Welcome to Session 5 of DS Foundation Course!

upx
Move up in life

To force their way over the enemy's surrounding wall :

1.They constructed ladders to reach the top of the enemy's wall, and they did this by calculating the height of the wall from the number of layers of bricks at a point which was facing in their direction and had not been plastered.

2.The layers were counted by a lot of people at the same time, and though some were likely to get the figure wrong, the majority would get it right, especially as they counted the layers frequently and were not so far away from the wall that they could not see it well enough for their purpose.

3.Thus, guessing what the thickness of a single brick was, they calculated how long their ladders would have to be ...

Which Central Tendency measure were they using to estimate the length of their ladders?

Mode

Crisis in Beaufort

Statistics to the rescue

Welcome to the Beaufort Police Department

Patrick

Data Scientist

Great with Programming and Statistics

Works in the Research Division

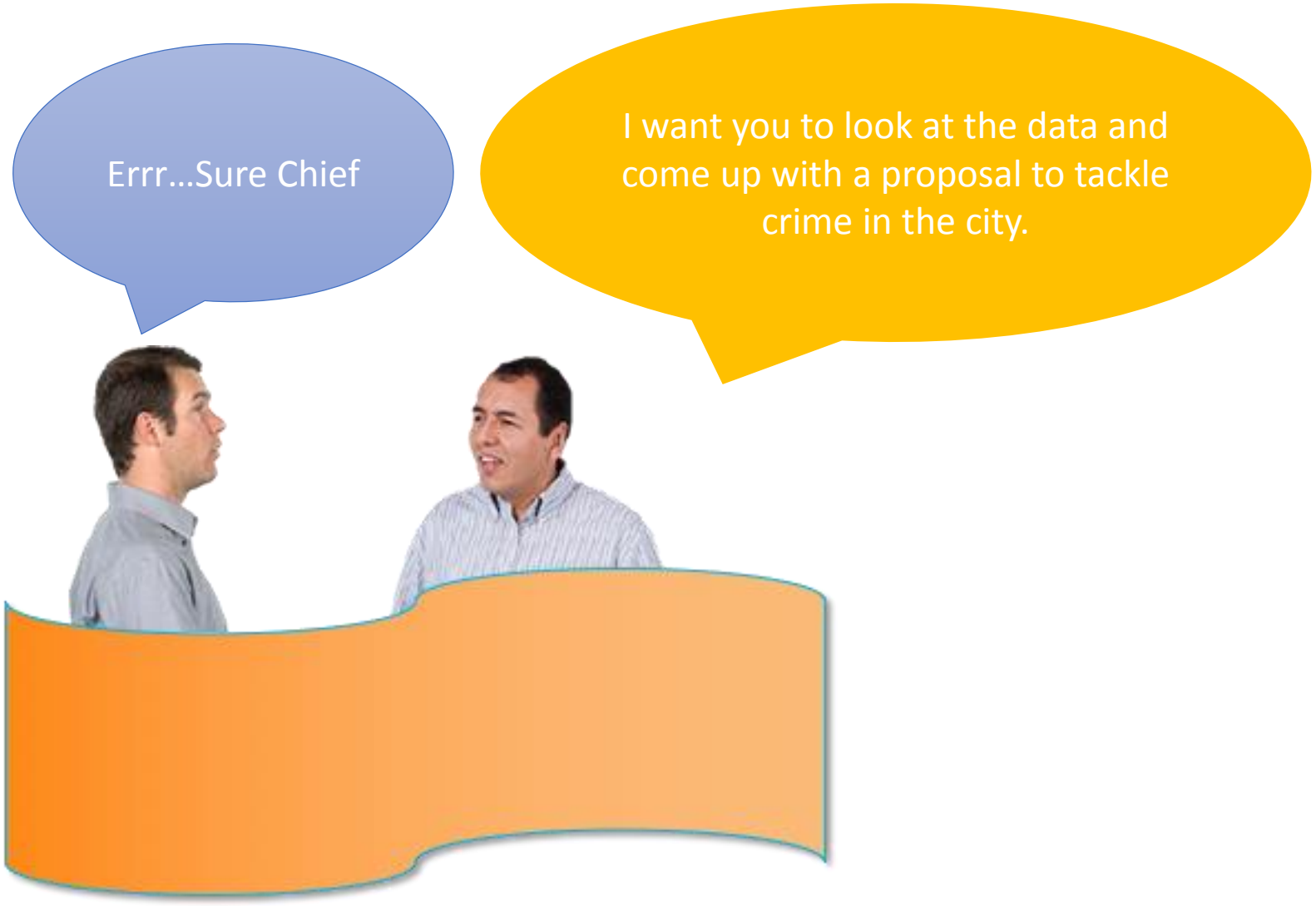
James

Chief of Police

Backs his instinct over numbers

Serving the state for 15+ years

The Crisis

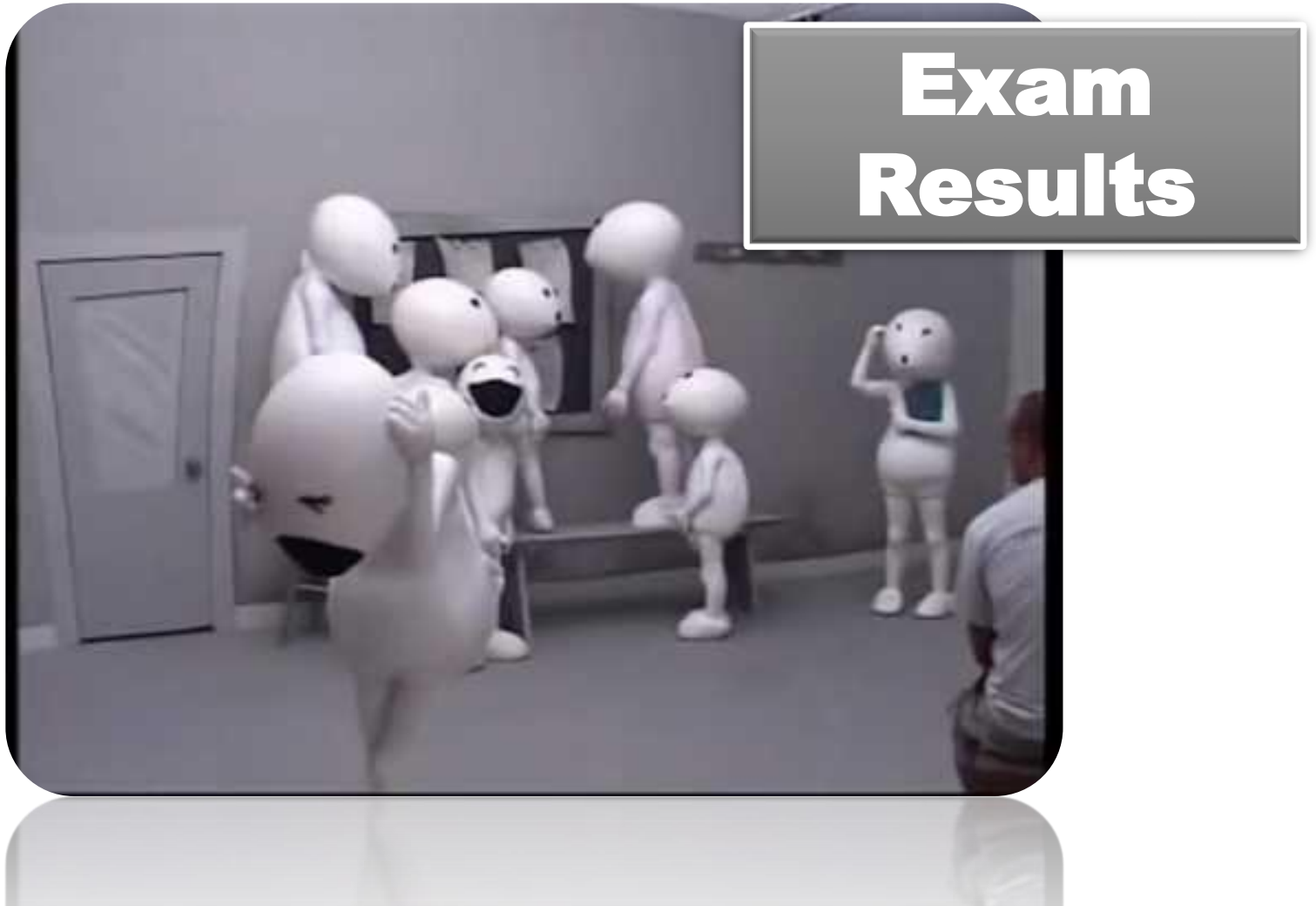

Errr…Sure Chief
I want you to look at the data and come up with a proposal to tackle crime in the city.

Patrick gets to work

- Identify the different ‘Crime’ categories
- Gather data for the past 10 years
- Analyze the data for any trends
- Report his findings to James

Step 1: Identify the different ‘Crime’ Categories

Central Tendency



How good is Ram?

Section A	Quiz 1	Quiz 2	Quiz 3
Ram	3	3	3
John	3	4	2
Rahman	3	4	2
Rivera	3	4	2
Lucy	3	5	1

Section B	Quiz 1	Quiz 2	Quiz 3
Ganesh	2	3	1
Alan	5	1	3
Sujatha	3	5	2
Murthy	3	2	5
Bindu	1	5	1

Step 3: Analyze the Car Theft data to identify trends

- To identify trends, Patrick decides to summarize the data
- To summarize data, he figured he would need to calculate a mean.

Step 3: Compare the monthly data with the historic data

- In an average year with 161 car thefts there may be an additional 16 cars stolen or 16 cars not stolen
- Between 145 – 177 cars are being stolen every year
- With this knowledge, Patrick could examine what is happening over time to see if any increases had appeared.
- Patrick decided to examine the monthly data for 2017.
- This would tell him what was happening this year, by month, and point out any increases that appear to be related to a crime pattern.
- Upon examination of the monthly reports, Patrick found that there did appear to be a rise in the number of cars stolen over the past six months.
- Average number of cars stolen this year every month was 20.

Step 4: Patrick presents his findings to James

- Patrick observes that the Car Thefts were indeed increasing quite alarmingly compared to the average (242 vs 161)
- Patrick repeats the exercise for other crime categories and finds incidents of assaults to be also increasing unexpectedly
- Based on Patrick's findings, James deploys specialized police investigators trained in handling car burglaries and assault incidents across Beaufort

Patrick analyzes other crime categories

Assaults: 63, 71,74, 78, 87, 98,246

Median

Extortion: 65, 70, 72, 74, 180, 201,203

Mean

Mean = 120
S.D = 60

Mean

What's up with all these symbols?

$$\text{Y-bar} = \frac{(Y1 + Y2 + \ldots + Yn)}{n}$$

$$\text{Y-bar} = \frac{\Sigma\, Yi}{n}$$

Y = your variable (could be X or Q or even "your name")

"-bar" or line over symbol of your variable = mean of that variable

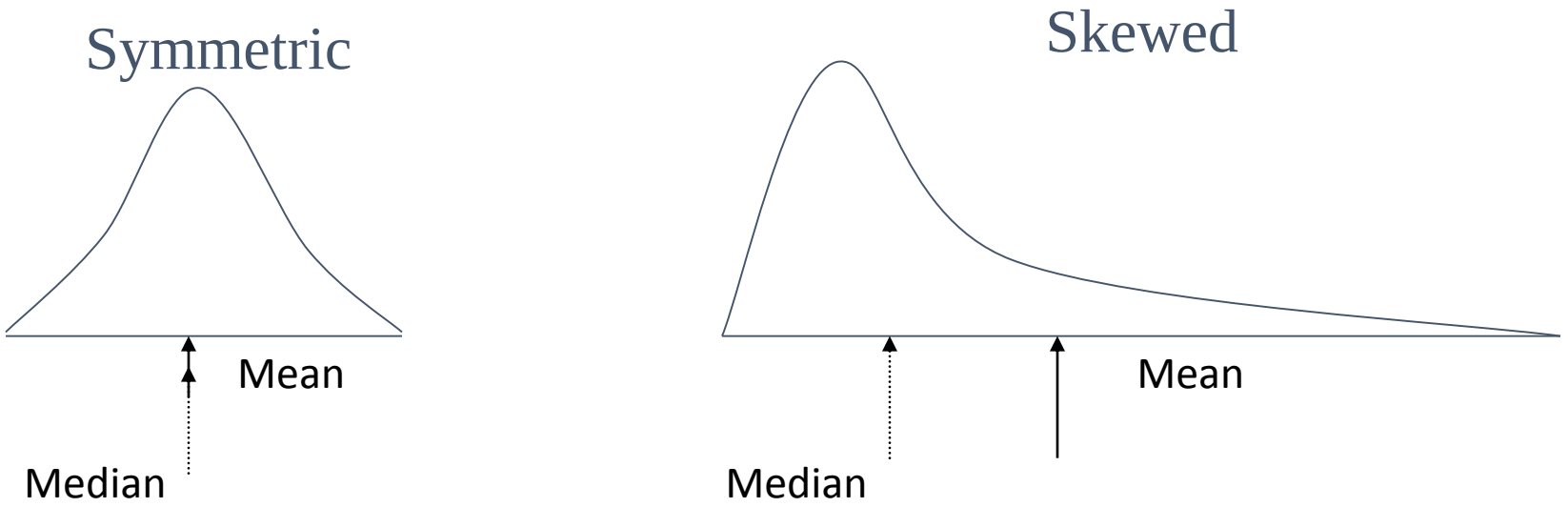
Y1 = first case's value on variable Y , ". . ." = ellipsis = continue sequentially

Yn = last case's value on variable Y, n = number of cases in your sample

Σ = Sum or add up what follows , *i* = a typical case or each case in the sample (1 through n)

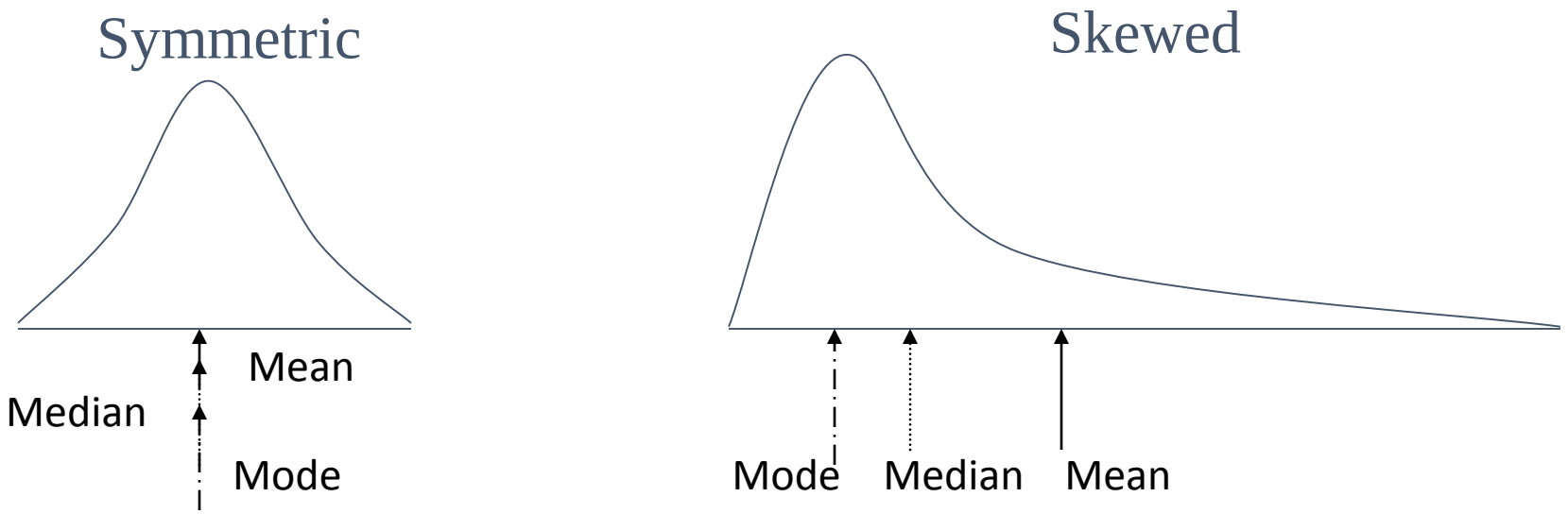
Median

- If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
- In skewed data, the mean lies further toward the skew than the median.



Mode – The most common Data Point

- It may give you the most likely experience rather than the “typical” or “central” experience.
- In symmetric distributions, the mean, median, and mode are the same.
- In skewed data, the mean and median lie further toward the skew than the mode.



Measures of Central Tendency

- Mean
- Median
- Mode

Measures of Variability

- Range
- Variance
- Standard Deviation

Range

The spread, or the distance, between the lowest and highest values of a variable.

To get the range for a variable, you subtract its lowest value from its highest value.

Class A—Scores of 13 Students	
102	115
128	109
131	89
98	106
140	119
93	97
110	

Class A Range = 140 - 89 = 51

Class B-- Scores of 13 Students	
127	162
131	103
96	111
80	109
93	87
120	105
109	

Class B Range = 162 - 80 = 82

Interquartile Range

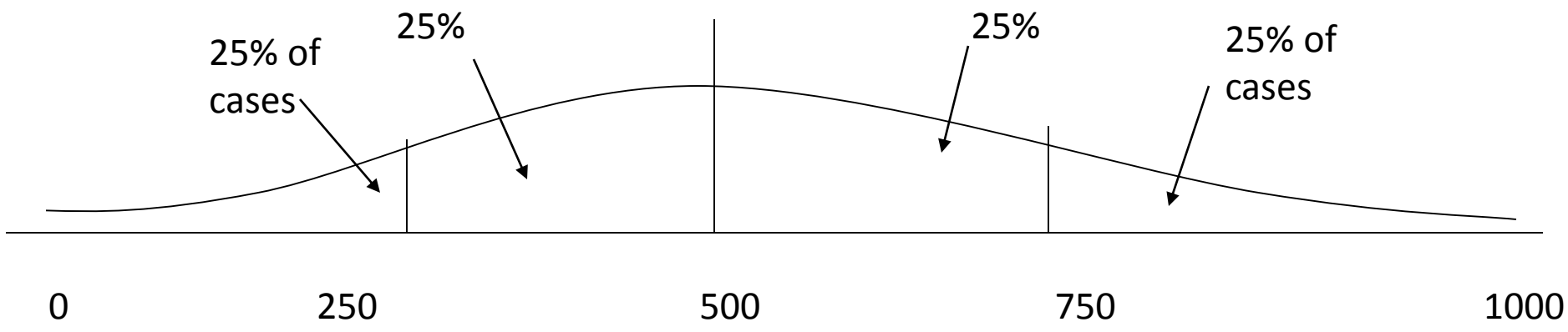
A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

The median is a quartile and divides the cases in half.

25th percentile is a quartile that divides the first ¼ of cases from the latter ¾.

75th percentile is a quartile that divides the first ¾ of cases from the latter ¼.

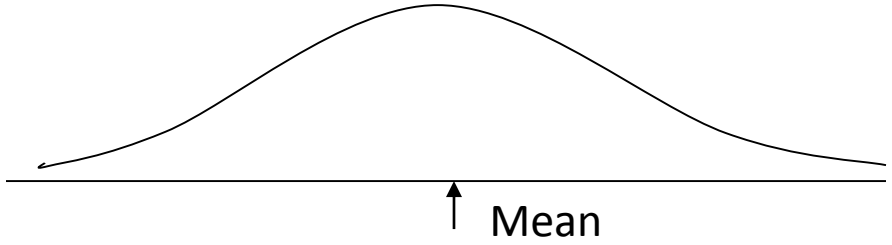
The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?



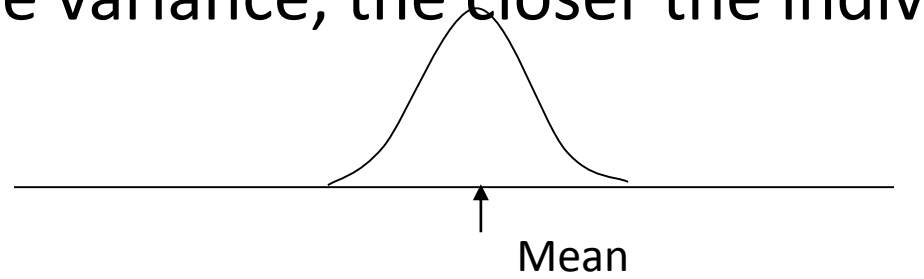
Variance

A measure of the spread of the recorded values on a variable. A measure of dispersion.

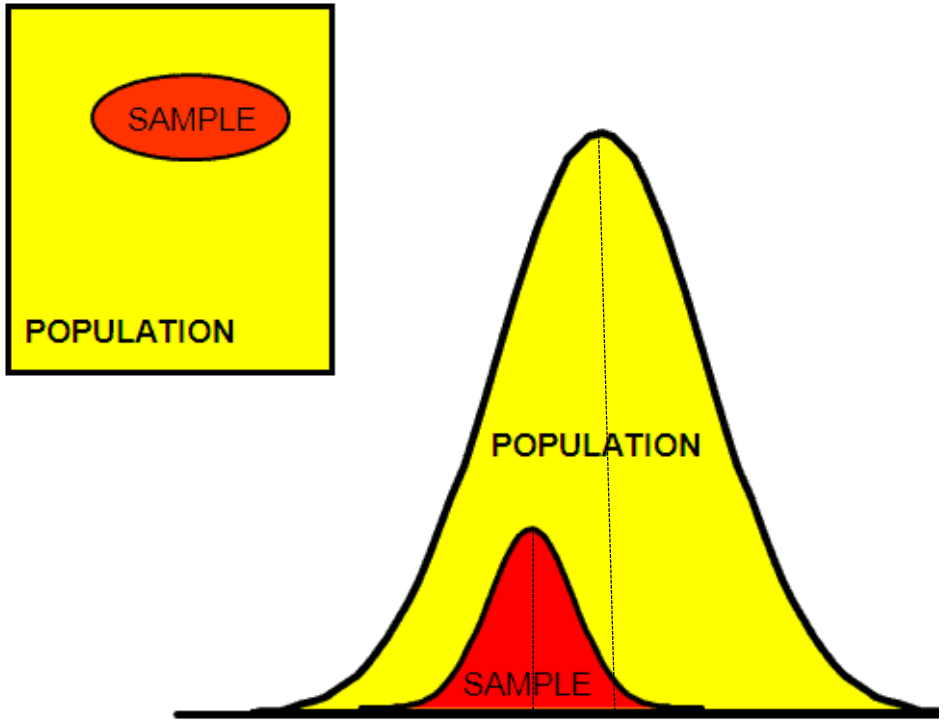
The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



Variance & Std. Deviation of **Sample vs Population**



For samples:

$$\text{variance} = s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

$$\text{standard deviation} = s = \sqrt{s^2}$$

Calculating Formula

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - 1}$$

For populations:

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

$$\text{standard deviation} = \sigma = \sqrt{\sigma^2}$$

Calculating Formula

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