# 第12章　カーネル密度推定法

## 10.1　ヒストグラム法
## 10.2　ノンパラメトリック法の枠組み

03T4027N　佐々　知広

# ヒストグラム法

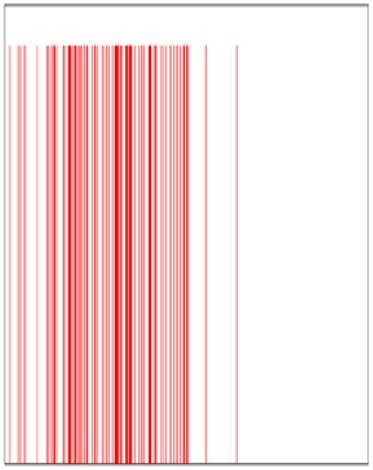
ノンパラメトリック法

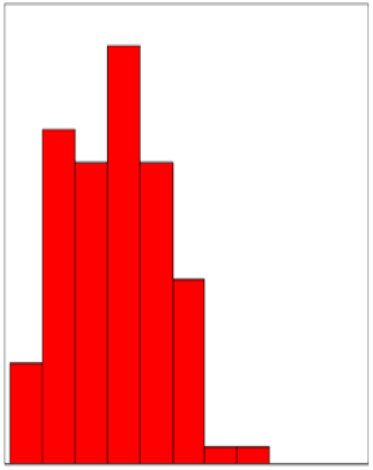
－真の確率関数が複雑なとき、適切なパラメトリック
モデルを構築するのが困難な為に用いる方法。

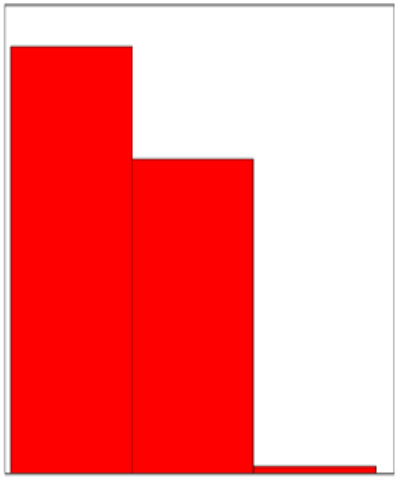
ヒストグラム法

－パターン空間Dを適当に分割し、
各分割内に入る訓練標本数を数え、
全体の積分が1になるように正規化したものを
確率密度関数の推定量とする方法。

# ヒストグラム法の問題点







- (a)幅小　　(b)適切　　(c) 幅大

# ヒストグラム法の問題点

- 領域の幅を適切に決定するのが難しい
- 推定した確率密度関数が領域の間で不連続
- 入力次元の増加に伴い領域の数が指数的に増加してしまい、領域に訓練標本数が含まれなくなる

# ノンパラメトリック法の表記

ある注目点 $x`$での確率密度 $p(x`)$を推定する

R：注目点 $x`$を含むパターン空間内のある領域

V：Rの体積　$V := \int_R dx$

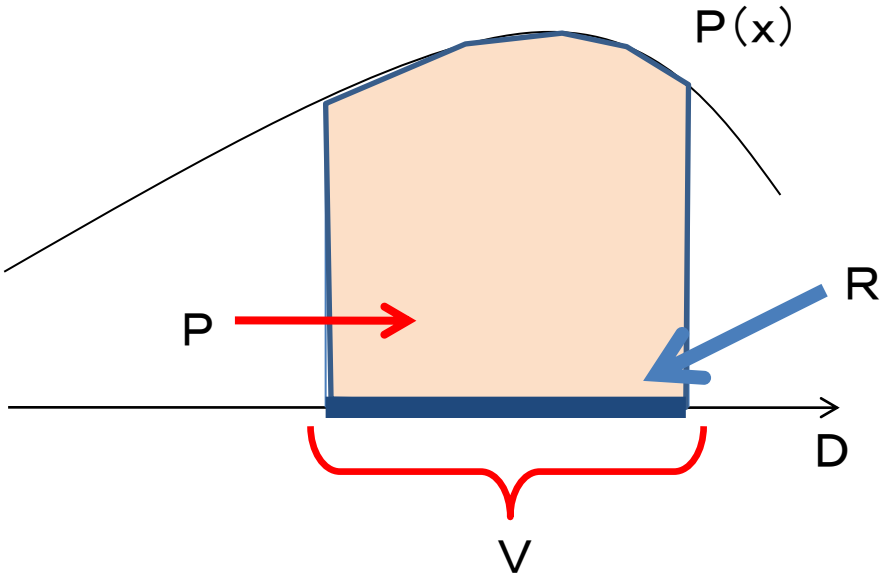
P：あるパターンxが領域Rに入る確率

$$P := \int_R p(x)dx$$

# ノンパラメトリック法の表記

k：n個の訓練標本のうち

領域Rに入っている個数



# ノンパラメトリック法の基礎

- 確率Pの値は以下のように近似できる。

(a)注目点 $x`$を用いて $P \approx Vp(x`)$

(b)n、kを用いて $P \approx \frac{k}{n}$

(c) (a),(b)を用いて $P(x`) \approx \frac{k}{nV}$

# 近似(a)

$$P \approx Vp(x`)$$

- （a）は積分の長方形近似。

- 領域R内でp（x）が定数関数に
  近いほうが近似精度が良い
  → 領域Rは小さいほうが良い

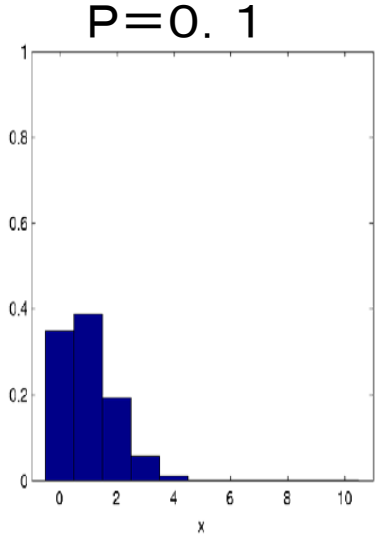
# 近似(b)

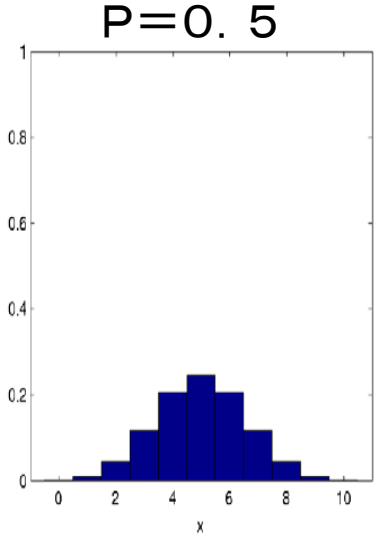
$$P \approx \frac{k}{n}$$

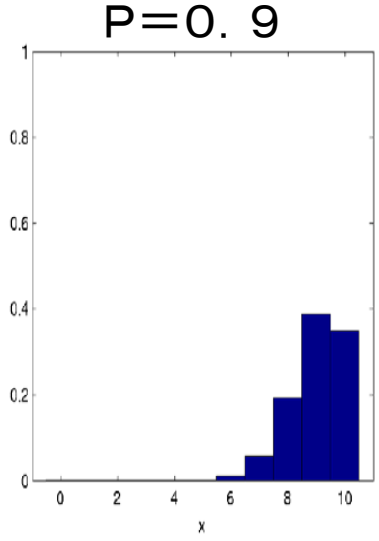
- n個の訓練標本のうちk個がRに入る確率は二項分布に従う。

確率: ${}_nC_k P^k (1-P)^{n-k}$　期待値: $nP$

分散: $nP(1-p)$



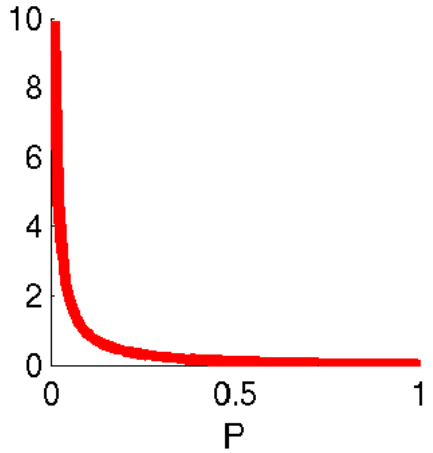
# 近似(b)

- 期待値は真のPと一致する。
- 分散が大きければ $\frac{k}{n}$ はPの推定量として適切でないので期待値を1に正規化した量Zを考える。

$$z := \frac{k}{nP} \qquad V[z] = \frac{1-P}{nP}$$

Pが大きいほうがzの分散は小さくなる

→ 領域Rが大きいほうが近似精度が良い




図、二項分布の正規化分散

# 近似(c)

- 近似(c)の精度は近似(a)(b)
  の近似の精度に依存する。

近似(a) ⇒ 領域Rを小さく

近似(b) ⇒ 領域Rを大きく

領域Rを程よい大きさに決める必要がある。