Mean, Median, Mode & Standard Deviation (Chapter 3)

Measure of **central tendency** is a value that represents a typical, or central, entry of a data set. The most common measures of central tendency are:

- **Mean (Average):** The sum of all the data entries divided by the number of entries.

  Population Mean: $\mu = \dfrac{\Sigma x}{N}$ Sample Mean: $\bar{x} = \dfrac{\Sigma x}{n}$
- **Median:** The value that lies in the **middle** of the data when the data set is **ordered.** If the data set has an odd number of entries, then the median is the middle data entry. If the data has an even number of entries, then the median is obtained by adding the two numbers in the middle and dividing result by two.
- **Mode**: The data entry that occurs with the **greatest frequenc**y. A data set may have one mode, more than one mode, or no mode. If no entry is repeated the data set has no mode.
- **Outliers** are not just greatest and least values, but values that are very different from the pattern established by the rest of the data. Outliers affect the mean. When outliers are present it is best to use the median as the measure of central tendency.

Measures of Variation:

- **Range:** The difference between the maximum and minimum data entries in the set. Range = (Max. data entry) – (Min. data entry)

- The **standard deviation** measure **variability** and **consistency** of the sample or population. In most real-world applications, consistency is a great advantage. In statistical data analysis, less variation is often better.

  Population Standard Deviation $= \sigma = \sqrt{\dfrac{\sum (x-\mu)^2}{N}}$ Sample Standard Deviation $= s = \sqrt{\dfrac{\sum (x-\bar{x})^2}{n-1}}$
- Example: Find Population Mean and Sample Standard Deviation for the following data set: 5, 10, 15, 20

$$\bar{x} = \frac{\Sigma x}{n} = \frac{5+10+15+20}{4} = \frac{50}{4} = 12.5$$

Data	$\boldsymbol{x - \bar{x}}$	$(\boldsymbol{x - \bar{x}})^2$
5	**5 — (12.5) = —7.5**	$(-7.5)^2 = 56.25$
10	**10 — (12.5) = —2.5**	$(-2.5)^2 = 6.25$
15	**15 — (12.5) = 2.5**	$(2.5)^2 = 6.25$
20	**20 — (12.5) = 7.5**	$(7.5)^2 = 56.25$
		$\boldsymbol{\Sigma(x - \bar{x})^2 = 125.01}$

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{125.01}{4-1}} \approx 6.455$$

Percentiles, Five-number Summary & Z-Scores (Chapter 3)

- **Percentiles**: Let p be any integer between 0 and 100. The *pth* percentile of data set is the data value at which p percent of the value in the data set are less than or equal to this value.
- **How to calculate percentiles**: Use the following steps for calculating percentiles for small data sets. For large data sets, we use computers to find percentiles.
  - Step 1: Sort the data in ascending order (from smallest to largest)
  - Step 2: Calculate $i^{th} = \left(\frac{p}{100}\right)n$, where p is the particular percentile you wish to calculate and n is the sample size.
  - Step 3: If i is an integer, the pth percentile is the mean of the data values in positions i and $i+1$. If i is not an integer, then round up to the next integer and use the value in this position.

Example: Use the following set of stock prices (in dollars): 10, 7, 20, 12, 5, 15, 9, 18, 4, 12, 8, 14 Find the 10th percentile and the 50th percentile

Solutions:

- First sort the data in ascending order: **4, 5, 7, 8, 9, 10, 12, 12, 14, 15, 18, 20**
- There are 12 scores so, n = 12.
- To find the 10th percentile, we use the formula

$$i^{th} = \left(\frac{p}{100}\right)n = \left(\frac{10}{100}\right)12 = 1.2 \approx \text{Round Up}(1.2) = 2$$

- The 10th percentile is the number in the 2nd position.

Position	1st	**2nd**	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th
Data	4	**5**	7	8	9	10	12	12	14	15	18	20

$$\text{10th Percentile} = P_{10} = 5$$

- To find the 50th percentile, we use the formula

$$i^{th} = \left(\frac{p}{100}\right)n = \left(\frac{50}{100}\right)12 = 6$$

- We need to find the 6th and 7th numbers in the sorted data set.
- Since the answer is an integer, we need to find the 6th and 7th number in the data set.

Position	1st	2nd	3rd	4th	5th	**6th**	**7th**	8th	9th	10th	11th	12th
Data	4	5	7	8	9	**10**	**12**	12	14	15	18	20

$$\text{50th Percentile} = P_{50} = \frac{10+12}{2} = 11$$

Box-and-whisker plot: Requires (**five-number summary**):

- Minimum entry
- First quartile= $Q_1 = P_{25}$
- Median= $Q_2 = P_{50}$
- Third quartile = $Q_3 = P_{25}$
- Maximum entry