Introduction to Probability and Statistics Fourteenth Edition

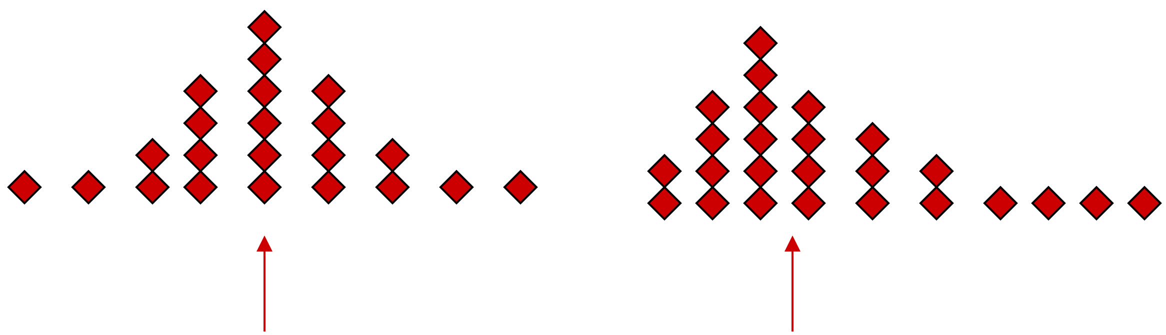
Chapter 2
Describing Data with Numerical Measures

Describing Data with Numerical Measures

- Graphical methods may not always be sufficient for describing data.
- **Numerical measures** can be created for both **populations** and **samples.**
  - A **parameter** is a numerical descriptive measure calculated for a population.
  - A **statistic** is a numerical descriptive measure calculated for a sample.

Measures of Center

- A measure along the horizontal axis of the data distribution that locates the center of the distribution.



Arithmetic Mean or Average

- The **mean** of a set of measurements is the sum of the measurements divided by the total number of measurements.

$$\boxed{\bar{x} = \frac{\sum x_i}{n}}$$

where $n =$ number of measurements

$\sum x_i$ = sum of all the measurements

Example

- The set: 2, 9, 1, 5, 6

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2+9+11+5+6}{5} = \frac{33}{5} = 6.6$$

If we were able to enumerate the whole population, the **population mean** would be called **μ** (the Greek letter “mu”).

Median

- The **median** of a set of measurements is the middle measurement when the measurements are ranked from smallest to largest.
- The **position of the median** is

$$\boxed{.5(n + 1)}$$

once the measurements have been ordered.

Example

- The set: $2, 4, 9, 8, 6, 5, 3 \quad n = 7$
- Sort: $2, 3, 4, 5, 6, 8, 9$
- Position: $.5(n + 1) = .5(7 + 1) = 4^{\text{th}}$

Median = 4^{th} largest measurement

- The set: 2, 4, 9, 8, 6, 5 $\quad n = 6$
- Sort: 2, 4, 5, 6, 8, 9
- Position: $.5(n + 1) = .5(6 + 1) = 3.5^{\text{th}}$

Median = (5 + 6)/2 = 5.5 — average of the 3^{rd} and 4^{th} measurements

Mode

- The **mode** is the measurement which occurs most frequently.
- The set: 2, 4, 9, 8, 8, 5, 3
  - The mode is **8**, which occurs twice
- The set: 2, 2, 9, 8, 8, 5, 3
  - There are two modes—**8** and **2** (bimodal)
- The set: 2, 4, 9, 8, 5, 3
  - There is no mode (each value is unique).

EXAMPLE

The number of quarts of milk purchased by 25 households:

0 0 1 1 1 1 1 2 2 2 2 2 2
2 2 2 3 3 3 3 3 4 4 4 5

- **Mean?**

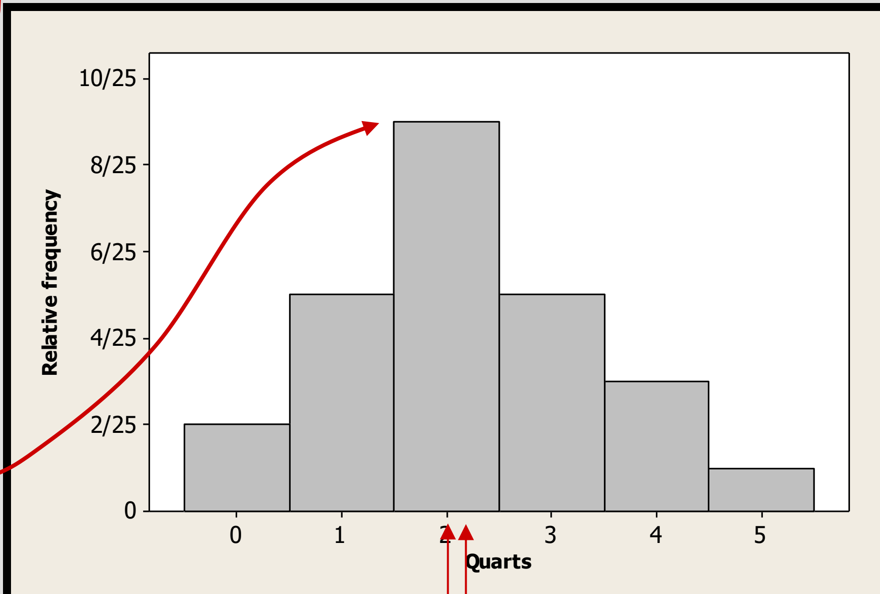
$$\boxed{\bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2}$$

- **Median?**

$$\boxed{m = 2}$$

- **Mode? (Highest peak)**

$$\boxed{\text{mode} = 2}$$

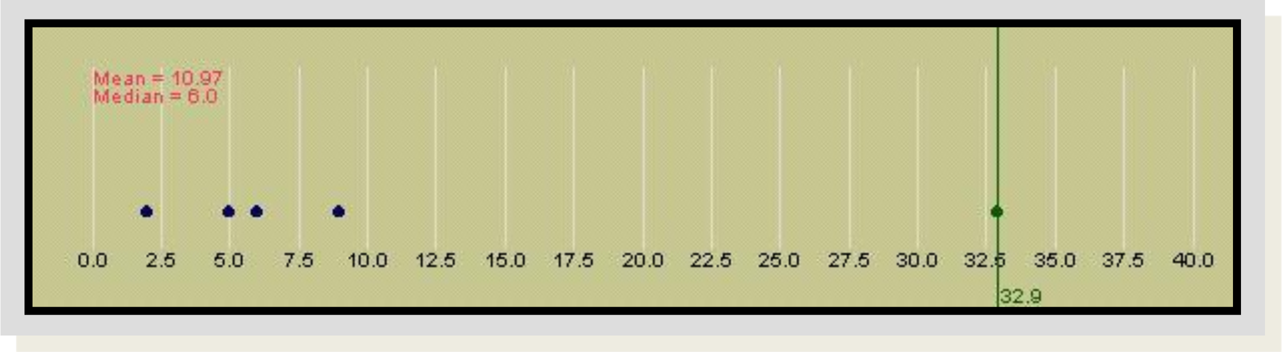


Frequency Table For Quarts of Milk Purchased

Quarts of Milk	Freq.
0	2
1	5
2	9
3	5
4	3
5	1
Total	25

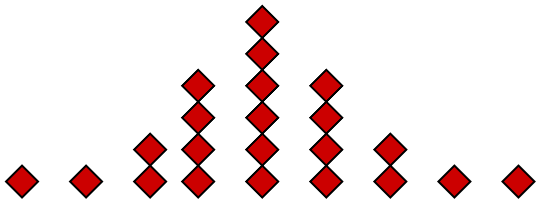
Extreme Values

- The mean is more easily affected by extremely large or small values than the median.

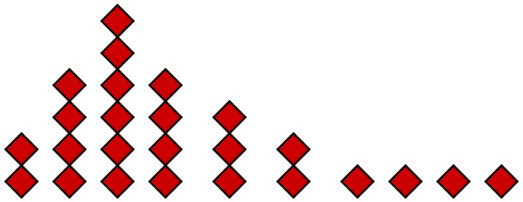


- The median is often used as a measure of center when the distribution is skewed.

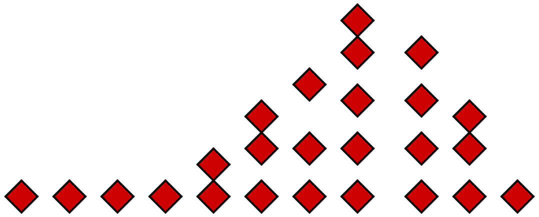
Extreme Values



Symmetric: Mean = Median



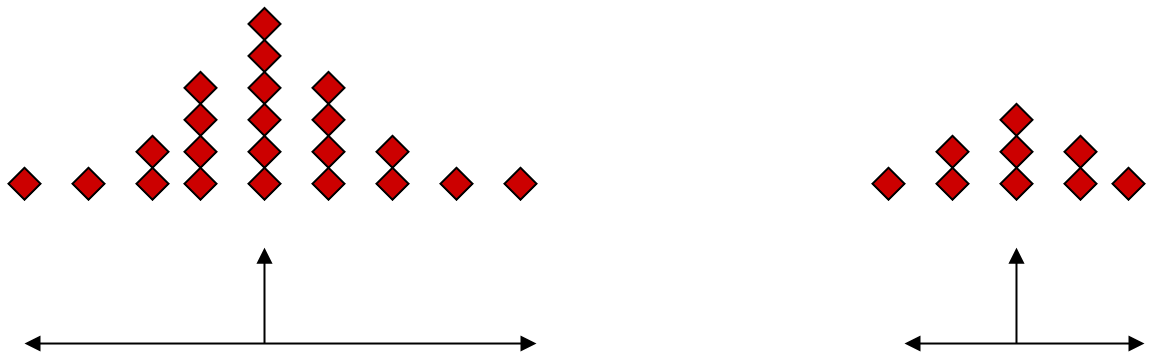
Skewed right: Mean > Median



Skewed left: Mean < Median

Measures of Variability

- A measure along the horizontal axis of the data distribution that describes the **spread** of the distribution from the center.



The Range

- The **range, R,** of a set of *n* measurements is the difference between the largest and smallest measurements.
- **Example:** A botanist records the number of petals on 5 flowers:

5, 12, 6, 8, 14

- The range is

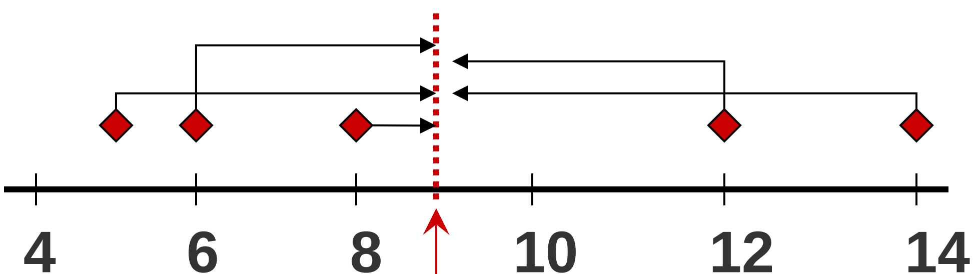
R = 14 – 5 = 9.

•Quick and easy, but only uses 2 of the 5 measurements.

The Variance

- The **variance** is measure of variability that uses all the measurements. It measures the average deviation of the measurements about their mean.
- **Flower petals: 5, 12, 6, 8, 14**

$$\bar{x} = \frac{45}{5} = 9$$



The Variance

- The **variance of a population** of N measurements is the average of the squared deviations of the measurements about their mean μ.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

- The **variance of a sample** of n measurements is the sum of the squared deviations of the measurements about their mean, divided by $(n - 1)$.

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

The Standard Deviation

- In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements.
- To return this measure of variability to the original units of measure, we calculate the **standard deviation**, the positive square root of the variance.

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

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

Two Ways to Calculate the Sample Variance

Use the Definition Formula:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	-4	16
12	3	9
6	-3	9
8	-1	1
14	5	25
Sum 45	0	60

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

$$= \frac{60}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

Two Ways to Calculate the Sample Variance

Use the Calculational Formula:

	x_i	x_i^2
	5	25
	12	144
	6	36
	8	64
	14	196
Sum	45	465

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

$$= \frac{465 - \frac{45^2}{5}}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

Some Notes

- The value of s is **ALWAYS** positive.
- The larger the value of s^2 or s, the larger the variability of the data set.
- **Why divide by n –1?**
  - The sample standard deviation s is often used to estimate the population standard deviation σ. Dividing by $n - 1$ gives us a better estimate of σ.