Preliminaries:
Descriptive Statistics

CS 418. Introduction to Data Science

Preliminaries
Descriptive vs. Inferential Statistics

- **Statistics** deals with the collection, analysis, interpretation, and presentation of data.
- There are two types of statistical analysis: **descriptive statistics** and **inferential statistics.**
- **Descriptive statistics** focuses on **summarizing data** about a sample drawn from a population.
  - **Measures of central tendency**:
    - Mean, median, mode.
  - **Measures of dispersion or spread**:
    - Range, standard deviation, variance.
- **Inferential statistics** focuses on using information from the sample to **make conclusions** about the population from which the sample was drawn.

Measures of Center
Mean

- The **arithmetic mean** (or **average**) of a dataset is the **sum of the values** in the dataset **divided by the number of values**.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

$a_1, a_2, \ldots, a_n$ are the values in the dataset and n is the number of values

- ***Example*:**

Age	24	35	32	21	28

What is the mean age?

$$\frac{24 + 35 + 32 + 21 + 28}{5} = 28$$

In Python (with **pandas**):

```
DataFrame.mean()
```

Click for documentation

Measures of Center
Median

- The **median** of a dataset is the **middle value** (if the number of values is odd) or the **mean of the two middle values** (if the number of values is even) in the sorted dataset.
- ***Example*:**

Grades
78
90
90
80
78
95
76
84

Sorted dataset:

76	78	78	80	84	90	90	95

What is the median grade? 82

What is the median grade if we remove 95? 80

In Python (with **pandas**):

```
DataFrame.median()
```

Click for documentation

Preliminaries
Exercise 2.1

THINK PAIR SHARE

Which measure of center (*mean* or *median*) better summarizes the following dataset? Why?

The **mean** salary is $146,750.

The **median** salary is $73,000.

Salaries
75,000
70,000
72,000
65,000
675,000
75,000
68,000
74,000

$675,000 is an **outlier**.

An **outlier** is a value that is much higher or much lower than the rest of the values in the dataset.

The **mean** is **not robust** to outliers.

The **median** is a better summary of this dataset.

Measures of Center
Mode

- The **mode** of a dataset is the **most frequent value** in the dataset.
- ***Example*:**

Grades
78
90
90
80
78
95
76
84

What is the mode? 78, 90

There can be more than one mode in a dataset.

A dataset with two modes is called **bimodal**

In Python (with **pandas**):

```
DataFrame.mode()
```

Click for documentation

Measures of Spread
Maximum, Minimum, and Range

- The **maximum** of a dataset is the **largest value** in the dataset. The **minimum** of a dataset is the **smallest value** in the dataset.
- The **range** of a dataset is the **difference between the maximum and minimum** of the dataset.
- ***Example*:**

Grades
78
90
90
80
78
95
76
84

What is the minimum grade? 76

What is the maximum grade? 95

What is the range of grades? 19

In Python (with **pandas**):

```
DataFrame.max()
```

```
DataFrame.min()
```

Click for documentation

Measures of Spread
Standard Deviation

- The **standard deviation** of a dataset is a **measure of the distance between the values** in the dataset **and the mean of the values**.

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

$a_1, a_2, \ldots, a_n$ are the values in the dataset, n is the number of values, and $\bar{x}$ is the mean of the values

- ***Example***:

Age	24	35	32	21	28

What is the standard deviation of the ages?

$$\sqrt{\frac{(24-28)^2 + (35-28)^2 + (32-28)^2 + (21-28)^2 + (28-28)^2}{4}} = 5.7$$

Measures of Spread
Variance

- The **variance** of a dataset is **the square of the standard deviation**.
- ***Example*:**

Age	24	35	32	21	28

What is the variance of the ages?

$$\frac{(24-28)^2+(35-28)^2+(32-28)^2+(21-28)^2+(28-28)^2}{4}=32.5$$

In Python (with **pandas**):

```
DataFrame.std()
```

```
DataFrame.var()
```

Click for documentation

Measures of Spread
Percentiles and Quartiles (I)

- **Percentiles** divide a dataset into **100 equal parts** such that **n% of the data** is less than or equal to the **n^{th} percentile**.
- **Quartiles** divide a dataset into **quarters**.
  - **25% of the data** is less than or equal to the **first quartile** (Q_1) or **25th percentile**.
  - **50% of the data** is less than or equal to the **second quartile** (Q_2) or **50th percentile**.
    - Q_2 is also the **median** of the dataset.
  - **75% of the data** is less than or equal to the **third quartile** (Q_3) or **75th percentile**.
- The **minimum** and **maximum** values, **Q_1**, the **median**, and **Q_3** form a set of descriptive statistics called the **five-number summary**.

Measures of Spread
Percentiles and Quartiles (II)

- ***Example*:**

Grades
78
90
90
80
78
95
76
84

Sorted dataset:

76	78	78	80	84	90	90	95

What is the five-number summary of this dataset?

Minimum = 76

Q_1 = 78

Median = 82

Q_3 = 90

Maximum = 95

In Python (with **pandas**):

```
DataFrame.quantile()
```

Click for documentation

Preliminaries
Exercise 2.2

THINK PAIR SHARE

Understanding percentiles.

1. **Suppose that you are waiting in line at the DMV. Your wait time is in the 85th percentile. Is that good or bad?**

   BAD

2. **Suppose that your salary is in the 25th percentile. Is that good or bad?**

   BAD

3. **Suppose that you are buying a house in a neighborhood. The most expensive house that you can afford is in the 34th percentile. Does that mean that you can afford 34% of the houses or 66% of the houses in that neighborhood?**

   34%

Preliminaries
References

- Joel Grus. *Data Science from Scratch* (2015).
- OpenStax. Introductory Statistics (2016).