Measures of Spread(Variability or Dispersion):

Sometimes the values in a data set are close together on a number line, while other times they are far apart from each other on a number line. Data sets that are close together are said to have less spread, while data sets that are far apart have a larger spread.

Our textbook considers two methods for measuring spread.

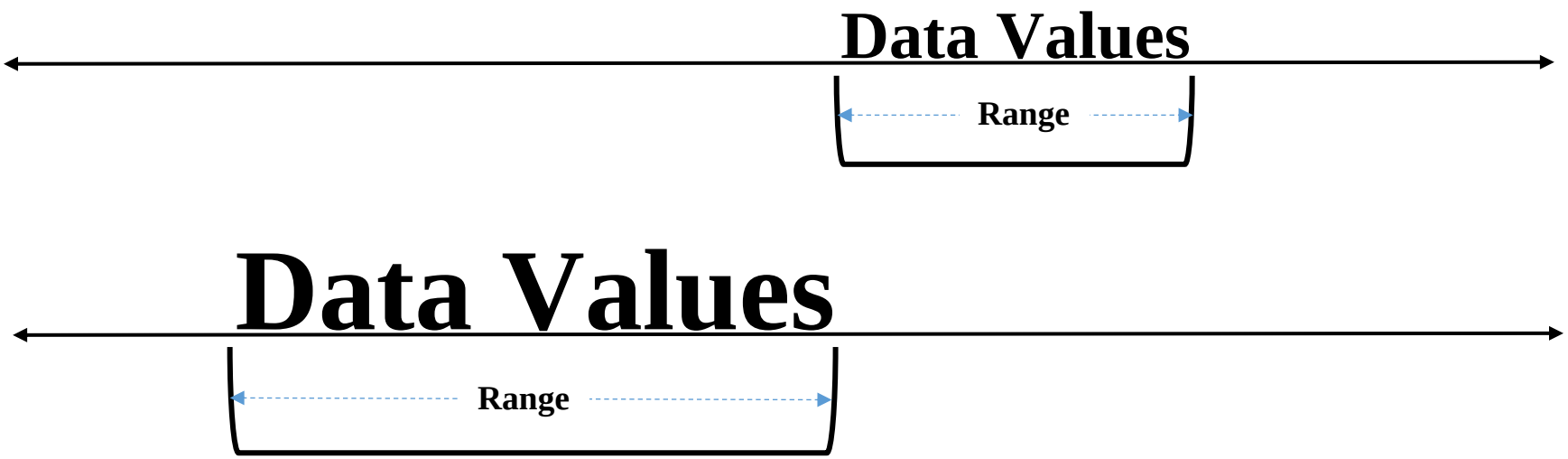
Range

Variance/Standard Deviation

Range:

The range of a data set is simply the difference between the largest value and the smallest value.

Geometrically, it represents the smallest width of an interval on a number line that could enclose all the values in the data set. The wider the interval needed to enclose all the values, the more spread the data set has. The narrower the interval needed to enclose all the values, the less spread the data set has.



Examples:

1. $\{5,3,7,2,2,11\}$

$$Range = 11 - 2 = \boxed{9}$$

2. $\{5,3,6,2,2,4,3\}$

$$Range = 6 - 2 = \boxed{4}$$

Which data set is considered to have more variability based upon the range values?

The first data set is considered to have more variability since it has a larger range.

The value of the range is determined by only two numbers in the data set!

Variance:

A center for the data set is established, and the variance is an average of the squared distances of the data values from the center.

The center used in the variance is the mean of the data values.

Every value of the data set contributes to the value of the variance!

If the data set represents a sample, then we have the sample variance.

If the data set represents a population, then we have the population variance.

Our textbook assumes that the data sets are samples.

Examples:

1. $\{1,2,3,6\}$

The mean of the data set is $\frac{1+2+3+6}{4}=\frac{12}{4}=3=\bar{x}$**.**

Associated with each data value, *x*, is its deviation from the mean. We're assuming that the data set represents measurements from a sample, so the deviations would be $x-\bar{x}$**.**

x	$\bar{x}$	$x-\bar{x}$
1	**3**	–2
2	**3**	–1
3	**3**	0
6	**3**	3
	Total	0

Why can't an average of the deviations be used to measure variability?

Their sum will always be zero, so their average will always be zero, as well.

To alleviate that the average of the deviations is always zero, they are squared to produce squared deviations whose average will be useful.

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
1	**3**	–2	4
2	**3**	–1	1
3	**3**	0	0
6	**3**	3	9
	Total	0	14

The general formula for the sample variance is

$$s^2 = \frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{14}{3} = \boxed{4.67}$$

Values of variances are usually reported to two decimal places beyond the data values.

The big problem with a variance is that its units are the square of the original units in the data set. To fix this problem, the square root of the variance is calculated, leading to the standard deviation.

Sample Standard Deviation:

$$s = \sqrt{s^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{14}{3}} = \boxed{2.16}$$

When calculating standard deviations, don't use a rounded variance! After calculating the square root of the unrounded variance, the standard deviation is usually reported to two decimal places beyond the data values.

2. $\{1,1,1,9\}$

The mean of the data set is $\frac{1+1+1+9}{4}=\frac{12}{4}=3=\bar{x}$**.**

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
1	**3**	–2	4
1	**3**	–2	4
1	**3**	–2	4
9	**3**	6	36
	Total	0	48

Sample Variance, s^2**:** $\frac{48}{3} = \boxed{16}$

Sample Standard Deviation, s**:** $\sqrt{\frac{48}{3}} = \boxed{4}$

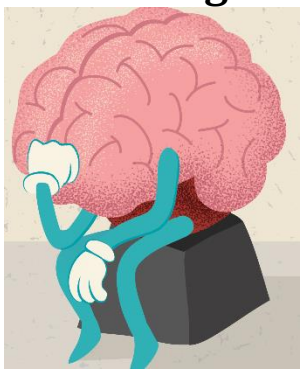
By comparing standard deviations which data set has more variability: $\{1,2,3,6\}$ **or** $\{1,1,1,9\}$**?** $\{1,1,1,9\}$ has the larger standard deviation, so it's considered to have more variability.

3. Without doing the calculations, which data set will have the larger standard deviation?

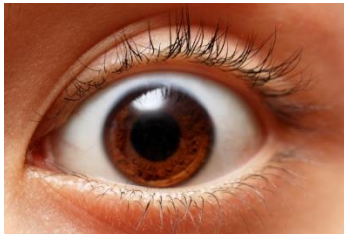
Set A: 35, 38, 41, 43, 50

Set B: 4, 19, 30, 40, 51

B



4. Without doing the calculations, which data set will have the smallest standard deviation?

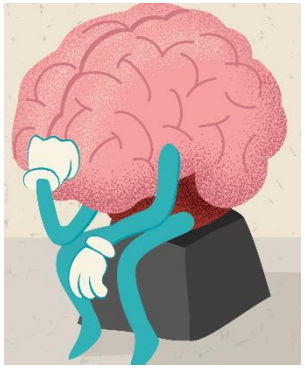


Set A: 46, 55, 72, 103, 103

Set B: 33, 99, 100, 100, 100

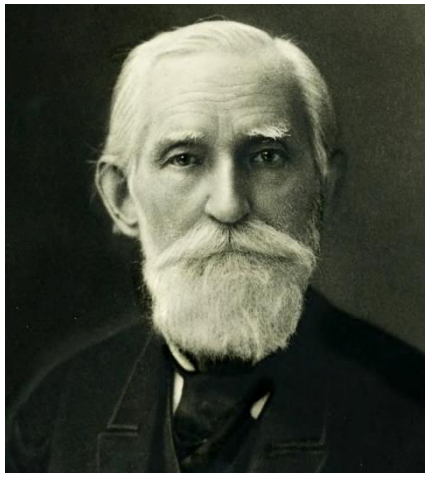
Set C: 54, 59, 61, 63, 64

C



Chebyshev's Theorem/Inequality:

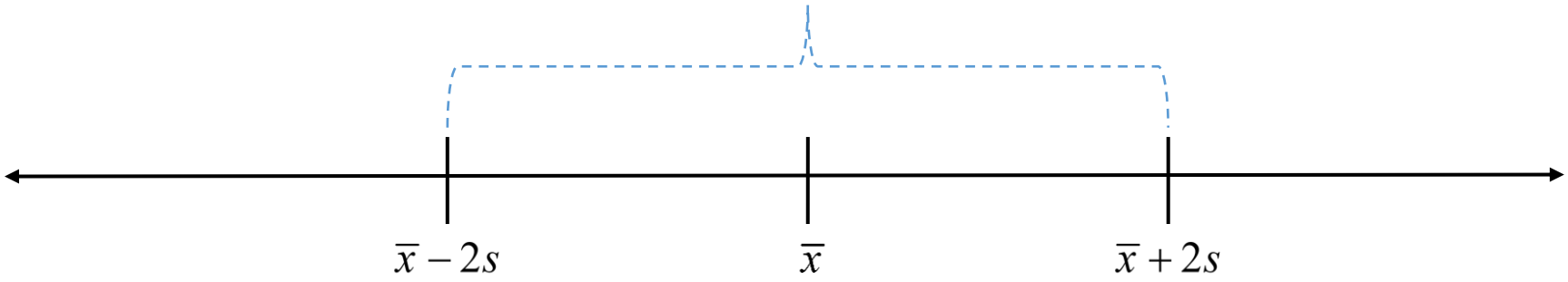
For any data set, the proportion of data values that are within *k* standard deviations from the mean is at least $1-\frac{1}{k^2}$.



Chebyshev's Theorem deals with the concentration of the data values in the vicinity of their mean value.

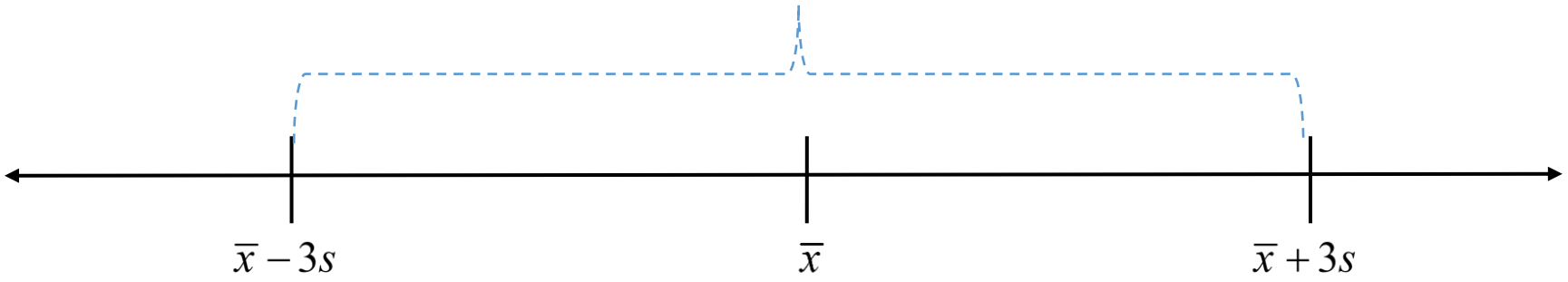
For $k = 2$, the proportion of data values within 2 standard deviations from the mean is at least $1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = .75 = 75\%$.

At least 75% of the data values are between these two numbers.



For $k = 3$, the proportion of data values within 3 standard deviations from the mean is at least $1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = .\overline{8} = 88.\overline{8}\%$.

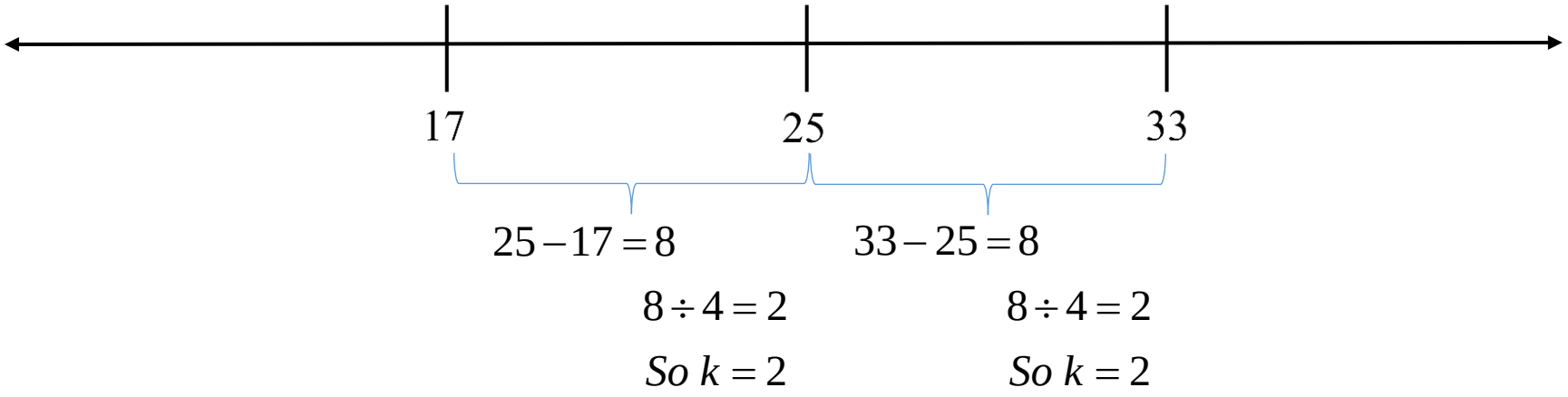
At least $88.\overline{8}\%$ of the data values are between these two numbers.



Examples:

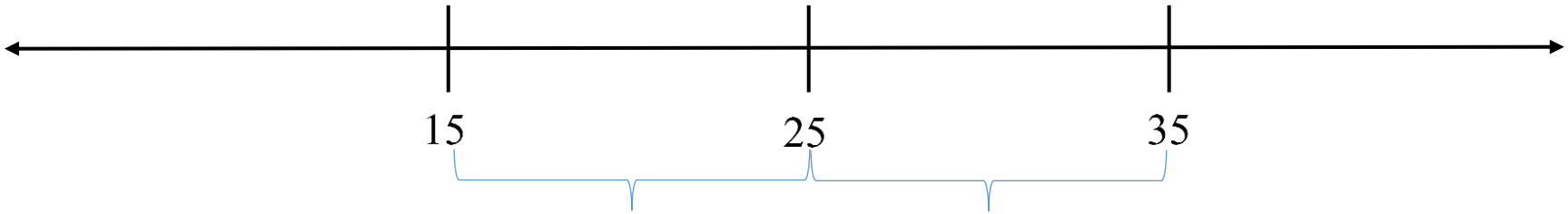
The flights for a certain airline are late by an average of 25 minutes with a standard deviation of 4 minutes.

At least what percentage of flights are between 17 and 33 minutes late?



$$1-\frac{1}{2^2}=1-\frac{1}{4}=\frac{3}{4}=\boxed{75\%}$$

At least what percentage of flights are between 15 and 35 minutes late?



$25-15=10$ $\qquad$ $35-25=10$

$10\div 4=2.5$ $\qquad$ $10\div 4=2.5$

$So\ k=2.5$ $\qquad$ $So\ k=2.5$

$$1-\frac{1}{(2.5)^2}=1-\frac{4}{25}=\frac{21}{25}=\boxed{84\%}$$

For the data set $\{1,5,6,7,8,9,10,15,25,34\}$, the sample mean is 12 and the sample standard deviation is 10.12. according to Chebyshev's Theorem, at least $55.\overline{5}\%$ of the values in the data set must be within 1.5 standard deviations from the mean. What's the actual percentage of values in the data set that are within 1.5 standard deviations from the mean?

$s = 10.12$**, so** $1.5s = 1.5(10.12) = 15.18$

$12 - 15.18 = -3.18$ **and** $12 + 15.18 = 27.18$

Nine of the 10 values in the data set fall between these two numbers, so the actual percentage of values that are within 1.5 standard deviations from the mean is 90%.