Exploratory data analysis (EDA)

Part 5: Measures of variation (variance and standard deviation)

By: Noureddin Sadawi, PhD

University of London

Measures of variation

- In the previous section we spoke about:
  - range
  - interquartile range
- In this section, we will consider the following measures of variation:
  - variance
  - standard deviation

Variance

The variance is the average of the squared deviations of values from the mean.

Population variance, denoted by σ^2 is given by:

$$\sigma^2 = \frac{\sum\left(x_i - \mu\right)^2}{N}$$

Where *N* denotes the population size the $\boldsymbol{\mu}$ denotes the population mean.

Variance

Sample variance, denoted by s^2 is given by:

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

Which simplifies to:2

$$s^2 = \frac{\sum x_1^2 - n\overline{x}^2}{n-1}$$

Where n denotes the sample size the $\overline{x}$ denotes the sample mean.

Notice we divide by n *-1* and not n.

This is because we are using the sample mean and not the population mean (more on this later in topic 3).

Standard deviation

- The standard deviation is the most commonly used measure of spread.
- This is because it is expressed in the same unit as the original data whereas the variance is in squared units.
- It is equal to the positive square root of the variance and is denoted by $\boldsymbol{\sigma}$ for population standard deviation and by $\boldsymbol{s}$ for sample standard deviation. That is:

Standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

And

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

The standard deviation shows variation about the mean.

Example

Suppose we want to compute the sample mean, variance and standard deviation of the following data:

11, 12, 13, 16, 16, 18, 19, 20, 21

Sample Mean $= \overline{x} =$

(11 + 12 + 13 + 16 + 16 + 18 + 19 + 20 + 21)/9 = 146/9

Example

To calculate the variance:

X	11	12	13	16	16	18	19	20	21	**Total**
X^2	121	144	169	256	256	324	361	400	441	**2472**

$$s^2 = \frac{\sum x_i^2 - n\overline{x}^2}{n-1} = \frac{2472 - 9 \times \left(\frac{146}{9}\right)^2}{9-1} = 12.94\ (2\ d.p.)$$

Standard deviation $s = \sqrt{12.94} = 3.60\ (2\ d.p.)$

Example

Standard deviation $s = \sqrt{12.94} = 3.60\ (2\ d.p.)$

If data is in frequency table

- If the data is discrete, then:

$$s^2 = \frac{\sum\left(f * (x - \overline{x})^2\right)}{\sum f - 1} \quad \text{and} \quad \overline{x} = \frac{\sum(f * x)}{\sum f}$$

Here, f denotes the frequency.

Example

x	$frequency(f)$	$f * x$	$x - \overline{x}$	$(x - \overline{x})^2$	$f * (x - \overline{x})^2$
5	5	25	-3.3	10.89	54.45
6	5	30	-2.3	5.29	26.45
10	2	20	1.7	2.89	5.78
11	5	55	2.7	7.29	36.45
12	3	36	3.7	13.69	41.07
Total	20	166		40.05	164.2

$\overline{x} = 8.3$

Example

$$\overline{x} = \frac{\sum(f*x)}{\sum f} = \frac{166}{20} = 8.3$$

$$s^2 = \frac{\sum(f * (x - \overline{x})^2)}{\sum f - 1} = \frac{164.2}{20 - 1}$$

$$s^2 \approx 8.642$$

$$s^2 = \sqrt{8.642} \approx 2.94$$

Coefficient of variation

- The Coefficient of variation (CV) measures how large the standard deviation is in relation to the mean.
- It it usually given as a percentage and is defined as:

$$\mathrm{CV} = \left(\frac{s}{\overline{x}} \times 100\right)\%$$

Coefficient of variation

- It is a dimensionless number, that is, it is independent of the unit of measurement.
- It also shows the extent of variability in relation to the mean of the population.

- For example, in finance investors calculate how much risk (i.e. volatility) they can assume in comparison to the amount of return they expect from investments.
- The risk-return trade-off: the lower the CV value the better!