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

Profa Carolina & Prof Gilberto

Instituto de Matemática e Estatística
Universidade Federal da Bahia

Estimação

Para encontrar o valor do parâmetro das distribuições de probabilidade, usamos as medidas de resumo da Tabela 1.

Tabela 1: Estimação pontual para distribuições de probabilidade.

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

Estimação pontual: exemplo simulado

Distribuição Bernoulli

```
m <- 1000 # tamanho da amostra
p <- 0.6 # probabilidade de sucesso
amostra <- rbern(m, p)
mean(amostra) # estimativa da probailidade de sucesso
```

```
## [1] 0.586
```

Distribuição binomial

```
m <- 1000 # tamanho da amostra
p <- 0.6 # probabilidade de sucesso
n <- 10 # número de casos
amostra <- rbinom(m, n, p)
mean(amostra) / n # estimativa da probabilidade sucesso
```

```
## [1] 0.604
```

Distribuição Poison

```
m <- 1000 # tamanho da amostra
media <- 5 # média populacional de ocorrência no intervalo de tempo
amostra <- rpois(m, media)
mean(amostra) # estimativa da média de ocorrência no intervalo de tempo
```

```
## [1] 5.089
```

Estimação pontual: exemplo simulado

Distribuição exponencial

```
m <- 1000 # tamanho da amostra
media <- 150 # média populacional
taxa <- 1 / media # taxa de decaimento
amostra <- rexp(m, rate = 1 / media)
1 / mean(amostra) # estimativa da taxa de decaimento
```

```
## [1] 0.006554166
```

Distribuição normal

```
m <- 1000 # tamanho da amostra
media <- 5 # média populacional
dp <- 1.25 # desvio padrão populacional
amostra <- rnorm(m, mean = media, sd = dp)
mean(amostra) # estimativa da média
```

```
## [1] 5.051186
```

```
sd(amostra) # estimativa do desvio padrão
```

```
## [1] 1.266737
```

Estimação pontual: aplicações (Bernoulli)

Distribuição Bernoulli: $X \sim Bernoulli(p)$

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

```
dados <- read_xlsx('dados.xlsx', sheet = 'PimaIndiansDiabetes')

dados <- dados %>%
  mutate(diabetes_binario = diabetes %>% recode("pos" = 1, "neg" = 0))

# Estimativa de proporção de sucesso
dados %>%
  summarise(prop_estimativa = mean(diabetes_binario))
```

```
## # A tibble: 1 x 1
##   prop_estimativa
##             <dbl>
## 1           0.349
```

Estimação pontual: aplicações (Binomial)

Dados sobre a estreia de filmes de Hollywood. Vamos contar o número de estreias de Dramas no dia do mês.

```
dados <- read_xlsx('dados.xlsx', sheet = 'hollywood')

# Criando duas variáveis: dia de lançamento, indDrama: 1 se for drama
dados <- dados %>%
  mutate(diaLancamento = ymd(dataLancamento) %>% day()) %>%
  mutate(indDrama = ifelse(genero %in% "Drama", 1, 0))

dadosDia <- dados %>%
  group_by(diaLancamento) %>%
  summarise(numDrama = sum(indDrama), numEstreias = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# Estimativa da proporção de Sucesso (Drama)
dadosDia %>%
  summarize(prop_estimativa = sum(numDrama) / sum(numEstreias))
```

```
## # A tibble: 1 x 1
##   prop_estimativa
##             <dbl>
## 1           0.363
```

Estimação pontual: aplicações (Poisson)

Dados sobre o número de visitas ao médico nos Estados Unidos em dois anos (1987-1988).

```
dados <- read_xlsx('dados.xlsx', sheet = 'demandaSaude')

# Número médio de visitas ao médico em dois anos
dados %>%
  summarise(media_lambda = mean(numMed))
```

```
## # A tibble: 1 x 1
##   media_lambda
##          <dbl>
## 1         5.77
```

Estimação pontual: aplicações (Exponencial)

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

```
dados <- read_xlsx('dados.xlsx', sheet = 'cancerPulmao')

# Estimativa do tempo média de vida
dados %>%
  summarise(estimativa_tempo = mean(tempo))
```

```
## # A tibble: 1 x 1
##   estimativa_tempo
##              <dbl>
## 1              283
```

```
# Estimativa da taxa de decaimento
dados %>%
  summarise(taxa_decaimento = 1 / mean(tempo))
```

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

Estimação pontual: aplicações (Normal)

Dados socio-econômicos de 36 funcionários da seção de orçamentos da companhia MB. Exemplo didático extraído do seguintes livro:

MORETTIN, Pedro Alberto; BUSSAB, Wilton Oliveira. **Estatística básica.** Saraiva Educação SA, 2017.

```
dados <- read_xlsx('dados.xlsx', sheet = 'companhia_MB')

# Estimativa para média e para o desvio padrão (em reais)
dados %>%
  summarise(media = mean(salario), desvio_padrao = sd(salario))
```

```
## # A tibble: 1 x 2
##   media desvio_padrao
##   <dbl>         <dbl>
## 1  11.1          4.59
```

Estimação intervalar

Objetivo

- parâmetro μ desconhecido
- encontrar a e b tal que $a < \mu < b$ com uma medida de *precaução ou prudência* γ
- chamamos γ de coeficiente de confiança
- chamamos (a, b) de intervalo de confiança
- μ pode estar entre a e b
- 100% dos intervalos contém μ

Estimação intervalar

Interpretação

```
dados <- read_xlsx('dados.xlsx', sheet = 'teor_alcoolico')

# média e desvio padrão populacionais
media <- with(dados, mean(teor_pop))
dp <- with(dados, sd(teor_pop))

# Intervalos de confiança para cada amostra
dados %>% group_by(amostra) %>%
  summarise(lower = MeanCI(teor_amostra, sd = dp)['lwr.ci'],
            upper = MeanCI(teor_amostra, sd = dp)['upr.ci'],
            media = media) %>%
  mutate(contemMedia = ifelse(lower < media & media < upper, "Sim", "Não")
```

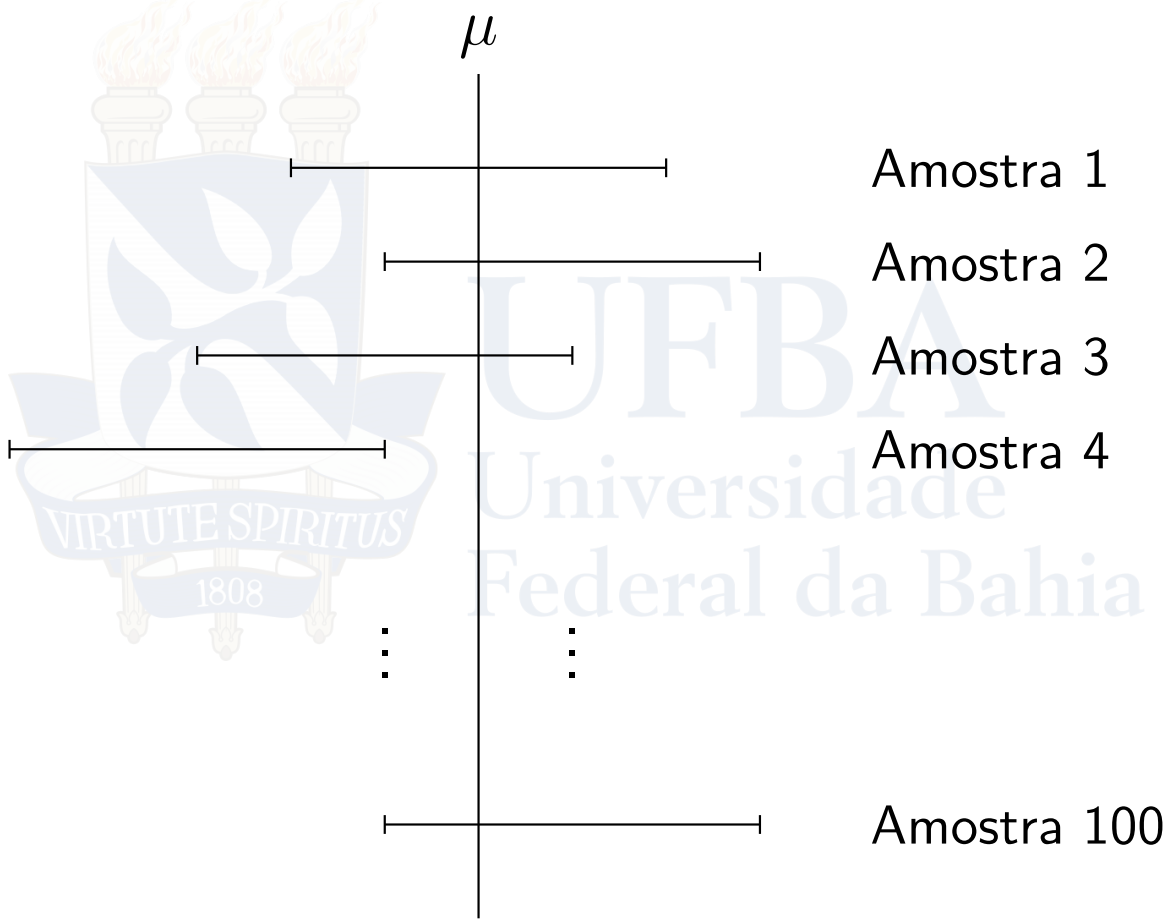
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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

Estimação intervalar

- $100 \cdot \gamma\%$ do intervalos contêm o parâmetro μ;
- Ilustramos essa ideia na Figura 1

Figura 1: Interpretação do coeficiente de confiança.



Estimação intervalar: Bernoulli (proporção)

Exemplo simulado

```
p <- 0.4 # proporção populacional de sucesso
m <- 1000 # tamanho da amostra
amostra <- rbern(m, p) # amostra

conf_bern(amostra, conf.level = 0.99)
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level proportion
##      <dbl>    <dbl>      <dbl>      <dbl>
## 1    0.368    0.450       0.99      0.409
```

Aplicação

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

```
dados <- read_xlsx('dados.xlsx', sheet = 'PimaIndiansDiabetes') %>%
  mutate(diabetes_binario = diabetes %>% recode("pos" = 1, "neg" = 0))

with(dados, conf_bern(diabetes_binario, conf.level = 0.95))
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level proportion
##      <dbl>    <dbl>      <dbl>      <dbl>
## 1    0.314    0.384       0.95      0.349
```

Estimação intervalar: Binomial (proporção)

Exemplo simulado

Aqui vamos assumir que ensaios em que o número de casos é balanceado.

```
n <- 10 # número de casos
p <- 0.65 # proporção populacional de sucesso
m <- 1000 # tamanho da amostra
amostra <- rbinom(m, n, p)

conf_binom(amostra, n)
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level proportion
##      <dbl>    <dbl>      <dbl>      <dbl>
## 1    0.634    0.654       0.95      0.644
```

Aplicação

```
dados <- read_xlsx('dados.xlsx', sheet = 'hollywood')

# Criando duas variáveis: dia de lançamento, indDrama: 1 se for drama
dados <- dados %>%
  mutate(diaLancamento = ymd(dataLancamento) %>% day()) %>%
  mutate(indDrama = ifelse(genero %in% "Drama", 1, 0))

dadosDia <- dados %>%
  group_by(diaLancamento) %>%
  summarise(numDrama = sum(indDrama), numEstreias = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
with(dadosDia, conf_binom(numDrama, numEstreias))
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level proportion
##      <dbl>    <dbl>      <dbl>      <dbl>
## 1    0.347    0.380       0.95      0.363
```

Estimação intervalar: Poisson

Exemplo simulado

```
m <- 1000 # tamanho da amostra
media <- 5 # média de ocorrência dentro de um intervalo
amostra <- rpois(m, media)

conf_pois(amostra)
```

```
## # A tibble: 1 x 4
##   lower upper conf_level  mean
##   <dbl> <dbl>      <dbl> <dbl>
## 1  4.99  5.28       0.95  5.13
```

Aplicação

Dados sobre o número de visitas ao médico nos Estados Unidos em dois anos (1987-1988).

```
dados <- read_xlsx('dados.xlsx', sheet = 'demandaSaude')

with(dados, conf_pois(numMed))
```

```
## # A tibble: 1 x 4
##   lower upper conf_level  mean
##   <dbl> <dbl>      <dbl> <dbl>
## 1  5.70  5.85       0.95  5.77
```

Estimação intervalar: Exponencial

Exemplo simulado

```
media <- 25 # tempo médio até a ocorrência
m <- 1000 # tamanho da amostra
amostra <- rexp(m, rate = 1 / media)

conf_exp(amostra, conf.level = 0.99)
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level  mean
##      <dbl>    <dbl>      <dbl> <dbl>
## 1     22.0     25.8       0.99  23.8
```

Aplicação

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

```
dados <- read_xlsx('dados.xlsx', sheet = 'cancerPulmao')

with(dados, conf_exp(tempo))
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level  mean
##      <dbl>    <dbl>      <dbl> <dbl>
## 1     244.     332.       0.95   283
```

Checando a distribuição normal

Gráfico de probabilidade normal

- $x_1, \ldots, x_n$ valores observados de X;
- estatísticas de ordem: $x_{(1)}, \ldots, x_{(n)}$;
- Média populacional: μ e desvio padrão populacional: σ;
- Na ausência de μ e σ, podemos usar $\bar{x}$ e s.
- Considere

$$z_{(i)} = \frac{x_{(i)} - \mu}{\sigma}, \ i = 1, \ldots, n,$$
$$\Phi\left(q_{(i)}\right) = \frac{i - 0,5}{n}, \ i = 1, \ldots, n.$$

  em que $P(Z < z) = \Phi(z)$ e $Z \sim N(0, 1)$.

- Para cada par $(z_{(i)}, q_{(i)}), \ i = 1, \ldots, n$, desenho no ponto um plano cartesiano.
- Chamamos este gráfico de quantis de gráfico de probabilidade normal.

Checando a distribuição normal

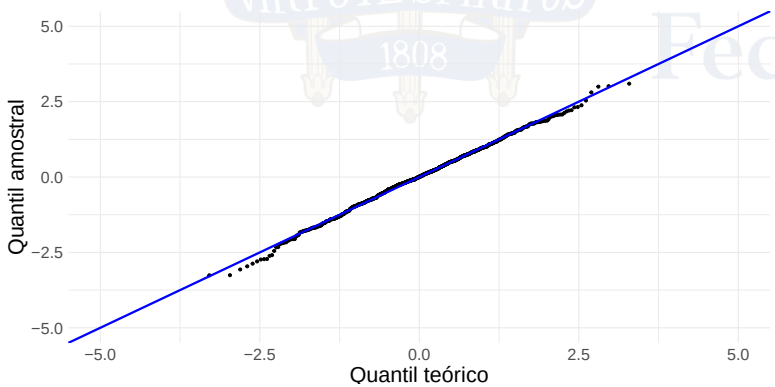
Gráfico de probabilidade normal

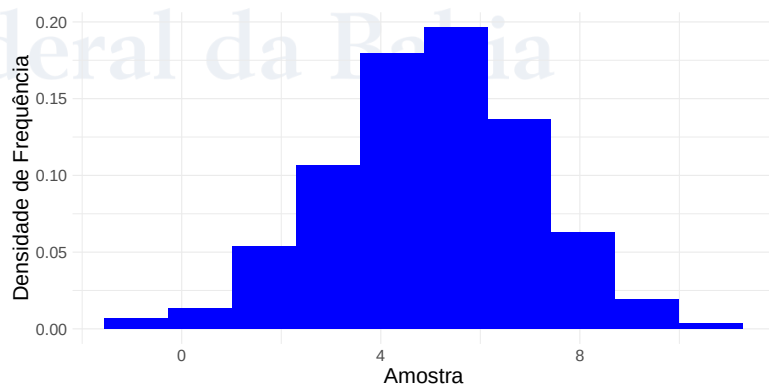
Exemplo simulado.

```
media <- 5 # média populacional
dp <- 2 # desvio padrão populacional
m <- 1000 # tamanho da amostra
amostra <- rnorm(m, mean = media, sd = dp)

qq_norm(amostra) +
  labs(x = "Quantil teórico", y = "Quantil amostral") +
  theme_minimal()

hist_sturge(amostra) +
  labs(y = "Densidade de Frequência", x = "Amostra") +
  theme_minimal()
```





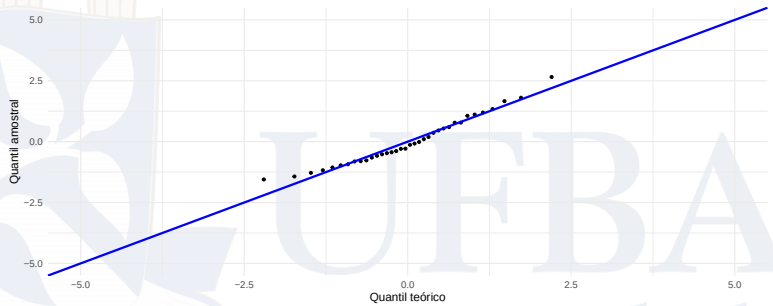
Checando a distribuição normal

Gráfico de probabilidade normal

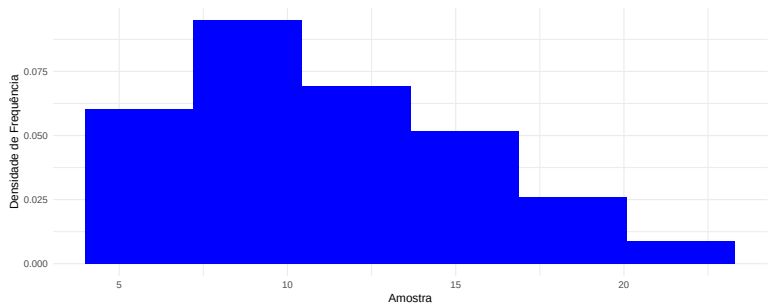
Aplicação: Dados socio-econômicos de 36 funcionários da companhia MB.

```
dados <- read_xlsx('dados.xlsx', sheet = 'companhia_MB')

qq_norm(dados$salario) +
  labs(x = "Quantil teórico", y = "Quantil amostral") +
  theme_minimal()
```




```
hist_sturge(dados$salario) +
  labs(y = "Densidade de Frequência", x = "Amostra") +
  theme_minimal()
```



Estimação intervalar: Normal (media, variância conhecida)

Exemplo simulado

```
media <- 5 # média populacional
dp <- 2 # desvio padrão populacional
m <- 50 # tamanho amostral
amostra <- rnorm(m, mean = media, sd = dp)

conf_norm(amostra, sd = dp)
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level  mean
##      <dbl>    <dbl>      <dbl> <dbl>
## 1     5.10     6.21       0.95  5.66
```

Aplicação

Dados socioeconômicos de 36 funcionários da seção de orçamentos da companhia MB. Vamos assumir que o desvio padrão do salário é $R\$\ 5,00$.

```
dp <- 5 # desvio padrão populacional
dados <- read_xlsx('dados.xlsx', sheet = 'companhia_MB')

with(dados, conf_norm(salario, sd = dp, conf.level = 0.99))
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level  mean
##      <dbl>    <dbl>      <dbl> <dbl>
## 1     8.98     13.3       0.99  11.1
```

Estimação intervalar: Normal (media, variância desconhecida)

Exemplo simulado

```
media <- 5 # média populacional
dp <- 2 # desvio padrão populacional
m <- 50 # tamanho amostral
amostra <- rnorm(m, mean = media, sd = dp)

conf_norm(amostra)
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level  mean
##      <dbl>    <dbl>      <dbl> <dbl>
## 1     4.47     5.74       0.95  5.10
```

Aplicação

Dados socioeconômicos de 36 funcionários da seção de orçamentos da companhia MB.

```
dp <- 5 # desvio padrão populacional
dados <- read_xlsx('dados.xlsx', sheet = 'companhia_MB')

with(dados, conf_norm(salario, conf.level = 0.99))
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level  mean
##      <dbl>    <dbl>      <dbl> <dbl>
## 1     9.04     13.2       0.99  11.1
```

Estimação intervalar: Normal (variância)

Exemplo simulado

```
media <- 5 # média populacional
dp <- 2 # desvio padrão populacional
m <- 50 # tamanho amostral
amostra <- rnorm(m, mean = media, sd = dp)

conf_var_norm(amostra, conf.level = 0.99)
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level   var
##      <dbl>    <dbl>      <dbl> <dbl>
## 1     1.58     4.53       0.99  2.52
```

Aplicação

```
dados <- read_xlsx('dados.xlsx', sheet = 'companhia_MB')

with(dados, conf_var_norm(salario, conf.level = 0.90))
```

```
## # A tibble: 1 x 4
##   lower_ci upper_ci conf_level   var
##      <dbl>    <dbl>      <dbl> <dbl>
## 1     14.8     32.8        0.9  21.0
```