MAKE SCHOOL

# Percentile, Correlation and Covariance

QL 1.1

# Learning Outcomes

MAKE SCHOOL

By the end of today, you should be able to...

1. Understand percentiles and how to use them to describe data
2. Understand correlation and covariance and to use them to describe the relationship between different variables
3. Implement these concepts algorithmically using various libraries

# Warm up

5 mins

1. Download the ‘ramen rating’ dataset from Kaggle.
2. Find the mean and standard deviation of the ‘Stars’ column.

Solution

# Percentile


You are at the **80th percentile**. It means 80 percent of people in this group are shorter than you.

# Percentile

1. Your **percentile** tells you how you did on the **SAT** compared with everyone else who took the test. For example, if you got a composite **percentile** of 76, this means what?
    - you scored higher than 76% of students on the whole **test**.
2. If you got a **percentile** of 47 on the Math section, you did
    - better than 47% of students on **SAT** Math

# Why we do care about percentile?

MAKE SCHOOL

- Percentiles report the relative standing of a particular value within a statistical data set.

- My current score on Stackoverflow is 6266. What does that mean to you? Does it mean I'm a prolific contributor in comparison with others or not?
  - It doesn't mean anything on its own. Even if I tell you the mean is 354, you still won't see where do I stand in comparison with others.
- What if I tell you my stackoverflow score is 94th percentile? 😎
  - (that means 94% of developers are below me).

# Percentile Definition

**Percentile**: the value below which a percentage of data falls.

Examples: Consider GPA of students in a school. If they ask you what value is for the 5th percentile, you would find the gpa below which 5 percents of students population falls. If that gpa is 2.3, it means 5% of students' gpa falls below 2.3.

# How to Calculate Percentile (nearest-rank method)

MAKE SCHOOL

Given a dataset we can calculate the nth percentile using the steps below:

1. **Sort** the data in **increasing order**
2. Find the index of the of the percentile value by calculating
   **index = ceiling((percent/100) * len(data))**

   $$i = [\frac{n}{100} \cdot N]$$

   **i: index of the target value in the array**
   If index become fractional, round it up (hence 'ceiling'). Ie 2.1 ➔ 3
   **n: nth percentile**
   **N: total number of elements**

3. Find the value that is located at the index
   **values[index]**

**Note**: There are other ways for calculating the percentile value leading to slightly different outcomes.

# Calculate by hand

(not programmatically)

5 mins

MAKE SCHOOL

Consider the list,

**{50, 15, 40, 20, 35}.**

1. What are the **5th, 30th, 40th, 50th and 100th percentiles** of this list?

2. What is the median value?

Answers are here.

# Percentile Algorithm (nearest-rank method)

**(15 mins)**

Please complete the following algorithm that calculates the percentile value of a dataset.

Exercise: Percentile Algorithm

# 10 mins break

# Correlation

# Correlation

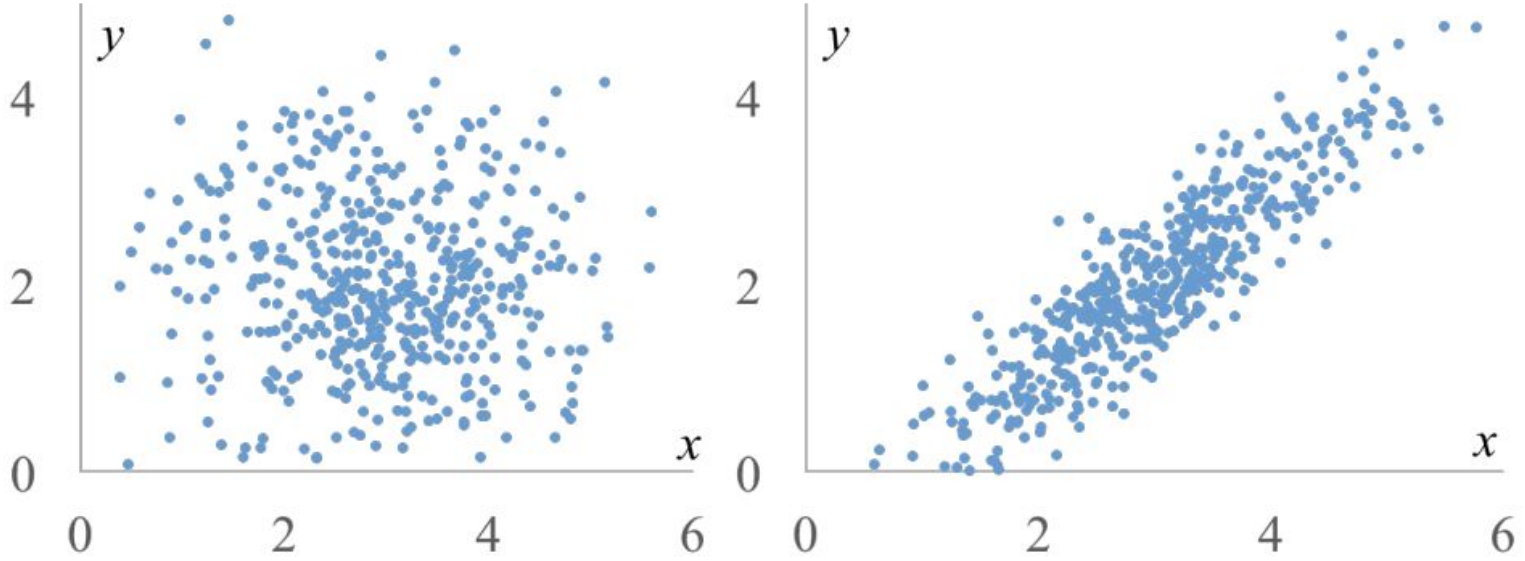
**Correlation** is the degree to which two or more attributes or measurements on the same group of elements show a tendency to vary together.

What moves with what? Which variables are "dependent" and which are "independent"? So what are some questions that we might want to find correlations?

1. Human weight correlates with human height. (Because bones are so heavy)
2. Size of vocabulary correlates with age up to adulthood, then the correlation evaporates.

# Correlation

**The mean and variance are the same in both the x and the y dimension. What is different?**



# Correlation vs. Causation

MAKE SCHOOL

Do not fall into the trap of mistaking correlation for causation.

The rain falls, the plants grow, thats causation because there is a causal link between the water of the rain and the plants growing. On the other hand, did you know that the rates of drowning have been known to jump when ice cream sales do?

# References


1. https://www.mathsisfun.com/data/percentiles.html
2. https://web.stanford.edu/class/archive/cs/cs109/cs109.1178/lectureHandouts/150-covariance.pdf