3.2: Measures of Variation (Measures of Dispersion)

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,

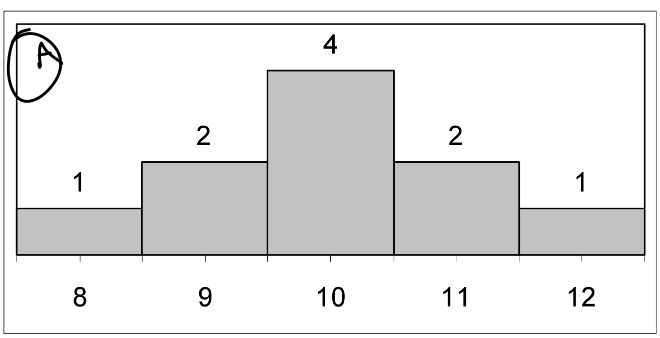
$$\text{Range} = \text{Maximum} - \text{Minimum}.$$

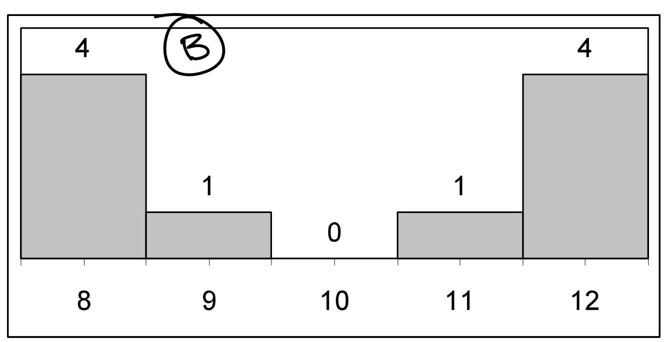
Example 1: Find the range. Range $= 1.3 - 0.2 = \boxed{1.1}$

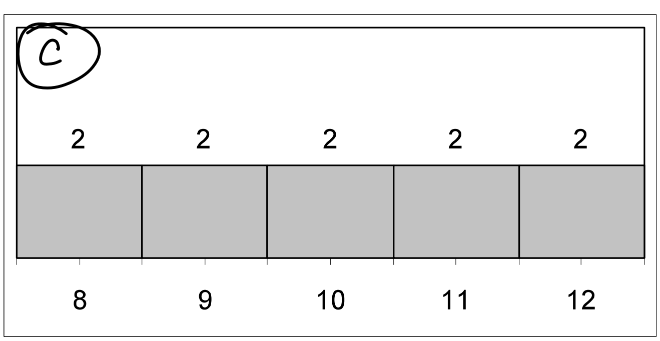
Commute Times							
0.3	0.7	0.2	0.5	0.7	1.2	1.1	0.6
0.6	0.2	1.1	1.1	0.9	0.2	0.4	1.0
1.2	0.9	0.8	0.4	0.6	1.1	0.7	1.2
0.5	1.3	0.7	0.6	1.1	0.8	0.4	0.8

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 = $\frac{100}{10} = 10$

Median = 10

Range = $12 - 8 = 4$

B is the most spread out (dispersed). Highest variance.

A is the least spread out (dispersed). Smallest Variance.

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.

In other words, the deviation of the ith data point, x_i is $x_i - \mu$. (or $x_i - \bar{x}$ if it's a sample)
(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.

<u>**Example 3:**</u> $A = \{12, 13, 7, 5, 9\}$

mean: $\mu = \dfrac{12+13+7+5+9}{5} = \dfrac{46}{5} = 9.2$

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 (true every time)

So the average of the deviations is $\frac{0}{5} = 0$. (always happens)

Instead, we square the deviations before adding them up.

Variance of a population: (Our book skips variance until a later chapter. But you need to know it!)

<u>Variance of a population</u>:

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

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

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

You are not responsible for this formula →

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

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

σ: Greek letter lowercase sigma

σ^2 is the population variance

without a table:

$$\frac{(12-9.2)^2+(13-9.2)^2+(7-9.2)^2+(5-9.2)^2+(9-9.2)^2}{5} = \boxed{8.96}$$

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

From Example 3, $\mu = 9.2$

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

Sum = 44.8

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

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 = 5th (missing) data point.

mean: $\bar{x} = 159$

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

Solve for x: $\frac{665+x}{5} = 159$

multiply by 5: $\cancel{5}\left(\frac{665+x}{\cancel{5}}\right) = (159)(5)$

$665 + x = 795$

$x = 795 - 665$

$= 130$

$\boxed{\text{The missing data point is 130.}}$

Variance of a sample:

If a sample has 5 data points and the sample mean is known, there are 4 degrees of freedom.

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

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

Alternative formula for the sample variance: (not responsible for this)
(sometimes known as the computational formula, computing formula or shortcut formula)

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

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

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
75	75−69 = 6	$6^2 = 36$
16	16−69 = −53	$53^2 = 2809$
50	50−69 = −19	$19^2 = 361$
88	88−69 = 19	$19^2 = 361$
79	79−69 = 10	$10^2 = 100$
95	95−69 = 26	$26^2 = 676$
80	80−69 = 11	$11^2 = 121$
Σ = 483		$\Sigma(x_i - \bar{x})^2 = 4464$

mean: $\bar{x} = \frac{75+16+50+88+79+95+80}{7} = \frac{483}{7} = 69$

Sample variance:

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{4464}{7-1} = \frac{4464}{6} = \boxed{744}$$

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.

For previous example (Example 6): the standard deviation is

$$s = \sqrt{s^2} = \sqrt{744} \approx \boxed{27.276}$$

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.

~~{70, 39, 54, 84, 68, 93, 75}~~ $\{8, 5, 19, 21, 17, 13, 12, 25\}$

$$\bar{x} = \frac{8+5+19+\cdots+25}{8} = \frac{120}{8} = 15$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
8	$8-15=-7$	$(-7)^2 = 49$
5	$5-15=-10$	$(-10)^2 = 100$
19	$19-15=4$	$4^2 = 16$
21	$21-15=6$	$6^2 = 36$
17	$17-15=2$	$2^2 = 4$
13	$13-15=-2$	$(-2)^2 = 4$
12	$12-15=-3$	$(-3)^2 = 9$
25	$25-15=10$	$10^2 = 100$
		Sum $= 318$

Variance: $s^2 = \frac{318}{8-1} = \frac{318}{7} \approx 45.429$

Standard Deviation:

$s = \sqrt{s^2} \approx \sqrt{45.429}$

≈ 6.740072064

$\approx \boxed{6.74}$

IMPORTANT:

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

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