Centre – The Middle of a Distribution

- An average is the single value that best represents the centre of a series of values.
- These values could be the shooting percentage of your favourite basketball team or the average rate at which convicted criminals re-offend.

{7, 8, 21, 22, 22, 23}

series of values

Centre – The Middle of a Distribution

- An average is the single value that best represents the centre of a series of values.
- These values could be the shooting percentage of your favourite basketball team or the average rate at which convicted criminals re-offend.
- An average is like the fulcrum of a seesaw – the point at which a seesaw is perfectly level.

Centre – The Middle of a Distribution

- Averages are known as **measures of central tendency** and have three variations, the
  1. Mean
  2. Median
  3. Mode

"Seesaw" by dianaholga: CC BY-NC-ND 2.0

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

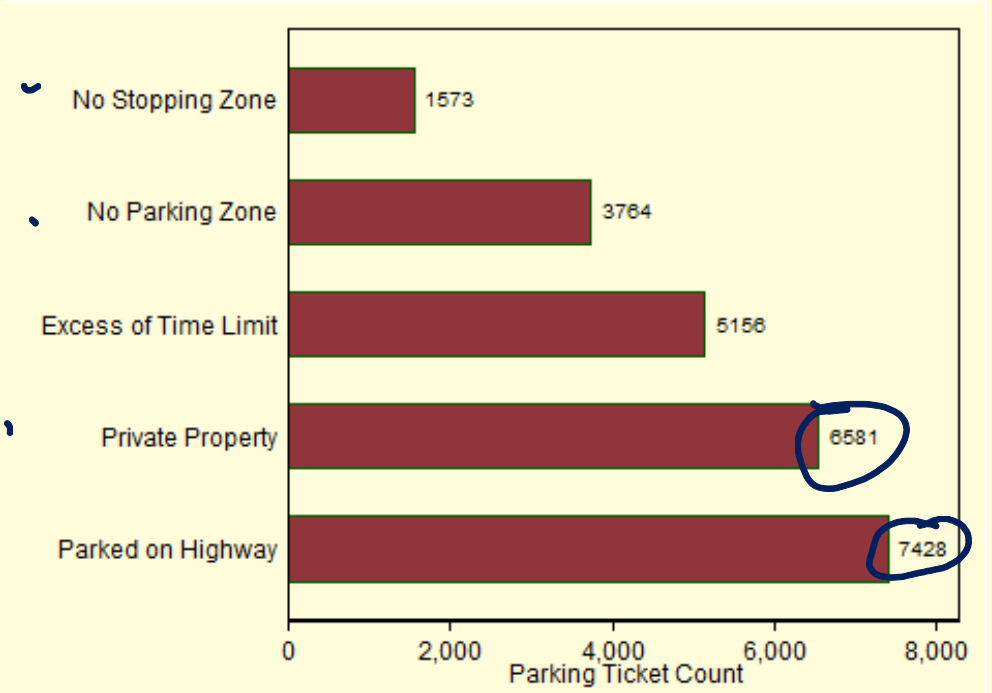
- $\bar{X}$ is the mean.
- $\sum$, the Greek letter sigma, instructs us to sum or add-up.
- X_i is the individual value.
- n is the number of values in a series, often called the **sample** or **population size**. number of observations

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

- $\bar{X}$ is the mean.
- $\sum$, the Greek letter sigma, instructs us to sum or add-up.
- X_i is the individual value.
- n is the number of values in a series, often called the **sample** or **population size**.

Parking tickets issued by the City of Guelph, Ontario

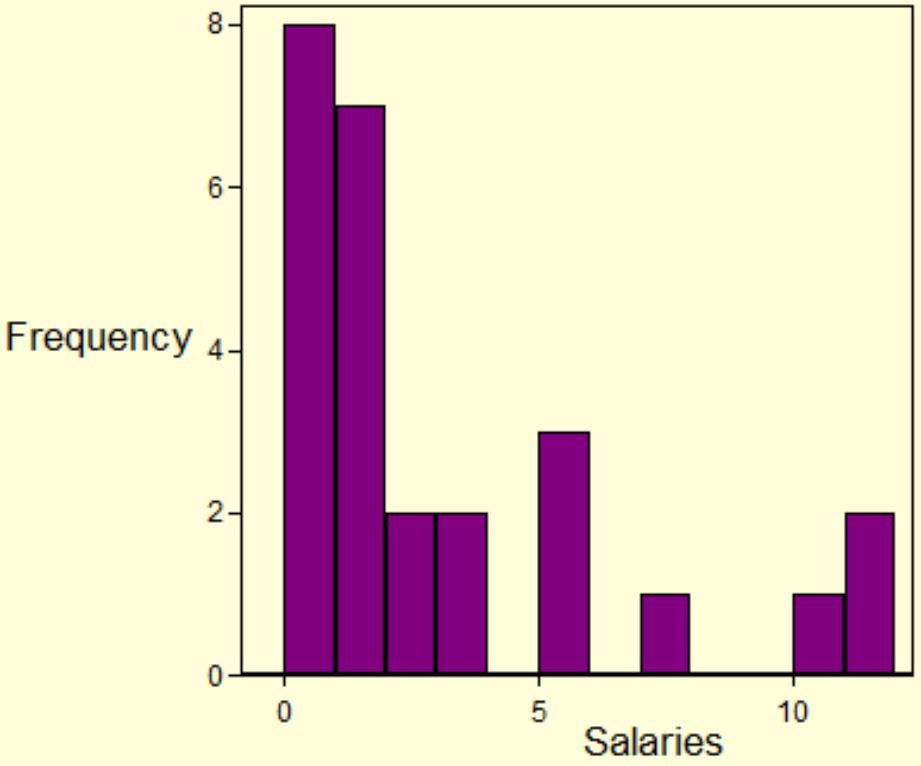


mean # of tickets issued $= \frac{1573 + 3764 + 5156 + 6581 + 7428}{5}$

$= 24,502 / 5 = 4,900.4$

The Mean – Maple Leafs Player Salaries 2021-22

The Mean – Maple Leafs Player Salaries 2021-22



Stem-and-Leaf Plot

stem	leaf
0	77789999
1	0225566
2	05
3	58
4	
5	006
6	
7	0
8	
9	
10	9
11	06

millions of US dollars

3.5 m

3.8 m

n = 26

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

- The Mean is the Average or most central score.

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

- The Mean is the Average or most central score.
- Sometimes a capital N represents the number of values. Other times, it is represented by a small n.

The Mean

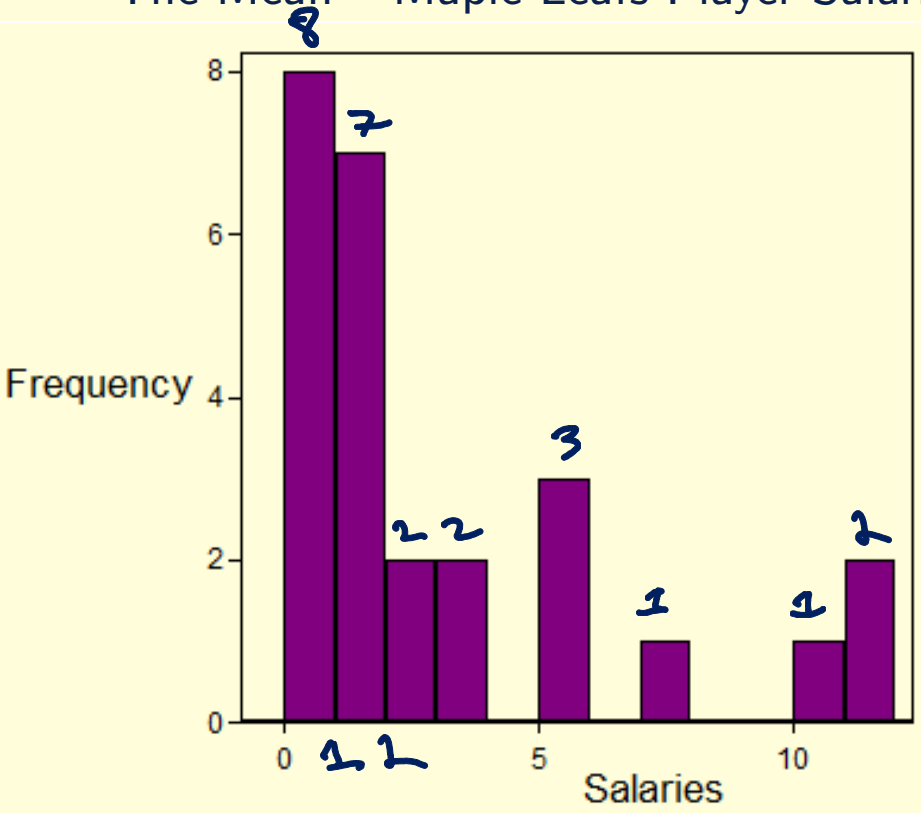
$$\bar{X} = \frac{\sum X_i}{n}$$

- The Mean is the Average or most central score.
- Sometimes a capital N represents the number of values. Other times, it is represented by a small n.

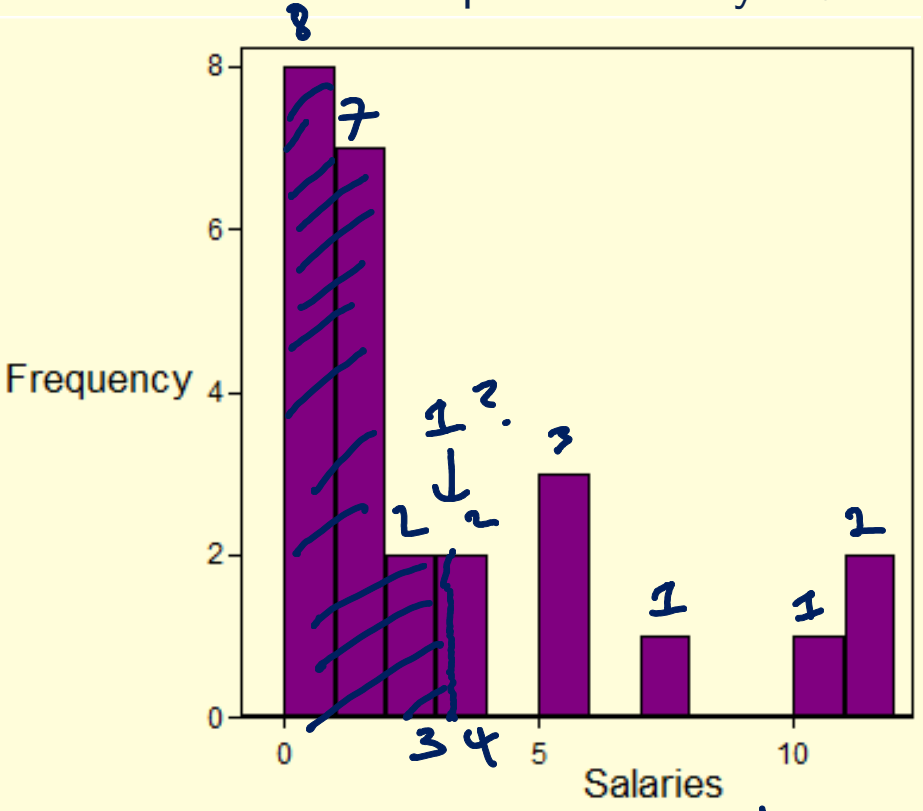
- The mean is like a fulcrum on a seesaw – the weight of the values below the mean equals the weight of the values above of the mean.

"Seesaw" by dianaholga: CC BY-NC-ND 2.0

The Mean – Maple Leafs Player Salaries 2021-22



The Mean – Maple Leafs Player Salaries 2021-22



Stem-and-Leaf Plot	
0	77789999
1	0225566
2	05
3	58
4	
5	006
6	
7	0
8	
9	
10	9
11	06

Mean = \$3.23 million

Top 9 players make as much as the other.
17 or 18 players.

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

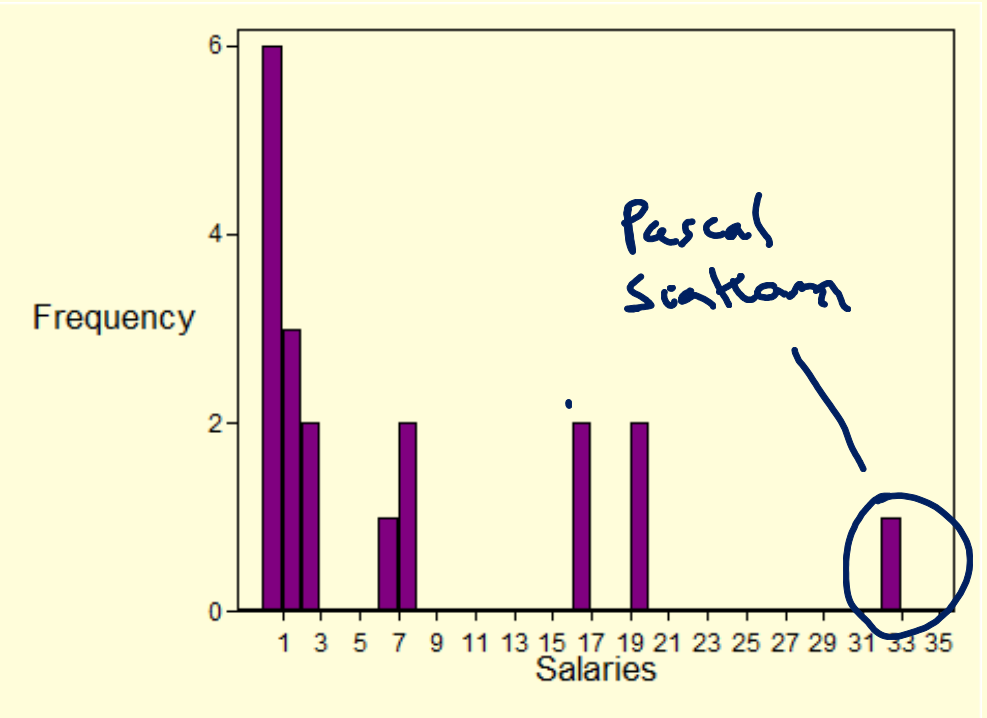
- A strength and weakness of the mean is its sensitive to extreme values (also known as **outliers**).

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

- A strength and weakness of the mean is its sensitive to extreme values (also known as **outliers**).

The Toronto Raptors Salary: 2021-22



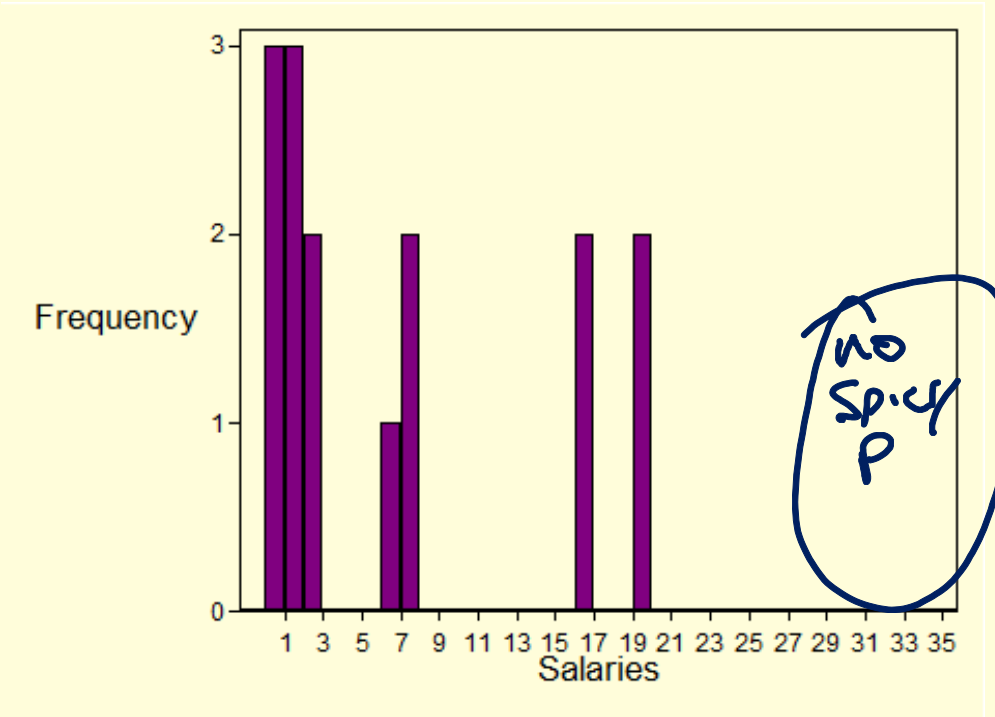
Mean = \$8.5 million

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

- A strength and weakness of the mean is its sensitive to extreme values (also known as **outliers**).

The Toronto Raptors Salary: 2021-22



Mean without Pascal Siakam = \$5.67 million

The Mean

Five Largest Toronto Raptors Salaries 2021-22 (rounded) – Stem-and-Leaf Diagram

1	669
2	0
3	3

in tens of millions of US dollars

The Mean

Five Largest Toronto Raptors Salaries 2021-22 (rounded) – Stem and Leaf Diagram

Tens of millions →

1	669
2	0
3	3

in tens of millions of US dollars

- Find the mean salary for the five NBA players.
- If we added a sixth NBA player, what salary would he need to have for the mean to become \$16 million?

$$\bar{x} = \frac{16m + 16m + 19m + 20m + 33m}{5}$$

$\bar{x} = 104/5$

$\bar{x} = 20.8$ million

$$\bar{x}_6 = \frac{104 + x_i}{5+1} = \$16$$

$104 + x_i = 96$

$x_i = -8$!

Not possible unless the salary is negative!

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

- The **arithmetic mean**: the sum of the deviations from the mean is equal to zero.

Deviation from the Mean $= X_i - \bar{X}$

The Mean

$$\bar{X} = \frac{\sum X_i}{n}$$

- The **arithmetic mean**: the sum of the deviations from the mean is equal to zero.

Deviation from the Mean $= X_i - \bar{X}$

Salaries	
Player 1	4,000,000
Player 2	5,000,000
Player 3	6,000,000

The Weighted Mean

How to Compute a Weighted Mean

- Find the value by multiplying the frequency and the parking fine.

The Weighted Mean

How to Compute a Weighted Mean

- Find the value by multiplying the frequency and the parking fine.

Parking Tickets – City of Guelph

Fine	Frequency	Value
\$30	2,209	
\$35	13,007	
\$40	5,410	
\$51	1,271	
\$60	1,488	
\$91	2,132	

Find the weight mean fine on a parking ticket.

The Weighted Mean

How to Compute a Weighted Mean

- Find the value by multiplying the frequency and the parking fine.

Parking Tickets – City of Guelph

Fine	Frequency	Value
\$30	2,209	66,270
\$35	13,007	455,245
\$40	5,410	216,400
\$51	1,271	64,821
\$60	1,488	89,280
\$91	2,132	194,012

Find the weight mean fine on a parking ticket.

The Weighted Mean

How to Compute a Weighted Mean

- Find the value by multiplying the frequency and the parking fine.
- Sum all the values and frequencies.

Parking Tickets – City of Guelph

Fine	Frequency	Value
\$30	2,209	66,270
\$35	13,007	455,245
\$40	5,410	216,400
\$51	1,271	64,821
\$60	1,488	89,280
\$91	2,132	194,012

Find the weight mean fine on a parking ticket.

The Weighted Mean

How to Compute a Weighted Mean

- Find the value by multiplying the frequency and the parking fine.
- Sum all the values and frequencies.

Parking Tickets – City of Guelph

Fine	Frequency	Value
\$30	2,209	66,270
\$35	13,007	455,245
\$40	5,410	216,400
\$51	1,271	64,821
\$60	1,488	89,280
\$91	2,132	194,012
Totals	25517	1086028

Find the weight mean fine on a parking ticket.

The Weighted Mean

How to Compute a Weighted Mean

- Find the value by multiplying the frequency and the parking fine.
- Sum all the values and frequencies.
- Divide the values by the total frequency to find the weighted average.

Parking Tickets – City of Guelph

Fine	Frequency	Value
\$30	2,209	66,270
\$35	13,007	455,245
\$40	5,410	216,400
\$51	1,271	64,821
\$60	1,488	89,280
\$91	2,132	194,012
	25,517	1,086,028

The Weighted Mean

How to Compute a Weighted Mean

- Find the value by multiplying the frequency and the parking fine.
- Sum all the values and frequencies.
- Divide the values by the total frequency to find the weighted average.

Parking Tickets – City of Guelph

Fine	Frequency	Value
\$30	2,209	66,270
\$35	13,007	455,245
\$40	5,410	216,400
\$51	1,271	64,821
\$60	1,488	89,280
\$91	2,132	194,012
	25,517	1,086,028

Find the weight mean fine on a parking ticket:

$$1{,}086{,}028 \;/\; 25{,}517 = \$42.6$$

The Weighted Mean: Weight on Each Fine Dollar Amount

- Find the weighting on each ticket fine by divide the value of each fine by the total value.

$\frac{216\,400}{1,086,028} = 19.9\%$

$\frac{64,821}{1086028} = 6\%$

Parking Tickets – City of Guelph

Fine	Frequency	Value	Weight
\$30	2,209	66,270	6.1%
\$35	13,007	455,245	41.9%
\$40	5,410	216,400	19.9%
\$51	1,271	64,821	6%
\$60	1,488	89,280	8.2%
\$91	2,132	194,012	17.9%
	25,517	1,086,028	100%

The Weighted Mean

- Find the weighting on each ticket fine by divide the value of each fine by the total value.

Parking Tickets – City of Guelph

Fine	Frequency	Value	Weight
\$30	2,209	66,270	6.1%
\$35	13,007	455,245	41.9%
\$40	5,410	216,400	19.9%
\$51	1,271	64,821	6.0%
\$60	1,488	89,280	8.2%
\$91	2,132	194,012	17.9%
	25,517	1,086,028	100%

$\bar{x} = 4,252.7$

The weight represents the relative contribution of each fine to total revenue collected.

The Median

The median is

- a different kind of **average**.

The Median

The median is

- a different kind of **average**.
- the **midpoint** of a series where half the values are below the median and half the values are above the median.

The Median

The median is

- a different kind of **average**.
- the **midpoint** of a series where half the values are below the median and half the values are above the median.

To find the median:

The Median

The median is

- a different kind of **average**.
- the **midpoint** of a series where half the values are below the median and half the values are above the median.

To find the median:

- List the values in order.

The Median

The median is

- a different kind of **average**.
- the **midpoint** of a series where half the values are below the median and half the values are above the median.

To find the median:

- List the values in order.
- Find the middle score.

The Median

The median is

- a different kind of **average**.
- the **midpoint** of a series where half the values are below the median and half the values are above the median.

To find the median:

- List the values in order.
- Find the middle score.
- If there are an even number of values, the two middle values are the ~~mean~~. median → {1.5, 1.6}

Maple Leafs Salaries 2021-22 – Stem-and-Leaf Plot

millions | hundreds of thousands

#		
8	0	77789999
7	1	0225566
2	2	05
2	3	58
	4	
3	5	006
	6	
1	7	0
	8	
	9	
1	10	9
2	11	06
26		

15 → 13, 14, 15

26/2 = 13 → 26+1/2 = 13.5

The Median

The median is

- a different kind of average.
- the **midpoint** of a series where half the values are below the median and half the values are above the median.

To find the median:

- List the values in order.
- Find the middle score. ~~Item~~ If there are an even number of values, the two middle values are the mean.

Five Largest Toronto Raptors Salaries 2021-22 (rounded) – Stem and Leaf

Tens of millions → | millions

1	669
2	0
3	3

{16, 16, 19, 20, 33} 5 values

↑ Median = 19 million

The Median v Mean

average
central tendency

- The mean is the middle point where
  - the value below and above the mean are equally weighted, and
  - the sum of the deviations from the mean equals zero.

fulcrum of a seesaw

$$\sum_{i=1} (x_i - \bar{x}) = 0$$

The Median v Mean

- The mean is the middle point where
  - the value below and above the mean are equally weighted, and
  - the sum of the deviations from the mean equals zero.
- The median is the middle point of a set of cases.

the midpoint in a series of values

The Median v Mean

- The mean is the middle point where
  - the value below and above the mean are equally weighted, and
  - the sum of the deviations from the mean equals zero.
- The median is the middle point of a set of cases.
- The median is insensitive to extreme values, while the mean is sensitive to extreme values.

The Median v Mean

- The mean is the middle point where
  - the value below and above the mean are equally weighted, and
  - the sum of the deviations from the mean equals zero.
- The median is the middle point of a set of cases.
- The median is insensitive to extreme values, while the mean is sensitive to extreme values.

$X = \{-10, 7, 9\}$

median is still 7!

$\bar{X} = \frac{-10+7+9}{3}$

$\bar{X} = 2$

Mean

median

$X = \{2, 7, 9\}$

$\bar{X} = \frac{2+7+9}{3} = \frac{18}{3} = 6$

$\bar{X} = 6$

$x - \bar{x} = \{2-6, 7-6, 9-6\}$

$\{-4, 1, 3\}$

$\sum(x - \bar{x}) = -4 + 1 + 3$

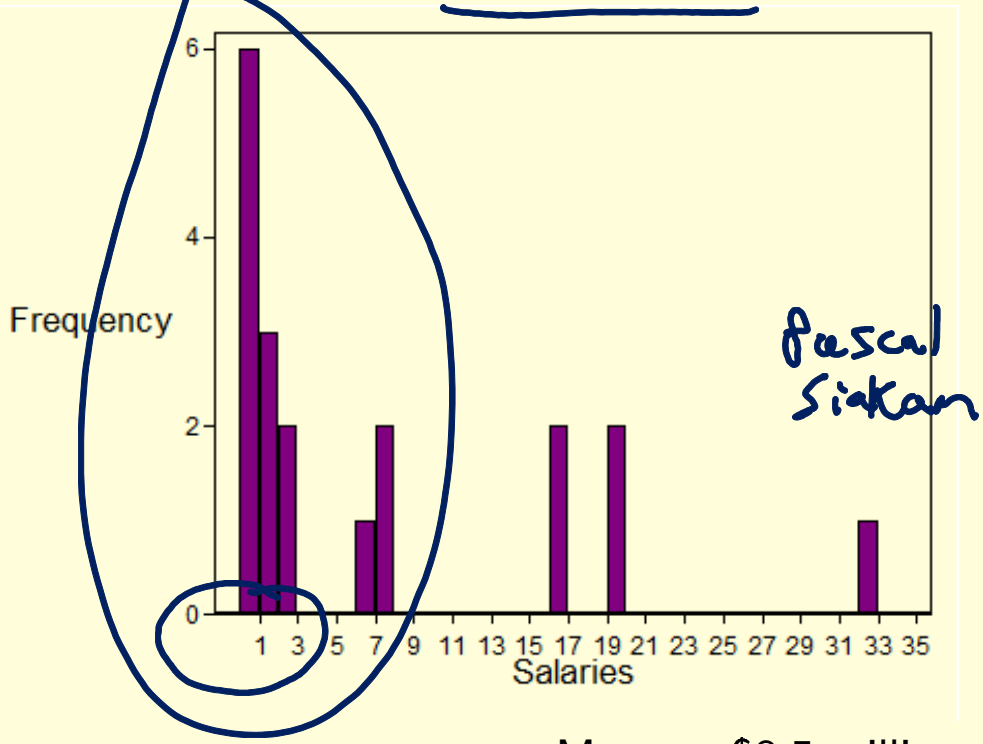
$= 0$

$\bar{x} = 6$ median $= 7$

The Median v Mean

- The mean is the middle point where
  - the value below and above the mean are equally weighted, and
  - the sum of the deviations from the mean equals zero.
- The median is the middle point of a set of cases.
- The median is insensitive to extreme values, while the mean is sensitive to extreme values.

The Toronto Raptors Salary: 2021-22

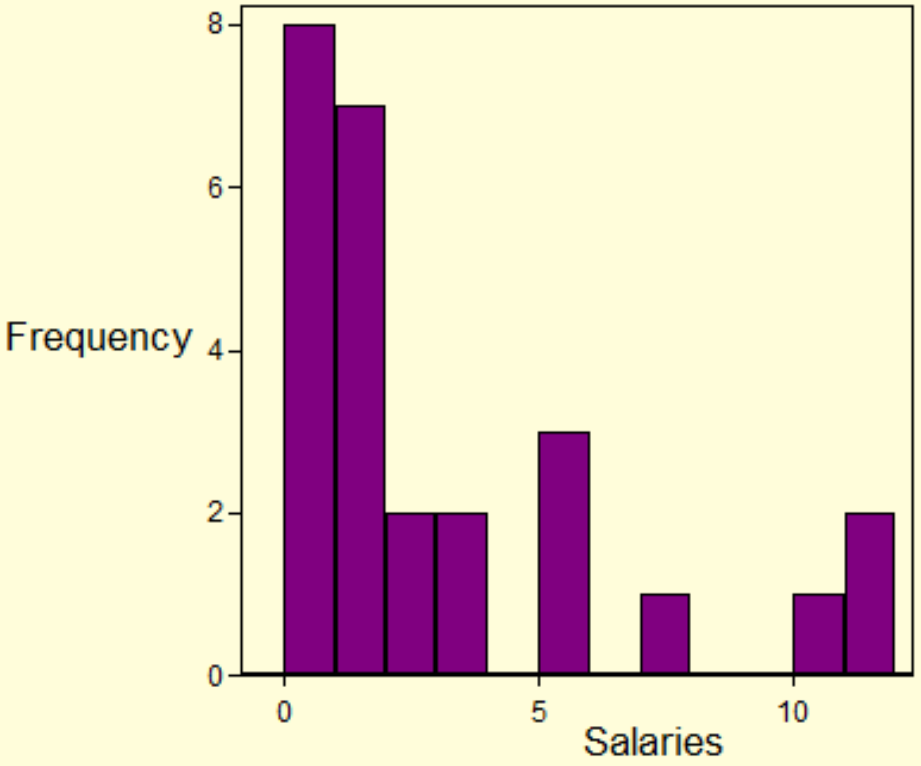


Pascal Siakam

Mean = $8.5 million

Median = ~~$4.5 million~~ $20 million

The Median v Mean



Stem-and-Leaf Plot for Maple Leafs Salaries 2021-22

0	77789999
1	0225566
2	05
3	58
4	
5	006
6	
7	0
8	
9	
10	9
11	06

The Median v Mean

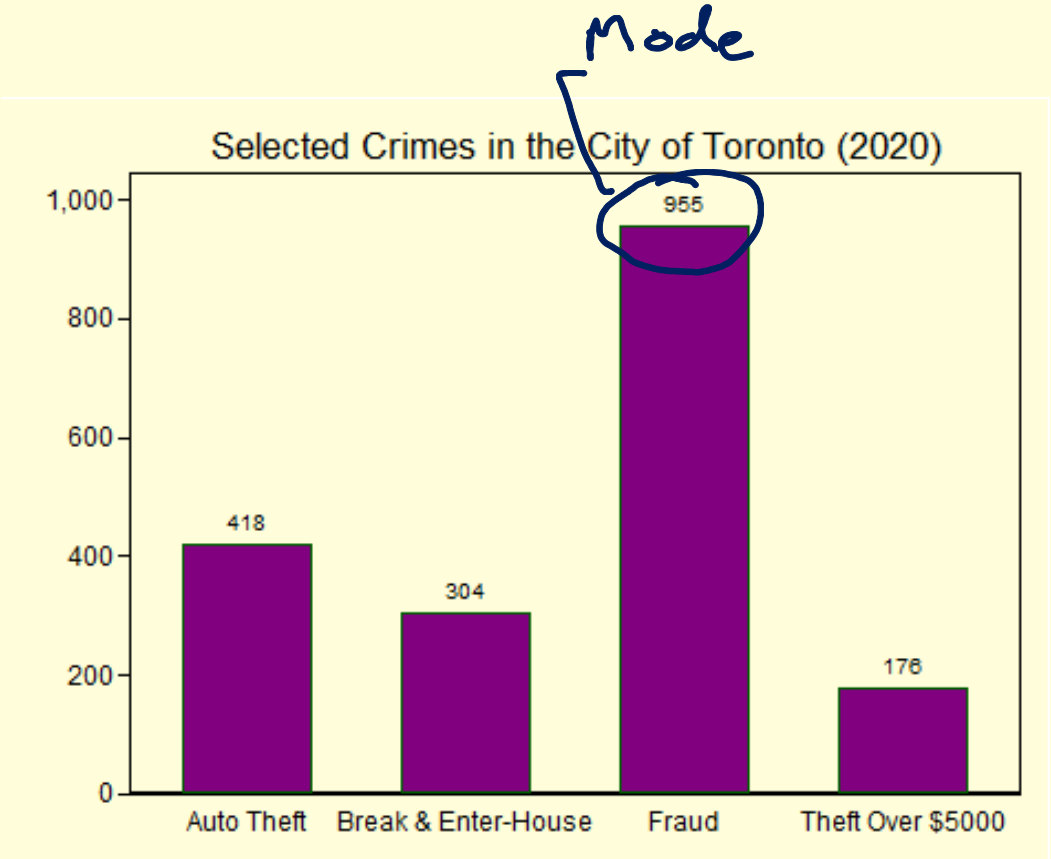
Stem-and-Leaf Plot for Maple Leafs Salaries 2021-22

0	77789999
1	0225566
2	05
3	58
4	
5	006
6	
7	0
8	
9	
10	9
11	06

Mean = \$3.23 million
Median = \$1.55 million

The Mode

- The mode is the value that occurs most often.
- Of the four crimes, which occurs most often?



What is the Mode?

Toronto Maple Leafs Salaries 2021-22

0	77789999
1	0225566
2	05
3	58
4	
5	006
6	
7	0
8	
9	
10	9
11	06

mode = 0.1 million

What is the Mode?

Toronto Maple Leafs Salaries 2021-22

0	77789999
1	0225566
2	05
3	58
4	
5	006
6	
7	0
8	
9	
10	9
11	06

Mean: \$3.23 million
Median: \$1.55 million
Mode: 0.9 million

What is the Mode?

Toronto Maple Leafs Salaries 2021-22

0	77789999
1	0225566
2	05
3	58
4	
5	006
6	
7	0
8	
9	
10	9
11	06

Mean: \$3.23 million
Median: \$1.55 million
Mode:

Find the Mode Response

Student Responses to a Test Question

	A	B	C	D	E
Frequency	22	31	44	45	9

What is Variability?

- **Variability** reflects the difference in values.

What is Variability?

- **Variability** reflects the difference in values.
- Variability is also known as **dispersion** or **spread**.

What is Variability?

- **Variability** reflects the difference in values.
- Variability is also known as **dispersion** or **spread**. $X = \{3, 3, 4, 4\}$ and $Y = \{0, 1, 6, 7\}$
- Compute the Mean of X and Y:
- Compute the Deviation from the Mean for X and Y.

$\bar{X} = \frac{3+3+4+4}{4} = \frac{14}{4} = 3.5$

$\bar{Y} = \frac{0+1+6+7}{4} = \frac{14}{4} = 3.5$

X	Y
$3 - 3.5 = -.5$	$0 - 3.5 = -3.5$
$3 - 3.5 = -.5$	$1 - 3.5 = -2.5$
$4 - 3.5 = .5$	$6 - 3.5 = 2.5$
$4 - 3.5 = .5$	$7 - 3.5 = 3.5$
	0

Y varies more than X.

What is Variability?

$X = \{3, 3, 4, 4\}$ and $Y = \{0, 1, 6, 7\}$
and $Z = \{3.5,\ 3.5,\ 3.5,\ 3.5\}$

- Do any values in Z deviate from the mean?

$\bar{Z} = 3.5 \qquad 3.5 - 3.5 = 0$

Z has no variability

What is Variability?

$X = \{3, 3, 4, 4\}$ and $Y = \{0, 1, 6, 7\}$
and $Z = \{3.5,\ 3.5,\ 3.5,\ 3.5\}$

- Do any values in Z deviate from the mean?
- Z has no variability at all!

Three measures of variability

- The Range IQR
- The Variance
- The Standard Deviation

The Range

- The range measures how far apart the values are.

The Range

- The range measures how far apart the values are.
- It is measured as the highest value in a series less the lowest value in a series.

The Range

- The range measures how far apart the values are.
- It is measured as the highest value in a series less the lowest value in a series.

Find the range for X, Y, Z:

$X = \{3, 3, 4, 4\}$

$Y = \{0, 1, 7, 8\}$

$Z = \{3.5,\ 3.5,\ 3.5,\ 3.5\}$

Range of $X = 4 - 3 = 1$

$Y = 8 - 0 = 8$

$Z = 3.5 - 3.5 = 0$

Find the Range

Toronto Maple Leafs Salaries 2021-22

0	77789999
1	0225566
2	05
3	58
4	
5	006
6	
7	0
8	
9	
10	9
11	06

$11.6 - 0.7$

$= 10.9$ million

Range

Selected Toronto Raptors Salaries 2021-22

0.	06,07,07
1.	669
2.	0
3*	3

Hundreds of Thousands

millions

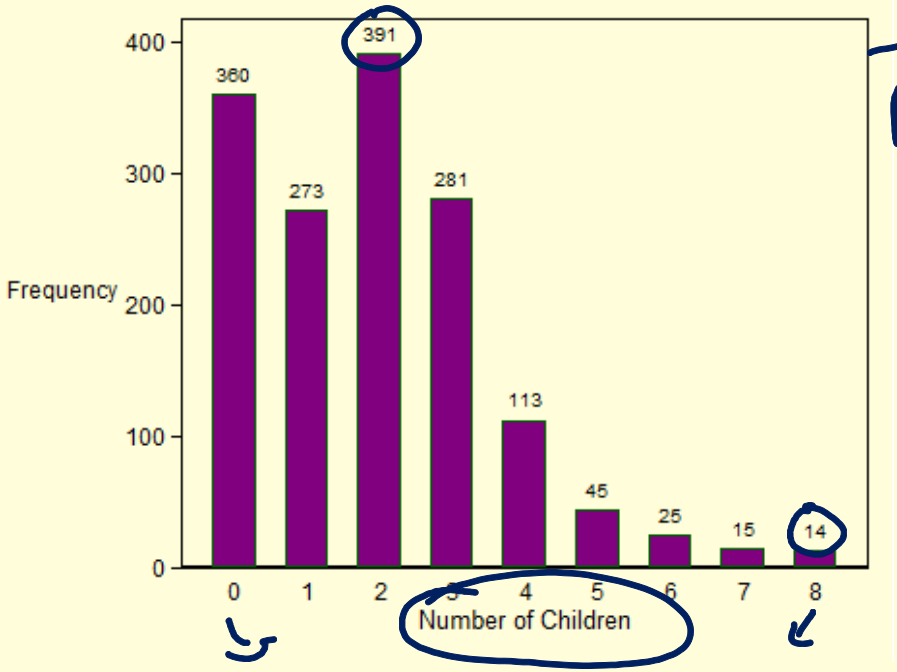
Tens of millions

Range: $33,000,000 - 600,000$

$32,400,000$

Range

Find the range



~~Range: 8 − 0 = 8~~

Range: 391 − 14 = 377

The Range

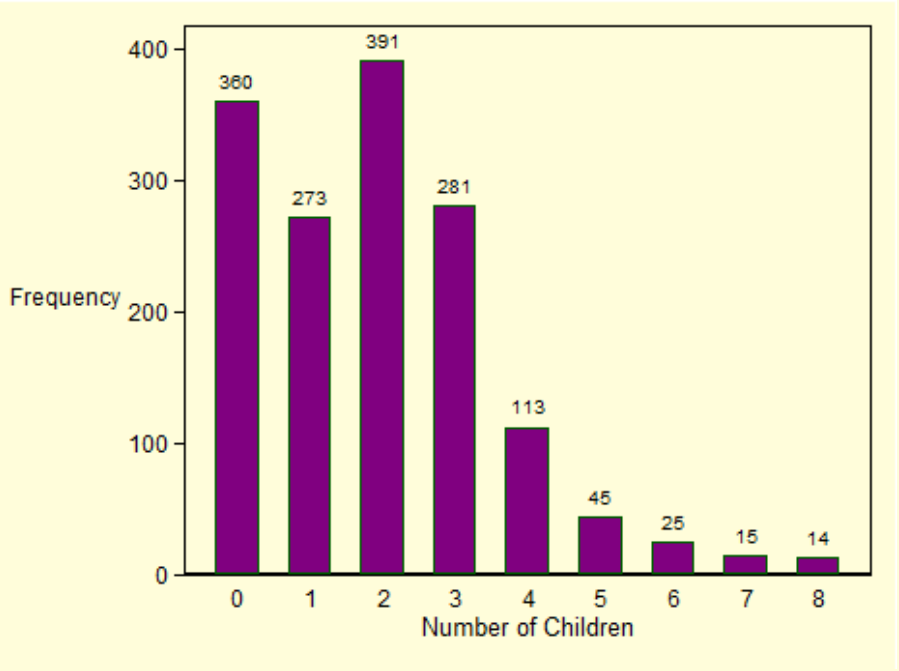
- The range is simple and only provides information about the end points which might be extremely large or small.

The Range

- The range is simple and only provides information about the end points which might be extremely large or small.
- The range provides no information on the interior values.

The Range

- The range is simple and only provides information about the end points which might be extremely large or small.
- The range provides no information on the interior values.
- Standard Deviation and Variance are more common measures of the spread (dispersion).

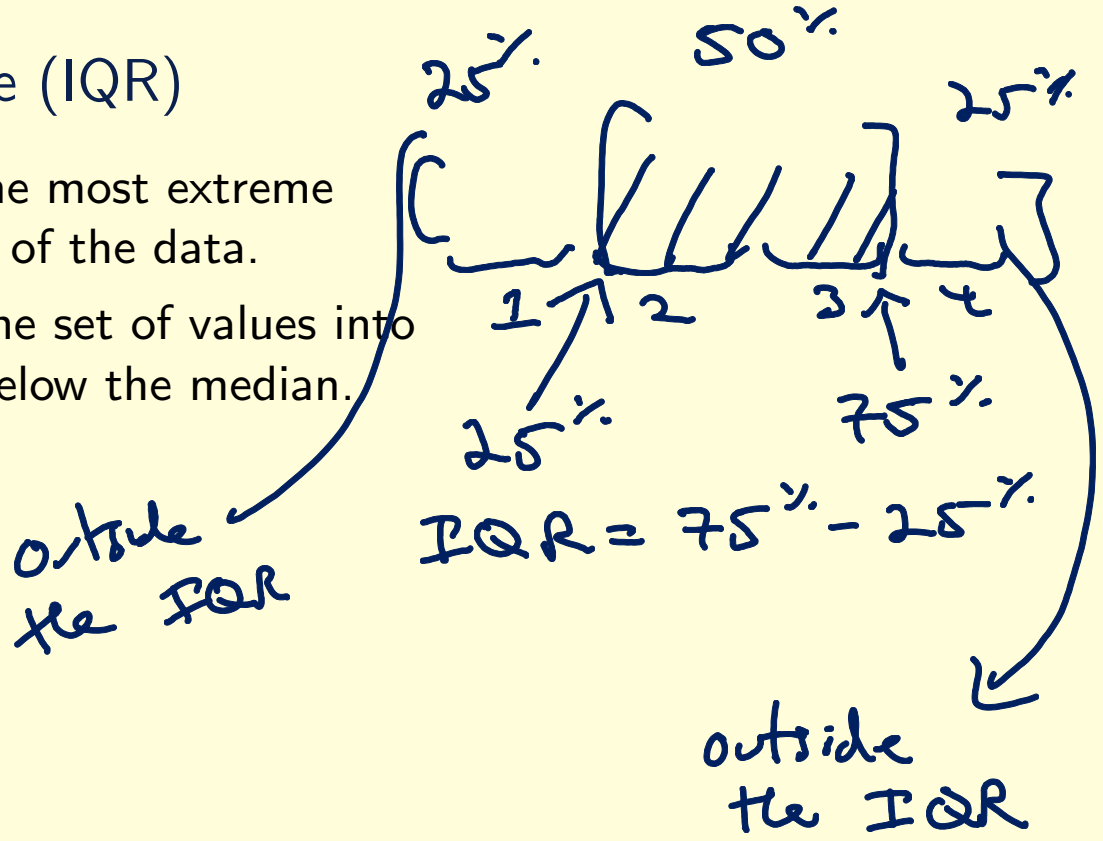


The Interquartile Range (IQR)

- The interquartile range ignores the most extreme values and focuses on the middle of the data.

The Interquartile Range (IQR)

- The interquartile range ignores the most extreme values and focuses on the middle of the data.
- Remember, the median divided the set of values into two halves: one above and one below the median.



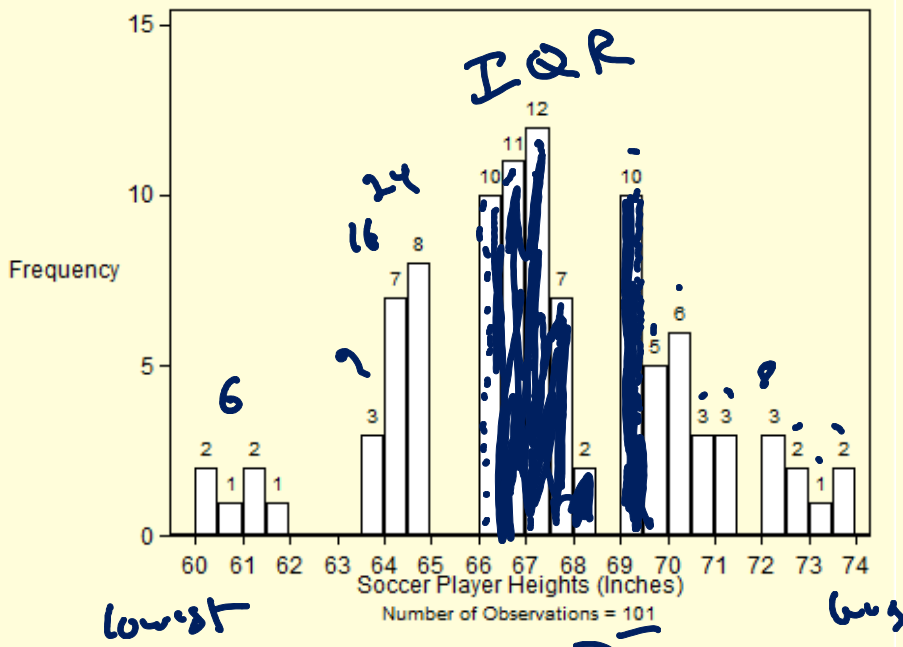
The Interquartile Range (IQR)

- The interquartile range ignores the most extreme values and focuses on the middle of the data.
- Remember, the median divided the set of values into two halves: one above and one below the median.
- The interquartile range divides the set of values into four quarters.

The Interquartile Range (IQR)

- The interquartile range ignores the most extreme values and focuses on the middle of the data.
- Remember, the median divided the set of values into two halves: one above and one below the median.
- The interquartile range divides the set of values into four quarters.
- One quarter of the values in the set are below the lower quartile and one quarter of the values in the set are above the upper quartile.

Find the Interquartile Range (IQR)



IQR

lowest

highest

1 inch = 2.54 centimeters

① Find the median

$\frac{101+1}{2} = 51$

67 inches

② Find the median of the lower half

50 obs. $\frac{50+1}{2} = 25.5$

lower median of 66 inches

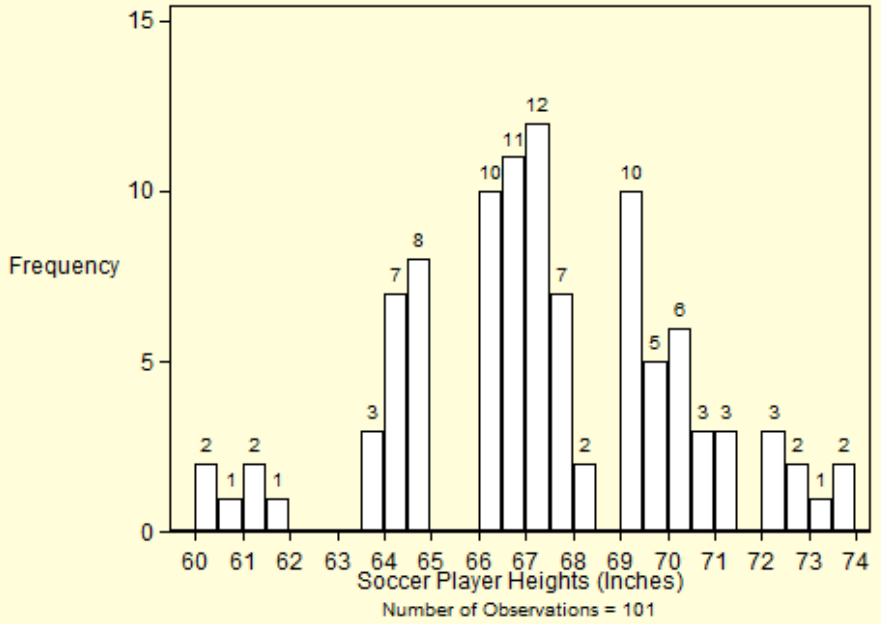
③ Find the median of the upper half → 69.5 and 70 inches → 70 − 66 = 4 in

IQR 69.5 − 66 = 3.5 in

How to Find the Interquartile Range (IQR)

How to Find the Interquartile Range (IQR)

- Find the median of all values in the set. 67
- Find the median of the lower half of values (Q1). 66
- Find the median of the upper half of values (Q3). 70 and 69.5



Why the Quartile Range?

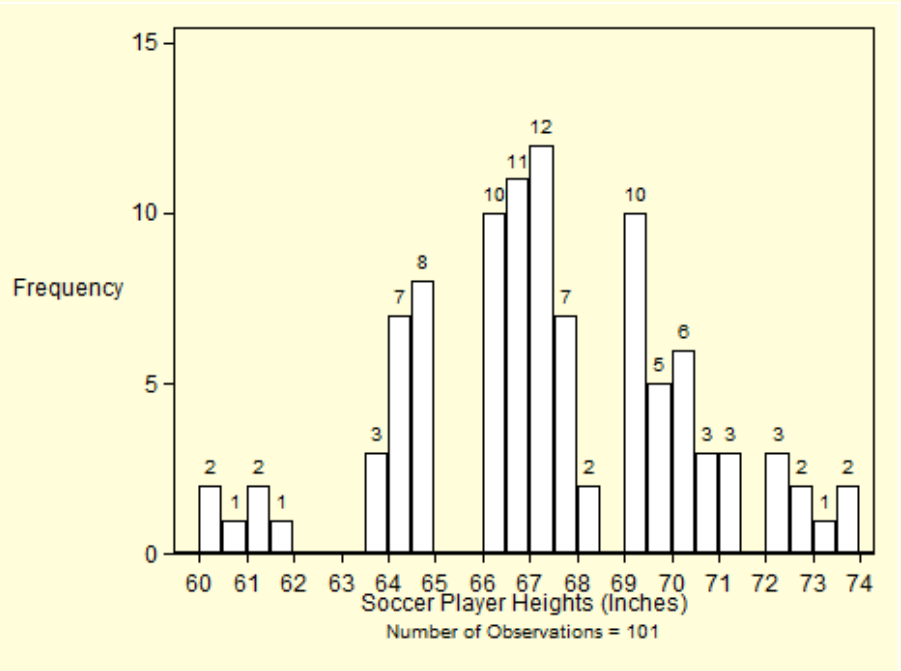
- Finding the IQR and the range give more information about the distribution of values than just the range.

Why the Quartile Range?

- Finding the IQR and the range give more information about the distribution of values than just the range.
- Sometimes, researchers study quintiles (5 equal parts) or deciles (10 equal parts).

Why the Quartile Range?

- Finding the IQR and the range give more information about the distribution of values than just the range.
- Sometimes, researchers study quintiles (5 equal parts) or deciles (10 equal parts).
- Software programs can easily find these for you.



The Standard Deviation is:

- A measure of *average variability* in a set of values.
- The average distance from the mean.

The Standard Deviation is:

- A measure of *average variability* in a set of values.
- The average distance from the mean.

If the standard deviation is larger, the set of values have more variability.

The Standard Deviation is:

- A measure of *average variability* in a set of values.
- The average distance from the mean.

If the standard deviation is larger, the set of values have more variability.

$$\sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

- σ is the standard deviation.

The Standard Deviation is:

- A measure of *average variability* in a set of values.
- The average distance from the mean.

If the standard deviation is larger, the set of values have more variability.

$$\sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

- σ is the standard deviation.
- Σ is sigma, the command to sum what follows.

The Standard Deviation is:

- A measure of *average variability* in a set of values.
- The average distance from the mean.

If the standard deviation is larger, the set of values have more variability.

$$\sigma = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n}}$$

- σ is the standard deviation.
- Σ is sigma, the command to sum what follows.
- X_i represents each individual value.

The Standard Deviation is:

- A measure of *average variability* in a set of values.
- The average distance from the mean.

If the standard deviation is larger, the set of values have more variability.

- σ is the standard deviation.
- Σ is sigma, the command to sum what follows.
- X_i represents each individual value.
- $X_i - \bar{X}$ is the difference between an individual value and the mean.

$$\sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

n = number of values in the series

Find the standard deviations:

$X = \{3, 3, 5, 5\}$

$\bar{X} = (3+3+5+5)/4 = 4$

$3-4 = -1$, $3-4 = -1$, $5-4 = 1$, $5-4 = 1$

$(-1)^2 = 1$, $(-1)^2 = 1$, $1^2 = 1$, $1^2 = 1$

$\frac{4}{4} = \frac{4}{4} = 1$

$\sqrt{1} = 1 = \sigma_X$

$$\sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

- σ is the standard deviation.
- Σ is sigma, the command to sum what follows.
- X_i represents each individual value.
- $X_i - \bar{X}$ is the difference between an individual value and the mean.

$Y = \{1, 1, 7, 7\}$

$\bar{Y} = 4$

$(1-4)^2 = 9$, $(1-4)^2 = 9$, $(7-4)^2 = 9$, $(7-4)^2 = 9$

$\frac{9+9+9+9}{4} = \frac{36}{4}$

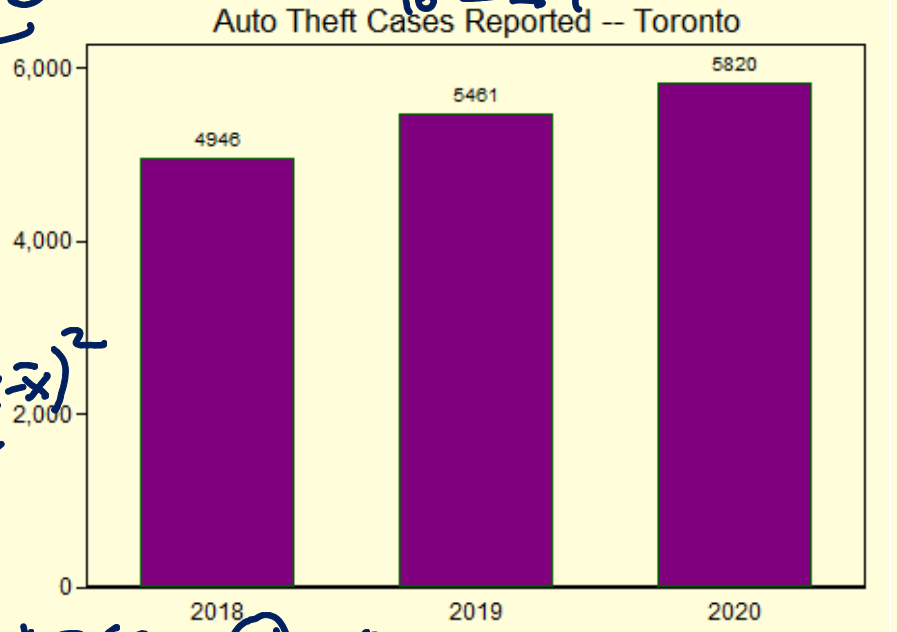
$\sqrt{\frac{\Sigma (x-\bar{x})^2}{n}} = \sqrt{\frac{36}{4}} = \sqrt{9} = 3 = \sigma_Y$

$Z = \{4, 4, 4, 4\}$

$\sigma_Z = 0$

How to Find the Standard Deviation

1. Compute the mean
2. Subtract the mean from each score
3. Square each individual difference.
4. Sum all the squared deviations.
5. Divide the sum by *n*.
6. Take the square root of the sum divided by *n*.



$\bar{x} = \frac{4946 + 5461 + 5820}{3} = \frac{16227}{3}$

$\bar{x} = 5409$

$4946 - 5409 = -463$

$5461 - 5409 = 52$

$5820 - 5409 = 411$

$(-463)^2 = 214,369$

$(52)^2 = 2,704$

$(411)^2 = 168,921$

$385,994$

$\frac{385,994}{3} = 128,664.67$

$\sqrt{128,664.67} = 359$

$\bar{x} = 1,223.3$

The Standard Deviation

Calgary to:	Distance	Deviation	Squared
Winnipeg	1327 km	103.7	10746.8
Bismarck	1286 km	62.7	3927.1
Vancouver	1057 km	−166.3	27666.8

The Standard Deviation

Calgary to:	Kilometers	Deviation	Squared
Winnipeg	1327	103.7	10746.8
Bismarck	1286	62.7	3927.1
Vancouver	1057	-166.3	27666.8
	Mean	1223.3	
	Sum		42340.7
	Sum ÷ 3		14113.6
	Squared-Root		118.8

$x - \bar{x}$ $(x-\bar{x})^2$ $\bar{x}$ $\Sigma(x-\bar{x})^2$ → variance standard deviation

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

The Standard Deviation

Calgary to:	Kilometers	Deviation	Squared
Winnipeg	1327	103.7	10746.8
Bismarck	1286	62.7	3927.1
Vancouver	1057	-166.3	27666.8
	Mean	1223.3	
	Sum		42340.7
	Sum $\div$ 3		14113.6
	Squared-Root		118.8

Why square each value?

The Standard Deviation

Calgary to:	Kilometers	Deviation	Squared
Winnipeg	1327	103.7	10746.8
Bismarck	1286	62.7	3927.1
Vancouver	1057	-166.3	27666.8
	Mean	1223.3	
	Sum		42340.7
	Sum $\div$ 3		14113.6
	Squared-Root		118.8

$\Sigma = 0$

Why square each value?

- Squaring each value removes negatives.

The Standard Deviation

Calgary to:	Kilometers	Deviation	Squared
Winnipeg	1327	103.7	10746.8
Bismarck	1286	62.7	3927.1
Vancouver	1057	-166.3	27666.8
	Mean	1223.3	
	Sum		42340.7
	Sum $\div$ 3		14113.6
	Squared-Root		118.8

Why square each value?

- Squaring each value removes negatives.
- If we simply added up the deviations from the mean, the sum would equal zero.

The Standard Deviation

Calgary to:	Kilometers	Deviation	Squared
Winnipeg	1327	103.7	10746.8
Bismarck	1286	62.7	3927.1
Vancouver	1057	-166.3	27666.8
	Mean	1223.3	
	Sum		42340.7
	Sum $\div$ 3		14113.6
	Squared-Root		118.8

Why square each value?

- Squaring each value removes negatives.
- If we simply added up the deviations from the mean, the sum would equal zero.

Why take the square root of the sum of squared-values?

The Standard Deviation

Calgary to:	Kilometers	Deviation	Squared
Winnipeg	1327	103.7	10746.8
Bismarck	1286	62.7	3927.1
Vancouver	1057	-166.3	27666.8
	Mean	1223.3	
	Sum		42340.7
	Sum ÷ 3		14113.6
	Squared-Root		118.8

in squared-kilometers

standard deviation is in kilometers

Why square each value?

- Squaring each value removes negatives.
- If we simply added up the deviations from the mean, the sum would equal zero.

Why take the square root of the sum of squared-values?

- The square root returns the values into the same units from which they started.

Statistical Bias

$$\sigma = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n}}$$

- σ is useful for summarizing information about a population.

Statistical Bias

$$\sigma = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n}}$$

- σ is useful for summarizing information about a population.
- However if you want to make a statistical inference about a population using a sample, σ is a **statistically biased** estimate of the variability in a population.

Statistical Bias

$$\sigma = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n}}$$

- σ is useful for summarizing information about a population.
- However if you want to make a statistical inference about a population using a sample, σ is a **statistically biased** estimate of the variability in a population.
- σ *underestimates* the true variability in a population.

Statistical Bias

$$\sigma = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n}}$$

- σ is useful for summarizing information about a population.
- However if you want to make a statistical inference about a population using a sample, σ is a **statistically biased** estimate of the variability in a population.
- σ *underestimates* the true variability in a population.

- For making statistical inferences, the **sample standard deviation**, s, is unbiased.

Statistical Bias

$$\sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

sigma

- σ is useful for summarizing information about a population.
- However if you want to make a statistical inference about a population using a sample, σ is a **statistically biased** estimate of the variability in a population.
- σ *underestimates* the true variability in a population.

under-estimate the true standard deviation

- For making statistical inferences, the **sample standard deviation**, s, is unbiased.

$$s = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}}$$

- where the denominator is now $n - 1$.

larger

$n-1$ is smaller its in denominator so $S > \sigma$

Statistical Bias

$$\sigma = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n}}$$

- σ is useful for summarizing information about a population.
- However if you want to make a statistical inference about a population using a sample, σ is a **statistically biased** estimate of the variability in a population.
- σ *underestimates* the true variability in a population.

- For making statistical inferences, the **sample standard deviation**, s, is unbiased.

$$s = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n-1}}$$

- where the denominator is now $n-1$.
- Dividing by a smaller value, $n-1$, makes the standard deviation larger.

Statistical Bias

$$\sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

- σ is useful for summarizing information about a population.
- However if you want to make a statistical inference about a population using a sample, σ is a **statistically biased** estimate of the variability in a population.
- σ *underestimates* the true variability in a population.

- For making statistical inferences, the **sample standard deviation**, s, is unbiased.

$$s = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}}$$

- where the denominator is now $n-1$.
- Dividing by a smaller value, $n-1$, makes the standard deviation larger.
- Allowing for a larger standard deviation is a *conservative* estimate.

to err on the side of caution assume variability is higher than it might appear based on a small sample

Statistical Bias

$$s = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

$n = 10$

Sample Size	$\Sigma (X_i - \bar{X})^2$	σ	s	$s - \sigma$
10	600	$\sqrt{\frac{600}{10}}$	$\sqrt{\frac{600}{10-1}}$	

Statistical Bias

$$s = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

Sample Size	$\Sigma (X_i - \bar{X})^2$	σ	s	$s - \sigma$
10	600	7.75	8.16	0.4190

Statistical Bias

$$s = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

Sample Size	$\Sigma (X_i - \bar{X})^2$	σ	s	$s - \sigma$
10	600	7.75	8.16	0.4190
100	600	2.45	2.46	0.0123

Statistical Bias

$$s = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

Sample Size	$\Sigma (X_i - \bar{X})^2$	σ	s	$s - \sigma$
10	600	7.75	8.16	0.4190
100	600	2.45	2.46	0.0123
1000	600	0.77	0.77	0.0004

Statistical Bias

$$s = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n-1}} \qquad \sigma = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n}}$$

Sample Size	$\Sigma \left(X_i - \bar{X}\right)^2$	σ	s	$s - \sigma$
10	600	7.75	8.16	0.4190
100	600	2.45	2.46	0.0123
1000	600	0.77	0.77	0.0004

- All else the same, as the sample size increases the difference between the standard deviation and the sample standard deviation becomes smaller.

The Standard Deviation

- If the standard deviation is larger, the values are more spread out.

The Standard Deviation

- If the standard deviation is larger, the values are more spread out.
- If the standard deviation is larger, the values are more different from each other.

The Standard Deviation

- If the standard deviation is larger, the values are more spread out.
- If the standard deviation is larger, the values are more different from each other.
- The standard deviation is sensitive to extremely large or small values.

The Standard Deviation

- If the standard deviation is larger, the values are more spread out.
- If the standard deviation is larger, the values are more different from each other.
- The standard deviation is sensitive to extremely large or small values.
- An extremely large or small value is one that is far away from the mean!

The Variance

- The variance is closely related to the standard deviation.
- The variance is equal to the standard deviation squared.

σ^2 = variance

The Variance

- The variance is closely related to the standard deviation.
- The variance is equal to the standard deviation squared.

The Variance is $\sigma^2 = \frac{\Sigma (X_i - \bar{X})^2}{n}$

$$\sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n}}$$

$$\sigma = \sqrt{\sigma^2}$$

The Variance

- The variance is closely related to the standard deviation.
- The variance is equal to the standard deviation squared.

The Variance is $\sigma^2 = \frac{\Sigma (X_i - \bar{X})^2}{n}$

A Previous Example:

Calgary to:	Kilometers	Deviation	Squared
Winnipeg	1327	103.7	10746.8
Bismarck	1286	62.7	3927.1
Vancouver	1057	-166.3	27666.8
	Mean	1223.3	
	Sum		42340.7
	Sum ÷ 3		14113.6
	Squared-Root		118.8

Variance σ^2

Std. dev. σ

The Variance

- Interpreting the variance can be more difficult than interpreting the standard deviation which is stated in the same units as the series.

The Variance

- Interpreting the variance can be more difficult than interpreting the standard deviation which is stated in the same units as the series.
- So,
  - If the values are stated in kilometers, the standard deviation is also stated in kilometers.

The Variance

- Interpreting the variance can be more difficult than interpreting the standard deviation which is stated in the same units as the series.
- So,
  - If the values are stated in kilometers, the standard deviation is also stated in kilometers.
  - If the values are stated in dollars, the standard deviation is also stated in dollars.

The Variance

- Interpreting the variance can be more difficult than interpreting the standard deviation which is stated in the same units as the series.
- So,
  - If the values are stated in kilometers, the standard deviation is also stated in kilometers.
  - If the values are stated in dollars, the standard deviation is also stated in dollars.
- However, variance does have useful applications in finance, economics, and statistics for the social sciences.

Data as a Distribution: Probability Density Function (PDF)

- Data can be represented according to frequency, a **frequency distribution**.

Data as a Distribution: Probability Density Function (PDF)

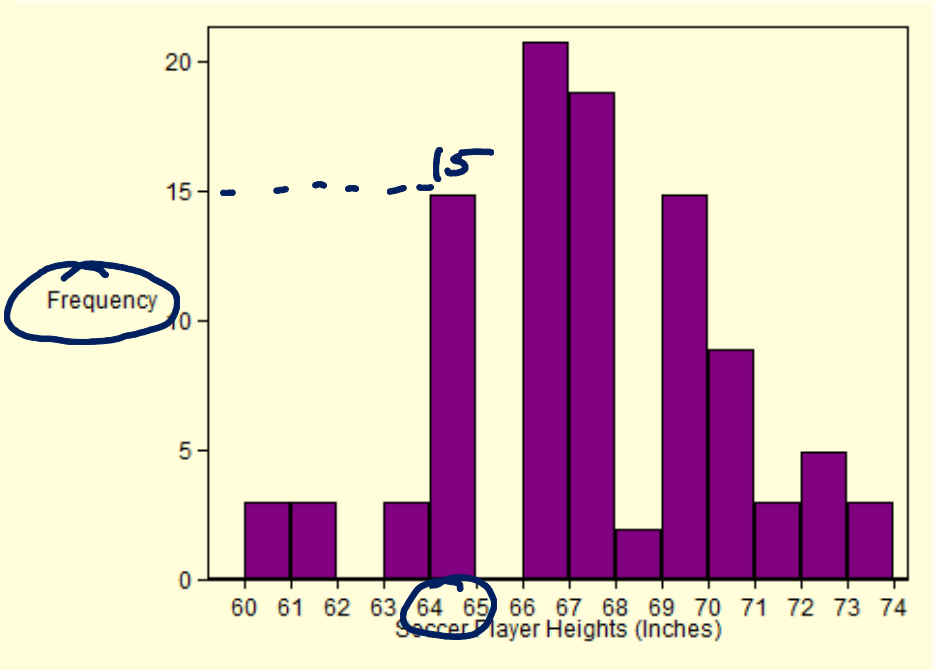
- Data can be represented according to frequency, a **frequency distribution**.
- If the data is continuous, values can be grouped into **class intervals**.

Data as a Distribution: Probability Density Function (PDF)

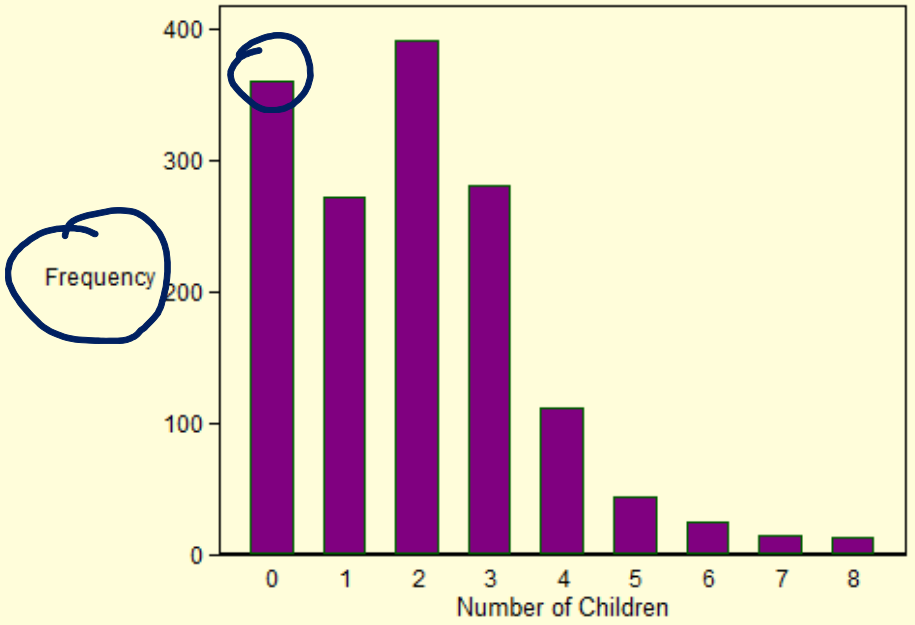
- Data can be represented according to frequency, a **frequency distribution**.
- If the data is continuous, values can be grouped into **class intervals**.
- A **class interval** is a range of values (numbers).

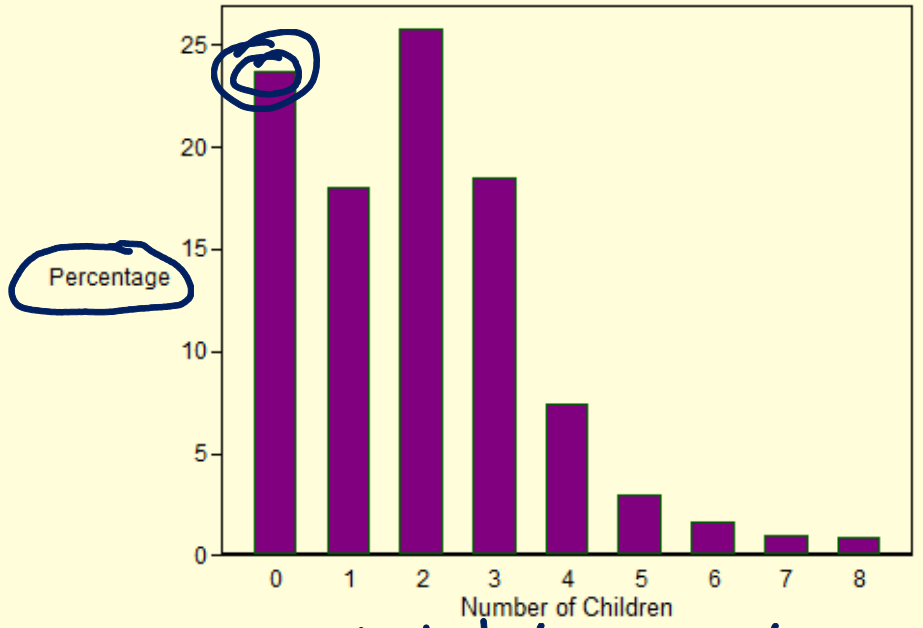
Data as a Distribution: Probability Density Function (PDF)

- Data can be represented according to frequency, a **frequency distribution**.
- If the data is continuous, values can be grouped into **class intervals**.
- A **class interval** is a range of values (numbers).
- Here, the class interval for soccer player heights is one-inch, for a total of 14 class intervals: 60-61, 61-62, ..., 73-74.



Data as a Distribution: Probability Density Function (PDF)

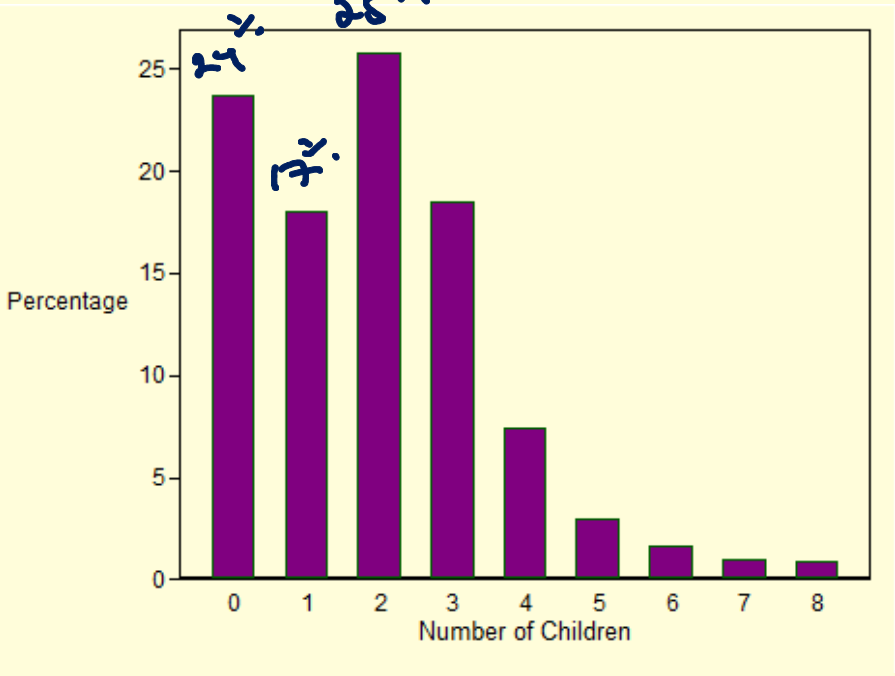




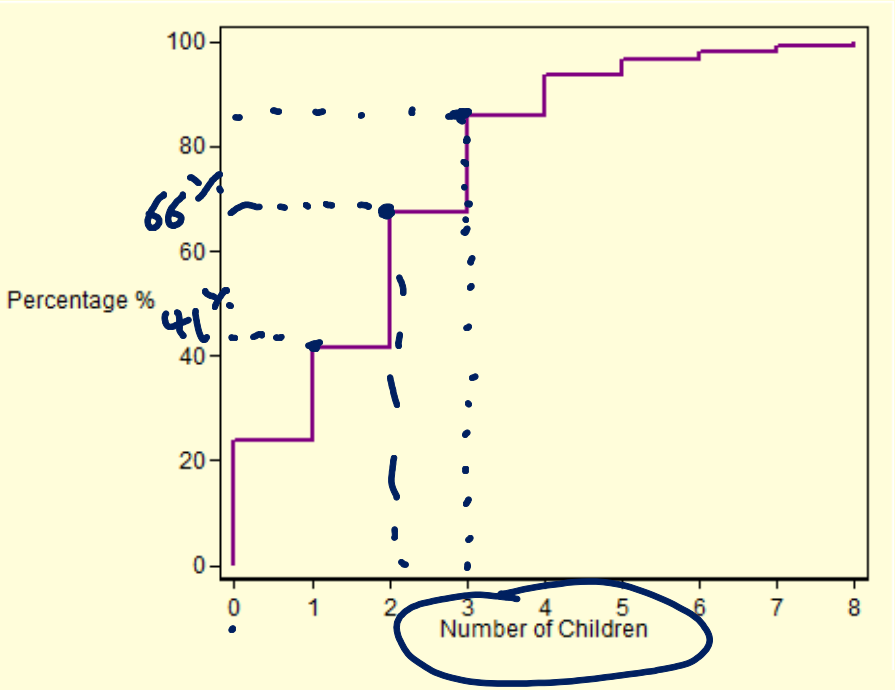
Probability Density Function

Data as a Distribution: Cumulative Density Function (CDF)

PDF

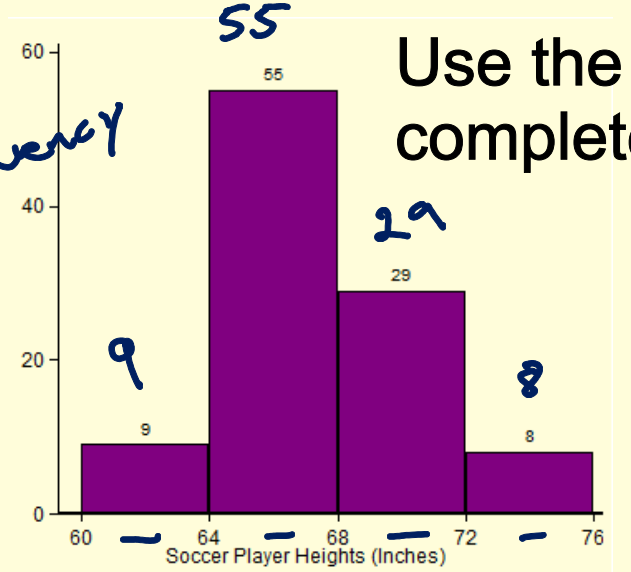


CDF



Cumulative Frequency & Cumulative Probability

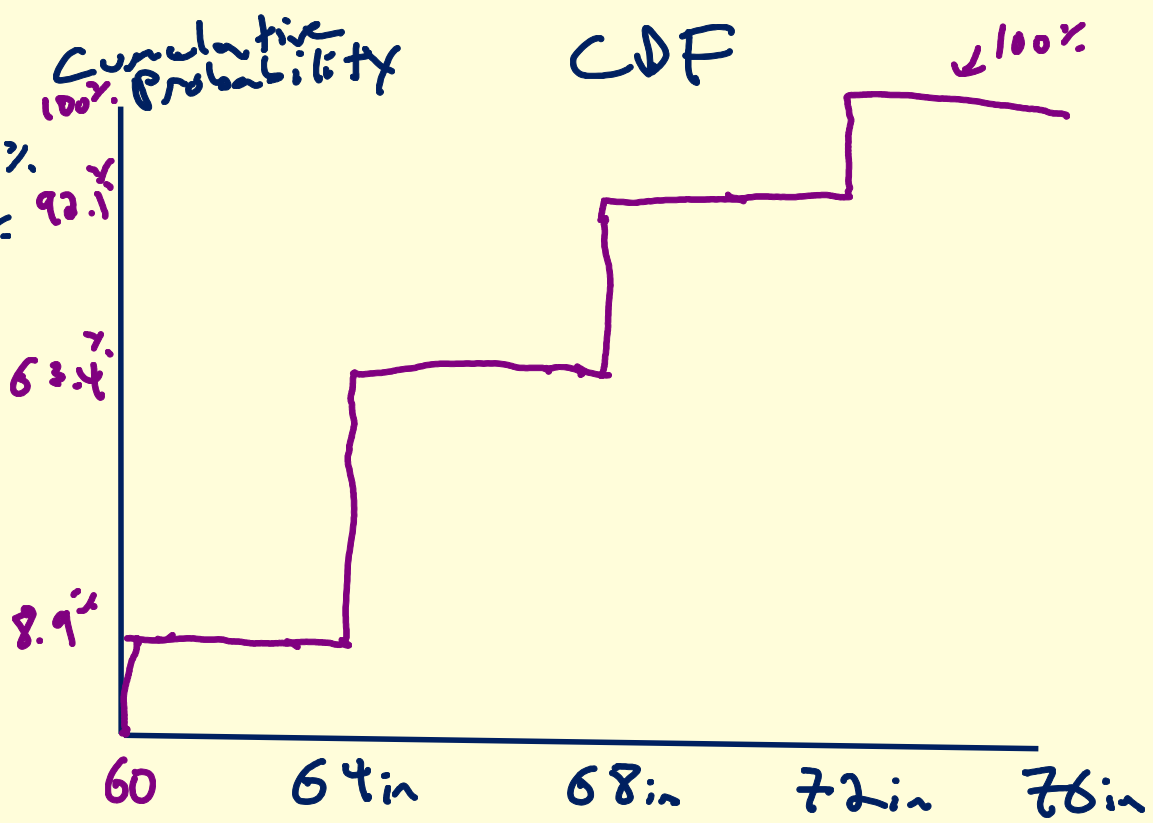
Range	Frequency	Cumulative Frequency	Probability (Fraction)	Cumulative Probability
60 - 64	9	9	9/101 = 8.9%	8.9%
64 - 68	55	55 + 9 = 64	55/101 = 54.5%	8.9% + 54.5% = 63.4%
68 - 72	29	64 + 29 = 93	29/101 = 28.7%	63.4% + 28.7% = 92.1%
72 - 76	8	93 + 8 = 101	8/101 = 7.9%	92.1% + 7.9% = 100%



Use the frequency histogram to complete the table.

9 + 55 + 29 + 8 = 101

60-64	8.9%
64-68	8.9 + 54.5 = 63.4%
68-72	63.4% + 28.7% = 92.1%
72-76	92.1% + 7.9% = 100%



CDF of a Soccer Player Heights With 0.5 Inch Intervals

more intervals => more steps

When working with continuous numeric data,

1. Sort the data from lowest to highest.
2. Divide the Data (values) into bins.
3. Graph the result.