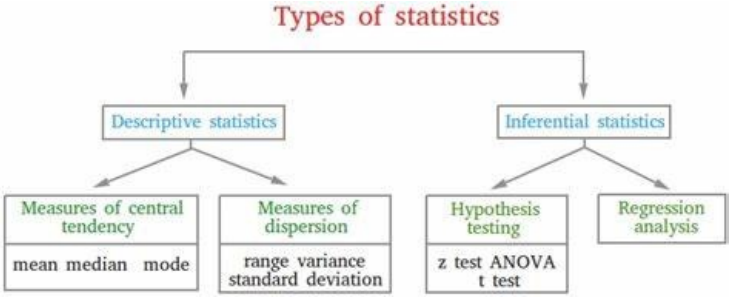
Statistics is the science of organizing, collecting and analyzing data.

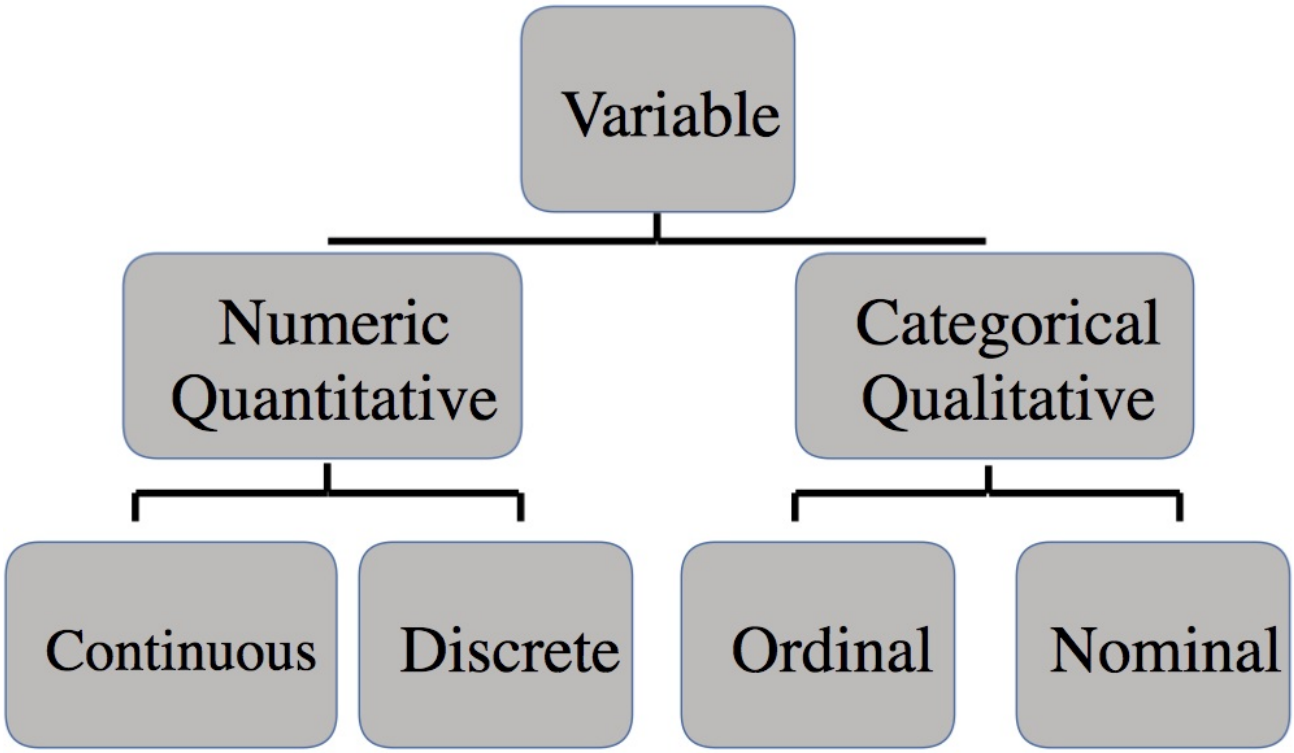
Type of Statistics

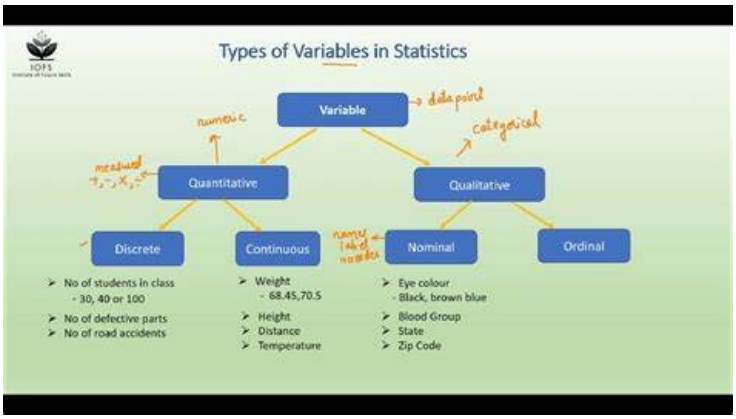


Difference between population (N) and sample (n)

Population	Sample
• A population is the entire collection of subjects affected by your research question.	• A sample is a subset of the population you study.
• Measurements taken from a whole population are called parameters.	• Measurements taken from a sample are called statistics.
• Data for an entire population is often very difficult or impossible to collect.	• When population data is unavailable, we use sample data to make inferences about the population.
• If you do have data for a whole population, your parameters will be "true" measures of some population characteristic.	• Sample data yield statistics, which can be used to estimate population parameters. These estimates will always involve some margin of error due to sampling bias and other errors.

Variable and Type of variable





Note: Mainly for continous value - we draw histogram and for discrete - we draw bar chart

Measure of central tendency

a. Mean b. Median c. Mode

Measure of central tendecy is used to measure the center of distribution of data

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

So, why there is various measure of central tendency like mean,median and mode.Because suppose we have data: 1,3,5,4,2 , what is mean for it? = (1+3+5+4+2)/5 = 3 but, if we add any other data, like 100 = 115/6 = 19.16 (SEE there is huge difference in mean by adding one data, exactly it is outlier, so looking oulier and distribution, we can choose different measure of central tendency like Mean, Mode and Median.

Now , if we do median in first case median is = sort it first (1,2,3,4,5) - median is 3 (whichever in middle) Similarly, in second case (1,2,3,4,5,100) - 3+4/2 = 3.5 is median Now see, there is not huge difference due to outlier like 100, thaswhy median works better with data which have outlier then mean.

Mode:Most frequent repetated number, mode does not work properly when there is outlier because suppose: our data= 1,2,3,4,5,200,200,200 (200 is outlier but it is repeating) Then what happen Mode will make it central tendency. OR measure it as center of distribution of data.

Mode can be mostly use in categorical variable , if there is some missing values, eg:

1. |Fruit|sale/day|
2. |Apple|20|
3. |orange|30|
4. |Apple|24|
5. |Mango|33|
6. |-||27|

- so now its better to use mode, and use apple inplace of missing categorical value

Measure of dispersion

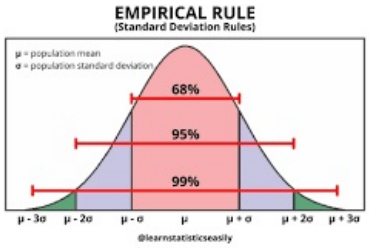
a. Variance b. Standard Deviation

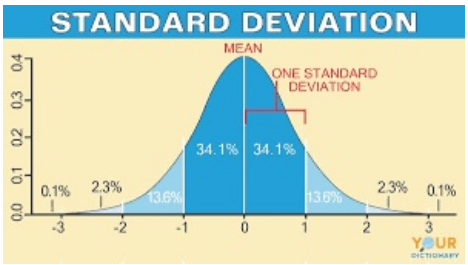
Dispersion is spread of data.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

Population	Sample
$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{n}}$	$S = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$
μ - Population Average xi - Individual Population Value n - Total Number of Population	x - Sample Average x_i - Individual Population Value n - Total Number of Sample

Standard deviation is square root of variance.





Variance mainly talk about spreadness of data, if variance is high -> spreadness is high

Example,calculating variance

Suppose we have a population dataset of exam scores:

Exam Score (X)
85
90
88
92
78
82

To calculate the variance of this population, we follow these steps:

1. Find the mean (average) of the dataset.
2. Subtract the mean from each data point and square the result.
3. Find the mean of the squared differences.

Let's calculate:

1. **Calculate the mean (μ):**

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\mu = \frac{85+90+88+92+78+82}{6} = \frac{515}{6} \approx 85.83$$

Exam Score (X)	Deviation (X - μ)	Deviation Squared ((X - μ)^2)
85	-0.83	0.6889
90	4.17	17.3289
88	2.17	4.7089
92	6.17	38.1289
78	-7.83	61.3689
82	-3.83	14.6689

1. **Find the mean of the squared differences (variance):**

$$\text{Variance} = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$$

$$\text{Variance} = \frac{0.6889+17.3289+4.7089+38.1289+61.3689+14.6689}{6}$$

$$\text{Variance} = \frac{137.8924}{6} \approx 22.98$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

$$\text{Standard Deviation} = \sqrt{22.98} \approx 4.79$$

So, the standard deviation of the population dataset is approximately 4.79.

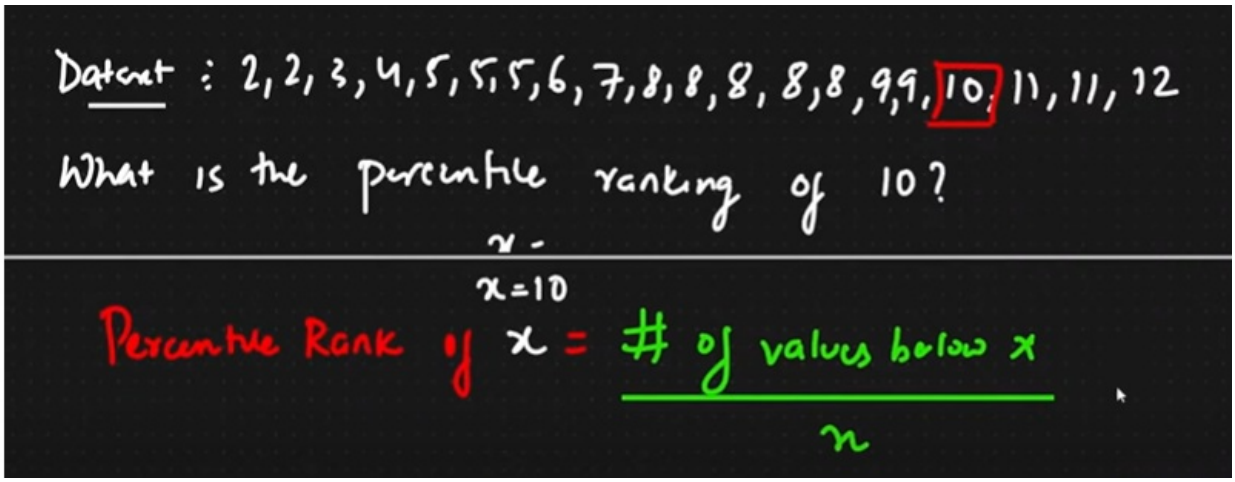
Now what exactly these values tell?

1. Variance: It quantifies the average squared difference between each data point and the mean of the dataset. A higher variance indicates that the data points are more spread out from the mean, while a lower variance suggests that the data points are closer to

the mean.

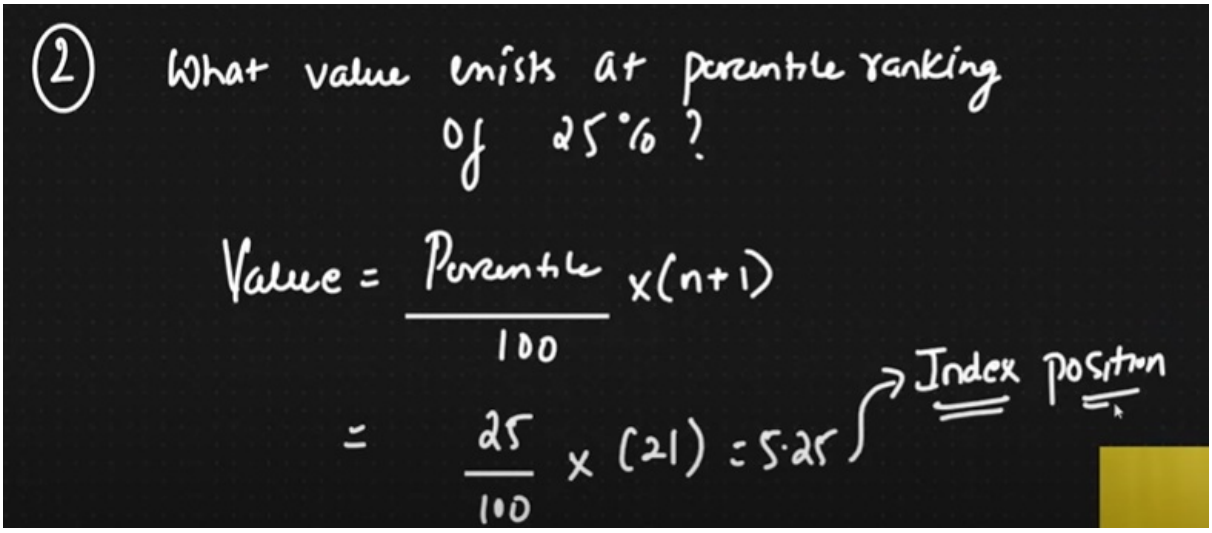
2. Standard Deviation: It is the square root of the variance and gives a measure of the average distance of each data point from the mean. A larger standard deviation means that the data points are more spread out from the mean, while a smaller standard deviation indicates that the data points are closer to the mean.

Percentile and Quartile - First step of finding outlier



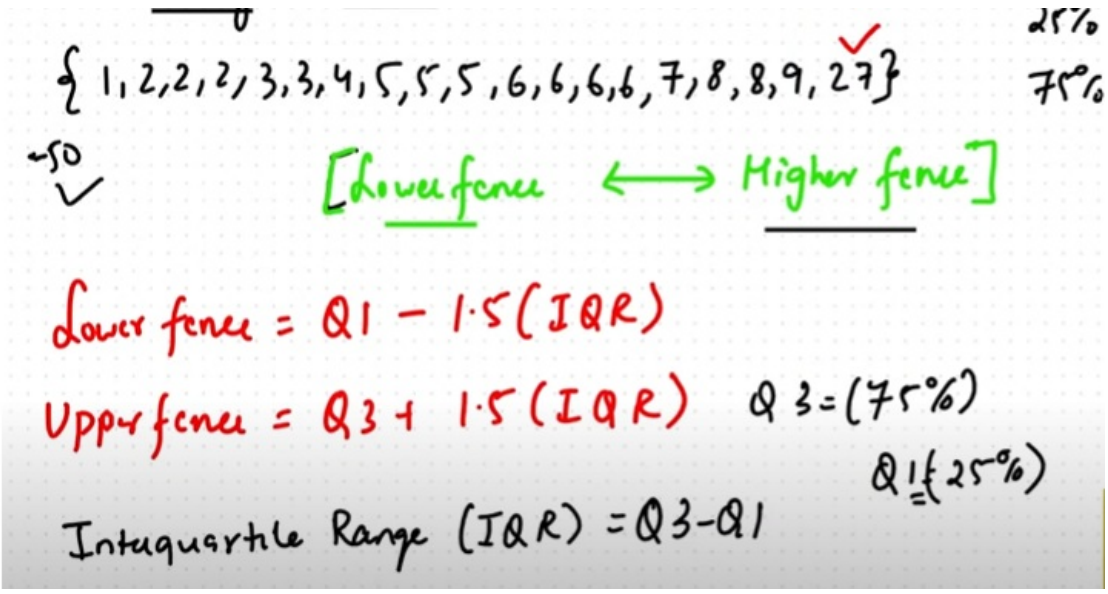
x=16/20*100 = 80% What is this mean?

- 80% of entire distribution is less than 10.



Five Number Summary

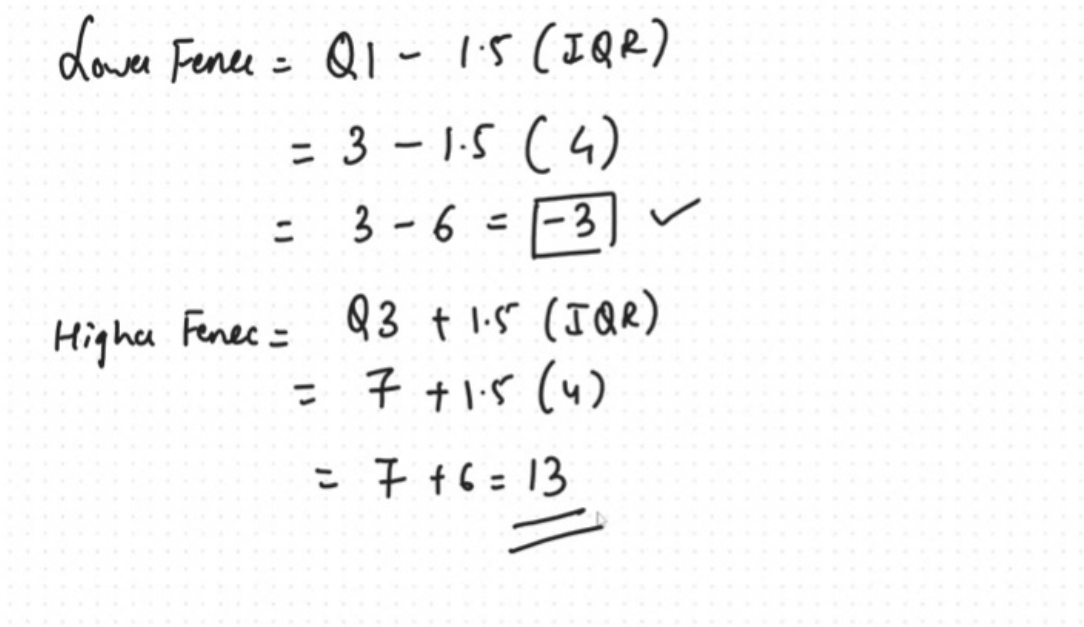
Note: By using 5 number summary, we can remove outlier



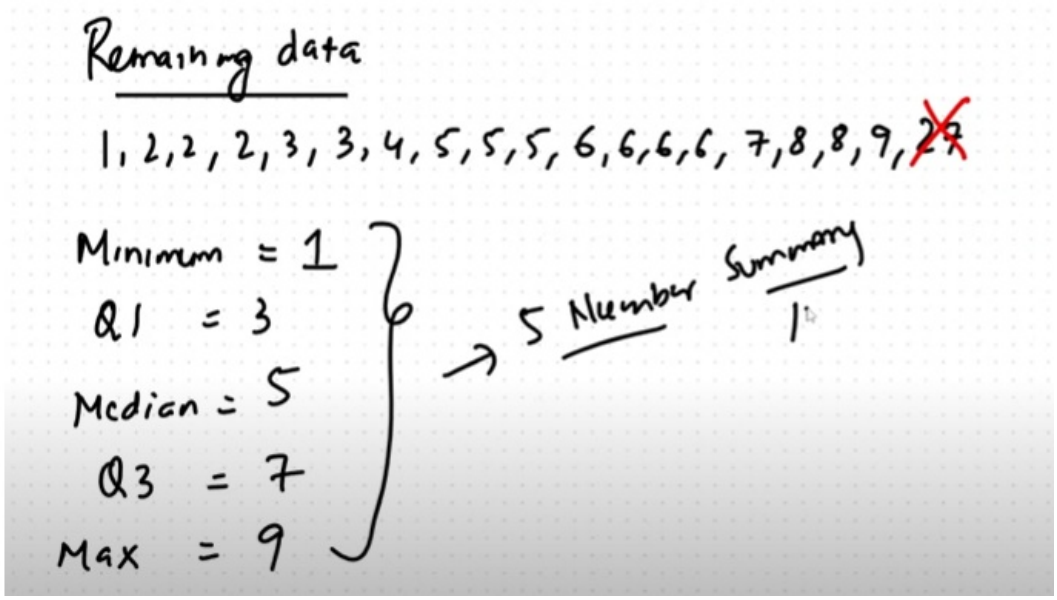
1. 25%= 25/100(19+1)=5th index poisition = 3
2. 75=15th = 7

Interquartile range = Q3-Q1=7-3=4

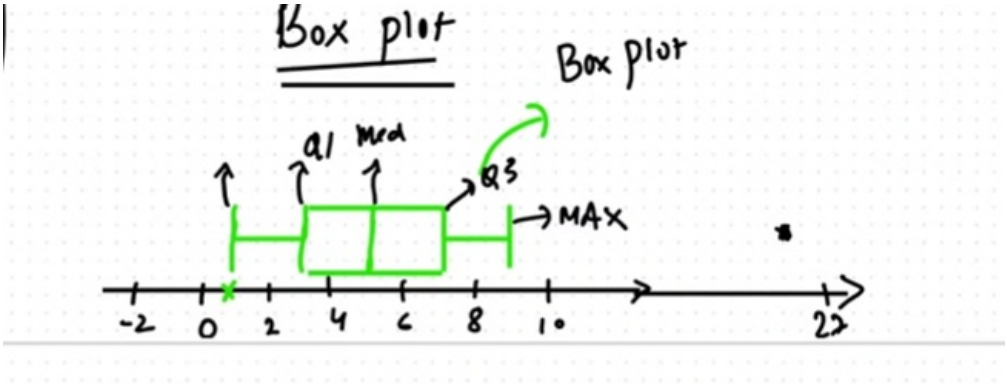
Finding lower and higher fence



Now all values below -3 and all values above 13 is outlier.



From 5 point summary, we can create box plot



Why n-1 in sample variance?

- Let's say you have a jar with 100 marbles. The marbles have different weights. This jar is our "population."
- Population Variance: You weigh all 100 marbles and use those weights to calculate the variance. You divide by 100 (the total number of marbles).
- Sample Variance: Now, let's say you randomly pick 5 marbles and weigh them. You calculate the variance for these 5 marbles.

But this time, you divide by 4 (n - 1), not 5. Why do you divide by 4?

- Remember, these 5 marbles are just a small piece of the 100 marbles in the jar. If you divide by 5, you might get a variance number that's too low. This could make you think the weights of all the marbles in the jar are closer together than they actually are. By dividing by "n - 1" (in this case, 4), you give a little "boost" to the variance number. This makes it a better estimate of how spread out the weights of all the marbles (the population) really are.

Example:

Suppose we have a sample of exam scores from five students:

Scores: 85 ,

90 ,

88 ,

92 ,

87 Scores: 85, 90, 88, 92, 87

Calculate the Sample Mean:

Sample Mean

85 + 90 + 88 + 92 + 87

5

442

5

88.4 Sample Mean= 5 85+90+88+92+87

5 442 =88.4

Calculate the Variance using "n":

Using n:

Variance

(85 − 88.4) 2 + (90 − 88.4) 2 + (88 − 88.4) 2 + (92 − 88.4) 2 + (87 − 88.4) 2 5 Variance= 5 (85−88.4) 2 +(90−88.4) 2 +(88−88.4) 2 +(92−88.4) 2 +(87−88.4) 2

= (− 3.4) 2 + (1.6) 2 + (− 0.4) 2 + (3.6) 2 + (− 1.4) 2

5

5 (−3.4) 2 +(1.6) 2 +(−0.4) 2 +(3.6) 2 +(−1.4) 2

= 11.56 + 2.56 + 0.16 + 12.96 + 1.96

5

5 11.56+2.56+0.16+12.96+1.96

= 29.2

5

5.84

5 29.2 =5.84

Calculate the Variance using "n-1":

Using n−1:

Variance

(85 − 88.4) 2 + (90 − 88.4) 2 + (88 − 88.4) 2 + (92 − 88.4) 2 + (87 − 88.4) 2 4 Variance= 4 (85−88.4) 2 +(90−88.4) 2 +(88−88.4) 2 + (92−88.4) 2 +(87−88.4) 2

= (− 3.4) 2 + (1.6) 2 + (− 0.4) 2 + (3.6) 2 + (− 1.4) 2

4

4 (−3.4) 2 +(1.6) 2 +(−0.4) 2 +(3.6) 2 +(−1.4) 2

= 11.56 + 2.56 + 0.16 + 12.96 + 1.96

4

4 11.56+2.56+0.16+12.96+1.96

= 29.2

4

7.3

4 29.2 =7.3

Comparing the two variances:

Variance using "n" = 5.84 Variance using "n-1" = 7.3 You can see that using "n" resulted in a smaller variance compared to using "n-1". This difference arises because dividing by "n" underestimates the true variability in the sample, whereas dividing by "n-1" provides a more accurate estimate of the population variance.

In [ ]: