Estatística descritiva usando R

bem-vinde ao tidyverse

Probabilidade usando R

Profa Carolina e Prof Gilberto

Parte 3

O que aprenderemos a seguir?

Estimação pontual: aproximação de parâmetro.
Exemplo: Estimar a nota média em matemática dos candidatos do ENEM na cidade de Salvador.

Estimação intervalar: estimativa intervalar para o parâmetro.
Exemplo: Encontrar a e b tal que a nota média de matemática esteja entre a e b com alguma probabilidade.

Teste de hipóteses: decisão entre duas hipóteses complementares.
Exemplo: Decidir entre duas hipóteses

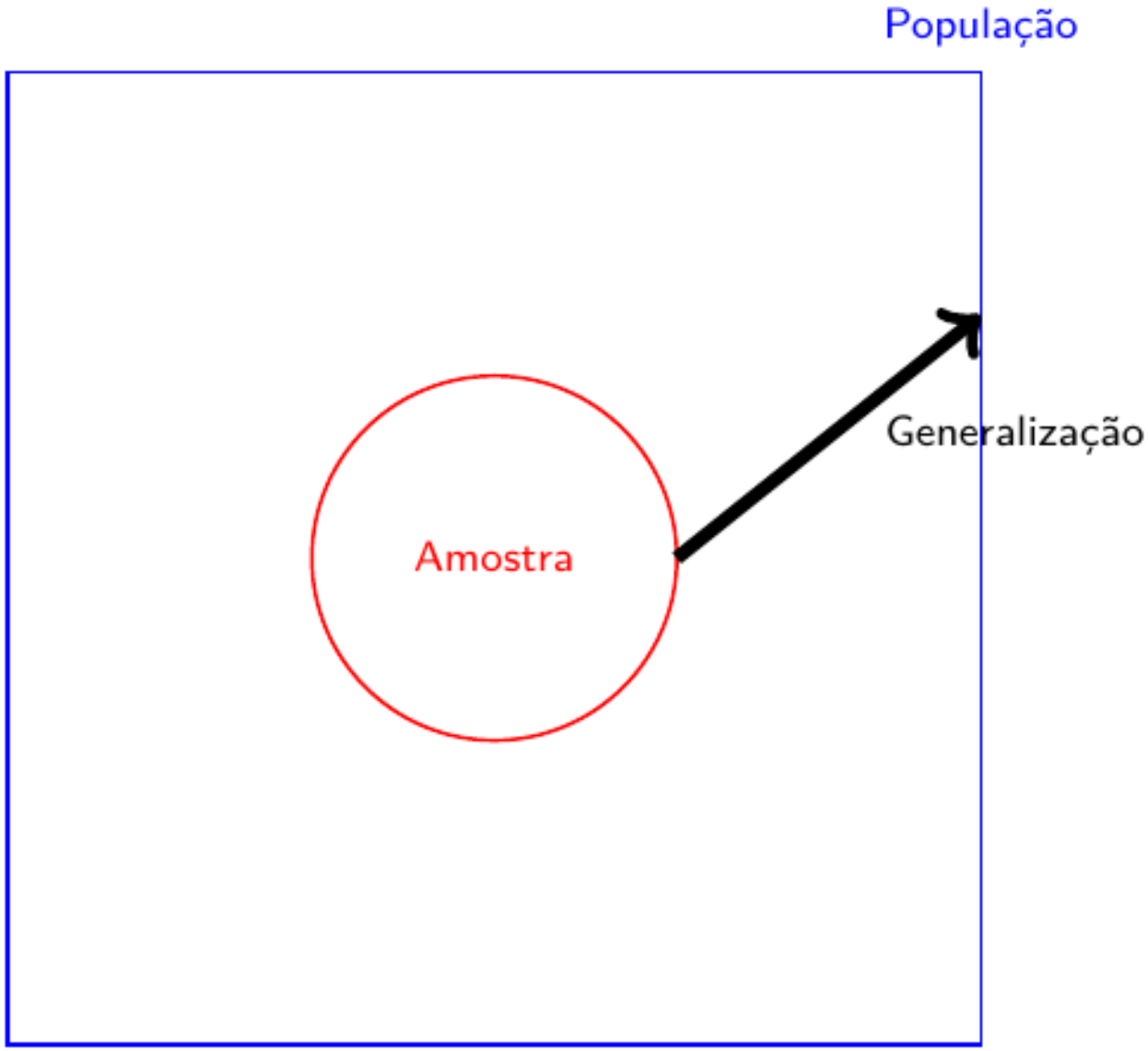
$$H_0 : \text{a média em matemática no enem em salvador é no máximo } 600$$
$$H_1 : \text{a média em matemática no enem em salvador é maior que } 600$$

Probabilidade

Probabilidade

Por que estudar probabilidade?

- **Exploração e visualização de dados** vale apenas para amostra
- **Inferência estatística** realiza *generalização* ~~indução~~ da amostra para população
    - Precisamos de probabilidade para fazer essa *generalização*


População
Amostra
Generalização

Probabilidade

Fenômero aleatório

Experimento (ou procedimento) cujo resultado é impossível de antecipar.
Por exemplo:

- Teremos uma guerra entre China e Estados Unidos da América?
- Quem ganhará a eleição em 2022 no Brasil?
- Qual o resultado de um lançamento de um dados?

Probabilidade

Nomenclatura básica

Considere o experimento aleatório que consite no lançamento de um dado.

- **Espaço amostral:** Conjunto de todos os resulados possíveis de um fenômeno aleatória.
  Exemplo: $\Omega = \{\text{face } 1, \text{face } 2, \text{face } 3, \text{face } 4, \text{face } 5, \text{face } 6\}$.
- **Evento:** Subconjunto de um espaço amostral.
  Exemplo: $A = \{\text{face par}\}$.
- **Probabilidade:** Plausibilidade de um ponto amostra ω de A ser o resultado do fenômeno aleatório.
  Exemplo: $P(A) = 0,5$.

Variável aleatória

Variável aleatória

Variável aleatória é uma **função**: $X : \Omega \to \mathbb{R}$.

Ideia:

- em Ω (e seus subconjuntos) conseguimos calcular probabilidade
- em $\mathbb{R}$ temos os dados (valores) que coletamos *na natureza* (*na prática*)

Classificação de variáveis aleatórias

- Se a imagem de X (valores possíveis de X) são números inteiros, temos uma **variável aleatória discreta**
- Se a imagem de X (valores possíveis de X) é um intervalo de valores, temos uma **variável aleatória contínua**

Variável aleatória

Uso

- Existem vários modelos (distribuições de probabilidade) de *variável aleatória*
- Analisamos o problema (com *exploração e visualização de dados*) e escolhemos a “melhor” distribuição de probabilidade para este problema

- Vamos apresentar várias distribuições de probabilidade
- Para cada distribuição de probabilidade, temos *estimação pontual*, *intervalo de confiança* e *teste de hipóteses* diferentes

Variável aleatória discreta

Valores possíveis são número s inteiros

Função de probabilidade:

$$f(x) = P(X = x)$$

Suporte
Valores possíveis de X.

$$\chi = \{x_1, \cdots, x_n\}.$$

Distribuição Bernoulli

Quando usar: temos um *fenômeno aleatório* com dois resultados possíveis.

O resultado mais relevante chamamos de **sucesso** (com probabilidade p).
O outro resultado chamamos de **fracasso** (com probabilidade $1-p$).

Sucesso é representado por **1**.
Fracasso é representado por **0**.

Exemplo:

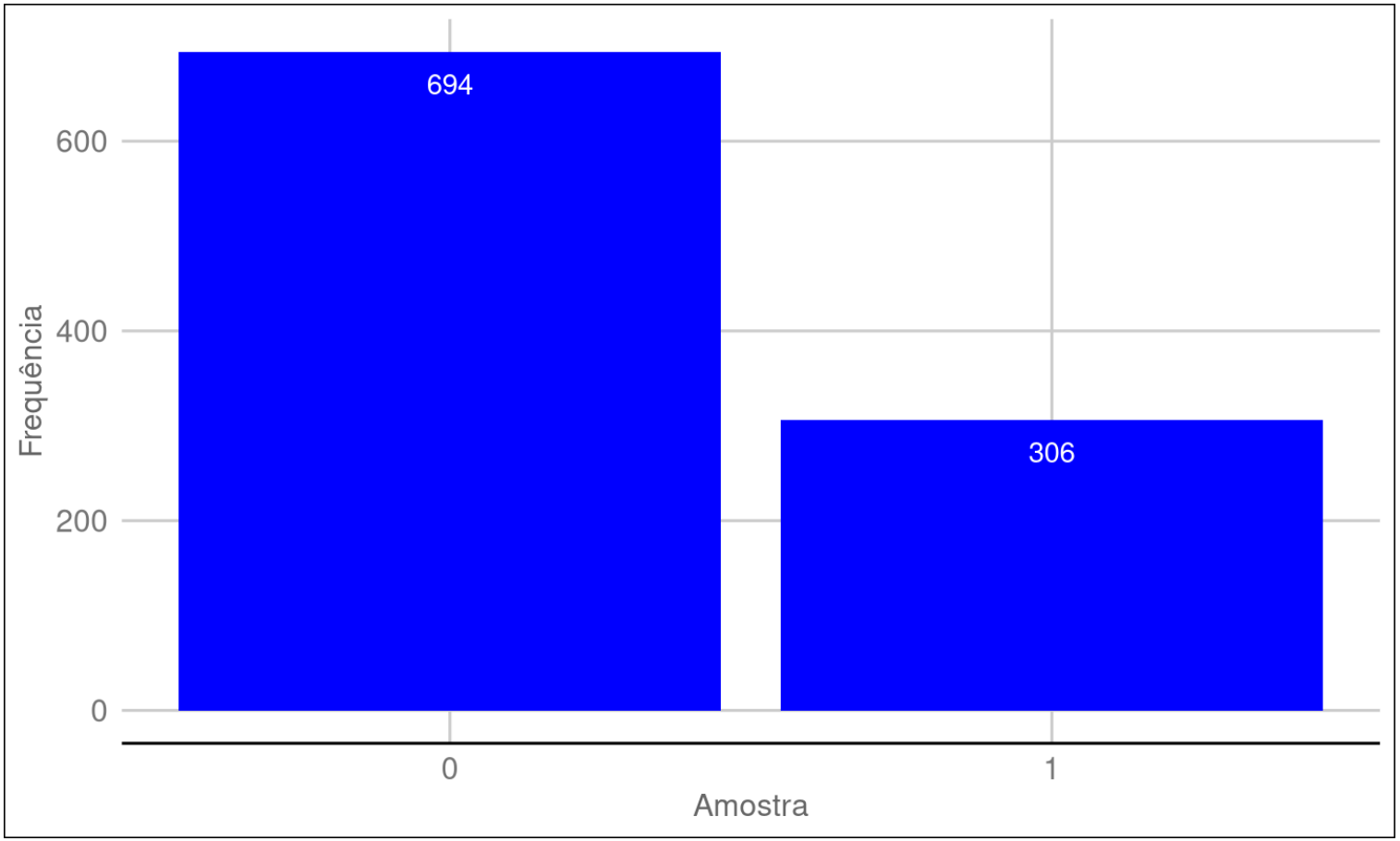
- Paciente pode estar *infectado com covid-19* ou *não infectado*
- Brasil pode *ganhar* ou *perder* a copa do mundo
- Lula pode *vencer* ou *perder* as eleições

Suporte: $\chi = \{0, 1\}$
Função de probabilidade: $f(0) = 1 - p$ e $f(1) = p$

Distribuição Bernoulli no R

Probabilidade de sucesso é $0,3$.

```
prob_sucesso <- 0.3
amostra <- rbinom(1000, 1, prob_sucesso)
ggplot(tibble(x = amostra)) +
    geom_bar(aes(x), fill = "blue") +
    geom_text(aes(x = x, label = ..count..),
            stat = "count", vjust = 2, colour = "white") +
    theme_gdocs()
```


694
306
600
400
200
0
Frequência
0
1
Amostra

Distribuição Bernoulli no R

Medidas resumo e quartis.

```
dados <- tibble(x = amostra)
dados |>
    summarise(media = mean(x), dp = sd(x), cv = dp / media,
              q1 = quantile(x, 0.25), q2 = quantile(x, 0.5), q3 = quantile(x, 0.75))
```

```
## # A tibble: 1 × 6
##   media    dp    cv    q1    q2    q3
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.306 0.461  1.51     0     0     1
```

Distribuição Binomial

Quando usar: temos n casos onde caso tem dois resultados possíveis, e queremos contar o número de sucesso.

O resultado mais relevante chamamos de **sucesso** (com probabilidade p).
O outro resultado chamamos de **fracasso** (com probabilidade $1-p$).

Sucesso é representado por **1**.
Fracasso é representado por **0**.

X: número de sucessos.

Exemplo:

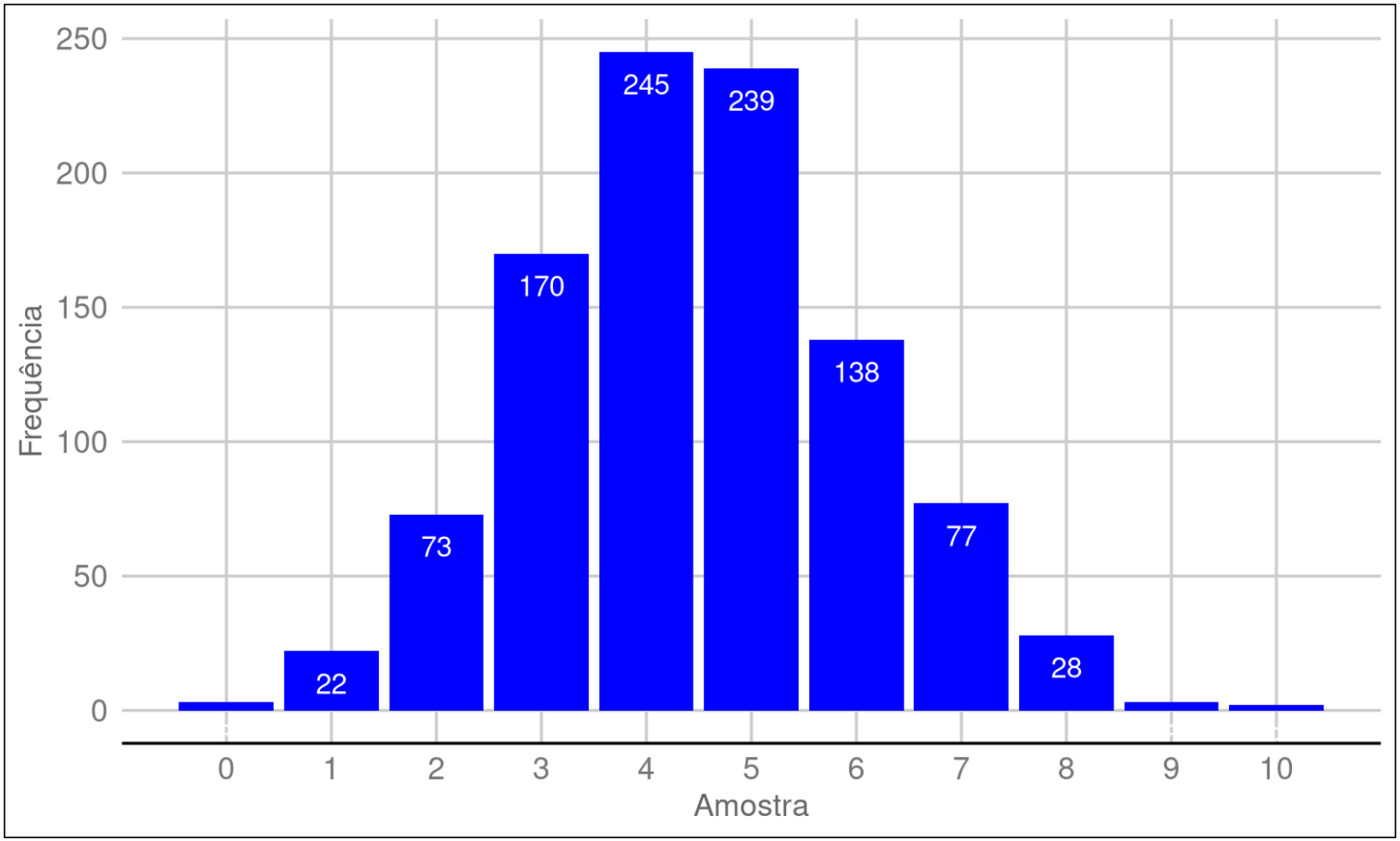
- Num grupo de 20 moscas, quantas sobrevivem ao agrotóxico?
- Num de 50 alunos, quantos serão aprovados?
- Num grupo de 10 funcionários, quantos estão infectados com covid-19?

Suporte: $\chi = \{0, 1, \cdots, n\}$
Função de probabilidade: $f(x) = \binom{n}{k} p^k (1-p)^{n-k}$, onde $x \in \{0, 1, \cdots, n\}$

Distribuição Binomial no R

10 casos e probabilidade de sucesso $0,45$.

```
prob_sucesso <- 0.45
amostra <- rbinom(1000, 10, prob_sucesso)
ggplot(tibble(x = amostra)) +
    geom_bar(aes(x), fill = "blue") +
    geom_text(aes(x = x, label = ..count..),
            stat = "count", vjust = 2, colour = "white") +
    theme_gdocs()
```


250
200
150
100
50
0
Frequência
245
239
170
138
77
73
28
22
0
1
2
3
4
5
6
7
8
9
10
Amostra

Distribuição Binomial no **R**

```
tibble(x = amostra) |>
    summarise(media = mean(x), dp = sd(x), cv = dp / media,
              q1 = quantile(x, 0.25), q2 = quantile(x, 0.5), q3 = quantile(x, 0.75))
```

```
## # A tibble: 1 × 6
##   media    dp    cv    q1    q2    q3
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  4.49  1.60 0.355     3     4     5
```

Distribuição Poison

Quando usar: temos um intervalo limitado e bem especificado, e queremos contar o número de *ocorrências* neste intervalo

X: número de ocorrência em um intervalo de tempo.

Exemplo:

- Número de partículas emitidas por um isótopo em um minuto
- Número de clientes que entram numa loja em um dia útil
- Número de mortes no trânsito em um mês
- Número de mortes diárias por covid-19

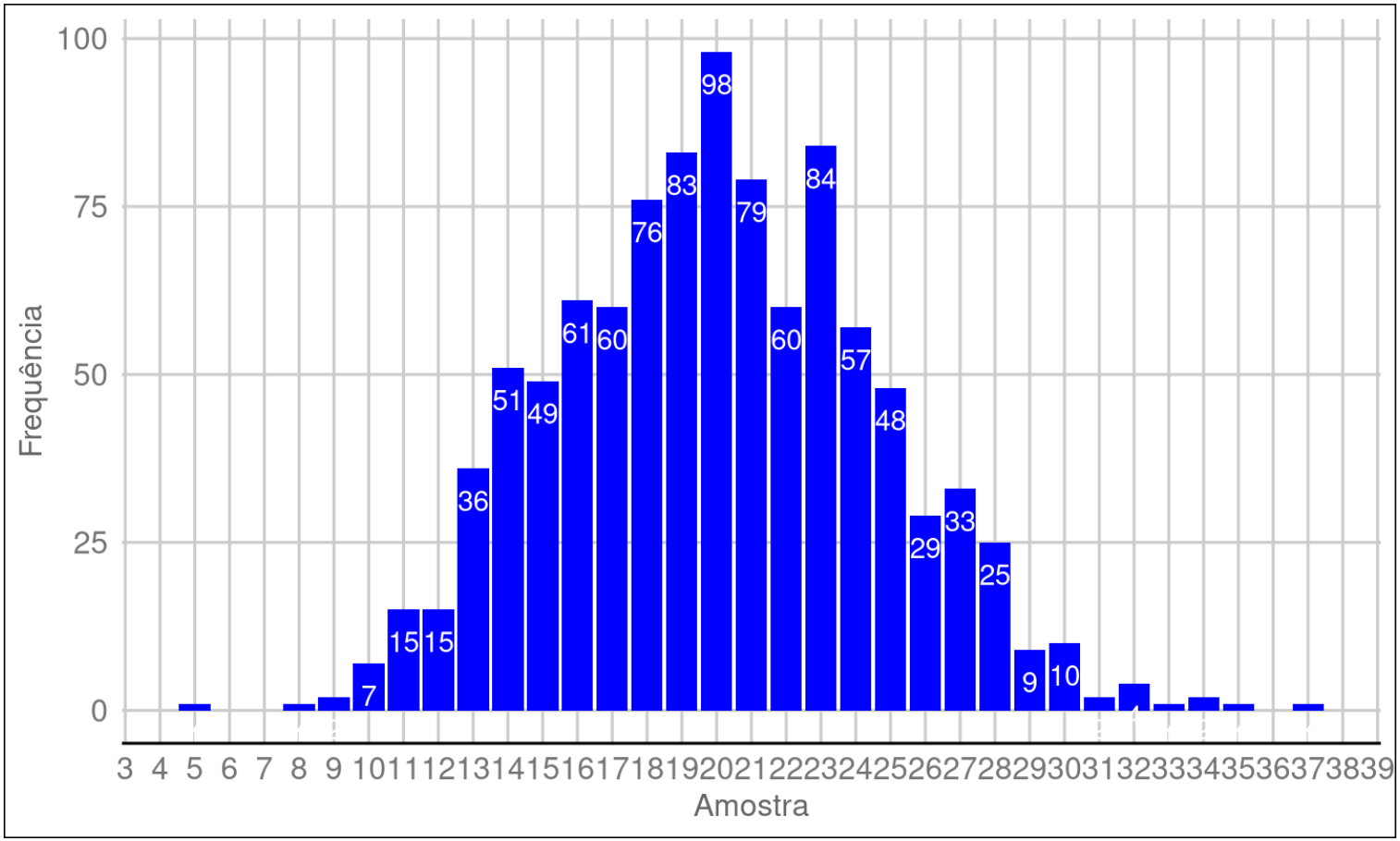
Suporte: $\chi = \{0, 1, \cdots, n, \cdots, \}$
Função de probabilidade: $f(x) = \frac{\exp(-\lambda)\lambda^x}{x!}$, para $x \in \{0, 1, \cdots, n, \cdots\}$

λ é a média de ocorrência entre todos os intervalos (possíveis e imagináveis).

Distribuição Poison no **R**

Número médio de ocorrência em 20 em um intervalo de tempo.

```
media <- 20
amostra <- rpois(1000, media)
ggplot(tibble(x = amostra)) +
    geom_bar(aes(x), fill = "blue") +
    geom_text(aes(x = x, label = ..count..),
            stat = "count", vjust = 2, colour = "white") +
    theme_gdocs()
```


100
75
50
25
0
Frequência
7
15
15
36
51
49
61
60
76
83
98
79
60
84
57
48
29
33
25
9
10
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Amostra

Distribuição Poison no **R**

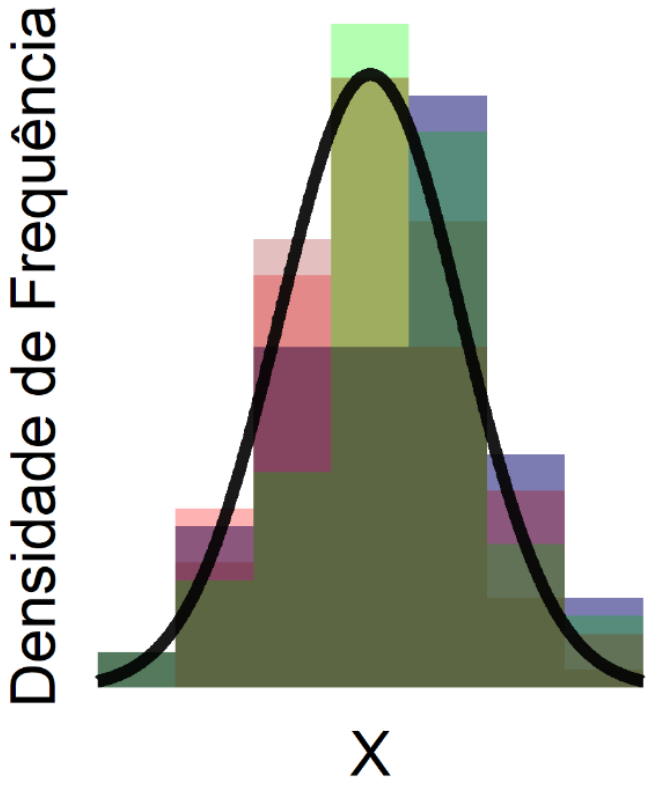
```
tibble(x = amostra) |>
    summarise(media = mean(x), dp = sd(x), cv = dp / media,
              q1 = quantile(x, 0.25), q2 = quantile(x, 0.5), q3 = quantile(x, 0.75))
```

```
## # A tibble: 1 × 6
##   media    dp    cv    q1    q2    q3
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  20.0  4.62 0.231    17    20    23
```

Variável aleatória contínua

Valores possíveis estão dentro de um intervalo (pode ser fracionado)

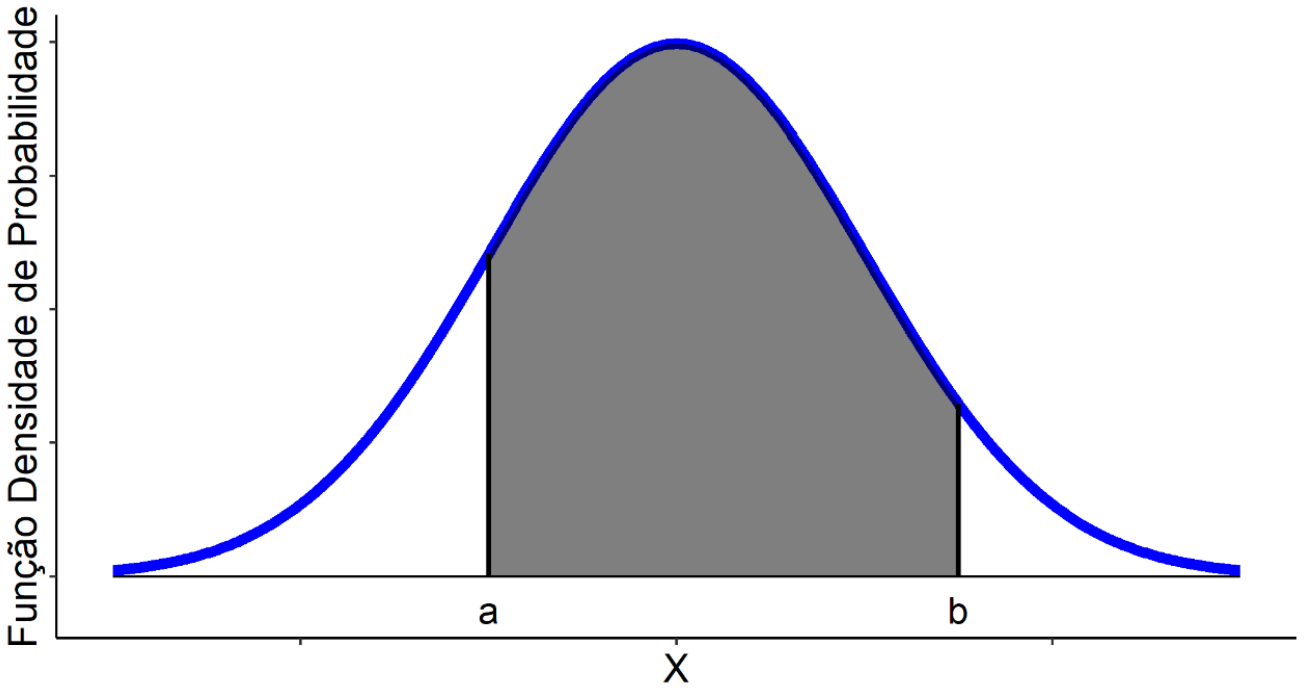
Função densidade de probabilidade: $f(x)$ para $x \in \chi$.



Variável aleatória contínua

Como calcular a probabilidade?

$$P(a < X < b) = \int_a^b f(u)du.$$



Distribuição exponencial

Quando usar: calculamos o tempo até ocorrer um evento de interesse (ocorrência)

X: tempo até a ocorrência

Exemplo:

- Tempo até *a morte* do paciente
- Tempo até *equipamento quebrar*
- Tempo até *a pessoa não pagar*

Suporte: $\chi = [0, \infty)$

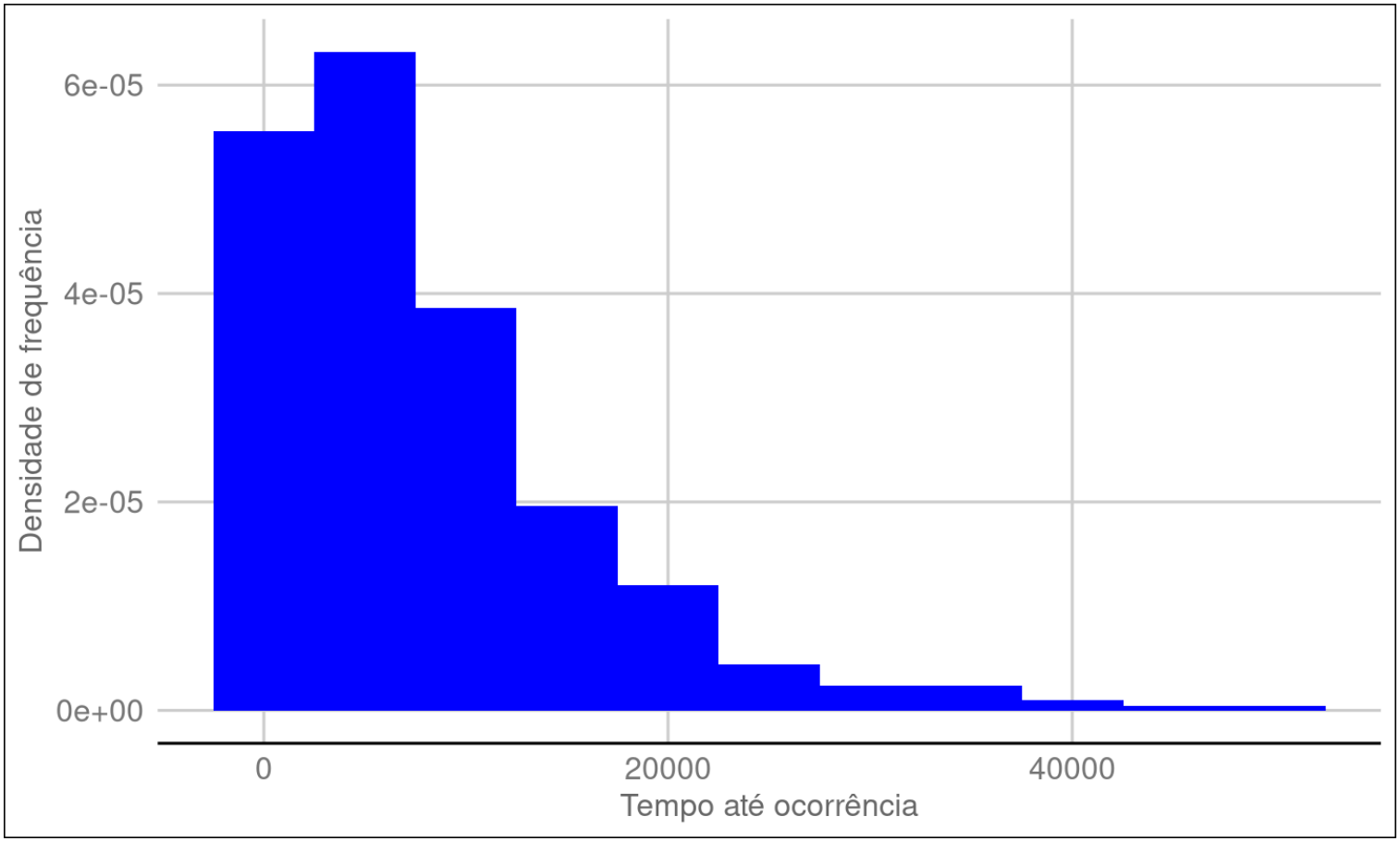
Função de probabilidade: $f(x) = \alpha \cdot \exp(\alpha \cdot x)$, para $x \in \chi$

- α: taxa de decaimento
- $\alpha = \frac{1}{\mu}$, onde μ é o tempo médio até a ocurrência

Distribuição exponencial

Tempo médio até a ocorrência é 8000 horas.

```
media <- 8000
k <- ceiling(1 + log2(1000))
taxa_decaimento <- 1 / media
amostra <- rexp(1000, taxa_decaimento)
ggplot(tibble(tempo = amostra)) +
    geom_histogram(aes(tempo, y = ..density..), fill = "blue", bins = k) +
    theme_gdocs()
```


6e-05
4e-05
2e-05
0e+00
Densidade de frequência
0
20000
40000
Tempo até ocorrência

Distribuição normal

Quando usar: valores ficam em torno de uma média e valores longe da média são *raros*

X: Valor *que está provavelmente perto da média*

Exemplo:

- altura de brasileiros
- peso de baianos
- nota de estudantes

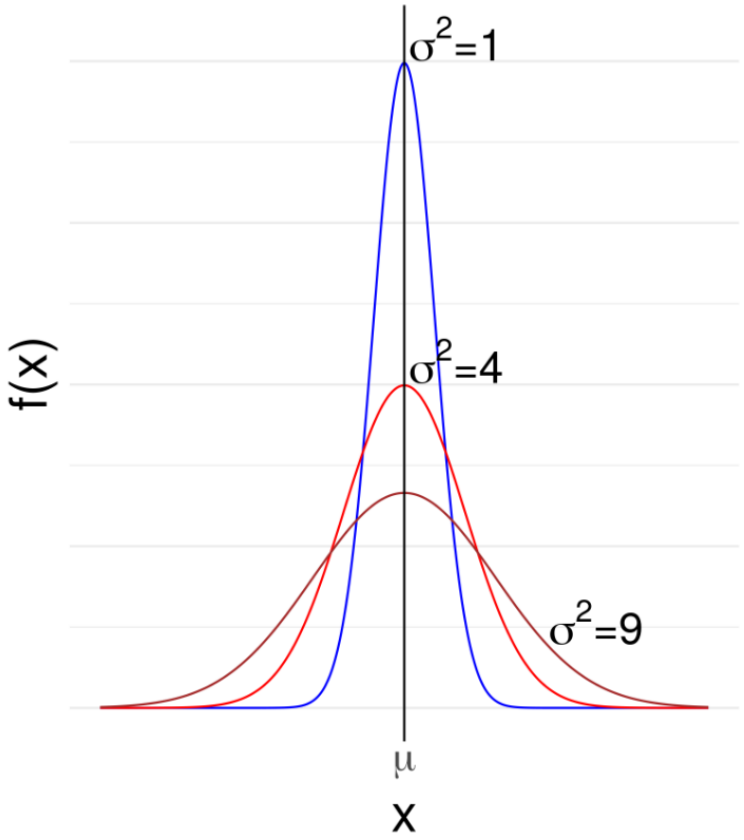
Suporte: $\chi = (-\infty, \infty)$

Função de probabilidade: $f(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2 \cdot \sigma^2}\right)$, para $x \in \chi$

- μ: média da população
- σ: desvio padrão da população

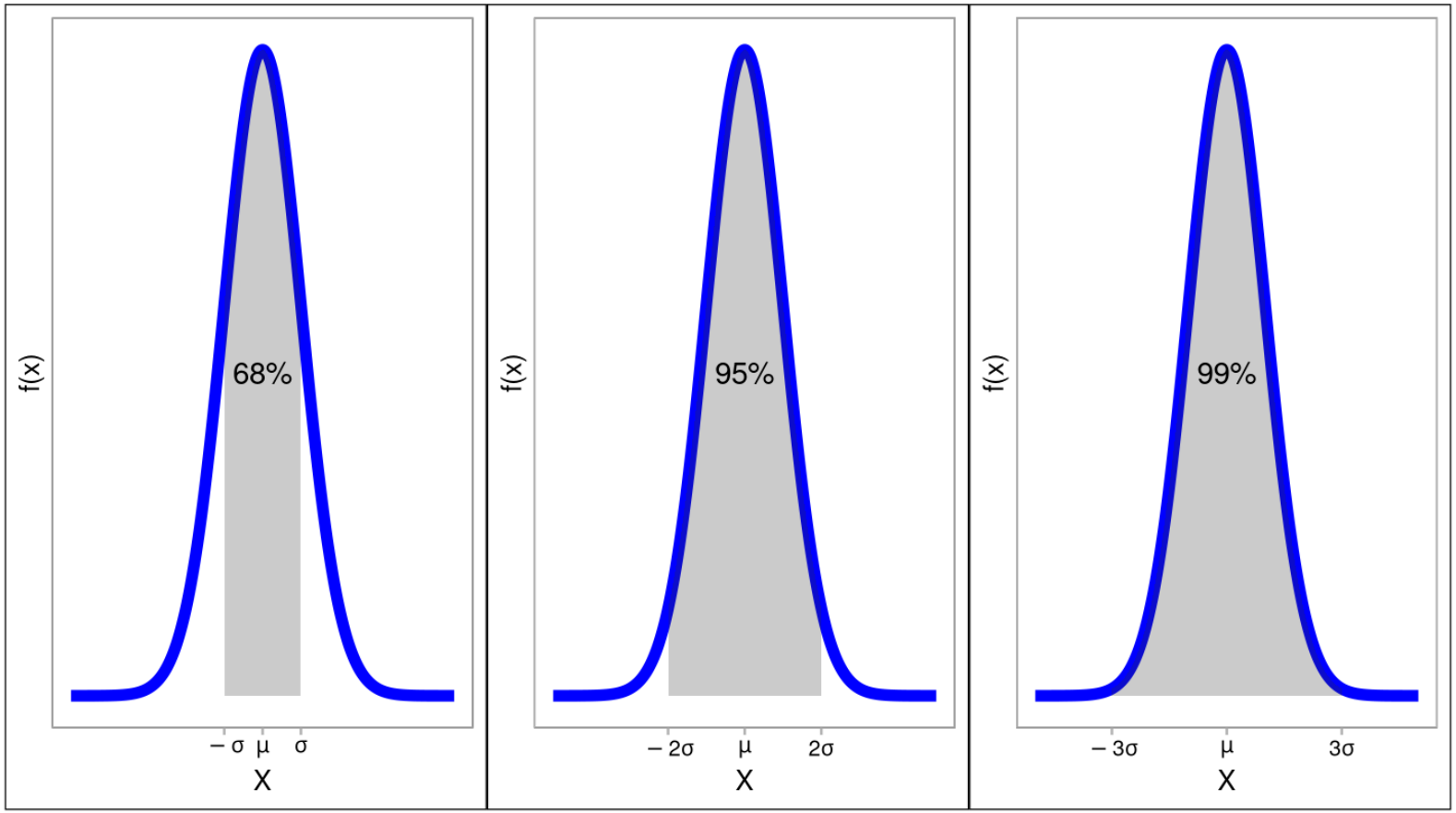
Distribuição normal

Função densidade de probabilidade



Distribuição normal

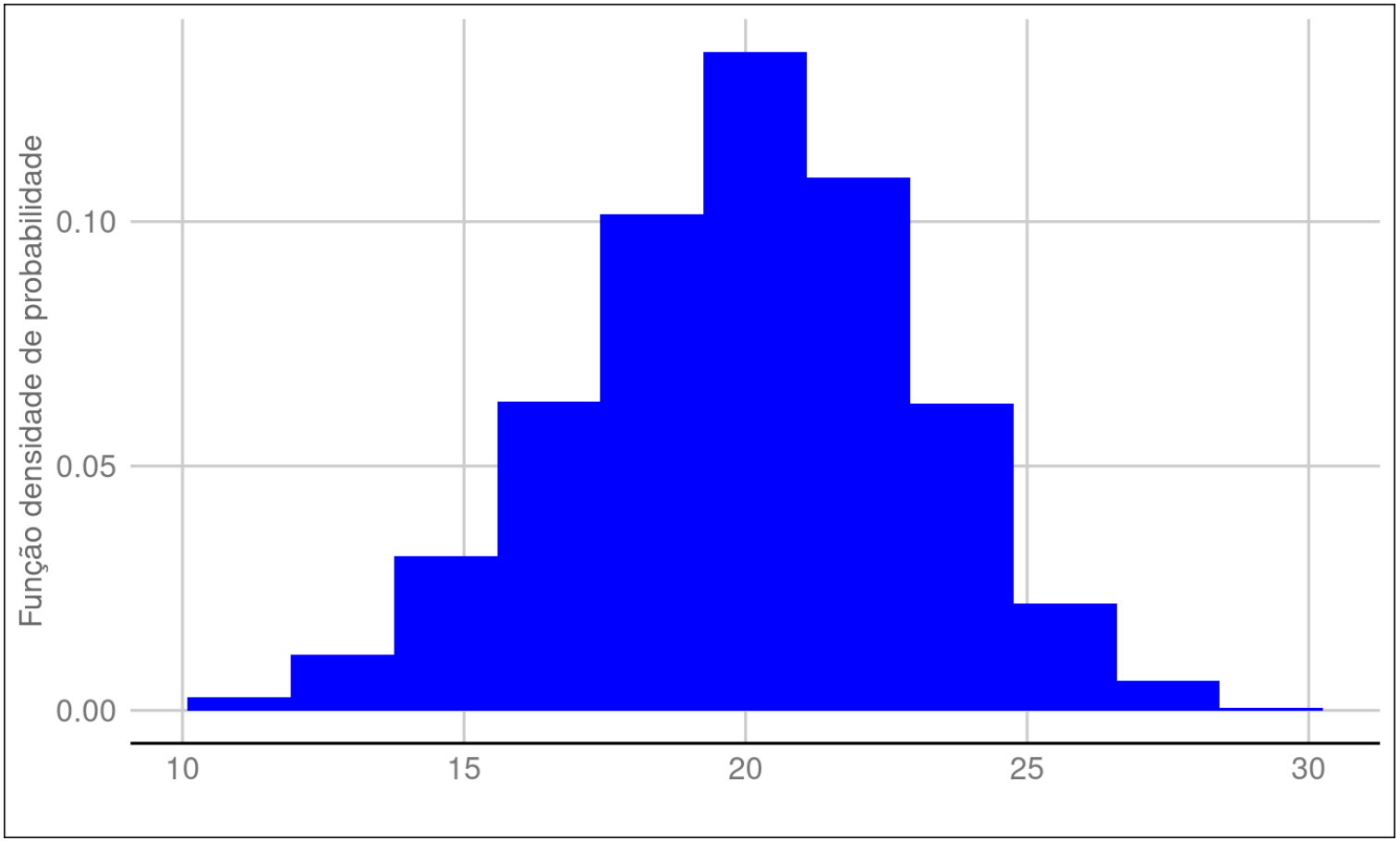
Área de acordo com o desvio padrão.



Distribuição normal

- Média: 20
- Desvio padrão: 3

```
media <- 20
dp <- 3
k <- ceiling(1 + log2(1000))
amostra <- rnorm(1000, mean = media, sd = dp)
ggplot(tibble(x = amostra)) +
    geom_histogram(aes(x = x, y = ..density..), bins = k, fill = "blue") +
    theme_gdocs()
```


Função densidade de probabilidade
0.10
0.05
0.00
10
15
20
25
30

Distribuição normal

```
tibble(x = amostra) |>
    summarise(media = mean(x), dp = sd(x), dv = dp / media,
              q1 = quantile(x, 0.25), q2 = quantile(x, 0.5), q3 = quantile(x, 0.75))
```

```
## # A tibble: 1 × 6
##   media    dp    dv    q1    q2    q3
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  20.0  3.03 0.151  18.0  20.0  22.1
```