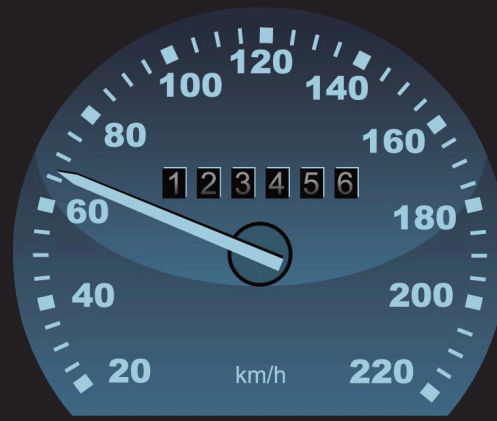


Statistics



1. “You are driving at 65 km/h”
2. “You will (most likely) reach destination in 30 mins”

Descriptive statistics

Summarise data

Central tendency, variability

Inferential statistics

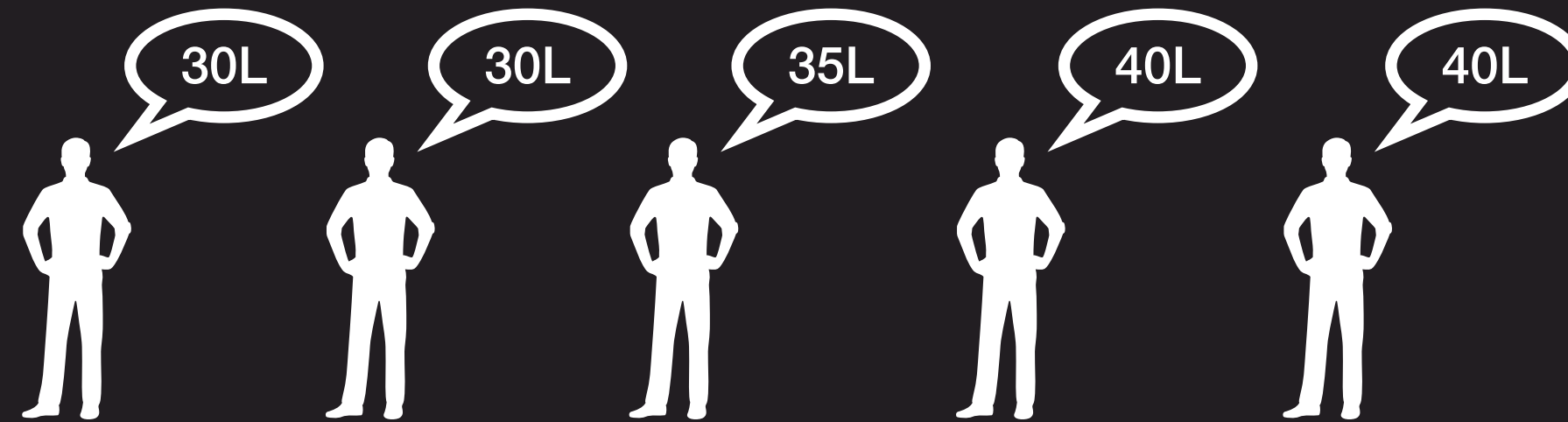
Drawing conclusions from observations

Confidence interval, hypothesis test, regression

1. “Vote share of candidate A was 70%”
2. “Our exit poll says candidate A will have 70% vote share”

Glassdoor/levels.fyi

Salary for Data Scientist at Google

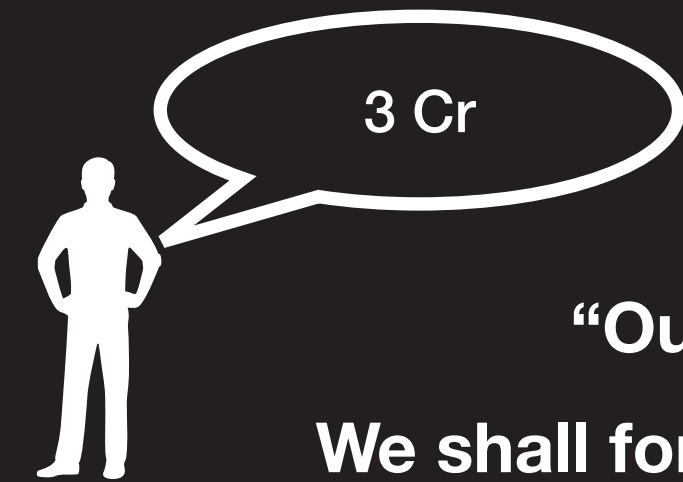


$$\text{Mean} = \frac{(30 + 30 + 35 + 40 + 40)}{5} = 35L$$

Another word for mean is “Average”

$$\text{New Mean} = \frac{30 + 30 + 35 + 40 + 40 + 300}{6} = 79L$$

Crucial observation: Median is more robust to outliers



We shall formalise this

$$\text{Median} = 35L$$

Central value (if unique)

$$N = 5, \text{ odd}$$

$$\text{New Median} = 37.5L$$

Average of 2 central values

$$\frac{35 + 40}{2}$$

$$N = 6, \text{ even}$$

Median

10, 20, 30, 40, 50, 60, 70

Middle number: 40; Median = 40

10, 20, 30, 40, 50, 60, 70, 80

Two middle numbers: 40, 50; Median = $(40 + 50)/2 = 45$

Quiz There are 4 people whose average age is 24.

We know the age of three people: 20, 22, and 28.

What is the median age of these 4 people?

$$\frac{20 + 22 + 28 + x}{4} = 24$$

$$x = 4 * 24 - (20 + 22 + 28)$$

$$x = 26$$

20, 22, 26, 28

$$\text{Median} = \frac{22 + 26}{2} = 24$$

Mode

90, 90, 90, 80, 90, 70, 95, 90

Mode = 90

Mode is the most frequently occurring number, if such a number exists

2, 2, 3, 3, 4

We call this bi-modal with 2 and 3 as the modes

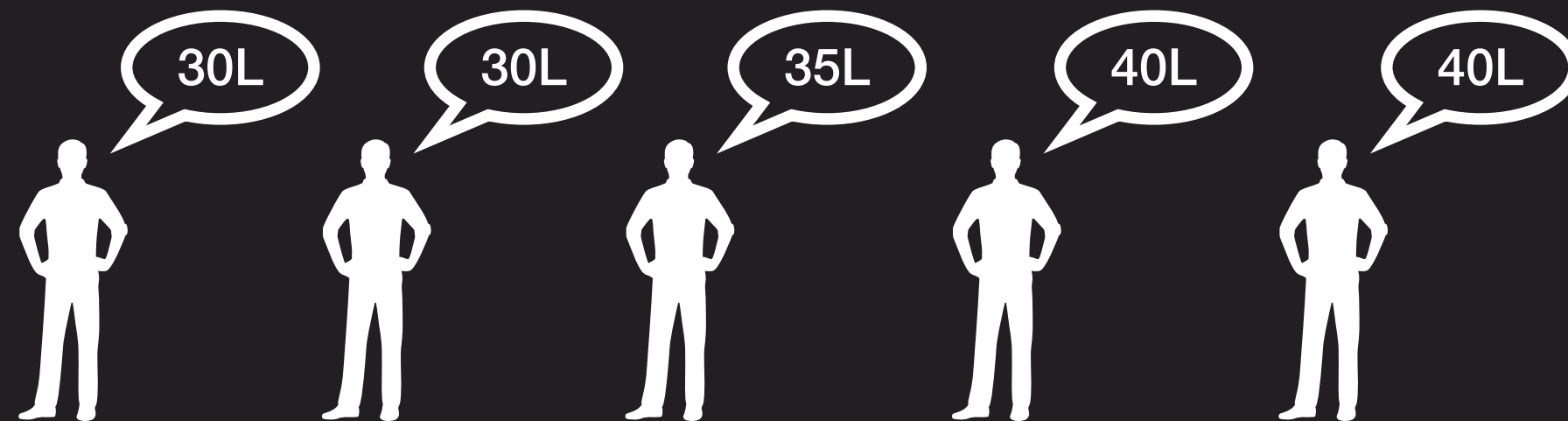
Range

Suppose a cricketer has scored as follows

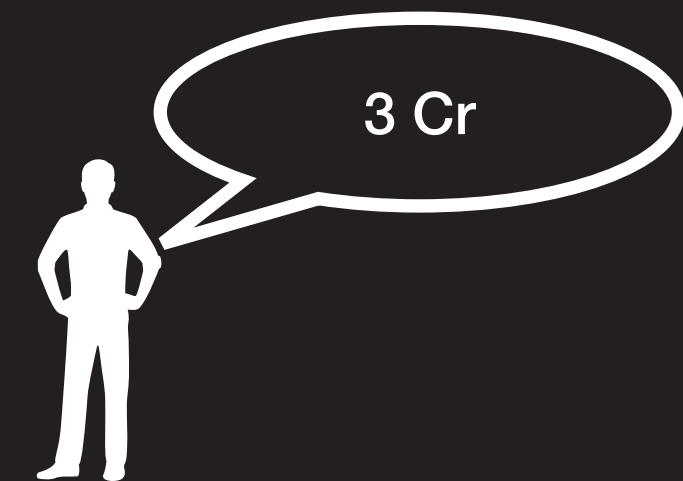
20, 25, 60, 100

We say the range = $100 - 20 = 80$

Consider again the example of salaries



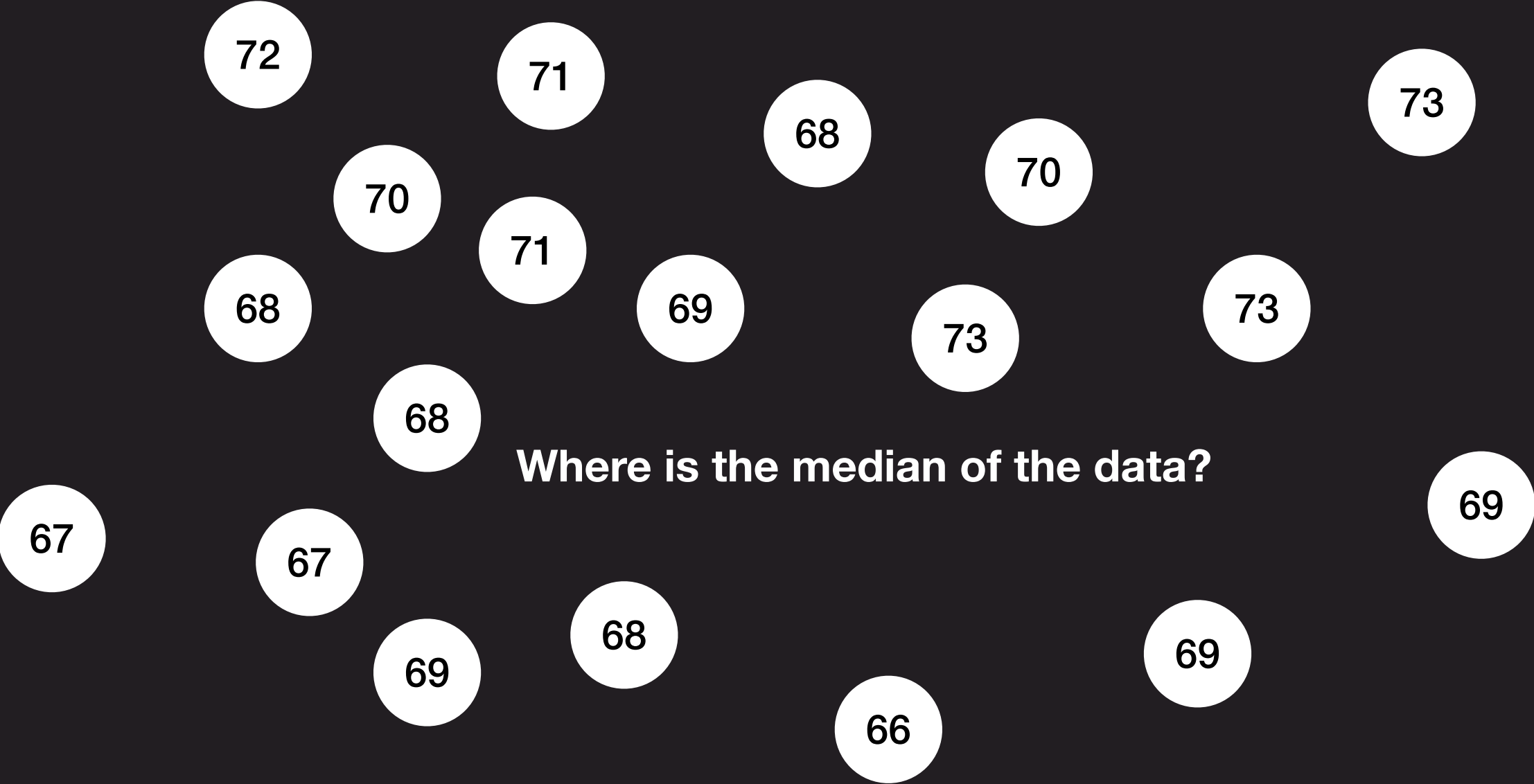
Here range = $300 - 30 = 270$ L



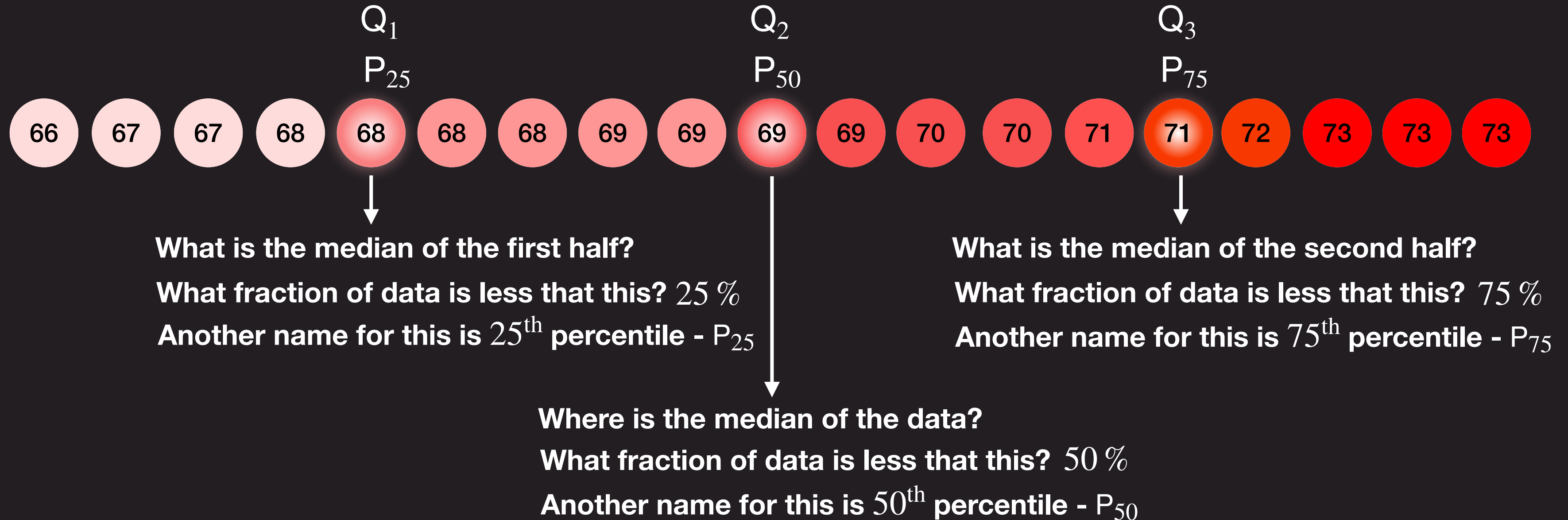
Sometimes, simply giving range may not make sense

Percentiles, Quartiles, Inter Quartile range (IQR)

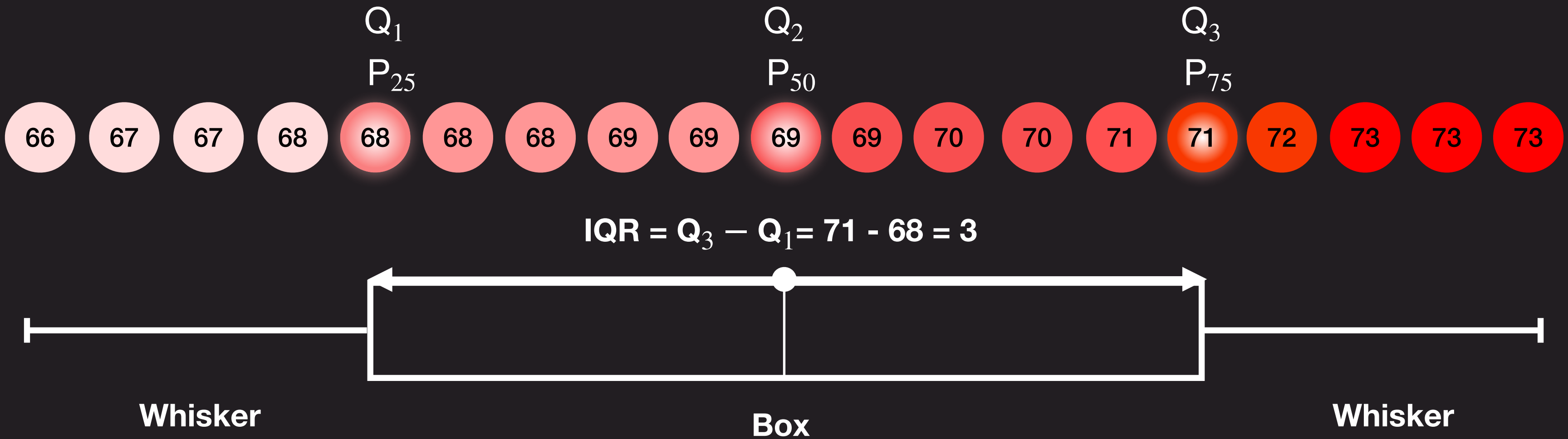
Sort the data!



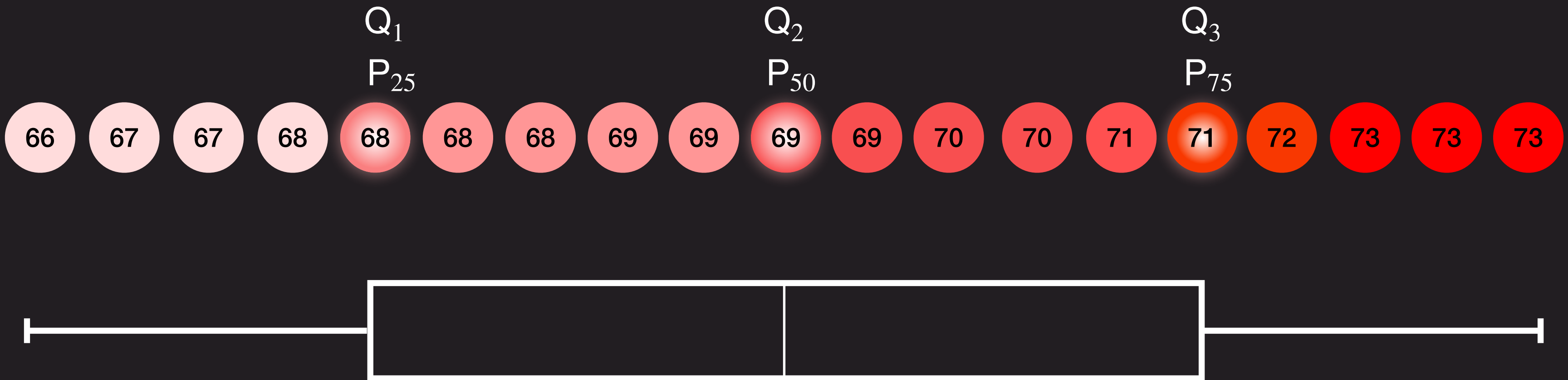
Percentiles, Quartiles, Inter Quartile range (IQR)



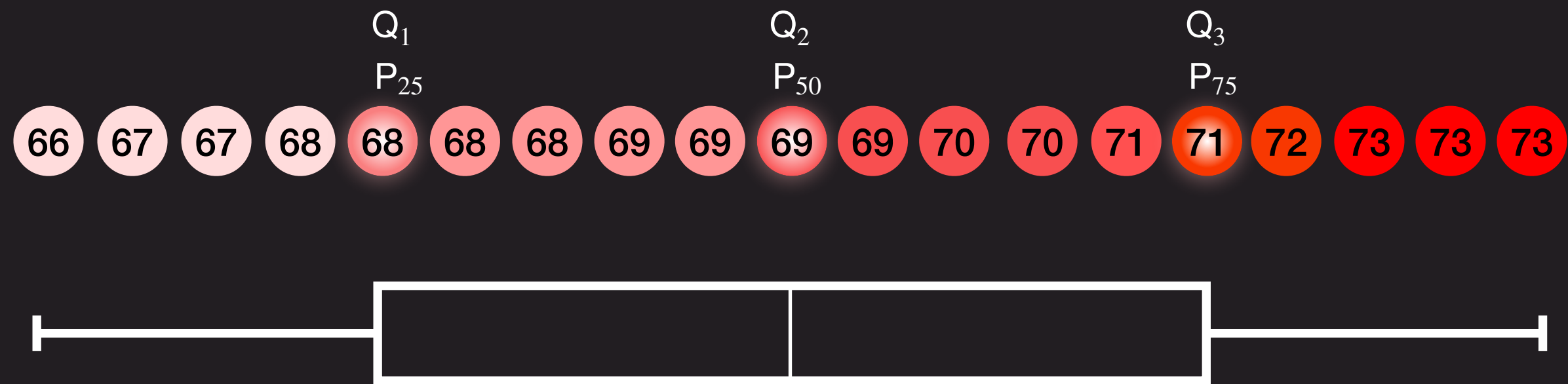
Percentiles, Quartiles, Inter Quartile range (IQR)



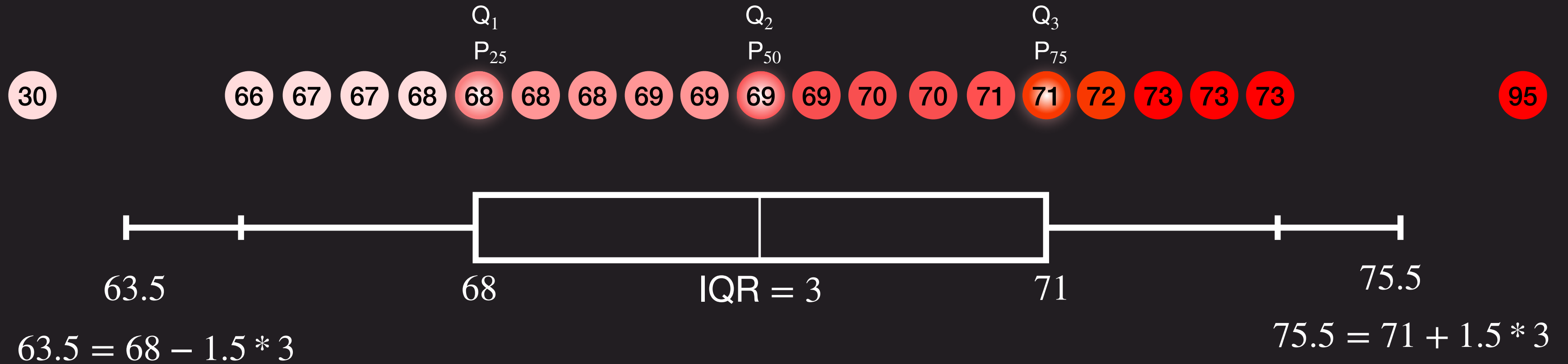
Percentiles, Quartiles, Inter Quartile range (IQR)



Percentiles, Quartiles, Inter Quartile range (IQR)

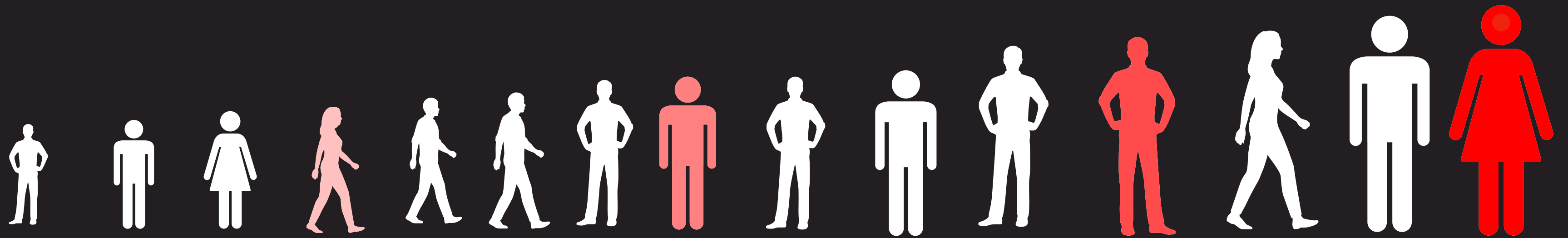


Percentiles, Quartiles, Inter Quartile range (IQR)



Outlier are points outside [63.5, 75.5]

Outlier are points outside $[Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR]$



0.25

What fraction of
people are shorter
than this lady?

0.5

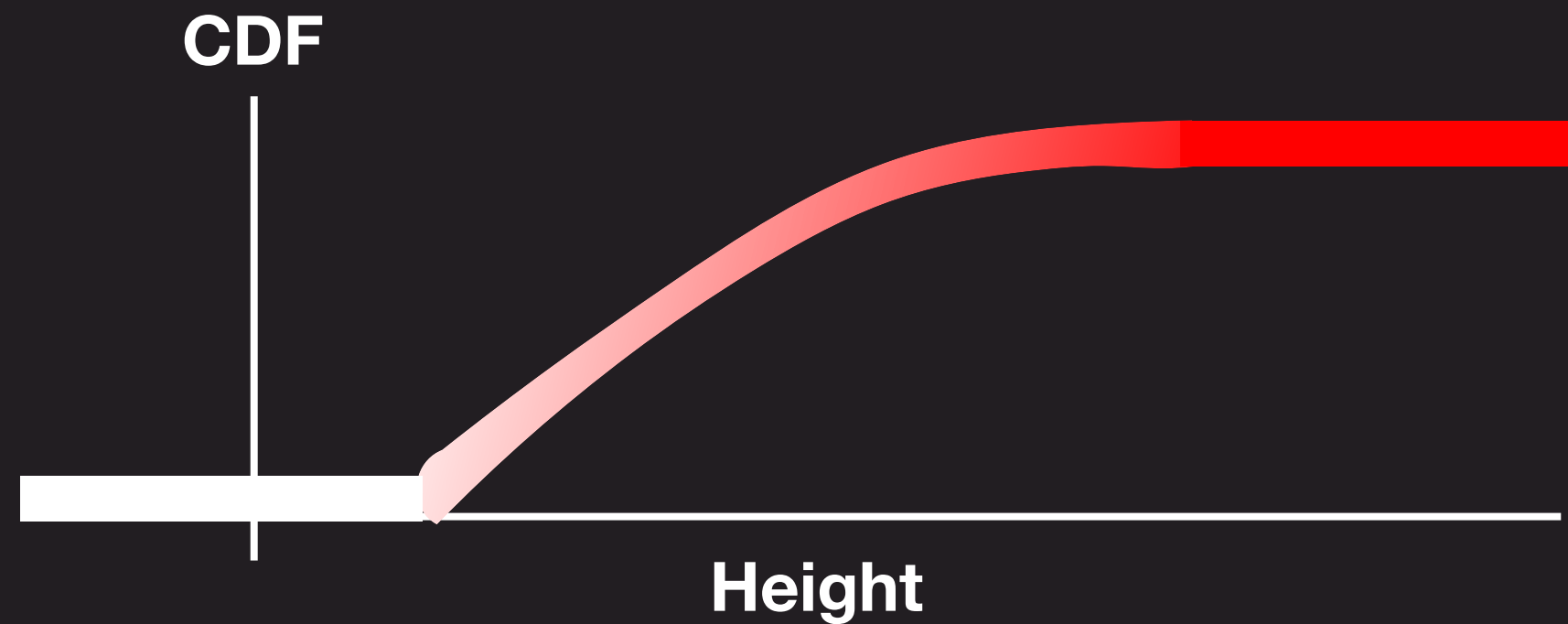
What fraction of
people are shorter
than this guy?

0.75

What fraction of
people are shorter
than this guy?

1

What fraction of
people are shorter
than this lady?



Simple Arithmetic

Original salary 30, 32, 35, 35, 38

Mean = 34

Median = 35

Mode = 35

Range = $38 - 30 = 8$

IQR = $36.5 - 31 = 5.5$

Effect of addition

After 5 L bonus

35, 37, 40, 40, 43

Mean = 39

Median = 40

Mode = 40

Range = $43 - 35 = 8$

IQR = $41.5 - 36 = 5.5$

Effect of multiplication

Salary in Yen: 1 Rs = 1.76 Yen

52.8 , 56.32, 61.6 , 61.6 , 66.88

Mean = 59.8

Median = 61.6

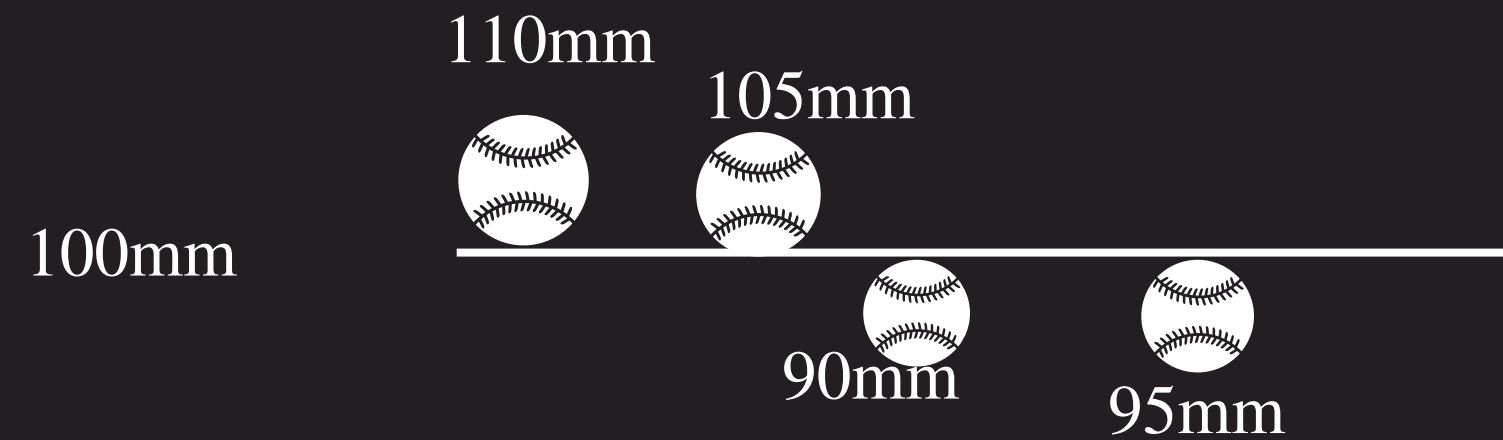
Mode = 61.6

Range = $66.88 - 52.8 = 14.08$

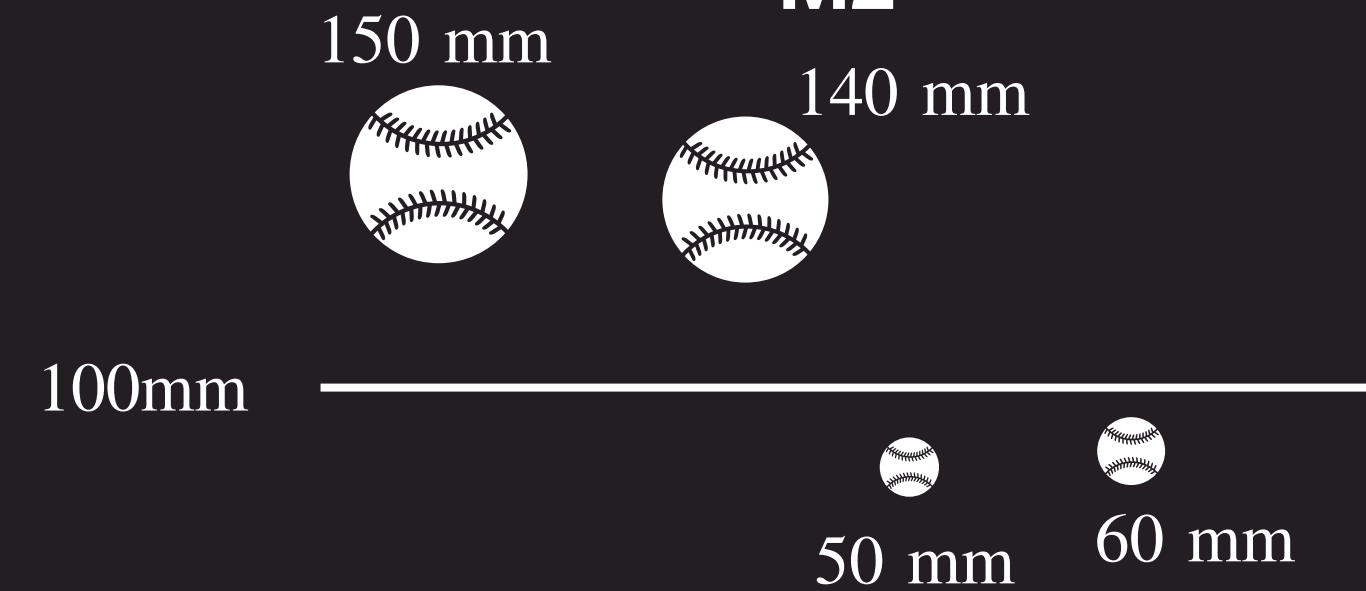
IQR = $64.24 - 54.56 = 9.68$

Variance

M1



M2



How to define Error?

$$10 \text{ mm} + 5 \text{ mm} + (-5 \text{ mm}) + (-10 \text{ mm}) = 0 \text{ mm} \quad \times$$

$$(10 \text{ mm})^2 + (5 \text{ mm})^2 + (-10 \text{ mm})^2 + (-5 \text{ mm})^2 = 250 \text{ mm}^2 \quad \checkmark$$

$$\text{Variance} = \frac{250}{4} \text{ mm}^2$$

$$\text{Std dev} = \sqrt{\frac{250}{4}} \text{ mm}$$

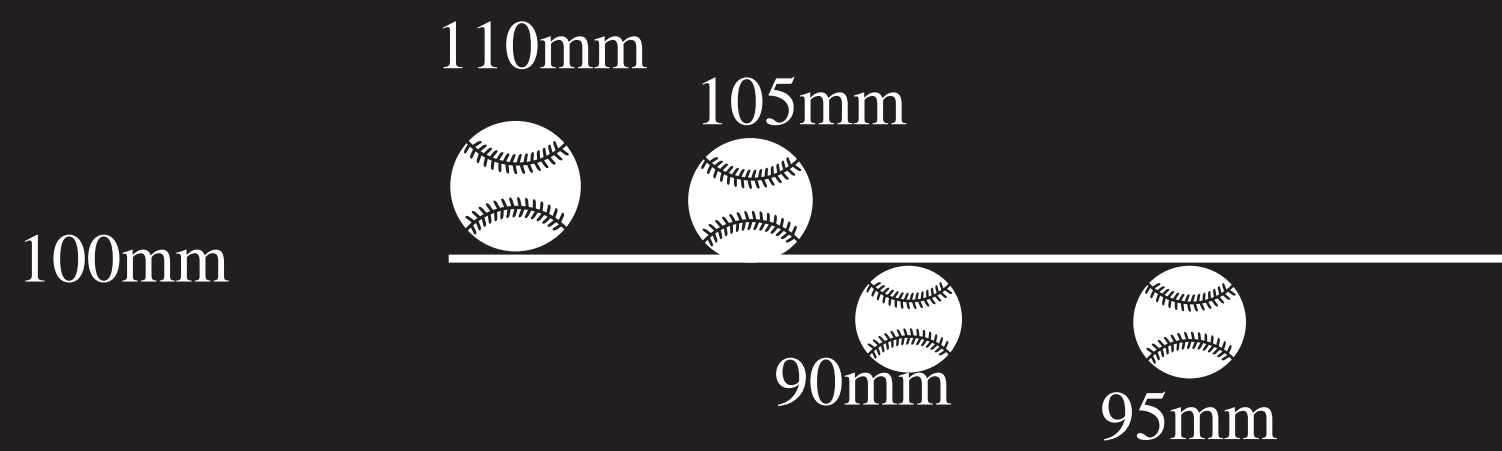
$$(50 \text{ mm})^2 + (40 \text{ mm})^2 + (-50 \text{ mm})^2 + (-40 \text{ mm})^2 = 8200$$

$$\text{Variance} = \frac{8200}{4} \text{ mm}^2$$

$$\text{Std dev} = \sqrt{\frac{8200}{4}} \text{ mm}$$

Variance

M1



x_1	110
x_2	105
x_3	95
x_4	90
\bar{x}	100

$$10 \text{ mm} + 5 \text{ mm} + (-5 \text{ mm}) + (-10 \text{ mm}) = 0 \text{ mm}$$

$$(10 \text{ mm})^2 + (5 \text{ mm})^2 + (-10 \text{ mm})^2 + (-5 \text{ mm})^2 = 250 \text{ mm}^2$$

$$\text{Variance} = \frac{250}{4} \text{mm}^2$$

$$\text{Std dev} = \sqrt{\frac{250}{4}} \text{mm}$$

$$\text{Variance} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2}{4}$$

$$\text{Std Dev} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2}{4}}$$

$$\text{Std Dev} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}} = \sigma$$

$$\text{Variance} = \frac{\sum_i (x_i - \bar{x})^2}{n} = \sigma^2$$