Descriptive Stat

Statistics

Descriptive statistics

Summarise data

Central tendency, variability

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

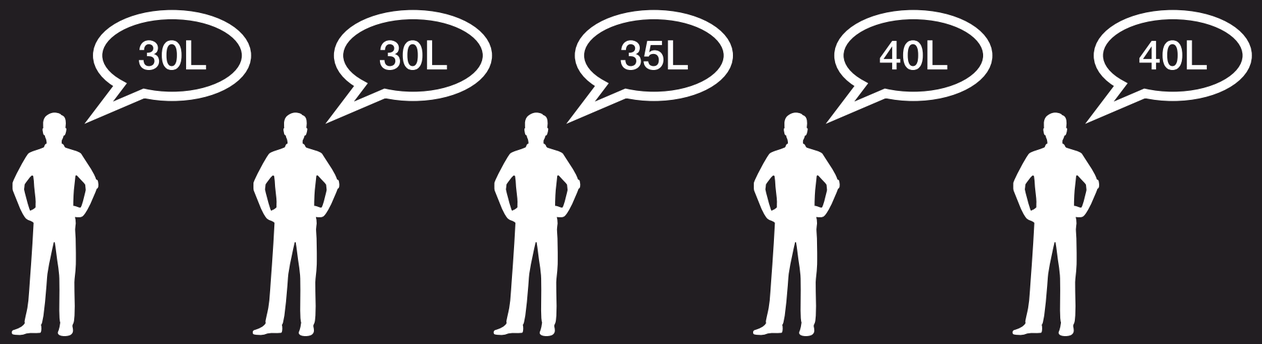
Inferential statistics

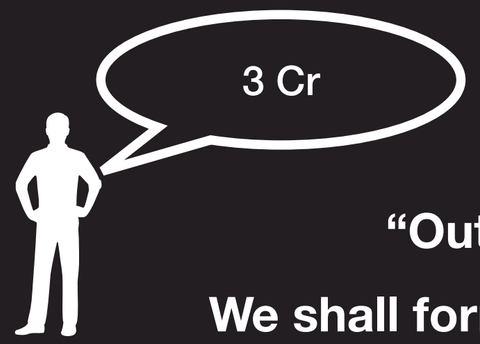
Drawing conclusions from observations

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





"Outlier"

We shall formalise this

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

Another word for mean is "Average"

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

Crucial observation: Median is more robust to outliers

Median = $35L$

Central value (if unique)

$N = 5$, odd

New Median = $37.5L$

Average of 2 central values

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

$N = 6$, 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?

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

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

$$x = 26$$

$$20,\ 22,\ 26,\ 28$$

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

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

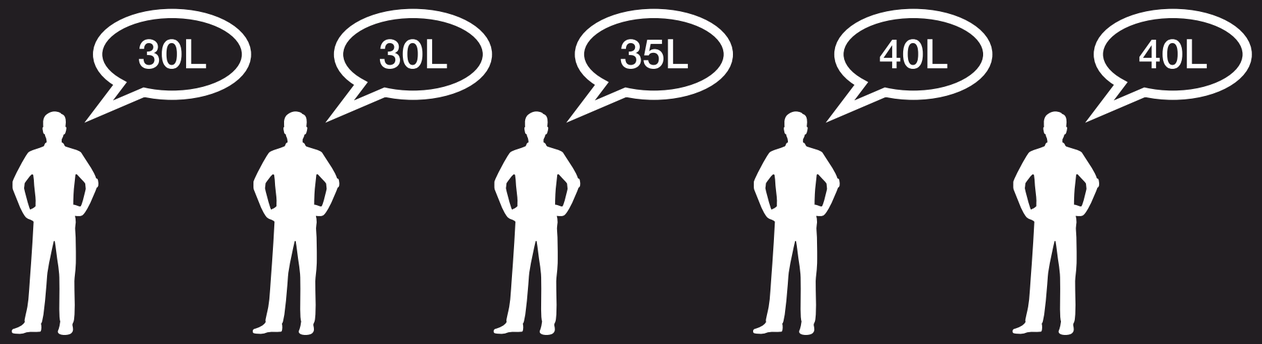
Range

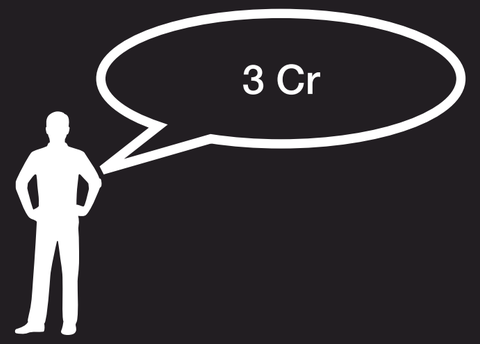
Suppose a cricketer has scored as follows

20, 25, 60, 100

We say the range = 100 - 20 = 80

Consider again the example of salaries



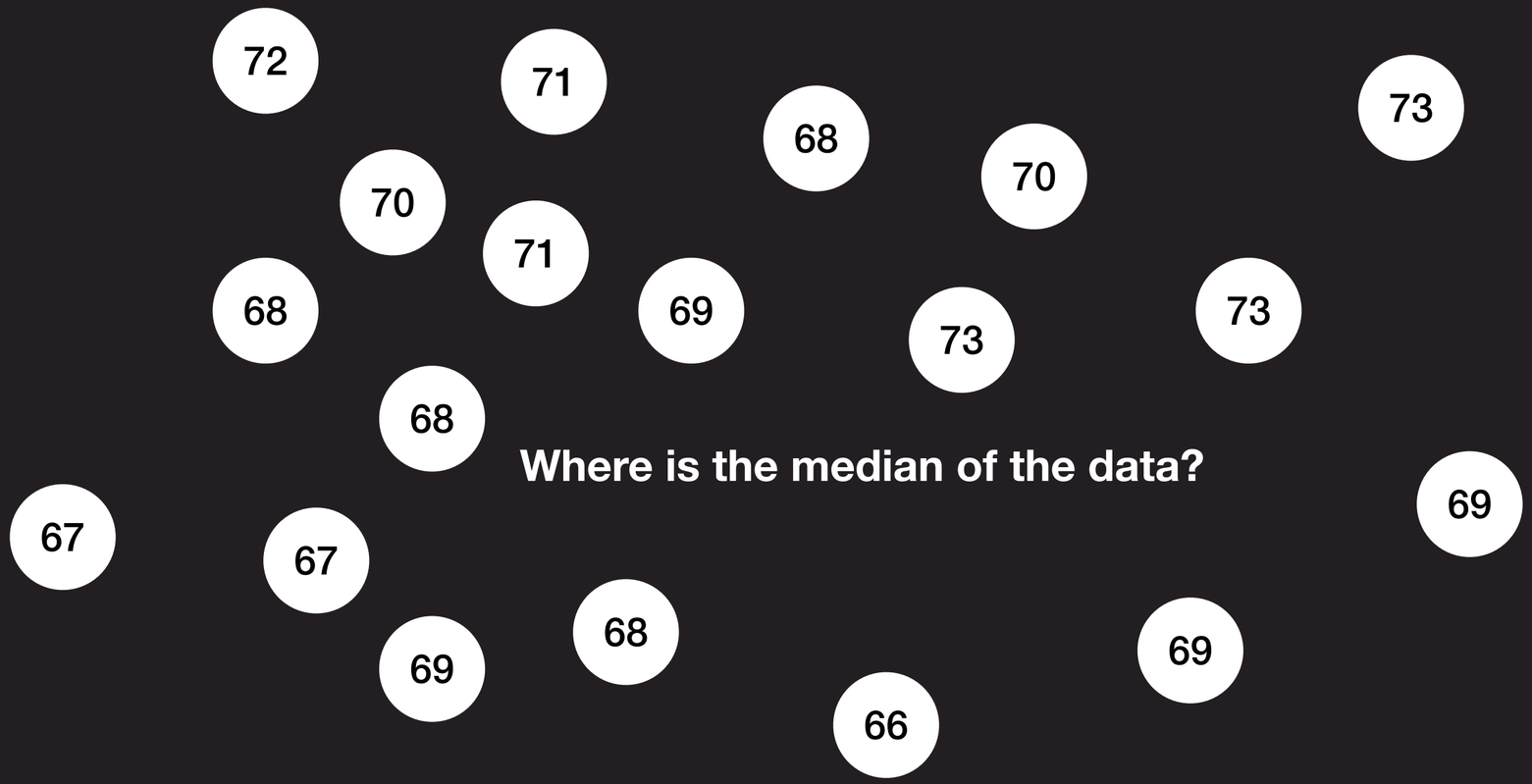


Here range = 300 - 30 = 270 L

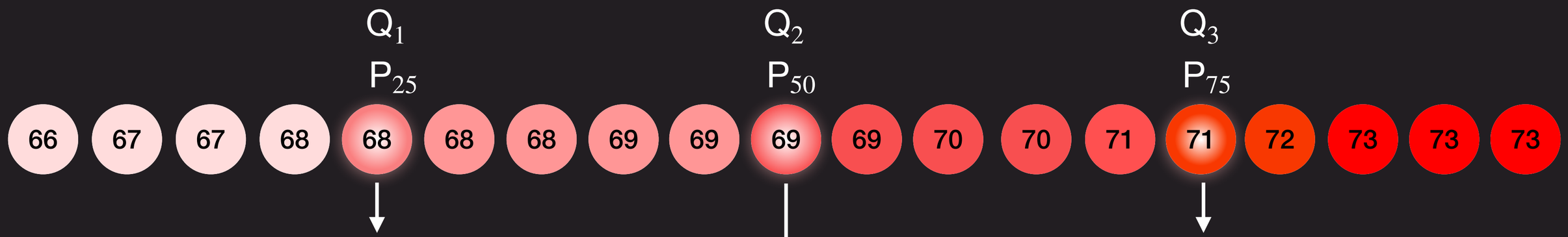
Sometimes, simply giving range may not make sense

Percentiles, Quartiles, Inter Quartile range (IQR)

Sort the data!



Percentiles, Quartiles, Inter Quartile range (IQR)



What is the median of the first half?

What fraction of data is less that this? $25\,\%$

Another name for this is 25^{th} **percentile -** P_{25}

Where is the median of the data?

What fraction of data is less that this? $50\,\%$

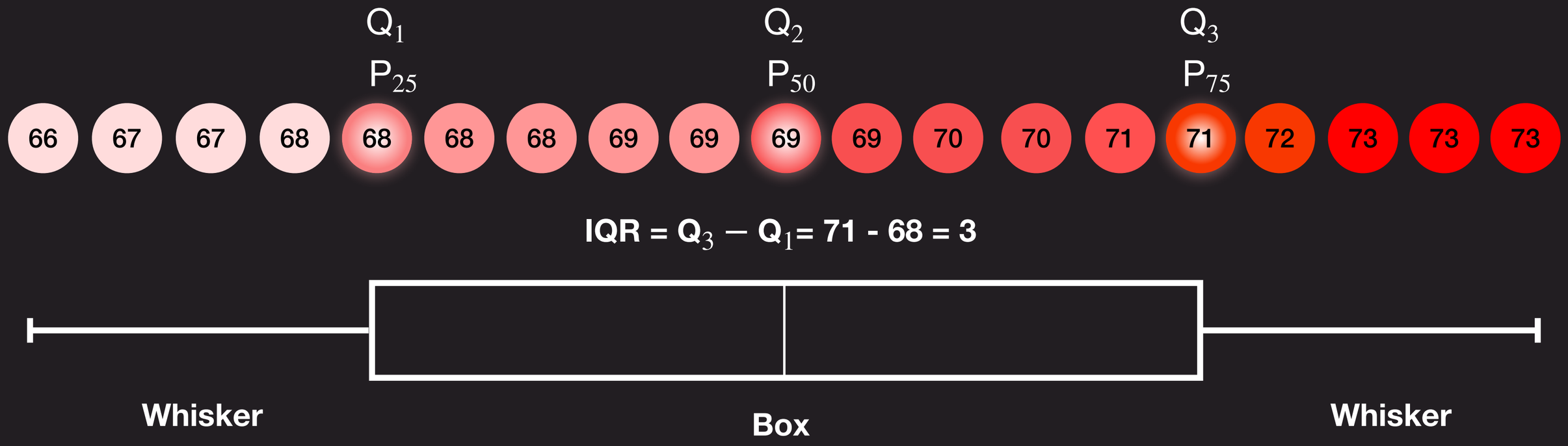
Another name for this is 50^{th} **percentile -** P_{50}

What is the median of the second half?

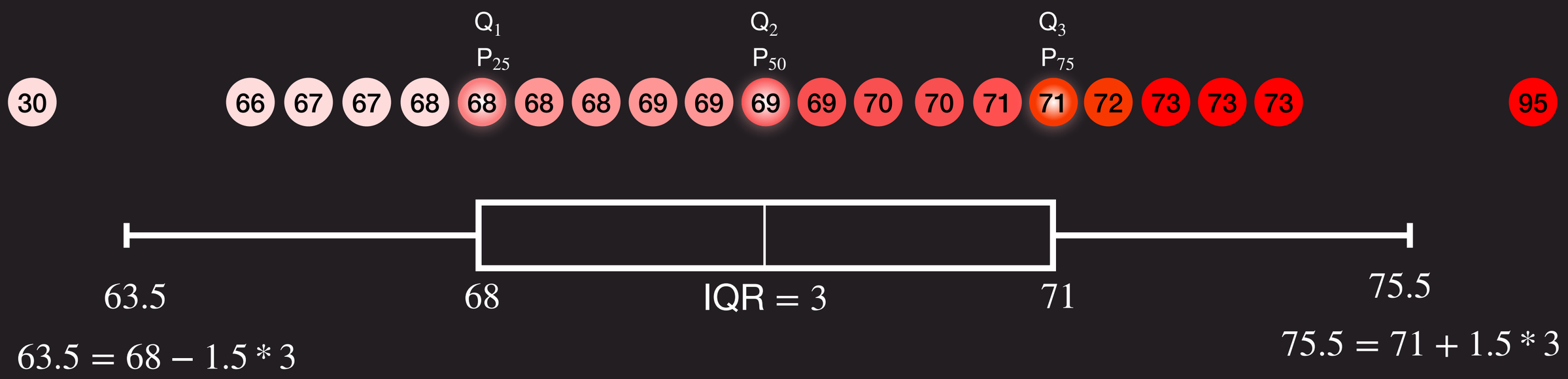
What fraction of data is less that this? $75\,\%$

Another name for this is 75^{th} **percentile -** P_{75}

Percentiles, Quartiles, Inter Quartile range (IQR)



Percentiles, Quartiles, Inter Quartile range (IQR)



$63.5 = 68 - 1.5 * 3$

$75.5 = 71 + 1.5 * 3$

Outlier are points outside $[63.5,\ \ 75.5]$

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

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

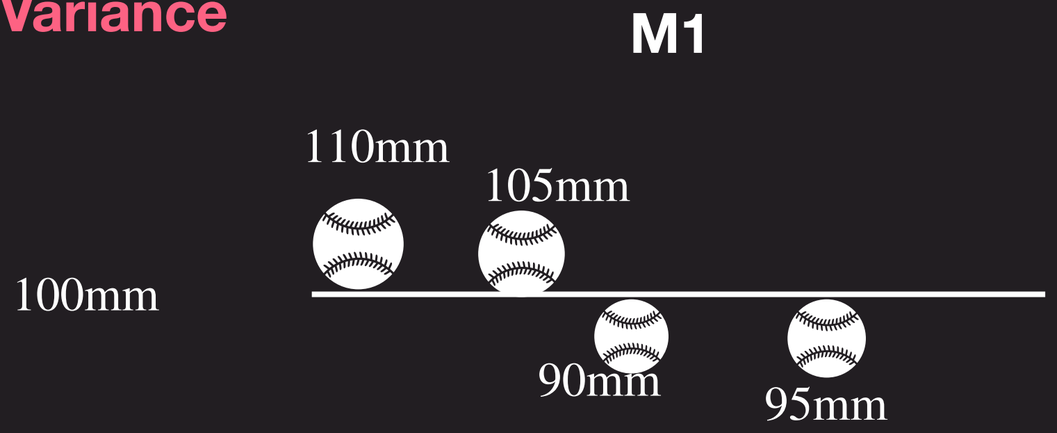
Median = 61.6

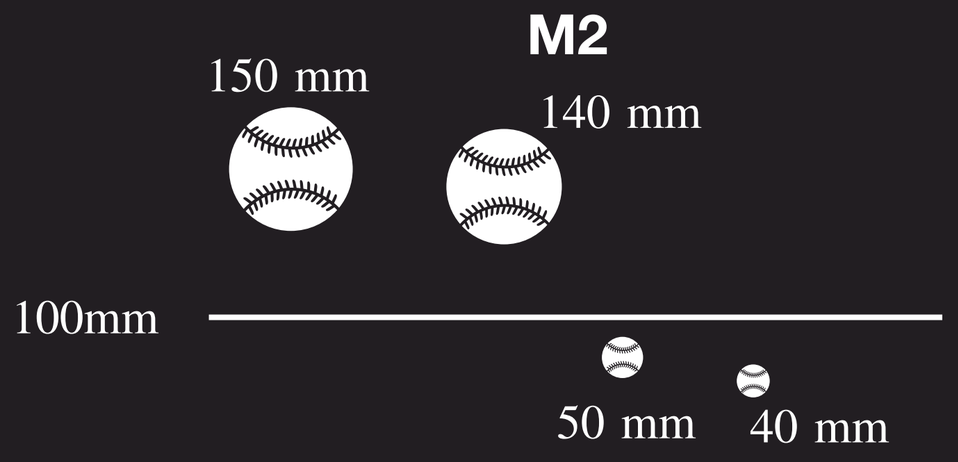
Mode = 61.6

Range = 66.88 - 52.8 = 14.08

IQR = 64.24 - 54.56 = 9.68

Variance





How to define Error?

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

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

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

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

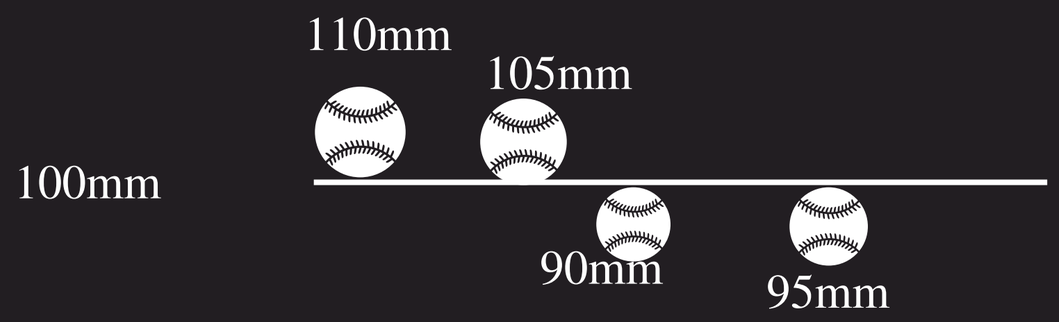
$(50\ \text{mm})^2 + (40\ \text{mm})^2 + (-50\ \text{mm})^2 + (-40\ \text{mm})^2 = 8200$

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

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

Variance

M1



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

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

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

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

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

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

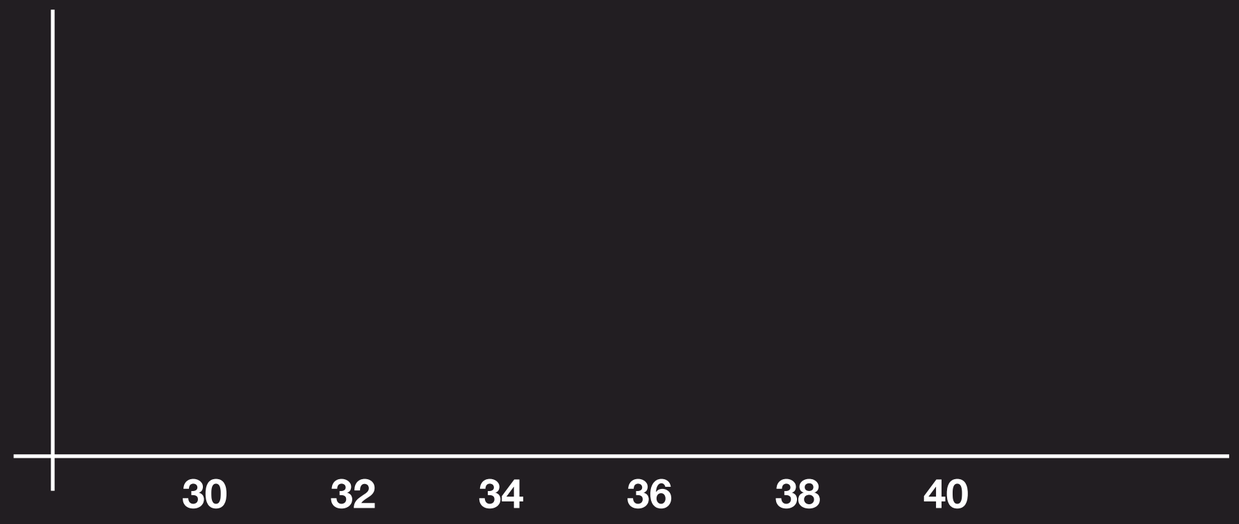
$$\text{Std Dev} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2}{4}}$$

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

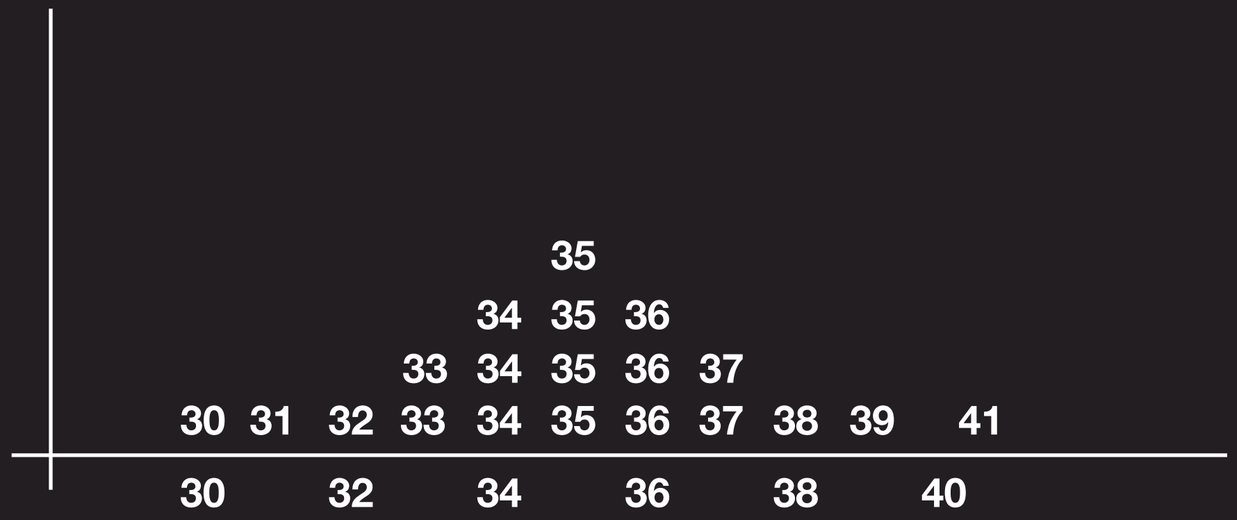
$$\text{Variance} = \frac{\sum_i (x_i - \bar{x})^2}{n} = \sigma^2$$

Histogram

32
36
33
34
38
34
35
36
39
36
33
35
37
34
35
37
30
35
41
31



Histogram


35
34 35 36
33 34 35 36 37
30 31 32 33 34 35 36 37 38 39 41
30 32 34 36 38 40

Histogram