Aula 7: Estimação Densidade de Probabilidade

Prof. Dr. Eder Angelo Milani

17/05/2023

Estimação da densidade de probabilidade

A Estimação da densidade é uma coleção de métodos para obter uma estimativa da densidade de probabilidade, como uma função da amostra observada.

Na abordagem paramétrica, adota-se algumas suposições sobre a forma funcional paramétrica dos dados. Entretanto, a suposição atribuída para a forma funcional pode ser muito restritiva ou até inadequada.

Nas aulas anteriores, utilizamos por diversas vezes o histograma para estimar a densidade, este é um exemplo clássico de estimador não-paramétrico para a densidade. A estimação da densidade utilizando métodos não-paramétrico fornece uma ferramenta flexível e poderosa para a visualização, exploração e análise dos dados.

Um problema de estimação da densidade requer uma abordagem não-paramétrica se não tivermos informações sobre a distribuição-alvo, além dos dados observados.

Histogramas

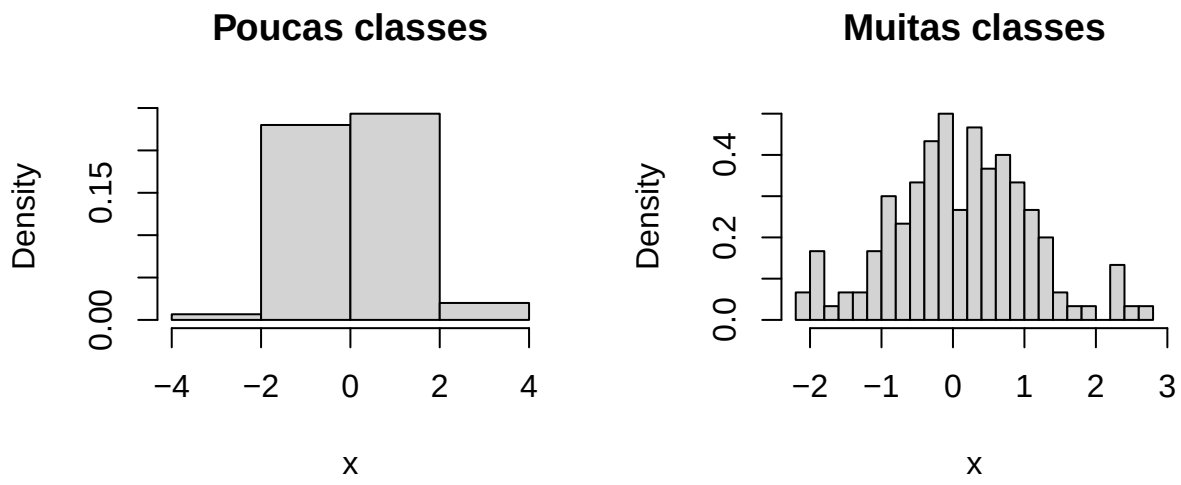
O histograma é apresentado em cursos elementares de estatística básica, além de estar disponível na maioria dos pacotes estatísticos. O histograma é a estimativa da densidade mais amplamente utilizada em estatística descritiva. No entanto, mesmo nos projetos de análise de dados elementares, deparamos com questões complicadas, tais como: determinar o melhor número de categorias, os limites e a largura dos intervalos de classe ou como lidar com larguras de intervalo de classe desiguais. Em muitos pacotes, essas decisões são tomadas automaticamente, mas às vezes produzem resultados indesejáveis.

Com o software R, o usuário tem controle sobre várias opções. O histograma é uma aproximação constante por partes da função densidade. Como os dados, em geral, estão contaminados por ruídos, o estimador que apresenta muitos detalhes (se encaixando mais de perto aos dados) não é necessariamente o "melhor". A escolha da largura das barras para um histograma é uma opção de parâmetro de suavização. Uma largura da barra muito estreita pode sub-suavizar os dados, apresentando muitos detalhes, enquanto uma largura da barra mais ampla pode super-suavizar os dados, obscurecendo características importantes. Várias regras são comumente aplicadas sugerindo uma escolha ideal da largura da barra.

A seguir é apresentado dois histogramas, responda:

(i) você acredita que os dois histogramas foram gerados da mesma amostra?

(ii) você acredita que os dados apresentam distribuição normal?



Suponha que a amostra aleatória $X_1, \ldots, X_n$ é observada. Para construir um histograma de frequência ou probabilidade da amostra, os dados devem estar classificados em categorias e a operação de categorização é determinada pelos limites dos intervalos de classe. Embora, em princípio, qualquer limite de classe possa ser usado, algumas escolhas são mais razoáveis do que outras em termos de qualidade da informação sobre a densidade populacional. Entre as regras comumente aplicadas para determinar os limites dos intervalos de classe de um histograma estão a regra de Sturges, a regra de referência normal de Scott, a regra Freedman-Diaconis (FD) e várias modificações dessas regras.

Adotando que os intervalos de classe são todos de comprimento h, o histograma de probabilidade estimado baseado em uma amostra de tamanho n é

$$\hat{f}(x) = \frac{\nu_k}{nh},\ t_k \leq x < t_{k+1},$$

sendo que ν_k é o número de pontos amostrais no intervalo de classse $[t_k; t_{k+1})$. Se a largura da barra for exatamente 1, a estimativa da densidade é a frequência relativa da classe que contém o ponto x.

Regra de Sturges

A regra de Sturges tende a suavizar demais os dados e a regra de Scott ou FD sejam geralmente preferíveis, a regra de Sturges é o padrão em muitos pacotes estatísticos.

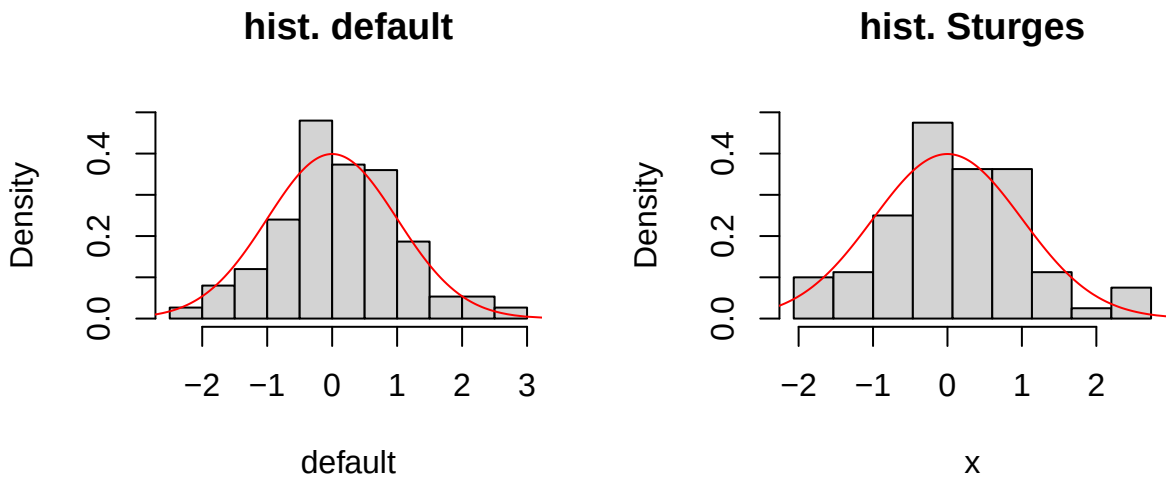
De acordo com Sturges, a largura ideal dos intervalos de classe é dado por

$$\frac{R}{1 + log_2 n}$$

sendo que R é a amplitude amostral. O número de barras depende apenas do tamanho da amostra n e não da distribuição. Essa escolha do intervalo de classe é projetada para dados amostrados de populações simétricas e unimodais, mas não é uma boa escolha para distribuições assimétricas ou com mais de uma moda. Para amostras grandes, a regra de Sturges tende a ser excessivamente suave.

```r
set.seed(2023)
n <- 150
x <- rnorm(n)
nclasses <- ceiling(1 + log2(n)) #quantidade de classes
larg <- diff(range(x) / nclasses) #largura das classes
breaks <- min(x) + larg * 0:nclasses # limites das classes

par(mfrow=c(1, 2))

h.default <- hist(x, freq = FALSE, xlab = "default", main = "hist. default", ylim=c(0,0.5))
z <- qnorm(ppoints(1000))
lines(z, dnorm(z), col="red")

hist_sturges <- hist(x, breaks = breaks, freq = FALSE, main = "hist. Sturges", ylim=c(0,0.5))
lines(z, dnorm(z), col="red")
```

Exemplo 1) Comparar o histograma obtido a partir da função `hist` com o obtido utilizando a regra de Sturges.




```r
print("Detalhes do histograma Sturges")
```

```
## [1] "Detalhes do histograma Sturges"
```

```r
cat("largura= ", larg, "\n")
```

```
## largura=  0.5334107
```

```r
cat("limite dos intervalos","\n", breaks, "\n")
```

```
## limite dos intervalos
##  -2.065457 -1.532046 -0.9986357 -0.4652251 0.06818561 0.6015963 1.135007 1.668418 2.201828 2.735239
```

Obs.: no R, o *default* da função `hist` é uma modificação da regra de Sturges

Referência Normal de Scott

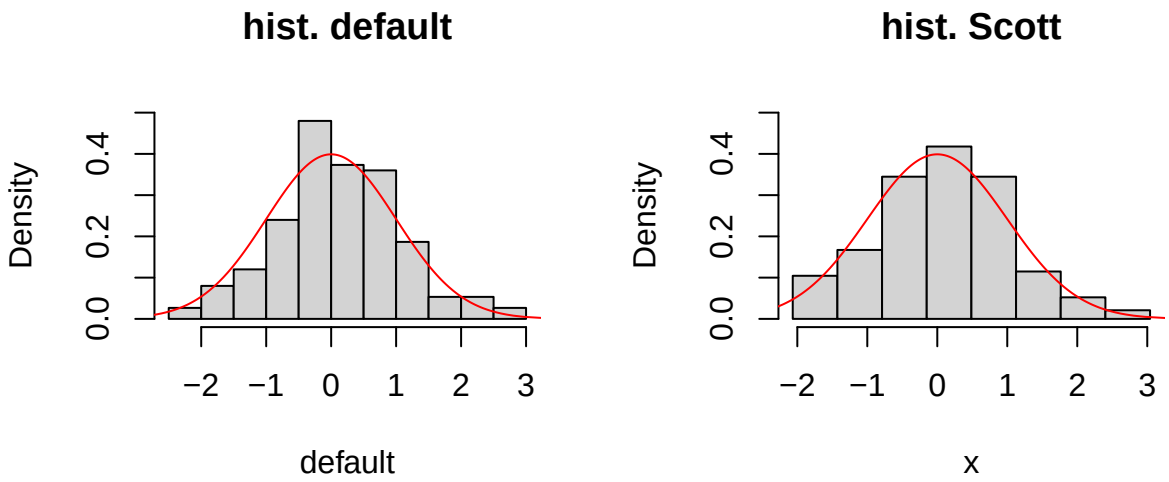
A regra de referência normal de Scott, que é calibrada para uma distribuição normal com variância σ^2, especifica uma largura de barra

$$\hat{h} = 3,49\hat{\sigma}n^{-1/3}$$

sendo que $\hat{\sigma}$ é uma estimativa do desvio padrão da população.

```
set.seed(2023)
n <- 150
x <- rnorm(n)
larg=3.49*sd(x)*n^(-1/3)
min_=min(x)
max_=max(x)
nclasses <- ceiling((max_-min_)/larg) #quantidade de classes
breaks <- min_ + larg * 0:nclasses # limites das classes

par(mfrow=c(1, 2))


h.default <- hist(x, freq = FALSE, xlab = "default", main = "hist. default", ylim=c(0,0.5))
z <- qnorm(ppoints(1000))
lines(z, dnorm(z), col="red")

hist_scott <- hist(x, breaks = breaks, freq = FALSE, main = "hist. Scott", ylim=c(0,0.5))
lines(z, dnorm(z), col="red")
```

Exemplo 2) Comparar o histograma obtido a partir da função `hist` com o obtido utilizando a regra de Scott.



```r
print("Detalhes do histograma Scott")
```

```
## [1] "Detalhes do histograma Scott"
```

```r
cat("largura= ", larg, "\n")
```

```
## largura=  0.638206
```

```r
cat("limite dos intervalos","\n", breaks, "\n")
```

```
## limite dos intervalos
##  -2.065457 -1.427251 -0.789045 -0.150839 0.487367 1.125573 1.763779 2.401985 3.040191
```

Regra de Freedman-Diaconis (FD)

A regra de Scott pertence a uma classe de regras que seleciona a largura da barra ideal de acordo com uma fórmula $\hat{h} = T_n^{-1/3}$, sendo que T_n é uma estatística. Estas regras estão relacionadas com o fato de que a taxa ótima de decaimento da largura da barra em relação às normas L_p é $n^{-1/3}$. Para a regra de FD, a estatística T_n é o dobro do intervalo interquartil da mostra. Isto é

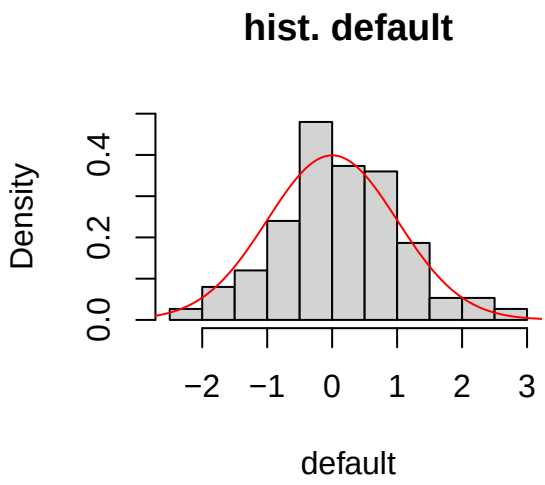
$$\hat{h} = 2(IQR)n^{-1/3}$$

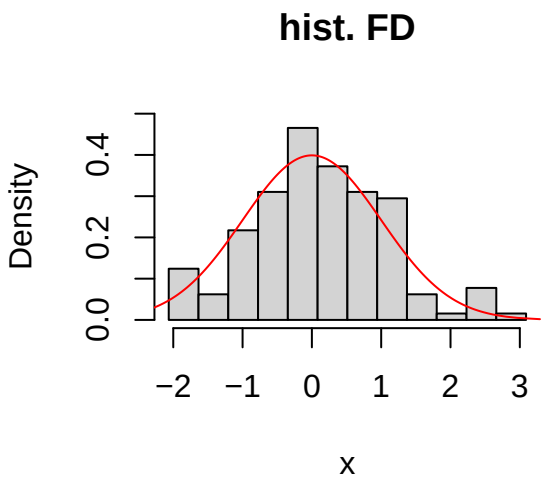
sendo que IQR é o intervalo interquartil da amostra.

Obs.: o IQR é menos sensível que o desvio padrão em relação a outliers nos dados.

```r
set.seed(2023)
n <- 150
x <- rnorm(n)
Q1 <- quantile(x, probs = 0.25)
Q3 <- quantile(x, probs = 0.75)
larg=2*(Q3-Q1)*n^(-1/3)
min_=min(x)
max_=max(x)
nclasses <- ceiling((max_-min_)/larg) #quantidade de classes
breaks <- min_ + larg * 0:nclasses # limites das classes

par(mfrow=c(1, 2))

h.default <- hist(x, freq = FALSE, xlab = "default", main = "hist. default", ylim=c(0,0.5))
z <- qnorm(ppoints(1000))
lines(z, dnorm(z), col="red")

hist_fd <- hist(x, breaks = breaks, freq = FALSE, main = "hist. FD", ylim=c(0,0.5))
lines(z, dnorm(z), col="red")
```

Exemplo 3) Comparar o histograma obtido a partir da função `hist` com o obtido utilizando a regra FD.






```
print("Detalhes do histograma FD")
```

```
## [1] "Detalhes do histograma FD"
```

```
cat("largura= ", larg, "\n")
```

```
## largura=  0.4296355
```

```
cat("limite dos intervalos","\n", breaks, "\n")
```

```
## limite dos intervalos
##  -2.065457 -1.635822 -1.206186 -0.7765507 -0.3469152 0.08272024 0.5123557 0.9419912 1.371627 1.80126
```

Exercícios

1 - Calcule a quantidade de classes a partir dos três métodos de construção de histogramas visto na aula de hoje, considerando: a distribuição normal padrão, tamanho amostral de 50, 100, 200 e 500, e 100 repetições. Apresentar resultado em uma tabela.

2 - Calcule a quantidade de classes a partir dos três métodos de construção de histogramas visto na aula de hoje, considerando: a distribuição exponencial($\lambda = 1$), tamanho amostral de 50, 100, 200 e 500, e 100 repetições. Apresentar resultado em uma tabela.

3 - Calcule a quantidade de classes a partir dos três métodos de construção de histogramas visto na aula de hoje, considerando: a distribuição qui-quadrado com 1 grau de liberdade, tamanho amostral de 50, 100, 200 e 500, e 100 repetições. Apresentar resultado em uma tabela.

Obs.: aqui pode ser utilizado `rnorm`, `rexp` e `rchisq`.