<u>Measures of Dispersion(variability or spread):</u>

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 variability, while data sets that are far apart have a larger variability.

Our textbook considers three methods for measuring variability.

IQR

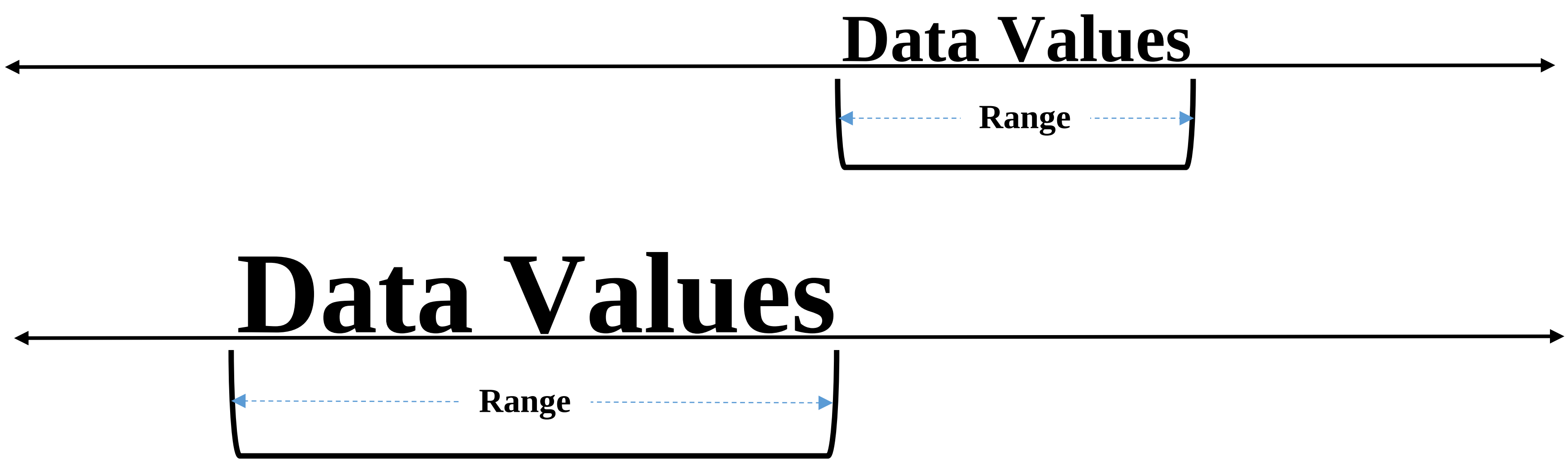
Range

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 = 9

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

Range = 6 – 2 = 4

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

The first data set has the larger range so it's considered to have more variability.

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!

Examples:

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

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

Associated with each data value, *x*, is its deviation from the mean, $x-\overline{x}$**. Our textbook generally assumes that the data set represents an entire population, so we'll be calculating a population variance.**

x	$\overline{x}$	$x-\overline{x}$
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?

The sum of the deviations will always equal zero, so their average would always be zero.

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

x	$\overline{x}$	$x-\overline{x}$	$\left(x-\overline{x}\right)^2$
1	**3**	-2	4
2	**3**	-1	1
3	**3**	0	0
6	**3**	3	9
	Total	0	14

The general formula for the population variance is

$$\sigma^2 = \frac{\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n} = \frac{\sum\left(x-\overline{x}\right)^2}{n} = \frac{14}{5} = \boxed{2.8}$$

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.

<u>Population Standard Deviation:</u>

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}} = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{14}{5}} = \boxed{1.67}$$

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

Population Variance, σ^2**:** $\frac{48}{4} = \boxed{12}$

Population Standard Deviation, σ**:** $\sqrt{\frac{48}{4}} = \boxed{3.46}$

By comparing standard deviations which data set has more variability: $\{1,2,3,6\}$ **or** $\{1,1,1,9\}$ **?**

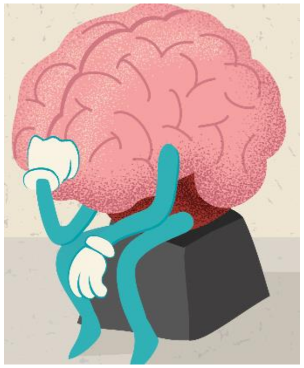
$\{1,1,1,9\}$ has the larger standard deviation, so it's considered to have more variability.

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

Set B



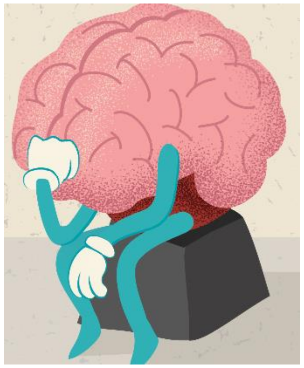
4. Without doing the calculations, which data set will have the smallest standard deviation?

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

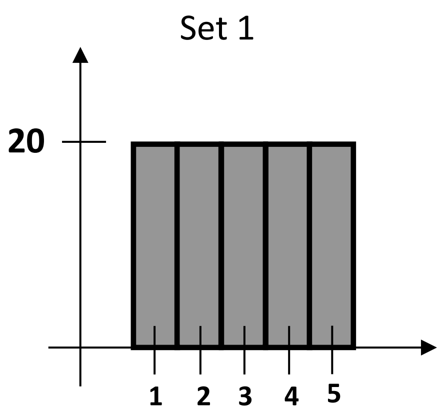
Set B: 33, 99, 100, 100, 100

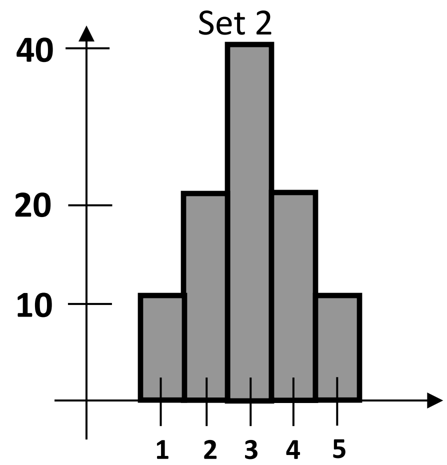
Set C: 54, 59, 61, 63, 64

Set C



5. **Given the following frequency histograms for the two data sets, Set 1 and Set 2. Determine which data set has the smaller standard deviation.**

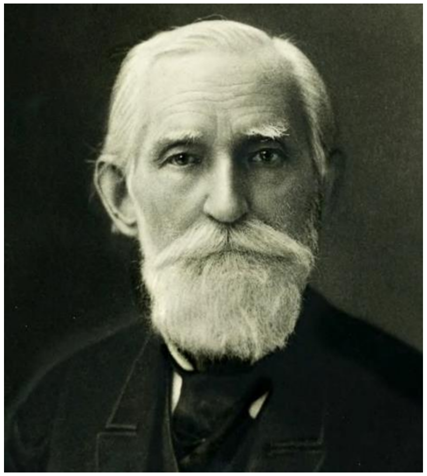




Set 2

Chebyshev's Theorem/Inequality:

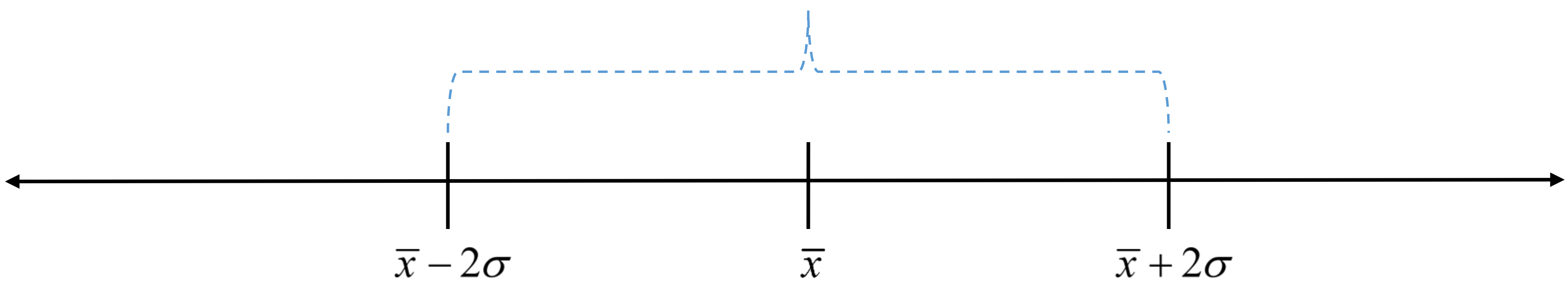
For any data set, the proportion of data values that are within *k* standard deviations from the mean is at least $1-\frac{1}{k^2}$.



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

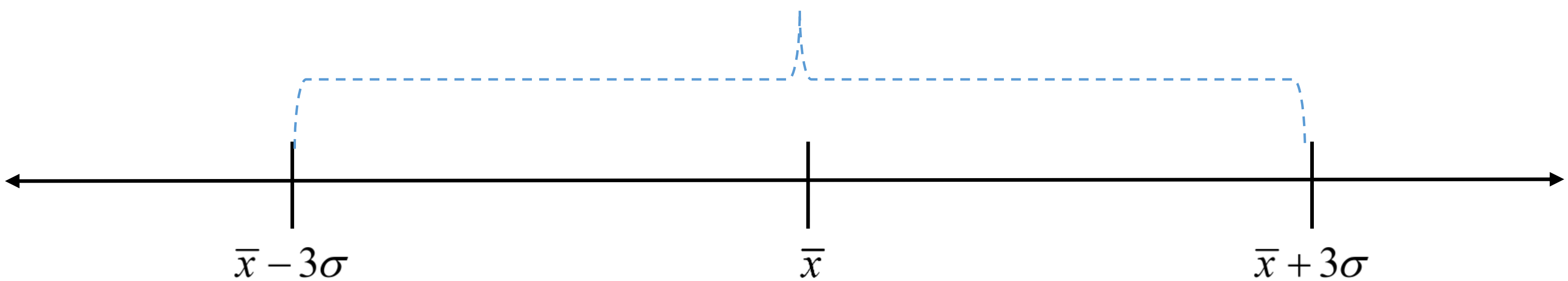
For $k = 2$, the proportion of data values within 2 standard deviations from the mean is at least $1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = .75 = 75\%$.

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 $1-\frac{1}{3^2}=1-\frac{1}{9}=\frac{8}{9}=.\overline{8}=88.\overline{8}\%$.

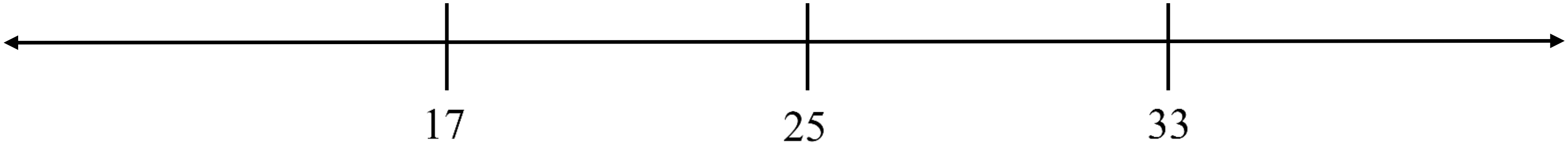
At least $88.\overline{8}\%$ of the data values are between these two numbers.



Example:

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?



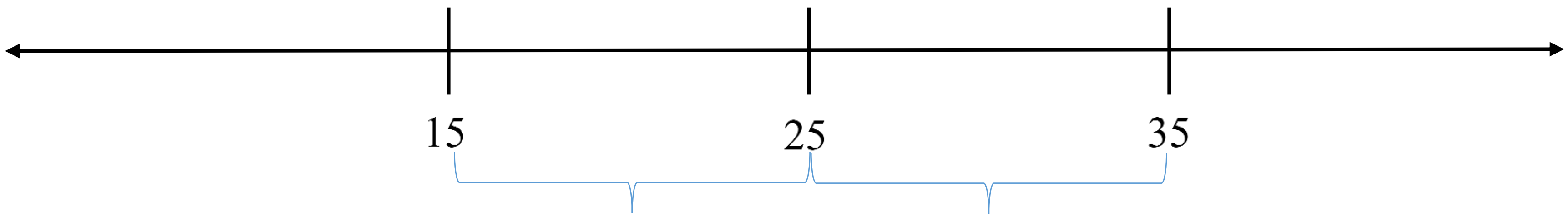
$25-17=8$ $\qquad$ $33-25=8$

$8\div 4=2$ $\qquad$ $8\div 4=2$

$So\ k=2$ $\qquad$ $So\ k=2$

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

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



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

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

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

$$1-\frac{1}{(2.5)^2}=1-.16=.84=\boxed{84\%}$$

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?

$1.5(10.12)=15.18$, $12-15.18=-3.18$ and $12+15.18=27.18$. 9 of the 10 values lie between the two outlier boundaries, so it's $\boxed{90\%}$.

<u>Percentiles:</u>

Percentiles are numbers that divide up a data set. An n^{th} percentile is a value that roughly divides the lower n% of the data values from the upper (100 – n)% of the data values. For example, a 5th percentile would divide the lower 5% of the data values from the upper 95% of the data values.

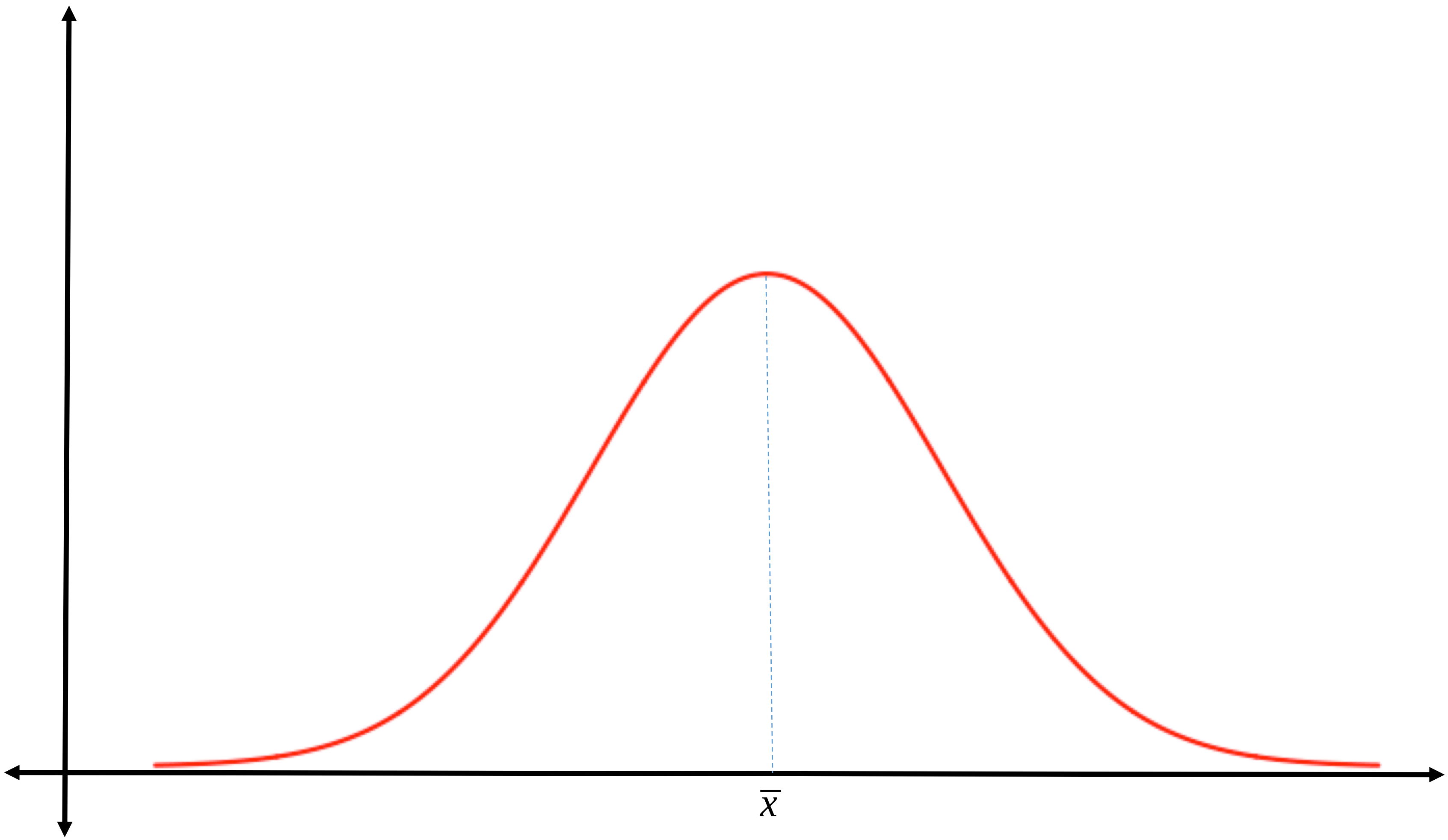
A number is considered to be in the n^{th} percentile of a data set if it's greater than or equal to n% of the data values.

Example:

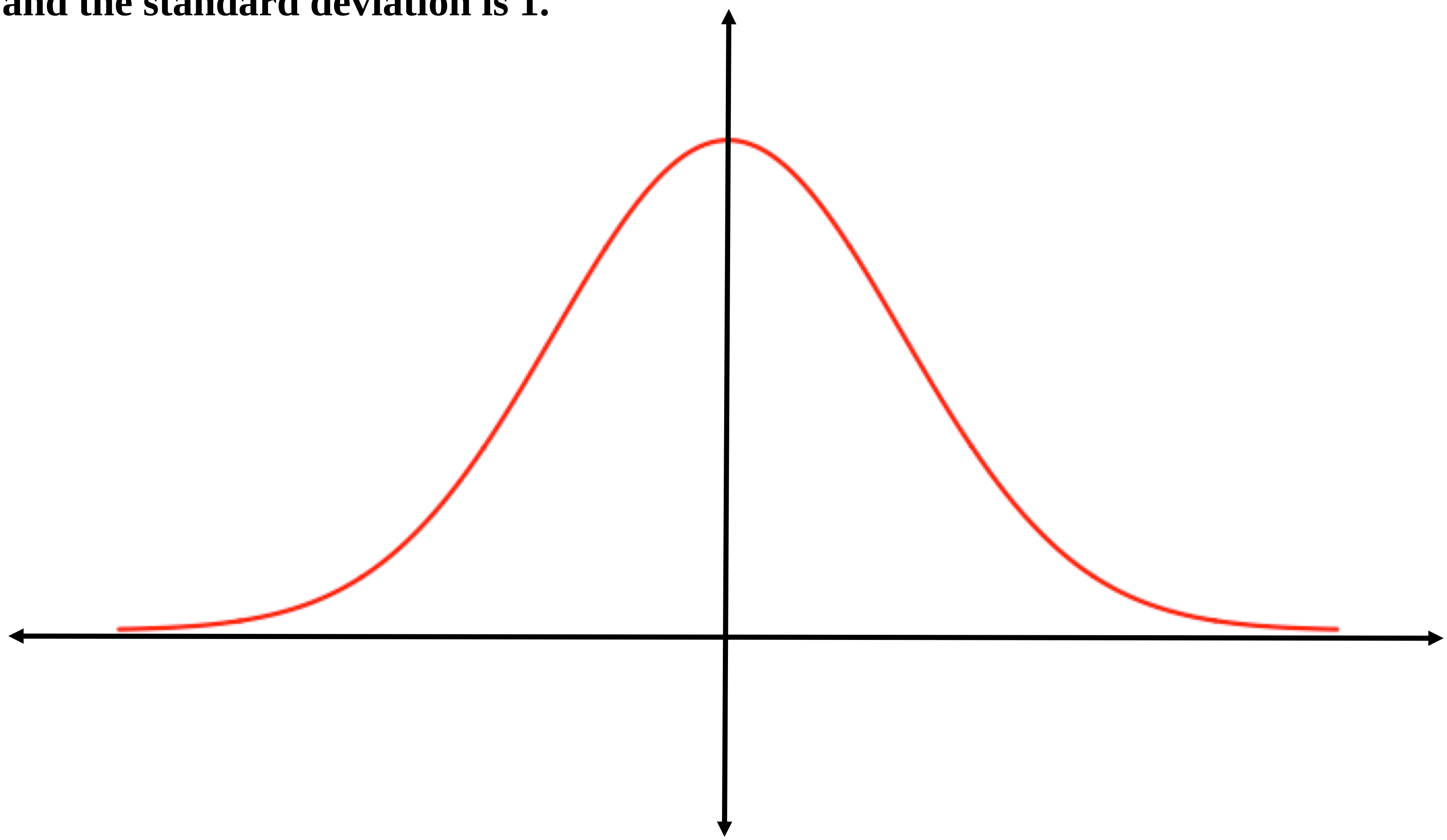
$\{1,5,8,2,6,9,11,8\}$

What percentile is 6 in? 6 is greater than or equal to 4 of the 8 values, so 6 is in the 50th percentile.

For large data sets, the percentages of values within given ranges can be determined from a graph called a distribution. In many common situations- heights, weights, test scores,…, the distribution has a bell shape and is called a normal distribution.



The curve is symmetric about the mean value, $\bar{x}$, and has its peak value there as well. The height of the peak depends on the value of the standard deviation, σ. The smaller the value of the standard deviation, the higher the peak. There is a special normal distribution called the standard normal distribution where the mean value is 0, and the standard deviation is 1.



The percentiles for the standard normal are well known and given in a table in the textbook on page 454. The values of a data set with a standard normal distribution are generally referred to as standard values or z-scores and just abbreviated by the letter z.

The good news is that information about any data set with a normal distribution can be determined by first converting the original data value into its equivalent standard value or z-score.

$$z = \frac{x - \bar{x}}{\sigma} \left\{ \frac{x - \text{mean}}{\text{standard deviation}} \right\}$$

Examples:

1. **The data value 11 comes from a normally distributed data set with a mean of 3 and a standard deviation of 4. What's the z-score of 11?**

$$z = \frac{11 - 3}{4} = \boxed{2}$$

2. Steve took the ACT and got a score of 22. If the mean of the test scores is 21.1 and the standard deviation is 5.2, then

a) What's the standard value rounded to three places?

$$z = \frac{22 - 21.1}{5.2} = \frac{.9}{5.2} = .1730769... = \boxed{.173}$$

b) What percentile is he in?

Check out the table.

.173 falls between the z-score values of .151 and .176 in the table which means that 22 is between the 56th and 57th percentiles. Since .173 isn't greater than or equal to .176, the best we can say from the table is that 22 is in the 56th percentile.

Percentile	z-score
1	-2.326
2	-2.054
3	-1.881
4	-1.751
5	-1.645
6	-1.555
7	-1.476
8	-1.405
9	-1.341
10	-1.282
11	-1.227
12	-1.175
13	-1.126
14	-1.080
15	-1.036
16	-0.994
17	-0.954
18	-0.915
19	-0.878
20	-0.842
21	-0.806
22	-0.772
23	-0.739
24	-0.706
25	-0.674
26	-0.643
27	-0.613
28	-0.583
29	-0.553
30	-0.524
31	-0.496
32	-0.468
33	-0.440

Percentile	z-score
34	-0.412
35	-0.385
36	-0.358
37	-0.332
38	-0.305
39	-0.279
40	-0.253
41	-0.228
42	-0.202
43	-0.176
44	-0.151
45	-0.126
46	-0.100
47	-0.075
48	-0.050
49	-0.025
50	0.000
51	0.025
52	0.050
53	0.075
54	0.100
55	0.126
56	0.151
57	0.176
58	0.202
59	0.228
60	0.253
61	0.279
62	0.305
63	0.332
64	0.358
65	0.385
66	0.412

Percentile	z-score
67	0.440
68	0.468
69	0.496
70	0.524
71	0.553
72	0.583
73	0.613
74	0.643
75	0.674
76	0.706
77	0.739
78	0.772
79	0.806
80	0.842
81	0.878
82	0.915
83	0.954
84	0.994
85	1.036
86	1.080
87	1.126
88	1.175
89	1.227
90	1.282
91	1.341
92	1.405
93	1.476
94	1.555
95	1.645
96	1.751
97	1.881
98	2.054
99	2.326

c) **What percentage of all the students who took the test had a better score than his?**

Approximately 44% had a better score.

d) **If 10,000 students took the test, about how many of them had a better score than his?**

44% of 10,000, or 4,400.