

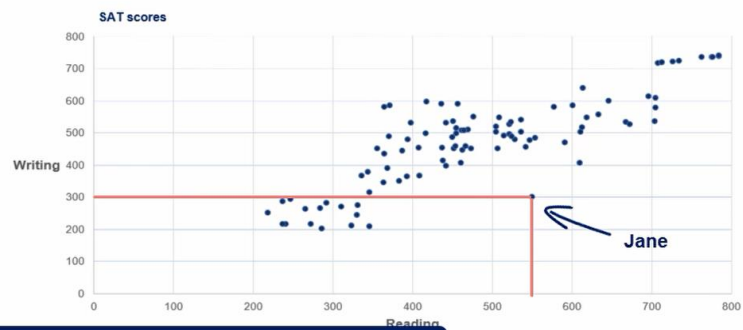
Scatter Plot

Scatter plot

| Student ID | Reading | Writing |
|------------|---------|---------|
| 1 | 273 | 216 |
| 2 | 292 | 282 |
| 3 | 219 | 250 |
| 4 | 241 | 217 |
| 5 | 284 | 266 |
| 6 | 247 | 294 |
| 7 | 237 | 215 |
| 8 | 286 | 203 |
| 9 | 237 | 286 |
| 10 | 266 | 263 |
| 11 | 311 | 270 |
| 12 | 324 | 243 |
| 13 | 330 | 243 |
| 14 | 331 | 275 |
| 15 | 336 | 367 |
| 16 | 344 | 378 |
| 17 | 346 | 315 |
| 18 | 364 | 208 |
| 19 | 356 | 445 |
| 20 | 364 | 346 |
| 21 | 365 | 435 |
| 22 | 365 | 579 |
| 23 | 369 | 390 |
| 24 | 436 | 589 |
| 25 | 393 | 365 |
| 26 | 394 | 445 |
| 27 | 417 | 499 |
| 28 | 438 | 414 |
| 29 | 398 | 530 |



| 1 | Graphs and tables for |
|----|---------------------------|
| 2 | Scatter plot |
| 3 | |
| 4 | Student I Reading Writing |
| 5 | 1 275 216 |
| 6 | 2 292 282 |
| 7 | 3 219 250 |
| 8 | 4 241 217 |
| 9 | 5 284 266 |
| 10 | 6 247 294 |
| 11 | 7 237 215 |
| 12 | 8 286 203 |
| 13 | 9 237 206 |
| 14 | 10 266 263 |
| 15 | 11 311 270 |
| 16 | 12 324 211 |
| 17 | 13 330 243 |
| 18 | 14 331 275 |
| 19 | 15 336 367 |
| 20 | 16 344 376 |
| 21 | 17 346 315 |
| 22 | 18 346 208 |
| 23 | 19 356 451 |
| 24 | 20 364 346 |
| 25 | 21 365 435 |
| 26 | 22 365 579 |
| 27 | 23 369 390 |
| 28 | 24 436 381 |
| 29 | 25 393 365 |
| 30 | 26 394 480 |
| 31 | 27 417 499 |



Scatter plots represent lots and lots of observations

Skewness

Skewness

Positive (right)

| <u>Dataset 1</u> | <u>Interval</u> | <u>Frequency</u> |
|------------------|-----------------|------------------|
| 1 | 0 to 1 | 4 |
| 1 | 1 to 2 | 6 |
| 1 | 2 to 3 | 4 |
| 2 | 3 to 4 | 2 |
| 2 | 4 to 5 | 2 |
| 2 | 5 to 6 | 0 |
| 2 | 6 to 7 | 1 |
| 2 | | |
| 2 | | |
| 2 | | |
| 3 | | |
| 3 | | |
| 3 | | |
| 3 | | |
| 4 | | |
| 4 | | |
| 5 | | |
| 5 | | |
| 7 | | |

| <u>Mean</u> | <u>Median</u> | <u>Mode</u> |
|-------------|---------------|-------------|
| 2.73 | 2.00 | 2.00 |

Zero (no skew)

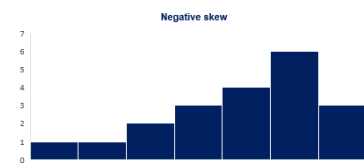
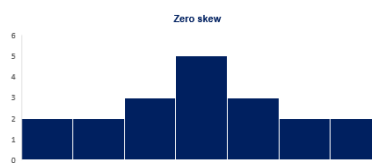
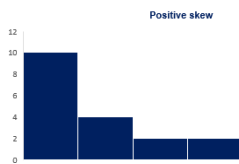
| <u>Dataset 2</u> | <u>Interval</u> | <u>Frequency</u> |
|------------------|-----------------|------------------|
| 1 | 0 to 1 | 2 |
| 1 | 1 to 2 | 2 |
| 2 | 2 to 3 | 3 |
| 2 | 3 to 4 | 5 |
| 3 | 4 to 5 | 3 |
| 3 | 5 to 6 | 2 |
| 3 | 6 to 7 | 2 |
| 4 | | |
| 4 | | |
| 4 | | |
| 4 | | |
| 5 | | |
| 5 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 7 | | |

| <u>Mean</u> | <u>Median</u> | <u>Mode</u> |
|-------------|---------------|-------------|
| 4.00 | 4.00 | 4.00 |

Negative (left)

| <u>Dataset 3</u> | <u>Interval</u> | <u>Frequency</u> |
|------------------|-----------------|------------------|
| 1 | 0 to 1 | 1 |
| 2 | 1 to 2 | 1 |
| 3 | 2 to 3 | 2 |
| 3 | 3 to 4 | 3 |
| 4 | 4 to 5 | 4 |
| 4 | 5 to 6 | 6 |
| 5 | 6 to 7 | 3 |
| 5 | | |
| 5 | | |
| 5 | | |
| 6 | | |
| 6 | | |
| 6 | | |
| 6 | | |
| 6 | | |
| 7 | | |
| 7 | | |
| 7 | | |

| <u>Mean</u> | <u>Median</u> | <u>Mode</u> |
|-------------|---------------|-------------|
| 4.90 | 5.00 | 6.00 |



Variance

VARIANCE



Variance measures the dispersion of a set of data points around their mean

VARIANCE

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

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



population
variance



sample
variance



10 SQUARED IS 100

➤ Dispersion is non-negative. Non-negative values don't cancel out

➤ Amplifies the effect of large differences



100 SQUARED IS 10,000!

| Variance | | | | | |
|------------|---------------------|------|----------------------|---------------------|------|
| Population | | | Imaginary population | | |
| 1 | Mean | 3.00 | 1 | Mean | 3.20 |
| 2 | Population variance | 2.00 | 1 | Population variance | 2.96 |
| 3 | Sample variance | 2.50 | 1 | | |
| 4 | | | 2 | | |
| 5 | | | 3 | | |
| | | | 4 | | |
| | | | 5 | | |
| | | | 5 | | |
| | | | 5 | | |
| | | | 5 | | |

Standard Deviation

STANDARD DEVIATION FORMULAS

$$\sigma = \sqrt{\sigma^2}$$

population standard deviation

$$S = \sqrt{S^2}$$

sample standard deviation

Coefficient of Variation (CV)

COEFFICIENT OF VARIATION (CV)

relative standard deviation/

standard deviation

mean

COEFFICIENT OF VARIATION (CV)

$$c_v = \frac{\sigma}{\mu}$$

Population formula

Sample formula

$$\hat{c}_v = \frac{s}{\bar{x}}$$

σ

Standard deviation is the most common measure of variability for a SINGLE DATASET

Comparing TWO OR MORE datasets

c_v

Comparing the standard deviations of two different data sets is meaningless but Comparing coefficient of coefficients is meaningful

Standard deviation and coefficient of variation
rice example

| NY Dollars | Pesos | | Dollars | Pesos |
|------------|------------|---------------------------|-----------------------|--------------------------|
| \$ 1.00 | MXN 18.81 | Mean | \$ 5.50 | MXN 103.46 |
| \$ 2.00 | MXN 37.62 | Sample variance | \$ ² 10.72 | MXN ² 3793.69 |
| \$ 3.00 | MXN 56.43 | Sample standard deviation | \$ 3.27 | MXN 61.59 |
| \$ 3.00 | MXN 56.43 | | | |
| \$ 5.00 | MXN 94.05 | | | |
| \$ 6.00 | MXN 112.86 | | | |
| \$ 7.00 | MXN 131.67 | | | |
| \$ 8.00 | MXN 150.48 | | | |
| \$ 9.00 | MXN 169.29 | | | |
| \$ 11.00 | MXN 206.91 | | | |

Sample standard deviation

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

Step 1: Sample or population?

Step 2: Find the mean

Step 3: Find the sample variance

Step 4: Find the sample standard deviation

Standard deviation and coefficient of variation

Price example

| NY Dollars | | Pesos |
|------------|-----|--------|
| \$ 1.00 | MXN | 18.81 |
| \$ 2.00 | MXN | 37.62 |
| \$ 3.00 | MXN | 56.43 |
| \$ 3.00 | MXN | 56.43 |
| \$ 5.00 | MXN | 94.05 |
| \$ 6.00 | MXN | 112.86 |
| \$ 7.00 | MXN | 131.67 |
| \$ 8.00 | MXN | 150.48 |
| \$ 9.00 | MXN | 169.29 |
| \$ 11.00 | MXN | 206.91 |

| | Dollars | | Pesos |
|---------------------------------|-----------------------|------------------|---------|
| Mean | \$ 5.50 | MXN | 103.46 |
| Sample variance | \$ ² 10.72 | MXN ² | 3793.69 |
| Sample standard deviation | \$ 3.27 | MXN | 61.59 |
| Sample coefficient of variation | 0.60 | | 0.60 |

- does not have a unit of measurement
- universal across datasets
- perfect for comparisons

Covariance and Liner corelation coefficient

Covariance

Housing data

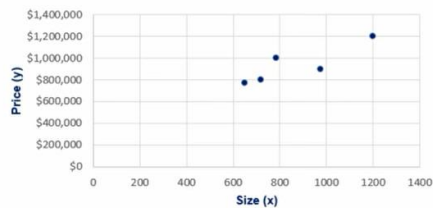
| Size (ft.) | Price (\$) |
|------------|------------|
| 650 | 772,000 |
| 785 | 998,000 |
| 1200 | 1,200,000 |
| 720 | 800,000 |
| 975 | 895,000 |

Sample formula

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Population formula

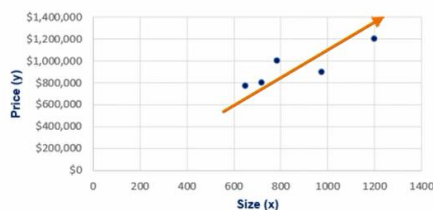
$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x) * (y_i - \mu_y)}{N}$$



Covariance

Housing data

| Size (ft.) | Price (\$) |
|------------|------------|
| 650 | 772,000 |
| 785 | 998,000 |
| 1200 | 1,200,000 |
| 720 | 800,000 |
| 975 | 895,000 |

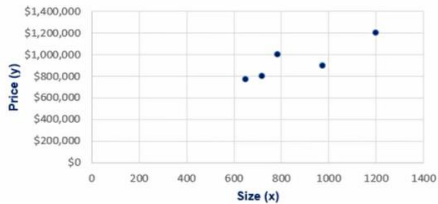


The two variables are correlated and the main statistic to measure this correlation is called covariance

Covariance

Housing data

| x | y | |
|-------------|------------|---------------------------|
| Size (ft.) | Price (\$) | $(x-\bar{x})*(y-\bar{y})$ |
| 650 | 772,000 | 34,776,000 |
| 785 | 998,000 | -5,265,000 |
| 1200 | 1,200,000 | 89,178,000 |
| 720 | 800,000 | 19,418,000 |
| 975 | 895,000 | -4,142,000 |
| Mean | 866 | 933,000 |
| Sum | | 133,965,000 |
| Sample size | | 5 |
| Cov. Sample | | 33,491,250 |



Covariance gives a sense of direction

> 0, the two variables move together

< 0, the two variables move in opposite directions

= 0, the two variables are independent

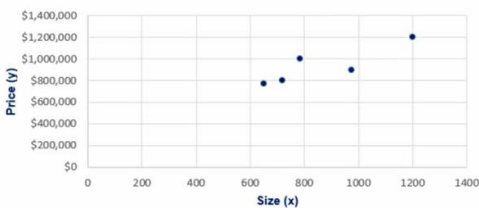
correlation coefficient

Correlation adjusts covariance, so that the relationship between the two variables becomes easy and intuitive to interpret

Correlation coefficient

Housing data

| Size (ft.) | Price (\$) | |
|---------------|------------|---------------------------|
| Size (ft.) | Price (\$) | $(x-\bar{x})*(y-\bar{y})$ |
| 650 | 772,000 | 34,776,000 |
| 785 | 998,000 | -5,265,000 |
| 1200 | 1,200,000 | 89,178,000 |
| 720 | 800,000 | 19,418,000 |
| 975 | 895,000 | -4,142,000 |
| Mean | 866 | 933,000 |
| Standard dev. | 222 | 173,615 |
| Sum | | 133,965,000 |
| Sample size | | 5 |
| Cov. Sample | | 33,491,250 |



$$\frac{Cov(x, y)}{Stdev(x) * Stdev(y)}$$

$$\downarrow \quad \downarrow$$

$$\frac{S_{xy}}{S_x S_y} \quad \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$-1 \leq \text{correlation coefficient} \leq 1$$

Correlation coefficient

Housing data



There is a **STRONG** relationship between the two variables

365 DataScience

PERFECT POSITIVE CORRELATION



Correlaton coeff. = 1

the entire variability of one variable is explained by the other

CORRELATION OF 0

Absolutely independent variables



Coffee in Brazil



Houses in London

They have nothing in common!

NEGATIVE CORRELATION

Perfect negative correlation of -1

Imperfect negative correlation: $(-1, 0)$

CORRELATION

The diagram illustrates the commutative property of correlation. It features two large teal U-shaped magnets. The left magnet has a white 'X' on its top-left pole and a white 'Y' on its top-right pole. To its right is a white equals sign. To the right of the equals sign is another large teal U-shaped magnet, which has a white 'Y' on its top-left pole and a white 'X' on its top-right pole. This visualizes that the correlation between X and Y is the same as the correlation between Y and X.