JAMES COOK
UNIVERSITY
AUSTRALIA

Basic Statistics

Describing Data – Measures of Spread

Describing Data

Learning Intentions

Today we will understand:

- Measures of Spread
  - * Calculate the range of a sample
  - * Determine quartiles and interquartile range
  - * Calculate variance
  - * Calculate standard deviation

Image accessed: http://intouchacquisitions.co.uk/in-touch-acquisitions-review-the-importance-of-business-statistics/

Describing Data

Two descriptions of data:

- Measures of Central Tendency
- Measures of Dispersion

Image accessed: http://www.fotosearch.com/illustration/two-people-looking-out-tower.html

Measures of Spread

- Also called measures of dispersion
- Describes variability in a sample or population
- Used in conjunction with a measure of central tendency to provide overall description of data

Image accessed: https://sites.google.com/a/clarkston.k12.mi.us/independence-third-grade/updates/math-curriculum

Range

- Simplest measure of spread
- Difference between the largest value and the smallest value of a dataset
- Range = maximum value – minimum value

23 46 33 12 44 31 29 15 47 37 22 34 35 41 36

Minimum Value

Maximum Value

Range = 47 – 12

= 35

Quartiles

- Ranked data – arranged into ascending order of magnitude
- Data can be divided into four groups – each with an equal number of data points

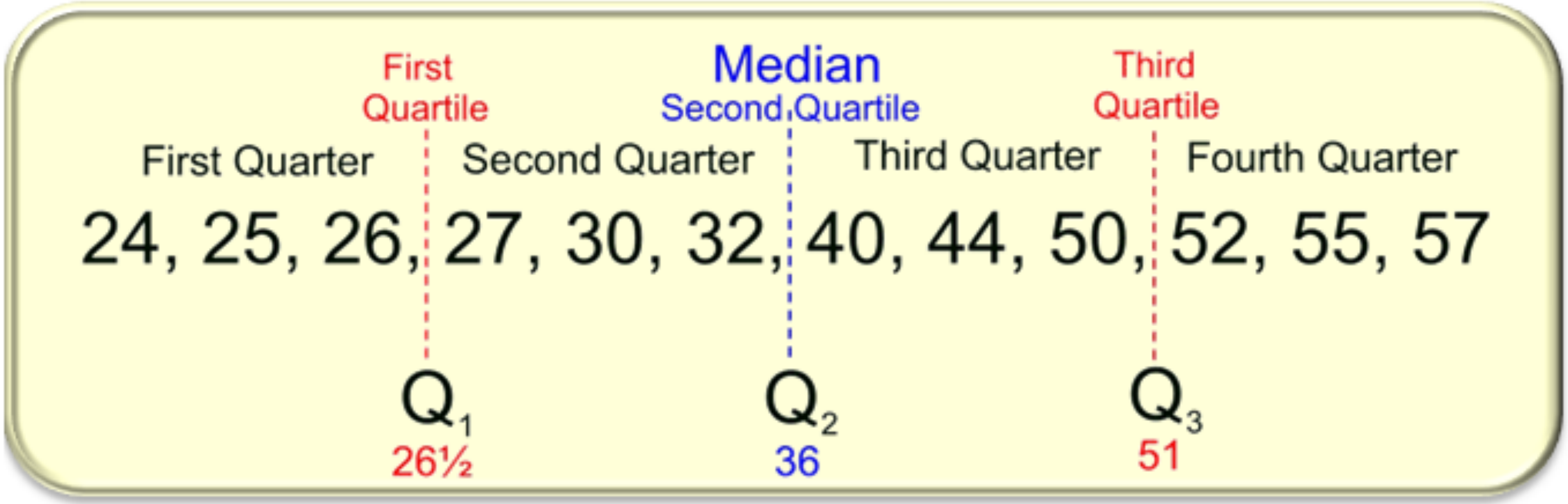


Image accessed: http://mathbitsnotebook.com/Algebra1/StatisticsData/STboxplot.html

Quartiles

Q_1: 1st Quartile/Lower Quartile	**Q_2: 2nd Quartile/Median**	**Q_3: 3rd Quartile/Upper Quartile**
• Median of the lower half of the data set • 25 % of data lies below • 75 % of data lies above	• Another name for the median • 50 % of data lies below • 50 % of data lies above	• Median of the upper half of the data set • 75 % of data lies below • 25 % of data lies above

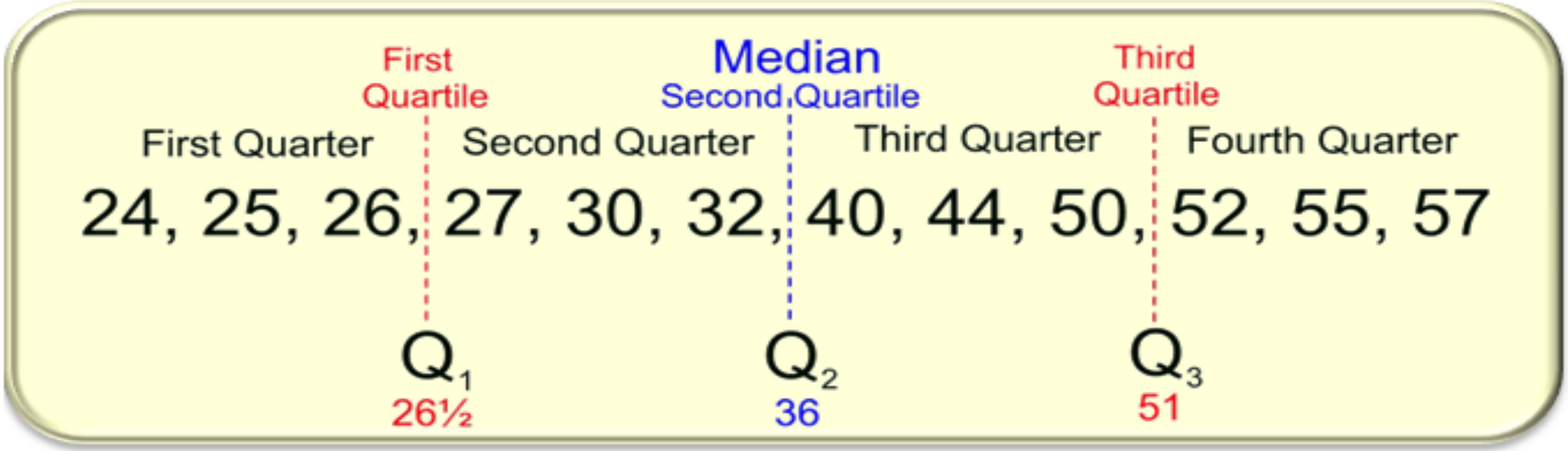
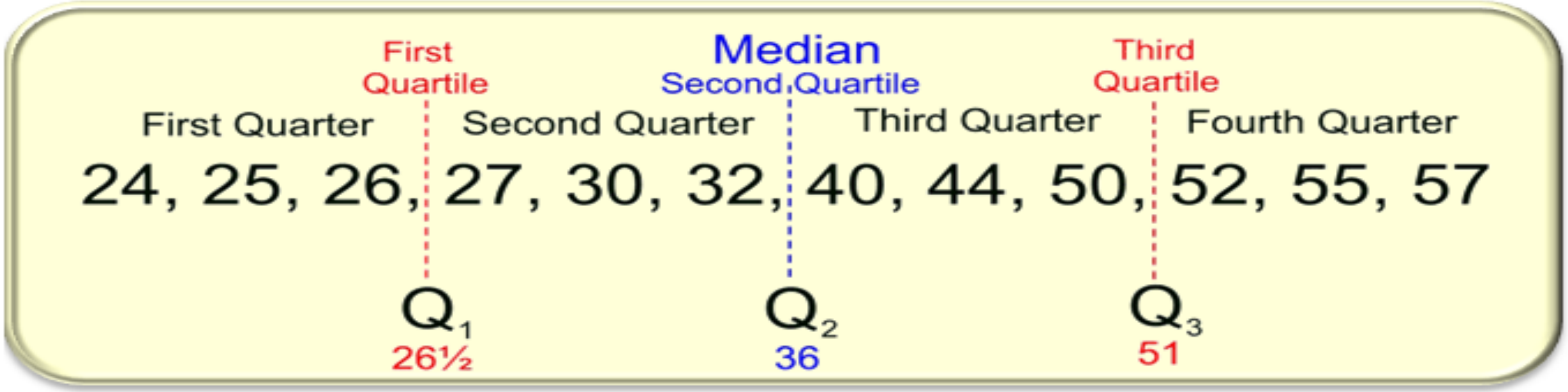


Image accessed: http://mathbitsnotebook.com/Algebra1/StatisticsData/STboxplot.html

Interquartile Range

- Difference between the third quartile, Q_3, and the first quartile, Q_1
- $IQR = Q_3 - Q_1$
- Range for the middle 50 % of data

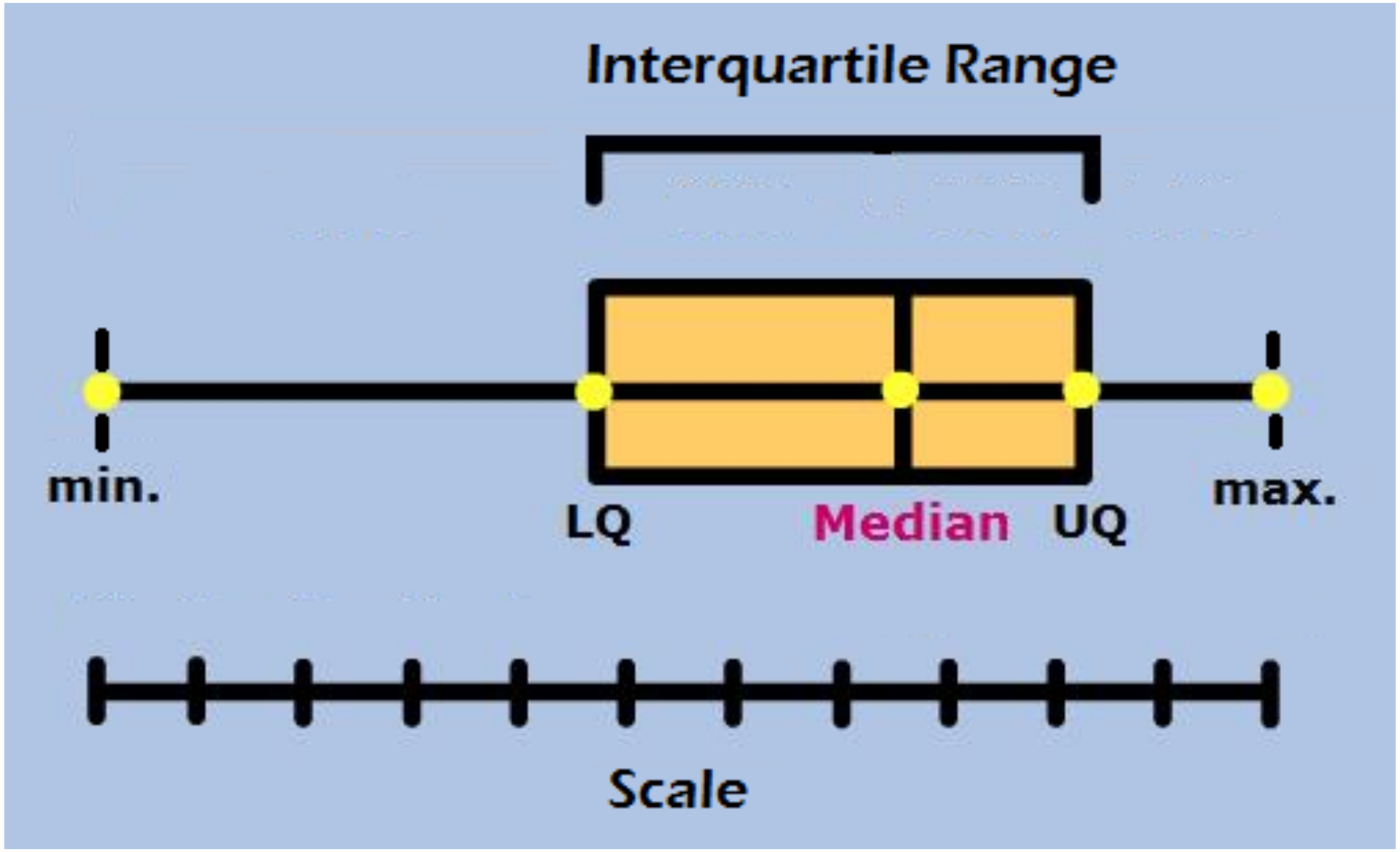


- IQR = 51 – 26.5

  = 24.5

Image accessed: http://mathbitsnotebook.com/Algebra1/StatisticsData/STboxplot.html

Quartiles and Interquartile Range



- Box and whisker plots are used to represent quartiles and the interquartile range

Variance

- Variance is a numerical value which indicates how 'spread out' a group of data points are

- Variance is derived from the difference between the value of each observation and the mean

- If individual observations vary greatly from the group mean, the variance is big; and vice versa

Image accessed: http://ramonlazatercerciclo.blogspot.com.au/

Population Variance

- If data is for a population

Remember:

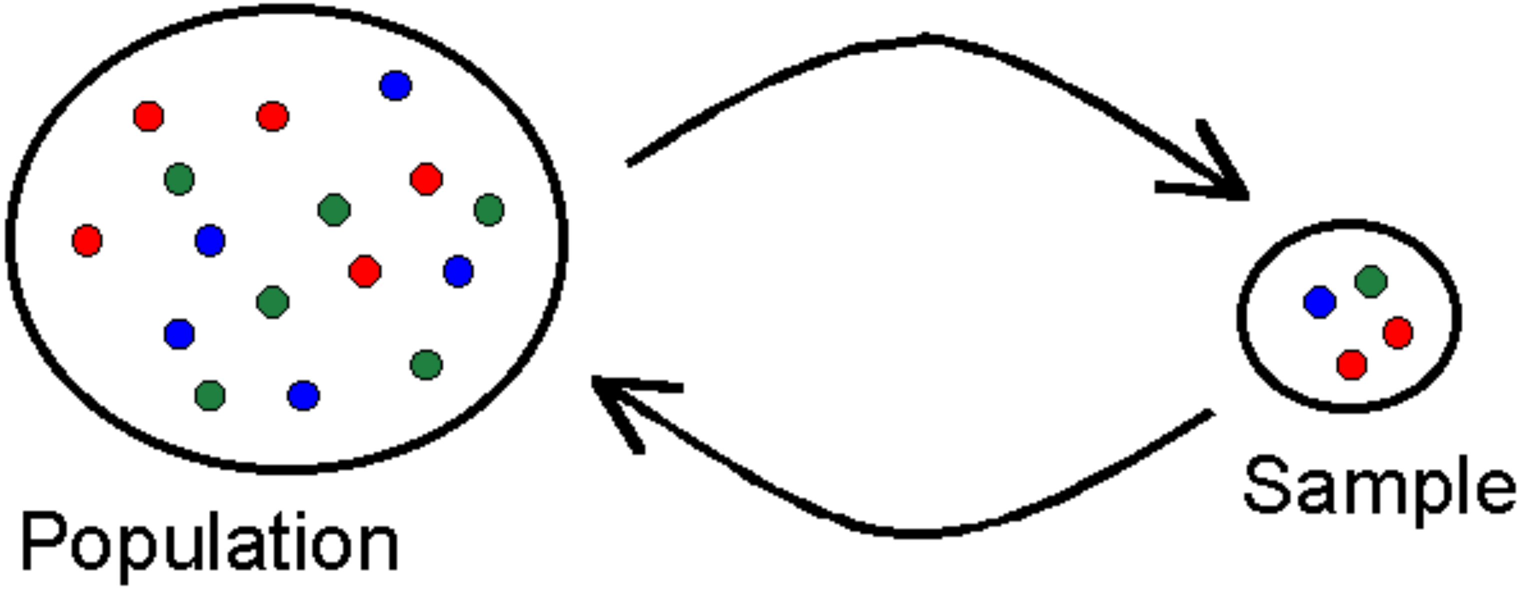


Image accessed: https://scholar.vt.edu/access/content/group/43c8db00-e78f-4dcd-826c-ac236fb59e24/STAT5605/normal01.htm

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Where:

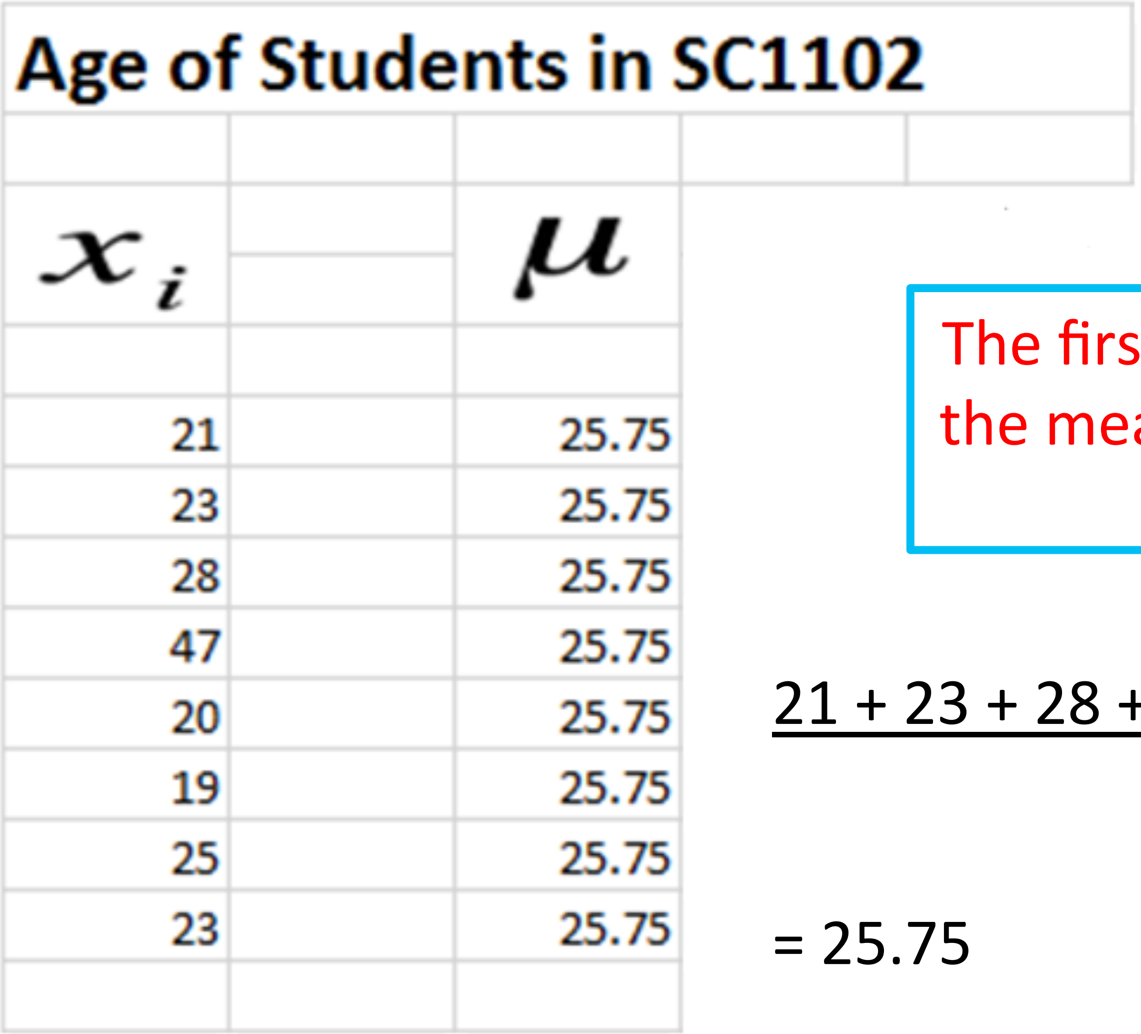
σ^2 = variance of the population (pronounced sigma squared)

x_i = the measurement of each data unit in the population

μ = the population mean

n = the size of the population

Population Variance



Age of Students in SC1102

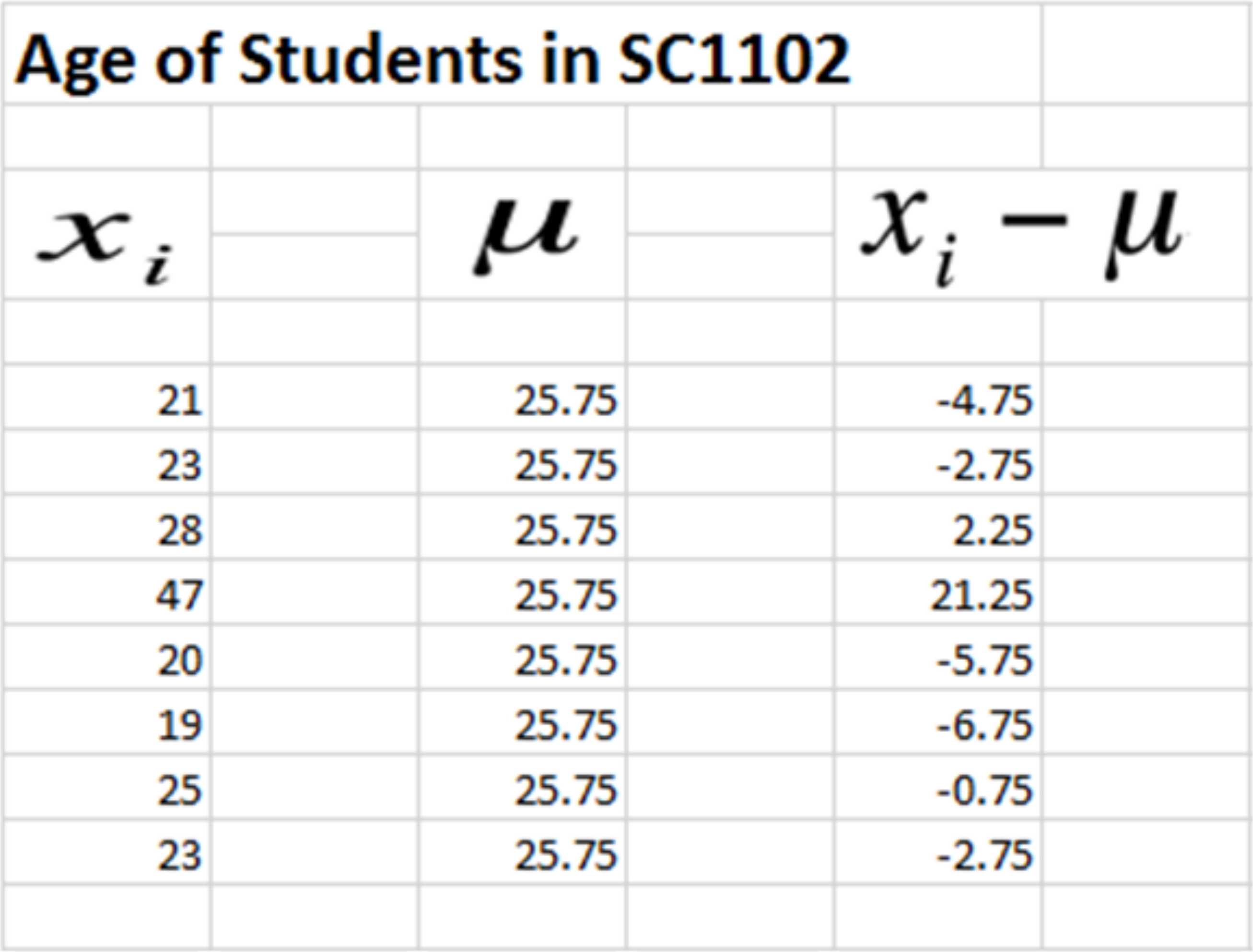
x_i		μ
21		25.75
23		25.75
28		25.75
47		25.75
20		25.75
19		25.75
25		25.75
23		25.75

The first step is to calculate the mean of the population (μ)

$$\frac{21 + 23 + 28 + 47 + 20 + 19 + 25 + 23}{8}$$

$= 25.75$

Population Variance

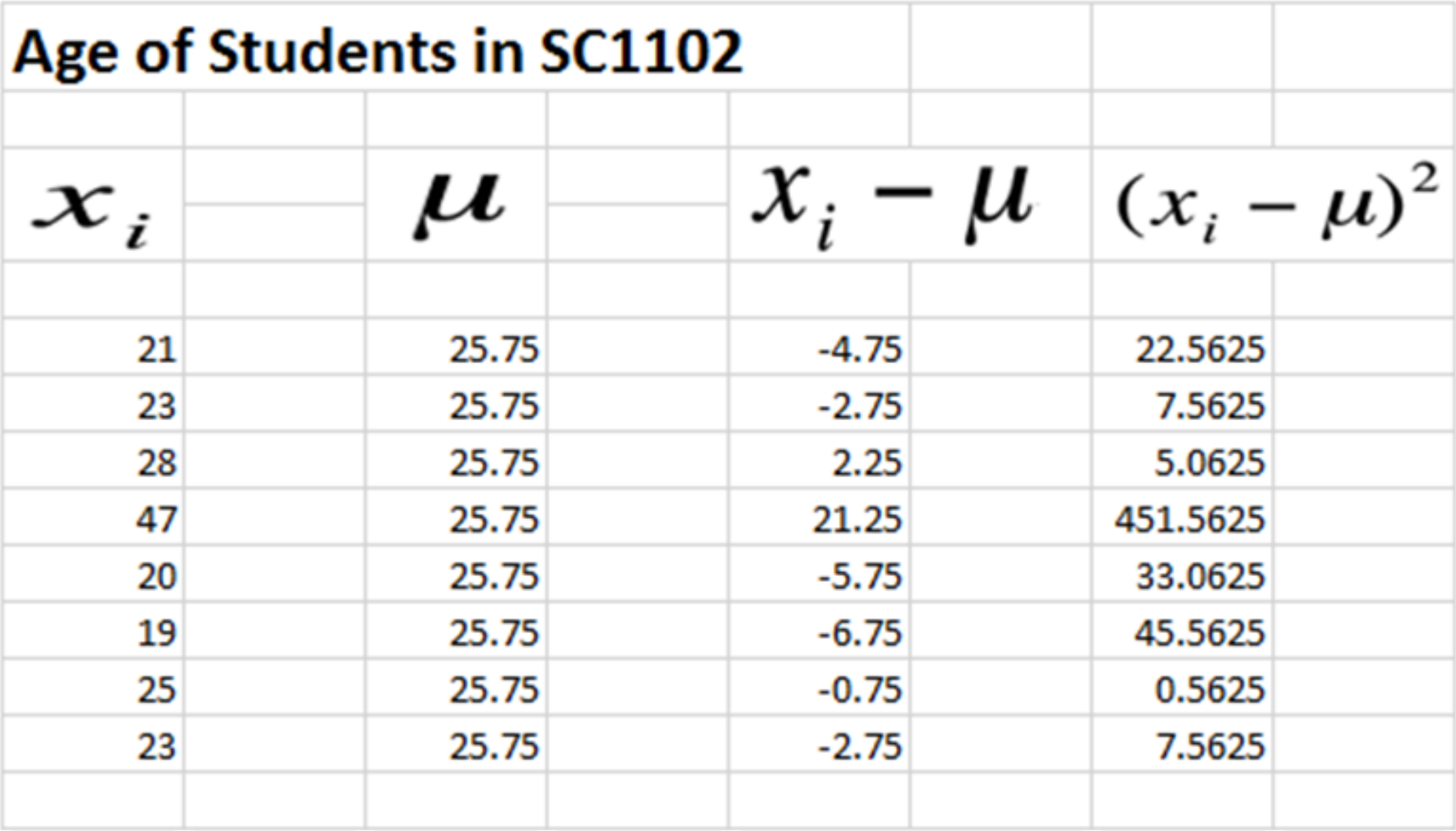


Age of Students in SC1102

x_i	μ	$x_i - \mu$
21	25.75	-4.75
23	25.75	-2.75
28	25.75	2.25
47	25.75	21.25
20	25.75	-5.75
19	25.75	-6.75
25	25.75	-0.75
23	25.75	-2.75

Subtract the mean of the population (μ) from the measurement of each data unit in the population

Population Variance

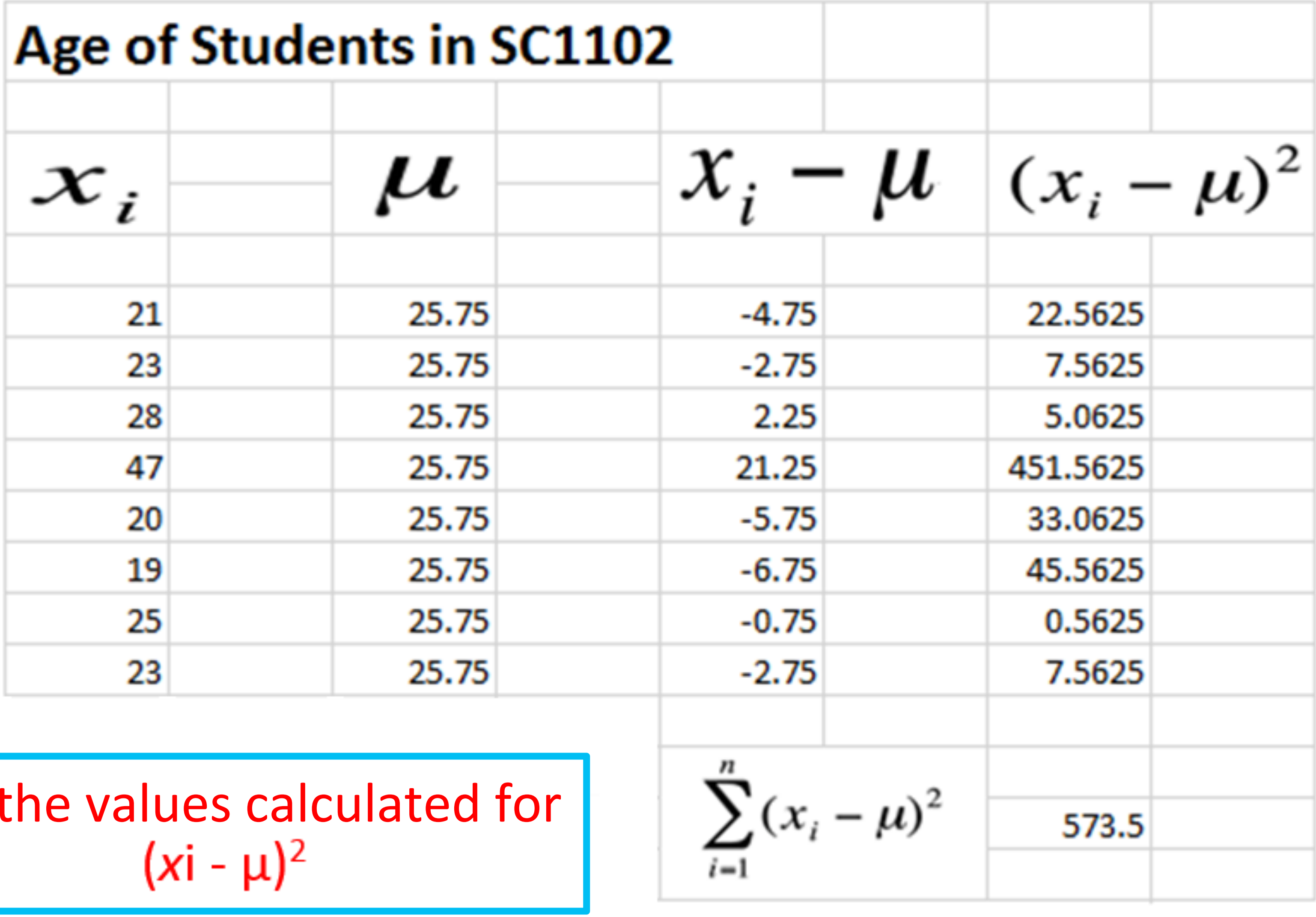


Age of Students in SC1102

x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
21	25.75	-4.75	22.5625
23	25.75	-2.75	7.5625
28	25.75	2.25	5.0625
47	25.75	21.25	451.5625
20	25.75	-5.75	33.0625
19	25.75	-6.75	45.5625
25	25.75	-0.75	0.5625
23	25.75	-2.75	7.5625

Square each value for x_i - μ

Population Variance



Age of Students in SC1102

x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
21	25.75	-4.75	22.5625
23	25.75	-2.75	7.5625
28	25.75	2.25	5.0625
47	25.75	21.25	451.5625
20	25.75	-5.75	33.0625
19	25.75	-6.75	45.5625
25	25.75	-0.75	0.5625
23	25.75	-2.75	7.5625
		$\sum_{i=1}^{n}(x_i - \mu)^2$	573.5

Add the values calculated for $(xi - \mu)^2$

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

$$= \frac{573.5}{8}$$

$$= 71.7$$

Divide

$\sum (xi - \mu)^2$

by the size of the population

Sample Variance

- If data is for a sample

Remember:

A subset of the population.

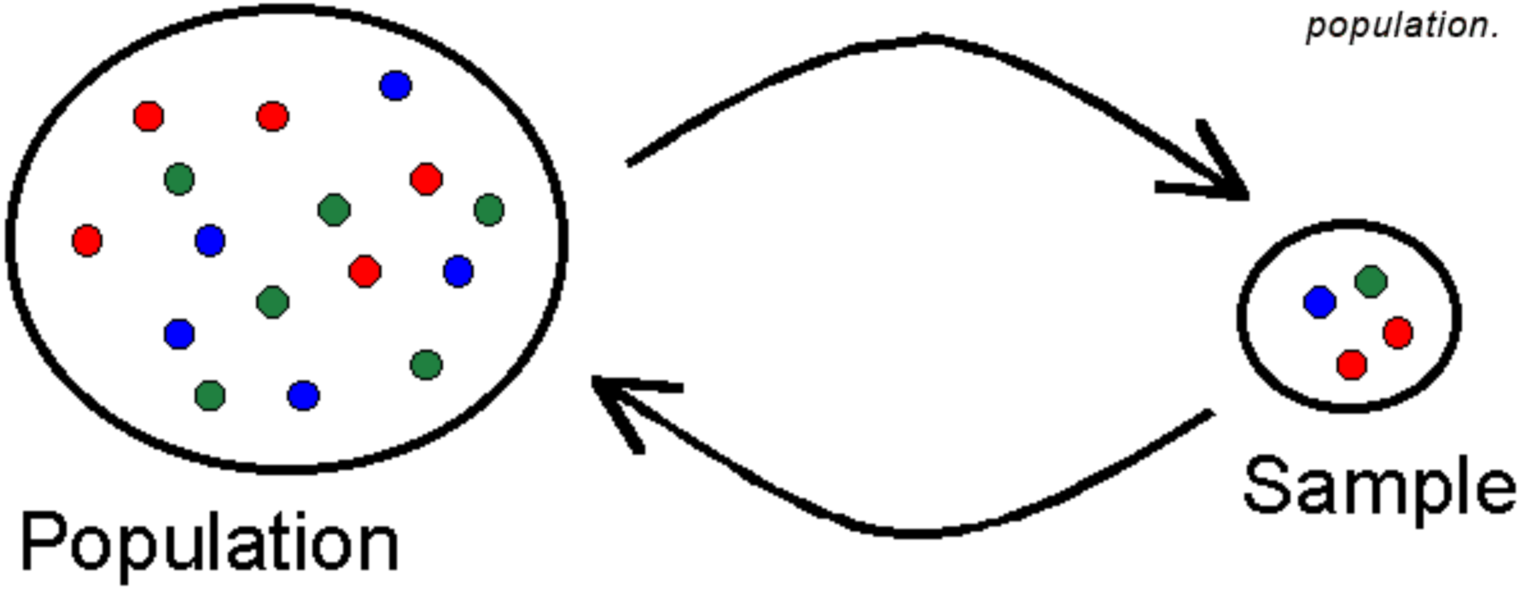


Image accessed: https://scholar.vt.edu/access/content/group/43c8db00-e78f-4dcd-826c-ac236fb59e24/STAT5605/normal01.htm

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Where:

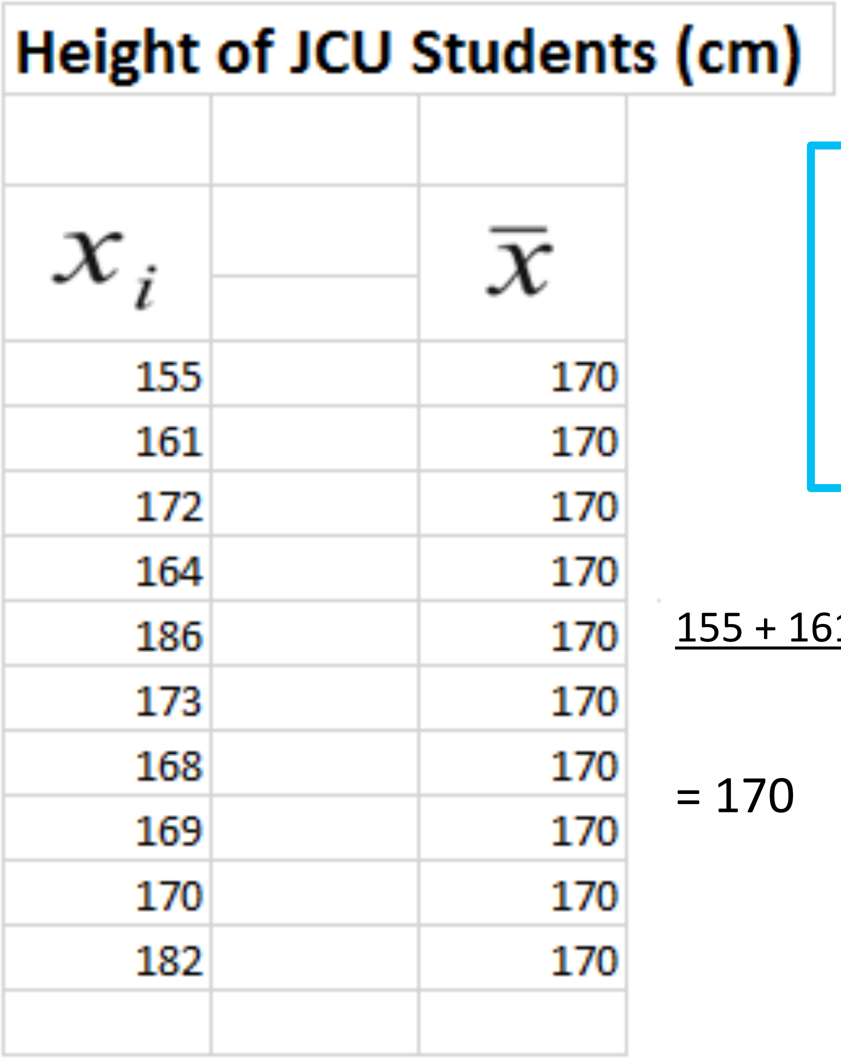
s^2 = the variance of the sample

x_i = the measurement of each data unit in the sample

$\bar{x}$ = the sample mean

n = the size of the sample (the number of data values)

Sample Variance



Height of JCU Students (cm)

x_i		$\bar{x}$
155		170
161		170
172		170
164		170
186		170
173		170
168		170
169		170
170		170
182		170

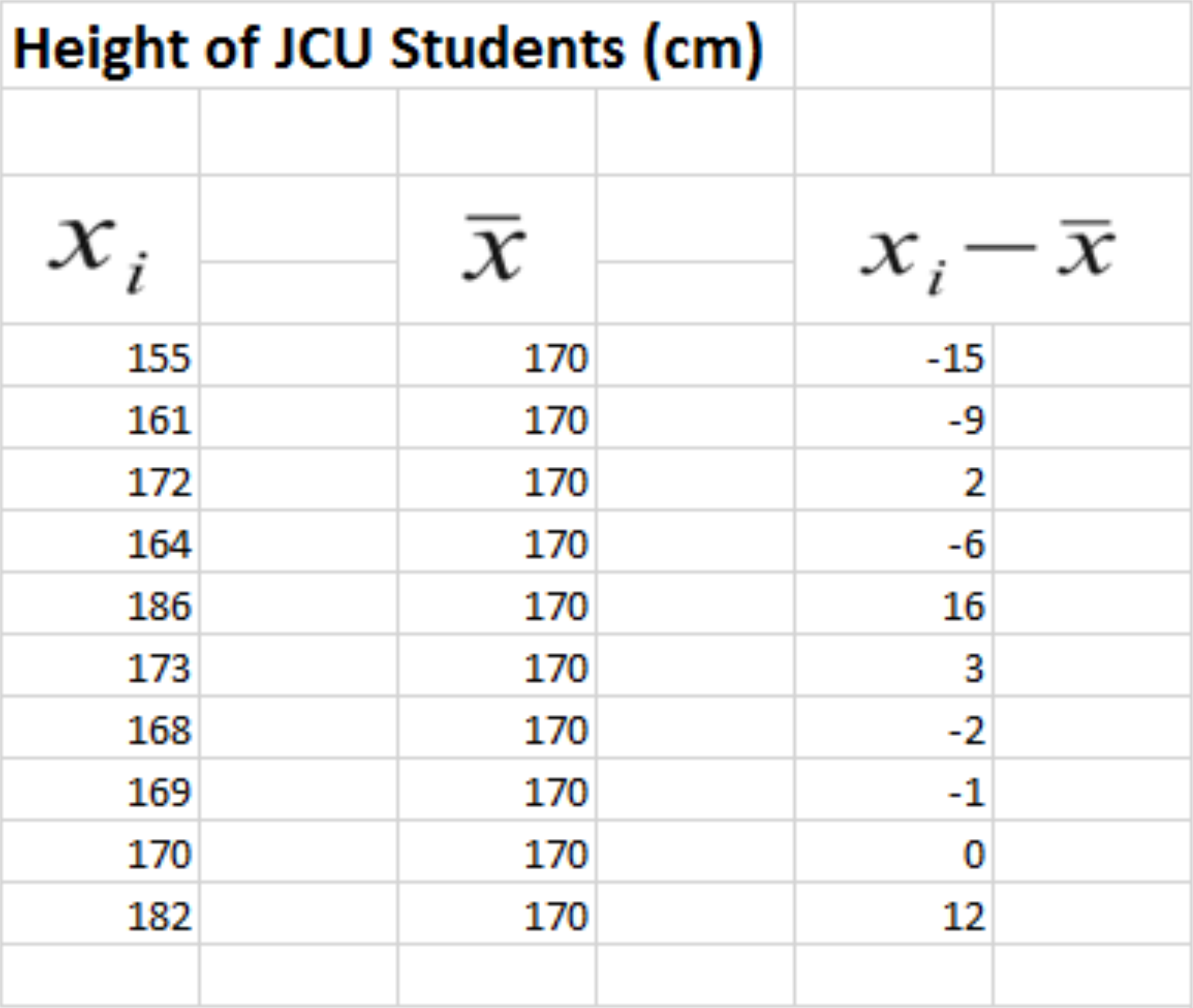
The first step is to calculate the mean of the sample:

$\bar{x}$

$$\frac{155 + 161 + 172 + 164 + 186 + 173 + 168 + 169 + 170 + 182}{10}$$

$= 170$

Sample Variance

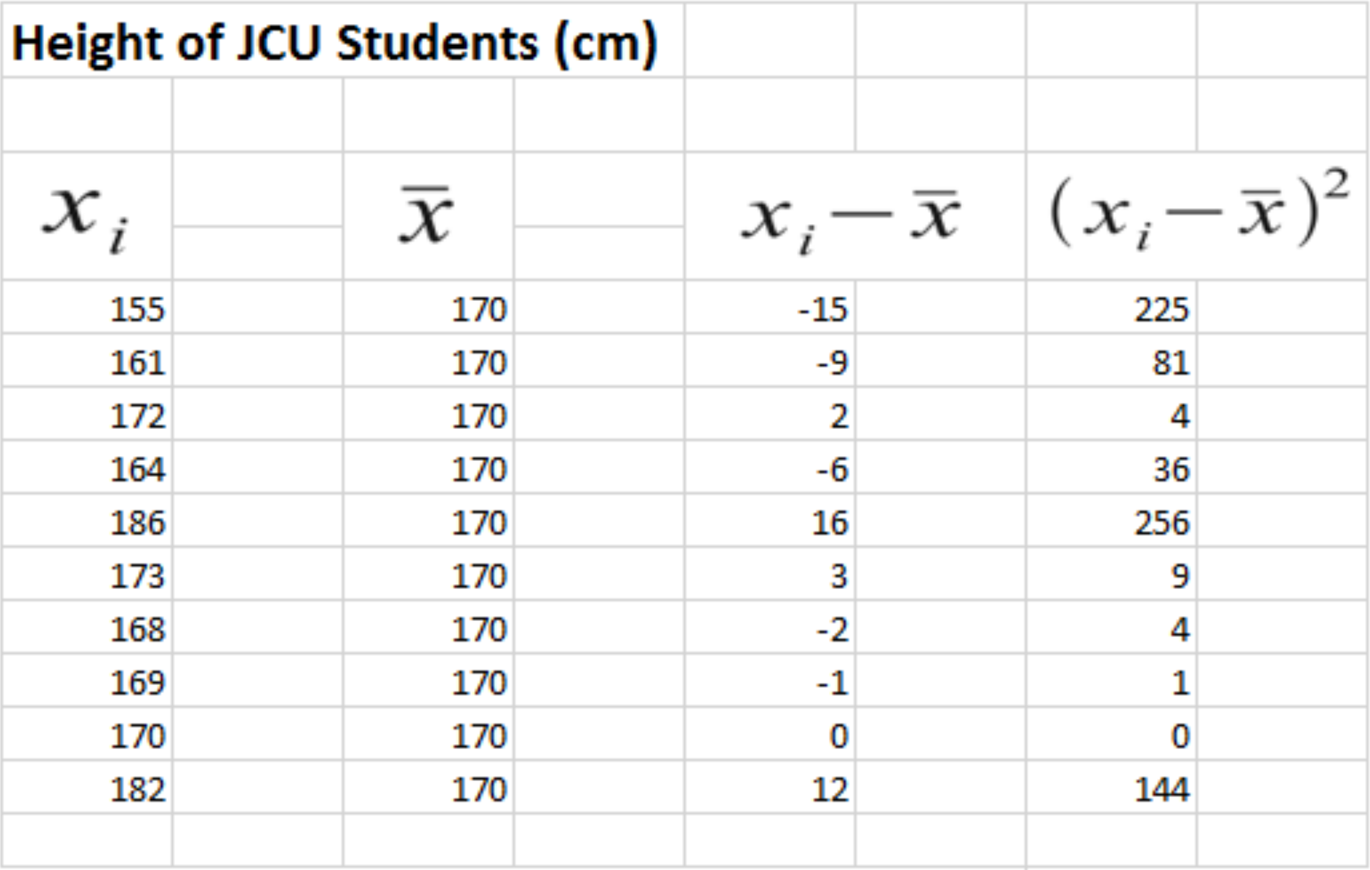


Height of JCU Students (cm)

x_i	$\bar{x}$	$x_i - \bar{x}$
155	170	-15
161	170	-9
172	170	2
164	170	-6
186	170	16
173	170	3
168	170	-2
169	170	-1
170	170	0
182	170	12

Subtract the mean of the sample ($\bar{x}$) from the measurement of each data unit in the sample

Sample Variance

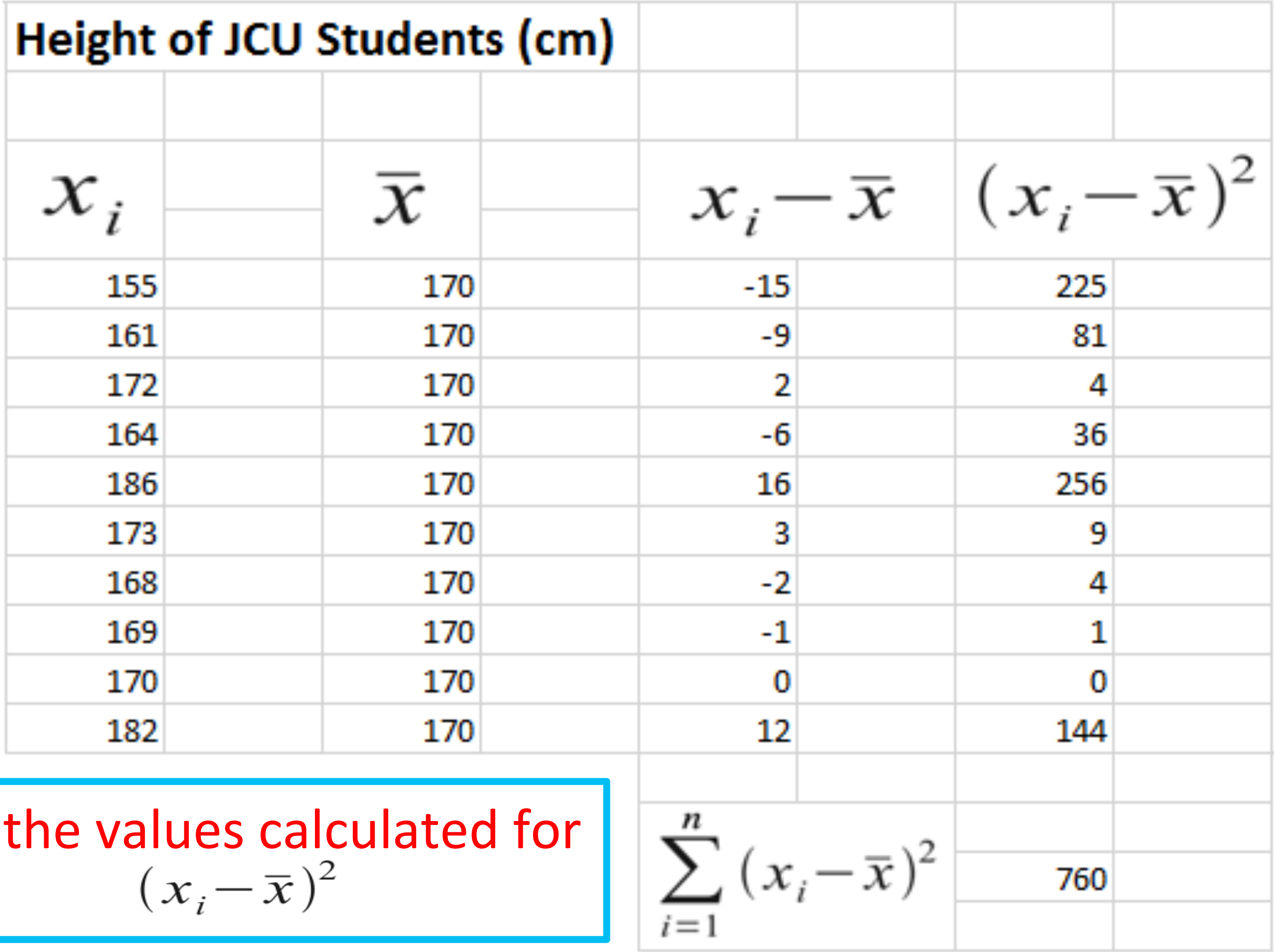


Height of JCU Students (cm)

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
155	170	-15	225
161	170	-9	81
172	170	2	4
164	170	-6	36
186	170	16	256
173	170	3	9
168	170	-2	4
169	170	-1	1
170	170	0	0
182	170	12	144

Square each value for $x_i - \bar{x}$

Sample Variance



Height of JCU Students (cm)

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
155	170	-15	225
161	170	-9	81
172	170	2	4
164	170	-6	36
186	170	16	256
173	170	3	9
168	170	-2	4
169	170	-1	1
170	170	0	0
182	170	12	144
		$\sum_{i=1}^{n}(x_i - \bar{x})^2$	760

Add the values calculated for $(x_i - \bar{x})^2$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

$$= \frac{760}{(10 - 1)}$$

$$= 88.4$$

Divide

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

by the size of the sample minus one

Standard Deviation

- Standard deviation is the square root of the variance
- There is SD for both the population and sample
- To calculate, first calculate the variance and then take the square root as the result
- A more useful measure than the variance as SD is in the units of the original data set

Standard Deviation - Population

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Where:

σ = the standard deviation of the population

x_i = the measurement of each data unit in the population

μ = the population mean

n = the size of the population

Standard Deviation - Sample

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

Where:

s = the standard deviation of the sample

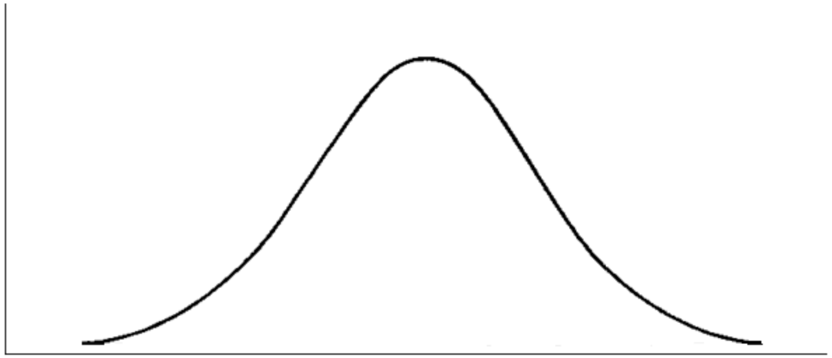
x_i = the measurement of each data unit in the sample

$\bar{x}$ = the sample mean

n = the size of the sample (the number of data values)

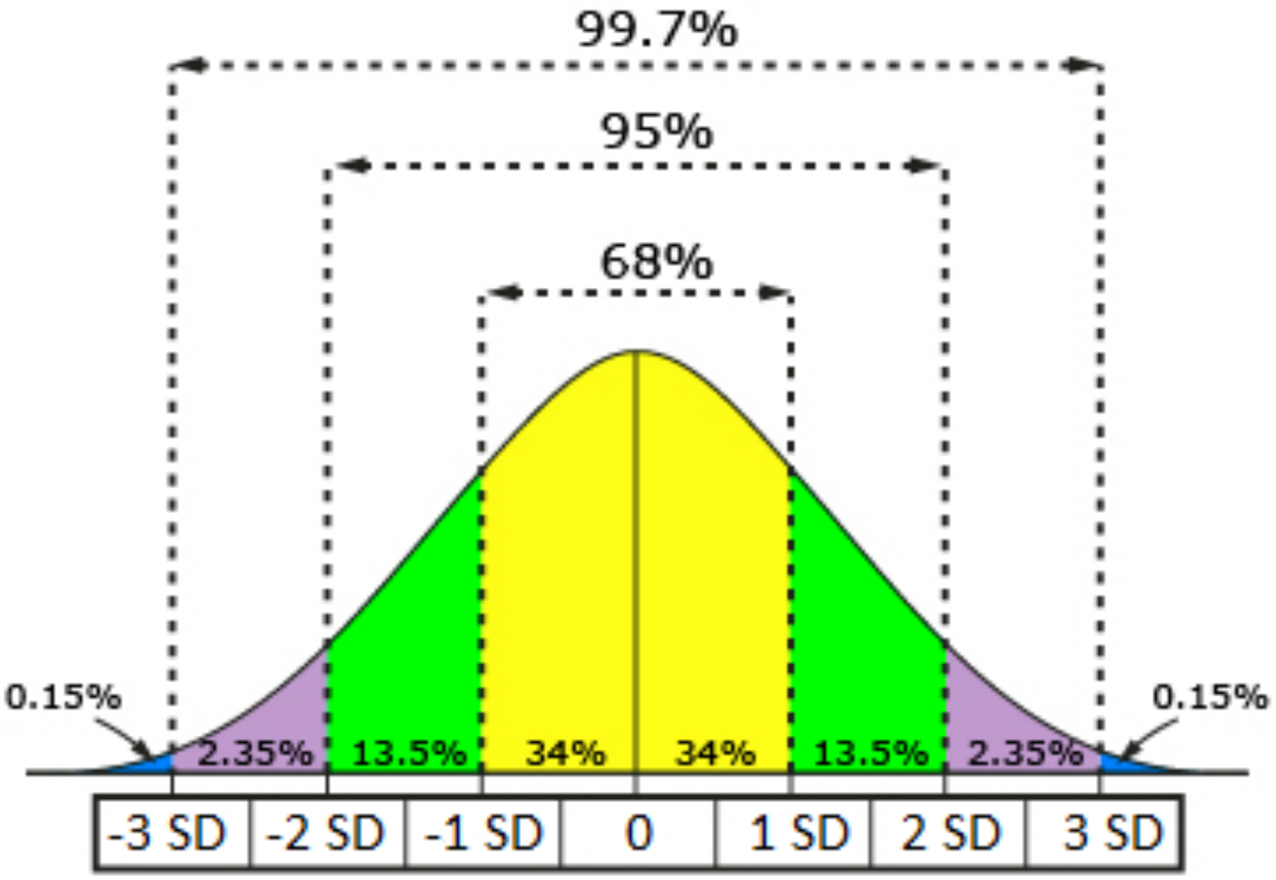
Standard Deviation – The Empirical Rule

- If the data distribution resembles a bell shape (ie. data is normally distributed)



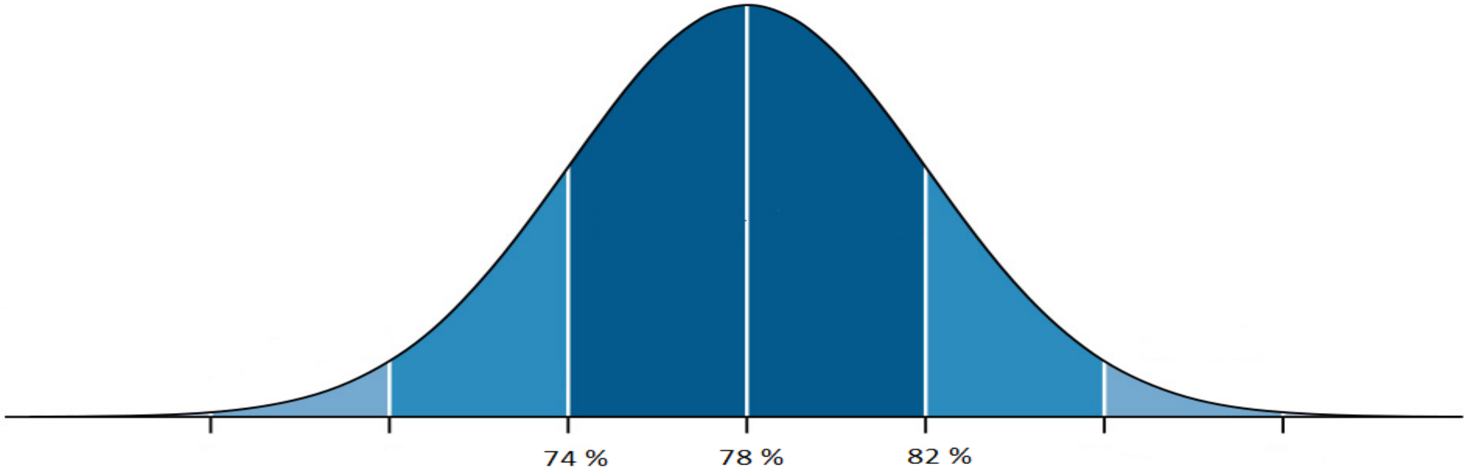
- The empirical rule tells us approximately 68 % of data values will fall within 1 standard deviation of the mean
- 95 % of data values will fall within 2 standard deviations of the mean
- 99.7 % of data values will fall within 3 standard deviations of the mean

Standard Deviation – The Empirical Rule



Standard Deviation – The Empirical Rule

- Mean exam score in a statistics class is 78 % and SD is 4 %
- Data normally distributed
- One SD above the mean is 82 % (78 + 4)
- One SD below the mean is 74 % (78 - 4)



- 68 % of the classes exam scores will fall between 74 % and 82 %