3.2: Measures of Variation

The mean, median, and mode can describe the "middle" of a data set, but none of them can describe how "spread out" the data is.

Range:

The *range* for ungrouped data is the difference between the largest and smallest values.
The *range* for grouped data (a frequency distribution) is the difference between the upper boundary of the highest class and the lower boundary of the lowest class.

In other words,

Range = Maximum – Minimum.

Example 1: Find the range.

Range = 1.3 − 0.2 = 1.1

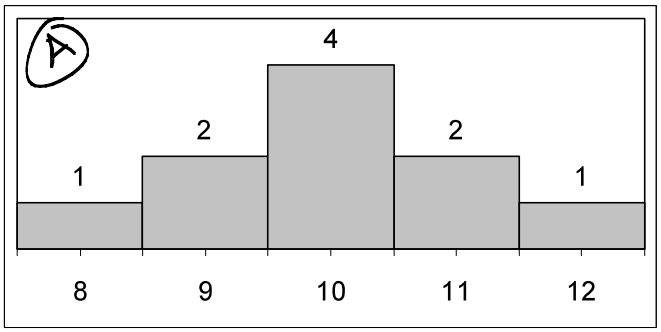
Commute Times							
0.3	0.7	0.2	0.5	0.7	1.2	1.1	0.6
0.6	0.2	1.1	1.1	0.9	0.2	0.4	1.0
1.2	0.9	0.8	0.4	0.6	1.1	0.7	1.2
0.5	1.3	0.7	0.6	1.1	0.8	0.4	0.8

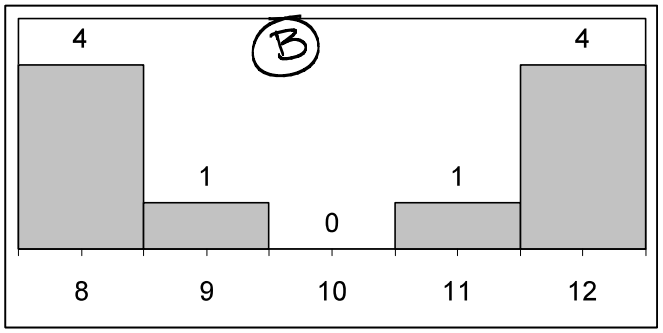
minimum

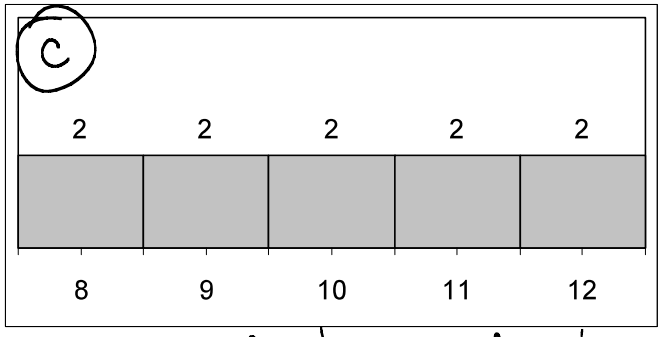
maximum

Example 2: Consider these data sets.

$A = \{8,9,9,10,10,10,10,11,11,12\}$, $B = \{8,8,8,8,9,11,12,12,12,12\}$,
$C = \{8,8,9,9,10,10,11,11,12,12\}$







In all:

Mean = 10

Median = 10

Range = 12 − 8 = 4

A is the least spread out (smallest variance, smallest std. deviation)

B is the most spread out (highest variance and highest standard deviation)

While the range is useful, it is dependent only on the extreme values of the data set. It doesn't tell you whether most of the data points are close to the mean, far from the mean, or evenly distributed. We need something else.

Deviation of a data point:

The deviation of a data point is the difference (i.e., the signed distance) between the data point and the mean. (negative if point is below the mean; positive if point is above the mean)

In other words, the deviation of the *i*th data point, x_i is $x_i - \mu$.

(Note that the deviation is positive if $x_i > \mu$; the deviation is negative if $x_i < \mu$.)

Let's average the deviations for a data set. (assume it's a population)

Example 3: $A = \{12, 13, 7, 5, 9\}$

$$\mu = \frac{12+13+7+5+9}{5} = 9.2$$

x_i	$x_i - \mu$
12	$12 - 9.2 = 2.8$
13	$13 - 9.2 = 3.8$
7	$7 - 9.2 = -2.2$
5	$5 - 9.2 = -4.2$
9	$9 - 9.2 = -0.2$
	sum = 0 (happens every time!)

Average of the deviations is always 0. Instead, we square the deviations and then average them.

Variance of a population: (our book postpones variance until a later section)

Variance of a population:

If $x_1, x_2, \ldots, x_n$ is a population with mean μ, then the *population variance* σ^2 is given by

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}.$$

Know this formula →

μ: mu = population mean

σ: sigma (lower-case Greek sigma)

In other words, the variance is the average (mean) of the squared deviations.

Alternative formula for the population variance:
(sometimes known as the computational formula, computing formula or shortcut formula)

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n}$$

(skip this formula)

Example 4: ~~Use the alternative (shortcut) formula to~~ Find the variance of the population $A = \{12, 13, 7, 5, 9\}$.

x_i	$x_i - \mu$	$(x_i - \mu)^2$
12	$12 - 9.2$	$(2.8)^2 = 7.84$
13	$13 - 9.2$	$(3.8)^2 = 14.44$
7	$7 - 9.2$	$(-2.2)^2 = 4.84$
5	$5 - 9.2$	$(-4.2)^2 = 17.64$
9	$9 - 9.2$	$(-0.2)^2 = 0.04$

Sum = 44.8

From earlier example, we know $\mu = 9.2$

Variance: $\sigma^2 = \frac{44.8}{5} = 8.96$

Variance: $\sigma^2 = 8.96$

Degrees of freedom:

The quantity known as *degrees of freedom* is the number of scores (data points) in a dataset that are free to vary in the presence of a statistical estimate.

If a sample has *n* data points and the sample mean $\bar{x}$ is specified, then $n-1$ of the data points can theoretically be anything; the *n*th data point is forced to be take on whatever value results in the specified mean $\bar{x}$. In other words, the first $n-1$ of the data points are free to vary; the *n*th data point is not free to vary.

Example 5: Suppose a sample has 5 data points and a mean of 159. Suppose also that the first four data points are 37, 203, 122, and 303. Calculate the fifth data point.

Let x = missing data point

$$\frac{37+203+122+303+x}{5} = 159$$

$$\frac{665+x}{5} = 159$$

→ multiply both sides by 5:

$$665 + x = 5(159)$$

$$665 + x = 795$$

$$x = 795 - 665$$

$$= 130$$

missing data point is 130.

Variance of a sample:

When we calculate the variance of a *sample* (not the whole population), we have no way to calculate the population mean. Therefore, we must use the sample mean (denoted $\bar{x}$) as an estimate of the population mean (denoted μ). Thus, in a sample of *n* data points, there are $n-1$ degrees of freedom.

When calculating the variance for a sample (not the entire population), we divide by $n-1$ (the degrees of freedom) instead of *n*. Dividing by *n* would underestimate the variance, because the points in the sample will be less spread out than those in the population. Using the degrees of freedom, $n-1$, in the denominator provides an unbiased estimate of the population variance.

Variance of a sample:

The *sample variance* s^2 of a set of *n* sample measurements $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$ is given by

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

← know this formula

Alternative formula for the sample variance:
(sometimes known as the computational formula, computing formula or shortcut formula)

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

← skip this formula

Example 6: Calculate the variance of this sample: {75, 16, 50, 88, 79, 95, 80}.

$$\bar{x} = \frac{75+16+50+88+79+95+80}{7} = 69$$

x_i	$(x_i - \bar{x})^2$
75	$(75-69)^2 = 6^2 = 36$
16	$(16-69)^2 = (-53)^2 = 2809$
50	$(50-69)^2 = (-19)^2 = 361$
88	$(88-69)^2 = (19)^2 = 361$
79	$(79-69)^2 = (10)^2 = 100$
95	$(95-69)^2 = (26)^2 = 676$
80	$(80-69)^2 = (11)^2 = 121$
	Sum = 4464

Variance:

$$s^2 = \frac{4464}{7-1} = \frac{4464}{6} = \boxed{744}$$

Standard deviation:

By squaring the deviations, we've changed the units (if there are any). In other words, if we started with "inches", we now have "square inches". This is easily fixed by taking square roots.

Standard deviation:

$$s = \sqrt{s^2} = \sqrt{744} \approx \boxed{27.28}$$

we can think of a "typical" data point as being about 27.28 units from the mean

Standard Deviation:

The *sample standard deviation* s of a set of n sample measurements $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$ is given by

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

If $x_1, x_2, \ldots, x_n$ is the whole population with mean μ, then the *population standard deviation* σ is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}.$$

Example 7: Given the following data sample, calculate the standard deviation to two decimal places.

{8, 5, 19, 21, 17, 13, 12, 24, 22}

$$\bar{x} = \frac{8+5+19+\ldots+22}{9} = \frac{141}{9} \approx 15.67$$

x_i	$(x_i - \bar{x})^2$
8	$(8-15.67)^2 = 58.7778$
5	113.7778
19	11.1111
21	28.4444
17	1.7778
13	7.1111
12	13.4444
24	69.4444
22	40.1111

Sum of Squares = 343.9999

Variance:

$$s^2 = \frac{343.9999}{9-1} = 42.99998\overline{75}$$

So variance is $\boxed{s^2 = 43}$

Standard deviation is

$$s = \sqrt{43} \approx \boxed{6.56}$$

IMPORTANT:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

$$(\text{Standard Deviation})^2 = \text{Variance}$$