USING STATISTICS TO SUMMARIZE DATA SETS

GONZALO G. PERAZA MUES

NUMERICAL QUANTITIES COMPUTED FROM A DATA SET ARE CALLED **STATISTICS**.

MEASURES OF CENTER

SAMPLE MEAN

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

EXAMPLE

The average fuel efficiencies, in miles per gallon, of cars sold in the United States in the years 1999 to 2003 were

28.2, 28.3, 28.4, 28.5, 29.0

If each data value is increased by a constant amount c, then this causes the sample mean also to be increased by c.

$$y_i = x_i + c$$

Then

$$\bar{y} = \bar{x} + c$$

EXAMPLE

The winning scores in the U.S. Masters Golf Tournament in the years from 1981 to 1990 were as follows:

280, 284, 280, 277, 282, 279, 285, 281, 283, 278

If each data value is multiplied by c, then so is the sample mean.

$$y_i = cx_i$$

Then

$$\bar{y} = c\bar{x}$$

WEIGHTED AVERAGE

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \sum_{i=1}^{n} w_i x_i$$

Where

$$w_i = \frac{f_i}{\sum_{i=1}^{n} f_i} = \frac{f_i}{n}$$

EXAMPLE

Age of members of a symphony orchestra for young adults.

Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

The deviations are the differences between the data values and the sample mean.

$$x_i - \bar{x}$$

The following identity is always true

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

EXAMPLE

Number of weeks after completion of a learn-to- drive course that it took a sample of seven people to obtain a driver's license.

2, 110, 5, 7, 6, 7, 3

Its value is greatly affected by extreme data values.

SAMPLE MEDIAN

Order the values of a data set of size n from smallest to largest. If n is odd, the sample median is the value in position $\frac{n+1}{2}$; if n is even, it is the average of the values in positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

EXAMPLE

Number of weeks after completion of a learn-to- drive course that it took a sample of seven people to obtain a driver's license.

2, 110, 5, 7, 6, 7, 3

EXAMPLES

Number of days it took 6 individuals to quit smoking after completing a course designed for this purpose.

1, 2, 3, 5, 8, 100

The sample mean makes use of all the data values.

The sample median is not affected by extreme values.

For roughly symmetric data sets the sample mean and sample median will have values close to each other.

EXAMPLE

4, 6, 8, 8, 9, 12, 15, 17, 19, 20, 22

EXAMPLE

A group of 5-week-old mice were each given a radiation dose of 300 rad.

Germ-Free Mice

1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

Conventional Mice

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

SAMPLES MODE

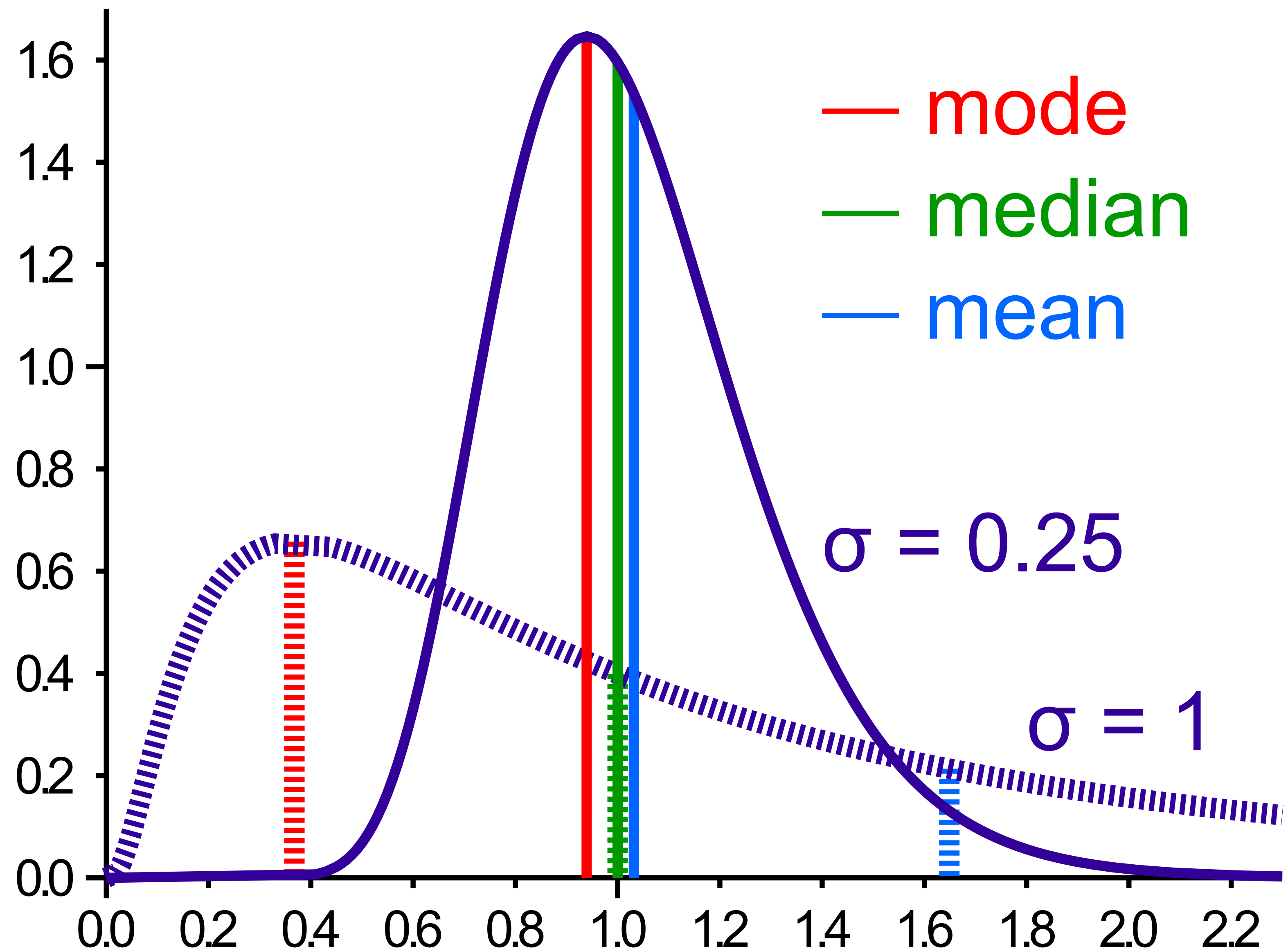
The data value that occurs with the greatest frequency.

EXAMPLES

40 roles of a dice.

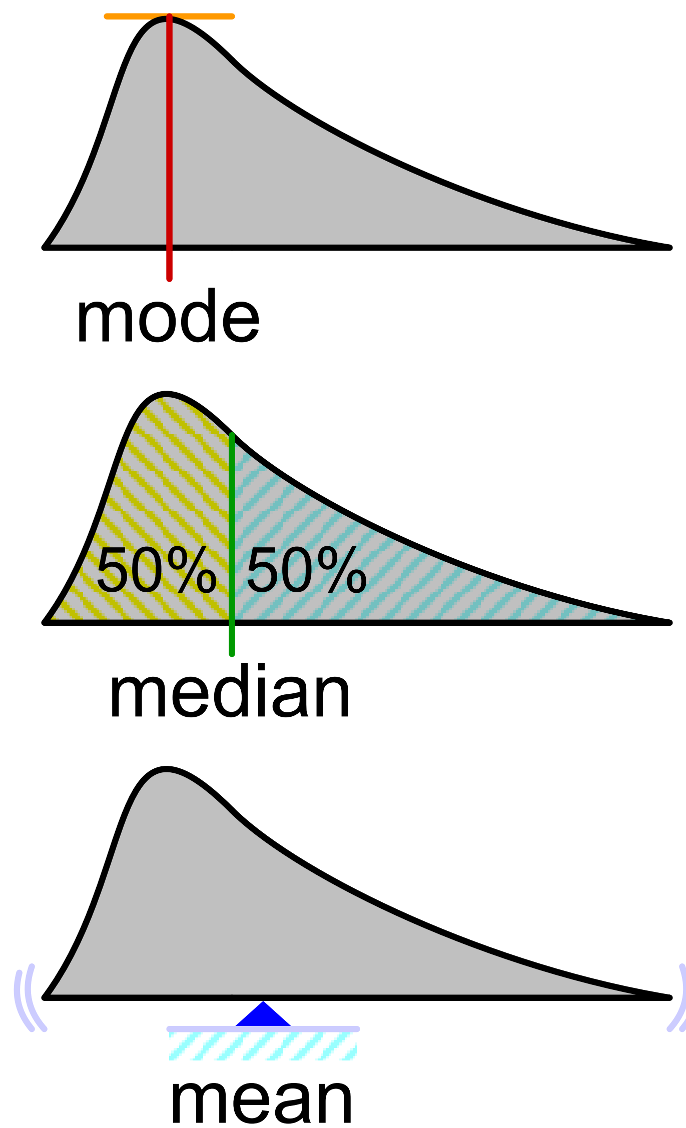
Value	Frequency
1	9
2	8
3	5
4	5
5	6
6	7

If no single value occurs most frequently, then all the values that occur at the highest frequency are called modal values.



mode
median
mean
σ = 0.25
σ = 1
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2

8.1


mode
50%
50%
median
mean

MEASURES OF SPREAD

RANGE

Difference between the largest and smallest values of a data distribution.

EXAMPLE

Apple weight values

5, 6, 7, 7, 6, 8, 6, 9, 10, 8

ROOT MEAN SQUARED (R.M.S.)

Provides an idea of the size of the values.

$$r.m.s. = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

EXAMPLE

How small or big are this values?

0, 5, -8, 7, -3

SAMPLES VARIANCE

Describe the spread or variability of the data values.

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

EXAMPLE

A: 3, 4, 6, 7, 10 B: -20, 5, 15, 24

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

The sample variance remains unchanged when a constant is added to each data value.

$$y_i = x_i + c$$

Then

$$y_i - \bar{y} = (x_i + c) - (\bar{x} + c) = x_i - \bar{x}$$

If each data value is multiplied by a constant c then the sample variance of the new data is the sample variance of the old data multiplied by c^2.

$$y_i = cx_i$$

Then

$$s_y^2 = c^2 s_x^2$$

EXAMPLE

Worldwide number of fatal airline accidents in the years from 1997 to 2005.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005
Accidents	25	20	21	18	13	13	7	9	18

SAMPLE STANDARD DEVIATION

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

$$y_i = cx_i$$

Then

$$s_y = |c|\, s_x$$

MEAN ABSOLUTE DEVIATION

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

SD VS MAD

- SD > MAD
- MAD is easier to understand
- SD is easier to work with
- SD is more widely used

Read: Revisiting a 90-year-old debate: the advantages of the mean deviation by Stephen Gorard

SAMPLE PERCENTILES

The sample $100p$ percentile is that data value such that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent are greater than or equal to it. If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these two values.

To find the sample $100p$ percentile of a data set of size n

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample $100p$ percentile.
3. If np is an integer, then the average of the values in positions np and $np + 1$ is the sample 100p percentile.

EXAMPLE

Which data value is the sample 90th percentile when the sample size is (a) 8, (b) 16, and (c) 100?

QUARTILES

- The sample 25th percentile is called the first quartile.
- The sample 50th percentile is called the median or the second quartile.
- The sample 75th percentile is called the third quartile.

The quartiles break up a data set into four parts.

EXAMPLE

36 noise levels outside of Grand Central Station in Manhattan in dB.

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

INTERQUARTILE RANGE

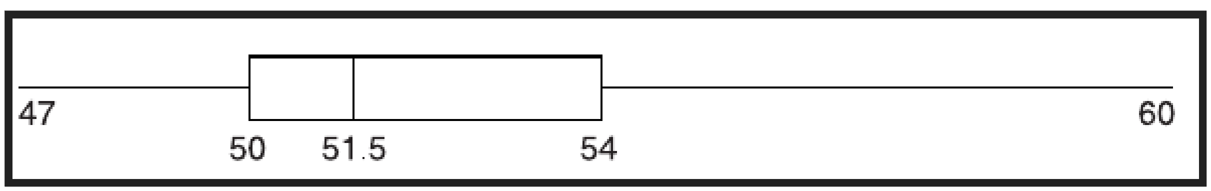
3rd quartile - 1st quartile

Length of the interval in which the middle half of the data values lie.

EXAMPLE

Starting salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

BOX PLOT



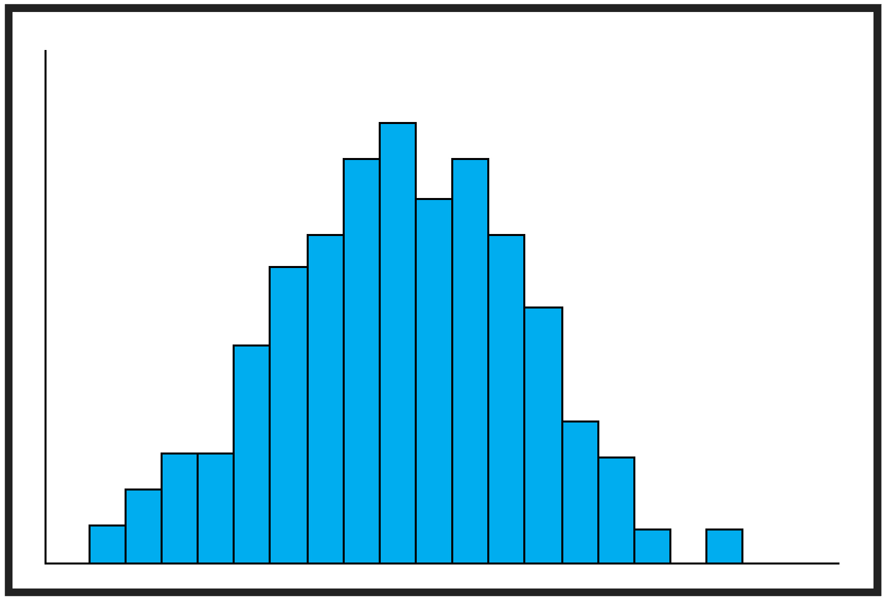
NORMAL DATA SETS

DEFINITION

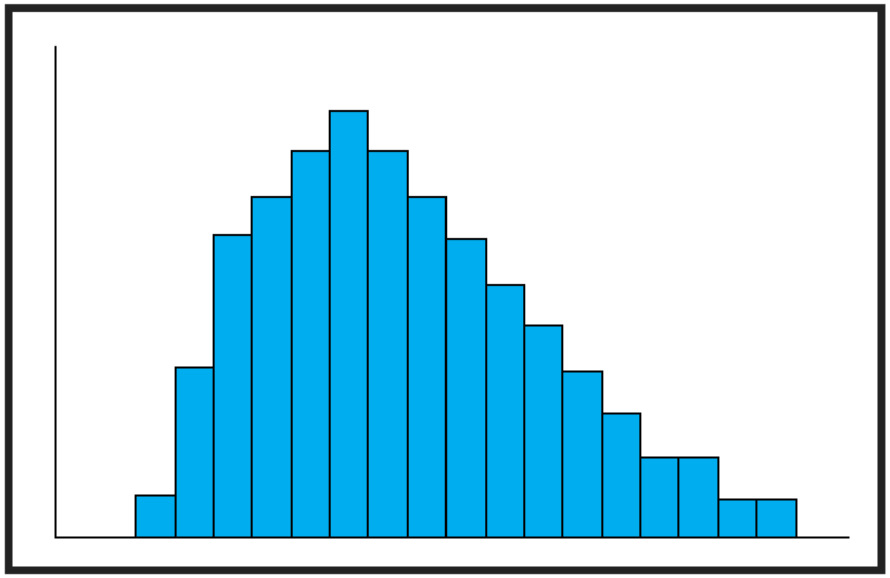
A data set is said to be normal if a histogram describing it has the following properties:

- It is highest at the middle interval.
- Moving from the middle interval in either direction, the height decreases in such a way that the entire histogram is bell-shaped.
- The histogram is symmetric about its middle interval.

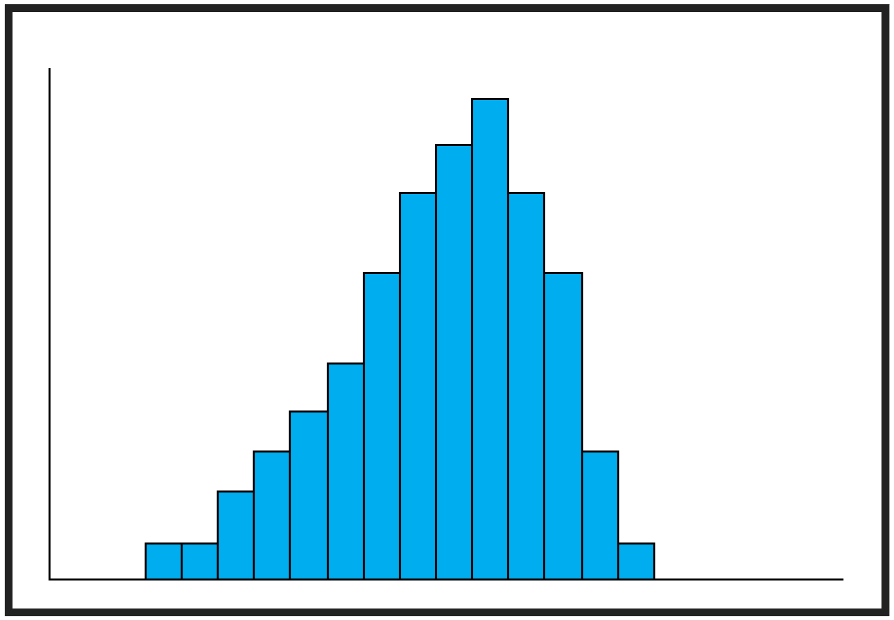
NORMAL DATA SET



SKEWED TO THE RIGHT



SKEWED TO THE LEFT



EMPIRICAL RULE

- Approximately 68% of the observations lie within $\bar{x} \pm s$
- Approximately 95% of the observations lie within $\bar{x} \pm 2s$
- Approximately 99.7% of the observations lie within $\bar{x} \pm 3s$

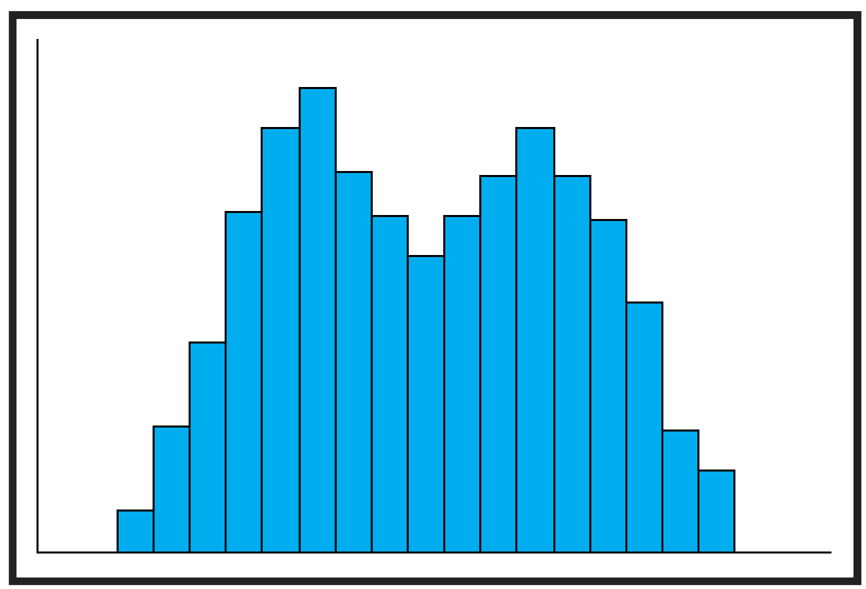
EXAMPLE

Scores on a statistics exam

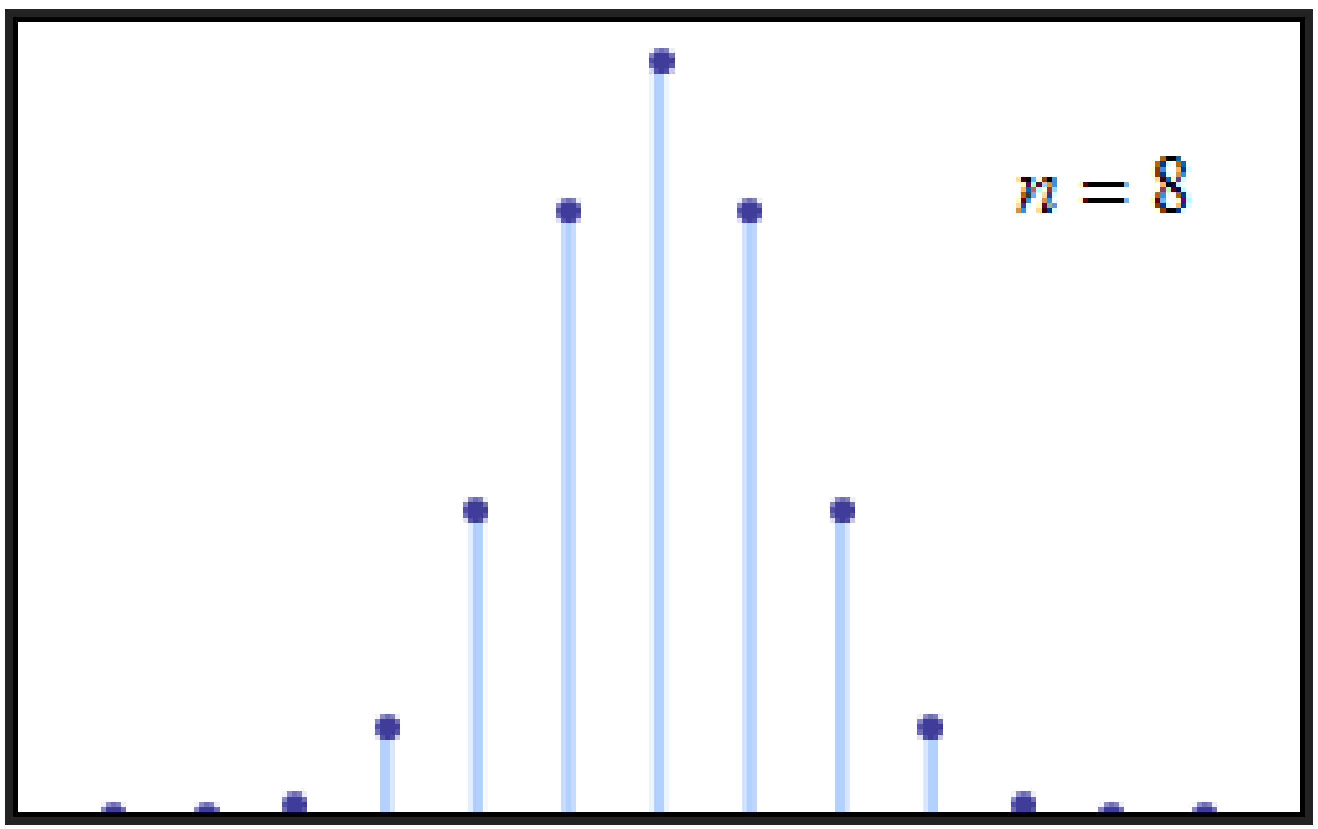
9	0, 1, 4
8	3, 5, 5, 7, 8
7	2, 4, 4, 5, 7, 7, 8
6	0, 2, 3, 4, 6, 6
5	2, 5, 5, 6, 8
4	3, 6

BIMODAL DATA SETS

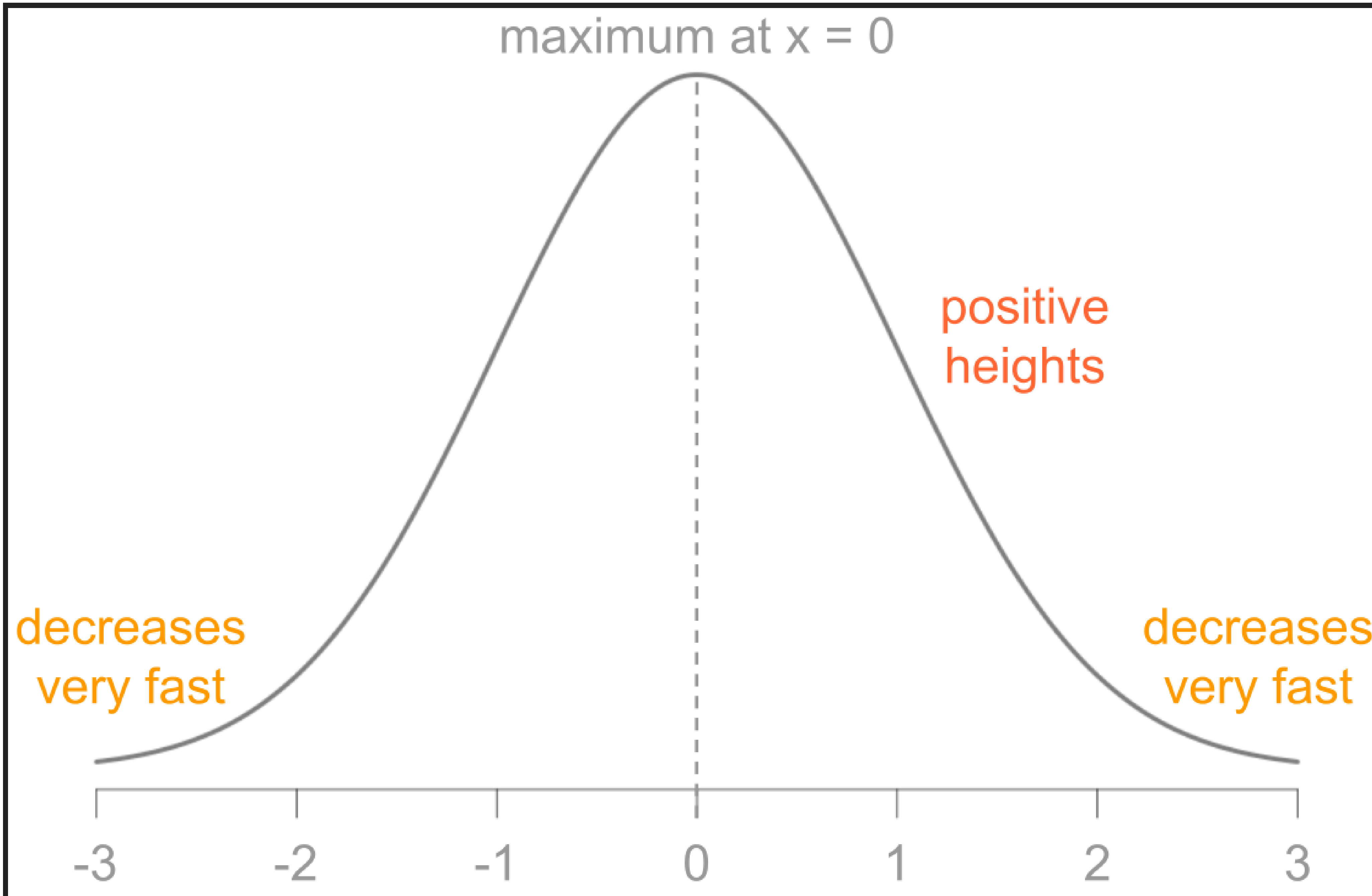
A data set that is obtained by sampling from a population that is itself made up of subpopulations.

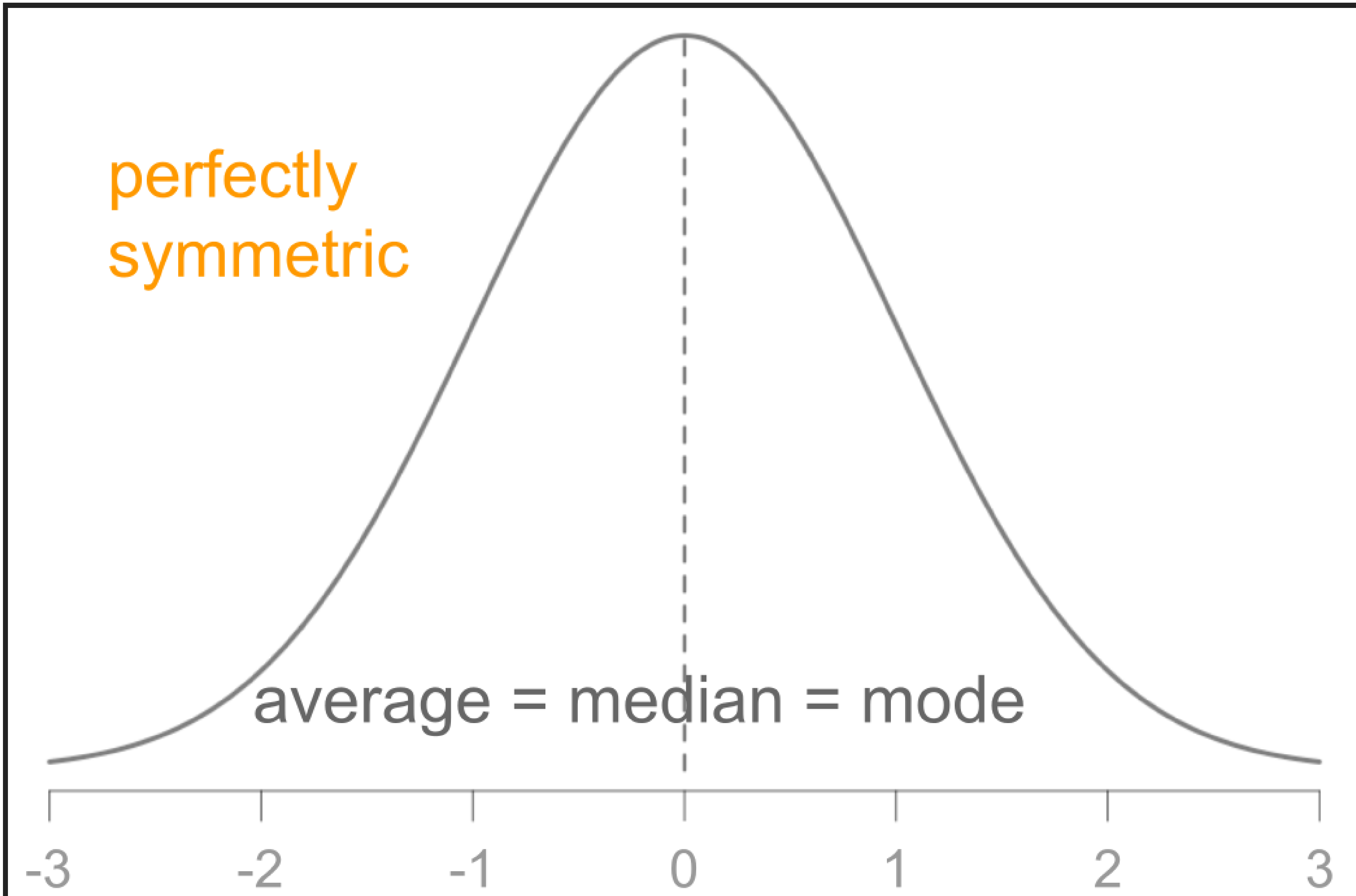


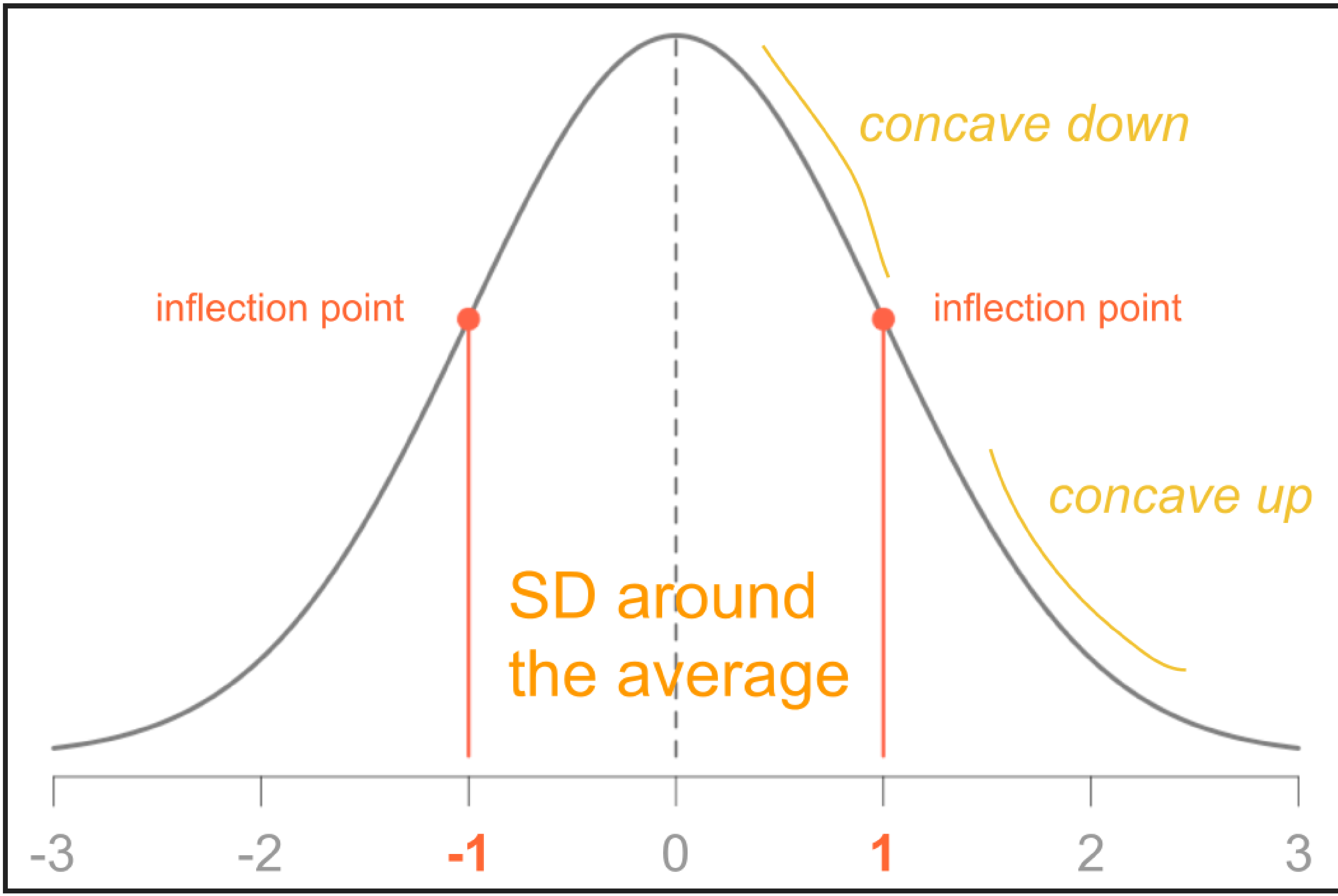
THE NORMAL CURVE



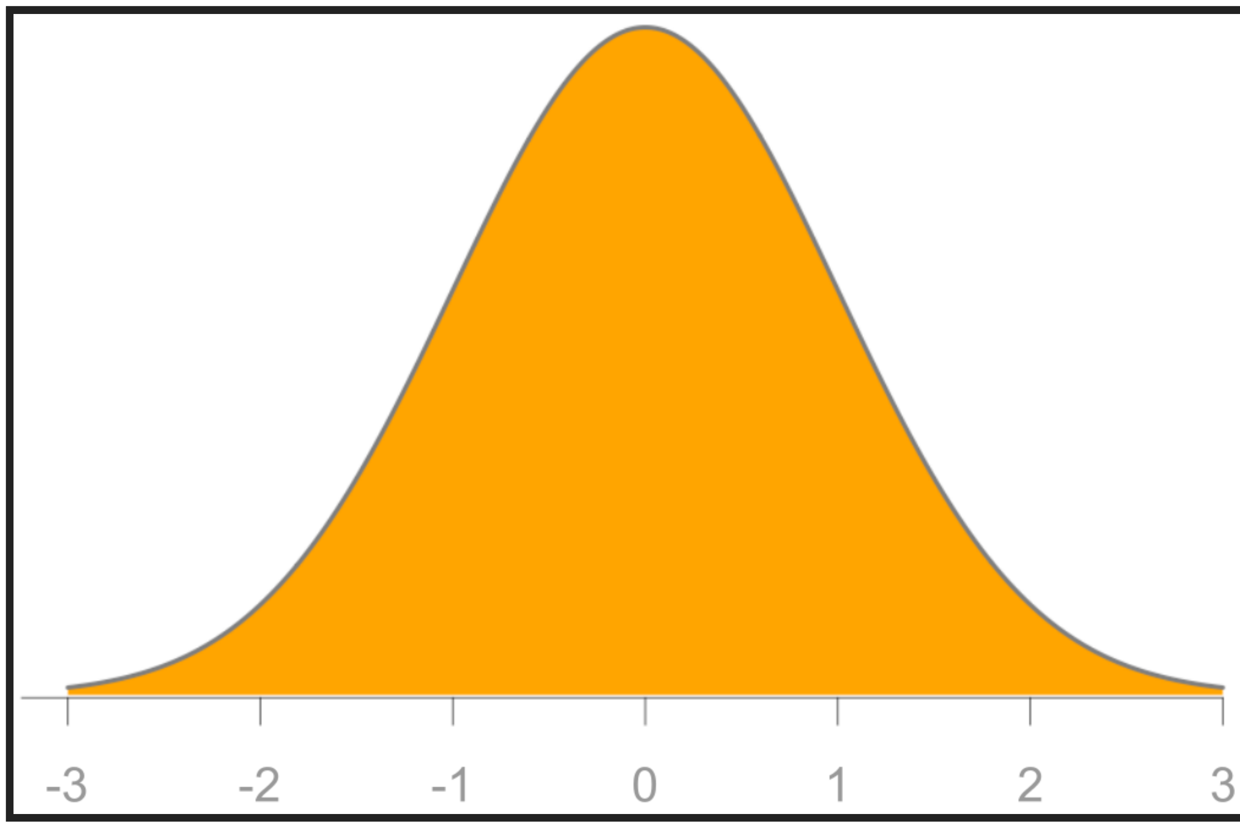
$$y = \frac{1}{2\pi} e^{-\frac{x^2}{2}}$$


maximum at x = 0
positive heights
decreases very fast
decreases very fast
-3
-2
-1
0
1
2
3


perfectly
symmetric
average = median = mode
-3
-2
-1
0
1
2
3

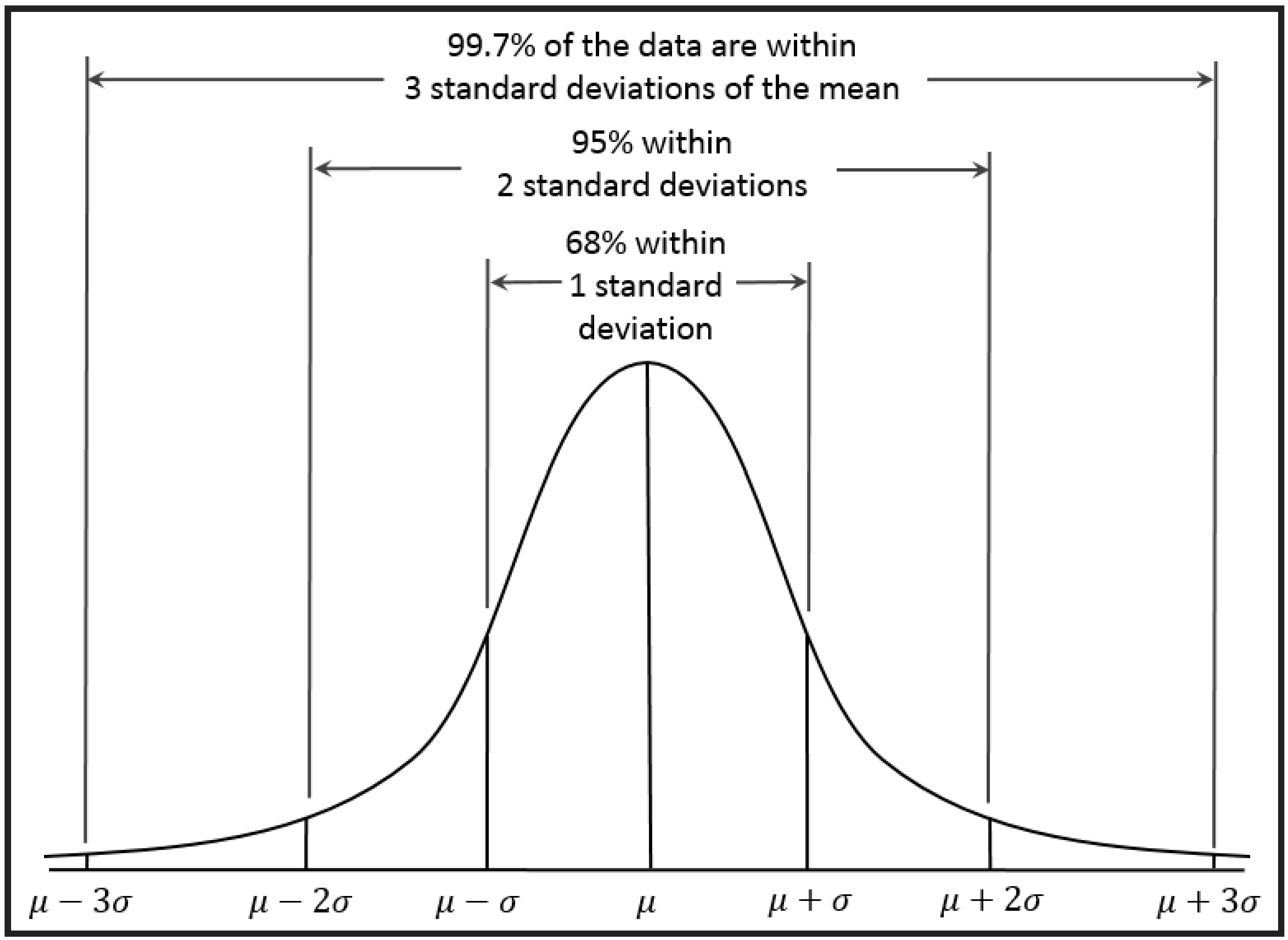

concave down
inflection point
inflection point
concave up
SD around
the average
-3
-2
-1
0
1
2
3

TOTAL AREA UNDER THE CURVE = 1



$$\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{x^2}{2}} dx = 1$$

EMPIRICAL RULE

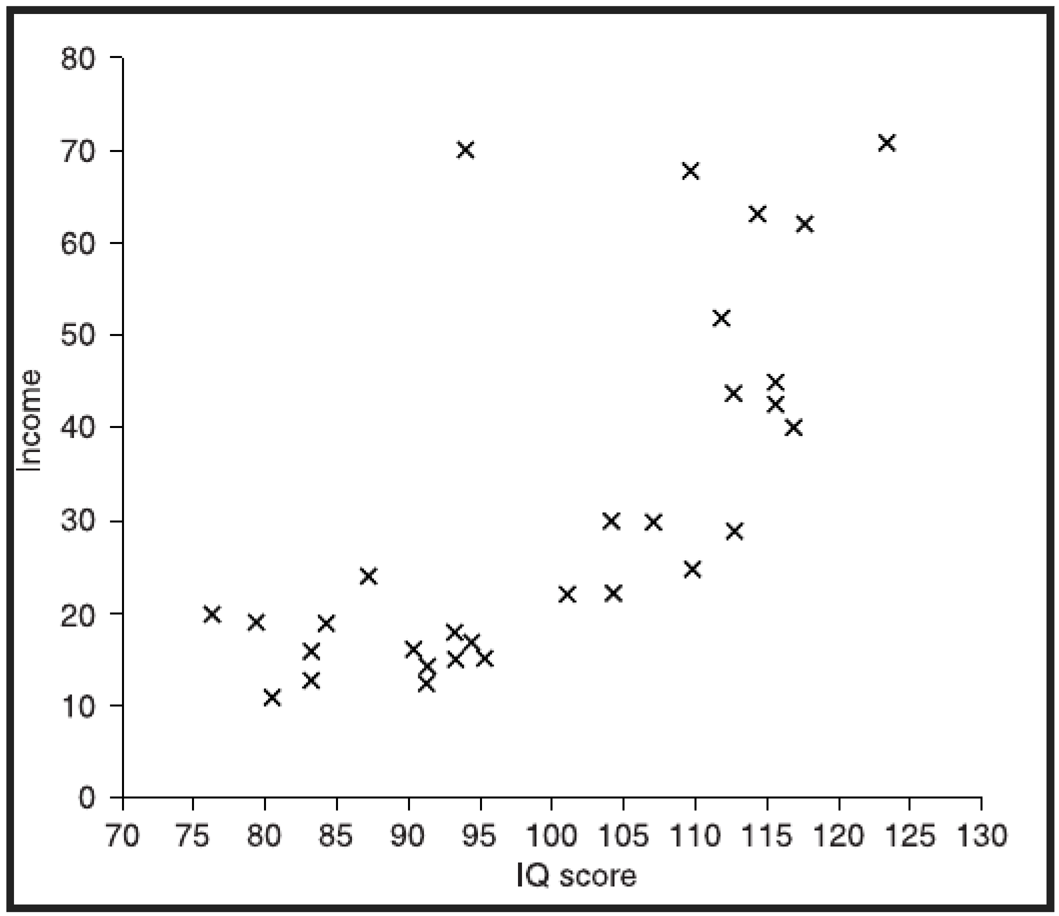


SETS OF PAIRED DATA

$$(x_i, y_i)$$

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

SCATTER DIAGRAM



$$r = 0.4189$$

SAMPLE CORRELATION COEFFICIENT

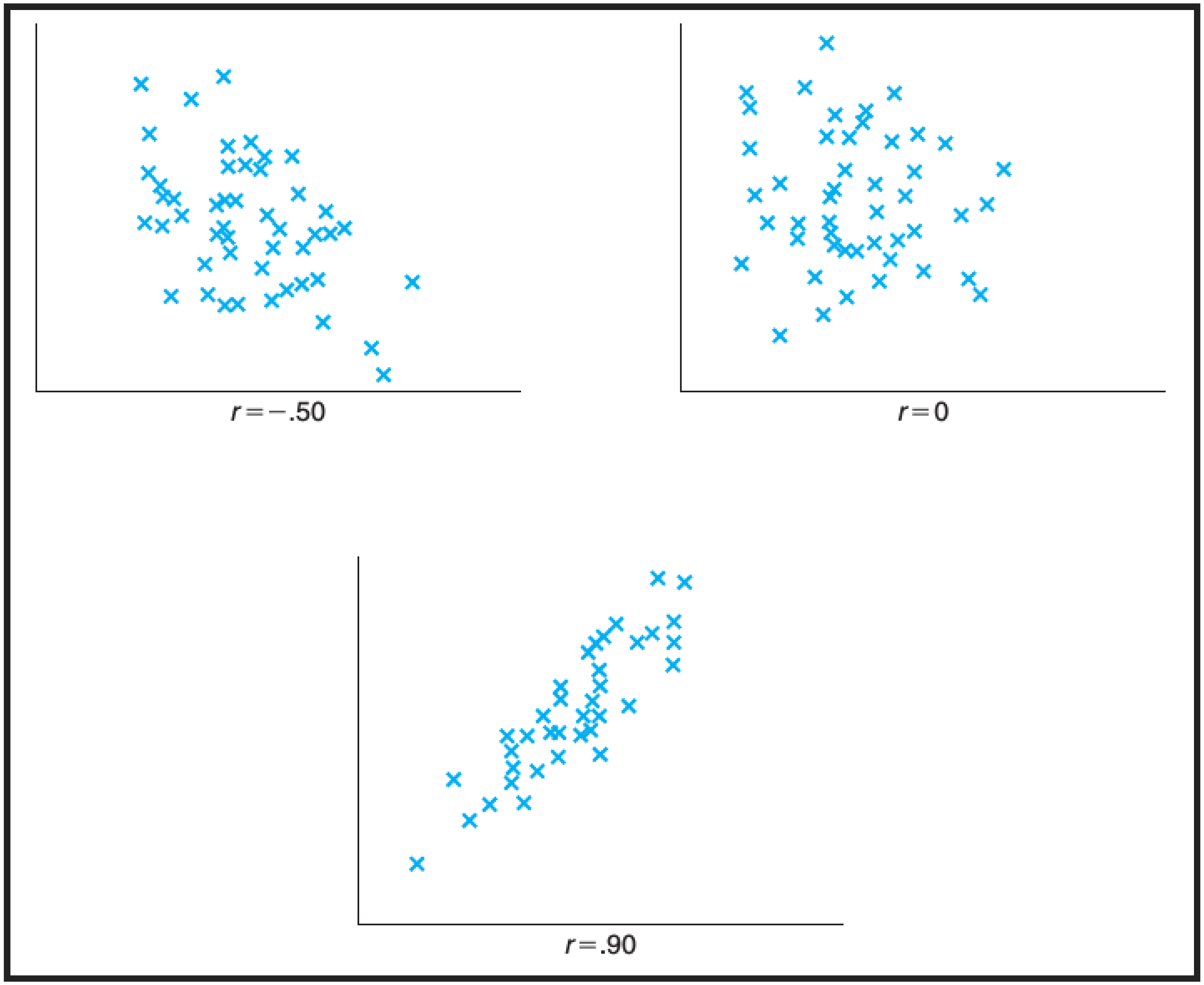
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

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

PROPERTIES OF R

- $-1 \leq r \leq 1$
- If for constants a and b, $y_i = a + bx_i$ then:
  - $r = 1$ if $b > 0$
  - $r = -1$ if $b < 0$
- $r(x_i, y_i) = r(a + bx_i, c + dy_i)$ if $sign(b) = sign(d)$

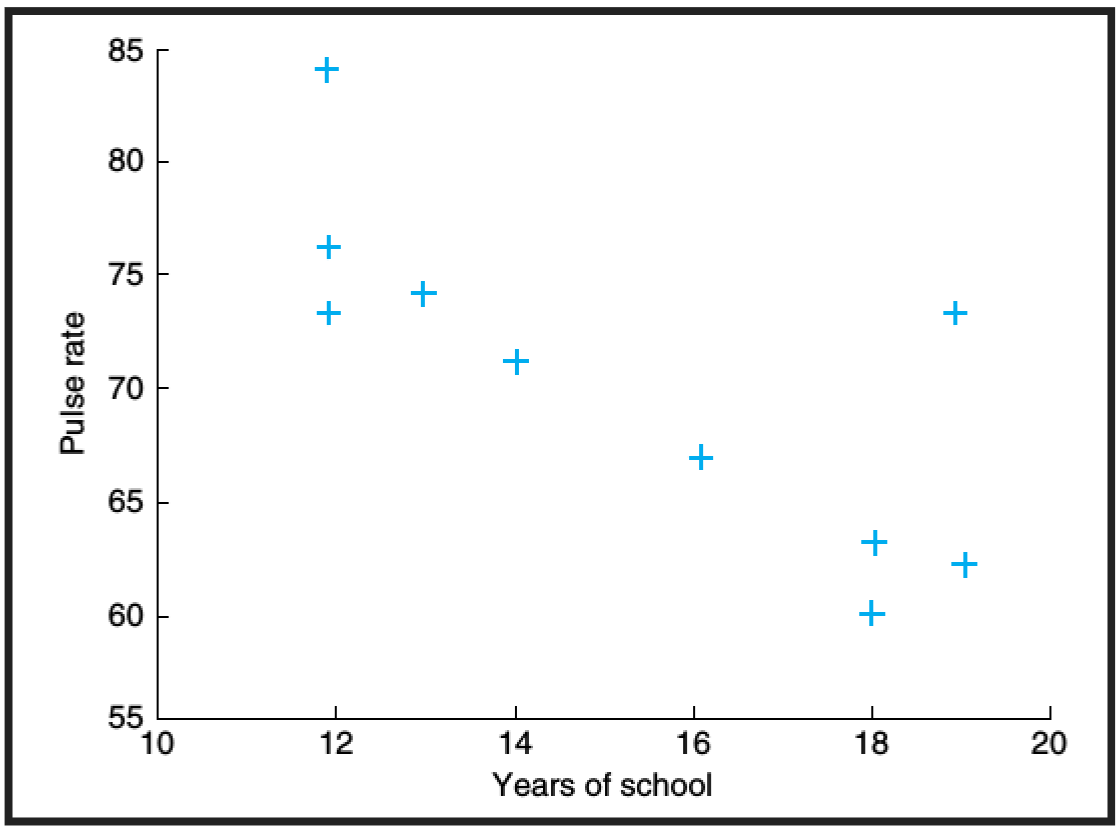
- $|r| = 1$ means a perfect linear relationship between data
- $|r| < 1$ gives the strength of the correlation
- $sign(r)$ gives the direction of the correlation.


$r = -.50$
$r = 0$
$r = .90$

EXAMPLE

Resting pulse rates (in beats per minute) and the years of schooling of 10 individuals. ($r = -0.7639$)

Person	1	2	3	4	5	6	7	8	9	10
Years of School	12	16	13	18	19	12	18	19	12	14
Pulse Rate	73	67	74	63	73	84	60	62	76	71



PROOF

- Start with $\sum \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0$

CORRELATION MEASURES ASSOCIATION, NOT CAUSATION

Often, the explanation for such an association lies with an unexpressed factor that is related to both variables under consideration.