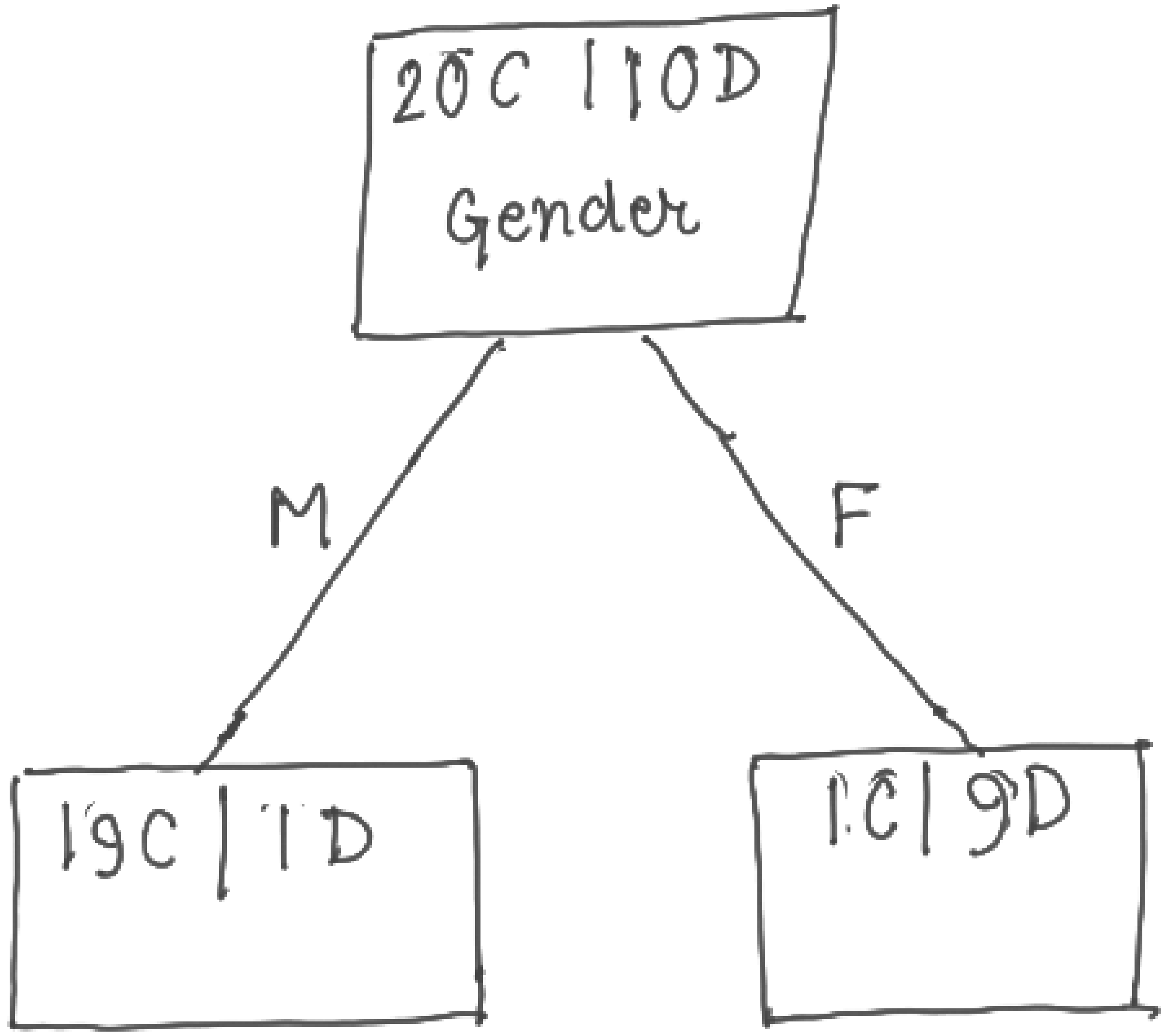
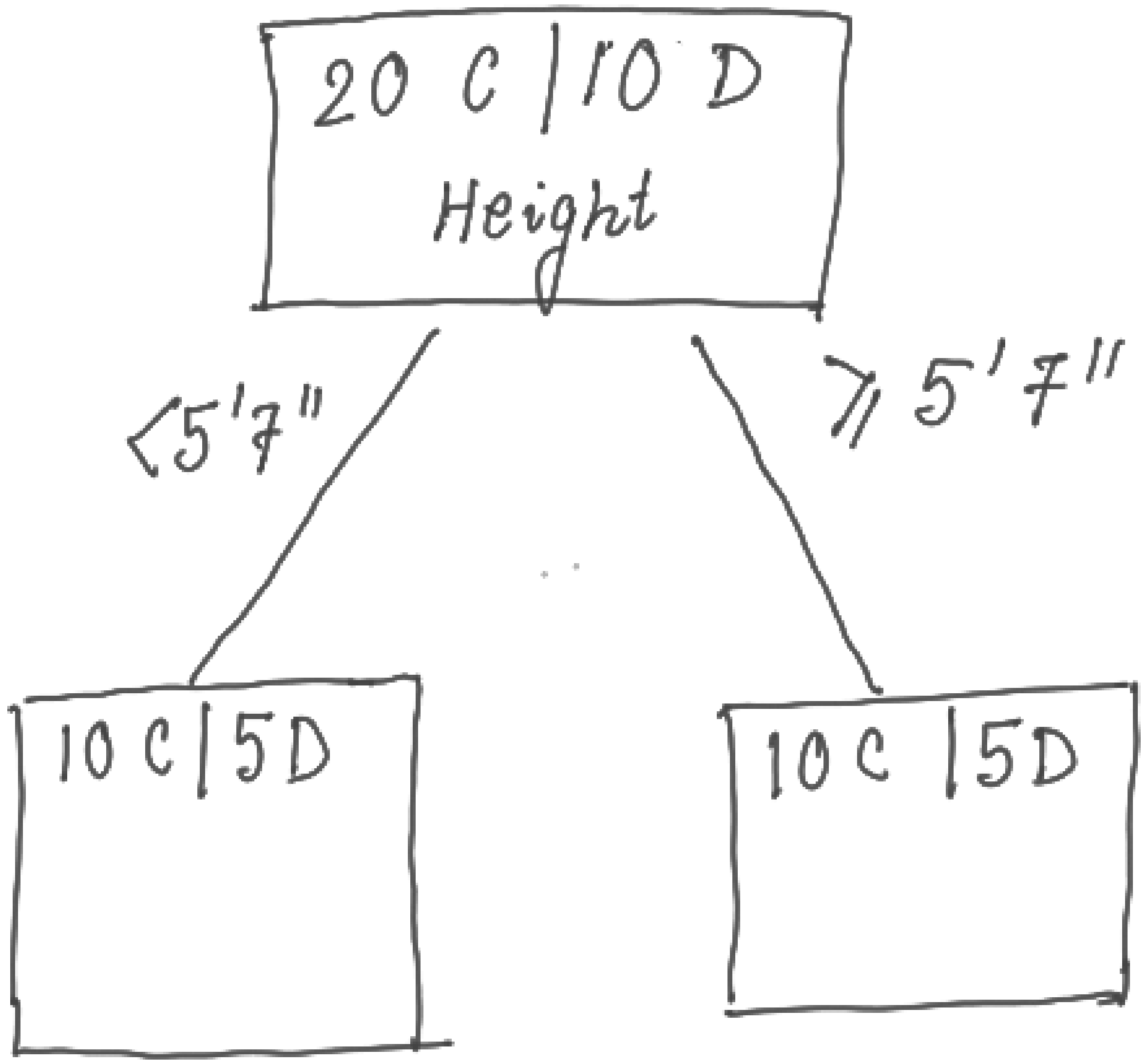


lesser entropy

Option 1

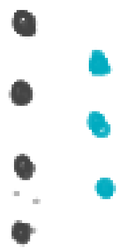


More entropy

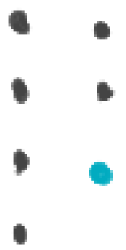
Option 2

Decision Trees are concerned with increase overall homogenity.

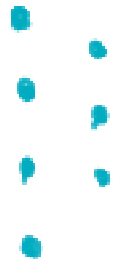
Entropy $\rightarrow$ way to measure non-uniformness of collection of items of various types

1.  $\rightarrow$ is more mixed ; hence more entropy $\quad p_B = \frac{4}{7} \quad p_B = \frac{3}{7}$

$$En = -\left[\underbrace{\frac{4}{7}\log\frac{4}{7}}_{0.1388} + \underbrace{\frac{3}{7}\log\frac{3}{7}}_{0.1577}\right] \qquad En = -\sum p_i \log p_i \qquad En = 0.2965$$

2.  $\rightarrow$ is less mixed ; hence lesser entropy $\quad p_B = \frac{6}{7} \quad p_B = \frac{1}{7}$

$$En = -\left[\frac{6}{7}\log\frac{6}{7} + \frac{1}{7}\log\frac{1}{7}\right] = 0.1781$$

3.  $\rightarrow$ is not mixed ; minimum possible entropy $\quad p_B = 1, p_B = 0$

$$En = -\left[\underbrace{1\log 1} + 0 \log 0\right] = 0$$

$$\underbrace{\overset{X}{x_1 \; x_2 \; \cdots \; x_n} \qquad y}_{\text{classification}}$$

- discrete features
- discrete response

ID3 algorithm

$30 \begin{cases} 15 \\ 15 \end{cases} \quad \begin{matrix} 1/2 \\ 1/2 \end{matrix}$

1. Entropy of the dataset $-\left[\frac{1}{2}\log\frac{1}{2}+\frac{1}{2}\log\frac{1}{2}\right]=1$

2. Identify the feature that leads to a maximum $\underbrace{\text{decrease in entropy}}_{\text{information gain.}}$

3. Using the most useful feature as parent node; we find another feature that leads to a max. decrease in entropy

4. Do (3) until all features are exhausted or you hit a stopping criteria.

Original DS : 15 C , 15 D

$En = 1$

Gender : $\underbrace{8D, 2C}$ | $\underbrace{13C, 7D}$

$En_g = 10 \times E_1 + 20 \times E_2 \,/\, (10+20)$

$$\underbrace{-\left(\frac{8}{10}\log\frac{8}{10} + \frac{2}{10}\log\frac{2}{10}\right)}_{E_1} \qquad \underbrace{-\left(\frac{13}{20}\log\frac{13}{20} + \frac{7}{20}\log\frac{7}{20}\right)}_{E_2}$$

$IG = En - En_{g/h/c}$

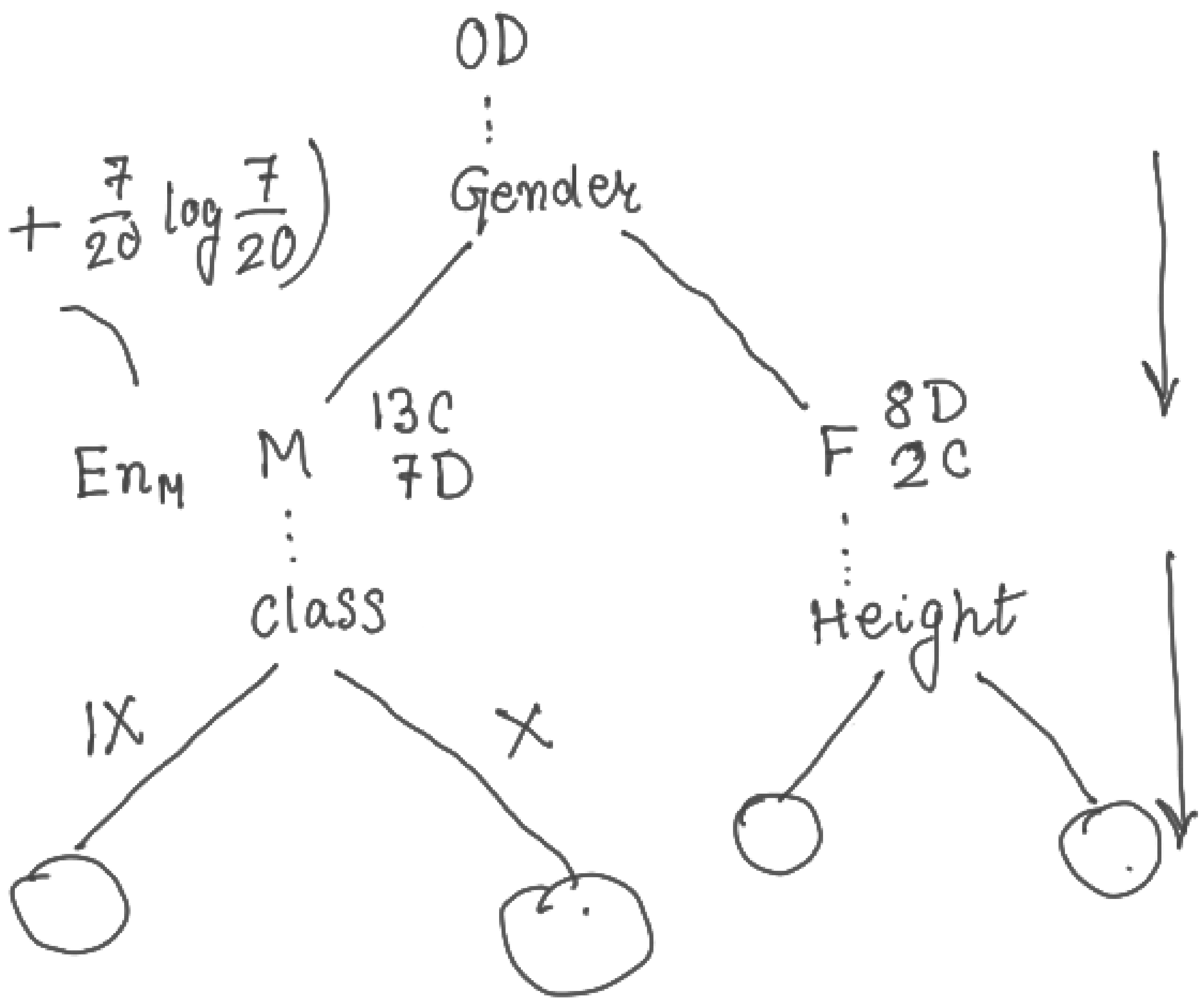
Height : $\underbrace{5C, 7D}_{E_3}$ | $\underbrace{10C, 8}_{E_4}$

$$En_h = \frac{12 \times E_3 + 18 \times E_4}{(12+18)}$$

Class :

En_c

$$-\left(\frac{13}{20}\log\frac{13}{20} + \frac{7}{20}\log\frac{7}{20}\right)$$



H	sort(H)	Splits
165	163	164
172	165	165.5
179	166	166.5
181	167	168
171	169	⋮
⋮	⋮	

$H < 164 \rightarrow En_{H<164}$

$H < 165.5 \rightarrow En_{H<165.5}$

$H < 166.5$

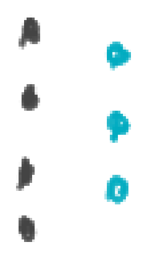
⋮

$avg(En_H)$

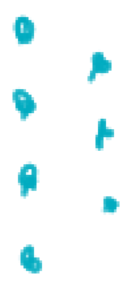
CART

- it uses Gini index to measure impurity / homogenity.

$$G = 1 - \sum p_i^2$$



$$G = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = \frac{24}{49} \sim 0.5$$



$$G = 1 - (1)^2 = 0$$

Algorithm uses GI instead of entropy; rest similar to ID3.

$x_1 \; x_2 \, .. \; x_p \quad y \to$ cont

y_1

y_2

$\vdots$

y_n

DT $\to$ minimizes MSE

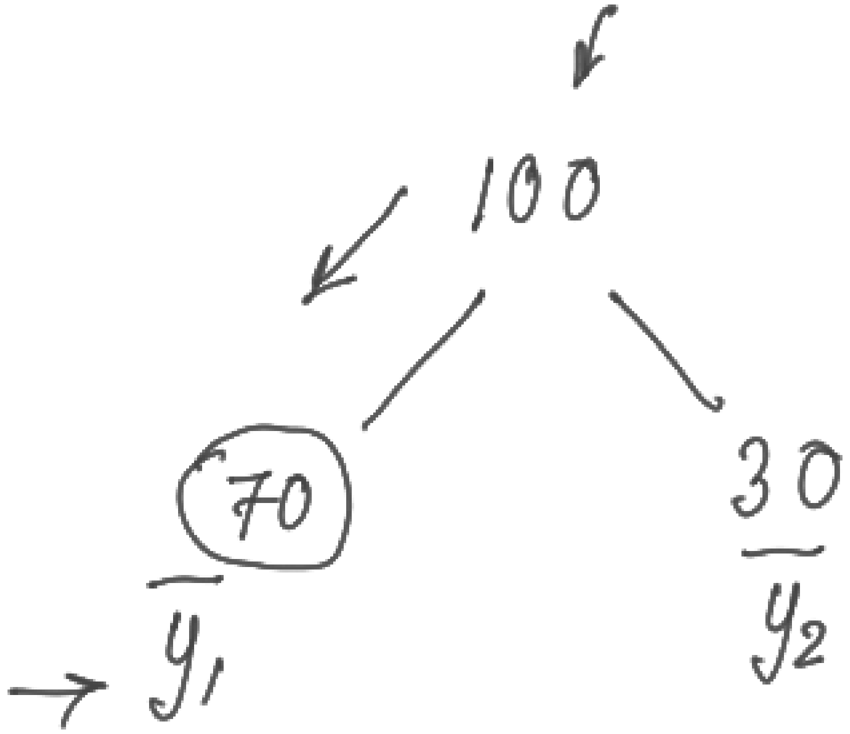
OD : $MSE = \underline{\underline{\frac{1}{n}}} \Sigma (\bar{y} - y_i)^2$

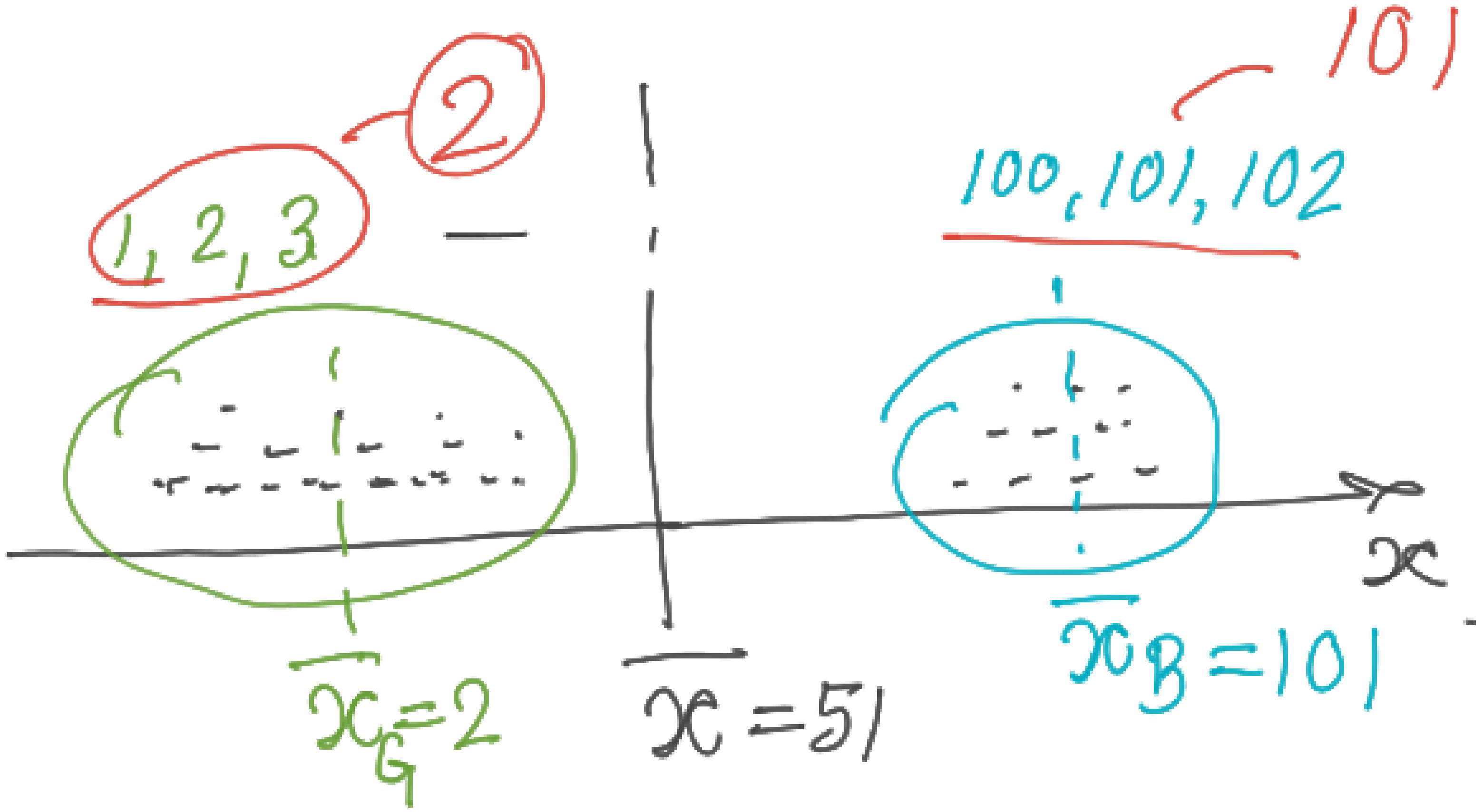
$$\begin{array}{ccc} & n_1 & n_2 \\ x_1 : & (D_1 , & D_2) \\ & \bar{y}_1 & \bar{y}_2 \end{array}$$

$$MSE_{x_1} : \underline{\underline{\frac{1}{n_1} \Sigma (y_i - \bar{y}_1)^2}} + \underline{\underline{\frac{1}{n_2} \Sigma (y_i - \bar{y}_2)^2}}$$

$n_1 + n_2 = n.$

MSE_{x_2} :





1. $\sum (x_i - \bar{x})^2 = (1-51)^2 + (2-51)^2 + (3-51)^2 + (100-51)^2 + (101-51)^2 + (102-51)^2$

$\approx 15000 = \text{MSE}$

$$\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{(100-101)^2 + (101-101)^2 + (102-101)^2} = 2. \begin{cases} \sum (x_{gi} - \bar{x}_G)^2 \\ \sum (x_{bi} - \bar{x}_B)^2 \end{cases} +$$

$= 4$ new MSE

14496 ΔMSE