ST 516: Foundations of Data Analytics

Variances

Variance of a Set of Numbers

Recall the definition of the **variance** of a set of numbers $\{x_1, x_2, \ldots, x_n\}$:

Definition

The **variance** of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is

$$\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

(Remember that $\overline{x}$ is the *mean* of the set of numbers.)

We use the notation s^2 to denote the variance, so we write

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

Variance of a Set of Numbers: Example

Example: What is the variance of the set of numbers $\{1, 4, -3, -10\}$?

$$\bar{x} = \frac{1 + 4 + (-3) + (-10)}{4} = \frac{-8}{4} = -2$$

$$\begin{aligned} s^2 &= \frac{(1-(-2))^2 + (4-(-2))^2 + (-3-(-2))^2 + (-10-(-2))^2}{3} \\ &= \frac{3^2 + 6^2 + (-1)^2 + (-8)^2}{3} \\ &= \frac{110}{3} \\ &= 36.667 \end{aligned}$$

Variance of a Set of Numbers: Example

Example: What is the variance of the set of numbers $\{2, 8, -6, -20\}$?

Notice that this new set of numbers is just 2 times each of the values in the previous set. Can you guess what the variance will be without calculating it?

$$\bar{x} = \frac{2 + 8 + (-6) + (-20)}{4} = \frac{-16}{4} = -4$$

$$\begin{aligned} s^2 &= \frac{(2-(-4))^2 + (8-(-4))^2 + (-6-(-4))^2 + (-20-(-4))^2}{3} \\ &= \frac{6^2 + 12^2 + (-2)^2 + (-16)^2}{3} \\ &= \frac{440}{3} \\ &= 146.667 \\ &= 2^2 \times 36.667 \end{aligned}$$

Variance of a Set of Numbers: Example

Example: What is the variance of the set of numbers $\{11, 14, 7, 0\}$?

Notice that this new set of numbers is just 10 plus each of the values in the previous set. Can you guess what the variance will be without calculating it?

$$\bar{x} = \frac{11 + 14 + 7 + 0}{4} = \frac{32}{4} = 8$$

$$\begin{aligned} s^2 &= \frac{(11-8)^2 + (14-8)^2 + (7-8)^2 + (0-8)^2}{3} \\ &= \frac{3^2 + 6^2 + (-1)^2 + (-8)^2}{3} \\ &= \frac{110}{3} \\ &= 36.667 \\ &= 36.667 \end{aligned}$$

Properties of Variances

The examples on the previous slides illustrated the following property:

If the variances of

$$\{x_1, x_2, \ldots, x_n\}$$

is s^2, then the variance of

$$\{ax_1 + b,\ ax_2 + b,\ \ldots, ax_n + b\}$$

is $a^2 s^2$.

That is, if we multiply each number in our set by the same value a, and then add to each number the same value b, the variance of the new set is equal to a^2 times the variance of the old set. *Note that adding the same value to each number in the set does not change the variance!*

Standard Deviation of a Set of Numbers

Note that the *units* for a variance will be the *square* of the units of our numbers.

For instance, if x_1, x_2, and x_3 are prices of books, in dollars, the variance will be a value with squared-dollars units.

For this reason, sometimes instead of working with the variance s^2 of a set of numbers we work with the square-root of the variance, which is called the **standard deviation** and denoted by s:

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Sample Variance

Suppose we perform an experiment where we gather n experimental units (members of a population of interest) and measure a variable X on each unit, so we have observations $X_1, X_2, \ldots, X_n$.

- X_1 is the value of the variable X for the first unit selected
- X_2 is the value of the variable X for the second unit selected
- ... and so on.

The **sample variance** for a the sample of observations $X_1, X_2, \ldots, X_n$ is the variance of the values observed in that particular sample.

We will continue to use the notation s^2 to denote the sample variance of a sample of random variables.

Sample Variance: Example

Suppose we randomly select an adult who lives in New York City, and we measure the variable

X = Amount (\$) that person spent on taxis during the past year

If we obtain a sample of $n = 30$ randomly selected New Yorkers and measure X for each one, the sample variance is the variance of the amounts that those 30 New Yorkers spent on taxis over the last year.

Note that if we repeat the experiment by getting a different sample of 30 New Yorkers, we would get a different sample variance.

Just like the sample mean, the sample variance is a *random variable*: it takes on different values depending on the specific random sample obtained.

Population Variance

The **population variance** of a random variable X is the variance of the values of that variable for the entire population.

We imagine that we could collect all members of the population of interest, and measure the value of the variable X for each individual member. Then we would take the variance of all of those values.

The Greek letter σ^2 is often used to denote the population variance.

(There is one minor difference in how we compute a population variance: we divide by n instead of $n-1$. This is a relatively unimportant difference in most settings, where n is large so $\frac{1}{n} \approx \frac{1}{n-1}$.)

Population Variance: Example

Suppose we randomly select an adult who lives in New York City, and we measure the variable

X = Amount ($) that person spent on taxis during the past year

The population variance σ^2 of X could be found by

- Collecting the amounts that *each adult* in New York City spent on taxis over the last year
- Finding the (population) mean of all of those amounts.
- Calculating the distance that individual amount is from that mean
- Adding up the square of those distances
- Dividing that sum by the *total number* of adults in New York City

Sample and Population Standard Deviation

Again, instead of working with the sample variance and population variance, we could instead talk about:

- The sample standard deviation s (the standard deviation of the values of the variable in our sample)
- The population standard deviation σ (the standard deviation of all of the values of the variable, from the entire population).

Note that the sample standard deviation is just the square root of the sample variance. Likewise, the population standard deviation is just the square root of the population variance.

Population Variance Examples

For certain population distribution families (like the Binomial, Poisson, Normal), we know the population variance: it is a function of the parameters that we specify to describe the distribution.

- If a random variable X has the Binomial(n, p) distribution, the population variance of X is

$$\sigma^2 = \text{Var}(X) = np(1-p)$$

- If a random variable X has the Poisson(λ) distribution, the population variance of X is

$$\sigma^2 = \text{Var}(X) = \lambda$$

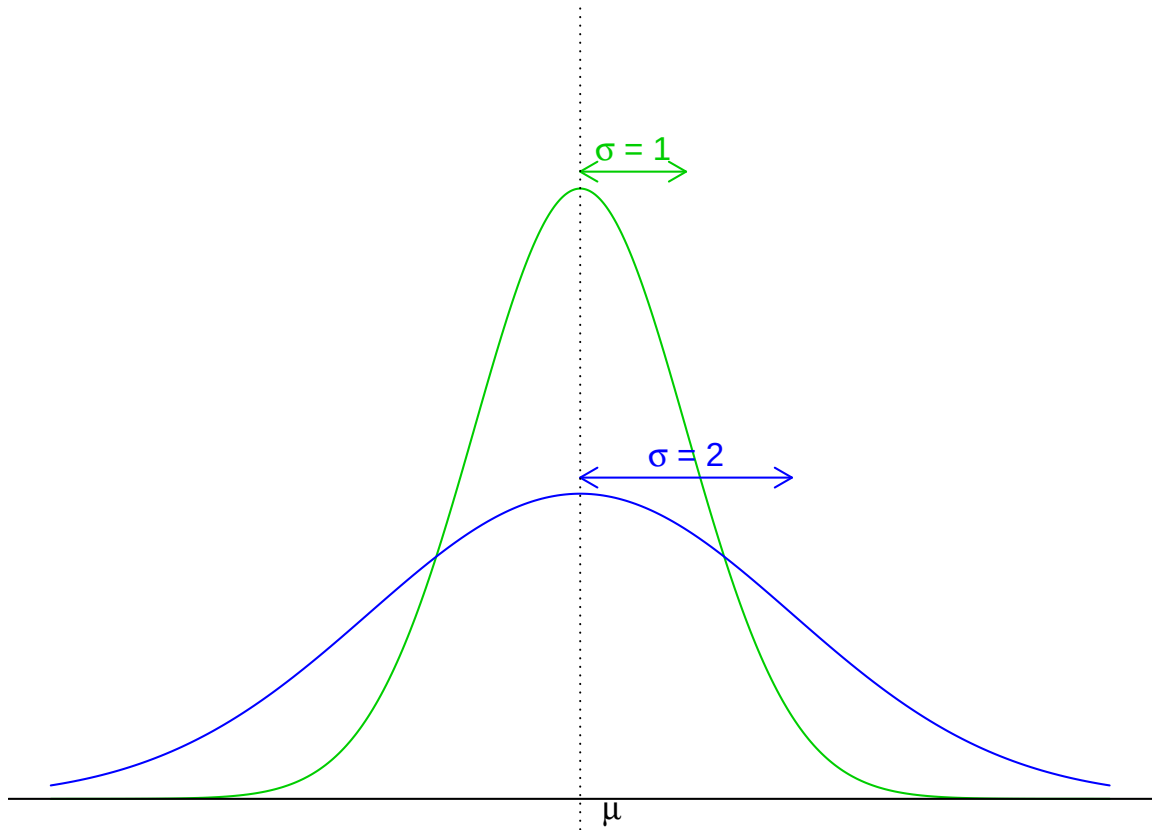
- If a random variable X has the Normal(μ, σ^2) distribution, the population variance of X is

$$\text{Var}(X) = \sigma^2$$

(This is why we use the notation 'σ^2' to denote the second parameter of a normal distribution!)

Population Variance Interpretation

The figure below illustrates two different normal distributions with the same mean, but different population variances (equivalently, different population standard deviations σ):



Population Variance Interpretation

The population variance (likewise, the population standard deviation) is a *scale parameter*: It tells us about the *spread* of values from that population distribution.

If a population distribution has a large variance, that means that the values of the variable in that population are very spread out: there are some that are much larger than the mean, and some that are much smaller than the mean.

In contrast, if a population distribution has a small variance, then the values of the variable in that population all tend to be quite close to the population mean.