Kaleb Burris
STAT F300

(§1.4) Measures of Variability or Spread

Why do we care how spread out a data set is? Isn't it enough to just know something about a (typical) middle value?

Example: Suppose you hear that three different instructors assign an average grade of "B" (3.0) in a course you're considering. How might you decide which instructor's class to sign up for?

The three instructors give out grades thusly:

- All B's
- C → A evenly spread out
- Half the class gets C's, half get A's

There are several measures of variability:

1. Range = max - min. Easy to calculate, but isn't very good as a measure of spread, for example: $1, 5, 5, 5, 9 : (9 - 1) = 8$. Even though there are three 5s, the range is higher than all but one member of the sample.

2. Population Variance = σ^2 = average of the squared deviations from the population mean.

Example: Amounts that all 5 of your friends owe you (You only have 5 friends):

$$\{20, 30, -10, 20, 90\}$$
$$N = 5 = \text{population or sample size}$$

---

x_i	$x_i - \mu$	$(x_i - \mu)^2$
20	−10	100
30	0	0
−10	−40	1600
20	−10	100
90	60	3600
150	0	5400

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{5}(150) = 30$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{5}(5400) = 1080$$

Population standard deviation:

$$\sigma = \sqrt{\sigma^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$
$$\sigma = \sqrt{\sigma^2} = \sqrt{1080} \approx 32.86$$

2'. Sample Variance: S^2

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

S^2 = average of the squared deviations from the sample mean, $(x_i - \overline{x})$

Example: Amounts that five of your 528 Facebook friends owe you:

$$\{20, 30, -10, 20, 90\}$$
$N = 5$ = population or sample size

---

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
20	−10	100
30	0	0
−10	−40	1600
20	−10	100
90	60	3600
150	0	5400

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(150) = 30$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{4}(5400) = 1350$$

Note: Sample variance gives a bigger value than what the population variance formula gave:

$$S^2 > \sigma \Rightarrow 1350 > 1080$$

An alternative formula for S^2, which is easier to compute by hand:

$$S^2 = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2\right]$$

This formula gives the same result, it's simply that the algebra has been reformatted for human computational ease. It's sometimes referred to as the "computational formula".

We sometimes prefer to use σ rather than σ^2, and S rather than S^2 because:

- Units for S^2 square the units of observation such as square gallons, square people, etc. S takes $\sqrt{S^2}$, thus we end up with the same units we started with.

R code for sample variance σ

```
> xx <- c(20, 30, -10, 20, 90)
> var(xx)
[1] 1350
> sd(xx)
[1] 36.74235
```

Some properties of the sample variance formula:

How is the variance affected if I add a constant c to all numbers in the data set?

Example: $c = 3$

$$\{1,3,5,11\} \Rightarrow S_1^2 = 18.\overline{666}$$
$$\{1,3,5,11\} + 3 = \{4,7,8,14\} \Rightarrow S_2^2 = 18.\overline{666}$$

Adding a number to all values doesn't change the variance.

How is the variance affected if I multiply all the numbers in the data set by some constant k?

Example:

$$\{1,3,5,11\} \Rightarrow S_1^2 = 18.\overline{666}$$
$$\text{for} \quad k = 2 : \{1,3,5,11\} * 2 = \{2,6,10,22\} \Rightarrow S_2^2 = 74.\overline{666} = 4(S_1^2)$$
$$\text{for} \quad k = -3 : \{1,3,5,11\} * -3 = \{-3,-9,-15,-33\} \Rightarrow S_4^2 = 168 = 9(S_1^2)$$

Multiplying the dataset changes the variance by k^2.

If we're asked to calculate the variance of some list of numbers, when should we use the formula for σ^2 and when should we use S^2?

You can only tell by looking at the context. Look for keywords:

- 5 of my Facebook friends $\Longrightarrow$ sample $\Longrightarrow$ S^2
- My 5 Facebook friends $\Longrightarrow$ population $\Longrightarrow$ σ^2

Box Plots: Another way to plot a data set. Box plots indicate the center, spread, symmetry, and outliers in the data set.

Definitions:

- Q_1 = First Quartile = median of smallest half of values
- Q_3 = Second Quartile = median of largest half of values
- $IQR = f_s$ = Fourth Spread = $Q_3 - Q_1$

If n is odd, the middle value is used when calculating both Q_1 and Q_3.

Example: $\{-1,1,4,7,\vert 12,13,20,22\}$ $n = 8:$ even

$$Q_1 = \frac{1+4}{2} = 2.5$$
$$Q_3 = \frac{13+12}{2} = 16.5$$
$$IQR = Q_3 - Q_1 = 16.5 - 2.5 = 14$$

Example: $\{9,4,6,11,21,23,55\}$ $n = 7:$ odd

$$Q_1 = \frac{4+6}{2} = 5$$
$$Q_3 = \frac{21+23}{2} = 22$$
$$IQR = Q_3 - Q_1 = 22 - 5 = 17$$

Here's the recipe for constructing a boxplot ("cat and whisker plot"):

1. Draw a horizontal line that extends from the smallest to largest values in your data set.
2. Draw a rectangle with vertical lines at Q_1, Q_2, and Q_3. (Q_2 = median)
3. If $x_i Q_1 - 1.5 * IQR$ *or* $x_i > Q_3 + 1.5 * IQR$, then x_i is considered an outlier. Put a dot at the locations of outliers.
4. Draw whiskers that extend from the rectangle to the most extreme non-outlying observation.

Example: $\{1,5,7,18,20,22,50\}$