Estatística descritiva usando R

bem-vinde ao tidyverse

Estimação pontual e intervalo de confiança usando R

Profa Carolina e Prof Gilberto

Parte 4

Inferência estatística

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

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

Estimação pontual

Estimação pontual

Amostra	Distribuição	Parâmetros	Estimativa
$x_1, \ldots, x_m$	$X \sim Bernoulli(p)$	p	$\hat{p} = \frac{x_1 + \cdots + x_m}{m}$
$x_1, \ldots, x_m$	Estudo balanceado: $X \sim b(n, p)$	p	$\hat{p} = \frac{x_1 + \cdots + x_m}{n \cdot m}$
$x_1, \ldots, x_m$	Estudo não balanceado: $X_1 \sim b(n_1, p), \ldots, X_m \sim b(n_m, p)$	p	$\hat{p} = \frac{x_1 + \cdots + x_m}{n_1 + \cdots + n_m}$
$x_1, \ldots, x_m$	$X \sim Exp(\alpha)$	α	$\hat{\alpha} = \frac{1}{\bar{x}}$
$x_1, \ldots, x_m$	$X \sim Poison(\lambda)$	λ	$\lambda = \bar{x}$
$x_1, \ldots, x_m$	$X \sim N(\mu, \sigma^2)$	μ, σ^2	$\hat{\mu} = \bar{x}$ $\sigma = s$

Distribuição Bernoulli

Dados sobre teste de diabetes para mulheres do povo Pima nos Estados Unidos.
Vamos considerar *sucesso* o teste de diabetes dar positivo.

Lendo os dados

```
df_prima <- read_xlsx("data/raw/dados.xlsx", sheet = "PimaIndiansDiabetes")
```

Definindo sucesso

```
df_prima <- df_prima |> mutate(diabetes_logical = diabetes == "pos")
```

Aproximando a probabilidade de sucesso p

```
df_prima |>
    summarise(prob_sucesso = mean(diabetes_logical))
```

```
## # A tibble: 1 × 1
##   prob_sucesso
##          <dbl>
## 1        0.349
```

Distribuição binomial

Dados sobre filmes estreados entre 1936 e 2019.
Vamos considerar *sucesso* se o filme é drama.

Lendo os dados

```
df_filmes <- read_xlsx("data/raw/filmes_drama.xlsx")
```

Aproximando a probabilidade de sucesso p

```
df_filmes |>
    summarise(prob_sucesso = sum(numero_drama) / sum(total_filmes))
```

```
## # A tibble: 1 × 1
##   prob_sucesso
##          <dbl>
## 1        0.363
```

Distribuição Poison

Dados sobre o número de visitas ao médico nos Estados Unidos da América em dois anos.

Lendo os dados

```
df_demanda_saude <- read_xlsx("data/raw/dados.xlsx", sheet = "demandaSaude")
```

Aproximando a média de ocorrências

```
df_demanda_saude |>
    summarise(media = mean(num_med))
```

```
## # A tibble: 1 × 1
##   media
##   <dbl>
## 1  5.77
```

Distribuição exponencial

Tempo até a morte de pacientes diagnosticados com câncer avançado no Pulmão.

Lendo os dados

```
df_cancer <- read_xlsx("data/raw/dados.xlsx", sheet = "cancerPulmao")
```

Tempo médio de ocorrências

```
df_cancer |> summarise(tempo_medio = mean(tempo))
```

```
## # A tibble: 1 × 1
##   tempo_medio
##         <dbl>
## 1         283
```

Taxa de decaimento

```
df_cancer |> summarise(taxa_decaimento = 1 / mean(tempo))
```

```
## # A tibble: 1 × 1
##   taxa_decaimento
##             <dbl>
## 1         0.00353
```

Distribuição normal

Dados socio-econômicos de 36 funcionários da companhia MB.

Lendo os dados

```
df_funcionarios <- read_xlsx("data/raw/companhia_MB.xlsx")
```

Média e desvio padrão salarial

```
df_funcionarios |>
    summarise(media = mean(salario), dp = sd(salario))
```

```
## # A tibble: 1 × 2
##   media    dp
##   <dbl> <dbl>
## 1  11.1  4.59
```

Intervalo de confiança

Intervalo de confiança

- encontrar a e b tal que $a < \mu < b$ (ou $a < \sigma < b$ ou $a < p < b$) com medida de *precaução ou prudência* γ
- chamamos γ de coeficiente de confiança
- Chamamos (a, b) de intervalo de confiança
- μ (ou σ ou p) pode ou não pode estar entre a e b
- $100 \cdot \gamma\%$ dos intervalos que contém μ

Intervalo de confiança

População: 30 garrafas raras de vinhos
Amostra: seis amostras com cinco garrafas
Variável: teor alcoolico

Lendo os dados

```
df_teor_alcoolico <- read_xlsx("data/raw/dados.xlsx", sheet = "teor_alcoolico")
```

Média e desvio padrão populacional

```
media_pop <- mean(df_teor_alcoolico$teor_pop)
media_pop
```

```
## [1] 6.373
```

```
dp_pop <- sd(df_teor_alcoolico$teor_pop)
dp_pop
```

```
## [1] 1.145565
```

Intervalo de confiança

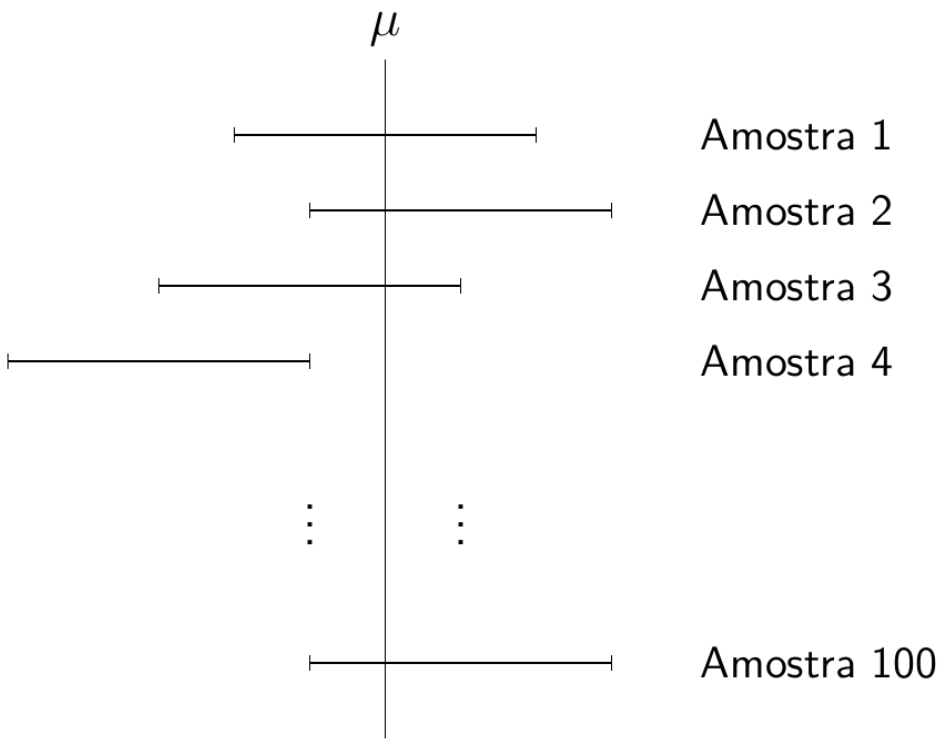
Média está no intervalo de confiança?

```
df_teor_alcoolico |>
    group_by(amostra) |>
    summarise(
        li = ci_1pop_norm(teor_amostra, sd_pop = dp_pop)$lower_ci,
        media = media_pop,
        ls = ci_1pop_norm(teor_amostra, sd_pop = dp_pop)$upper_ci
    ) |>
    mutate(contem_media = ifelse(li < media & media < ls, "Sim", "Não"))
```

```
## # A tibble: 6 × 5
##   amostra     li media    ls contem_media
##   <chr>    <dbl> <dbl> <dbl> <chr>
## 1 amostra1  6.49  6.37  8.50 Não
## 2 amostra2  5.13  6.37  7.14 Sim
## 3 amostra3  5.79  6.37  7.80 Sim
## 4 amostra4  4.30  6.37  6.31 Não
## 5 amostra5  4.48  6.37  6.48 Sim
## 6 amostra6  5.35  6.37  7.36 Sim
```

Intervalo de confiança

- $100 \cdot \gamma\%$ dos intervalos *estão corretos* (contém o parâmetro)



Intervalo de confiança – Bernoulli

Tipo de língua na prova do ENEM.
Vamos considerar *sucesso* se a prova for *Inglês*.

Lendo os dados

```
df_enem <- read_xlsx("data/raw/amostra_enem_salvador.xlsx")
```

Acrescento indicador de sucesso

```
df_enem <- df_enem |> mutate(ingles_logico = TP_LINGUA == "Inglês")
```

Construindo o intervalo de confiança

Coeficiente de confiança: 99%.

```
ci_1pop_bern(df_enem$ingles_logico, conf_level = 0.99)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci conf_level
##      <dbl>    <dbl>      <dbl>
## 1    0.641    0.677       0.99
```

Intervalo de confiança – Bernoulli

Pesquisa eleitoral IPEC – 29/08/2022
Vamos considerar *sucesso* o eleitor votar no Lula.

Na matéria temos: 2.512 pessoas entrevistadas e 1.105 pessoas falaram que votariam em Lula.

Intervalo de confiança

Coeficiente de confiança: 95%.

```
ci_1pop_bern(1105, 2512)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci conf_level
##      <dbl>    <dbl>      <dbl>
## 1    0.420    0.459       0.95
```

Intervalo de confiança – Poison

Dados sobre o número de visitas ao médico nos Estados Unidos da América em dois anos.

Lendo os dados

```
df_demanda_saude <- read_xlsx("data/raw/dados.xlsx", sheet = "demandaSaude")
```

Intervalo de confiança

Coeficiente de confiança: 97%.

```
ci_1pop_general(df_demanda_saude$num_med, conf_level = 0.97)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci conf_level
##      <dbl>    <dbl>      <dbl>
## 1     5.55     6.00       0.97
```

Intervalo de confiança – exponencial

Tempo até a morte de pacientes diagnosticados com câncer avançado no Pulmão.

Lendo os dados

```
df_cancer <- read_xlsx("data/raw/dados.xlsx", sheet = "cancerPulmao")
```

Intervalo de confiança

Coeficiente de confiança: 99,9%.

```
ci_1pop_exp(df_cancer$tempo, conf_level = 0.999)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci conf_level
##      <dbl>    <dbl>      <dbl>
## 1     222.     371.      0.999
```

Intervalo de confiança – normal

Checar a distribuição normal

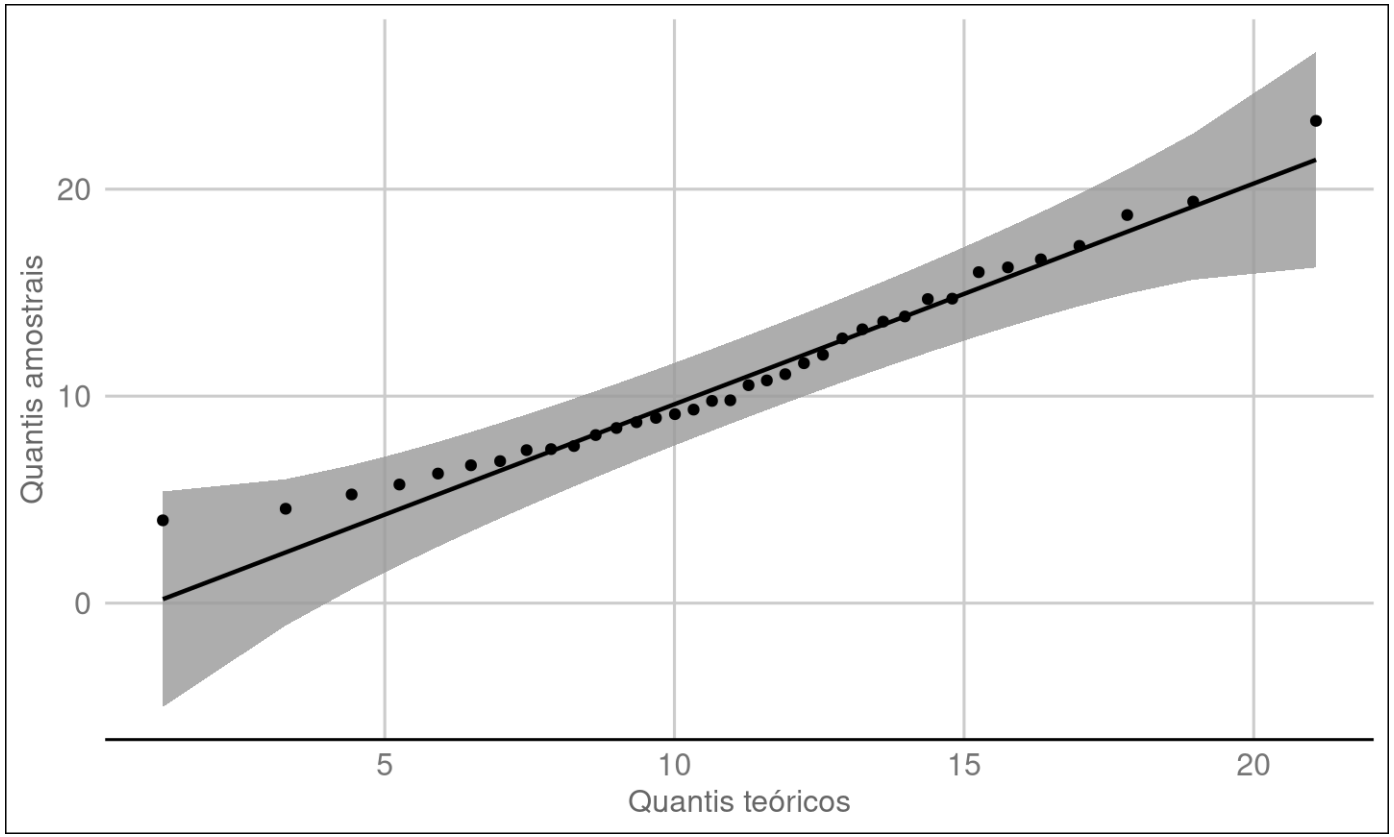
Vamos analisar o salário dos 36 funcionários da companhia MB.

Lendo os dados

```
df_funcionario <- read_xlsx("data/raw/companhia_MB.xlsx")
```

QQPlot

```
library(qqplotr)
ggplot(df_funcionario, aes(sample = salario)) +
    stat_qq_band() + stat_qq_line() + stat_qq_point() +
    labs(x = "Quantis teóricos", y = "Quantis amostrais") +
    theme_gdocs()
```


20
10
0
Quantis amostrais
5
10
15
20
Quantis teóricos

Intervalo de confiança – normal

Coeficiente de confiança: 99%.

```
ci_1pop_norm(df_funcionario$salario, conf_level = 0.99)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci conf_level
##      <dbl>    <dbl>      <dbl>
## 1     9.04     13.2       0.99
```