Inferência aproximada: o valor de uma premissa

Motivação

Em Inferência Estatística, em geral temos como ponto de partida um modelo estatístico e um conjunto de dados, e nossa tarefa é produzir inferências ótimas. A otimalidade dos procedimentos e estimativas obtidos é contingente na adequação das premissas feitas à realidade. Em Modelagem Estatística, vamos primeiro desafiar a ideia de um modelo fixo e construir vários modelos para o mesmo fenômeno, buscando sempre confrontar os ajustes obtidos aos dados no sentido de checar a adequação das premissas feitas. Nesta primeira lição, veremos o que pode ser feito sob poucas premissas acerca do processo gerador dos dados (*data generating process*, DGP).

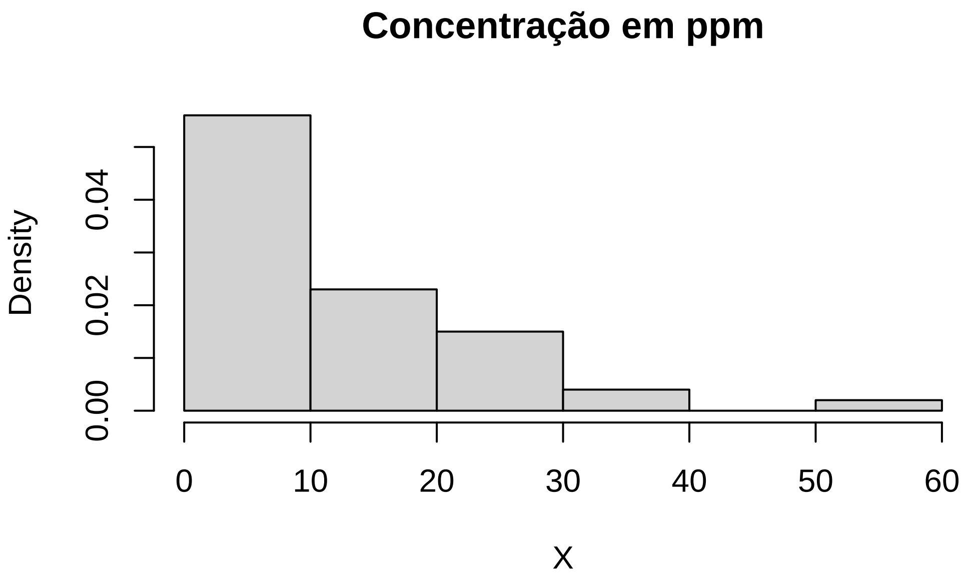
Os dados

Vamos analizar medições da concentração (em partes por milhão, ppm) de um composto químico em $n = 100$ amostras de bateladas de um determinado produto.

```
conc <- read.csv("../data/chem.csv", header = FALSE)$V1
```

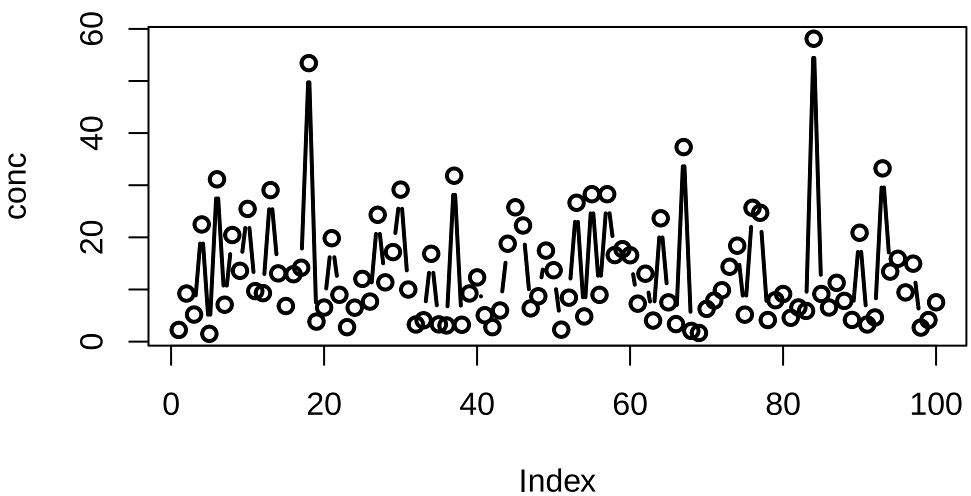
Vamos explorar um pouco os dados plotando um histograma.

```
hist(conc, probability = TRUE,
     main = "Concentração em ppm",
     xlab = "X")
```



Será que os dados apresentam alguma tendência em relação ao número da batelada (índice da observação)?

```
plot(conc, type = "b", lwd = 2)
```



Estatísticas descritivas

Vamos olhar quartis, amplitude e desvio padrão:

```r
summary(conc)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.490   5.185   9.230  12.975  17.530  58.110 
```

```r
sd(conc)
```

```
[1] 10.58646
```

Inferência

Vamos raciocinar sobre algumas quantidades, a saber a média ($\bar{X}_n$) e a mediana ($\hat{M}$) amostrais. Em particular, vamos pensar sobre como construir intervalos de confiança para essas quantidades.

```r
## Ver exercícios ao final
med_app <- function(x, gamma = 0.95){
  ## Aproximação normal usando o método Delta.
  dens <- density(x)
  app.pdf <- approxfun(dens)
  med.hat <- median(x)
  n <- length(x)
  sd.approx <- 1/(4 * n * (app.pdf(med.hat))^2)
  pars <- c(med.hat, sqrt(sd.approx))
  return(
  c(med.hat, qnorm(p = c(1 - gamma, 1 + gamma)/2,
        mean = pars[1], sd = pars[2]))
  )
}
med_np <- function(x, gamma = 0.95){
  ## método não-paramétrico, baseado na binomial
  med.hat <- median(x)
  return(
    c(med.hat, sort(x)[qbinom(p = c(1 - gamma, 1 + gamma)/2, size = length(x), prob = 0.5)])
  )
}
```

Agora, vamos aplicar estas funções aos nossos dados. Primeiro, a média amostral:

```r
Nivel <- 0.95
xbar <- mean(conc)
c(xbar, xbar + c(-1, 1) * qnorm(p = (1 + Nivel)/2) * sd(conc)/length(conc))
```

```
[1] 12.97450 12.76701 13.18199
```

Depois, a mediana:

```r
med_np(conc)
```

```
[1]  9.23  7.68 12.32
```

```r
med_app(conc)
```

```
[1]  9.230000  7.188329 11.271671
```

Investigando a função de distribuição empírica

Um ótimo descritor de uma distribuição de probabilidade é a função de distribuição acumulada (f.d.a.) também chamada de *cumulative distribution function*, CDF:

$$F(x) := \Pr(X \leq x).$$

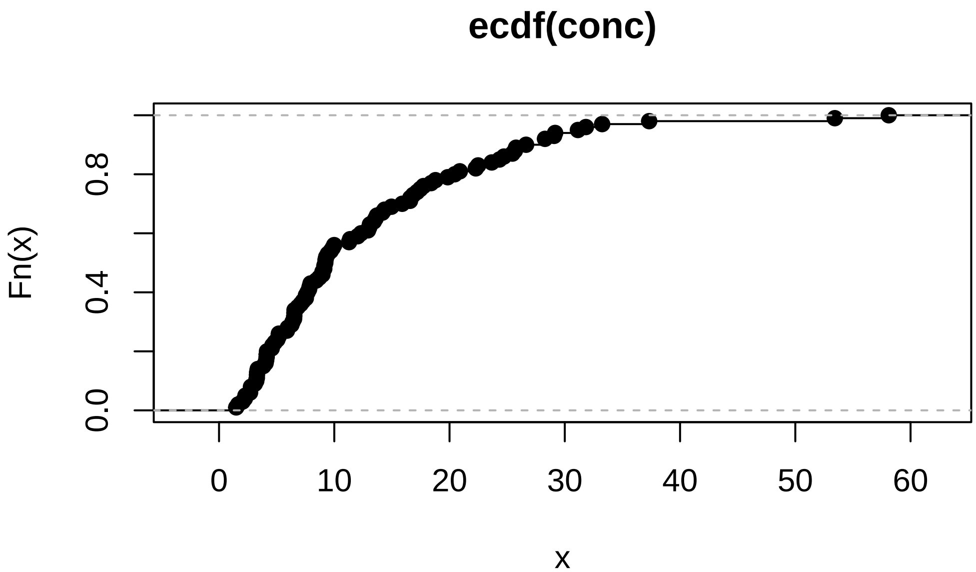
Como discutido nos exercícios abaixo, podemos aproximar F a partir da amostra através do estimador

$$Y_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(X_i \leq x\right)$$

Vamos ver como ficam as estimativas usando os dados sob análise:

```r
fda.empirica <- ecdf(conc)

plot(fda.empirica)
```



Desta forma, se o alvo inferencial é $\Pr(X \leq 30)$, podemos obter uma estimativa fazendo

```
( P30.hat <- fda.empirica(30) )
```

```
[1] 0.94
```

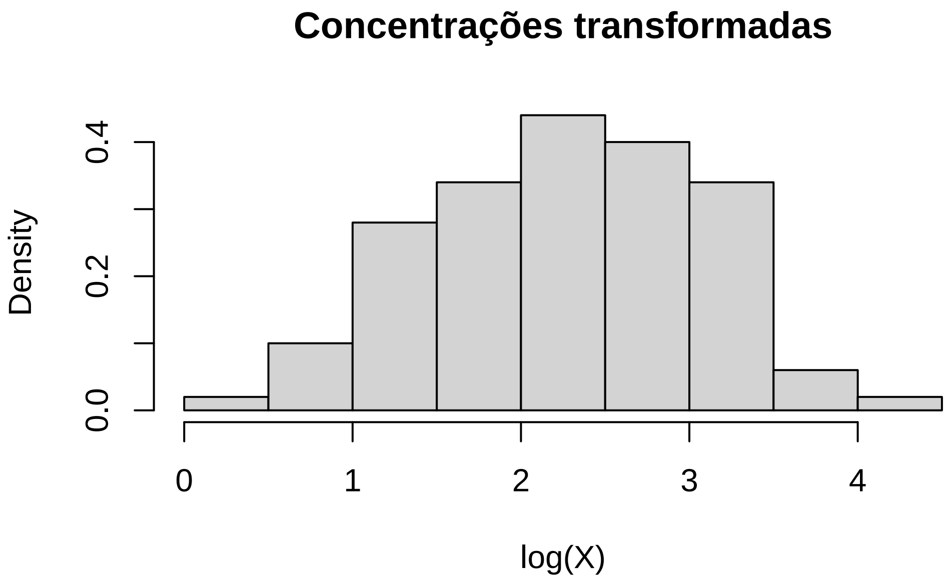
Para quantificar a incerteza, podemos fazer (porquê?)

```
P30.hat + c(-1, 1) * qnorm(p = (1 + Nivel)/2) * (P30.hat * (1-P30.hat))/length(conc)
```

```
[1] 0.9388946 0.9411054
```

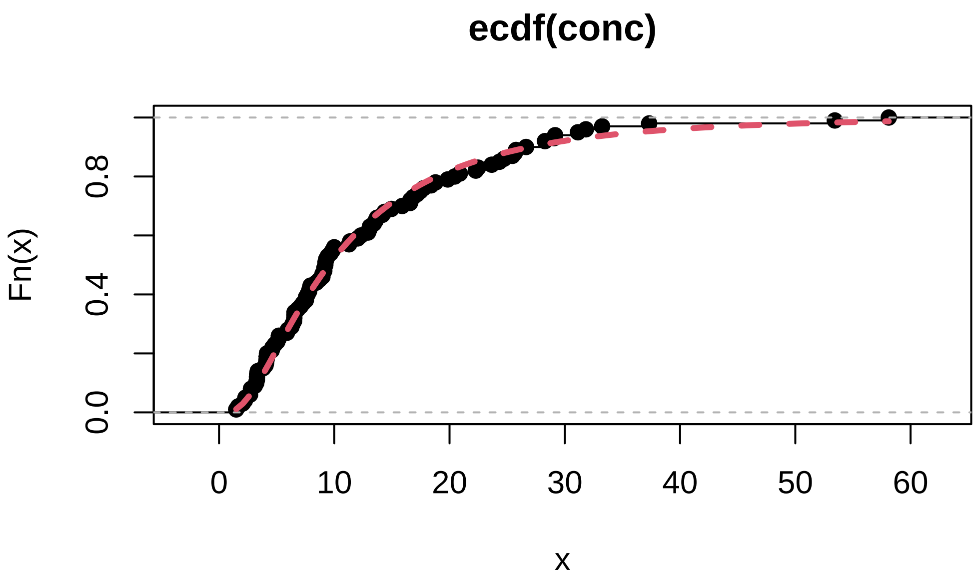
Para finalizar, vamos olhar o que acontece se fizermos (i) uma transformação e (ii) uma premissa paramétrica:

```
lg.conc <- log(conc)
hist(lg.conc, probability = TRUE,
     main = "Concentrações transformadas",
     xlab = "log(X)")
```




```
mu.hat <- mean(lg.conc)
sd.hat <- sd(lg.conc)

plot(fda.empirica)
curve(plnorm(x, meanlog = mu.hat, sdlog = sd.hat), min(conc), max(conc), lwd = 3, lty = 2, co
```



Exercícios de fixação

1. Tome $X_1, X_2, \ldots, X_n$ uma amostra de uma distribuição conjunta F_n. Sejam $\theta_i := E[X_i]$ e $v_i := \mathrm{Var}(X_i)$. Considere a média amostral $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Deduza que para $\varepsilon > 0$:
$$\Pr\left(\left|\bar{X}_n - \frac{1}{n}\sum_{i=1}^n \theta_i\right| \geq \varepsilon\right) \leq \frac{\sum_{i=1}^n v_i + 2\sum_{i<j}\mathrm{Cov}(X_i, X_j)}{(n\varepsilon)^2},$$
e argumente sobre o que acontece à medida que $n \to \infty$ sob as premissas de (a) independência (b) indentidade de distribuição. O que precisamos assumir sobre v_i? E sobre as covariâncias?

2. **O método Delta**. Suponha que $Y_1, Y_2, \ldots$ é uma sequência de variáveis aleatórias i.i.d. para as quais vale um teorema central do limite, isto é,
$$\sqrt{n}\frac{(Y_n - \mu)}{\sigma} \implies \mathrm{Normal}(0, 1),$$
onde $\mu := E[Y_1]$ e $\sigma := \sqrt{\mathrm{Var}(Y_1)}$ e a convergência é em distribuição. Tome $g : \mathcal{Y} \to \mathbb{R}$ uma função[1] tal que $g'(\mu) \neq 0$. Prove que
$$\sqrt{n}\frac{(g(Y_n) - g(\mu))}{\sigma|g'(\mu)|} \implies \mathrm{Normal}(0, 1).$$
Dica: use o teorema de Taylor, o teorema do mapeamento contínuo e o teorema de Slutsky.

3. Tome $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma distribuição com f.d.a. (cdf) comum F. Para $x \in \mathbb{R}$, defina $Y_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i \leq x)$. Mostre que

   a. $E[\mathbb{I}(X_i \leq x)] = F(x)$ e que $\mathrm{Var}(\mathbb{I}(X_i \leq x)) = F(x)[1 - F(x)]$.
   b. $\sqrt{(n)}(Y_n(x) - F(x)) \implies \mathrm{Normal}(0, F(x)[1 - F(x)])$.
   c. Considere a função $g(t) = F^{-1}(t)$ para $t \in (0, 1)$, isto é a inversa generalizada de F. Calcule $g'(t)$ em termos de F e da densidade f.
   d. Finalmente, use o resultado do item 2 para deduzir que
   $$\sqrt{n}\left(F^{-1}(Y_n(x)) - x\right) \implies \mathrm{Normal}\left(0, \frac{p(1-p)}{[f(x)]^2}\right),$$
   para $p = F(x)$.

4. Use o resultado anterior para construir um intervalo de confiança de $\gamma \times 100\%$ aproximado para a mediana amostral.

Dica: $f(x)$ pode ser bem estimada em qualquer ponto x usando o método do histograma. Seja $\{B_k := (t_k, t_{k+1}) : t_k = t_0 + hk, k \in \mathbb{Z}\}$ e defina para $x \in B_k$ uma função constante em B_k $\hat{f}(x, t_0, h) := \frac{v_k}{nh}$, onde v_k é o número de observações que estão no intervalo B_k. Sob condições de regularidade[2], temos que quando $h \to 0$, $\hat{f}(x, t_0, h) \to f(x)$, isto é, o estimador

[1] mensurável.
[2] Basicamente queremos que $nh \to \infty$

da densidade é consistente[3]. No R, podemos fazer

```
amostra <- rnorm(100)
dens <- density(amostra)
app.pdf <- approxfun(dens)
```

para obter uma pdf aproximada usando como amostra um conjunto de v.a.s. normal padrão, por exemplo.

5. (**Desafio**) A discussão dos itens anteriores supõe amostras grandes, para as quais faça sentido falar em teorema central do limite. Agora vamos estudar uma maneira de construir um intervalo de confiança para a mediana que seja válido para amostras finitas. Suponha uma amostra aleatória de tamanho n ímpar de uma distribuição F e considere sua versão ordenada:
$$X_{(1)}, X_{(2)}, \ldots, X_{\left(\frac{n+1}{2}\right)}, \ldots, X_{(n)}.$$
Seja $\tilde{X}$ a mediana de F, i.e., $\tilde{X} = \inf\{x : F(x) \geq 1/2\}$. Defina $\pi_m = \Pr(X \leq \tilde{X})$, onde X tem f.d.a. F.

   a. Mostre que
   $$\Pr(X_{(i)} > \tilde{X}) = \sum_{j=0}^{i-1} \binom{n}{j} \pi_m^j \left(1 - \pi_m\right)^{n-j}.$$

   b. Argumente que $\pi_m \geq 1/2$. Em seguida, escreva $\pi_m = 1/2 + \varepsilon$ e mostre que para $2j \leq n$ vale que
   $$\pi_m^j \left(1 - \pi_m\right)^{n-j} \leq 2^{-n},$$
   e que, então,
   $$\Pr(X_{(i)} > \tilde{X}) \leq 2^{-n} \sum_{j=0}^{i-1} \binom{n}{j}.$$

c. Use os resultados anteriores para mostrar que *sempre* existem $l \leq u \in \{1, \ldots, n\}$ tais que
$$\Pr\left(X_{(l)} \leq \tilde{X} \leq X_{(u)}\right) \geq 2^{-n} \sum_{j=0}^{i-1} \binom{n}{j}.$$

d. Para terminar, use os resultados anteriores para construir um intervalo de confiança de $\gamma \times 100\%$ para $\tilde{X}$.

Referências

- Seção 6.4.3 de Evans & Rosenthal (2023).
- Capítulo 5 de Hahn & Meeker.

[3] E assintoticamente não-viesado. Para mais detalhes, veja essas notas