

# MATHEMATICAL FOUNDATIONS FOR DATA SCIENCE

## FUNCTIONS

Set - unordered collection, unique items

$x \in S, x \notin S, \{x \in S : P(x)\}, [a, b], (a, b)$

$\downarrow$                $\downarrow$                $\downarrow$   
 cond<sup>n</sup>      closed      open  
 interval

$$\sum_{x=a}^b f(x) = f(a) + f(a+1) + \dots + f(b-1) + f(b)$$

$$\prod_{x=a}^b f(x) = f(a) \times f(a+1) \times \dots \times f(b-1) \times f(b)$$

$f: X \rightarrow Y$   
 $\downarrow$                $\downarrow$   
 domain      co-domain  
 $\downarrow$   
 all possible outputs

$\{f(x) : x \in X\}$   
 $\downarrow$   
 image (set of outputs)  
 $\downarrow$   
 for given range of inputs

Zeros of  $f(x) = \{x \in X : f(x) = 0\}$   
 $\downarrow$   
 roots

Inverse :  $f^{-1} \quad \Gamma f: X \rightarrow Y$

$\text{L } f^{-1}: \mathcal{Y} \rightarrow \mathcal{X}$

$$f^{-1}(f(x)) = x \quad f^{-1}(y) \neq \frac{1}{x}$$

## TYPES OF FUNCTION

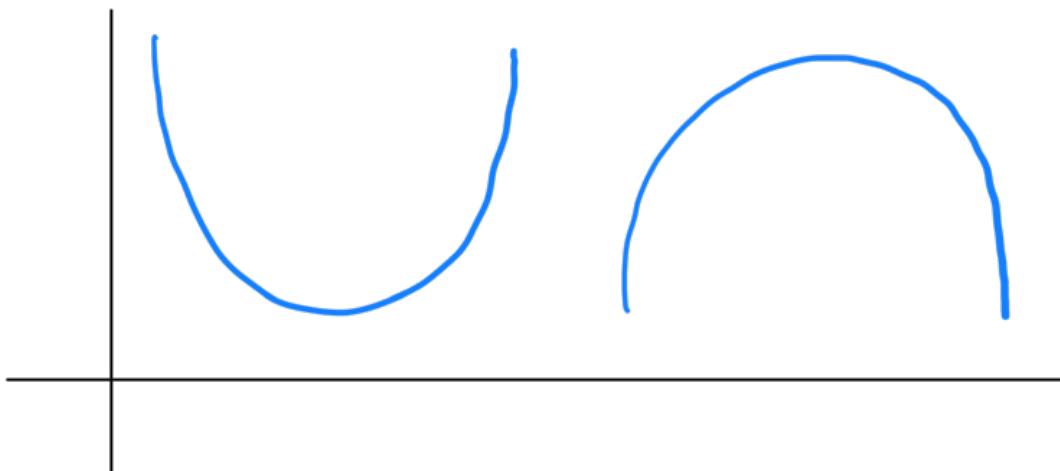
(A) Injective - One to One ①

Surjective - for every  $f^{-1}(y)$  there exists a  $x \in \mathcal{X}$  ②

Bijective - BOTH ① and ②

(B) Convex

Concave



(C) Linear

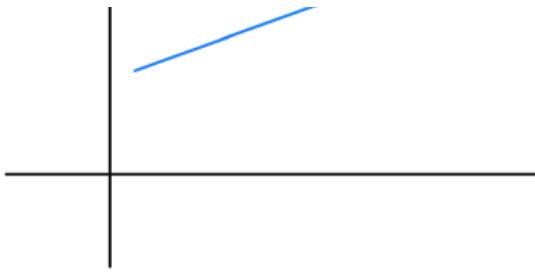
$$f(x) = mx + b$$

Only 1 zero at  $x = -b/m$

Bijective if  $m \neq 0$

both convex and concave





Polynomial

$$f(x) = \sum_{i=0}^n a_i x^i$$

$$= a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$$



$$x^0 = 1$$

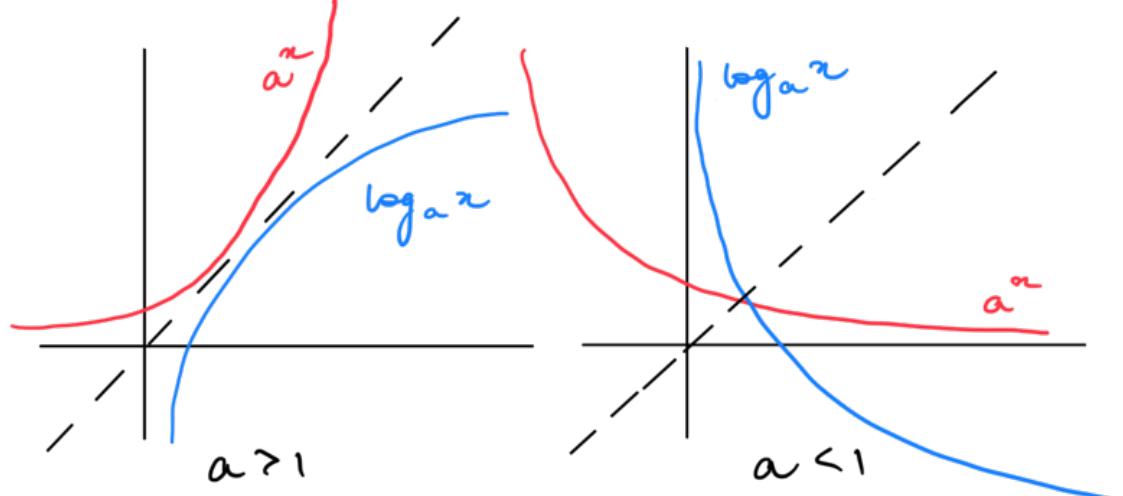
Exponential

$$f(x) = b \cdot a^x$$

Logarithmic

$$f(x) = \log_a(x)$$

} inverse of each other  
if  $f(x) = a^x$   
 $f^{-1}(x) = \log_a(x)$



$$\log_a(mn) = \log_a m + \log_a n$$

$$\log_a(m/n) = \log_a m - \log_a n$$

$$1 - m = 1 - m$$

$$\log_a x^n$$

$$\frac{\log_b x^n}{\log_b a}$$

$$\log_a x^m = m \log_a x \rightarrow \log_a (a^n) = n$$

$$a^{\log_a x} = x$$

Power law

$$f(x) = b \cdot x^{-a}$$

exp  $f(x)$  decay faster than power laws

$$\downarrow \quad \log \rightarrow \text{power law} \rightarrow \text{st. line}$$

## DIFFERENTIATION

### LINEAR TRANSFORMATION

\* log  $\log_a(x)$

\* log - log power

data  $(x_1, y_1) \dots (x_n, y_n)$

plot  $(\ln x_1, \ln y_1) \dots (\ln x_n, \ln y_n)$

$$f(x) = b x^{-a} \quad [f(x) = y]$$

$$\ln(y) = \ln b + (-a) \ln x$$

$$\hat{y} = \ln b - a \hat{x}$$

$$m = (-a) \quad \text{and} \quad b = \ln(b)$$

\* semi-log



log lin    lin log  
 exp              log

log lin

$$(x_1, \ln(y_1) \dots x_n, \ln(y_n))$$

$$f(x) = b \cdot a^x$$

$$\ln(y) = \ln b + x \ln a$$

$$\hat{y} = (\ln a)x + \ln b$$

$$m = \ln a \quad \text{and} \quad b = \ln(b)$$

lin log

$$(\ln(x_1), y_1 \dots \ln(x_n), y_n)$$

$$\begin{aligned} f(x) &= b \cdot \log_a x \\ &= b \cdot \frac{\ln x}{\ln a} \end{aligned}$$

$$y = \frac{b}{\ln a} (\ln x)$$

$$\hat{y} = \frac{b}{\ln a} \cdot \tilde{x} \qquad m = b/\ln(a)$$

## DERIVATIVE

$f(x) \longrightarrow f'(x)$  : slope of tangent

$$x^n$$

$$a^n$$

$$\log_a n$$

$$n x^{n-1}$$

$$\ln a \cdot a^n$$

$$\frac{1}{\ln a \cdot n}$$

$$\begin{aligned} e^x &\rightarrow e^x \\ \ln x &\rightarrow \frac{1}{x} \end{aligned}$$

$$f(x) + g(x)$$

$$f'(x) + g'(x)$$

$$\begin{array}{ll} \text{if } u \\ f(x) = g(x) & f'(x) = g'(x) \\ f(x)g(x) \text{ prod rule} & f'(x) \cdot g(x) + f(x) \cdot g'(x) \\ f(g(x)) \text{ chain rule} & g'(x) \cdot f'(g(x)) \end{array}$$

$|x|$  is not differentiable

$$f(x) = x^3 \quad f'(x) = 3x^2 \quad f''(x) = 6x \quad f'''(x) = 6$$

## OPTIMISING FUNCTION

### INCREASING FUNCTION

$x \uparrow$  and  $f(x) \uparrow$

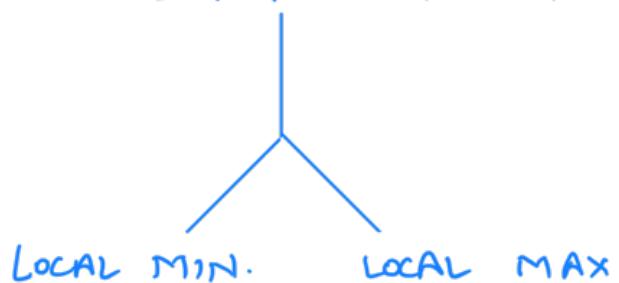
$$f'(x) > 0$$

### DECREASING FUNCTION

$x \downarrow$  and  $f(x) \downarrow$

$$f'(x) < 0$$

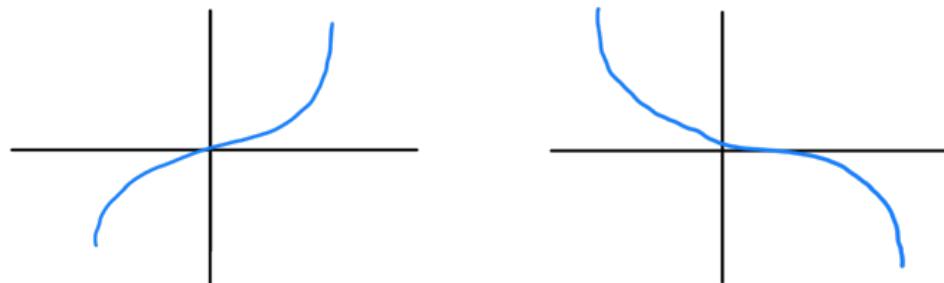
$x$  where  $f'(x) = 0$  : STATIONARY POINT



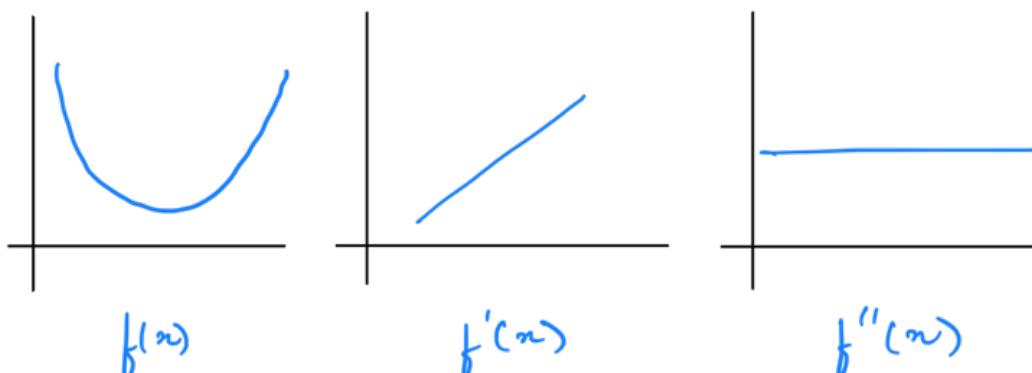
$f'(x)$  changes from -ve to +ve      +ve to -ve



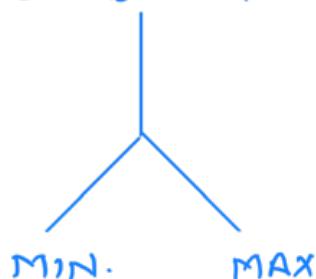
Inflection point - st. pt. but neither local max nor local min



ex-



### GLOBAL EXTREMUM



$$f(a) \leq f(x) \quad f(a) \geq f(x) \quad \forall x \in X$$

for a  $f(x)$  with  $x \in [a, b]$ , we may get global extrema at :

end pt. 1.  $a$  or  $b$

st. pt. 2.  $c$  where  $f'(c) = 0$

critical pt. 3.  $c$  where  $f'(c)$  does not exist

ii.

$$f''(x) > 0$$

Convexity

$$f'(x) < 0 \quad \text{concavity}$$

$\therefore$  if a  $f(x)$  is completely concave or completely convex, then its local extrema is also the global extrema.

↪ if NOT, break into sub domains

## RESIDUAL SUM OF SQUARES

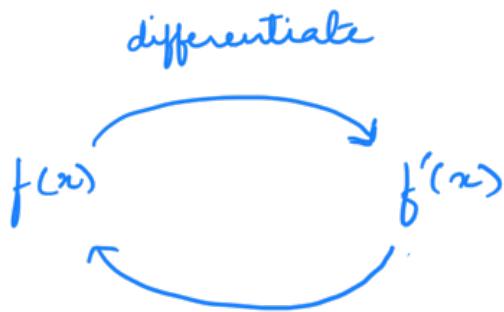
for data  $(x_1, y_1) \dots (x_n, y_n)$

$$\text{RSS} = \sum_{i=1}^n (y_i - f(x_i))^2$$

Squaring to measure deviations?

$\therefore$  it allows finding derivatives and penalise [large] deviations heavily

## INTEGRATION



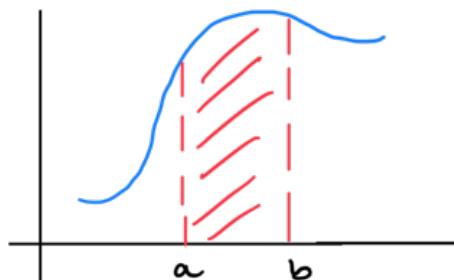
integrate



anti-derivative

$$\int_a^b f(x) \cdot dx = F(b) - F(a)$$

$$\frac{1}{a+1} x^{a+1}$$



-area under the curve



area below x-axis is -ve

$\frac{1}{x}$

$\dots$

$e^{\alpha n}$

$\frac{1}{a} e^{\alpha n}$

## LINEAR ALGEBRA

### VECTORS

$$\mathbb{R}^d = \left\{ \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} : x_1, x_2, \dots, x_n \in \mathbb{R}^d \right\}$$

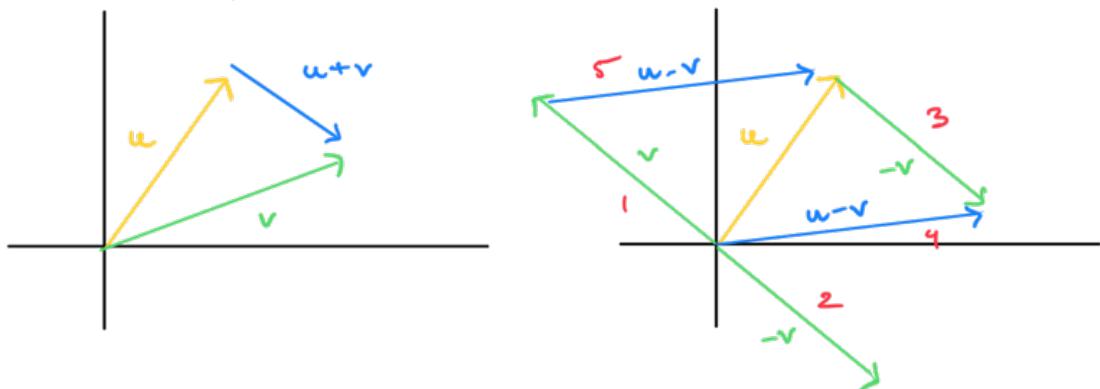
↓  
set of d tuples

ex-  $\mathbb{R}^3$  contains  $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ ,  $\begin{pmatrix} 2 & 5 \\ 4 \\ 4 \end{pmatrix}$  and  $\begin{pmatrix} c \\ z \\ x \end{pmatrix}$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

$$c \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{pmatrix}$$

a vector  $(v, \vec{v}, \tilde{v})$  has length and dir<sup>n</sup>  
but no pos<sup>n</sup>



A line joining points  $u$  and  $v$

contains v contains the points corresponding to  $\alpha u + (1-\alpha)v$  where  $\alpha \in [0,1]$

$$w = a_1v_1 + a_2v_2 + \dots + a_nv_n$$



linear comb<sup>n</sup> of  $v_1 \dots v_n$

→ linear dependence leads to redundancy

If not, then linearly Independent

ex-  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$  &  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$  are linearly independent

but  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ ,  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$  and  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$  are not.

as  $\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

if  $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$        $w = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$

$$v \cdot w = (v, w)$$

$$= v_1w_1 + v_2w_2 + \dots + v_nw_n$$

if  $v \cdot w = 0$  then v and w are orthogonal  
i.e. perpendicular

### EUCLIDEAN NORM

$$\|v\| = \sqrt{v_1^2 + v_2^2 + v_3^2 + \dots + v_n^2}$$

$$= \sqrt{v \cdot v}$$

$\vdots \quad \dots \quad \sim \quad \sim \quad \sim \quad \sim \quad \sim$

we can't use the term length in case of multiple dimensions therefore it is referred to as NORM

## MATRICES

$(m \times n)$  matrix where  $m = \text{rows}$   
 $n = \text{columns}$ .

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{pmatrix}$$

$$c \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} ca_{11} & ca_{12} \\ ca_{21} & ca_{22} \end{pmatrix}$$

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$A \cdot B = \begin{pmatrix} a_{11} \cdot b_{11} + a_{12} \cdot b_{21} & a_{11} \cdot b_{12} + a_{12} \cdot b_{22} \\ a_{21} \cdot b_{11} + a_{22} \cdot b_{21} & a_{21} \cdot b_{12} + a_{22} \cdot b_{22} \end{pmatrix}$$

$$M_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad M_2 = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

$$\det M_1 = ad - bc$$

$$\det M_2 = a(ei - fh) - b(di - gf) + c(de - gf)$$

$$\text{Identity Matrix (I)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\text{Zero Matrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

if  $B \cdot A = I$  then  $B = A^{-1}$  and  $A = B^{-1} \rightarrow AB = I$

$$A \cdot B \neq B \cdot A$$

$$\det(AB) = \det(A) \cdot \det(B)$$

$$\det(I) = 1 \\ \det(A) \neq 0 \quad \text{if } A^{-1} \text{ exists}$$

→ let  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$

then  $A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

$$Ax = b \Rightarrow x = A^{-1}b.$$

## GAUSSIAN ELIMINATION

- Steps -
1. Swap 2 rows
  2. Mul a row by non zero num.
  3. Add a multiple of one row to another.

to achieve

$$\begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} * \\ * \\ * \end{pmatrix}$$

for ex-  $R_1 \leftrightarrow R_2$

$$R_2 \rightarrow 5 \cdot R_2$$

$$R_3 \rightarrow 2R_3 - R_2$$

Types of Solution :

1. Exactly ONE solution
2. NO solution

$$\begin{pmatrix} * & * \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} * \\ * \end{pmatrix}$$

$*$  = non zero num.  $0x + 0y = *$

$0 \neq *$

### 3. MULTIPLE Solutions

$$\text{ex} - \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\begin{aligned} x + z &= 1 \\ y - 2z &= 0 \end{aligned}$$

$$\Rightarrow x = 1 - z \quad \text{and} \quad y = 2z \\ \text{put } z = t$$

$$\text{Sol}^n = (1-t, 2t, t)$$

→ free variable : INFINITE sol<sup>n</sup>(s)

### EIGEN VALUES, VECTOR

$Ax = \lambda x \rightarrow \text{eigen vector (can't be zero)}$



eigen value

$$\Rightarrow (A - \lambda I)x = 0 \quad ①$$

$$\det(A - \lambda I) = 0 \quad \text{characteristic eqn} \quad ②$$

$n \times n$  matrix  $\rightarrow n$  sol<sup>n</sup>(s)  $\rightarrow n$  eigen values

Eigen values aren't unique (multiple of d)

Steps -

1. find  $\lambda$  from  $\det(A - \lambda I) = 0$

2. use  $\lambda$  to find  $x$  from  $(A - \lambda I)x = 0$

$A$  is diagonalisable if  $A = PDP^{-1}$  and

where  $\begin{pmatrix} x & 0 & 0 \\ 0 & y & 0 \\ 0 & 0 & z \end{pmatrix}^n = \begin{pmatrix} x^n & 0 & 0 \\ 0 & y^n & 0 \\ 0 & 0 & z^n \end{pmatrix}$  we get.

$$A^n = P D^n P^{-1}$$

↓      |      ↗  
 nxn matrix       $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$  upto  $\lambda_n$   
 ↓

$(v_1, v_2, \dots, v_n)$

## BINARY RELATIONS

a binary rel<sup>n</sup> relates 2 parameters

e.g. points in xy plane (1, 2)

every  $f(x) \xrightarrow{\text{GIVES}}$  rel<sup>n</sup> but NOT conversely

$$\text{eqn of circle : } (x-a)^2 + (y-b)^2 = r^2$$

$$\text{ellipse : } \frac{(x-a)^2}{a^2} + \frac{(y-b)^2}{b^2} = r^2$$

centre at  $(a, b)$

## MULTIVARIATE FUNCTION

$$z = f(x, y)$$

$$f_x = \frac{\partial}{\partial x}(f)$$

$$f_y = \frac{\partial}{\partial y}(f)$$

⇒ PARTIAL DERIVATIVE

$$\nabla f = \begin{bmatrix} f_x \\ f_y \end{bmatrix} \rightarrow \text{GRADIENT VECTOR}$$

## OPTIMISING MULT. FUNCTIONS

$$f(x + \Delta x) = f(x) + f'(x) \Delta x$$

$$f(x + \Delta x, y + \Delta y) \approx f(x, y) + \\ f_x(x, y) \Delta x + \\ f_y(x, y) \Delta y$$

## STATIONARY POINT

$f(x, y)$  where  $f_x = 0$  and  $f_y = 0$

implying  $\nabla f(x, y) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$  and is either local extrema or saddle pt.

## HESSIAN MATRIX

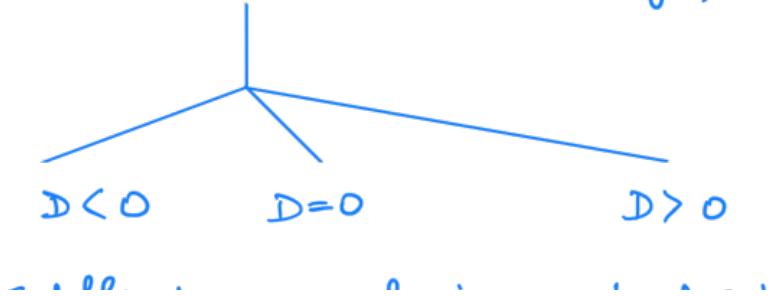
$$H(x, y) = \begin{bmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{bmatrix}$$

$\downarrow$

Second partial derivative

if  $f_{xy} = f_{yx}$  the nice funct

$$D = \det(H(x, y))$$



Saddle pt. inconclusive local extremum



$$f_{xx} > 0 \quad f_{xx} < 0$$

local min local max.

## GLOBAL EXTREMA

possible candidates - 1. stationary pts  
 $f_x$  &  $f_y$  are undefined ← 2. singular pts  
3. boundary

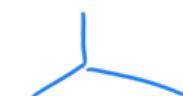
## COMBINATORICS

Multiplication Rule  $|S| = \prod_{j=1}^n k_j$   $k \neq 1$

Addition Rule  $|S| = \sum_{j=1}^n s_j$   $k + l$

Complement Rule  $S = S_g \cup S_b \quad \left\{ \begin{array}{l} |S| \\ S_g \cap S_b = \emptyset \end{array} \right. \quad \left\{ \begin{array}{l} 1 - k \\ S_g = S - S_b \end{array} \right.$

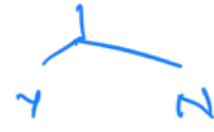
## TYPES OF SELECTION (4)



order matters

repn allowed

$\downarrow$   
 $AB \neq BA$



ordered sel<sup>n</sup> w/o rep<sup>n</sup>

$$\frac{n!}{(n-r)!}$$

unordered sel<sup>n</sup> w/o rep<sup>n</sup>  $\frac{n!}{r!(n-r)!}$

ordered sel<sup>n</sup> with rep<sup>n</sup>  $n^r$

unordered sel<sup>n</sup> with rep<sup>r</sup>  $\frac{(n+r-1)!}{r!(n-1)!}$

## PASCAL'S TRIANGLE

$$\binom{n}{r} = \binom{n}{n-r}$$

## PIGEON HOLE PRINCIPLE

1. n items placed in m containers
2.  $n > m$
3. at least one container has  $\lceil \frac{n}{m} \rceil$  items

↓  
rounding up.

## PROBABILITY

Sample Space : set of possible outcomes

Pr:  $S \rightarrow [0, 1]$  prob. funct

sum of prob.(s) of outcomes = 1

uniform prob. space : each event has equal prob.

Event - subset of Sample Space.

$$\text{ex - } \Pr(A) = \frac{\text{no. of outcomes in } A}{\text{total no. of outcomes}}$$
$$= \frac{|A|}{|S|}$$

$$\Pr(\bar{A}) = 1 - \Pr(A)$$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

if  $\Pr(A \cap B) = 0 \rightarrow$  mutually exclusive

i.e. A and B cannot occur together

Independent Events :  $\Pr(A \cap B) = \Pr(A)\Pr(B)$

if 1 true then all true :

1. A and B are independent
2.  $\bar{A}$  and  $\bar{B}$  are independent
3. A and  $\bar{B}$  are independent
4.  $\bar{A}$  and B are independent

## INDEPENDENT REPEATED TRIALS

Sample Space =  $S_1 \times S_2$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

$$= \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}$$

if independent,  $\Pr(A|B) = \Pr(A)$

## BAYES THEOREM

$$\begin{aligned}\Pr(A|B) &= \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B|A) \cdot \Pr(A) + \Pr(B|\bar{A}) \cdot \Pr(\bar{A})} \\ &= \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}\end{aligned}$$

## ADVANCED PROBABILITY

### RANDOM VARIABLE

$P(X=a)$  a funct<sup>n</sup> from sample space to R

Random vars are independent if & only if

$$\Pr(X=a \text{ and } Y=b) = \Pr(X=a) \cdot \Pr(Y=b)$$

$\vdots \quad \vdots \quad p_1 \quad \vdots \quad \vdots \quad p_n$

→ Expected Value

$$E[x] = p_1 x_1 + p_2 x_2 + \dots + p_n x_n$$



weights / probabilities

law of large no.s -

$$\lim_{n \rightarrow \infty} \frac{1}{n} (x_1 + x_2 + \dots + x_n) = \mu$$

if  $E[x] = \mu$

$$\text{Var}[x] = E[(x-\mu)^2] = E[x^2] - \mu^2$$

↓  $\sigma^2 = \text{Var}[x]$  → std. dev.  
variance

If  $x$  and  $y$  are independent :

$$E[x+y] = E[x] + E[y] \rightarrow \text{linearity of Expect.}$$

$$E[x \cdot y] = E[x] \cdot E[y]$$

$$E[kx] = k E[x]$$

$$\text{Var}[kx] = k^2 \cdot \text{Var}[x]$$

$$\text{Var}[x+y] = \text{Var}[x] + \text{Var}[y]$$

### DISCRETE UNIFORM DISTRIBUTION

set of cons. integers (equally likely)

$$P(x=k) = \frac{1}{b-a+1} \quad \text{for } k \in \{a, a+1, \dots, b\}$$

$$E[x] = \frac{a+b}{2} \quad \text{Var}[x] = \frac{(b-a+1)^2 - 1}{12}$$

## BERNOULLI DISTRIBUTION

prob. that a process succeeds or fails

$$P(X=k) = \begin{cases} p & k=1 \text{ Success } p \in [0,1] \\ 1-p & k=0 \text{ fail} \end{cases}$$

$$E[X] = p \quad \text{Var}[X] = p(1-p)$$

## GEOMETRIC DISTRIBUTION

Seq of Bernoulli trials -  $k$  failures before first success.

$$P(X=k) = p(1-p)^k \text{ for } k \in \mathbb{N}$$

$$E[X] = (1-p)/p \quad \text{Var}[X] = (1-p)/p^2$$

## BINOMIAL DISTRIBUTION

Seq of  $n$  independent Bernoulli Trials with  $k$  success.

$$n \in \mathbb{Z}^+ \text{ and } p \in [0,1]$$

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k \in \{0, n\}$$

## POISSON DISTRIBUTION

with an avg of  $\lambda$  events per unit time, it gives the prob. that  $k$  events occur in unit time.

$$P(X=k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \quad \text{for } k \in \mathbb{N}$$

$$E[X] = \lambda \quad \text{Var}[X] = \lambda$$

## CONTINUOUS DISTRIBUTION

Probability Density Function

→ PDF can not be negative

→ Integral of the PDF = 1 (over whole domain)

$$P(a \leq x \leq b) = \int_a^b f \cdot dx$$

$$E[x] = \int_{-\infty}^{\infty} x \cdot f dx \quad \text{Var. stays same.}$$

A continuous uniform dist has following PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = (a+b)/2 \quad \text{Var}[x] = (b-a)^2/12$$

## EXPONENTIAL DISTRIBUTION

Returns the time b/w events in a Poisson process.

→ randomly spread out with even density

$$\text{PDF: } \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

$$E[x] = 1/\lambda \quad \text{Var}[x] = 1/\lambda^2$$

## NORMAL DISTRIBUTION

$$\text{PDF: } f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$E[x] = \mu \quad \text{Var} = \sigma^2$$

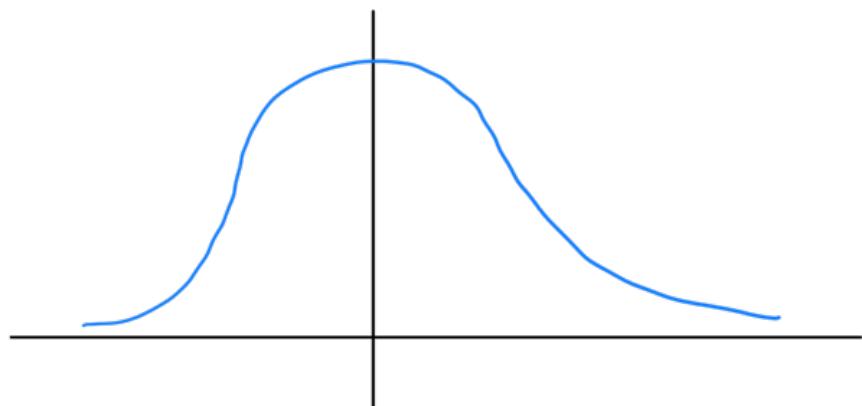
many discrete dist(s) tend towards a normal dist

- - . . . /

$$z = (x - \mu)/\sigma$$

how many std. deviations above or below the mean  $x$  is.

$$\Pr(a < x < b) = \Pr\left(\frac{a-\mu}{\sigma} < z < \frac{b-\mu}{\sigma}\right)$$



## GRAPH THEORY

Graph - vertices + edges

Vertices - may or may not be labelled

vertices sharing an edge are labelled as ADJACENT

Edge - connects b/w 2 vertices

loop - vertex related to itself

Parallel Edges - multiple routes b/w same vertices

loop



parallel

A  B edges.

Multi-graph: loops & parallel allowed

Simple graph: NO loops / parallel edges

Digraph - directed graph (dir<sup>n</sup> on edges)

Walks - Sequence of vertices (adjacent)

Length of walk - no. of steps.

Path - walk with distinct vertices



Connected graph



Disconnected graph

Cycle - start vertex is the end ver.

Degree (of a vertex) - no. of edges that include vertex

Regular graph - If every vertex has same degree  $k$ .  
k-regular graph.

### HANDSHAKING LEMMA

In any graph,

Sum of degrees =  $2 \times$  (no. of edges)

$\Rightarrow$  Even no. of vertices of odd degree.

### TREE

Connected graph with no cycles

Prop(s) - If one true, all true:

(i) T is a tree

(ii) Any 2 vertices - linked by unique path

(iii) Deleting an edge - DISCONNECTED

(iv) Adding an edge - CYCLE

n vertices - (n-1) edges

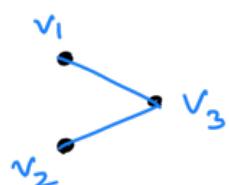
## SPANNING TREE

a tree contained in graph (includes all vertices)



Every connected graph contains a ST.

## ADJACENCY MATRIX



$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

$$a_{ij} = \begin{cases} 0 & \text{adjacent} \\ 1 & \text{not adjacent} \end{cases}$$

## EULER CIRCUIT

closed walk - uses every edge once

A connected graph is Eulerian iff every vertex has even degree

Euler Trail - start and end are diff

A connected graph has Euler Trail iff at most 2 vertices have odd degrees

## STATISTICAL DATA MODELLING

model = rep<sup>"</sup> of real world prob.  
= allowed exp<sup>"</sup>, analysis  
probabilistic

Keywords - probability  
expected value  
variance

Chebychev ineq.  
central limit thm.

hypothesis  
predictive modelling

frequentist approach - counting

median =  $x_{(n+1)/2}$  or  $\frac{1}{2}(x_{n/2} + x_{n/2+1})$   
 $n$  is odd                       $n$  is even

quartile ( $Q_k$ ) =  $x_p + \frac{q}{4}(x_{p+1} - x_p)$   
 $p = \text{floor}((k(n+1))/4)$   
 $q = (k(n+1)) \bmod 4$

variance ( $s_x^2$ ) =  $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

standard deviation =  $s_x$

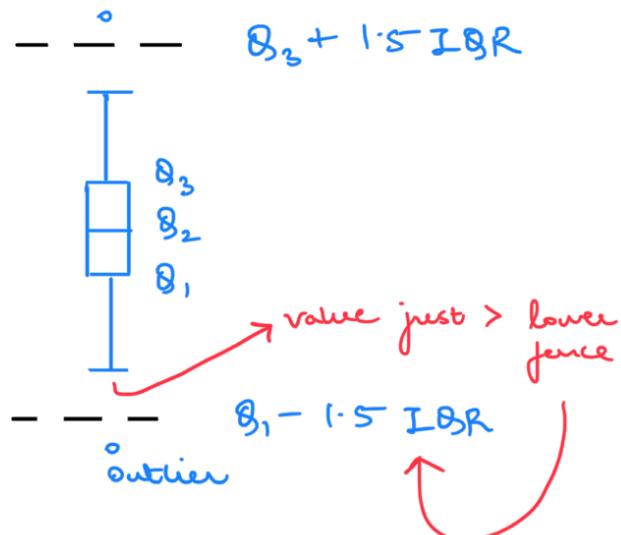
range =  $\max(x_i) - \min(x_i)$

$$IQR = Q_3 - Q_1$$

$$\text{covariance } (q_{xy}) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{correlation coeff. } (r_{xy}) = \frac{q_{xy}}{s_x s_y}$$

Boxplot.



## PROBABILITY

random variables (experiment output)

randomness due to unmeasured factor  
ex- reading error.

confounding var — hidden factor that contributes in computing output  
(unsupervised)

majority of times unable to trace  
but if found it can be used  
to predict output  $\rightarrow$  hypothesis testing

$$P(x = x), x \in X$$

prob that a RV  $x$  takes on the value  $x$  from  $X$

$$\dots \cap P(x = x) \in \Gamma_{0.17} \vee x \in X$$

also  $\cup : (\cap - \cup) = \cup \cap \rightarrow \cup \cup$

$$\text{and } \sum_{x \in X} P(x = x) = 1$$

$$\textcircled{2} P(X \in A_1 \cup A_2) = P(X \in A_1)$$



Similarly for 2 RVs :

$$P(X = x) = \sum_{y \in Y} P(X = x, Y = y)$$

marginal pr.    joint pr.  
irrespective of Y

$$\sum_{x \in X, y \in Y} P(X = x, Y = y) = 1 \quad \text{from } \textcircled{1}$$

Conditional Probability

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Bayes Theorem

$$P(x | A) = \frac{P(A | x) P(x)}{P(A)}$$

$$\textcircled{1} \text{ if } P(X = x, Y = y) = P(X = x) P(Y = y)$$

$\Rightarrow X$  and  $Y$  are independent

$$\Rightarrow P(X = x | Y = y) = P(X = x)$$

\textcircled{2} a Special case :

Independent & identically dist. (IID) :

$$x_1 \in X, x_2 \in X$$

$$P(x_1 = x) = P(x_2 = x) \quad \forall x \in X$$

Continuous RV

- prob. density functn (pdf)

$$P(x = x) = f(x)$$

$$f(x) > 0 \quad \forall x \in X$$

$$P(a < x < b) = \int_a^b f(x) \cdot dx$$

$$P(X \in A) = \int_A f(x) \cdot dx$$

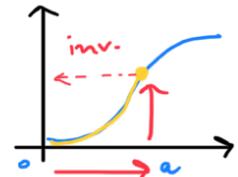
$$P(x = x) = 0 \quad (\text{slide 39})$$

$$\int_X f(x) \cdot dx = 1$$

- cumm. density functn (cdf)

$$P(x \leq x) = \int_{-\infty}^x f(x) \cdot dx$$

$$F(x) = \int_0^x f(x) \cdot dx$$



- inverse cdf (quantile functn)

$$g(x) = F^{-1}(x)$$

## EXPECTATION

Expected value of a dist<sup>n</sup>

↳ some char. known

$$E[x] = \sum_{x \in X} x \cdot p(x) \quad \text{discrete}$$

$$= \int x \cdot p(x) \cdot dx \quad \text{cont.}$$

$E[x^n]$  —  $n^{\text{th}}$  moment

$E[(x - E[x])^n]$  —  $n^{\text{th}}$  central moment

Variance — variat<sup>m</sup> around the mean.

↳ expected squared dev<sup>n</sup>

$$\begin{aligned} V[x] &= E[(x - E[x])^2] = E[x^2] - E[x]^2 \\ &= \sum_{x \in X} (x - E[x])^2 p(x) \quad \begin{matrix} \text{positive only} \\ (\sim \text{mean}) \end{matrix} \end{aligned}$$

Standard Deviation

$$\sigma_x = \sqrt{V[x]}$$

Linearity of Expectation :

$$E[cx] = cE[x]$$

$$V(cx) = c^2 V[x]$$

$$E[c] = c$$

$$V[c] = 0$$

$$E[af(x) + bg(y)] = aE[f(x)] + bE[g(y)]$$

$$E[g(x)E[f(x)]] = E[f(x)]E[g(x)]$$

Covariance  $(-\infty, \infty)$

$$\begin{aligned} \text{cov}(x, y) &= E[(x - E[x])(y - E[y])] \\ &= E[xy] - E[x]E[y] \end{aligned}$$

→ joint variability

0                    1

Correlation       $[\pm 1]$

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{v}[x] \text{v}[y]}}$$

$\rightarrow$  rel<sup>n</sup> b/w the 2

$$z_x = \frac{x - E[x]}{\sqrt{v[x]}}$$

$$\Rightarrow \text{corr}(x, y) = \text{cov}(z_x, z_y)$$

- strength of linearity
- +ve or -ve
- slope not relevant (due to standardising)

if  $x$  and  $y$  independent then  $\left\{ \begin{array}{l} \text{cov}(x, y) = \text{corr}(x, y) = 0 \\ \text{Converse} \\ \text{NOT true} \end{array} \right.$

$$\Rightarrow E[xy] = E[x] E[y]$$

$$v[x+y] = v[x] + v[y]$$

### Chebychev's Inequality

if  $x$  is a RV with mean( $\mu$ ) and var( $\sigma^2$ )  
then for any  $k > 0$ :

$$P\left(\frac{|x - \mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}$$

$\hookrightarrow$  no. of std-dev<sup>n</sup> from  $\mu$

only mean & var. must be known

$\rightarrow$  weak inference

general but not always correct  
(slide 36)

### Weak law of large no.(s)

mean of a sample of RVs converges into expected value as sample size grows larger

$$P\left(\left|\frac{x_1 + x_2 + \dots + x_n}{n} - \mu\right| > \varepsilon\right) \xrightarrow{\text{as } n \rightarrow \infty} 0$$

where  $E[x_i] = \mu$  &  $\varepsilon > 0$

lot of data — no need to look for  
TRUE mean & variance.

## DISTRIBUTIONS

prob. dist<sup>m</sup> as models

$$P(x = x | \theta) = p(x | \theta)$$

$\downarrow$   
parameters

$\theta$  changes — dist<sup>m</sup> changes

— Gaussian

$$x = r$$

$$p(x | \theta = \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

mean var.

$$x \sim N(\mu, \sigma^2)$$

Std. Normal —  $N(0, 1)$   
if  $z \sim N(0, 1) \Rightarrow x = \mu + \sigma z$

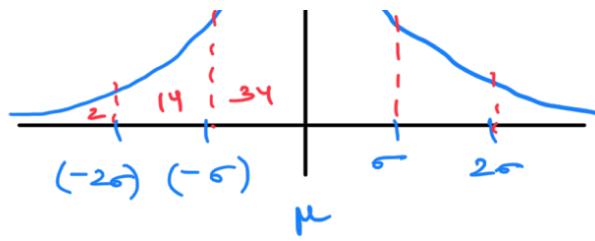
$$E[x] = \mu = \int x \cdot p(x) dx$$

$$V[x] = \sigma^2$$

mode =  $\mu$  = median

almost all dist<sup>m</sup> converge to gaussian  
as sample size increases





### - Bernoulli

$$X = \{0, 1\}$$

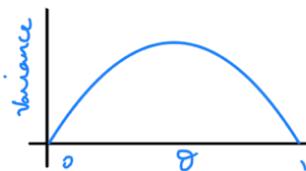
$$P(X=1 | \theta) = \theta$$

$\theta \in [0, 1]$   
Success pr.

$$P(X|\theta) = \theta^x (1-\theta)^{1-x}$$

$$E[X] = \theta$$

$$\sqrt{V[X]} = \sqrt{\theta(1-\theta)}$$



Variance is highest when  $\theta = 1/2$

### - Binomial

$n$  Bernoulli trials —  $m$  successes  
 $m \in \{0, 1, \dots, n\}$

$$P(m|n, \theta) = \binom{n}{m} \theta^m (1-\theta)^{n-m}$$

$$E[X] = n\theta$$

$$V[X] = n\theta(1-\theta)$$

conditions :  $n$  is finite  
 $\theta$  is constant  
independent events

dist $\sim$  is skewed towards  $n\theta$

### - Uniform

when outcomes are equally likely

a) discrete

$$P(X = k | a, b) = \frac{1}{b-a+1}$$

$$X \in \{a, \dots, b\} \quad b \geq a$$

$$E[X] = \frac{a+b}{2}$$

$$V[X] = \frac{(b-a+1)^2 - 1}{12}$$

b) continuous

$$p(x|a,b) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \leq x \leq b \\ 0 & x \geq b \end{cases}$$

$$E[X] = \frac{a+b}{2}$$

$$V[X] = \frac{(b-a)^2}{12}$$

- Poisson

$$x \in \mathbb{Z}_+$$

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \lambda = \text{rate}$$

$$E[X] = \lambda$$

$$V[X] = \lambda$$

conditions

- events are independent
- 2 events do not occur simultaneously
- $\lambda$  is constant

## STATISTICAL INFERENCE

while we may have idea about the shape of dist<sup>n</sup>  $\theta$  is generally unknown

so we estimate pop<sup>n</sup> parameters ( $\theta$ )

Sum of Squared Errors — SSE

$$\hat{\mu} = \arg \min \left\{ \sum_{i=1}^n (u_i - \mu)^2 \right\}$$

estimator  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$

$$SSE(\mu) = \sum_{i=1}^n (y_i - \mu)^2$$

$$\frac{dSSE(\mu)}{d\mu} = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$$

sample mean  
but no sol<sup>n</sup> for  $\sigma$

Max. Likelihood Est<sup>n</sup>—MLE

$$\hat{\theta} = \arg \max \{ p(y|\theta) \}$$

$$\textcircled{1} \quad \frac{\partial}{\partial \mu} L(y|\theta; \mu, \sigma) = 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i \text{ (same as sample mean)}$$

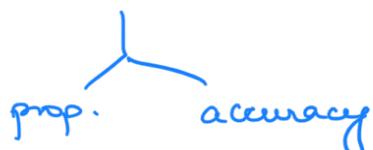
$$\textcircled{2} \quad \frac{\partial}{\partial \sigma} L(y|\theta; \mu, \sigma) = 0$$

$$\Rightarrow \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2} \quad \text{DIFF from sample var.}$$

while  $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \mu)^2}$

unbiased

$\mu_{MLE}$  vs  $\mu_{unbiased}$ ?  
→ Comparing estimators



estimation is data dependent

- ∴ we do Sampling — confidence interval  
— hypo. testing  
— compare estimators

assuming  $y_1, y_2, \dots, y_n$  follow a parametric dist<sup>n</sup>  $p(y|\theta)$

an estimator  $\hat{\theta}(y_1, \dots, y_n)$  is a funct<sup>n</sup> of  $y_1, \dots, y_n$  called  $\hat{\theta}$  an op<sup>n</sup>er

→ popul<sup>n</sup> parameters

if  $y_1, y_2, \dots, y_n$  are iid from  $f(y|\theta)$

$\Rightarrow$  the estimator follows a dist<sup>n</sup>  
determined by  $p(y|\theta)$

for ex- if  $y \sim N(\mu, \sigma^2)$

$$\text{then } \bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\downarrow$   
sample means.

Estimator Bias

$$B_\theta(\hat{\theta}) = E[\hat{\theta}(y)] - \theta$$

if bias = 0 then estimator is unbiased

Variance of estimator

$$V_\theta(\hat{\theta}) = E[(\hat{\theta}(y) - E[\hat{\theta}(y)])^2]$$

high bias low var vs high var low bias ?  
net effect - MSE

Mean Squared Error

$$\begin{aligned} \text{MSE}_\theta(\hat{\theta}) &= E[(\hat{\theta}(y) - \theta)^2] \\ &= b_\theta(\hat{\theta})^2 + V_\theta(\hat{\theta}) \end{aligned}$$

an estimator is consistent if

$$\begin{cases} b_\theta(\hat{\theta}) \rightarrow 0 \\ V_{\theta,\hat{\theta}}(\hat{\theta}) \rightarrow 0 \end{cases} \quad \left. \right\} \text{as } n \rightarrow \infty \text{ & } \theta$$

[if  $y_1, \dots, y_n$  has  $E[y_i] = \mu$  &  $V[y_i] = \sigma^2$

$$b_\mu(\hat{\mu}_{\bar{y}}) = 0 \quad V_{\mu,\hat{\mu}_{\bar{y}}}(\hat{\mu}_{\bar{y}}) = \frac{\sigma^2}{n}$$

and thus  $\text{MSE} = \sigma^2/n$

## CENTRAL LIMIT THEOREM

Central Limit Theorem

$$\sum_{i=1}^n Y_i \xrightarrow{d} N(n\mu, n\sigma^2)$$

and  $\bar{Y} \xrightarrow{d} N(\mu, \sigma^2/n)$  as  $n \rightarrow \infty$

a lot of distributions converge into normal

$$\text{bin}(\theta, n) \approx N(n\theta, n\theta(1-\theta))$$

$$\text{Pois}(\lambda) \approx N(\lambda, \lambda)$$

any Poi( $\lambda$ ) RV is sum of  $\lambda$  Poi(1) RV

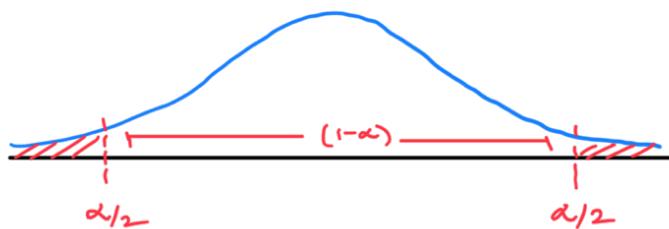
Confidence Interval

as point estimation is not highly accurate due to randomness we use interval estm.

$$\hat{\theta}_{\text{int}}(y) \rightarrow (\hat{\theta}_{\text{point}}^-(y), \hat{\theta}_{\text{interval}}^+(y))$$

i.e. 2 sided CI:  $[Q(\alpha/2), Q(1-\alpha/2)]$

we are  $(1-\alpha)$  confident  $Z \sim (0, 1)$  falls inside  $(-Z_{1-\alpha/2}, Z_{1-\alpha/2})$



$$\Rightarrow [-Z_{1-\alpha/2}, Z_{1-\alpha/2}]$$

CI for Normal Means

① unknown  $\mu$ , known  $\sigma^2$

—      " —

$$\hat{\mu}_{ML} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (\text{sample mean})$$

$$\left( \hat{\mu}_{ML} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \hat{\mu}_{ML} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

(2) unknown  $\mu$ , unknown  $\sigma^2$

we can not assume sample mean to be equivalent to population mean

Sample variance.

$$\left\{ \begin{array}{l} \hat{\sigma}_\mu^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{Y})^2 \\ \text{but is no longer Normally dist.} \end{array} \right.$$

so we use Student-t dist<sup>n</sup>

$$\left( \hat{\mu}_{ML} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \hat{\mu}_{ML} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right)$$

## HYPOTHESIS TESTING

- propose a null hypothesis ( $H_0$ ) and an alternate one ( $H_a$ )
- try to reject  $H_0$  using p-value
- low p-value : evidence against Null

Goal — Reject Null.

$$p = \begin{cases} 2P(-|z|) & H_0 \text{ is } = \\ 1 - P(z) & \leq \\ P(z) & \geq \end{cases}$$

just like in CI,  $\sigma$  is either

① known

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

calculate p-value under the curve :

$p < 0.01$  STRONG

$0.01 < p < 0.05$  MODERATE

$p > 0.05$  WEAK evidence against NULL

② unknown

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

unlike above, we don't care about p-value here

if  $|t_{\hat{\mu}}| >$  required t-score  
then Reject  $H_0$

else we can't reject  $H_0$

## REGRESSION & CORRELATION

### Linear Regression

$$y = \beta_0 + \beta_1 x$$

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = n(\bar{x^2} - \bar{x}^2)$$

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = n(\bar{xy} - \bar{x}\bar{y})$$

$$SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = n(\bar{y^2} - \bar{y}^2)$$

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = \frac{\bar{y}\bar{x^2} - \bar{x}\bar{y}\bar{x}}{\bar{x^2} - \bar{x}^2}$$

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \frac{SS_{yy} SS_{xx} - SS_{xy}^2}{SS_{xx}}$$
$$= \dots - \dots \hat{\sigma}^2$$

$$R^2 = 1 - \frac{RSS}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} = r_{xy}^2$$

$$s.e. (\beta_0) = \sqrt{\frac{RSS}{n(n-2)} \frac{\bar{x}^2}{\bar{x}^2 - \bar{x}^2}}$$

$$s.e. (\beta_1) = \sqrt{\frac{RSS}{n(n-2)} \frac{1}{\bar{x}^2 - \bar{x}^2}}$$

Multi linear Reg.

$$E[y_i | x_{i1}, \dots, x_{ip}] = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

## CLASSIFICATION

Naive Bayes.

$$P(y|x_1, \dots, x_p) = \frac{P(y) \prod_{j=1}^p P(x_j|y)}{P(x_1, \dots, x_p)}$$

Logistic Regression

$$\alpha_i = w_0 + w_1 x_{i1} + \dots + w_p x_{ip}$$

$$\alpha_i = w_0 + \sum_{j=1}^p w_j x_{ij}$$

$$\hat{y}_i = \frac{1}{1 + e^{-\alpha_i}} \quad \text{bound to } [0, 1]$$

$$\text{cost } (\omega) = -\frac{1}{n} \sum_{i=1}^n y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i)$$

-ve log likelihood / log-loss

$$d\omega_j = \sum_{k=1}^p (\hat{y}_i - y_i) x_{ik}$$

$$\omega_j = \omega_j - d\omega_j \times \eta$$

learning rate

$\omega_0$  — bias parameter

if  $\omega_0 < 0$  : class 0  
 $\omega_0 > 0$  : class 1

$$\Rightarrow p(Y_i=1 | x_1, \dots, x_p) = \frac{1}{1 + e^{-\gamma_i}}$$

$$\gamma_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$$

Decision Trees

$$H(x) = E \left[ \log_2 \frac{1}{p(x)} \right]$$

$$H(x|y) = \sum_y p(y) H(x|y=y)$$

$$\text{Information Gain} = H(x) - H(x|y)$$

initial after split

maximise IG by minimising  $H(x|y)$

if  $x, y$  are independent :

$$H(x,y) = H(x) + H(y)$$