Estatística Computacional

Universidade Federal da Bahia

Gilberto Pereira Sassi

Tópico 6

FDA empírica

Definição 1 (FDA empírica.) A *função de distribuição empírica* é dada por

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

em que $I(X_i \leq x) = \begin{cases} 1, & X_i \leq x, \\ 0, & X_i > x. \end{cases}$

Teorema 1 (Convergência em probabilidade.) Seja $X_1, \ldots, X_n \overset{iid}{\sim} F(\cdot)$. Para todo x, temos que

- $\hat{F}_n(x) \overset{P}{\rightarrow} F(x)$
- $E(F_n(x)) = F(x)$
- $Var(F_n(x)) = \frac{F(x)\cdot(1-F(x))}{n}$
- $EQM = \frac{F(x)\cdot(1-F(x))}{n} \rightarrow 0$

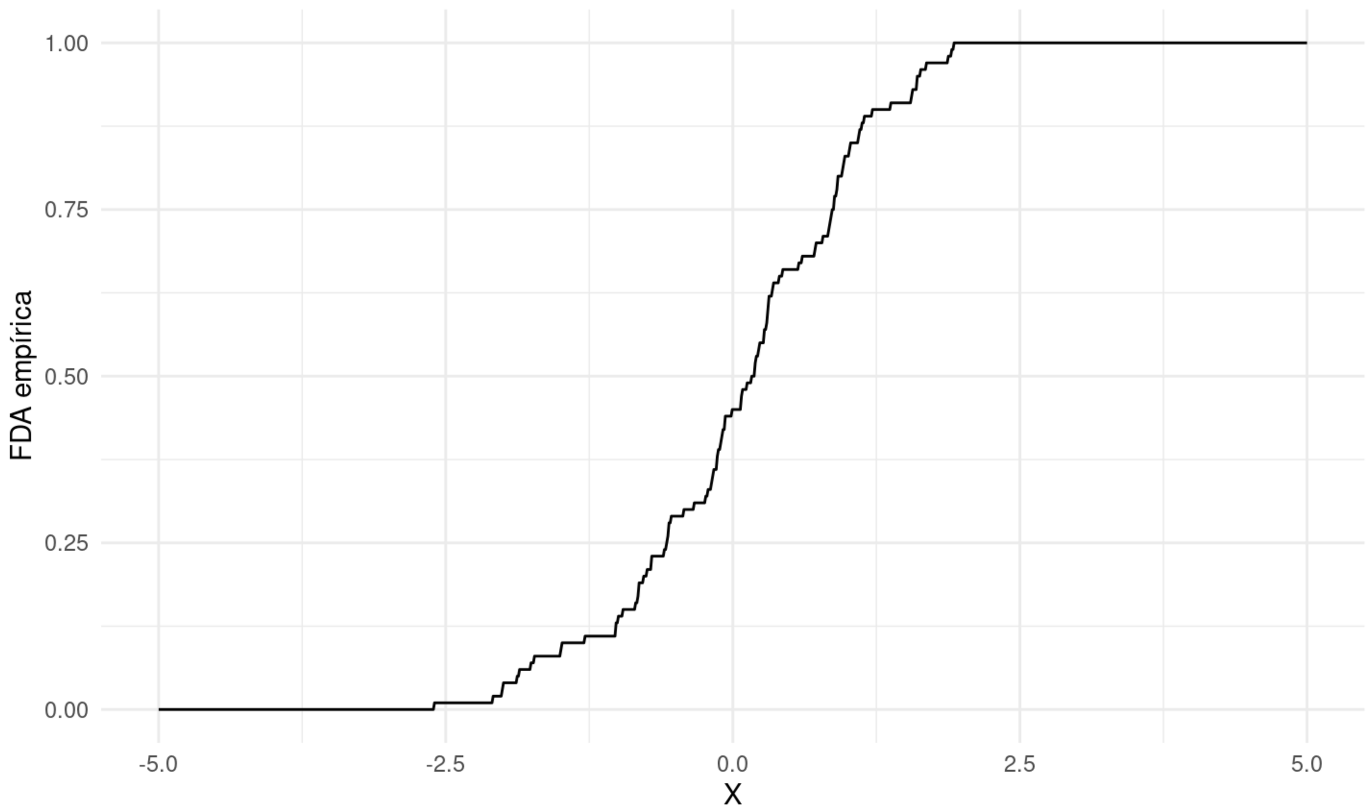
FDA empírica

```r
fda_empirica <- function(amostra) {
  \(x) seq_along(x) |> map_dbl(\(i) (amostra <= x[i]) |> mean())
}

amostra <- rnorm(100)
fda <- fda_empirica(amostra)
x <- seq(from = -5, to = 5, length.out = 1000)
dados <- tibble(x = x, y = fda(x))

ggplot(dados) +
  theme_minimal() +
  geom_line(aes(x = x, y = y)) +
  labs(x = "X", y = "FDA empírica")
```

FDA empírica



Desigualdades e convergência

Teorema 2 (Teorema de Glivenko-Cantelli) Seja $X_1, \dots, X_n \overset{idd}{\sim} F(\cdot)$, então

$$\sup_n |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

Teorema 3 (Desigualdade de Dvoretzky-Kiefer-Wolfowitz.) Seja $X_1, \dots, X_n \overset{iid}{\sim} F(\cdot)$, então para $\epsilon > 0$ temos que

$$P(\sup_x |F_n(x) - F(x)| > \epsilon) \leq 2^{-2 \cdot n\epsilon^2}.$$

Intervalo de confiança

Note que

$$P(\sup_x |F_n(x) - F(x)| > \epsilon) + P(\sup_x |F_n(x) - F(x)| \leq \epsilon) = 1$$

e, usando a desigualdade de Dvoretzky-Kiefer-Wolfowitz, temos que

$$\begin{aligned} P(\sup_x |F_n(x) - F(x)| \leq \epsilon) &= 1 - P(\sup_x |F_n(x) - F(x)| > \epsilon), \\ &\geq 1 - 2^{-2 \cdot n\epsilon^2}, \\ [\sup_x |F_n(x) - F(x)| \leq \epsilon] &\subset [|F_n(x) - F(x)| \leq \epsilon], \end{aligned}$$

logo

$$1 - 2^{2n\epsilon^2} \leq P(\sup_x |F_n(x) - F(x)| \leq \epsilon) \leq P(|F_n(x) - F(x)| \leq \epsilon).$$

Para n fixo e nível de significância α, considere $1 - \alpha = 1 - 2^{2n\epsilon_n^2}$ e, consequentemente, $\epsilon_n = \sqrt{\frac{\log_2(\alpha)}{-2n}}$

Intervalo de confiança

Para α e n fixos, considere $\epsilon_n = \sqrt{\frac{\log_2(\alpha)}{-2n}}$. Então, o intervalo de confiança para $F(x)$ é dado por

$$-\epsilon_n + F_n(x) \leq F(x) \leq \epsilon_n + F_n(x),$$

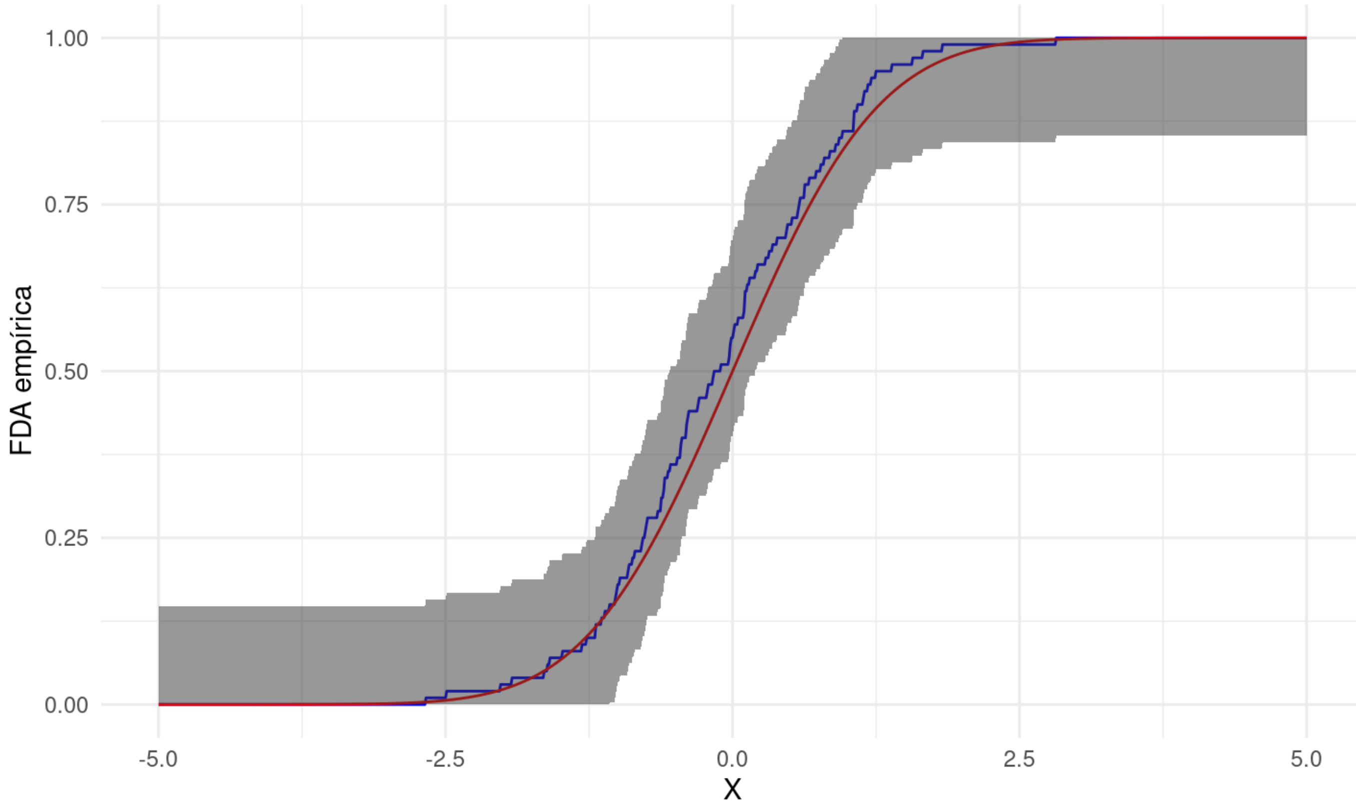
ou seja,

$$IC(F(x), 1-\alpha) = \left(\max\left(-\sqrt{\frac{\log_2(\alpha)}{-2n}} + F_n(x); 0\right); \min\left(\sqrt{\frac{\log_2(\alpha)}{-2n}} + F_n(x); 1\right)\right).$$

Intervalo de confiança

```r
fda_empirica_2 <- function(amostra, sig_level = 0.05) {
   \(x) {
     Fn <- seq_along(x) |> map_dbl(\(i) (amostra <= x[i]) |> mean())
     en <- sqrt(log2(sig_level) / (-2 * length(amostra)))
     list(est = Fn,
          lower = Fn |> map_dbl(\(Fn) max(Fn - en, 0)),
          upper = Fn |> map_dbl(\(Fn) min(Fn + en, 1)))
   }
}
amostra <- rnorm(100)
fda <- fda_empirica_2(amostra)
x <- seq(from = -5, to = 5, by = 0.01)
y <- fda(x)
dados <- tibble(x = x, est = y$est, lower_ic = y$lower, upper_ci = y$upper,
                fda = pnorm(x))
ggplot(dados, mapping = aes(x = x)) +
  theme_minimal() +
  geom_line(aes(y = est), color = "blue") +
  geom_line(aes(y = fda), color = "red") +
  geom_ribbon(aes(ymin = lower_ic, ymax = upper_ci), alpha = 0.5)
```

Intervalo de confiança



Funcional estatístico

Definição 2 (Funcional estatística.) Funcional estatística é uma função $T : P \to \mathbb{R}$, em que P é o conjunto das função de distribuição de probabilidade.

Exemplo 1 (Média e variância.) Alguns exemplos de funcionais estatísticos:

- **Média:** $\mu = \int x dF(x)$
- **Variância:** $\sigma^2 = \int (x - \mu)^2 dF(x)$

Definição 3 (Estimador plug-in) O estimador *plug-in* de um funcional estatístico $\theta = T(F)$ é dado por $\hat{\theta} = T(F_n)$, em que F_n é a função de distribuição acumulada empírica.

Definição 4 (Funcional linear) Se existe uma função $r : \mathbb{R} \to \mathbb{R}$ tal que $T(F) = \int r(x) dF(x)$, em que $T(F)$ é um funcional estatístico, então $T(F)$ é chamado de **funcional linear**.

Funcional estatístico

Teorema 4 O estimador *plug-in* para um funcional linear dado por $\theta = T(F) = \int r(x)dF(x)$ é dado por

$$\hat{\theta} = T(F_n) = \int r(x)dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i).$$

Funcional estatístico

Exemplos

Exemplo 2 (Média.) $\hat{\mu} = \int x d_F n(x) = \frac{1}{n} \sum_{i=1}^{n} X_i$

```
amostra <- rnorm(1000)
media_hat <- seq_along(amostra) |>
  map_dbl(\(i) amostra[i] / length(amostra)) |>
  sum()
media_hat
```

```
## [1] 0.01597413
```

Exemplo 3 (Variância) $\hat{\sigma}^2 = \int (x-\mu)^2 dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$

```
amostra <- rnorm(1000, mean = 1)
var_hat <- seq_along(amostra) |>
  map_dbl(\(i) (amostra[i] - 1)^2 / length(amostra)) |>
  sum()
var_hat
```

```
## [1] 1.000338
```