

# Introduction to statistics

INTRODUCTION TO STATISTICS IN SPREADSHEETS

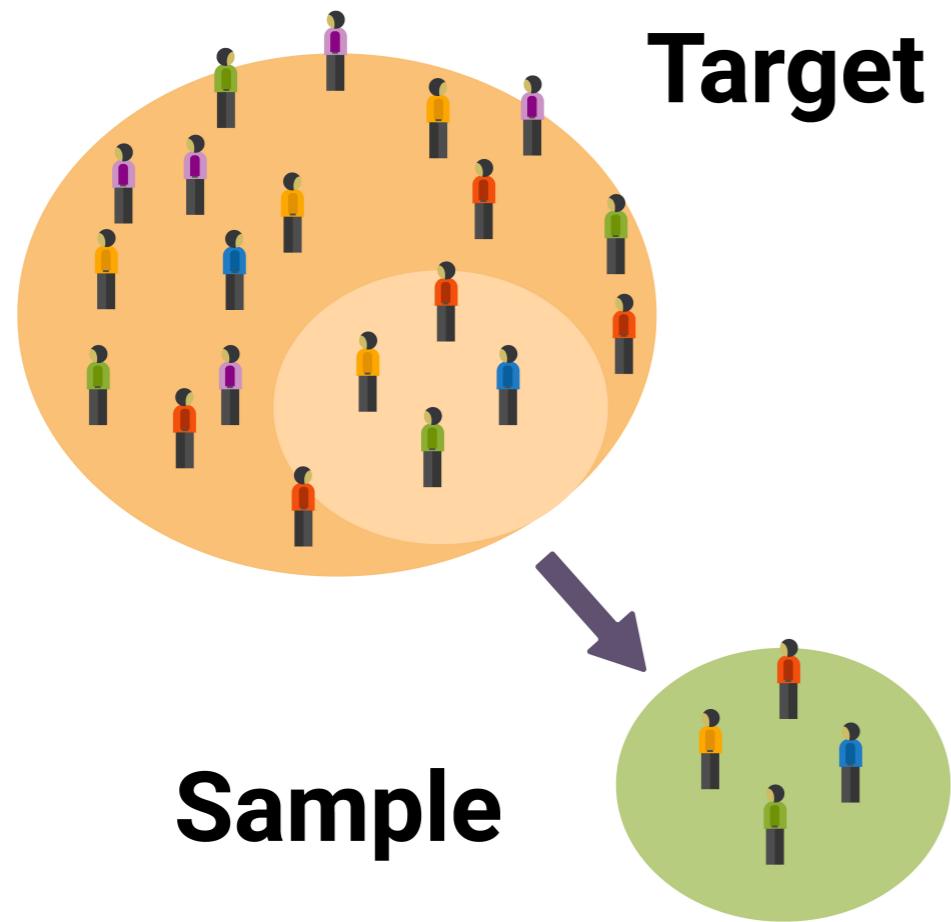


Ted Kwartler

Data Dude

# Sometimes it's ok to be mean (average)!

- Averages reduce information



# What it takes to be mean (average)?

- Mean = Sum of all observations / Number of observations

# Let's calculate an average

Person	Age
1	40
2	2
3	47
4	48
5	35
6	26
7	76
8	55
9	48
10	96

- Sum of all observations: 473
- Number of observations: 10
- Mean: 47.3

# Let's calculate another average

Person	Age
1	15
2	17
3	21
4	13
5	10
6	16
7	12
8	21
9	19
10	14

- Sum of all observations: 158
- Number of observations: 10
- Mean: 15.8

# Median averages

Person	Age
2	2
7	26
6	35
1	40
3	47
4	48
5	48
9	55
8	76
10	96

- Middle number of a dataset
- Half the numbers are less than the median
- Half the numbers are above the median

# Median averages

Person	Age
2	2
7	26
6	35
4	40
3	47
4	48
5	48
9	55
8	76
10	96

- Middle number of a dataset
- Half the numbers are less than the median
- Half the numbers are above the median

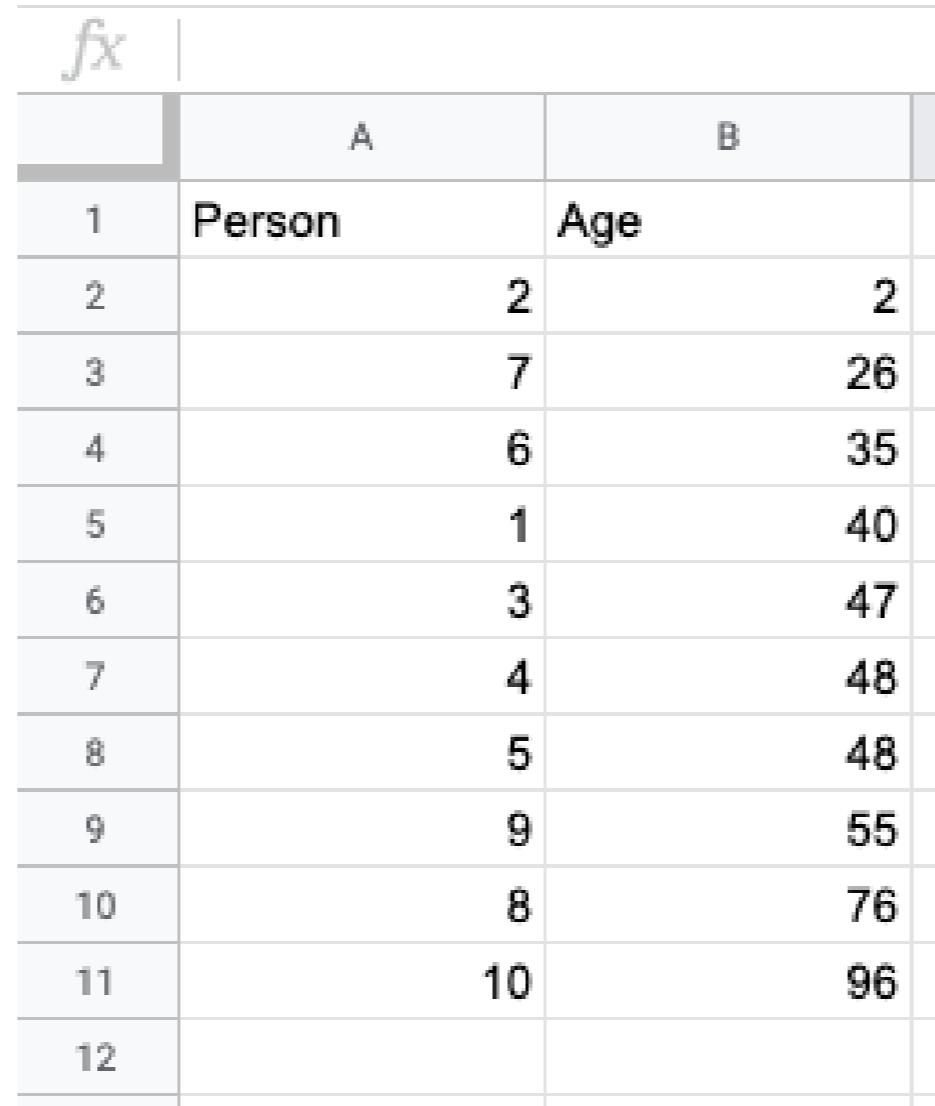
# Modal average

Person	Age
1	40
2	2
3	47
4	<b>48</b>
5	35
6	26
7	76
8	55
9	<b>48</b>
10	96

- Person 4 & 9 are both 48
- All other observations occur only once

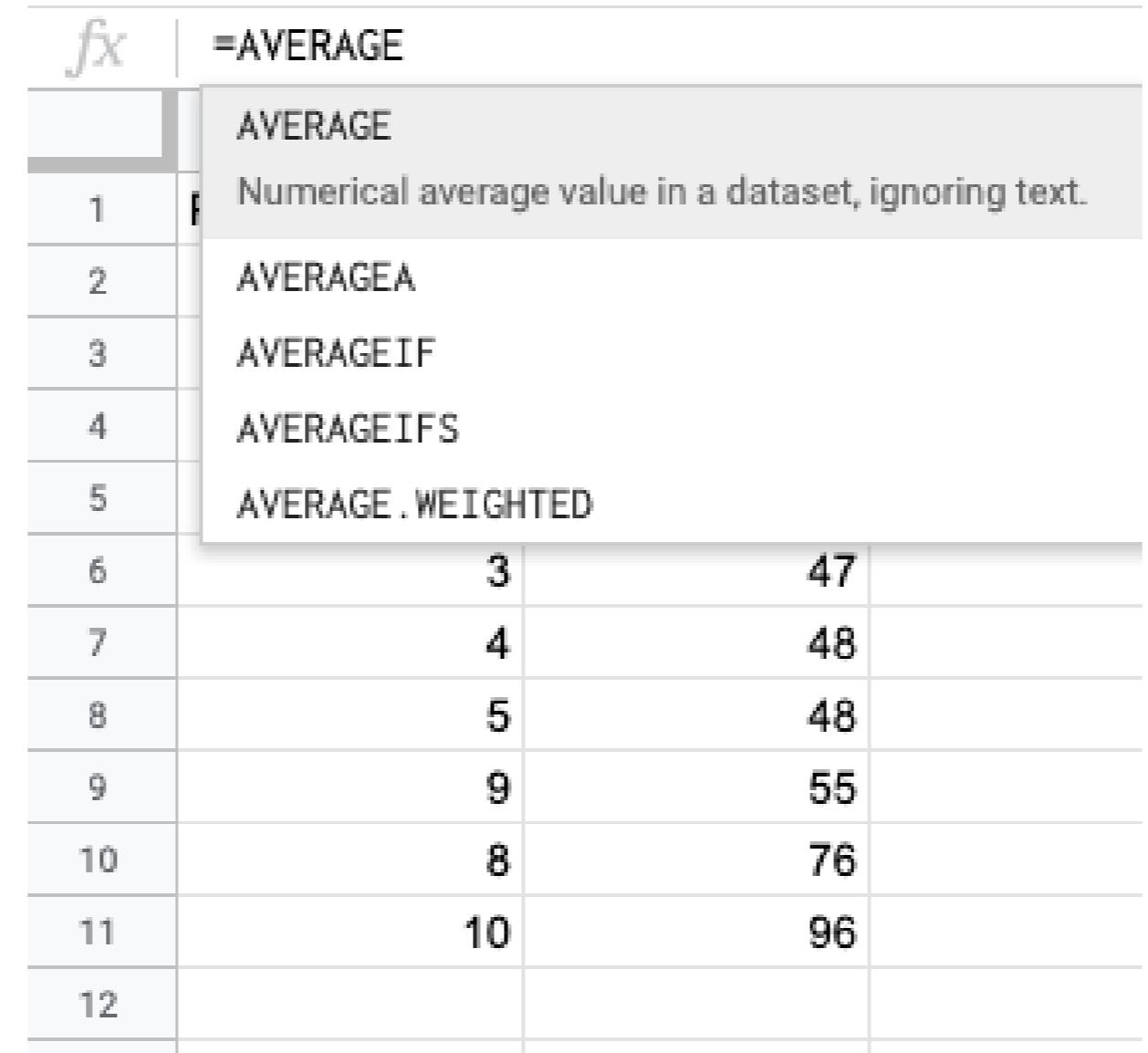
**Therefore the mode is 48**

# Mean average in Spreadsheets



	A	B
1	Person	Age
2	2	2
3	7	26
4	6	35
5	1	40
6	3	47
7	4	48
8	5	48
9	9	55
10	8	76
11	10	96
12		

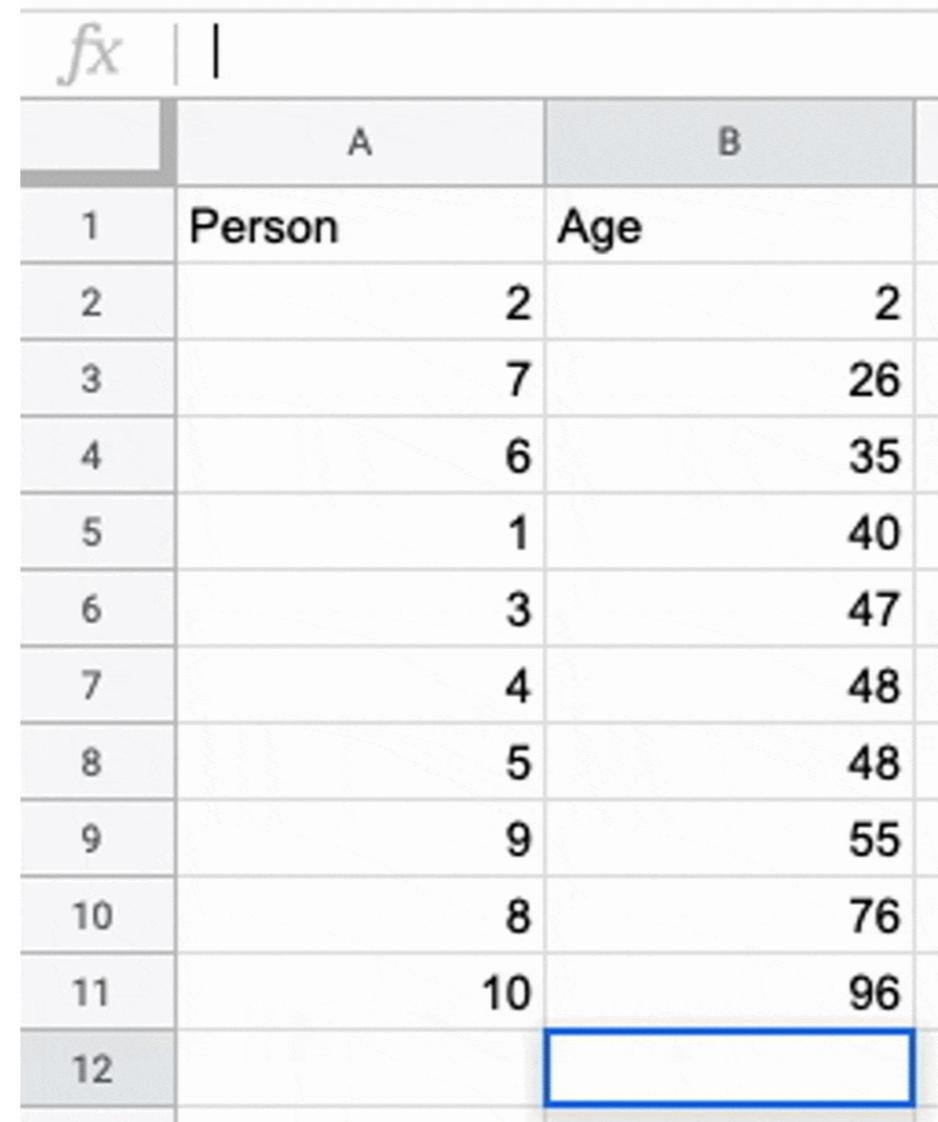
# Mean average in Spreadsheets



A screenshot of a spreadsheet application showing the context menu for the AVERAGE function. The menu is titled '=AVERAGE' and includes options: AVERAGE, AVERAGEA, AVERAGEIF, AVERAGEIFS, and AVERAGE.WEIGHTED. The AVERAGE option is highlighted. Below the menu, there is a table with 12 rows and 3 columns. The first column contains row numbers from 1 to 12. The second column contains numerical values: 3, 4, 5, 9, 8, 10, and an empty cell for row 12. The third column contains numerical values: 47, 48, 48, 55, 76, 96, and an empty cell for row 12.

1		
2		
3		
4		
5		
6	3	47
7	4	48
8	5	48
9	9	55
10	8	76
11	10	96
12		

# Mean average in Spreadsheets



	A	B
1	Person	Age
2	2	2
3	7	26
4	6	35
5	1	40
6	3	47
7	4	48
8	5	48
9	9	55
10	8	76
11	10	96
12		

# Median and Mode in Spreadsheets

*fx* | =MEDIAN(B2:B11)

	A	B
1	Person	Age
2	2	2
3	7	26
4	6	35
5	1	40
6	3	47
7	4	48
8	5	48
9	9	55
10	8	76
11	10	96
12		47.5

*fx* | =MODE(B2:B11)

	A	B
1	Person	Age
2	2	2
3	7	26
4	6	35
5	1	40
6	3	47
7	4	48
8	5	48
9	9	55
10	8	76
11	10	96
12		48

# **Let's practice some averages!**

**INTRODUCTION TO STATISTICS IN SPREADSHEETS**

# How far from average?

INTRODUCTION TO STATISTICS IN SPREADSHEETS



Ted Kwartler

Data Dude

# All aboard!



# First stop, Variance Station

- Measures how dispersed a dataset is
- Smaller variance indicates a dataset is less spread
- Large differences between data points increase the variance

# First stop, Variance Station

A
347
347
347
347

# First stop, Variance Station

A
347
347
347
347

- Column A Variance: 0

# First stop, Variance Station

A	B
347	10
347	<b>14</b>
347	10
347	10

- Column A Variance: 0

# First stop, Variance Station

A	B
347	10
347	14
347	10
347	10

- Column A Variance: **0**
- Column B Variance: **3**

# First stop, Variance Station

A	B	C
347	10	10
347	14	14
347	10	<b>100</b>
347	10	10

- Column A Variance: **0**
- Column B Variance: **3**

# First stop, Variance Station

A	B	C
347	10	10
347	14	14
347	10	100
347	10	10

- Column A Variance: **0**
- Column B Variance: **3**
- Column C Variance: **1476.75**

# Calculating Variance: Step 1

	A	B	C	D
1				
2			10	
3			14	
4			10	
5			10	
6	Mean:	11		

# Calculating Variance: Step 2

fx	A	B	C	D
1			Subtract Mean from Each Value	
2			10	-1
3			14	3
4			10	-1
5			10	-1
6	Mean:		11	

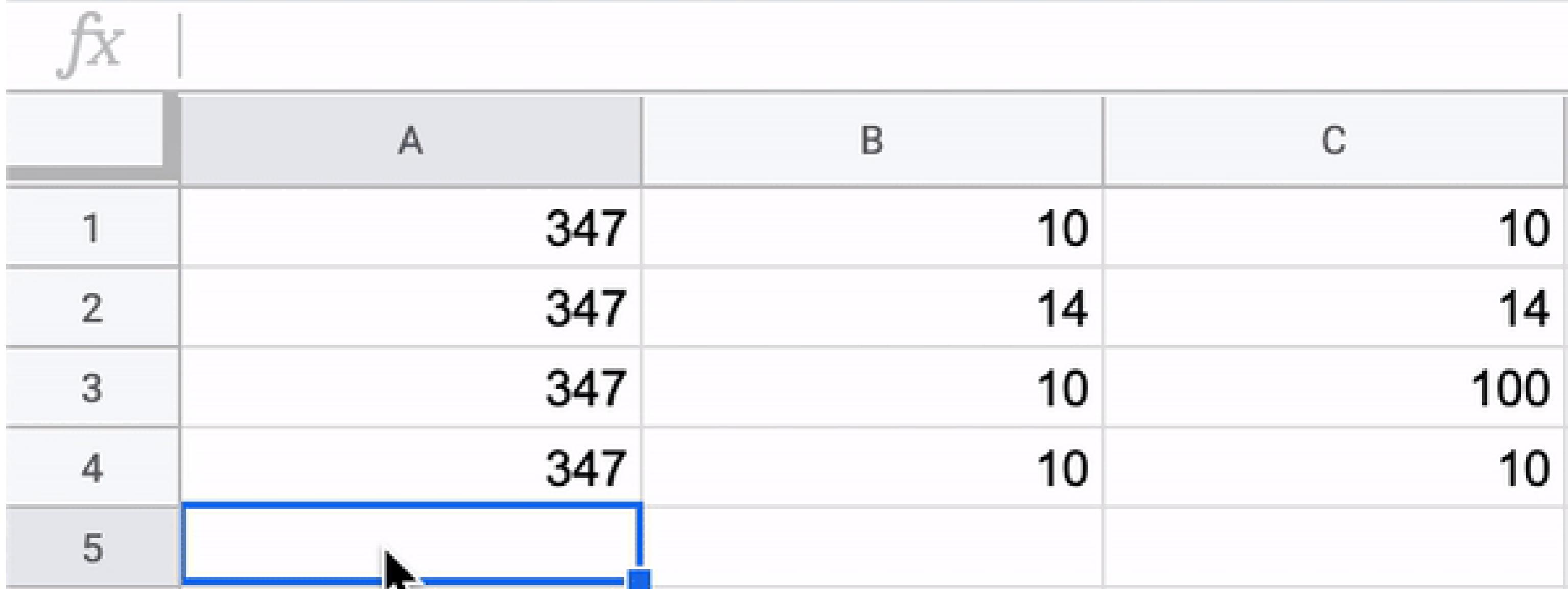
# Calculating Variance: Step 3

fx	A	B	C	D
1			Subtract Mean from Each Value	Square the Differences
2		10	-1	1
3		14	3	9
4		10	-1	1
5		10	-1	1
6	Mean:	11		

# Calculating Variance: Step 4

	A	B	C	D
1			Subtract Mean from Each Value	Square the Differences
2		10	-1	1
3		14	3	9
4		10	-1	1
5		10	-1	1
6	Mean:	11		3

# Calculating variance in spreadsheets



A screenshot of a spreadsheet application showing a table of data. The table has columns labeled A, B, and C, and rows labeled 1 through 5. The data in column A is all 347. The data in row 1 is 347, 10, and 10. The data in row 2 is 347, 14, and 14. The data in row 3 is 347, 10, and 100. The data in row 4 is 347, 10, and 10. In row 5, the cell in column A is selected, and a blue selection handle is visible at the bottom right corner of the cell. The formula bar at the top shows the text "fx".

	A	B	C
1	347	10	10
2	347	14	14
3	347	10	100
4	347	10	10
5			

# Pulling into, Standard Deviation!



# Standard Deviation

## Manual Standard Deviation:

	<i>fx</i>	=SQRT(B5)
	A	B
1		10
2		14
3		10
4		10
5	Variance	3
6	Standard Deviation	1.732050808

## Spreadsheets Standard Deviation:

	<i>fx</i>	=STDEVP(B1:B4)
	A	B
1		10
2		14
3		10
4		10
5	Variance	3
6	Standard Deviation	1.732050808

# Standard Deviation as a unit of measure

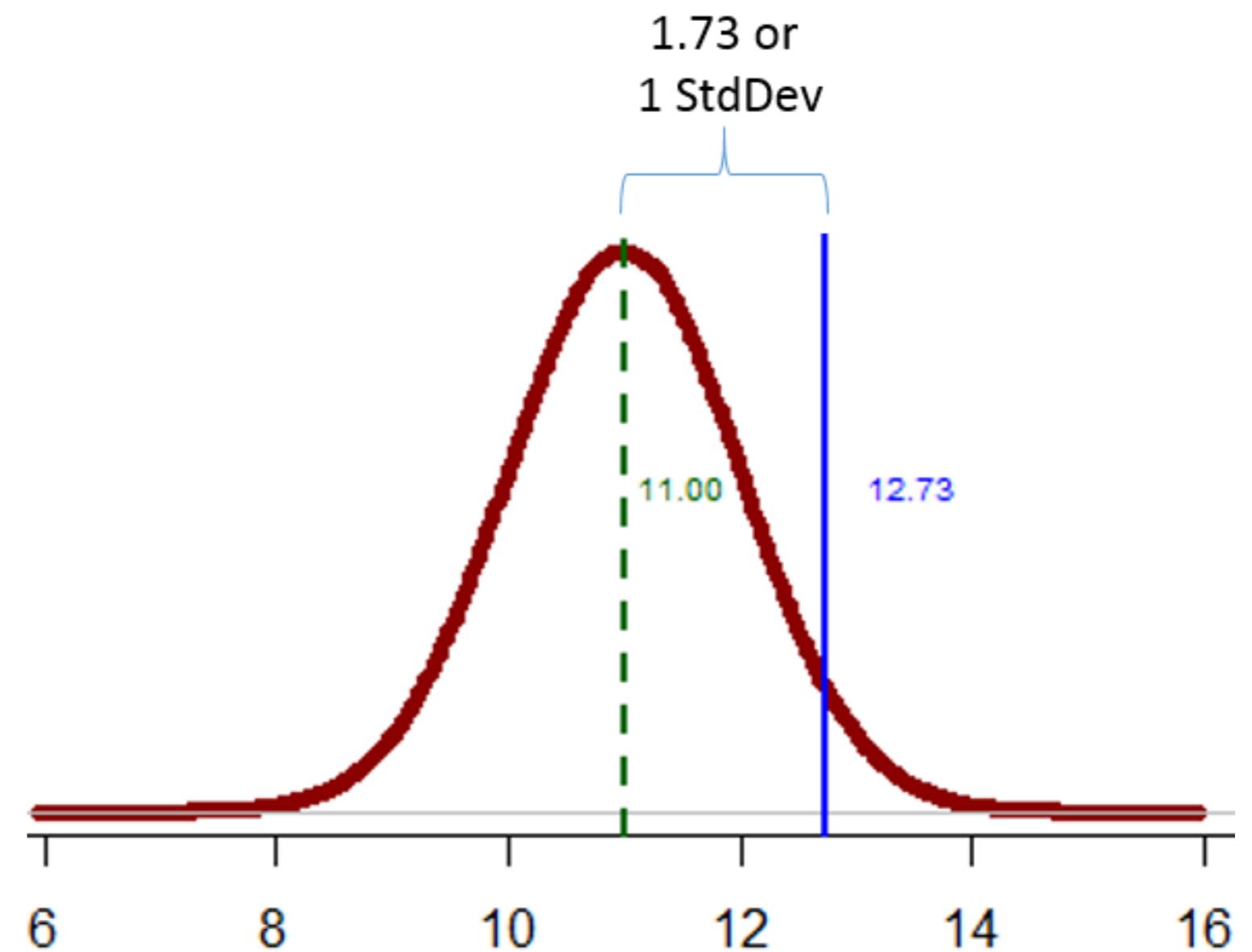
Sample average & standard deviation

=AVERAGE(10,14,10,10) = 11

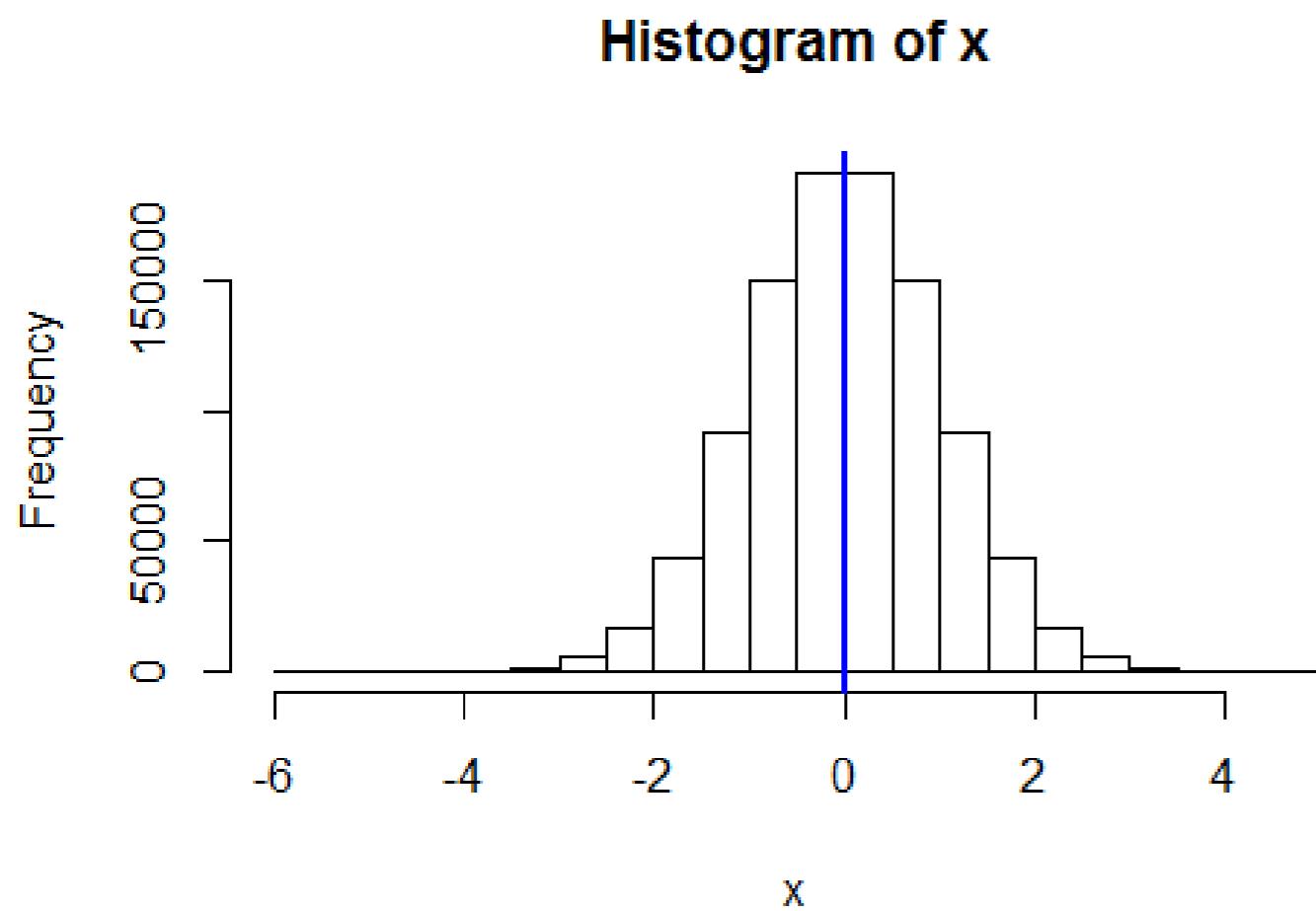
=STDEVP(10,14,10,10) = 1.73

New Data Point: 12.73

12.73 - 1.73 = 11

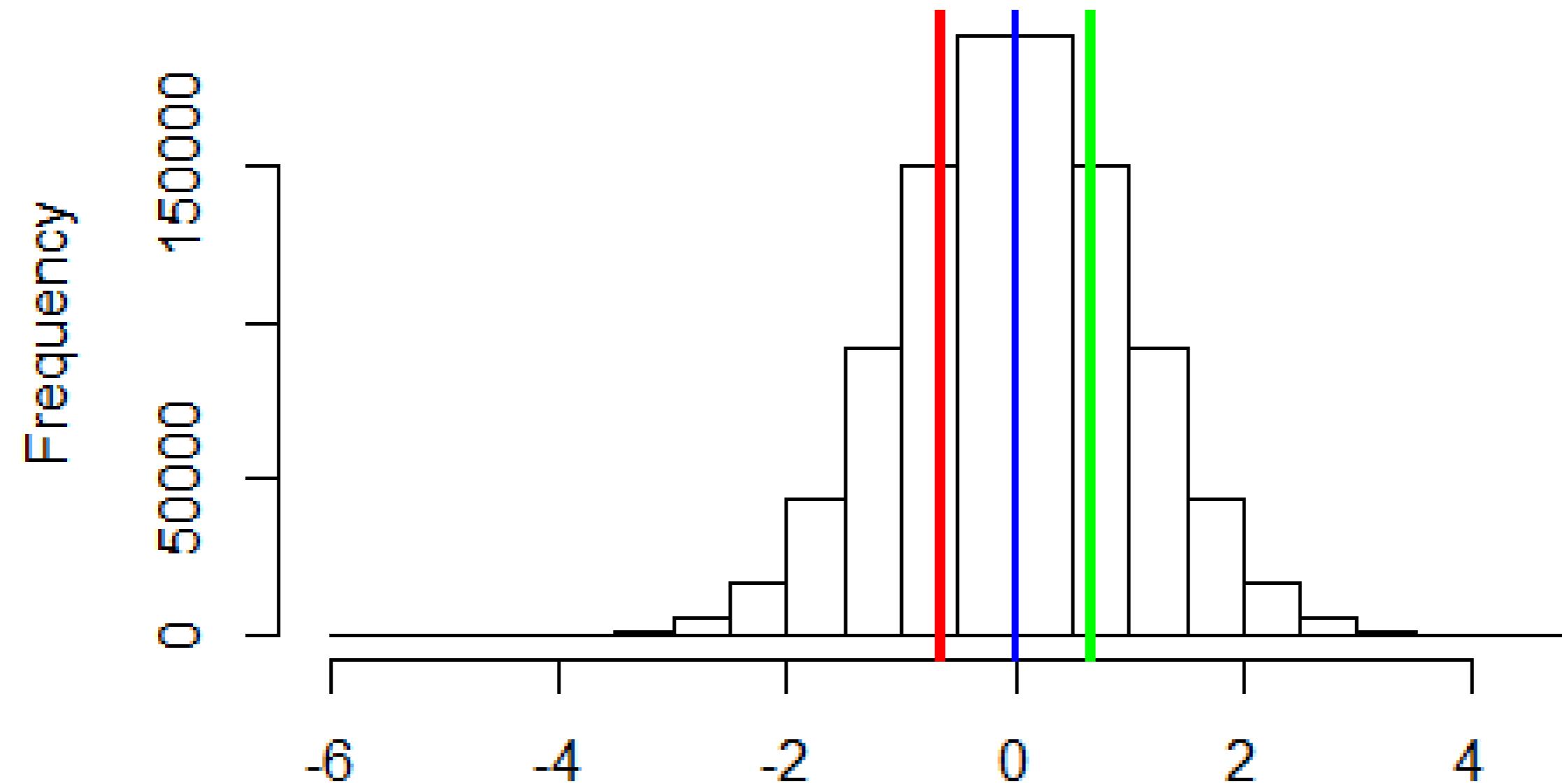


# Last distribution stop on the stats train: quartiles



- Percentile: % of values below a specific point
- 50th percentile: Splits the data evenly
- Here, 0 is the 50th percentile

## Histogram of x



*fx* =QUARTILE(B1:B5, 1)

	A	B
1		123
2		456
3		789
4		234
5		567
6	First Quartile	234

*fx* =QUARTILE(B1:B5, 2)

*fx* =QUARTILE(B1:B5, 2)

	A	B
1		123
2		456
3		789
4		234
5		567
6	Second Quartile	456

*fx* =QUARTILE(B1:B5, 3)

	A	B
1		123
2		456
3		789
4		234
5		567
6	Third Quartile	567

*fx* =QUARTILE(B1:B5, 4)

	A	B
1		123
2		456
3		789
4		234
5		567
6	Fourth Quartile	789

# **Let's practice!**

**INTRODUCTION TO STATISTICS IN SPREADSHEETS**

# Standardizing data

INTRODUCTION TO STATISTICS IN SPREADSHEETS



Ted Kwartler

Data Dude

# Why standardize?

- Variables are sometimes measured on different scales
- Makes it harder to compare
- Easier to misinterpret variable importance
- Solution: standardize your data
- All variables on the same scale

*fx*

	A	B	C	D	E
1		Data			
2			1		
3			2		
4			3		
5					
6					

*fx*

=AVERAGE(B2:B4)

	A	B	C	D	E
1		Data			
2			1		
3			2		
4			3		
5	Mean:	2			
6					

*fx*

=STDEVP(B2:B4)

	A	B	C	D	E
1		Data			
2			1		
3			2		
4			3		
5	Mean:		2		
6	Standard Deviation:	0.8164965809			

*fx*

	A	B	C	D	E
1	Data	Mean	Std. Deviation		
2		1	2	0.8164965809	
3		2	2	0.8164965809	
4		3	2	0.8164965809	
5	Mean:	2			
6	Standard Deviation:	0.8164965809			

*fx* | =(B2-C2)/D2

	A	B	C	D	E
1		Data	Mean	Std. Deviation	Z-Score
2			1	2	0.8164965809
3			2	2	0.8164965809
4			3	2	0.8164965809
5	Mean:		2		
6	Standard Deviation:	0.8164965809			

*fx* | =(B3-C3)/D3

	A	B	C	D	E
1		Data	Mean	Std. Deviation	Z-Score
2			1	2	0.8164965809 -1.224744871
3			2	2	0.8164965809 0
4			3	2	0.8164965809
5	Mean:		2		
6	Standard Deviation:	0.8164965809			

*fx* | =(B4-C4)/D4

	A	B	C	D	E
1		Data	Mean	Std. Deviation	Z-Score
2		1	2	0.8164965809	-1.224744871
3		2	2	0.8164965809	0
4		3	2	0.8164965809	1.224744871
5	Mean:	2			
6	Standard Deviation:	0.8164965809			

*fx* ? =STANDARDIZE()

	A	B	C	D	E	F
1		Data	Mean	Std. Deviation	Z-Score	
2			1	2	0.8164965809	=STANDARDIZE( )
3			2	2	0.8164965809	
4			3	2	0.8164965809	
5	Mean:		2			
6	Standard Deviation:	0.8164965809				

*fx* ? =STANDARDIZE(B2)

	A	B	C	D	E	F
1		Data	Mean	Std. Deviation	Z-Score	
2		1		0.8164965809	=STANDARDIZE(B2)	
3		2		0.8164965809		
4		3		0.8164965809		
5	Mean:	2				
6	Standard Deviation:	0.8164965809				

*fx*

? =STANDARDIZE(B2, C2)

	A	B	C	D	E	F
1		Data	Mean	Std. Deviation	Z-Score	
2		1	2	0.8164965809	=STANDARDIZE(B2, C2)	
3		2	2	0.8164965809		
4		3	2	0.8164965809		
5	Mean:	2				
6	Standard Deviation:	0.8164965809				

*fx* ? =STANDARDIZE(B2, C2, D2)

	A	B	C	D	E	F
1		Data	Mean	Std. Deviation	Z-Score	
2		1	2	0.8164965809	=STANDARDIZE(B2, C2, D2)	
3		2	2	0.8164965809		
4		3	2	0.8164965809		
5	Mean:	2				
6	Standard Deviation:	0.8164965809				

*fx*

=STANDARDIZE(B2, C2, D2)

	A	B	C	D	E	F
1		Data	Mean	Std. Deviation	Z-Score	
2		1	2	0.8164965809	-1.224744871	
3		2	2	0.8164965809		
4		3	2	0.8164965809		
5	Mean:	2				
6	Standard Deviation:	0.8164965809				

*fx*

	A	B	C	D	E	F
1		Data	Z-Score		Data	
2			1 -1.224744871			10
3			2 0			20
4			3 1.224744871			30
5	Mean:		2			
6	Standard Deviation:	0.8164965809				

*fx*

	A	B	C	D	E	F
1		Data	Z-Score		Data	
2			1 -1.224744871			10
3			2 0			20
4			3 1.224744871			30
5	Mean:		2			20
6	Standard Deviation:	0.8164965809			8.164965809	

*fx*

	A	B	C	D	E	F
1		Data	Z-Score		Data	Z-Score
2			1 -1.224744871			10 -1.224744871
3			2 0			20 0
4			3 1.224744871			30 1.224744871
5	Mean:		2			20
6	Standard Deviation:	0.8164965809			8.164965809	

# **Almost there, let's standardize!**

**INTRODUCTION TO STATISTICS IN SPREADSHEETS**