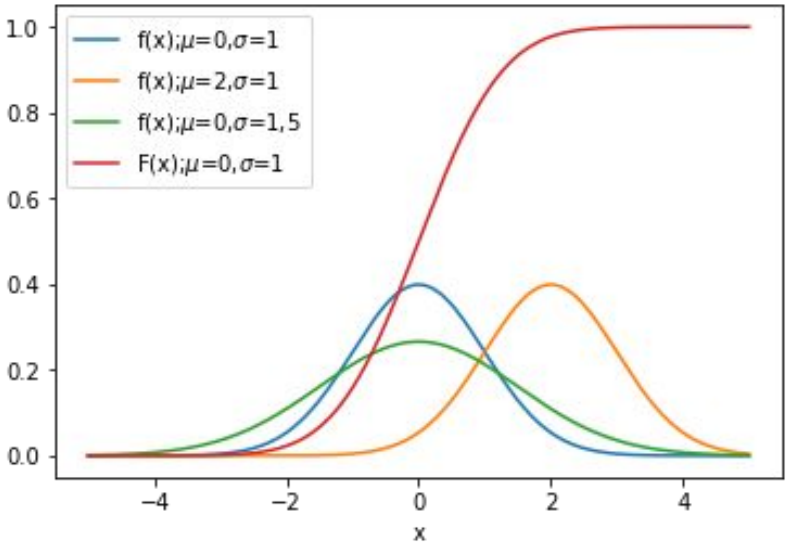
Modelo probabilístico mais usual na prática (distribuição normal ou gaussiana)



A. de Moivre
(1667-1754)

P.S. La Place
(1749-1827)

C.F. Gauss
(1777-1855)



X ~ Normal(μ; σ²), $\mu \in \Re, \sigma^2 > 0$

$E(X) = \mu$

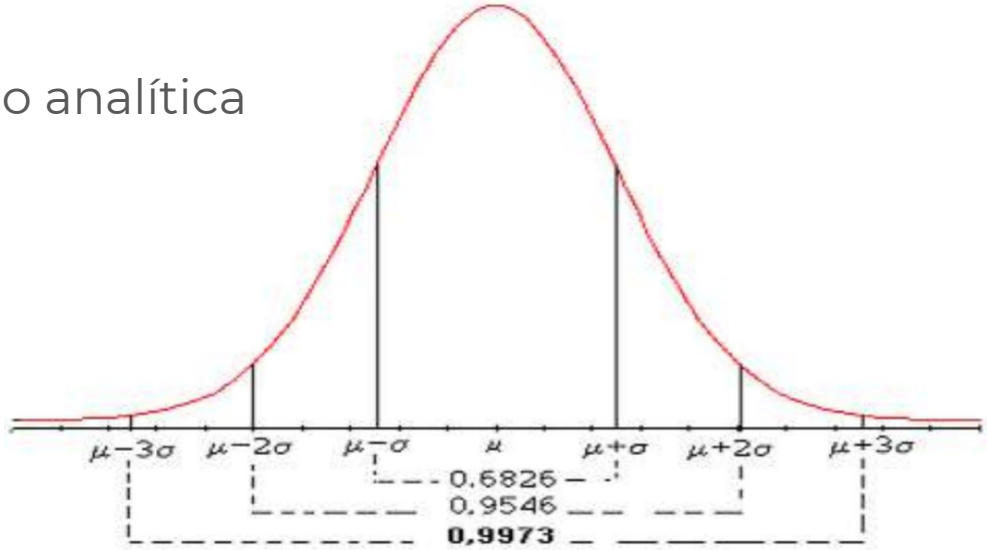
$V(X) = \sigma^2$

$-\infty < x < \infty$: $f(x) = \sqrt{\frac{1}{2\pi\sigma^2}} exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$

Modelos contínuos: Normal

Propriedades:

- $F(x) = \int_{-\infty}^{x} f(t)dt$: não tem solução analítica
- mediana = moda = média (μ)
- simétrica em torno de μ
- P(μ - σ < X < μ + σ) = 0,683
- P(μ - 2σ < X < μ + 2σ) = 0,955
- P(μ - 3σ < X < μ + 3σ) = 0,997



- $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$: normal padrão ou reduzida (transf. linear de X)
- $X_1, X_2, ..., X_n$ independentes e Normais, então ΣX_i também é Normal

Modelos contínuos: exemplo 1

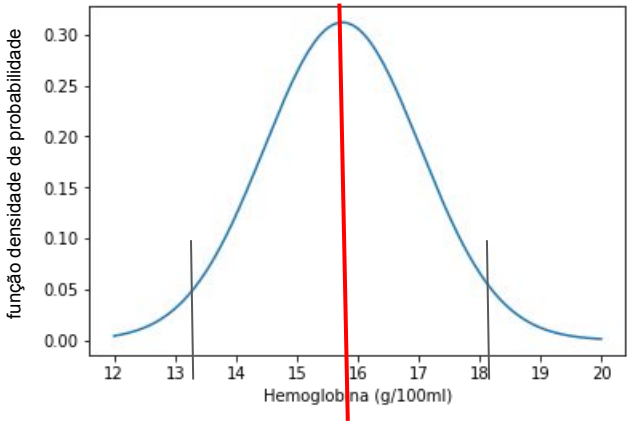
Suponha que o intervalo de referência para a quantidade de hemoglobina (g/100mL) no sangue de homens adultos é construindo supondo distribuição Normal para tal variável, de forma simétrica, de tal forma que 95% de indivíduos saudáveis estejam dentro desse intervalo.

Qual a média e o desvio padrão da quantidade de hemoglobina, se o intervalo é dado por [13,2; 18,3]?

Para compensar o pouco oxigênio, a concentração de hemoglobina costuma ser mais elevada em pessoas que vivem em regiões altas. Quantos % de homens saudáveis espera-se encontrar com níveis acima de 19 g/100mL?

Modelos contínuos: exemplo 1

Valores de referência para hemoglobina



$$f(x) = \frac{1}{\sqrt{2\pi 1{,}28^2}} exp\{-\frac{(x-15{,}75)^2}{2.1{,}28^2}\}dx$$

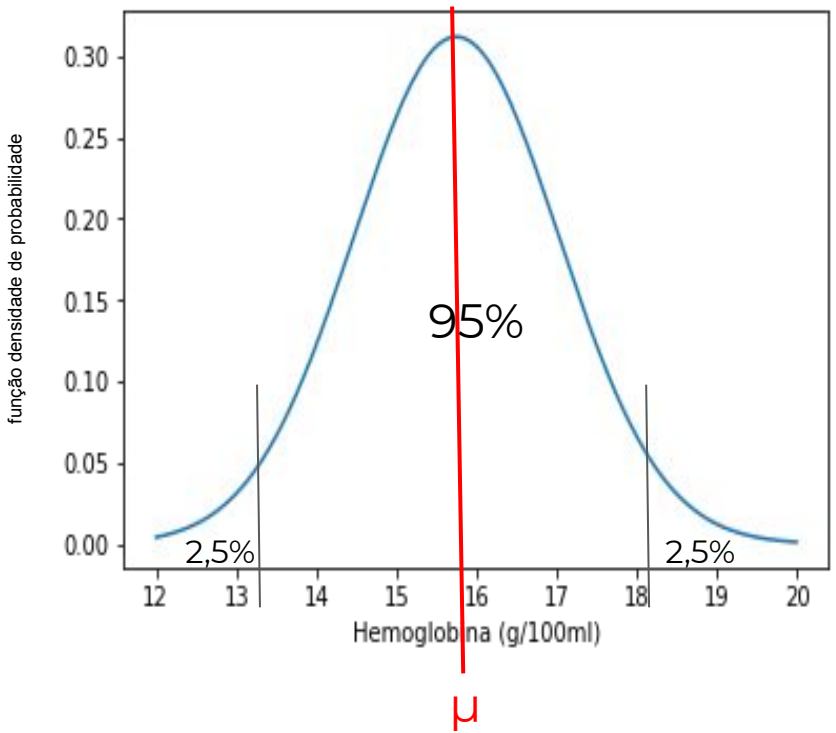
$$P(13,2 \leq X \leq 18,3) = \int_{13,2}^{18,3} \frac{1}{\sqrt{2\pi 1{,}28^2}} exp\{-\frac{(x-15{,}75)^2}{8^2}\}dx$$

≃ 95%

$$E(X) = \int_{-\infty}^{\infty} x. \frac{1}{\sqrt{2\pi 1{,}28^2}} exp\{-\frac{(x-15{,}75)^2}{2.1{,}28^2}\}dx = 15{,}75$$

Intervalo de normalidade: [13,2; 18,3]

Modelos contínuos: exemplo 1



$18,3 - \mu = \mu - 13,2$

$\mu = 15,75$

$F_X(18,3) - F_X(13,2) = 0,95$

$F_X(18,3) = 0,975$

$F_Z(1,96) = 0,975$

Pela relação linear com Z ~N(0,1):

$z = (X-\mu)/\sigma$

$\sigma = (18,3 - 15,75)/1,96 = 1,3$

$P(X > 19) = 0,006 = 0,6\%$

Resultados assintóticos: n → ∞

$X_1, X_2, ..., X_n$ variáveis aleatórias independentes com mesma distribuição de probabilidades (inclusive os parâmetros dessa distribuição),

$E(X_i) = \mu < \infty$, i = 1, 2, ... n

$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

para todo $\varepsilon > 0$

Lei Fraca dos Grandes Números

$P(|\bar{X} - \mu| > \epsilon) \to 0$, quando $n \to \infty$

$\bar{X}$ converge para μ

Lei Forte dos Grandes Números

Com prob. 1, $\bar{X} \to \mu$, quando $n \to \infty$

Resultados assintóticos: Teorema Central do Limite (TCL)

$X_1, X_2, ..., X_n$ variáveis aleatórias independentes com mesma distribuição de probabilidades (inclusive os parâmetros dessa distribuição),

$V(X_i) = \sigma^2 < \infty$, para i = 1, 2, ... n

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

n ↑ , melhor a aproximação
X pode ser discreta ou contínua

$$\frac{\sum X_i - n\mu}{\sqrt{n}\sigma} \to N(0,1)$$

normal padrão

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \longrightarrow N(0,1), \text{ quando } n \to \infty$$

Repare que $E(\bar{X}) = E(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$

e, pela independência: $V(\bar{X}) = V(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$

Resultados assintóticos: exemplo TCL

X: tempo de execução de um serviço de assistência técnica oferecido por uma empresa

X ~ Exponencial(λ=¼=0,25), com média de 4h por serviço

(a) Qual a probabilidade do tempo de execução de um serviço ultrapassar 5h?

$$P(X>5)=1-0,7135 = 0,2865$$

(b) Se a empresa realiza 150 serviços de assistência técnica ao mês, qual a probabilidade do tempo médio de execução desses serviços ultrapassar 5h?

Pelo TCL:

$$P(\bar{X} > 5) \approx P(Z > \frac{5-4}{\sqrt{16/150}}) = P(Z > 3,06) = 0,0011$$