Density estimation

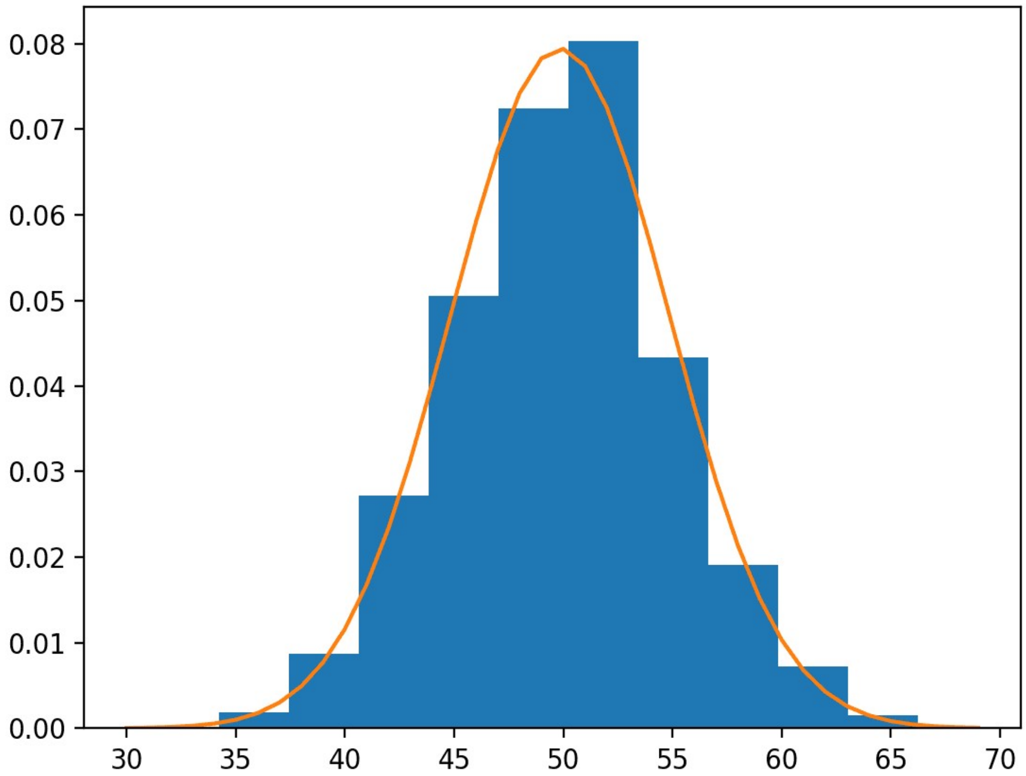
Parametric

Non-parametric

1. Kernel density (Parzen)
2. Nearest - neighbourhood

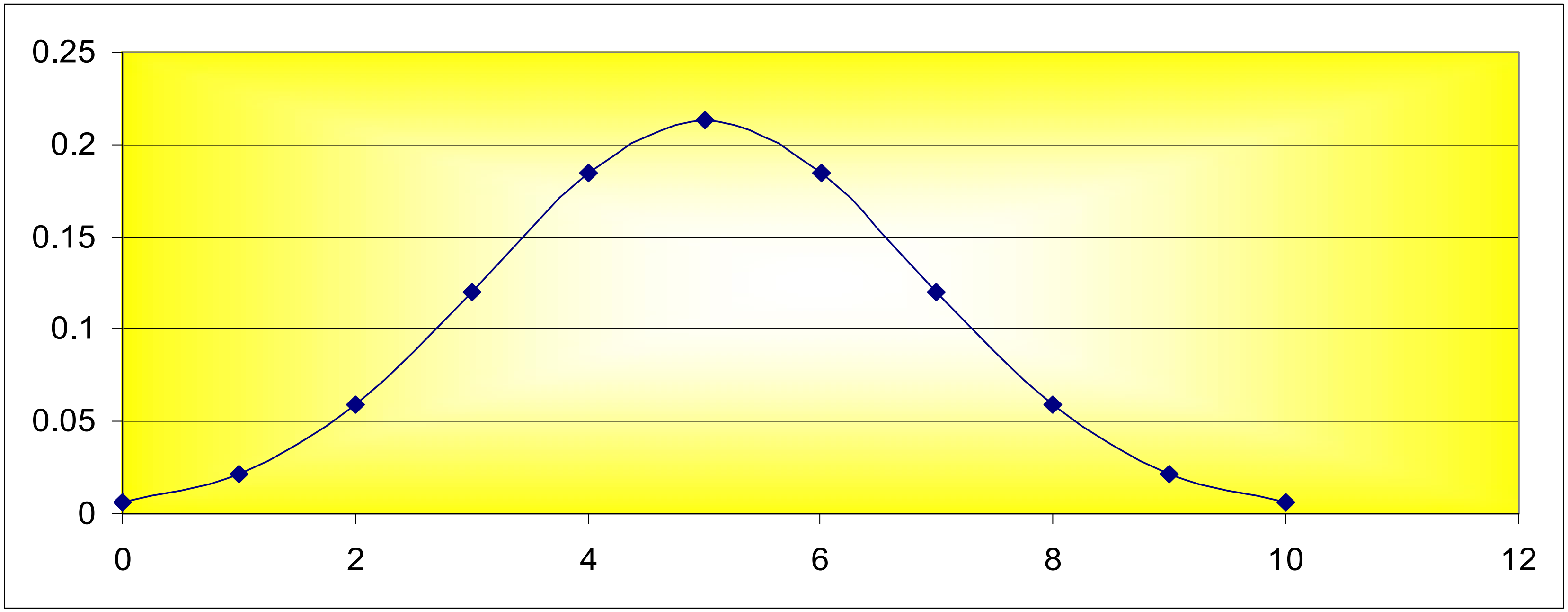
Parametric estimation

Parametric density estimation



- Generate histogram from the given data
- Look @ the shape
- Try to guess the distributions
- Popular distributions
  - Gaussian
  - Poisson
  - Uniform

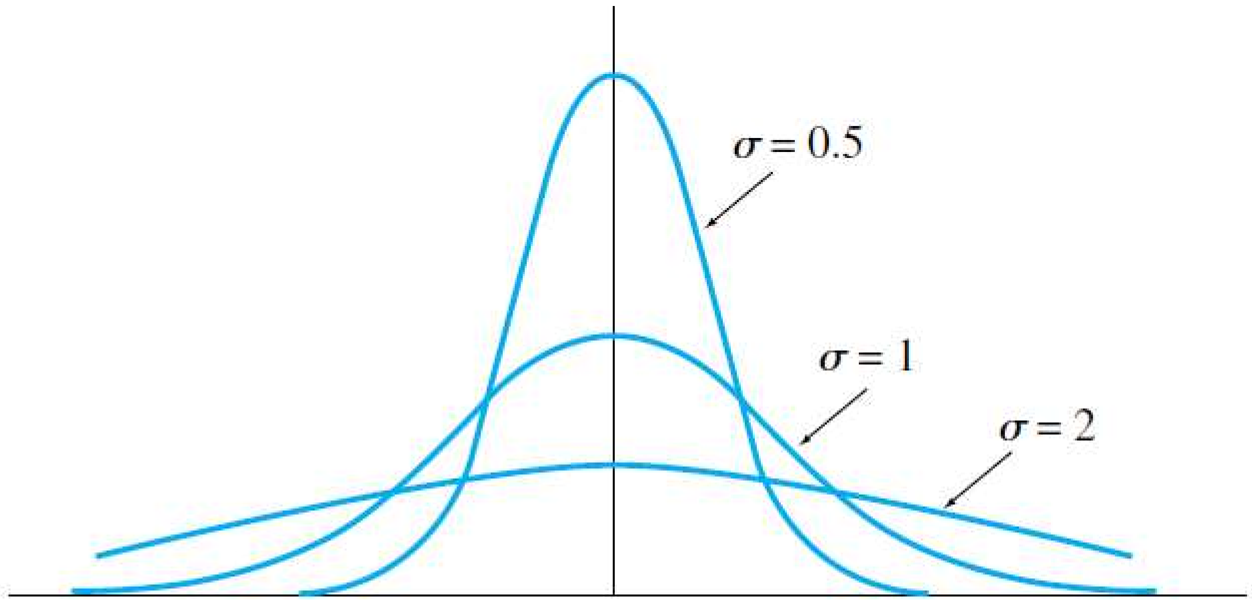
Gaussian distribution – mean and standard deviation



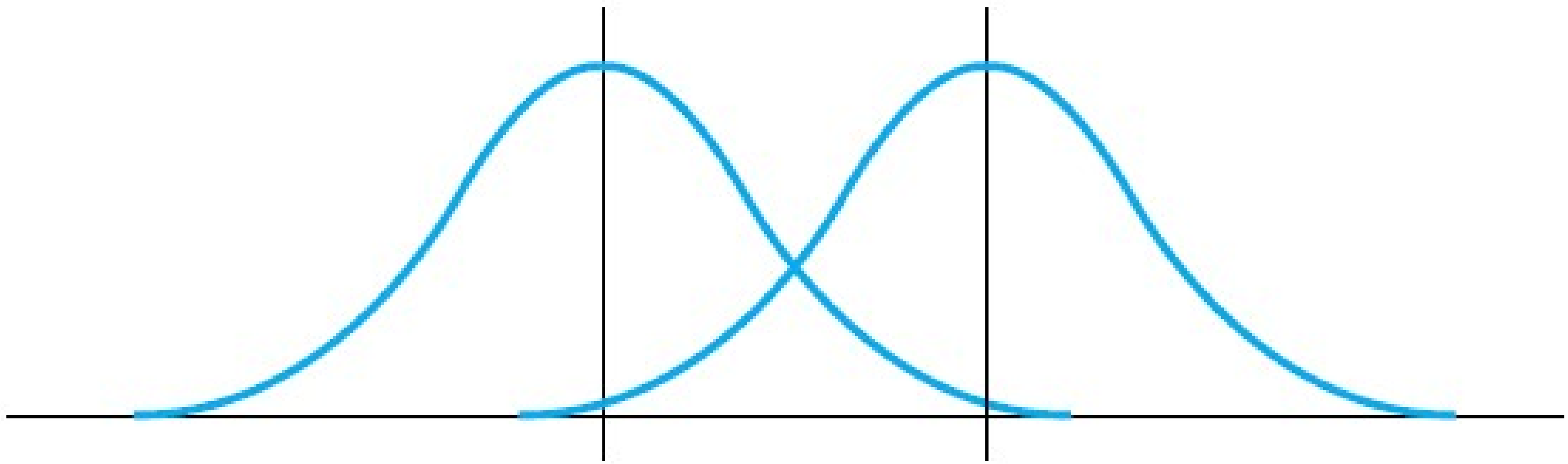
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Defined from $-\infty$ to $+\infty$

Same mean – different standard deviations



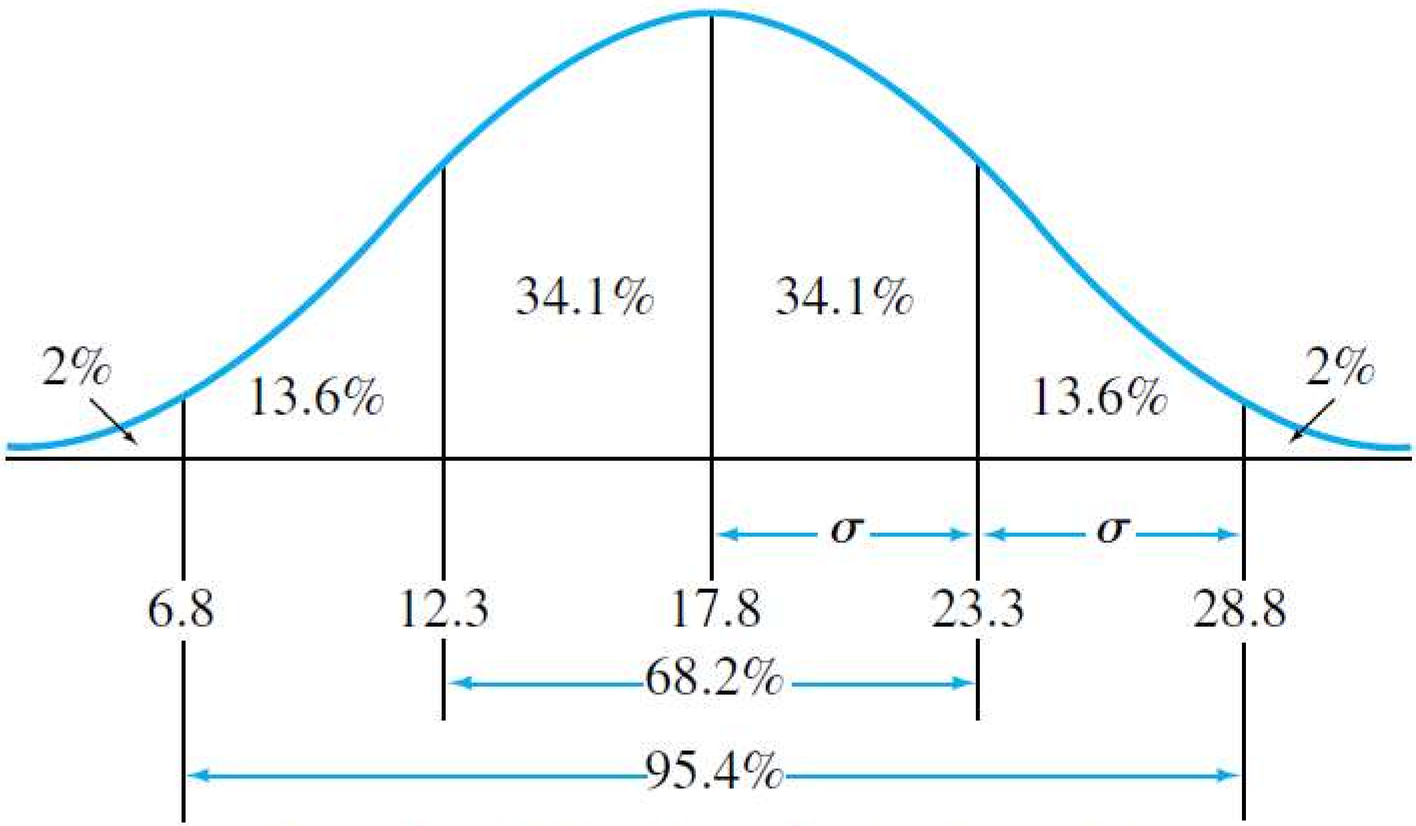
Same standard deviation – different means



Standard normal deviation

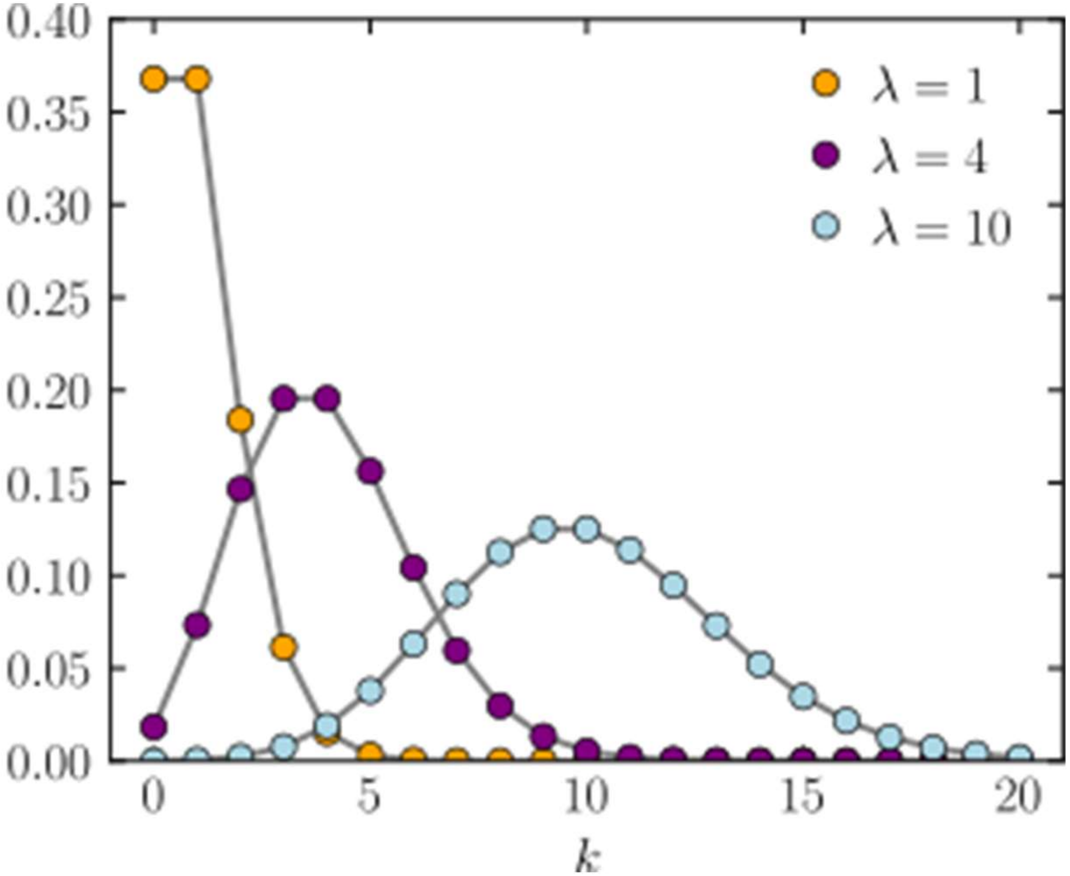
- Make, $\sigma = 1$
- Area under the curve becomes 1
- can be used as probability measure

Area under the normal curve



Normal curve with $\mu = 17.8$ and $\sigma = 5.5$

Poisson's distribution



$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

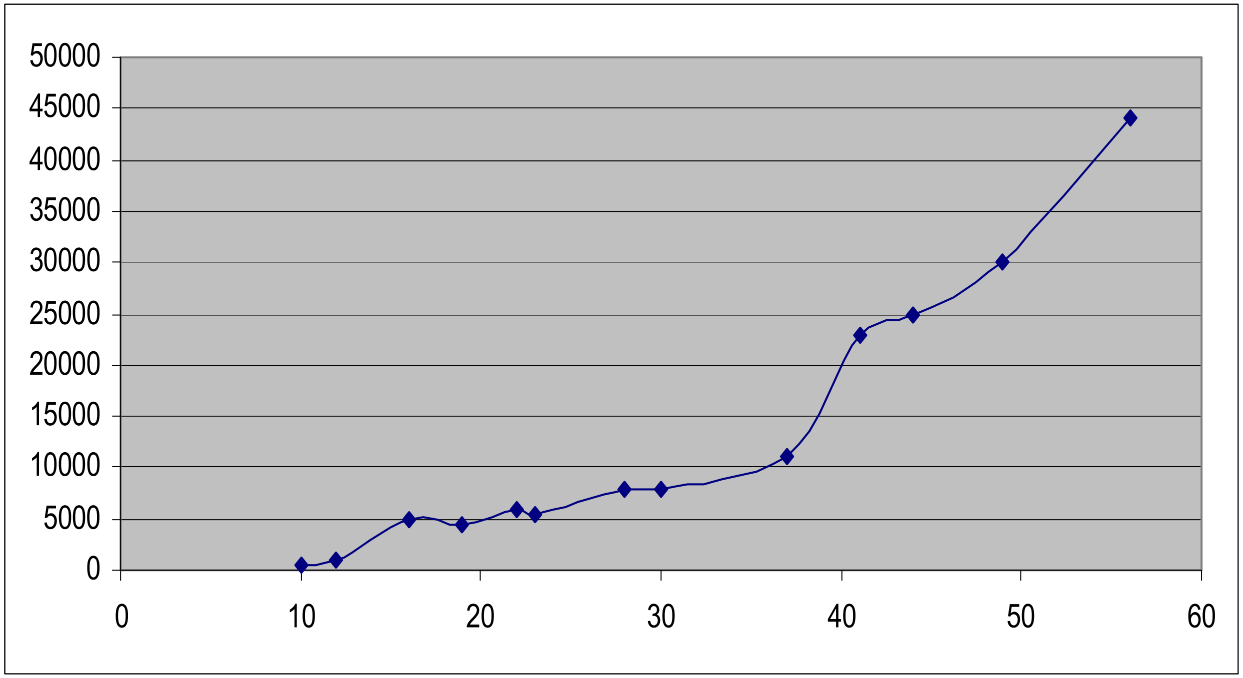
Estimate the distribution parameters

- After guessing distribution, estimate the distribution parameters

- Gaussian
  - Estimate μ and σ
- Poisson
  - Estimate λ

Data visualization

AGE	Ca - **deficiency**
9	500
12	1000
16	5000
19	4500
22	6000
23	5500
28	8000
30	7800
37	11000
41	23000
44	25000
49	30000
56	44000



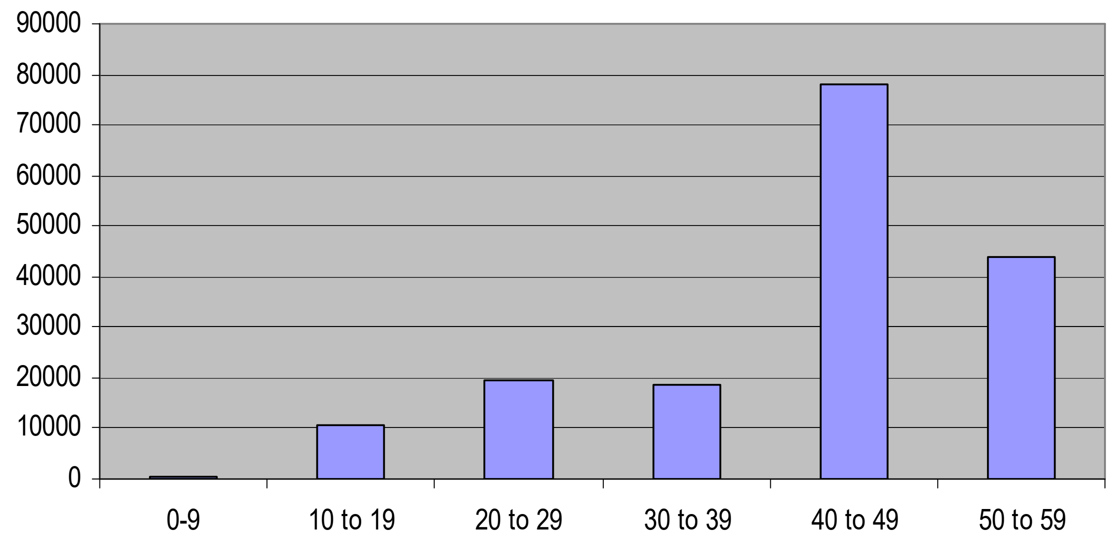
Conclusion from the previous slide

- As age increases people more prone for Calcium deficiency

If you are > 50 years

"Drink Bournvita"

9	500
12	1000
16	5000
19	4500
22	6000
23	5500
28	8000
30	7800
37	11000
41	23000
44	25000
49	30000
56	44000

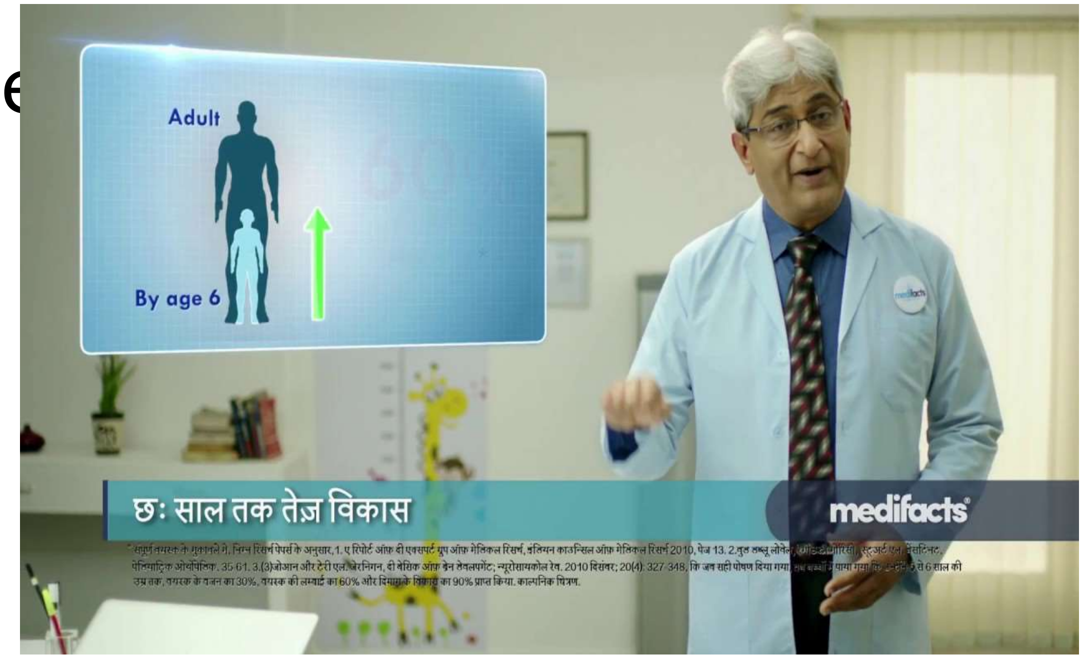


0-9	500
10 to 19	10500
20 to 29	19500
30 to 39	18800
40 to 49	78000
50 to 59	44000

Histogram

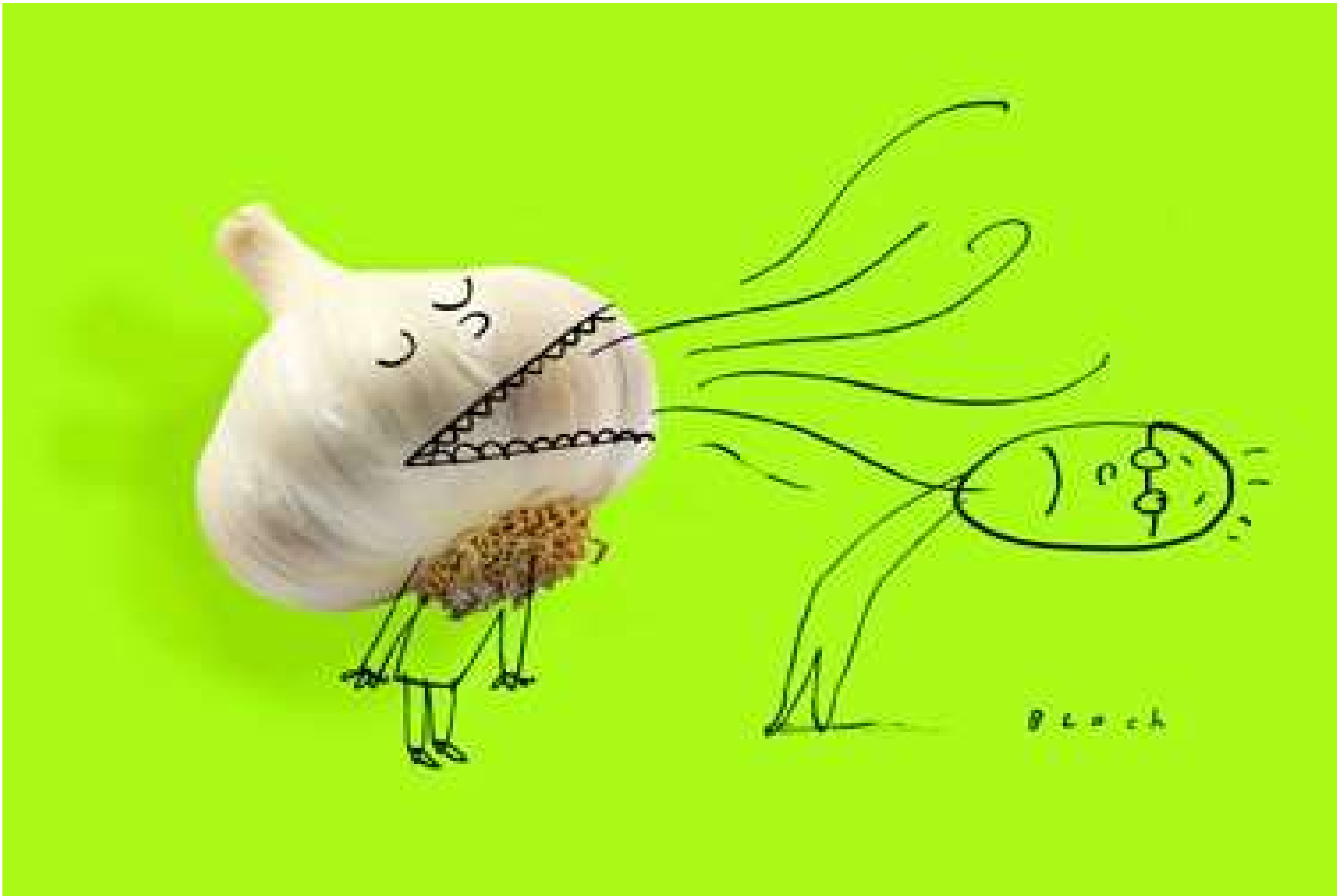
Horlicks Advertisement: from the previous slide

- People in 40-49 age group more prone for Calcium deficiency

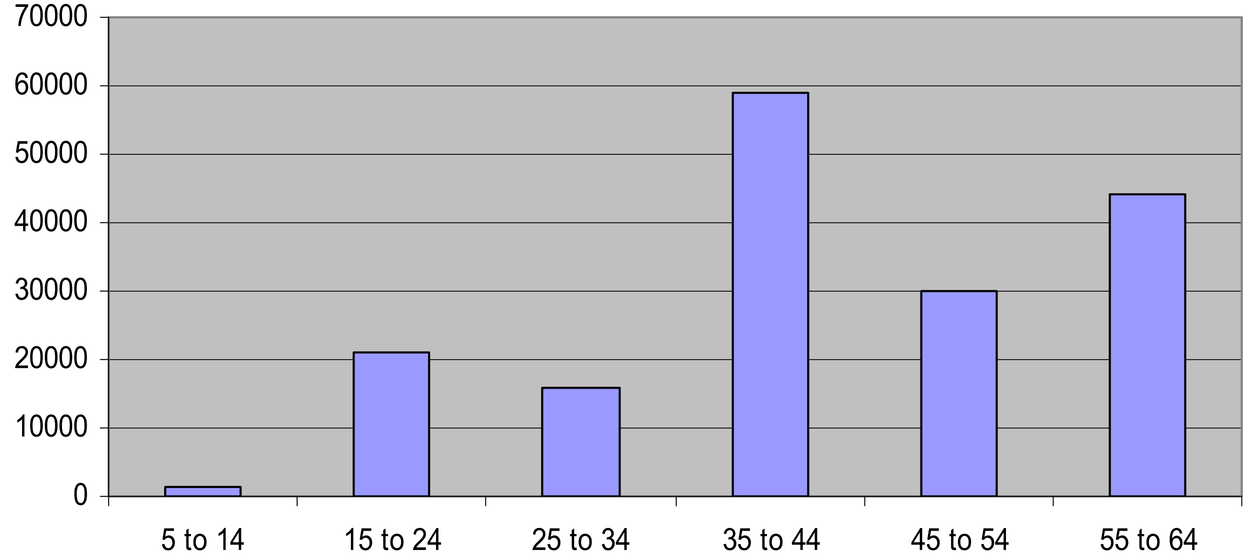


If you are 40-49
“Drink Horlicks”

Let us start a company “Garlicks”



9	500
12	1000
16	5000
19	4500
22	6000
23	5500
28	8000
30	7800
37	11000
41	23000
44	25000
49	30000
56	44000

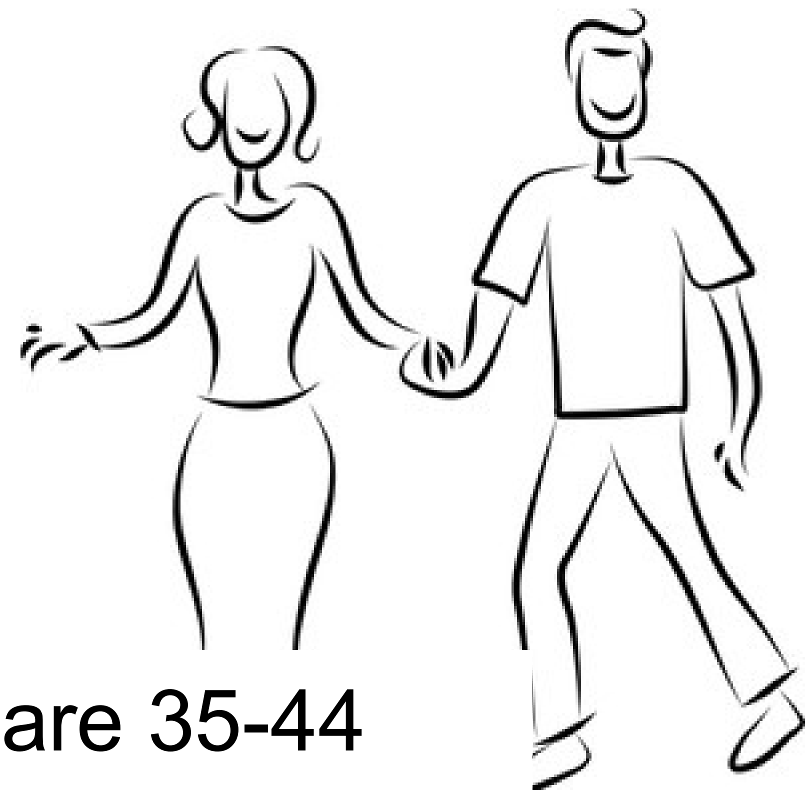


5 to 14	1500
15 to 24	21000
25 to 34	15800
35 to 44	59000
45 to 54	30000
55 to 64	44000

Histogram

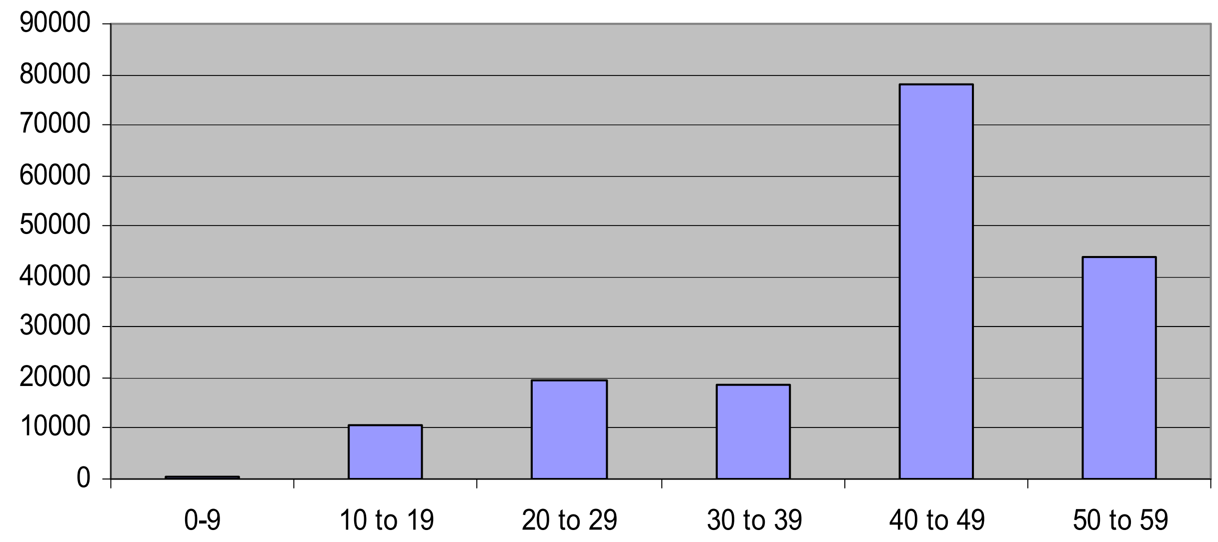
Garlicks advertisement: from the previous slide

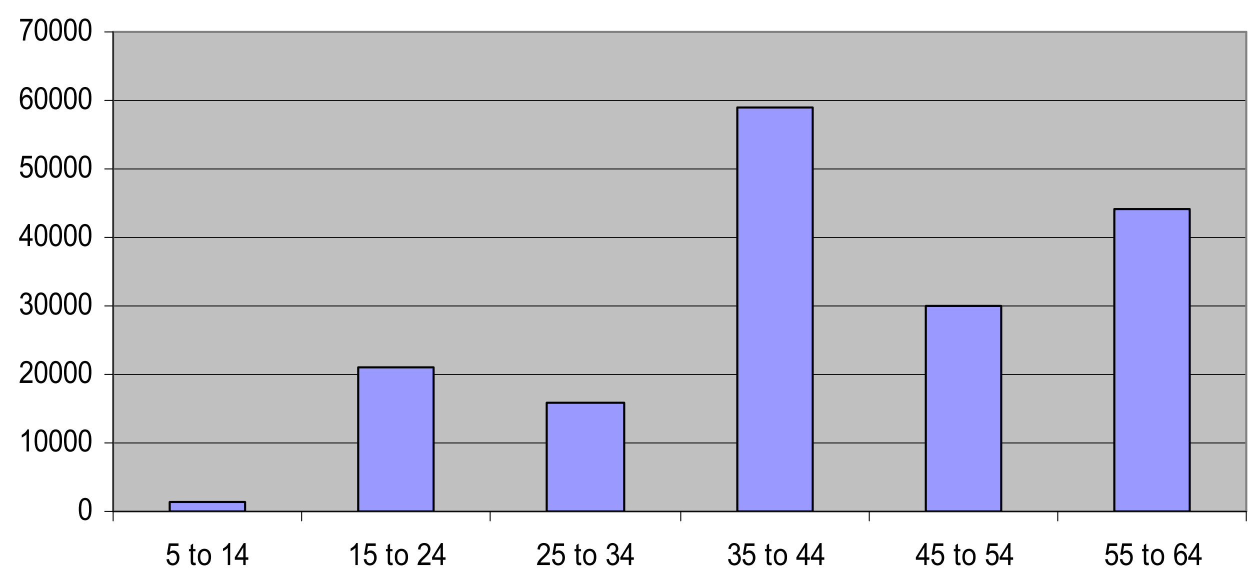
- A recent study shows that people in 35-44 age group also more prone for Calcium deficiency



If you are 35-44
"Drink Garlicks"

9	500
12	1000
16	5000
19	4500
22	6000
23	5500
28	8000
30	7800
37	11000
41	23000
44	25000
49	30000
56	44000





Which is correct?

Histogram - problems

- Histogram shape depends on the bin width
- Change the bin width – shape also changes
- Suppose bin width is constant. Is histogram unique? – NO
  – 0-2; 2-4; 4-6; … are bins & bin width = 2
  – -1-+1; 1-3; 3-5; … are bins & bin width=2
- Bin origin is another problem

Bin width

- Smoothing parameter
  - A smaller binwidth leads to a relatively jagged histogram
  - A larger binwidth results in a smoother looking histogram

Bin edge

- Sensitivity of the histogram to the placement of the bin edge
- Average shifted histogram:
  - Averages several histograms based on shifts of the bin edges

Sample No.	Value
1	-2.1
2	-1.3
3	-0.4
4	1.9
5	5.1
6	6.2

If the histogram is pdf then

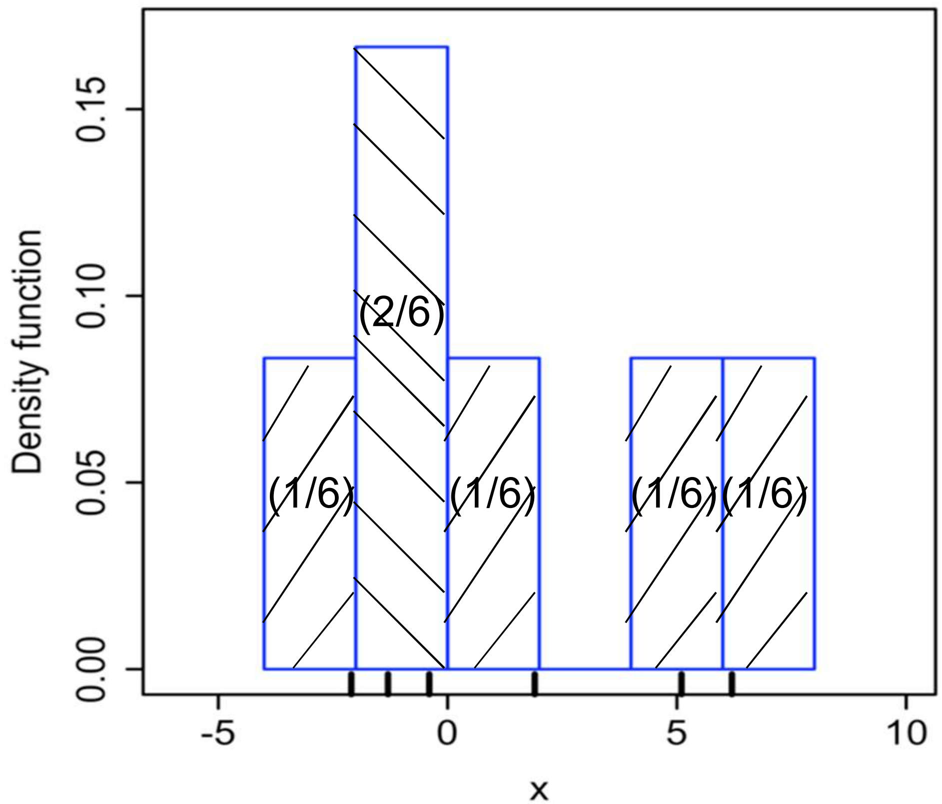
- $f(x) = \text{limit } (h \rightarrow 0)\ 1/(2h).\ P(x-h < X < x+h)$
- Area under the bins should add up to 1
- Window size
  - More the window width more number of points will fall
- Weight for one point in the window
  - Inversely proportional to the number of points (n)
  - Also inversely proportional to window size
    - Bigger the window size many points will fall within window i.e. lesser weight for one point
    - Smaller the window size few points will fall within window i.e. lesser weight for one point

- Histogram $\equiv$ stacking boxes
  - One box width is 2h
  - One box height 1/(2hn)
- Total box width = 2h
- Total box height = N*[1/(2hn)] where N is the number of points within the window
- If there are M boxes then
  - Area of all the M boxes = 1
- Box height is the density estimate

Histogram

-4 to -2; -2 to 0; 0 to 2; 2 to 4; & 4 to 6

- Horizontal axis is divided into sub-intervals or bins which cover the range of the data
- E.g. 6 bins each of width 2
  - One data point falls inside this interval
  - n = 6 and 2h = 2 i.e. h = 1
  - Height = 1/(2nh) = 1/12
  - 1 point in a bin → Place 1 box of height 1/12
  - 2 points in a bin → Place two boxes (2 x 1/12)
  - And so on…
  - Area of one box = 2 x 1/12 = 1/6
  - Area of six boxes = 6x1/6 = 1



Algorithm

- n data points $\{X_1, X_2, \ldots, X_n\}$
- Box center is x
- Box width $x-h$ to $x+h$
- When a data point X_i will fall with in this window?
- If $(|x-X_i|/h) < 1$ then X_i falls within the window otherwise does not fall
  - i.e. all the data points are weighted by a weighting function $w(|x-X_i|/h)$
    - If the argument is <1 then X_i is given weightage of $1/(2nh)$ otherwise zero

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right)$$

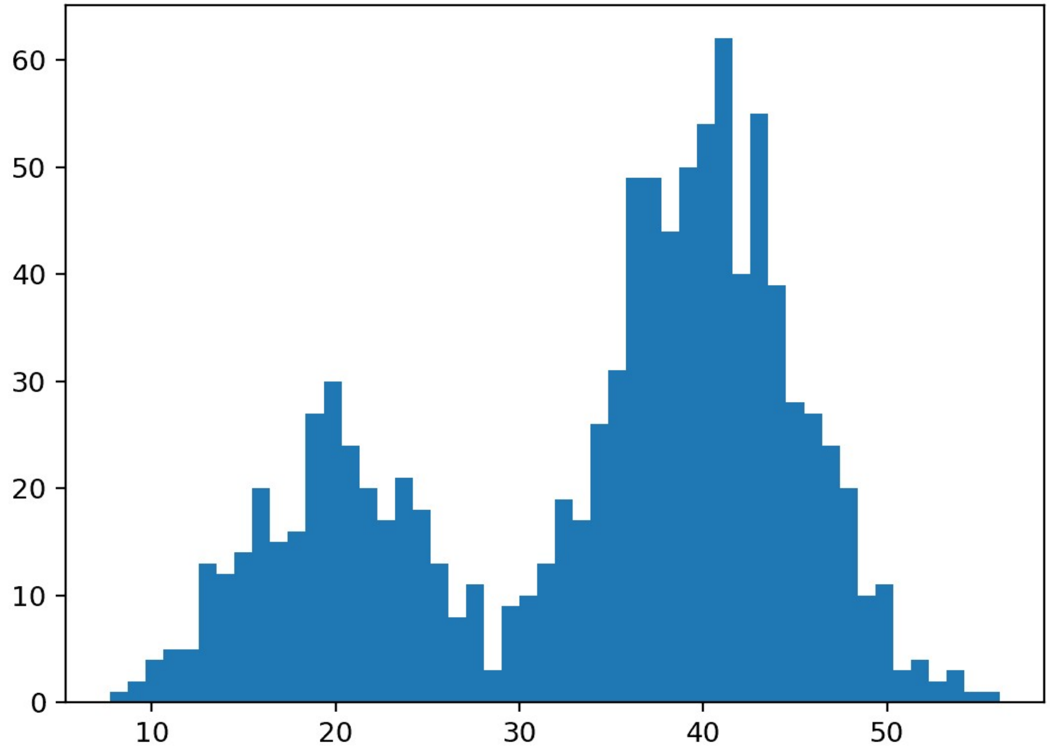
$$w(x) = \frac{1}{2} \quad if \quad |x| < 1$$

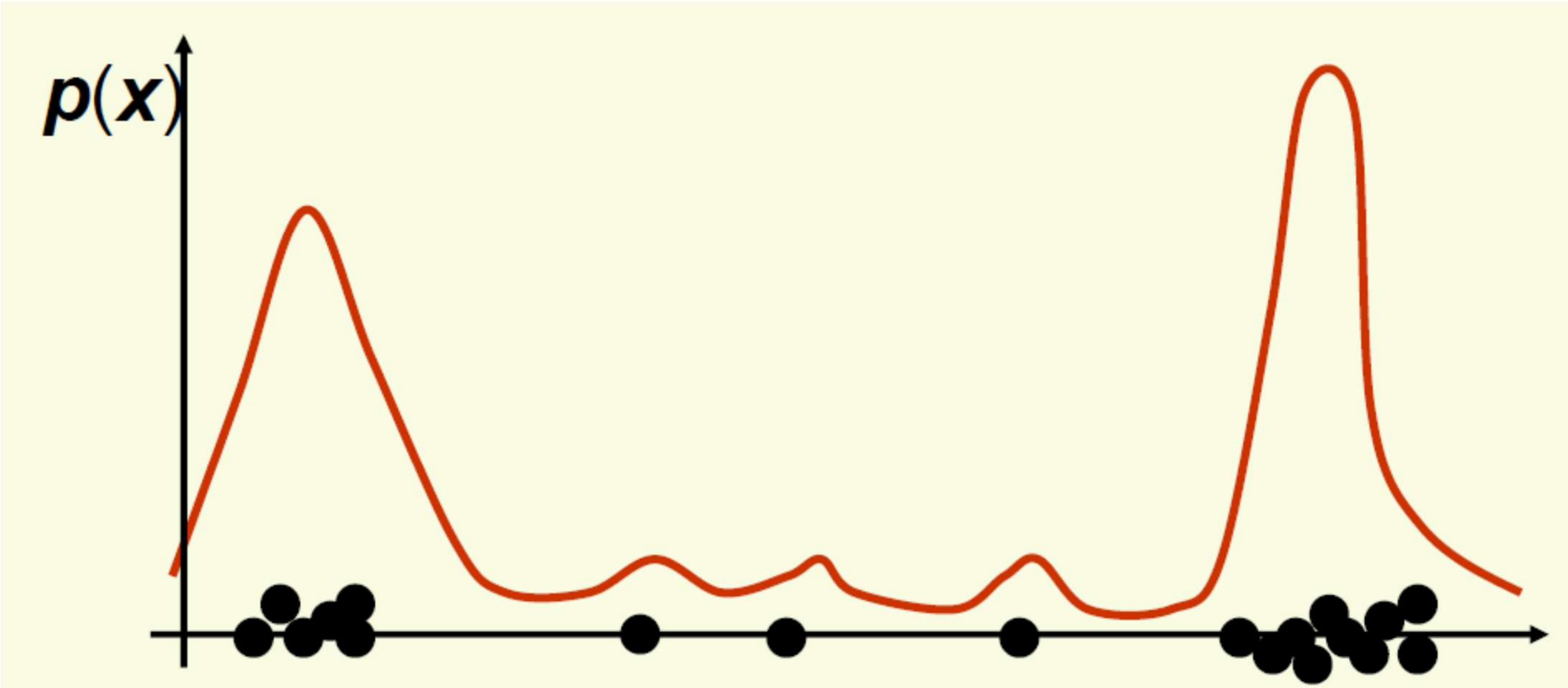
$$= 0 \quad otherwise$$

Non-parametric estimation

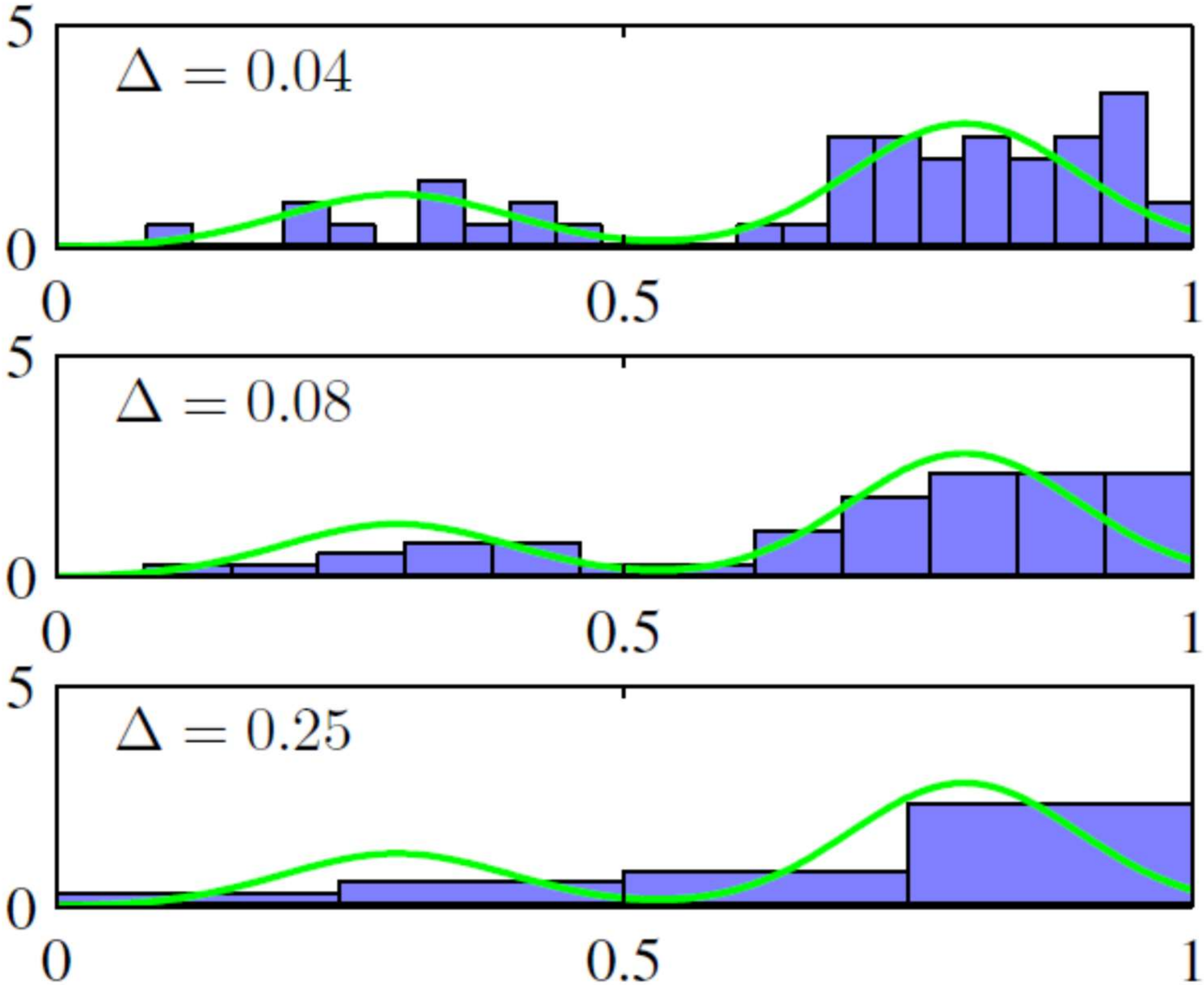
Non-parametric estimation

- If we are not able to guess the distribution
  - May be we have two peaks or more than two peaks
- Kernel density estimation




p(x)

Impact of Bin width (Δ)



- Green is the correct distribution
- When Δ = 0.04 or 0.25, Histogram do not reflect the green

Density estimate

- Imagine continuous distribution
- We got certain points
- Interested in estimating the distribution of data from the given data

Kernel estimator

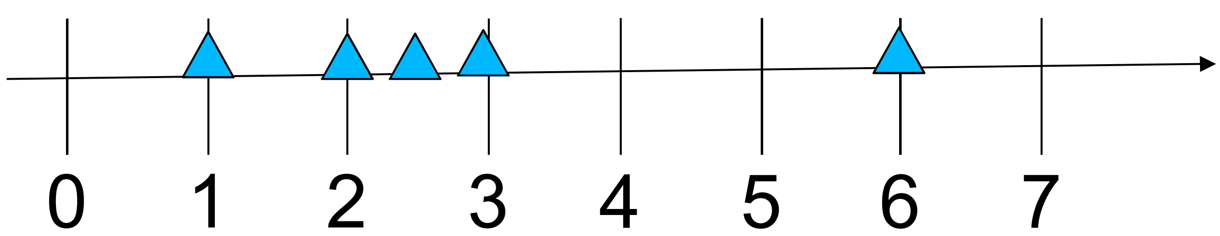
$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)$$

Parzen window estimator: Use normal Gaussian

Example

Given a set of five data points x1 = 2, x2 = 2.5, x3 = 3, x4 = 1 and x5 = 6

Find **Parzen probability density function** (pdf) estimates at x = 3, using the Gaussian function with $\sigma = 1$ as window function

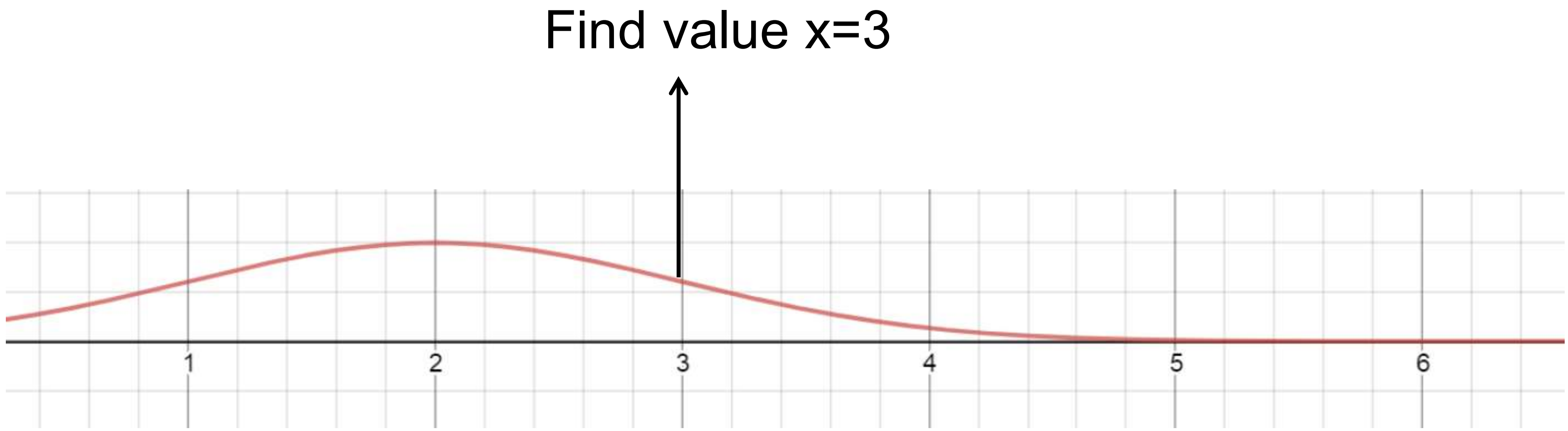


Algorithm

<u>x1 = 2, x2 = 2.5, x3 = 3, x4 = 1 and x5 = 6</u>

1. Place a Gaussian at x=2 i.e. μ = 2
2. Find its value @ x=3
3. Place a Gaussian at x=2.5 i.e. μ = 2.5
4. Find its value @ x=3
5. ..
6. ..
7. ..
8. ..
9. Place a Gaussian at x=6 i.e. μ = 6
10. Find its value @ x=3

Gaussian with μ = 2

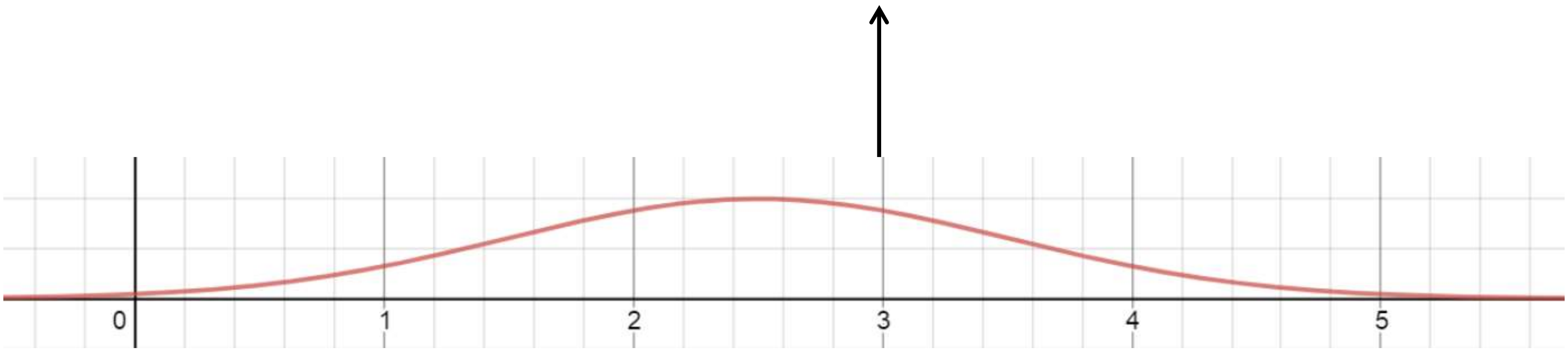


$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_1 - x)^2}{2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(2-3)^2}{2}\right) = 0.2420$$

Gaussian with μ = 2.5

Find value x=3

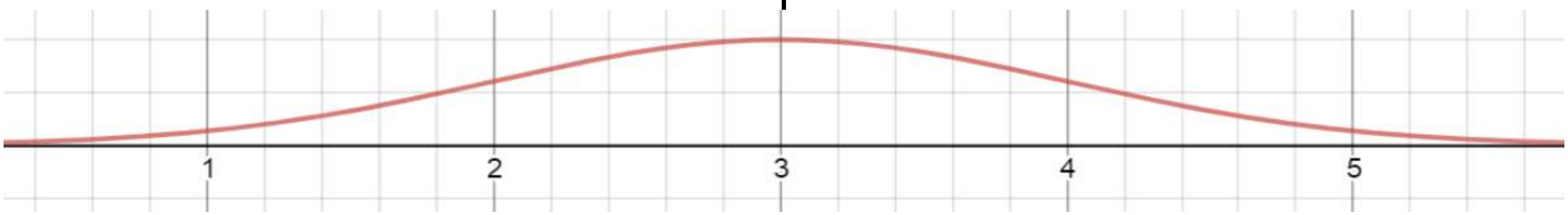


$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_2 - x)^2}{2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(2.5 - 3)^2}{2}\right) = 0.3521$$

Gaussian with $\mu = 3$

Find value x=3 (we get maximum value)

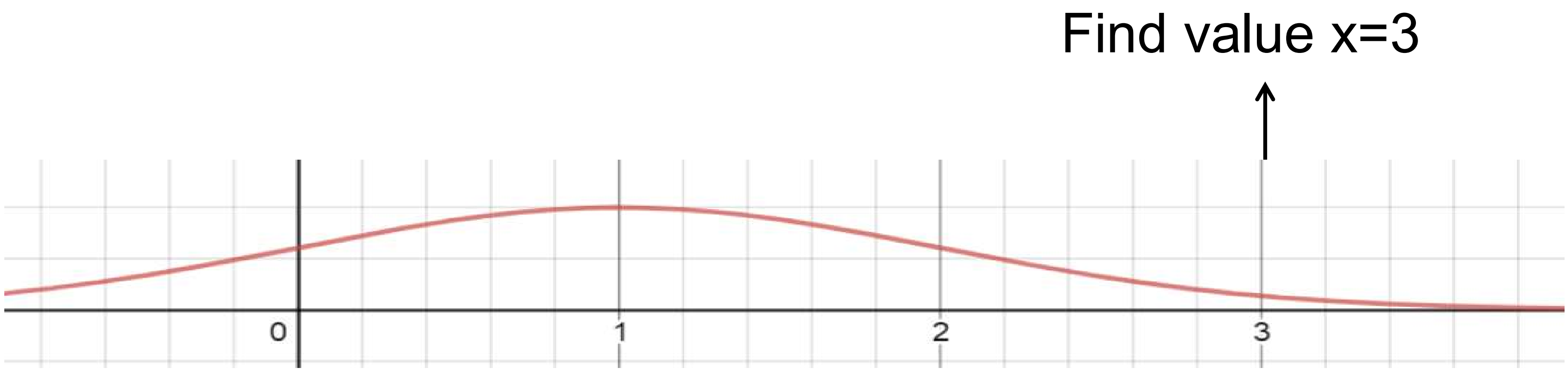


$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_3 - x)^2}{2}\right)$$

$$= \frac{}{\sqrt{2\pi}}\exp\left(-\frac{}{3 \quad 2}\right) = 0.3521$$

0.3989

Gaussian with μ = 1



$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_4 - x)^2}{2}\right)$$

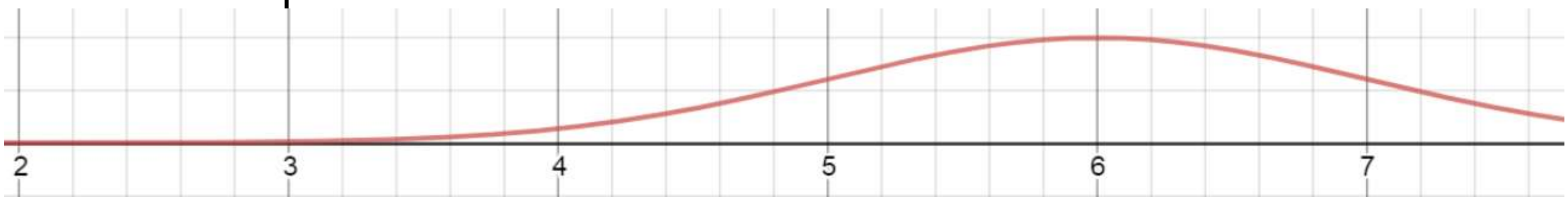
$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(2.5 - 3)^2}{2}\right) = 0.3521$$

1

0.054

Gaussian with $\mu = 6$

Find value x=3



$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_{5} - x)^2}{2}\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(2.5 - 3)^2}{2}\right) = 0.3521$$

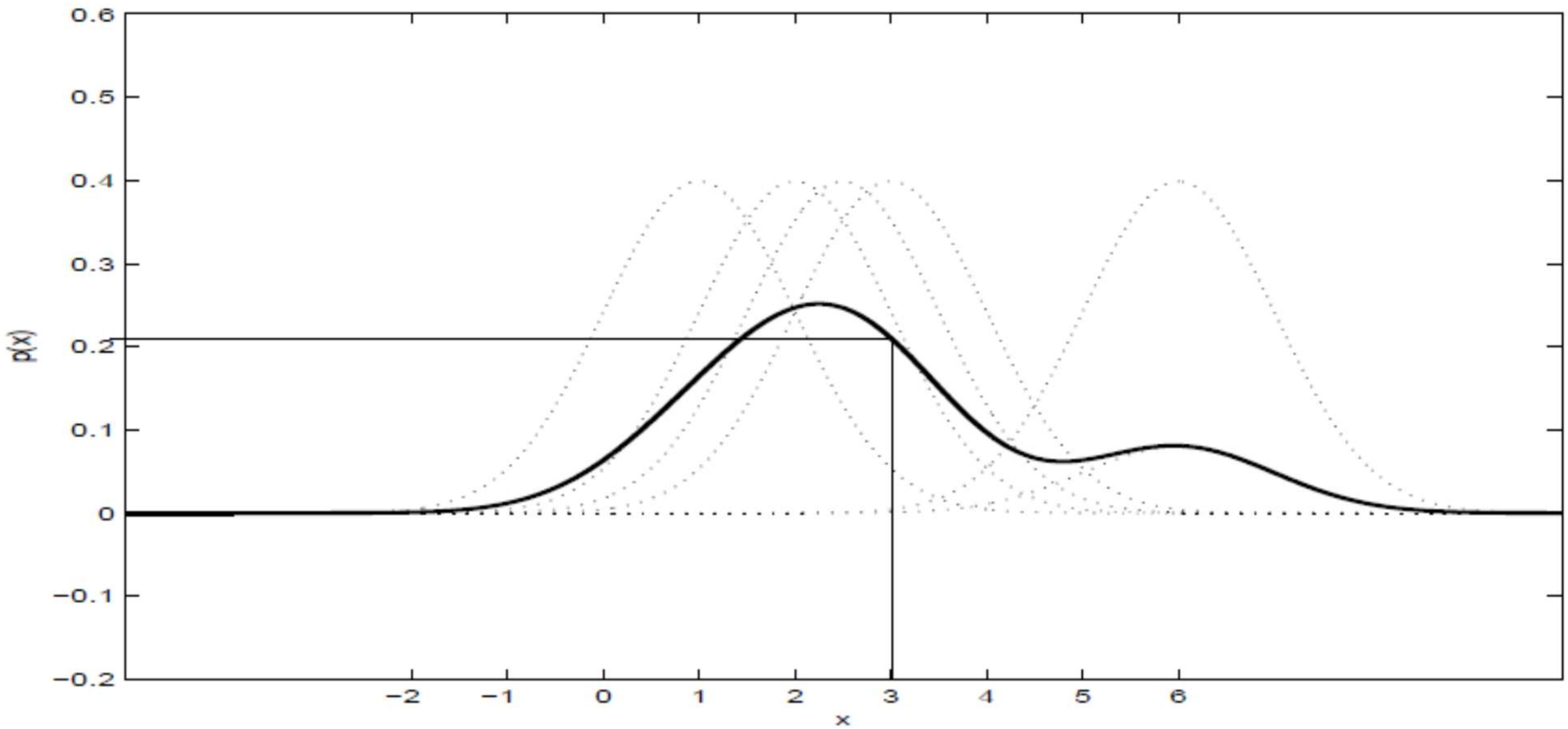
6

0.0044

$$p(x=3) = (0.2420 + 0.3521 + 0.3989$$

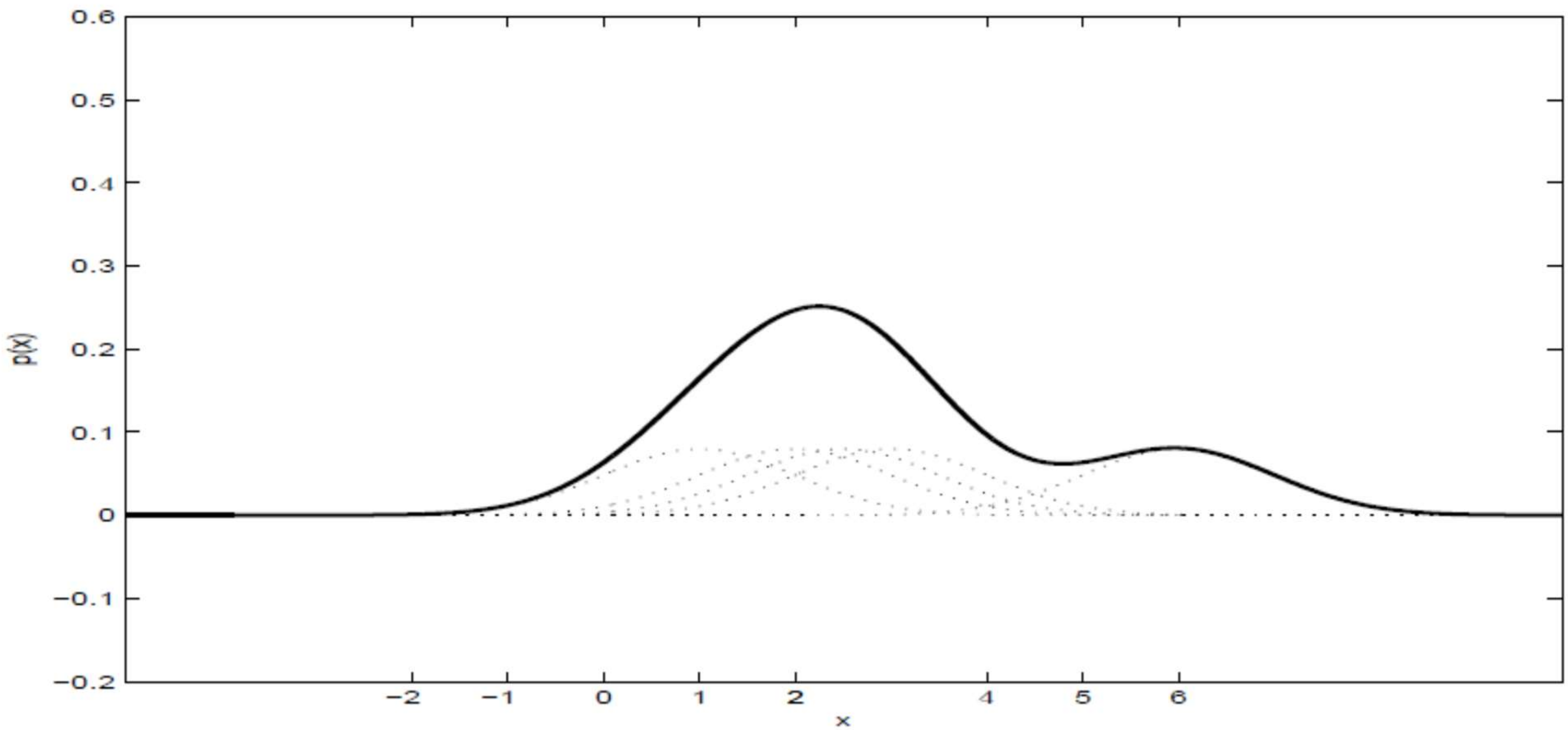
$$+0.0540 + 0.0044)/5 \quad = 0.2103$$

x1 = 2, x2 = 2.5, x3 = 3, x4 = 1 and x5 = 6



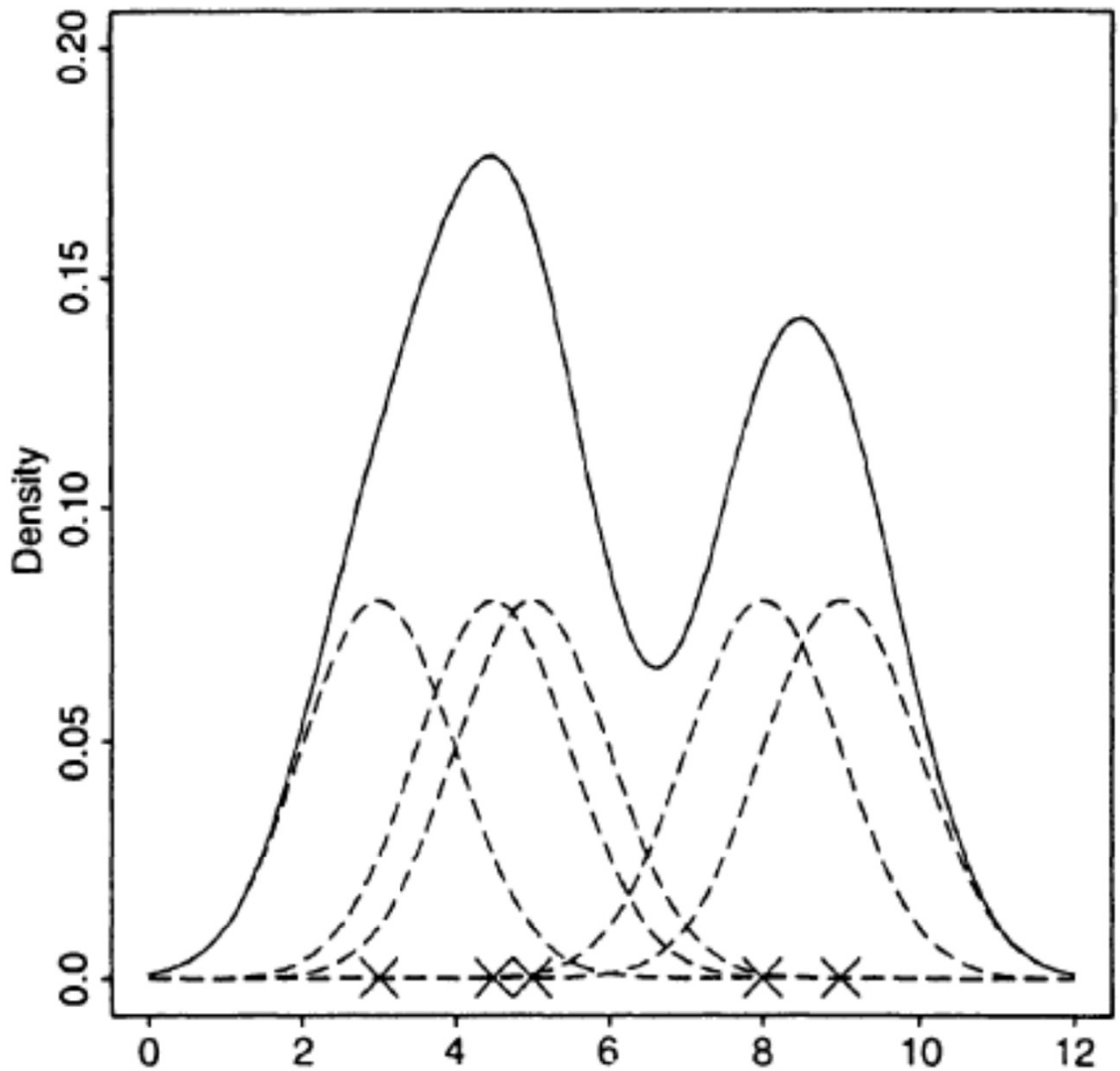
What is p(3)? Answer is 0.21

Given: $x1 = 2$, $x2 = 2.5$, $x3 = 3$, $x4 = 1$ and $x5 = 6$



What is p(x) in general?
Answer is the estimated curve

Kernel density estimate based on five observations

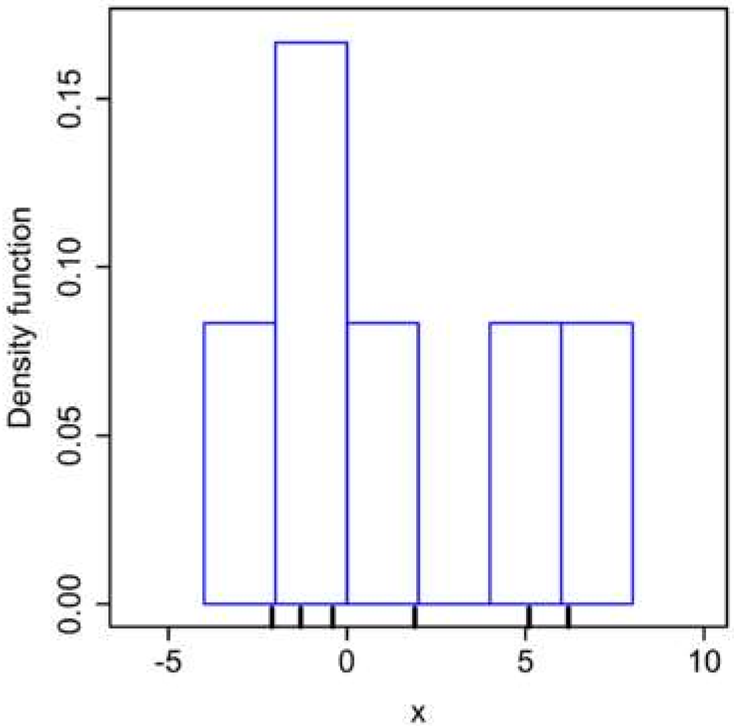


Parzen Window

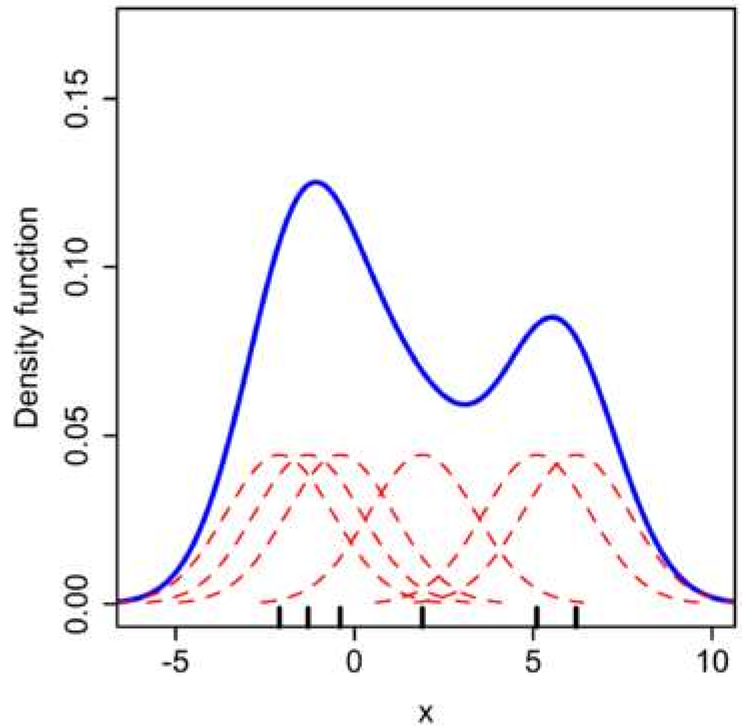
- Number of data points (n)
- n Gaussian computations to calculate the density @ a point
- To find density @ n points we need n x n i.e. n^2 calculations
- Instead of 1 dimensional data if we have d d dimensional then we need d times more computations.

Sample No.	Value
1	-2.1
2	-1.3
3	-0.4
4	1.9
5	5.1
6	6.2

Histogram-based Density

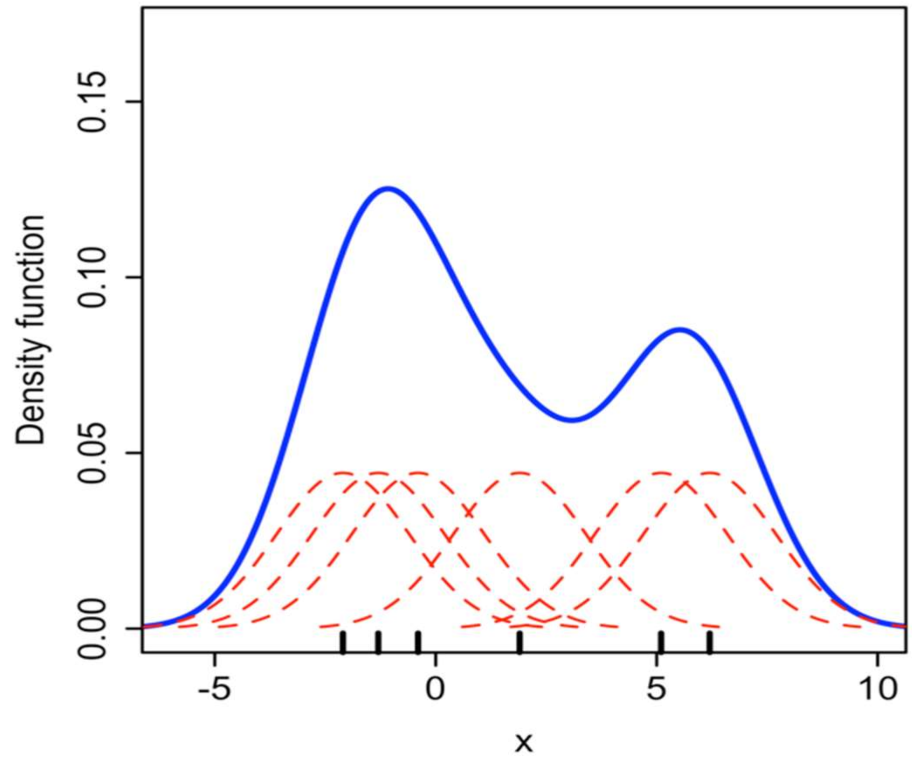


Parzen window-based Density



Kernel density estimation

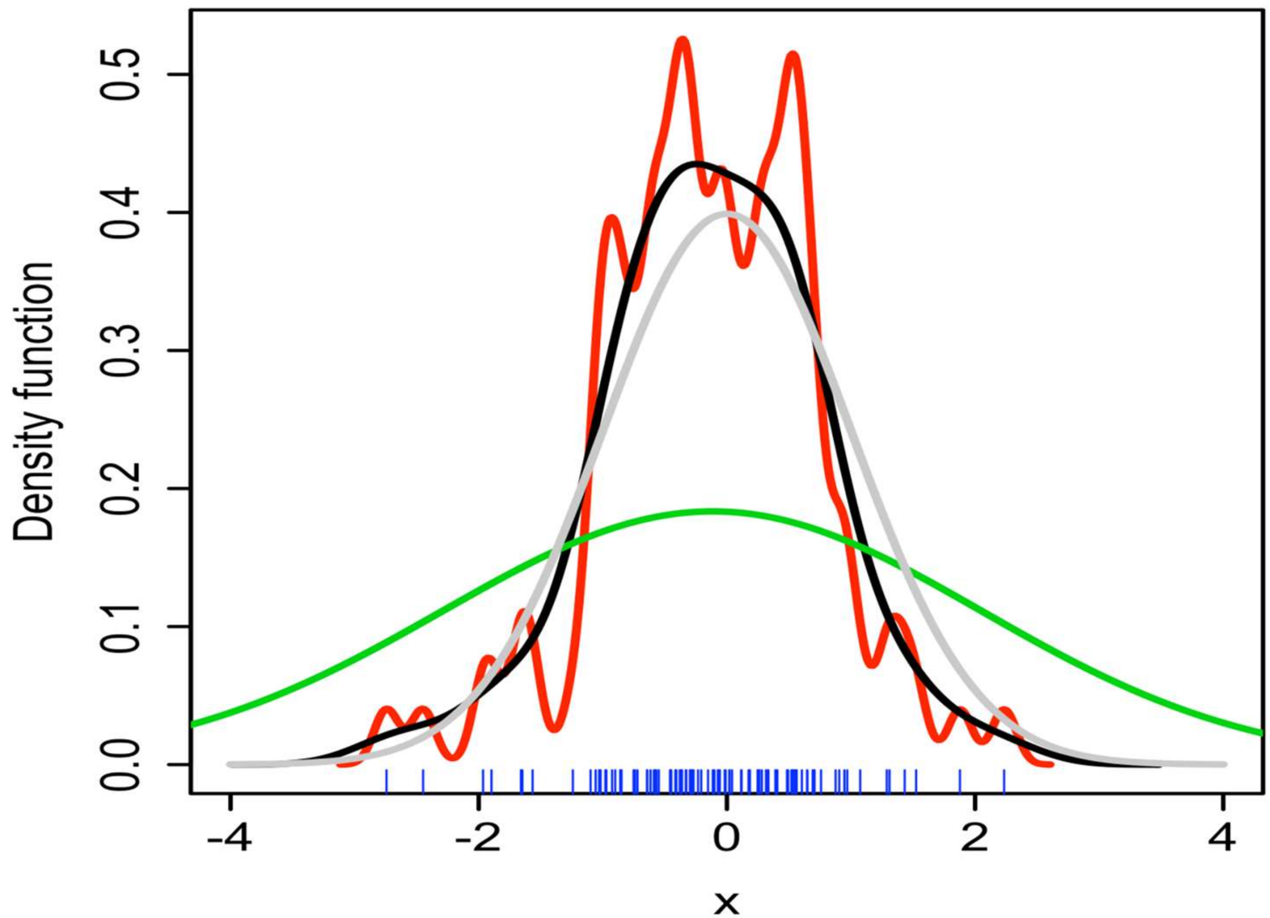
- Normal kernel with variance 2.25 on each of the data points x_i
- Kernels are summed to make the kernel density estimate



Bandwidth

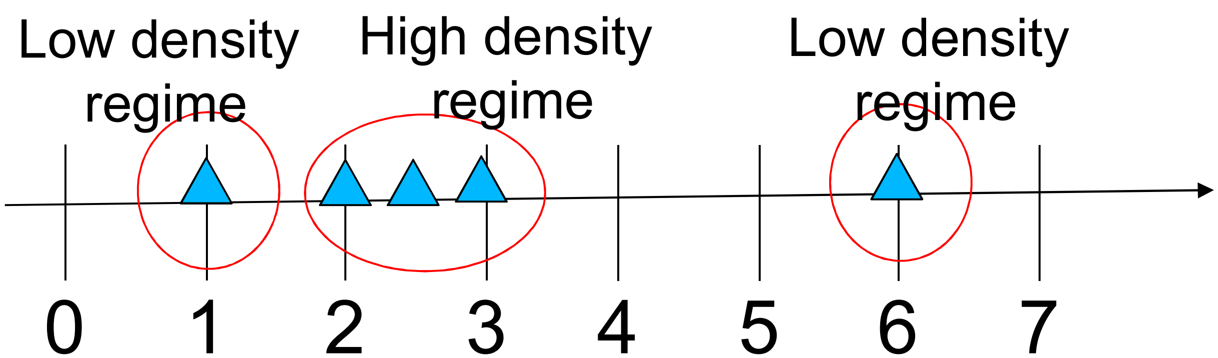
- Variance/bandwidth of kernel – free parameter
- Important parameter – decides the smoothness of estimation

Gray: True distribution
Black: appropriate estimation
red: too small bandwidth
green: too high bandwidth



Parzen density estimation - problems

- kernel width (h) is fixed in all regimes
  - High density regime
  - Low density regime
- Large h in high density regime – over smoothing
- Low h in low density region – noisy estimates



Nearest-neighbour methods

- kernel width (h) NOT fixed
  - High density regime – low h
  - Low density regime – high k
- In other words – choose h to accommodate fixed points