Introdução à Estatística usando o R: Seja bem-vind@ ao tidyverse

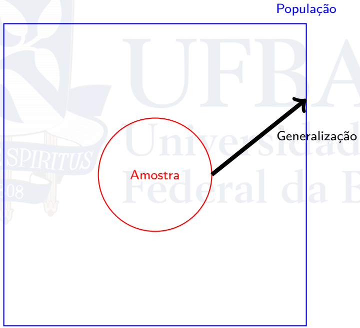
Profa Carolina & Prof Gilberto

Instituto de Matemática e Estatística
Universidade Federal da Bahia

Probabilidade: Motivação

Com estatística descritiva podemos fazer afirmações válidas para amostra, mas queremos fazer afirmações válidas para toda a população. Com esse objetivo, vamos usar inferência estatística (ou estatística inferencial) para fazer generalizações da amostra para a população, conforme ilustrado na Figura 1. As técnicas de inferência estatística, usam probabilidade para fazer as generalizações como apresentado a seguir.

Figura 1: Ilustração da estatística inferencial.



O que faremos nesse curso?

- Estimação pontual: Aproximar um parâmetro.

  Exemplo: Estimar o teor alcóolico de uma bebida.

- Intervalo de confiança: Encontrar uma estimativa intervalar para um parâmetro.

  Exemplo: Encontrar números a e b tal que o teor alcóolico verdadeiro está entre a e b com uma probabilidade estabelecida pelo pesquisador.

- Teste de hipóteses: Decidir entre duas hipóteses H_0 e H_1 : negação de H_0.

  Exemplo: Decidir entre duas hipóteses:

$$H_0 : \text{O teor alcóolico da bebida é } 10\%,$$
$$H_1 : \text{O teor alcóolico da bebida não é } 10\%.$$

Em todos esses casos, precisamos usar probabilidade.

Probabilidade

Fenômeno Aleatório

Procedimento ou evento cujo resultado não é possível antecipar de forma determinística. Por exemplo:

- Teremos uma guerra total na Venezuela envolvendo o Brasil, Colômbia e Estados Unidos da América?
- Qual o resultado do lançamento de um dado "justo"?

Notação e nomes

- **Espaço amostral:** O conjunto de todos os resultados de um fenômeno aleatório.
  Notação: Ω
- **Evento:** Subconjunto de um espaço amostral.
  Notação: $A, B, C, \cdots$.
- **Ponto amostral:** Um resulto possível de um fenómeno aleatório.
  Notação: ω.
- **Probabilidade:** A plausibilidade de um ponto amostral ω de A ser o resultado do fenômenos aleatório.
  Notação: $P(A)$.

Classificação de variáveis aleatórias

- Dizemos que X é uma variável aleatória discreta, se os valores possíveis desta variável são números inteiros, geralmente resultado de contagem;
- Dizemos que X é uma variável aleatória contínua, se os valores possíveis desta variável pode ser qualquer número (incluindo aqueles por parte decimal);
- O conjunto dos valores possíveis de X representamos por χ.

Variável aleatória discreta

Seja X uma variável aleatória discreta. Então, podemos definir

- **Função de probabilidade (FP):**
$$f(x) = P(X = x)$$
- **Função de distribuição acumulada (FDA):**
$$F(x) = f(x_1) + \cdots + f(x_k),$$
em que $x_k \leq x$ e $x_{k+1} > x$.

Medidas de resumo para variável aleatória discreta

Seja X uma variável aleatória discreta com suporte $\chi = \{x_1, \ldots, x_n\}$ e função de probabilidade $f(x)$. Então

- **Média:** $\mathrm{E}(X) = x_1 \cdot f(x_1) + \cdots + x_n \cdot f(x_n) = \mu$;

- **Variância:** $\mathrm{Var}(X) = (\mu - x_1)^2 f(x_1) + \cdots + (\mu - x_n)^2 f(x_n)$;

- **Mediana:** Md é um número satisfazendo $P(X \leq Md) \geq 0,5$ e $P(X \geq Md) \geq 0,5$;

- **Desvio Padrão:** $\mathrm{DP}(X) = \sqrt{\mathrm{Var}(X)}$.

Distribuição Bernoulli

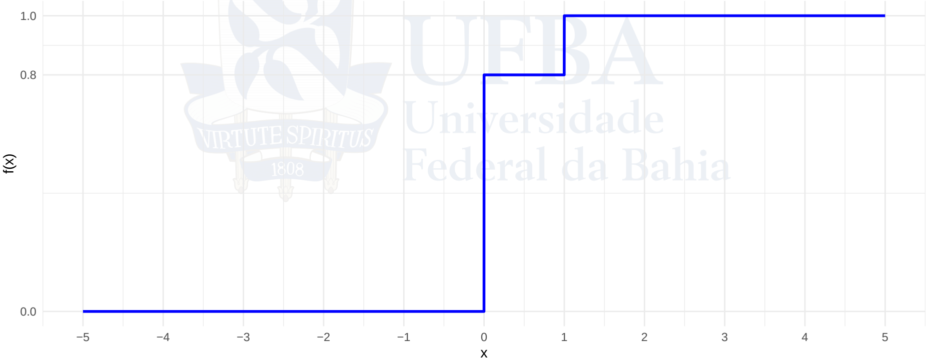
- Cada elemento da população pode ser **sucesso** ou **fracasso**;
- $P(\text{sucesso}) = p$ e $P(\text{fracasso}) = 1 - p$;
- $X(\omega) = \begin{cases} 1, & \text{se } \omega \text{ é sucesso} \\ 0 & \text{se } \omega \text{ é fracasso} \end{cases}$;
- Valores possíveis de X: $\chi = \{0, 1\}$;
- **Função de probabilidade:** $f(0) = 1 - p, f(1) = p$;
- **Função de distribuição acumulada:** $F(x) = \begin{cases} 0, & \text{se } x < 0 \\ 1 - p, & \text{se } 0 \leq x < 1 \\ 1, & \text{se } x \geq 1 \end{cases}$;
- $\mathrm{E}(X) = n \cdot p$;
- $\mathrm{Var}(X) = n \cdot p \cdot (1 - p)$;
- $X \sim \mathit{Bernoulli}(p)$.

Distribuição Bernoulli – função de probabilidade

```r
# gráfico da função de probabilidade
p <- 0.3 # probabilidade de sucesso
x <- c(0,1)
y <- dbinom(x, 1, p)
tibble(x = x, `f(x)`=y) %>%
  ggplot(aes(x, `f(x)`)) +
  geom_point(colour = 'blue', size=4)+
  scale_x_continuous(breaks = c(0,1)) +
  scale_y_continuous(breaks = c(1-p,p), limits = c(0,1)) +
  labs(y = 'f(x)') + theme_minimal()
```

Distribuição Bernouli – função de distribuição acumulada

```r
# Função de distribuição acumulada
p <- 0.2
x <- seq(from = -5, to = 5, by = 0.001)
y <- pbinom(x, 1, p)
# gráfico -- FDA
tibble(x=x, `f(x)`=y) %>% ggplot()+
  geom_line(aes(x, `f(x)`), color = 'blue', size = 1) +
  scale_x_continuous(breaks = seq(from = -5, to = 5, by = 1)) +
  scale_y_continuous(breaks = c(0, 1-p, 1)) +
  theme_minimal()
```



Distribuição Bernoulli: simulando uma amostra

```
# Variável Bernoulli: X ~ Bernoulli(p)
p <- 0.3
# Função densidade
(dbinom(c(0,1),1,p))
```

```
## [1] 0.7 0.3
```

```
# simular valores da variável Bernoulli
n <- 1000 # tamanho da amostra
amostra <- rbinom(n, 1, p)
tibble(x = amostra) %>%
  summarise(media = mean(x), mediana = median(x),
            dp = sd(x), cv = sd(x) * 100 / mean(x),
            q1 = quantile(x, probs = 0.25),
            q3 = quantile(x, probs= 0.75))
```

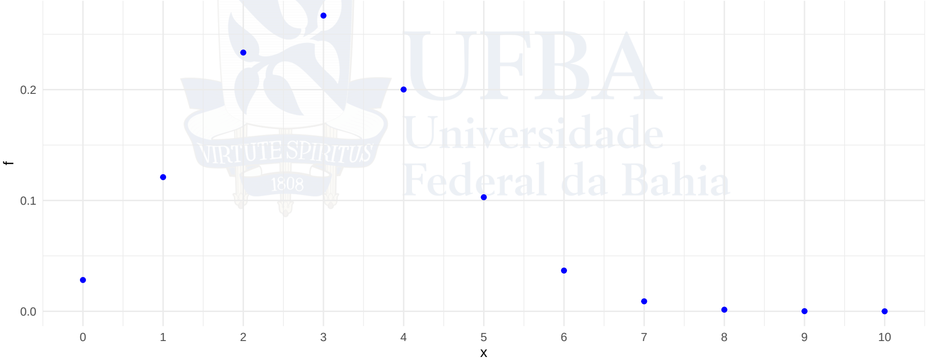
```
## # A tibble: 1 x 6
##   media mediana    dp    cv    q1    q3
##   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.290       0 0.454  157.     0     1
```

Distribuição Binomial

- Temos n casos em que cada caso pode ser **sucesso** ou **fracasso**;
- $P(\text{sucesso}) = p$ e $P(\text{fracasso}) = 1 - p$;
- X : número de sucessos em n casos;
- Valores possíveis de X: $\chi = \{0, 1, 2, \cdots, n\}$;
- **Função de probabilidade:** $f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \forall x \in \chi$;
- **Função de distribuição acumulada:** $F(x) = f(x_1) + \cdots + f(x_k)$, em que $x \leq x_k$ e $x_{k+1} > x$;
- $\mathrm{E}(X) = n \cdot p$;
- $\mathrm{Var}(X) = n \cdot p \cdot (1-p)$;
- $X \sim b(n, p)$.

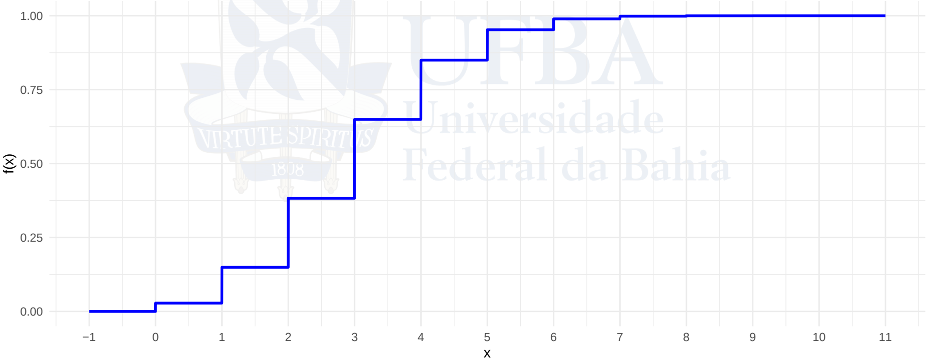
Distribuição binomial – função de probabilidade

```r
# gráfico da função de probabilidade
n <- 10
x <- 0:n
p <- 0.3
f <- dbinom(x, n,  p)
tibble(x=x, f = f) %>%
  ggplot()+
  geom_point(aes(x=x, y=f), color = 'blue') +
  scale_x_continuous(breaks = 0:10) + theme_minimal()
```



Distribuição binomial – função de distribuição acumulada

```
# gráfico da função de distribuição acumulada
n <- 10
p <- 0.3
x <- seq(from = -1, to = 11, by = 0.001)
y <- pbinom(x, n, p)
tibble(x = x, `f(x)`=y) %>%
  ggplot()+
  geom_line(aes(x, `f(x)`), stat = 'identity', size = 1, color = 'blue')+
  scale_x_continuous(breaks = seq(from = -1, to = 11, by = 1)) +
  theme_minimal()
```



Distribuição binomial: simulando uma amostra

```
# Amostra da distribuição binomial
m <- 100 # tamanho da amostra
n <- 10 # número de casos
p <- 0.3
amostra <- rbinom(m,n,p)
tibble(x = amostra) %>%
  summarise(media = mean(x), mediana = median(x), Var = var(x),
            dp = sd(x), cv = sd(x) * 100 / mean(x),
            q1 = quantile(x, probs = 0.25),
            q3 = quantile(x, probs = 0.75))
```

```
## # A tibble: 1 x 7
##   media mediana   Var    dp    cv    q1    q3
##   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  3.19       3  2.32  1.52  47.7     2     4
```

Distribuição Poison

- X: número de ocorrências num intervalo de tempo. Exemplo: número de partículas emitidas por um isótopo em um minuto; número de chamadas em um *call center* durante um período de 24 horas;

- λ é a média de ocorrência durante o intervalo de tempo;

- Valores possíveis de X: $\chi = \{0, 1, 2, 3, \dots, \}$;

- **Função de probabilidade:** $f(x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \forall x \in \chi$, em que
  $x! = \begin{cases} 1, & \text{se } x \in \{0, 1\} \\ 1 \cdot 2 \cdot \dots \cdot (x-1) \cdot x, & \text{se } x \in \{2, 3, 4, \dots\} \end{cases}$;

- **Função de distribuição acumulada:** $F(x) = f(x_1) + \dots + f(x_k)$, em que $x \leq x_k$ e $x_{k+1} > x$;

- $\text{E}(X) = n \cdot p$;

- $\text{Var}(X) = n \cdot p \cdot (1 - p)$;

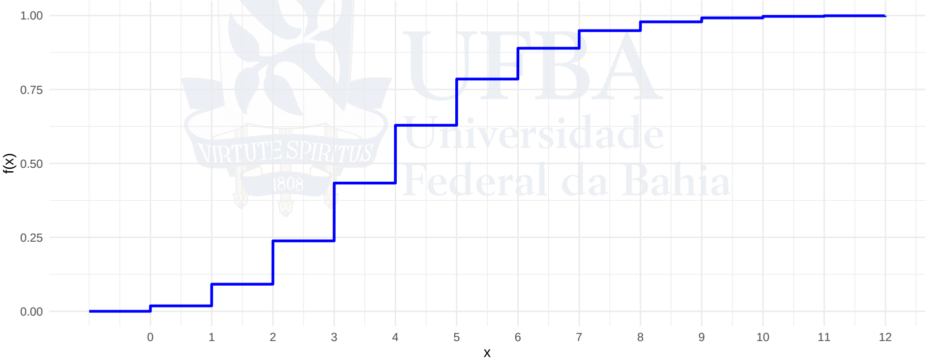
- $X \sim Poison(\lambda)$.

Distribuição Poison – função de probabilidade

```r
# gráfico da função de probabilidade
lambda <- 4 # média de ocorrência no intervalo de tempo
x <- 0:12
y <- dpois(x, lambda)
tibble(x = x, `f(x)`=y) %>%
  ggplot(aes(x, `f(x)`)) +
  geom_point(colour = 'blue', size=1)+
  scale_x_continuous(breaks = x) +
  labs(y = 'f(x)') + theme_minimal()
```

Distribuição Poison – função de distribuição de probabilidade

```r
# gráfico da função de distribuição acumulada
lambda <- 4 # média de ocorrência no intervalo de tempo
x <- seq(from = -1, to = 12, by = 0.001)
y <- ppois(x, lambda)
tibble(x = x, `f(x)`=y) %>%
  ggplot()+
  geom_line(aes(x, `f(x)`), stat = 'identity', size = 1, color = 'blue')+
  scale_x_continuous(breaks = 0:12) +
  theme_minimal()
```



Distribuição Poison: simulando uma amostra

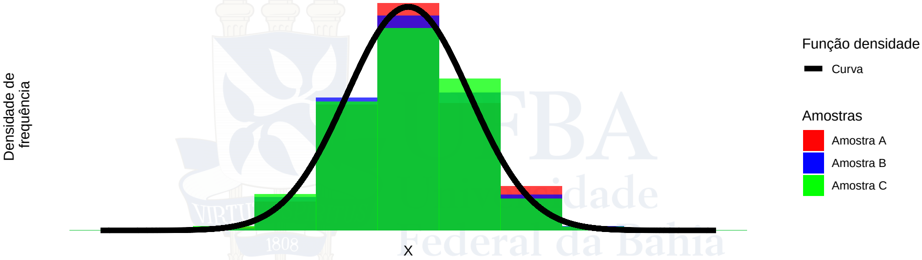
```
# Amostra da distribuição Poison
m <- 100 # tamanho da amostra
lambda <- 4 # média de ocorrência no intervalo de tempo
amostra <- rpois(m, lambda = lambda)
tibble(x = amostra) %>%
  summarise(media = mean(x), mediana = median(x), Var = var(x),
            dp = sd(x), cv = sd(x) * 100 / mean(x),
            q1 = quantile(x, probs = 0.25),
            q3 = quantile(x, probs = 0.75))
```

```
## # A tibble: 1 x 7
##   media mediana   Var    dp    cv    q1    q3
##   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  3.93       4  2.93  1.71  43.6     3     5
```

Variável aleatória contínua

Motivação:

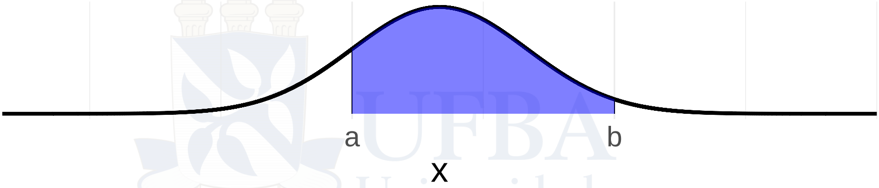
- Para cada amostra, temos um histograma;
- Queremos encontrar uma curva que aproxima bem todos os histogramas possíveis;



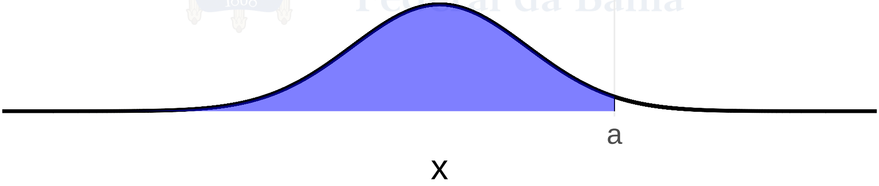
Chamamos a curva preta de função densidade.

Propriedades de uma variável aleatória contínua

- Suporte de uma variável aleatória: intervalo(s) de números reais;
- $P(X = a) = 0$;
- $P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b)$;
- **Notação:** probabilidade de X estar entre a e b – $P(a < X < b)$;



- **Notação:** probabilidade de X estar entre a e b – $P(a < X < b)$;



Distribuição normal

- Valores da variável aleatória concentrados em torno da média populacional μ;
- Valores da variável aleatória afastados da média populacional μ são pouco prováveis;
- Valores possíveis da variável: todos os números reais $\chi = \mathbb{R}$;
- Função densidade (fd): curva em formato de sino;
- μ é a média da população e σ^2 é a variância da população;
- Função de distribuição acumulada (fda): $F(x) = P(X \leq x)$;
- Usamos a notação: $X \sim N(\mu, \sigma^2)$;
- Seja $\Phi(x)$ a fda de uma variável $Z \sim N(0, 1)$, então se $X \sim N(\mu, \sigma^2)$ temos que

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Distribuição normal – continuação

- Se $X \sim N(\mu, \sigma^2)$, então

$$P(a < X < b) = F(b) - F(a)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

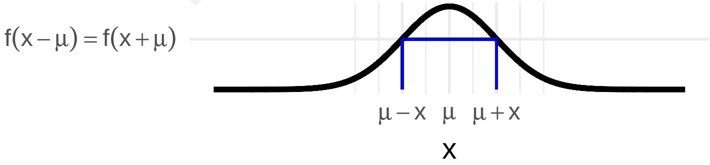
- Seja $\Phi(x)$ a fda de uma variável $Z \sim N(0,1)$, então

$$\Phi(a) = 1 - \Phi\left(|a|\right), \quad \text{se } a < 0$$

- Média, moda, mediana, variância para $X \sim N(\mu, \sigma^2)$:

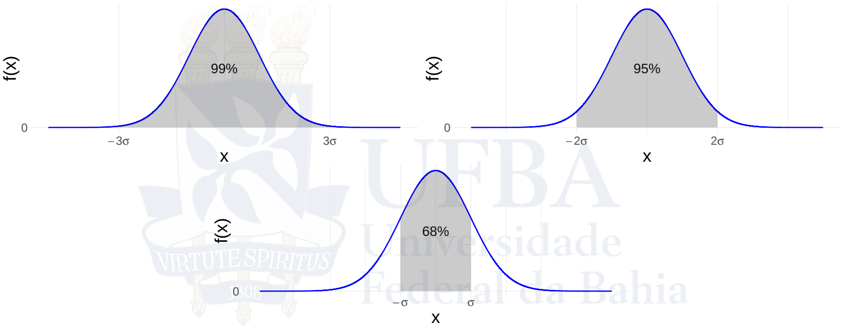
$$\mathrm{E}(X) = \mu, \quad Mo(X) = \mu, \quad Md(X) = \mu, \quad \mathrm{Var}(X) = \sigma^2.$$

- A função densidade é simétrica em torno de μ.
  - Se $X \sim N(\mu, \sigma^2)$, então $f(\mu - x) = f(x - \mu)$.



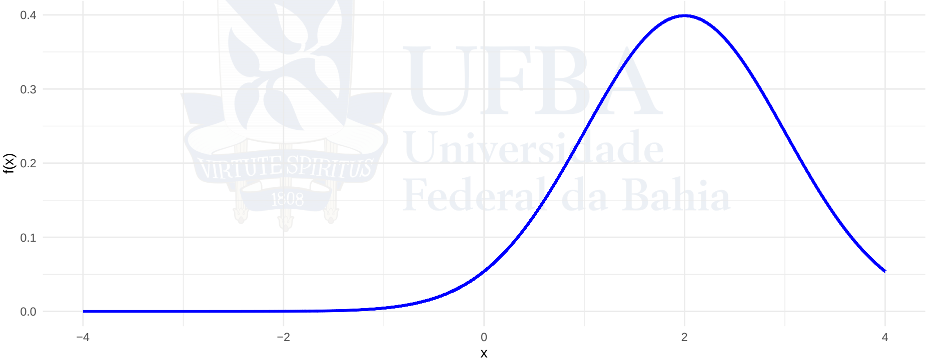
Distribuição normal – função densidade

Função densidade tem formato de sino.

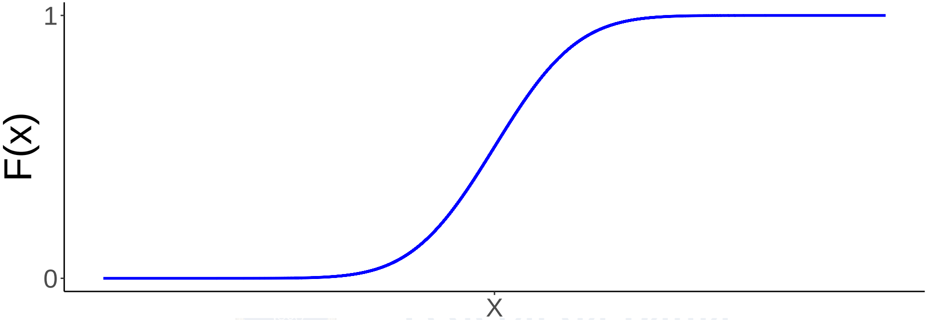


Distribuição normal – função densidade

```
media <- 2 # média populacional
s2 <- 1 # variância populacional
x <- seq(from = -4, to = 4, by = 0.0001)
y <- dnorm(x, mean = media, sd = sqrt(s2))
dados <- tibble::tibble(x, y)
ggplot(dados) +
  geom_line(aes(x, y), color = "blue", size = 1) +
  theme_minimal() + labs( y = "f(x)")
```

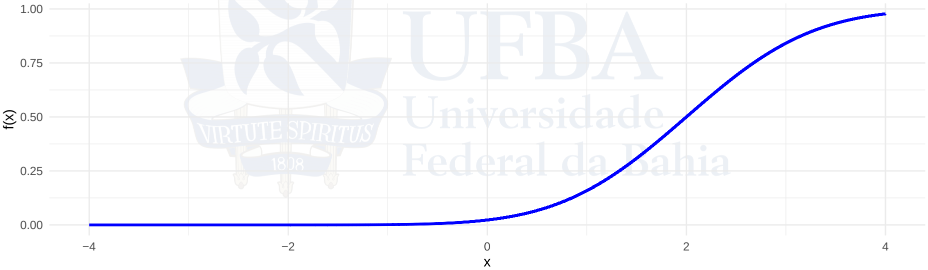


Distribuição normal – função de distribuição acumulada



Distribuição normal – função de distribuição acumulada

```r
media <- 2 # média populacional
s2 <- 1 # variância populacional
x <- seq(from = -4, to = 4, by = 0.0001)
y <- pnorm(x, mean = media, sd = sqrt(s2))
dados <- tibble(x, y)
ggplot(dados) +
  geom_line(aes(x, y), color = "blue", size = 1) +
  labs(y = "f(x)") + theme_minimal()
```



Distribuição normal – simulando uma amostra

```r
# Amostra da distribuição Poison
m <- 100 # tamanho da amostra
media <- 2 # média populacional
s2 <- 1 # variância populacional
amostra <- rnorm(m, mean = media, sd = sqrt(s2))
tibble(x = amostra) %>%
  summarise(media = mean(x), mediana = median(x), Var = var(x),
            dp = sd(x), cv = sd(x) * 100 / mean(x),
            q1 = quantile(x, probs = 0.25),
            q3 = quantile(x, probs = 0.75))
```

```
## # A tibble: 1 x 7
##   media mediana   Var    dp    cv    q1    q3
##   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  1.99    1.97 0.874 0.935  47.0  1.39  2.63
```

Distribuição Exponencial

- X: tempo até a ocorrência. Exemplo: tempo até a morte de um cidadão brasileiro; tempo de vida de um equipamento;

- μ: média de tempo até a ocorrência;

- $\alpha = \frac{1}{\mu}$; taxa de decaimento;

- A função densidade é uma curva em formato de decaimento exponencial, ou seja, tempo de ocorrência grandes são menos prováveis;

- Usamos a notação: $X \sim Exp(\alpha)$.

- Média, Moda, Mediana e Variância para $X \sim Exp(\alpha)$:

$$\mathrm{E}(X) = \mu = \frac{1}{\alpha}, \quad Mo(X) = 0, \quad Md(X) = \mu \ln(2) = \frac{\ln(2)}{\alpha}, \quad \mathrm{Var}(X) = \mu^2 = \frac{1}{\alpha^2};$$

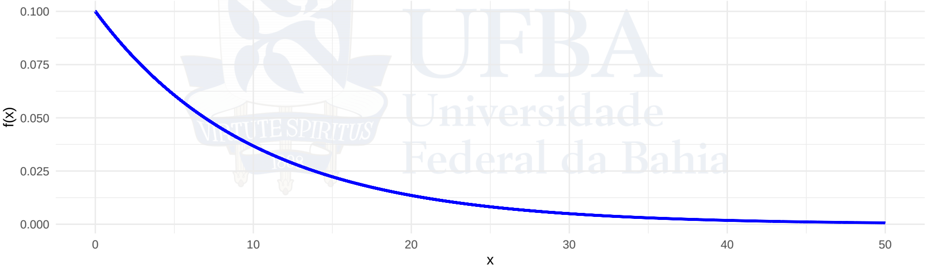
- $X \sim Exp(\alpha)$.

Distribuição Exponencial – função densidade

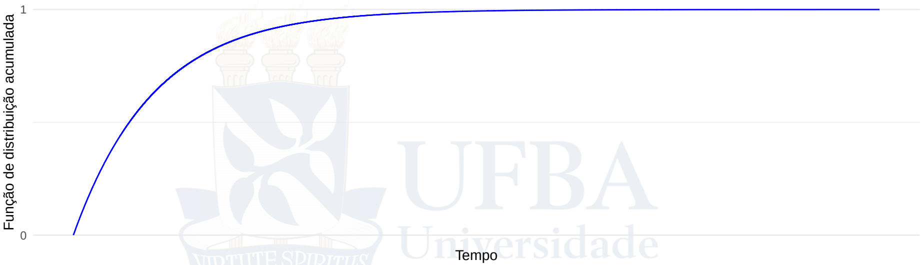
Função densidade tem formato de decaimento exponencial.

Distribuição Exponencial – função densidade

```r
media <- 10 # tempo médio para ocorrência
alpha <- 1 / media # taxa de decaimento
x <- seq(from = 0, to = 50, by = 0.0001)
y <- dexp(x, rate = alpha)
dados <- tibble(x, y)
ggplot(dados) +
  geom_line(aes(x, y), color = "blue", size = 1) +
  theme_minimal() + labs( y = "f(x)")
```

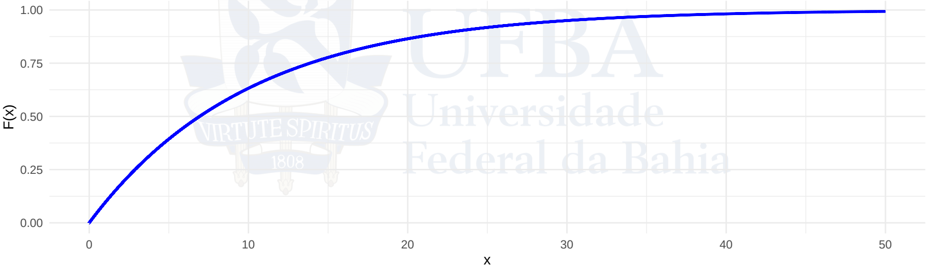


Distribuição Exponencial – função de distribuição acumulada



Distribuição Exponencial – função de distribuição acumulada

```r
media <- 10 # tempo médio para ocorrência
alpha <- 1 / media # taxa de decaimento
x <- seq(from = 0, to = 50, by = 0.0001)
y <- pexp(x, rate = alpha)
dados <- tibble(x, y)
ggplot(dados) +
  geom_line(aes(x, y), color = "blue", size = 1) +
  theme_minimal() + labs( y = "F(x)")
```



Distribuição Exponencial – simulando uma amostra

```
m <- 1000 # tamanho da amostra
media <- 10 # tempo médio para ocorrência
alpha <- 1 / media # taxa de decaimento
amostra <- rexp(m, rate = alpha)
tibble(x = amostra) %>%
  summarise(media = mean(x), mediana = median(x), Var = var(x),
            dp = sd(x), cv = sd(x) * 100 / mean(x),
            q1 = quantile(x, probs = 0.25),
            q3 = quantile(x, probs = 0.75))
```

```
## # A tibble: 1 x 7
##   media mediana   Var    dp    cv    q1    q3
##   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  10.4    7.60  98.4  9.92  95.5  3.20  13.8
```