Measures of Dispersion

MATH 130, *Elements of Statistics I*

J Robert Buchanan

Department of Mathematics

Fall 2023

Introduction

- Recall that a measure of central tendency is a number which is "typical" of all the values of a variable in a data set.
- Another important measure is the amount of "spread" in the values of the variable. A **measure of dispersion** is a number which describes the spread of the data.

Example

A sample of the times (in minutes) spent waiting in line at two different banks is shown below. Find the mean, median, mode, and midrange for each bank.

Bank A	Bank B
6.5	4.2
6.6	5.4
6.7	5.8
6.8	6.2
7.1	6.7
7.3	7.7
7.4	7.7
7.7	8.5
7.7	9.3
7.7	10.0

Example

A sample of the times (in minutes) spent waiting in line at two different banks is shown below. Find the mean, median, mode, and midrange for each bank.

Bank A	Bank B
6.5	4.2
6.6	5.4
6.7	5.8
6.8	6.2
7.1	6.7
7.3	7.7
7.4	7.7
7.7	8.5
7.7	9.3
7.7	10.0

$$\bar{x}_A = 7.15$$
$$\bar{x}_B = 7.15$$

Example

A sample of the times (in minutes) spent waiting in line at two different banks is shown below. Find the mean, median, mode, and midrange for each bank.

Bank A	Bank B
6.5	4.2
6.6	5.4
6.7	5.8
6.8	6.2
7.1	6.7
7.3	7.7
7.4	7.7
7.7	8.5
7.7	9.3
7.7	10.0

$$\begin{aligned} \overline{x}_A &= 7.15 \\ \overline{x}_B &= 7.15 \\ M_A &= 7.20 \\ M_B &= 7.20 \end{aligned}$$

Example

A sample of the times (in minutes) spent waiting in line at two different banks is shown below. Find the mean, median, mode, and midrange for each bank.

Bank A	Bank B
6.5	4.2
6.6	5.4
6.7	5.8
6.8	6.2
7.1	6.7
7.3	7.7
7.4	7.7
7.7	8.5
7.7	9.3
7.7	10.0

$$
\begin{aligned}
\overline{x}_A &= 7.15 \\
\overline{x}_B &= 7.15 \\
M_A &= 7.20 \\
M_B &= 7.20 \\
\text{mode}_A &= 7.7 \\
\text{mode}_B &= 7.7
\end{aligned}
$$

Example

A sample of the times (in minutes) spent waiting in line at two different banks is shown below. Find the mean, median, mode, and midrange for each bank.

Bank A	Bank B
6.5	4.2
6.6	5.4
6.7	5.8
6.8	6.2
7.1	6.7
7.3	7.7
7.4	7.7
7.7	8.5
7.7	9.3
7.7	10.0

$$\begin{aligned} \overline{x}_A &= 7.15 \\ \overline{x}_B &= 7.15 \\ M_A &= 7.20 \\ M_B &= 7.20 \\ \text{mode}_A &= 7.7 \\ \text{mode}_B &= 7.7 \\ \text{midrange}_A &= 7.10 \\ \text{midrange}_B &= 7.10 \end{aligned}$$

The Range

Definition

The **range**, denoted R, is the difference between the largest and smallest data values.

$$R = \text{largest data value} - \text{smallest data value}$$

The Range

Definition

The **range**, denoted R, is the difference between the largest and smallest data values.

$$R = \text{largest data value} - \text{smallest data value}$$

Example

Find the range in waiting times for each of the banks in the previous example.

The Range

Definition

The **range**, denoted R, is the difference between the largest and smallest data values.

$$R = \text{largest data value} - \text{smallest data value}$$

Example

Find the range in waiting times for each of the banks in the previous example.

$$\begin{aligned} R_A &= 7.7 - 6.5 = 1.20 \\ R_B &= 10.0 - 4.2 = 5.80 \end{aligned}$$

The Variance

Idea: the variance of a variable is its deviation about the mean.

The **population variance** will be denoted σ^2.

The **sample variance** will be denoted s^2.

Formula for Population Variance

Definition

The **population variance** of a variable is the sum of the squared deviations about the population mean divided by the number of observations in the population.

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N} = \frac{\sum (x_i - \mu)^2}{N}$$

Example

Suppose the following table contains the total number of hours students spent studying per week for their courses.

28.4	44.4	36.4	33.5
34.9	30.5	32.7	24.6

Find the population variance. *Hint:* the population mean is $\mu = 33.18$.

Solution

Hours x_i	Deviation $x_i - \mu$	Squared Deviation $(x_i - \mu)^2$
28.4		
44.4		
36.4		
33.5		
34.9		
30.5		
32.7		
24.6		

$$\sigma^2 =$$

Solution

Hours x_i	**Deviation** $x_i - \mu$	**Squared Deviation** $(x_i - \mu)^2$
28.4	-4.78	
44.4	11.22	
36.4	3.22	
33.5	0.32	
34.9	1.72	
30.5	-2.68	
32.7	-0.48	
24.6	-8.58	

$$\sigma^2 =$$

Solution

Hours x_i	**Deviation** $x_i - \mu$	**Squared Deviation** $(x_i - \mu)^2$
28.4	-4.78	22.8484
44.4	11.22	125.888
36.4	3.22	10.3684
33.5	0.32	0.1024
34.9	1.72	2.9584
30.5	-2.68	7.1824
32.7	-0.48	0.2304
24.6	-8.58	73.6164

$$\sigma^2 =$$

Solution

Hours x_i	**Deviation** $x_i - \mu$	**Squared Deviation** $(x_i - \mu)^2$
28.4	-4.78	22.8484
44.4	11.22	125.888
36.4	3.22	10.3684
33.5	0.32	0.1024
34.9	1.72	2.9584
30.5	-2.68	7.1824
32.7	-0.48	0.2304
24.6	-8.58	73.6164
		$\sum(x_i - \mu)^2 = 243.195$

$$\sigma^2 =$$

Solution

Hours x_i	Deviation $x_i - \mu$	Squared Deviation $(x_i - \mu)^2$
28.4	-4.78	22.8484
44.4	11.22	125.888
36.4	3.22	10.3684
33.5	0.32	0.1024
34.9	1.72	2.9584
30.5	-2.68	7.1824
32.7	-0.48	0.2304
24.6	-8.58	73.6164
		$\sum(x_i - \mu)^2 = 243.195$

$$\sigma^2 = \frac{243.195}{8} = 30.40$$

Sample Variance

Definition

The **sample variance** of a variable is the sum of the squared deviations about the sample mean divided by $n - 1$.

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

The quantity $n - 1$ is called the **degrees of freedom**.

Alternative Formula for s^2

Remark: the formula for variance needs the sample mean. An alternative formula for calculating s^2 which does not require prior knowledge of the sample mean is

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

Example

The percentage of on-time arrivals for a sample of airports is tabulated below. Find the sample variance.

Atlanta	79.9
Dallas/Fort Worth	84.9
Los Angeles	81.0
Orlando	82.2
San Diego	78.4
Tampa	79.9

Solution (1 of 2)

Data x_i	**Data Squared** x_i^2
79.9	
84.9	
81.0	
82.2	
78.4	
79.9	

Solution (1 of 2)

Data x_i	**Data Squared** x_i^2
79.9	6384.01
84.9	7208.01
81.0	6561.00
82.2	6756.84
78.4	6146.56
79.9	6384.01

Solution (1 of 2)

Data x_i	Data Squared x_i^2
79.9	6384.01
84.9	7208.01
81.0	6561.00
82.2	6756.84
78.4	6146.56
79.9	6384.01
$\sum x_i = 486.3$	

Solution (1 of 2)

Data x_i	**Data Squared** x_i^2
79.9	6384.01
84.9	7208.01
81.0	6561.00
82.2	6756.84
78.4	6146.56
79.9	6384.01
$\sum x_i = 486.3$	$\sum x_i^2 = 39440.43$

Solution (2 of 2)

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

$$\sum x_i = 486.3$$

$$\sum x_i^2 = 39440.43$$

Solution (2 of 2)

$$\sum x_i = 486.3$$
$$\sum x_i^2 = 39440.43$$

$$\begin{aligned} s^2 &= \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} \\ &= \frac{39440.43 - \frac{(486.3)^2}{6}}{6-1} \\ &= \frac{39440.43 - \frac{236487.69}{6}}{5} \\ &= \frac{39440.43 - 39414.615}{5} \\ &= \frac{25.815}{5} \\ &= 5.16 \end{aligned}$$

Standard Deviation

Definition

The **population standard deviation** denoted σ is the square root of the population variance.

$$\sigma = \sqrt{\sigma^2}$$

The **sample standard deviation** denoted s is the square root of the sample variance.

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

Example

Find the standard deviations of the two previous samples (repeated below).

Bank A	Bank B
6.5	4.2
6.6	5.4
6.7	5.8
6.8	6.2
7.1	6.7
7.3	7.7
7.4	7.7
7.7	8.5
7.7	9.3
7.7	10.0

Example

Find the standard deviations of the two previous samples (repeated below).

Bank A	Bank B
6.5	4.2
6.6	5.4
6.7	5.8
6.8	6.2
7.1	6.7
7.3	7.7
7.4	7.7
7.7	8.5
7.7	9.3
7.7	10.0
$s_A = 0.48$	$s_B = 1.82$

Interpreting Standard Deviation

Assume the distribution of values for a variable is bell-shaped and symmetric.

- The mean measures the center of the distribution.
- The standard deviation measures the spread of the distribution.
- The larger the standard deviation, the greater the spread in the values of the variable.

Example

Recall the bank waiting times presented earlier.

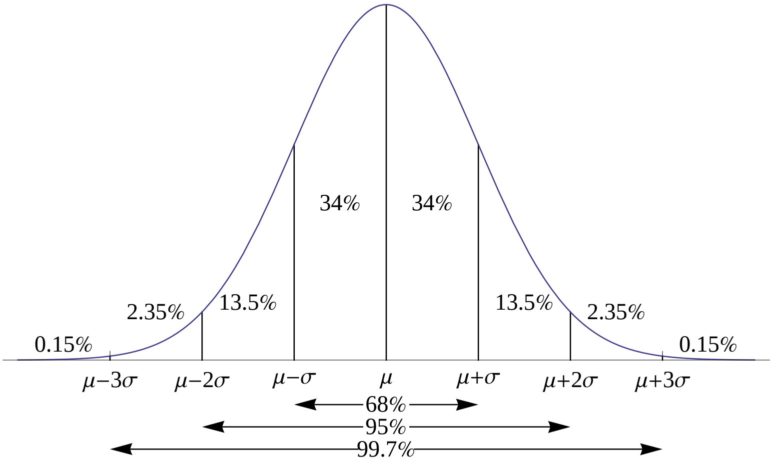
Bank A	**Bank B**
6.5	4.2
6.6	5.4
6.7	5.8
6.8	6.2
7.1	6.7
7.3	7.7
7.4	7.7
7.7	8.5
7.7	9.3
7.7	10.0
$s_A = 0.48$	$s_B = 1.82$

The Empirical Rule (1 of 2)

If a distribution is roughly bell-shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. In other words, approximately 68% of the data lie between $\mu - \sigma$ and $\mu + \sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. In other words, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. In other words, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

The Empirical Rule (2 of 2)



Example

If college students sleep an average of 7.3 hours per night with a standard deviation of 0.9 hours in the number of hours of sleep,

1. determine the percentage of students whose sleep hours are within 2 standard deviations of the mean.
2. determine the percentage of students who get between 6.4 and 9.1 hours of sleep per night.
3. determine the percentage of students who get less than 5.5 hours of sleep per night.

Chebyshev's Inequality

For any data set, regardless of the shape of the distribution, at least $\left(1 - \frac{1}{k^2}\right)$ 100% of observations will lie within k standard deviations of the mean where k is any number greater than 1. In other words, at least $\left(1 - \frac{1}{k^2}\right)$ 100% of observations will lie between $\mu - k\sigma$ and $\mu + k\sigma$.

Example

1. Determine the minimum percentage of data which will lie within 2 standard deviations of the mean.

2. Determine the minimum percentage of data which will lie within 3 standard deviations of the mean.

3. Determine the minimum percentage of data which will lie within 1.5 standard deviations of the mean.

Example

1. Determine the minimum percentage of data which will lie within 2 standard deviations of the mean.

$$\left(1 - \frac{1}{2^2}\right) 100\% = \left(1 - \frac{1}{4}\right) 100\% = 75\%$$

2. Determine the minimum percentage of data which will lie within 3 standard deviations of the mean.

3. Determine the minimum percentage of data which will lie within 1.5 standard deviations of the mean.

Example

1. Determine the minimum percentage of data which will lie within 2 standard deviations of the mean.

$$\left(1 - \frac{1}{2^2}\right) 100\% = \left(1 - \frac{1}{4}\right) 100\% = 75\%$$

2. Determine the minimum percentage of data which will lie within 3 standard deviations of the mean.

$$\left(1 - \frac{1}{3^2}\right) 100\% = \left(1 - \frac{1}{9}\right) 100\% = 88.9\%$$

3. Determine the minimum percentage of data which will lie within 1.5 standard deviations of the mean.

Example

1. Determine the minimum percentage of data which will lie within 2 standard deviations of the mean.

$$\left(1 - \frac{1}{2^2}\right) 100\% = \left(1 - \frac{1}{4}\right) 100\% = 75\%$$

2. Determine the minimum percentage of data which will lie within 3 standard deviations of the mean.

$$\left(1 - \frac{1}{3^2}\right) 100\% = \left(1 - \frac{1}{9}\right) 100\% = 88.9\%$$

3. Determine the minimum percentage of data which will lie within 1.5 standard deviations of the mean.

$$\left(1 - \frac{1}{(1.5)^2}\right) 100\% = \left(1 - \frac{1}{2.25}\right) 100\% = 55.6\%$$