4 Teoria da Probabilidade

Apresentam-se neste capítulo conceitos de probabilidade e de estimação de funções densidade de probabilidade necessários ao desenvolvimento e compreensão do modelo proposto (capítulo 5).

A probabilidade é uma medida de incerteza associada ao resultado futuro de um sorteio aleatório. É importante perceber que, uma vez ocorrido o sorteio, do ponto de vista probabilístico não há mais dúvida sobre o resultado e, portanto, o valor da probabilidade tem um valor preciso [25]. Por exemplo, quando uma moeda é lançada, devido ao conhecimento parcial sobre sua estrutura física e suas condições iniciais, não é possível prever com exatidão se o resultado será cara ou coroa. Porém, uma vez pousada, não há mais dúvida quanto ao resultado: ou é cara ou é coroa.

4.1. Definição

Seja o universo Ω de todos os possíveis eventos elementares; por definição, a probabilidade *P* de um evento *E*, denotada por *P*(*E*), deve seguir os 3 axiomas de Kolmogorov [26]:

i) Para qualquer evento *E*, tem-se $P(E) \geq 0$. Isto é, a probabilidade de um evento é um número real não negativo.

ii) $P(\Omega) = 1$: a probabilidade de todos os eventos possíveis é um; ou mais especificamente, não há evento elementar fora do universo Ω.

iii) Todo conjunto de eventos incompatíveis enumeráveis E1, E2,..., satisfaz a $P(E_1 \cup E_2 \cup \ldots) = \sum P(E_i)$. Ou seja, a probabilidade de um conjunto de eventos formado a partir da união de eventos disjuntos é a soma das probabilidades destes eventos. Este axioma também é conhecido como a propriedade da aditividade.

A partir destes axiomas, são enunciadas as seguintes propriedades:

- Para qualquer evento *E*, tem-se $0 \leq P(E) \leq 1$. Ou seja, a probabilidade é um número entre 0 e 1.
- Para qualquer evento *E*, define-se o evento contrário, $\overline{E}$, por $P(\overline{E}) = 1 - P(E)$.
- $P(\varnothing) = 0$.
- Para quaisquer dois eventos *A* e *B*, tem-se $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

4.2. Distribuição de probabilidade

Uma função distribuição de probabilidade é uma função que associa probabilidades a eventos e pode ser classificada em discreta ou contínua. No primeiro caso, a função é definida para um conjunto discreto e contável, tal como um subconjunto dos números inteiros; no segundo caso, a distribuição possui uma função definida para um conjunto contínuo, como, por exemplo, um subconjunto dos números reais.

Uma forma de definir uma função distribuição de probabilidade, *F*(*x*), é por meio de uma função densidade de probabilidade (*pdf*), *f*(*x*), conforme o exemplo da equação (4.1) para o caso contínuo:

$$F(x) = \Pr[X \leq x] = \int_{-\infty}^{x} f(X)\, dX \tag{4.1}$$

A partir da função densidade de probabilidade, também é possível expressar a probabilidade de obter um valor no intervalo [*c*, *d*]:

$$P[c,d] = \int_{c}^{d} f(x)\, dx \tag{4.2}$$

4.2.1. A Distribuição Normal

Entre as diversas distribuições de probabilidade contínuas, a distribuição Normal se destaca por modelar vários fenômenos naturais, entre os quais a incerteza de medição [13]. A distribuição Normal é definida para $-\infty < x < +\infty$ por sua *pdf*:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \tag{4.3}$$

onde μ representa a média e σ^2 a variância.

4.2.2. Intervalo de Confiança

No caso da incerteza de medição, é comum que ela seja dada, simplesmente, como um intervalo em torno do resultado de uma medição, com o qual se espera abranger uma grande fração da distribuição de valores, que poderiam razoavelmente ser atribuídos ao mensurando [13]. Sendo assim, a incerteza de medição não é, necessariamente, dada como um múltiplo de um desvio padrão.

Para uma grandeza *z* descrita por uma distribuição normal, com média μ_z e desvio padrão σ, o fator de abrangência k_p fornece o intervalo $\mu_z \pm k_p\sigma$ que corresponde ao intervalo de confiança com um nível de confiança *p*. Valores típicos para níveis de confiança são 90, 95 e 99 por cento, com fatores de abrangência 1,64; 1,96 e 2,58; respectivamente [13]. É comum que o nível de confiança seja expresso pelo valor (1 - α) (onde este valor é um número fixo, positivo e menor do que 1), correspondente à probabilidade associada com um intervalo de confiança.

4.3. Métodos de Estimação de Probabilidade

Quando a função densidade de uma quantidade aleatória **x** não é conhecida, uma estimativa desta densidade pode ser obtida utilizando-se amostras provenientes de *n* observações $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ de **x**. Os métodos de estimação podem ser classificados como paramétricos ou não-paramétricos: no primeiro caso, um vetor de parâmetros de uma função é estimado, enquanto que, no segundo caso, a função $p(\mathbf{x})$ é estimada sem que nenhum modelo específico seja adotado. A Figura 14 apresenta a taxonomia dos métodos de estimação de funções de densidade de probabilidade.

A estimação da densidade de probabilidade através de métodos paramétricos supõe que as formas das funções de densidade de probabilidade estudadas são conhecidas.

Contudo, as fórmulas paramétricas usuais nem sempre se ajustam nas densidades encontradas na prática. Além disso, a maioria das densidades paramétricas clássicas é unimodal (têm um único máximo), enquanto que muitos dos problemas práticos envolvem densidades multimodais.

Por outro lado, métodos não-paramétricos podem ser utilizados com distribuições arbitrárias e sem a suposição que as formas das densidades estudadas sejam conhecidas.

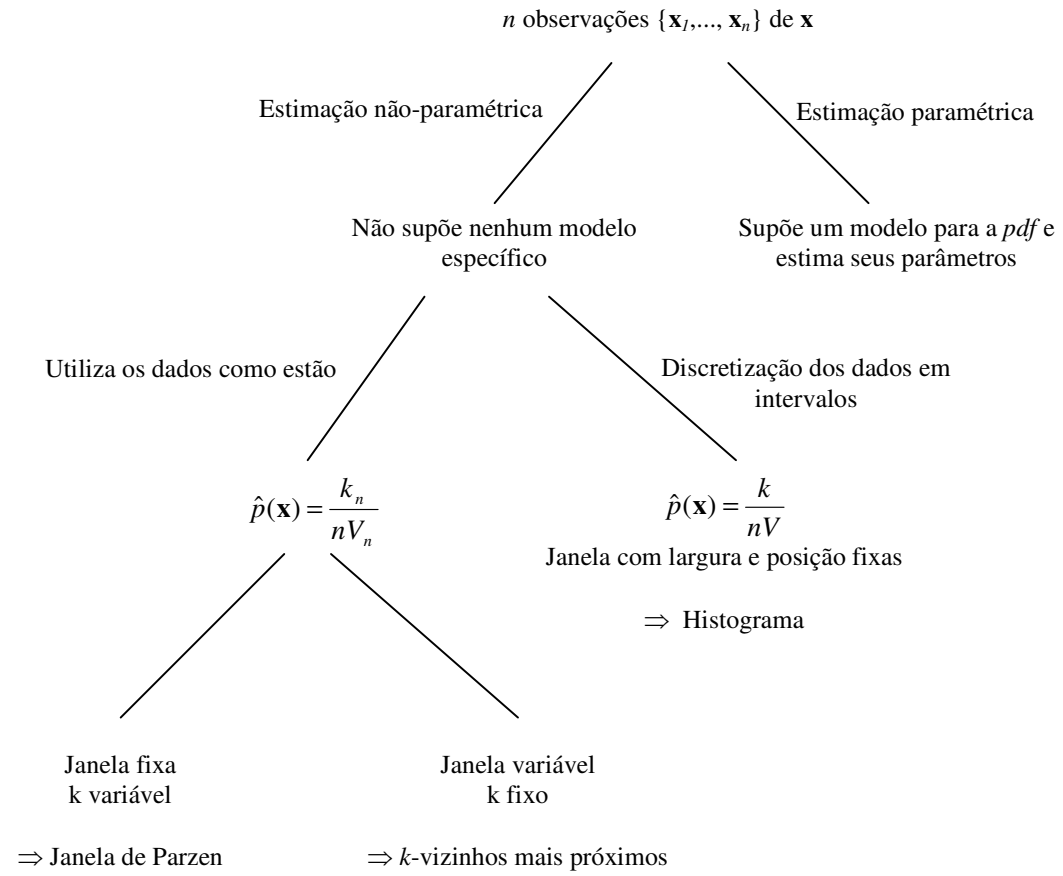

Figura 14 - Classificação dos métodos de estimação de densidade de probabilidade

4.4. Métodos de Estimação Não-Paramétricos

4.4.1. Definição

As técnicas não-paramétricas fundamentais se baseiam no fato de que a probabilidade *P* de que um vetor **x** pertença à região ***R*** é dada pela equação:

$$P = \int_R p(\mathbf{x}) d\mathbf{x} \tag{4.4}$$

Conseqüentemente P é uma versão suavizada da função de densidade $p(\mathbf{x})$ e assim é possível estimar este valor suavizado de p através da estimação da probabilidade P.

Sejam n amostras $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ independentes e identicamente distribuídas de acordo com a distribuição $p(\mathbf{x})$. A probabilidade de que k das n amostras caiam na região $\boldsymbol{R}$ é dada pela lei binomial:

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \tag{4.5}$$

cujo valor esperado de k é $E[k] = nP$ e a melhor estimativa para P é $\hat{P} = \frac{k}{n}$.

Considerando que $p(\mathbf{x})$ é contínua e que a região $\boldsymbol{R}$ é suficientemente pequena de modo que $p(\mathbf{x})$ não varia muito dentro dela, pode-se escrever:

$$\int_R p(\mathbf{x}) d\mathbf{x} \approx p(\mathbf{x}) V \tag{4.6}$$

onde $\mathbf{x}$ é um ponto dentro de $\boldsymbol{R}$ e V é o volume da região $\boldsymbol{R}$. Combinando as equações (4.4), (4.5) e (4.6), a estimativa para $p(\mathbf{x})$ é:

$$p(\mathbf{x}) \approx \frac{k/n}{V} \tag{4.7}$$

4.4.2. Histograma

O histograma é o estimador de densidade mais antigo e mais utilizado para representar e observar dados unidimensionais. A construção de um histograma consiste em dividir um intervalo de referência $\Omega = [x_{\min}, x_{\max}]$ em K células (ou compartimentos) C_k e contar o número a_k de observações pertencentes a cada célula C_k. O número a_k é o acumulador associado à célula C_k. Seja χ_{C_k} a função característica de C_k:

$$a_k = \sum_{i=1}^{n} \chi_{C_k}(x_i)$$

Quando todas as células do histograma têm a mesma largura, é dito que o histograma é uniforme ou regular. A largura de cada célula, Δ, mais comum é:

$$\Delta = \frac{x_{\max} - x_{\min}}{K}$$

Uma probabilidade empírica $P(C_k)$ pode ser associada a cada célula C_k:

$$P(C_k) = \frac{a_k}{n}$$

Tomando-se como hipótese que a probabilidade é uniforme em cada célula, uma estimativa $\hat{p}(x)$ da função de densidade de probabilidade estudada, $p(x)$, para qualquer valor real do intervalo Ω, pode ser avaliada por:

$$\hat{p}(x) = \sum_{k=1}^{K} \frac{a_k}{n\Delta} \chi_{C_k}(x) = \frac{1}{n\Delta} \sum_{k=1}^{K} a_k \chi_{C_k}(x) \tag{4.8}$$

que corresponde à equação (4.7), onde Δ representa o volume V e o número de amostras em cada célula é $k = \sum_{k=1}^{K} a_k \chi_{C_k}(x)$.

Contudo, esta estimativa possui algumas fraquezas que fazem com que ela seja raramente utilizada como uma ferramenta estatística. Primeiramente, a aproximação $\hat{p}(x)$, definida na equação (4.8), é uma função não-contínua cuja estimação é limitada pela dualidade precisão/confiança. Esta dualidade reside no fato de que, quanto menor for a distância desejada entre $\hat{p}(x)$ e $p(x)$, menor deve ser a largura Δ; porém, como n é um número finito, também menor será o valor do acumulador de cada célula e conseqüentemente menor será a convergência local de $\hat{p}(x)$ em $p(x)$. Por outro lado, quanto maior for a largura da célula, menor é a habilidade da densidade estimada de responder apropriadamente a variações de $p(x)$.

Além da dificuldade na escolha apropriada da largura da célula, a escolha do intervalo de referência Ω também pode influenciar no resultado encontrado. A Figura 15 ilustra o efeito da translação do intervalo de referência na forma dos histogramas construídos a partir de 100 amostras obtidas de uma densidade normal $N(0, 1)$.

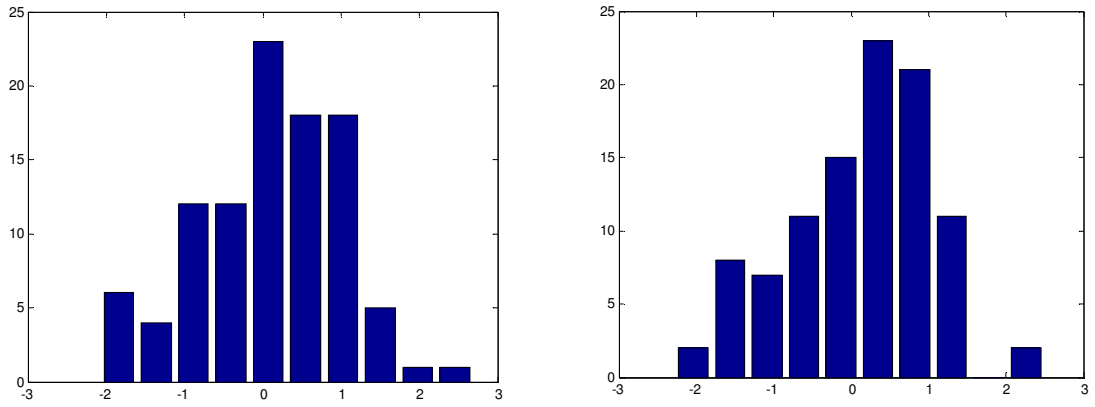

Figura 15 - Dependência da forma do histograma em função da escolha da origem das células.

Outro pronto fraco dos histogramas é a necessidade de um alto número de amostras. Esta deficiência fica ainda mais em evidência em problemas com espaços em alta dimensão, já que o número de células aumenta e conseqüentemente muitas observações são necessárias para evitar que a estimativa seja nula em uma grande região.

4.4.3. Métodos de Kernel

Os métodos de *kernel* tentam solucionar o problema da escolha da posição inicial das células e ao mesmo tempo obter uma função contínua. Para isso, diferentemente dos histogramas que utilizam células com volumes fixos (com largura Δ) e posições pré-determinadas, o volume V das células varia em função do número de amostras e cada célula é centrada em cada amostra.

Re-escrevendo a estimativa dada pelas eqs. (4.6) e (4.7), nota-se que ela também corresponde a uma média espacial de $p(\mathbf{x})$:

$$\frac{P}{V} = \frac{\int_R p(\mathbf{x})d\mathbf{x}}{\int_R d\mathbf{x}} \tag{4.9}$$

Para obter uma estimativa para $p(\mathbf{x})$, e não de valores médios, V deve se aproximar de zero. No entanto, se o número n de amostras é fixo e V tende a zero, o volume da região pode eventualmente ficar "pequeno demais" e assim pode não conter nenhuma amostra, levando a estimativa $p(\mathbf{x}) \approx 0$ a ser inútil.

Do ponto de vista prático, o número de amostras é sempre finito e será necessário considerar algum nível de suavização na estimativa de $p(\mathbf{x})$ e aceitar alguma variância na razão k/n.

Do ponto de vista teórico, pode-se considerar um número infinito de amostras. Para estimar a densidade $p(\mathbf{x})$ em $\mathbf{x}$, constrói-se a seqüência de regiões $\boldsymbol{R_1}$, $\boldsymbol{R_2}$,..., contendo $\mathbf{x}$, onde a primeira região contém uma amostra, a segunda duas, e assim adiante. Seja V_n o volume da região $\boldsymbol{R_n}$ que contém k_n amostras, e seja $p_n(\mathbf{x})$ a n-ésima estimativa para $p(\mathbf{x})$ dada por:

$$p_n(\mathbf{x}) \approx \frac{k_n/n}{V_n} \tag{4.10}$$

Pode-se provar que $p_n(\mathbf{x})$ converge para $p(\mathbf{x})$, ou seja $\lim_{n\to\infty} p_n(\mathbf{x}) = p(\mathbf{x})$, se as três condições abaixo forem satisfeitas [27]:

i) $\lim_{n\to\infty} V_n = 0$

ii) $\lim_{n\to\infty} k_n = \infty$

iii) $\lim_{n\to\infty} k_n/n = 0$

A primeira condição assegura que a média espacial P/V converge para $p(\mathbf{x})$. A segunda garante que a razão de freqüência (em probabilidade) converge para a probabilidade P. A terceira condição afirma que o número de amostras caindo na região

$\boldsymbol{R_n}$ é sempre uma pequena parcela desprezível do número total de amostras. Ela é necessária para que $p_n(\mathbf{x})$ definida pela eq. (4.10) convirja.

Um dos caminhos para se obter seqüências de regiões que satisfazem estas condições é, a partir de um volume inicial, encolhê-lo à medida que *n* aumenta; por exemplo: $V_n = 1/\sqrt{n}$. É em seguida necessário mostrar que k_n e k_n/n têm comportamento apropriado para que $p_n(\mathbf{x}) \to p(\mathbf{x})$. Este é basicamente o método conhecido como "Janela de *Parzen*" [27], que será examinado a seguir.

4.4.4. Janela de Parzen

Supondo que a região $\boldsymbol{R_n}$ é um hipercubo *d*-dimensional com lado igual a h_n, o seu volume é dado por $V_n = h_n^d$. Seja a função que assume 1 para os pontos dentro do hipercubo unitário centrado na origem e 0 para os pontos externos; esta função é chamada de *função de janela* e é definida por:

$$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \le \frac{1}{2} \quad j = 1,\ldots,d \\ 0 & \textit{caso contrário} \end{cases} \tag{4.11}$$

Conseqüentemente, $\varphi((\mathbf{x}-\mathbf{x}_i)/h_n)$ será igual a 1 se $\mathbf{x}_i$ estiver dentro do hipercubo de volume V_n centrado em $\mathbf{x}$, e zero caso contrário. Portanto, o número de amostras dentro deste hipercubo é dado por:

$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) \tag{4.12}$$

Substituindo na eq. (4.10), obtém-se:

$$p_n(\mathbf{x}) \approx \frac{k_n/n}{V_n} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) \tag{4.13}$$

A estimativa $p_n(\mathbf{x})$ definida acima é uma média de funções (de janela). Tipicamente a função de janela tem seu máximo na origem e seus valores caem à medida

que se distanciam da origem. Desta forma, cada amostra está contribuindo com a estimativa conforme sua distância de **x**.

Para que a estimativa $p_n(\mathbf{x})$ seja uma função de densidade legítima, isto é, que seja não-negativa e integre em 1, as três condições abaixo devem ser atendidas:

i) $$\varphi(\mathbf{u}) \geq 0 \tag{4.14}$$

ii) $$\int \varphi(\mathbf{u})\, d\mathbf{u} = 1 \tag{4.15}$$

iii) $$V_n = h_n^d \tag{4.16}$$

Uma escolha comum para a função de janela é a Normal de média $\mathbf{x}_i$ e variância h_n:

$$\varphi(\mathbf{u}) = \frac{1}{(2\pi)^{d/2}} \exp[-0{,}5\mathbf{u}^T\mathbf{u}]$$

que produz a estimativa:

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi h_n{}^2)^{d/2}} \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h_n{}^2}\right) \tag{4.17}$$

É importante observar que, se h_n for muito grande, pontos muito afastados da amostra $\mathbf{x}_i$ serão afetados de maneira importante pela janela. Assim a estimativa será composta pela superposição de funções lentas, ficando suave demais e com uma visão "fora de foco" da densidade de probabilidade. Por outro lado, se h_n for muito pequeno, apenas os pontos muitos próximos a $\mathbf{x}_i$ serão afetados de maneira importante pela janela. Neste caso a estimativa será uma superposição de "picos agudos" centrados nas amostras e $p_n(\mathbf{x})$ será muito "ruidosa".

Na prática, deve ser procurada uma concessão aceitável, já que o número de amostras é sempre limitado. É comum escolher um valor para h_1 e definir $h_n = h_1/\sqrt{n}$. Infelizmente, a escolha do valor de h_1 pode ser problemática.

A Figura 16 ilustra 3 estimativas, com janelas de *Parzen* com diferentes larguras, a partir de 100 amostras geradas através de uma mistura de duas distribuições do tipo

Normal. Percebe-se claramente a influência da largura da janela na estimação: para h_1=1, a janela é estreita demais e a estimativa é muito ruidosa, apresentando vários picos; para h_1=16, a janela é larga demais e praticamente não são notados os dois picos da distribuição original; e h_1=4 parece ser um valor adequado, sem grandes ruídos nem suavizado em excesso, sendo a qualidade de sua estimação comprometida pelo baixo número de amostras.

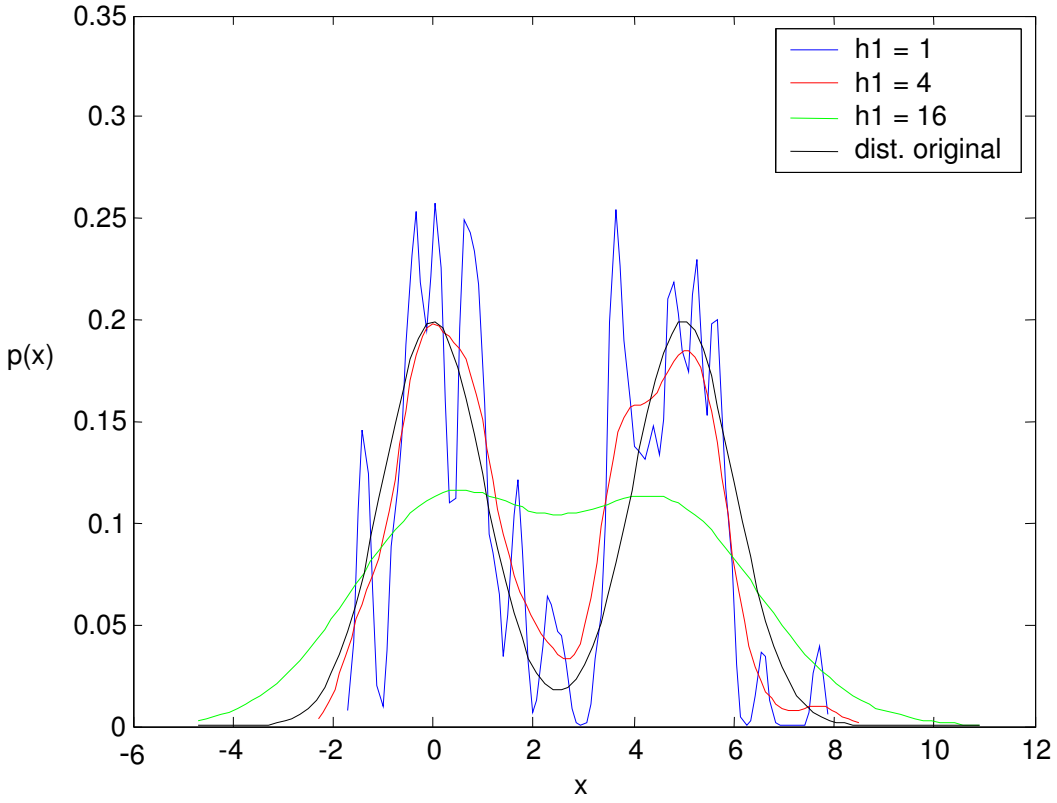

Figura 16 – Influência da largura da janela na estimativa por Janela de *Parzen*