Introduction to Core Data Concepts

28.04.2025

What is data?

* Data is collected observations and information about something.

What are the common ways of distinguishing data?

Continuous vs Discrete (Categorical)

Continuous

* Can take any value.
* There are an "infinite" amount of values in-between any two values.
* Always numeric
* Height, weight...

Discrete (Categorical)

* Can only take certain values.
* There are no "in-between" values
* Can be numeric or string.
* Playing card values, car models, possible values of rolling a dice...

Nominal vs Ordinal

Nominal

* Has no natural order or rank.
* Can not be sorted.
* Animal categories: Dogs, cats, horses...

Ordinal

* Has a natural order.
* Can be sorted.
* Passenger classes on a plane, weather data such as "cold/mild/hot.

Structured vs Unstructured

Structured

* Highly specific and stored in a predefined format.
* Excel spreadsheets, XML files, JSON files...

Unstructured

* Not stored in a particular format.
* Audio, video, or text data.

---

* Data is collected observations and information on something.
* Data can be continuous vs. discrete, nominal vs. ordinal, structured vs. unstructured, on population vs. on a sample of the population.

Population	vs	Sample
* Consists of every member of a group (dependent on the context)		* Subset of the members of the group.
* Entire available data set.		* Needs to be representative of the population.
* List of all student names.		* Survey with a group of students in a school.

Measurements of Central Tendency

28.04.2025

Mean

* Where is the "center" of this data set?

* $$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{\text{sum of all data points}}{\text{number of data points}}$$

* μ (mu) = Mean of the population
* $\bar{x}$ (x-bar) = Mean of a sample of the population
* Formula is the same for both μ and $\bar{x}$

* Mean: The balancing point of the data

* Weighted Mean: when some data points are more significant

Dept Size	20	7	13	25
mean sal.	8.4	6.1	9.1	7.8

* $$\frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i} = \frac{20(8.4) + 7(6.1) + 13(9.1) + 25(7.8)}{20+7+13+25} \cong 8.1$$

* Truncated Mean: When we have data with outliers.

~~16~~ 18 21 27 32 32 33 ~~91~~ ???

↓ because we deleted "91"

* we need to state that we took away 25% of our dataset.

* Not always a good way to deal outliers.

---

* Mean, median, and mode are the measures of central tendency.
* Mean (μ for the population, $\bar{x}$ for the sample) = $\frac{\sum_{i=1}^{n} x_i}{n}$
* Weighted mean = when some data points are more significant: $\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Mode

* The value that occurs the most often in the data set.

Ex: 16 18 21 27 32 32 33 91

"32 is the mode of the data set"

* Bimodal: When we have two modes:

16 18 18 21 27 32 32 33 91

* No-mode: Every value occurs the same amount of time

16 18 21 27 32 33 91 ⇒ No mode!

Median

* When we order the data set, the value in the middle. If data set has even number of data points, median is the average of the two in the middle.

16 18 21 (27) 32 33 91

median

16 18 21 27 32 32 33 91

$\frac{27+32}{2}$ = median = 29.5

When can we use which measure of central tendency?

	mean	median	mode
Continuous	✓	✓	✓
Discrete	✓	✓	✓
Nominal	maybe	✗	✓
Ordinal	maybe	✓	✓
Numeric	✓	✓	✓
Non-numeric	✗	✓	✓

maybe? If numeric, yes. If non-numeric, no.

* Mode: The value that occurs the most often in the dataset.
* Median: When we order the data set, median is the value in the middle (or mean of the two). Helpful when the data is highly-skewed.

When should we use which measure of central tendency

* Not very easy to tell. In general, median is better than mean if there are outliers, e.g. household income.
* Remember, we're trying to answer the question "Where is the center of this data?" So we should assess if the measure we choose makes sense in this context. Imagine there's a marathon and there's a time limit. We might not even know the times of the participants who didn't make under the limit. If we take the mean of those who finished under the limit, it doesn't tell us the "center" for all participants. In this case, since we know the number of participants, we can easily find the median (assuming people who finished under the time limit are not out-numbered 🙂)

Quiz (1) Dataset = {329, 479, 459, 269, 369, 799, 649}

What's the median?

Dataset_ordered = {269, 329, 369, 459, 479, 649, 799}

Measures of Dispersion

04.05.2025

* "How much is the data spread out around the mean?"

	Population N	Sample n
mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$ (deviation)	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ (deviation)

* "mean distance from the mean itself"
* "on average, how far is each data point away from the center of the dataset?"

(?) Why squared?

(1) So $(x_i - \mu)^2$ is always positive. When we add them, we don't worry about "accidentally subtracting."

(2) So that bigger deviations are more significant than smaller deviations. (3 becomes 9, 9 becomes 81)

(?) Why N vs $n-1$? * To try to unbalance the bias we introduced when we minimized the deviation in our sample by calculating $\bar{x}$. ($\bar{x}$ is biased because it depends on the sample we use.)

	Population N	Sample n
standard deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

* Standard Deviation = $\sqrt{\text{variance}}$ =

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}}$$

* we will use standard deviation instead of variance almost always b.c. the units of std. dev. will match the units of the data.

* Heights = cm ⇒

mean = cm, variance = cm^2, std. dev. = cm

Ex:

ind.	A	B	C	D	E
heights	1	2	3	4	5

⇒ sample data $n=5$

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

$$s = \sqrt{\frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5-1}} = \sqrt{\frac{4+1+0+1+4}{4}} = \frac{\sqrt{10}}{2} \cong 1.58$$

Quiz: top = [220.1, 220.0, 220.0, 220.2, 220.1]

bottom = [220.1, 220.4, 220.2, 220.0, 220.1]

a.) is $s_{top} \leq 0.1$? b) is $s_{bottom} \leq 0.01$

$$\bar{x}_{top} = \frac{220.1, 220.0, 220.0, 220.2, 220.1}{5} = \frac{1100,4}{5} = 220.08$$

$$s_{top} = \sqrt{\frac{(0.02)^2+(0.08)^2+(0.08)^2+(0.12)^2+(0.02)^2}{5-1}} = 0.0837$$

≤ 0.1 ✓

$$\bar{X}_{bottom} = \frac{220.1, 220.4, 220.2, 220.0, 220.1}{5} = 220.16$$

$$s_{bottom} = \sqrt{\frac{(0.06)^2 + (0.24)^2 + (0.04)^2 + (0.16)^2 + (0.06)^2}{4}} = 0.1517$$

≤ 0.1 X

Quartiles and IQR

05.05.2025

* Golf scores of 18 golfers:

Q_1 = 68, 69 = median, Q_3 = 71

66, 67, 67, 68, **68**, 68, 68, 69, **69**, 69, 69, 70, 70, **71**, 71, 72, 73, 75

- Q_1: First quartile, Lower quartile, 25th percentile
- Q_2: Second quartile, Median, 50th percentile
- Q_3: Third quartile, Upper quartile, 75th percentile

* No universal method to calculate Q_1 and Q_3. Commonly:
  * If n is even: Find the median of lower and upper halves.
  * If n is odd: Exclud the median and find the median of lower and upper halves.
* Range = max − min = 75 − 66 = 9
* IQR (Interquartile Range) = $Q_3 - Q_1$ = 71 − 68 = 3

Five number summary

*

Min	Q_1	Median	Q_3	Max
66	68	69	71	75

IQR = 3

* Gives us an idea of the center and the spread of our data set at one glance.

Outliers

* One common method to classify a data point as an outlier:

$$\text{Outlier}(x) = \begin{cases} \text{True} & \text{if } x < Q_1 - 1.5 \times IQR \text{ or } x > Q_3 + 1.5 \times IQR \\ \text{False} & \text{otherwise} \end{cases}$$

---

* Five-Number Summary = Min − Q_1 (25%) − Median (Q_2) (50%) − Q_3 (75%) − Max
* $IQR = Q_3 - Q_1$
* $$\text{Outlier}(x) = \begin{cases} \text{True} & \text{if } x < Q_1 - 1.5 \times IQR \text{ or } x > Q_3 + 1.5 \times IQR \\ \text{False} & \text{otherwise} \end{cases}$$

Quiz ① Find the Five-Number summary and the outliers (if any)

$a = [10, 12, 15, 18, 20, 21, 22, 24, 25, 27, 28, 30, 32, 35, 38, 40, 42, 45]$

$n = 19$ $\quad Q_1 = 20$ $\quad IQR = 38 - 20 = 18$

median $= 27$ $\quad Q_3 = 38$ $\quad 1.5 \times IQR = 27$

$$\text{Outliers}(x) = \begin{cases} \text{True} & \text{if } x < -7 \text{ or } x > 65 \\ \text{False} & \text{otherwise} \end{cases}$$

NO OUTLIERS!