Z-Score :- Assumes that Feature is normally distributed

Z-Score :- Assumes that Feature is normally distributed

100

34.1   34.1

2.1%
0.15%   2.1%   13.6   13.6   0.15%

-3σ -2σ -1σ   μ   1σ   2σ 3σ

68.2%

95.4%

99.7%

obs

Age = 100

mean = 36

68

(25 , 35)

132

(35, 40)

60

60 - 30

= 30

= 3b 36

= 1

20 25 35 40 ni

13 30 15

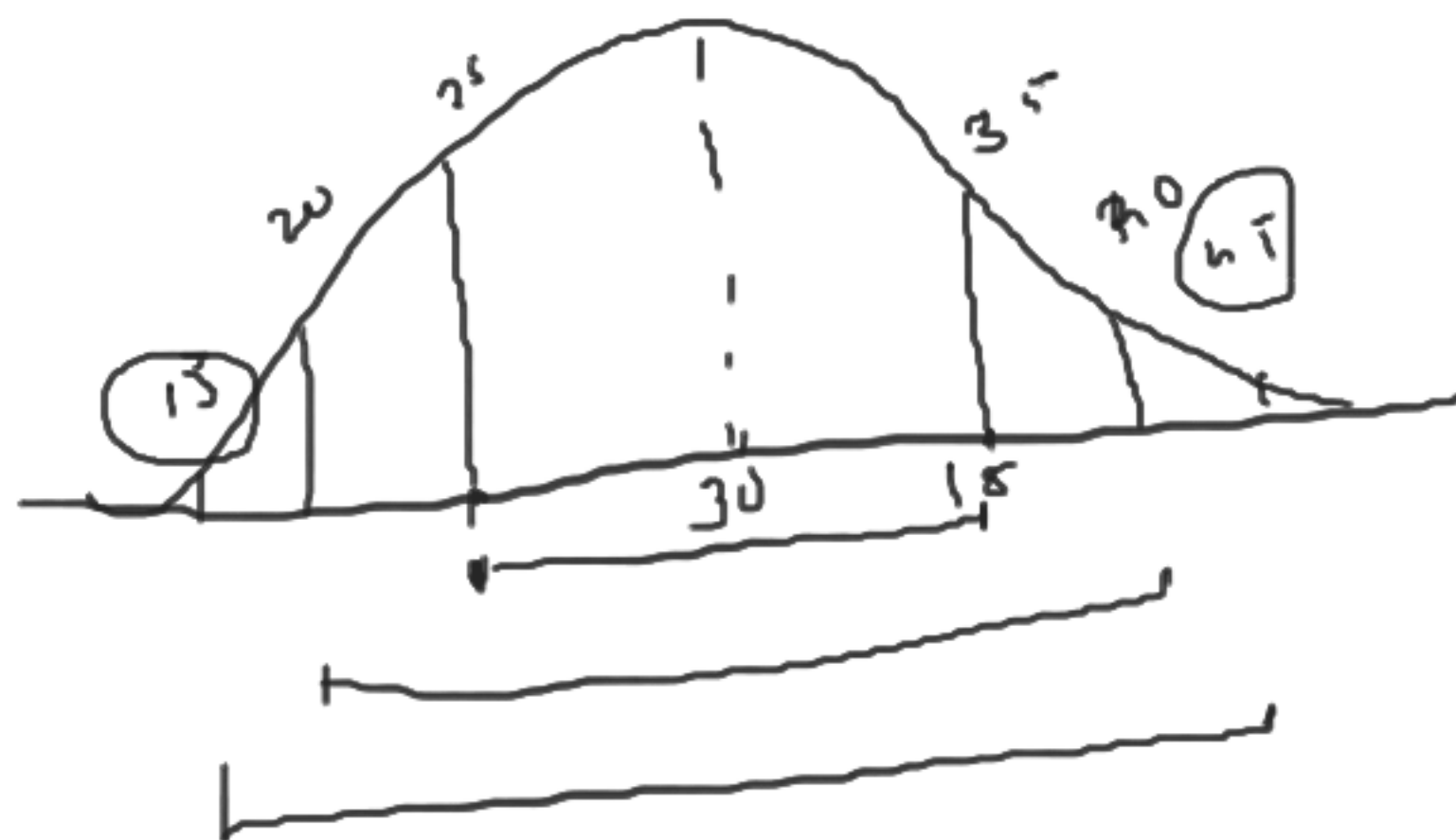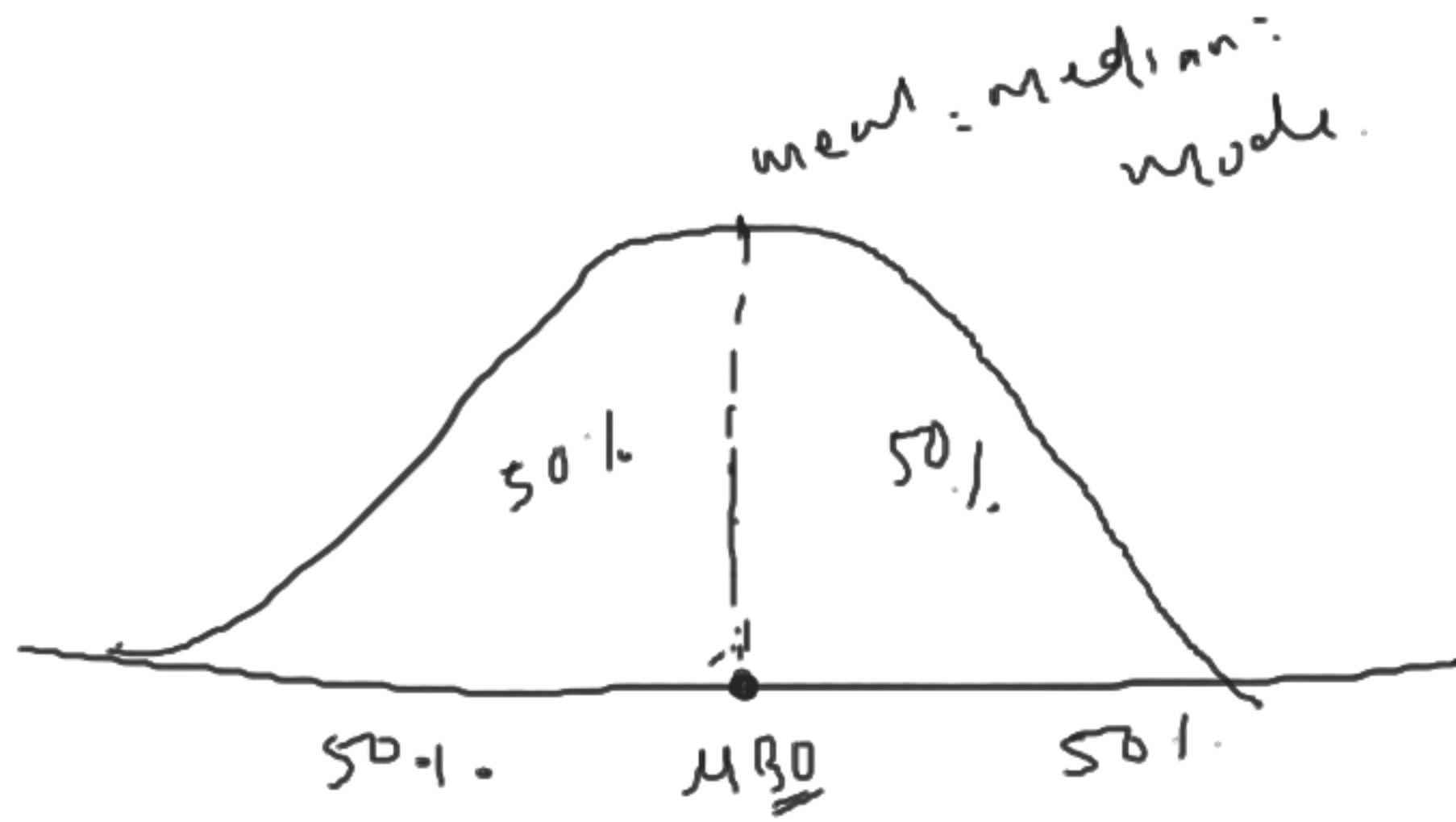Most of the values lies near contre / mean

#1   Age   2) heights   3) marks   60-80% 80-90

                                                              90-100

4) errors in measurement   5) blood pressure

mean = median = mode

50%   50%

50%   μ 30   50%

Std = How far away a datapoint is from mean value

industry ⟹ 3 std ⟹ outlier

2 std

4 std } slight
5 std

$\underline{Z\text{-score}} \Rightarrow$ how many std away a datapoint is

$\underline{Z\text{-score}}$

$$77 = 66.37 + 3 \times 3.7$$

mark

$$\underline{66.37} = 11.5$$

$$\boxed{77.0}$$

ex

$\underline{\text{std}}$

63.

68

74

$\begin{array}{l} 0.1 \\ 0.5 \\ 1.2 \\ 7.0h \\ 3.7 \\ 4.6 \\ 5.1 \end{array}$

78

$\underline{66.37}$

$\underline{3.7}$

58

72

$$= 66.37 + \mathbf{2} \times 3.7$$

65

69

$$= 73.77$$

66

$\boxed{73.7} \rightarrow \boxed{Z}$

$$66.37 + 2 \cdot 8 \cdot 7$$

$$68.2 \, \text{'/.} = \begin{array}{l} \mu + 1\sigma \rightarrow 34.1 \\ \mu - 1\sigma \rightarrow 34.1 \end{array}$$

$$95 \cdot 4 \, \text{'/.} = \begin{array}{l} \mu + 2\sigma \longrightarrow 13.6 \\ \mu - 2\sigma \longrightarrow 17.6 \end{array}$$

$$99 \cdot 7 \, \text{'/.} = \begin{array}{l} \mu + 3\sigma \\ \mu - 3\sigma \end{array}$$

$$x \qquad \mu \qquad \sigma$$

$$80 = 66.37 + \underline{\underline{z}} \times 3.7$$

$$\frac{80 - 66.37}{3.7} = \underline{\underline{z}} = \frac{x - \mu}{\sigma}$$

$$Z_{score} = 3.68$$

Outlier treatment

threshold = 3

$\mu + 3\sigma =$
$\mu - 3\sigma =$

drop
Trimming

Impute
capping

3.7
$66.37 + 3\sigma = \boxed{77.4}$
$66.37 - 3\sigma = 55.2$
3.87

$100 =$
5 = outliers
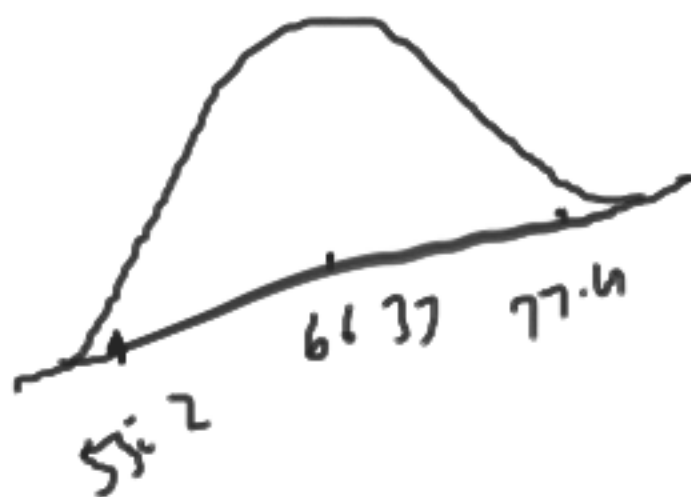
$100 = s$
$5 = $ out.
$100 = $ Remaining

66.37 77.4

55.2

80   83   48   49   52

77.4 77.4 55.2 55.2 55.2

95

$IQR =$

IQR

Q_1 Q_2 Q_3

Min

$1, 5$

min

$Q_1 - 1.5 \times IQR$

median

25%

75%

max

whiskers

$Q_3 + 1.5 \times IQR$

$= max$

outliers

Q_1 Q_3

0   25%   50%   75%   100%
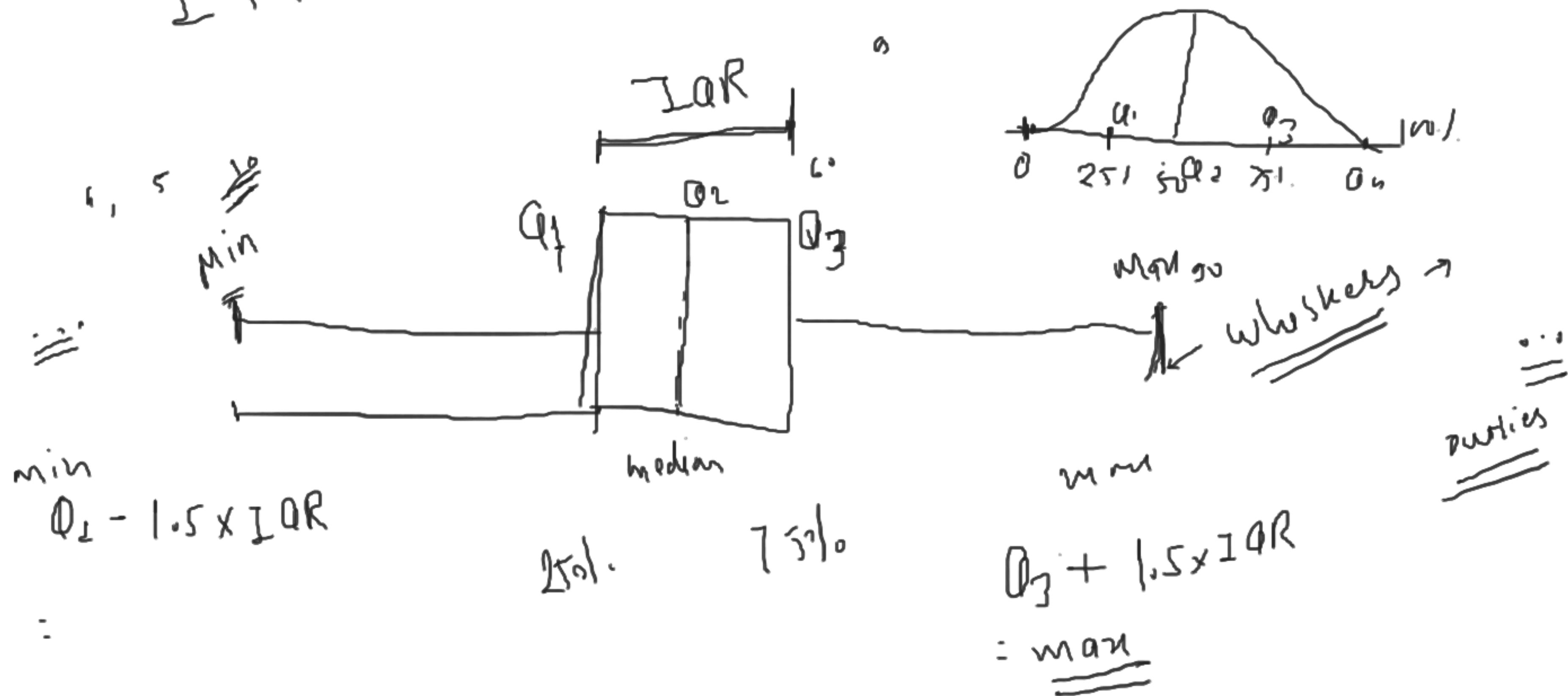
median = (Q2 | 50% percentile) :

the middle value of dataset