Lecture 24

Data : Collection of numbers obtained as an observation of specific system

Consider • "free" data

$y_1, y_2, \ldots, y_n$ n numbers

Here, all you have is n observations

• "parameterized" data

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ n pair of numbers

OR $(\underline{x}_1, \underline{y}_1), (\underline{x}_2, \underline{y}_2), \ldots, (\underline{x}_n, \underline{y}_n)$ n pairs of vectors

where $\underline{x}_i = (x_i^1, x_i^2, \ldots, x_i^l)$ l element vector

$\underline{y}_i = (y_i^1, y_i^2, \ldots, y_i^m)$ m element vector

Here, data y_i (or data vector $\underline{y}_i$) is obtained under parameters x_i (or parameter vector $\underline{x}_i$)

I.e. each data have associated parameter

Example: (1.) flip coins and record head (0) or tail (1)

So $(y_1, y_2, \ldots, y_n)$

where y_i is either 0 or 1

(2.) Consider m different coins manufactured (so each coin will be slightly different from another) then record

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

where
x_i is either coin 1, coin 2, ..., or coin m

y_i is either 0 or 1.

(3.) Measure drag coefficient by throwing an object of fixed shape & size n times and measuring C_d

$C_{d_1}, C_{d_2}, \ldots, C_{d_n}$ n observed drag coefficients

(4.) Measure drag coefficient by throwing an object of fixed shape & size n times

$(m_1, C_{d_1}), (m_2, C_{d_2}), \ldots, (m_n, C_{d_n})$

where
m_i = mass of an object in i^{th} experiment

Set containing data values A set of numbers from which each observation value is drawn.

think of more examples ←

Discrete set : for coin flipping, the set is $\{0, 1\}$

Continuous / Continuum / Real set : for drag coefficient, set is $\{x > 0\}$ of any positive number

• Statistics of the data

Consider $y_1, y_2, \ldots, y_n$ n data

• Mean (arithmetic mean)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

for drag coefficients $C_{d_1}, \ldots, C_{d_n}$

$$\bar{C_d} = \frac{1}{n} \sum_i C_{d_i}$$

but (m_i, C_{d_i})

$$\tilde{C_d} = \frac{\sum_i m_i C_{d_i}}{\sum_i m_i}$$

this is weighted mean

$m_1 = m_2 = \cdots = m_n = 1$

$$\bar{C_d} = \frac{\sum_i C_{d_i}}{n}$$

• Median (50th percentile of data)

arrange in increasing order

$$a_1 < a_2 < a_3 \cdots < a_n$$

then if n is odd a_i where $i = \frac{n+1}{2}$ is median

if n is even $\frac{a_i + a_{i+1}}{2}$ where $i = \frac{n}{2}$

• Mode

value in data that appears most frequently

• Spread of data

while mean, mode, median etc. inform about the "center" or "key" value of data, we also want to know how large the values in data can vary from each other.

Example: two samples have same mean (zero mean)

(i) 0, -0.1, 0.1, 0.2, -0.25, 0.35, 0.3

(ii) 0, −1, 1, 4, −6, 8, −6

standard deviation (std)

$$\sigma_y = \sigma = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

Variance : $$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

$\Rightarrow y_1$

$\bar{y} = y_1$

σ X

$\Rightarrow y_1, y_2$

$\bar{y} = \frac{y_1 + y_2}{2}$

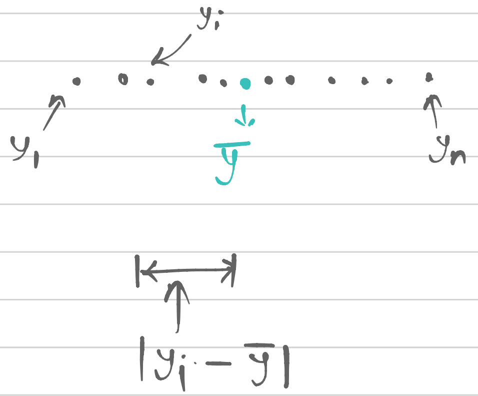
$\sigma = \sqrt{(y_1-\bar{y})^2 + (y_2-\bar{y})^2}$

$= \sqrt{2(y_1-\bar{y})^2}$

$= |y_1 - \bar{y}|\sqrt{2}$

$= \frac{|y_2 - y_1|}{2}\sqrt{2}$

$= \frac{|y_2-y_1|}{\sqrt{2}}$



$$\frac{\sqrt{\sum_{i=1}^{n} |y_i - \bar{y}|^2}}{\sqrt{(n-1)}}$$

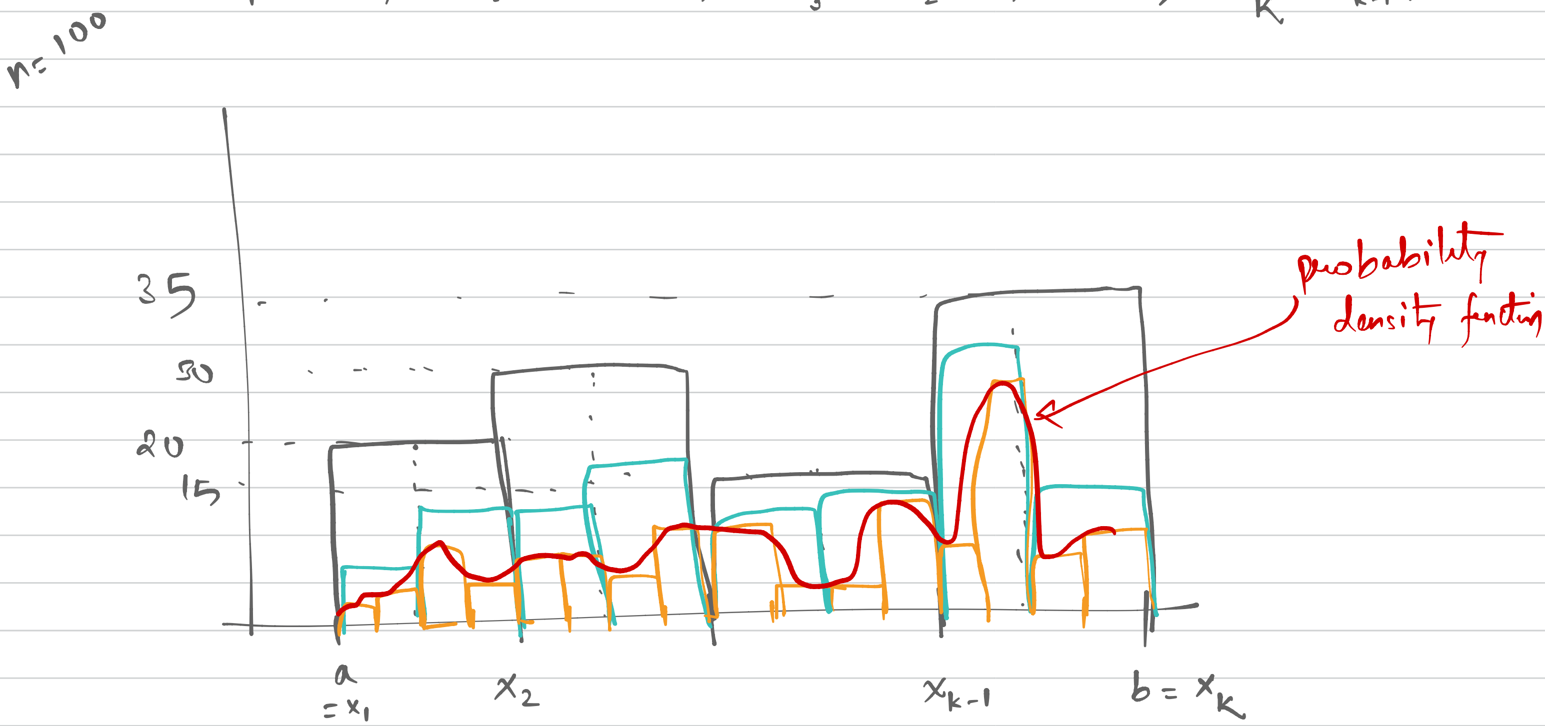
$C_{d_1}, \quad C_{d_2}, \quad \ldots, \quad C_{d_n}$

$C_{d_i} \in [a, b]$

<u>Goal</u>: Given $x \in [a,b]$, what is a probability that I would observe $C_d = x$ as drag coefficient

<u>Step 1</u>: Take smaller intervals in $[a, b]$

$x_1 = a, \quad x_2 = x_1 + h, \quad x_3 = x_2 + h, \quad \ldots, \quad x_K = x_{K-1} + h$



<u>for Given ith observation</u>, C_{d_i}, find interval that C_{d_i} belongs to.

I.e. find l, $1 \le l \le K-1$, s.t.

$C_{d_i} \in [x_l, x_{l+1})$

Set 1 $[x_1, x_2)$

Set 2 $[x_2, x_3)$

⋮

set k-1 $[x_{k-1}, x_k]$

Given d_i, find set l s.t. $d_i \in [x_l, x_{l+1})$

for all $d_1, \ldots, d_n$

Count

how many times d_i was in set 1

——//—— set 2

⋮

——//—— set k-1

Probability density function

$$f : [a, b] \to [0, \infty)$$

$f(x)$ is the probability of "x" being observed

$$\int_a^b f(x)\,dx = 1$$