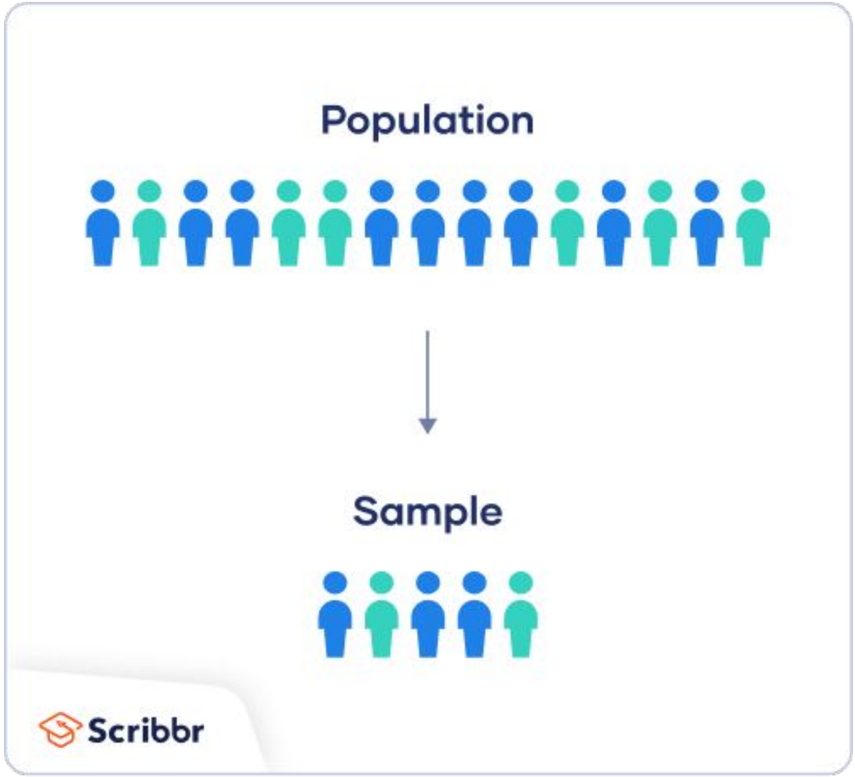
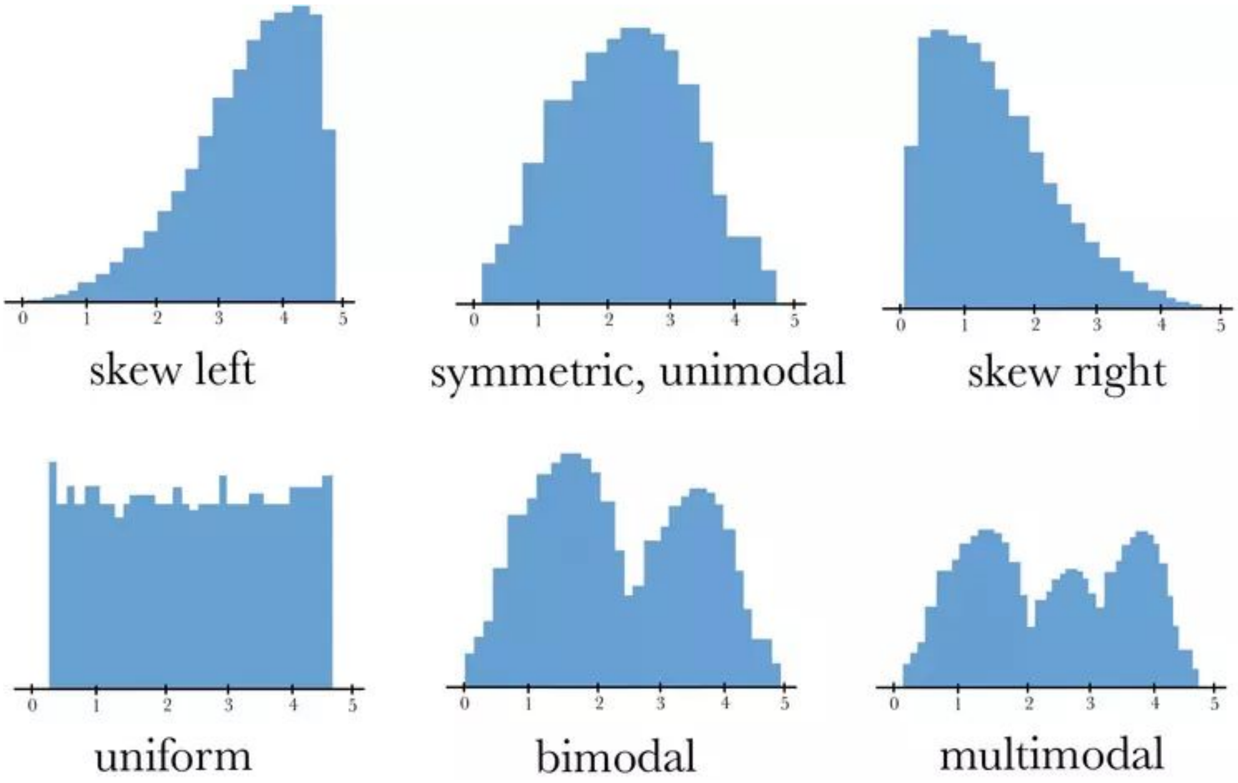
TUT206 Sep27

Recap



Recap



skew left

symmetric, unimodal

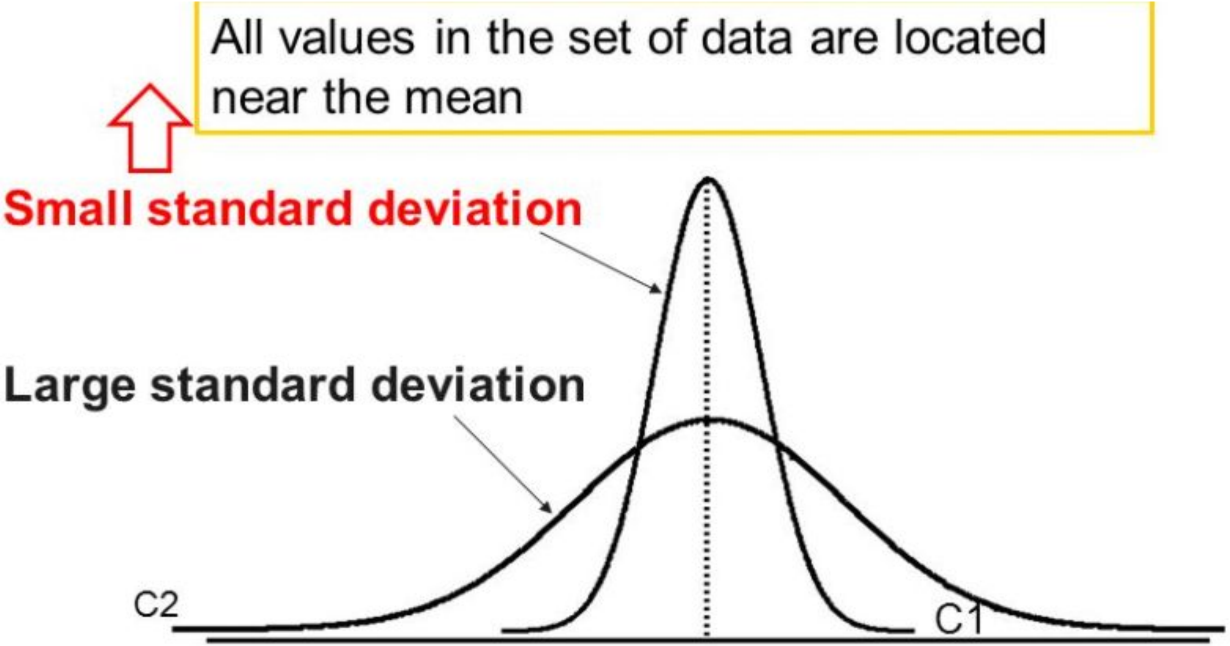
skew right

uniform

bimodal

multimodal

Recap

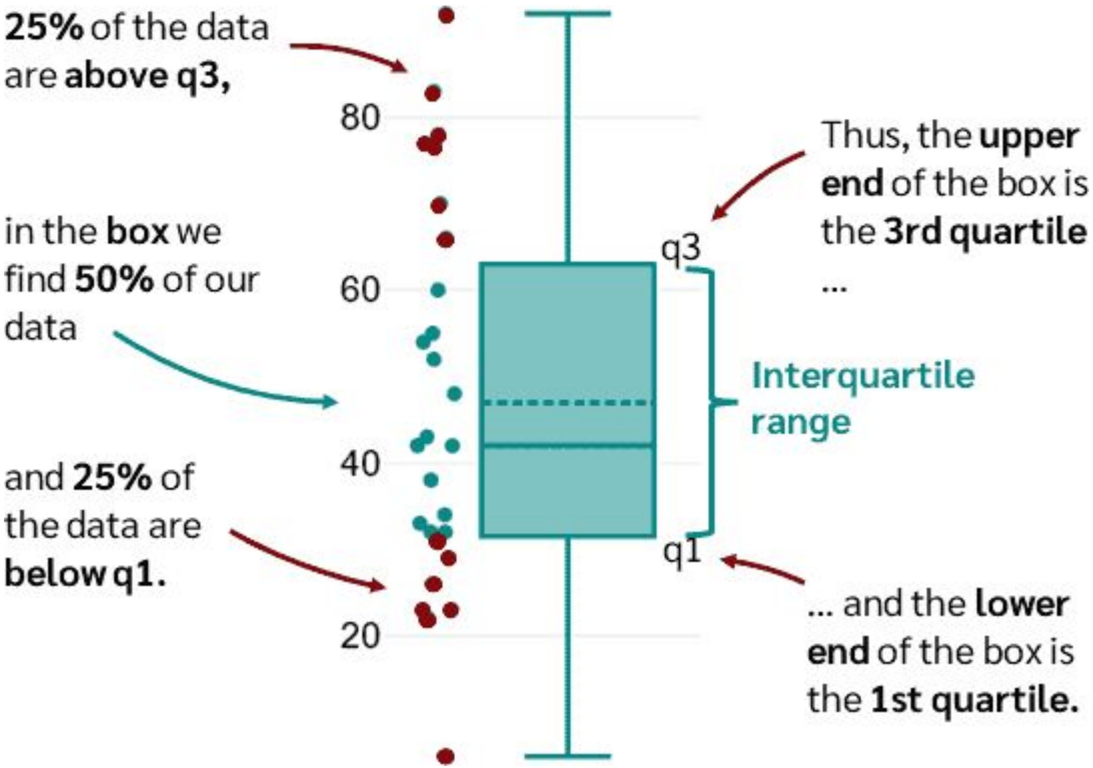


All values in the set of data are located near the mean

Small standard deviation

Large standard deviation

Recap



25% of the data are **above q3,**

in the **box** we find **50%** of our data

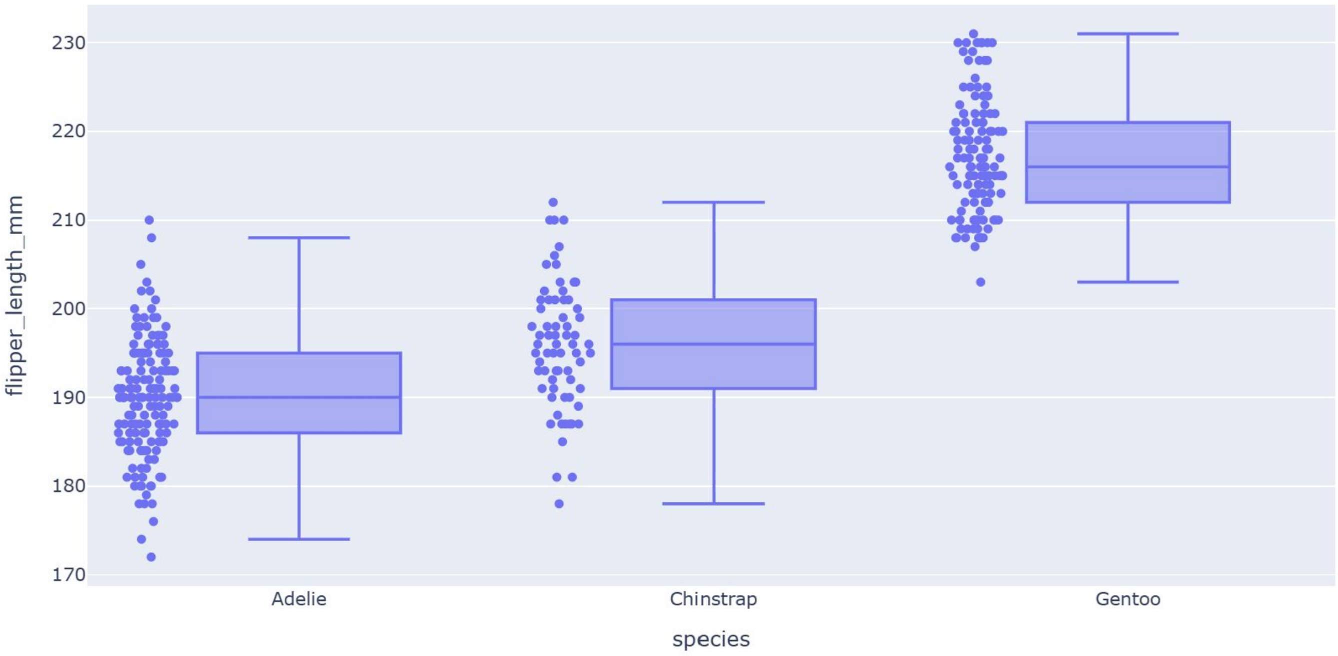
and **25%** of the data are **below q1.**

Thus, the **upper end** of the box is the **3rd quartile** ...

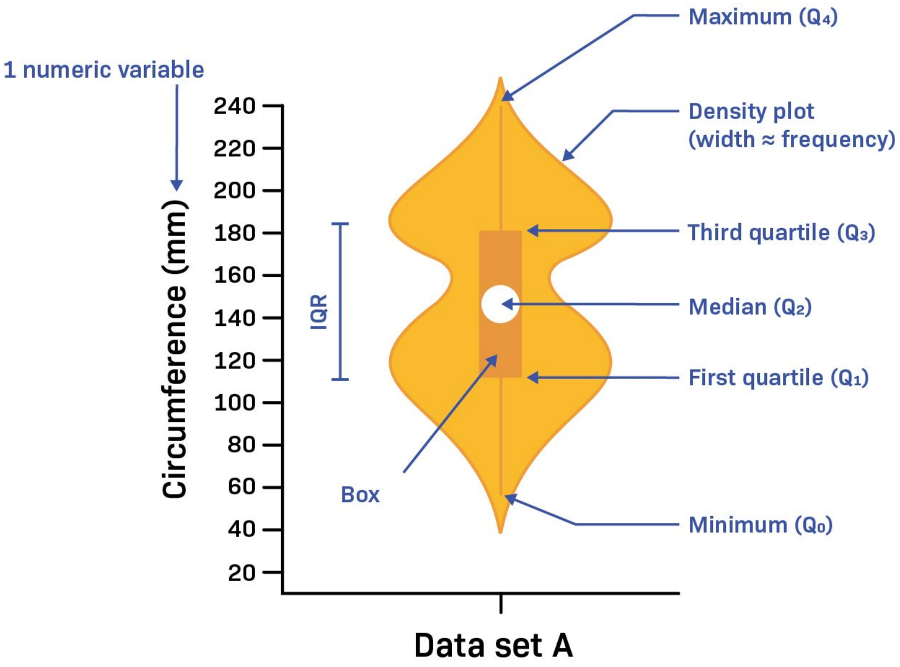
... and the **lower end** of the box is the **1st quartile.**

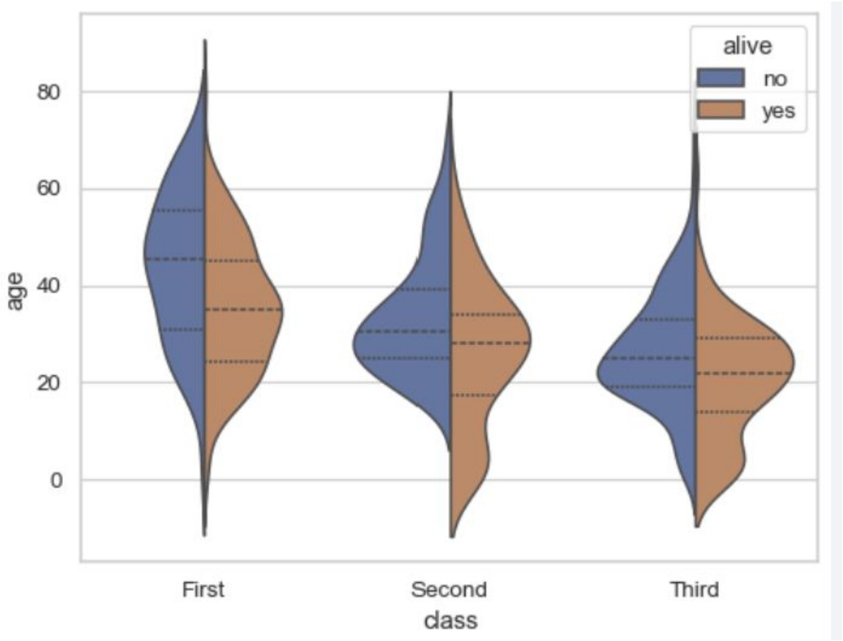
Recap

```
import plotly.express as px
fig.show()  # USE `fig.show(renderer="png")
```

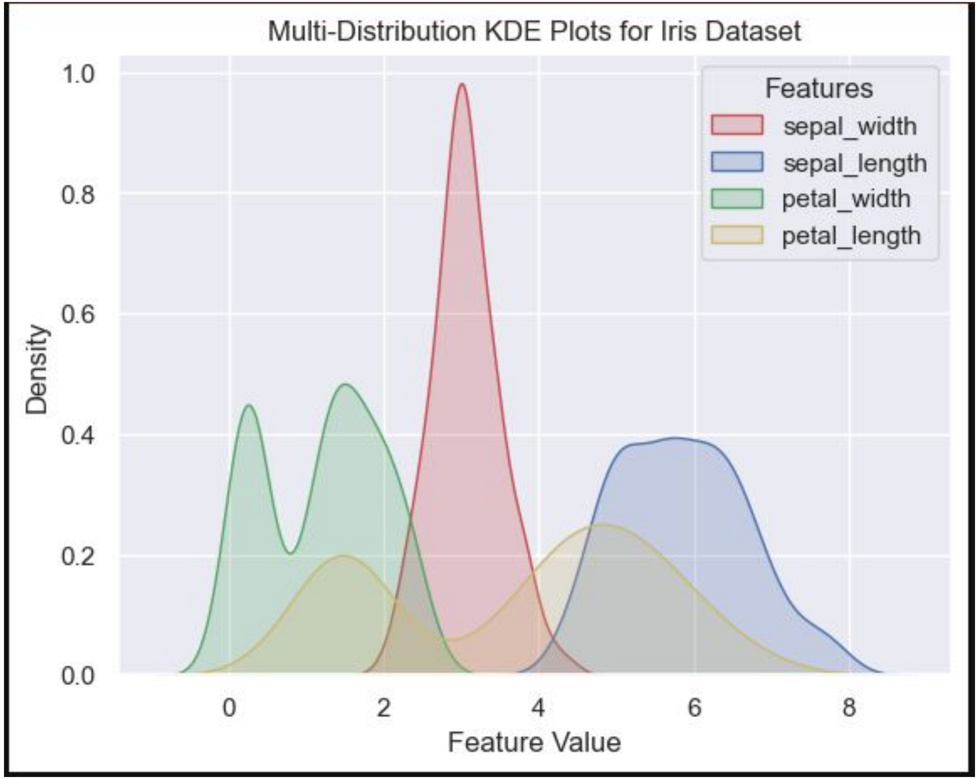


Recap


1 numeric variable
Circumference (mm)
240
220
200
180
160
140
120
100
80
60
40
20
IQR
Box
Data set A
Maximum (Q_4)
Density plot
(width ≈ frequency)
Third quartile (Q_3)
Median (Q_2)
First quartile (Q_1)
Minimum (Q_0)


alive
no
yes
age
80
60
40
20
0
First
Second
Third
class

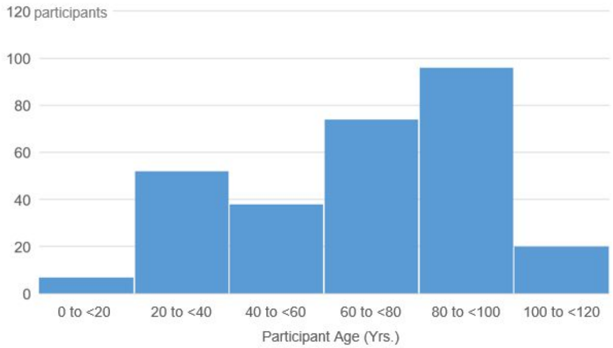
Recap



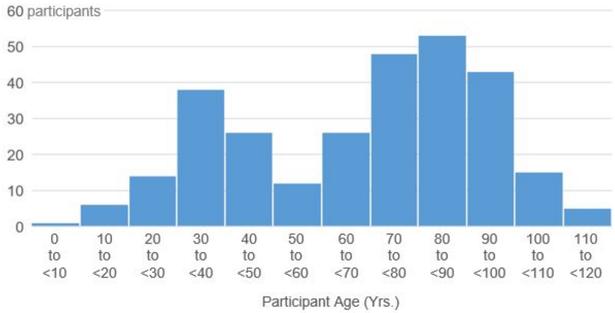
Multi-Distribution KDE Plots for Iris Dataset
Features
🟥 sepal_width
🟦 sepal_length
🟩 petal_width
🟨 petal_length
Density
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
Feature Value

Recap

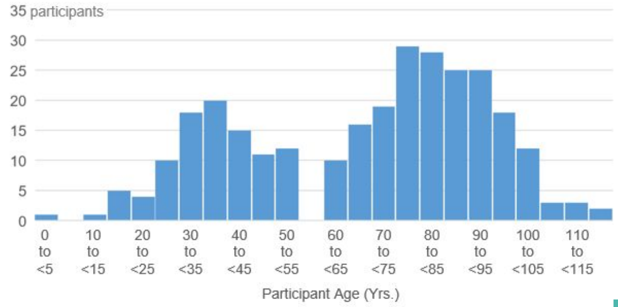
6 bins:



12 bins:



24 bins:

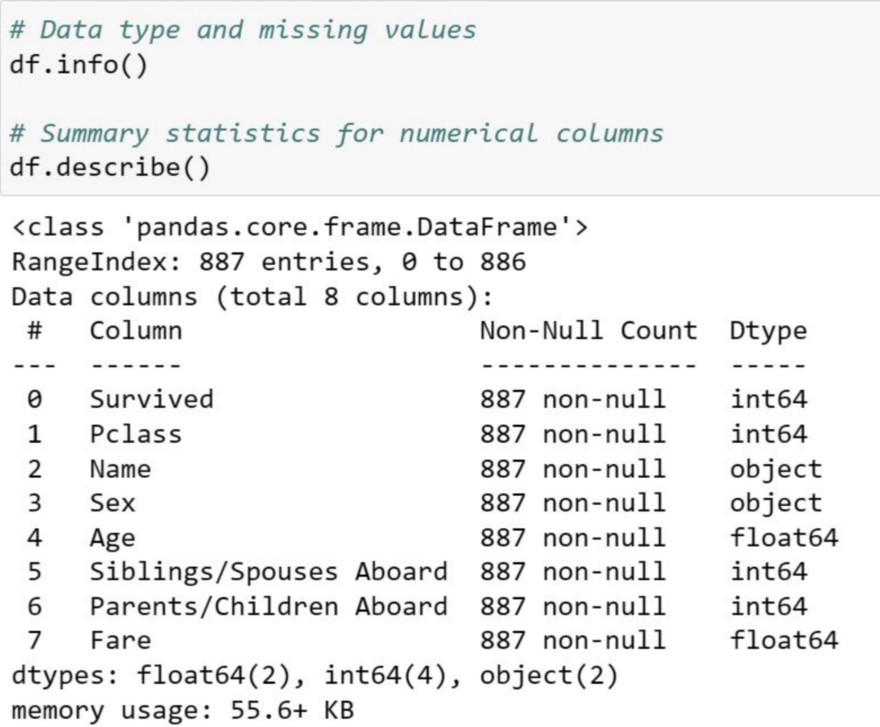


Discussion cont.

Last week: Break into 4 groups of 6 students and prepare a speech describing the **generic strategy or general sequence of steps you would take to understand a dataset**

This week: Go find an **interesting dataset** and use **summary statistics and visualizations** to understand and demonstate some interesting aspects of the data

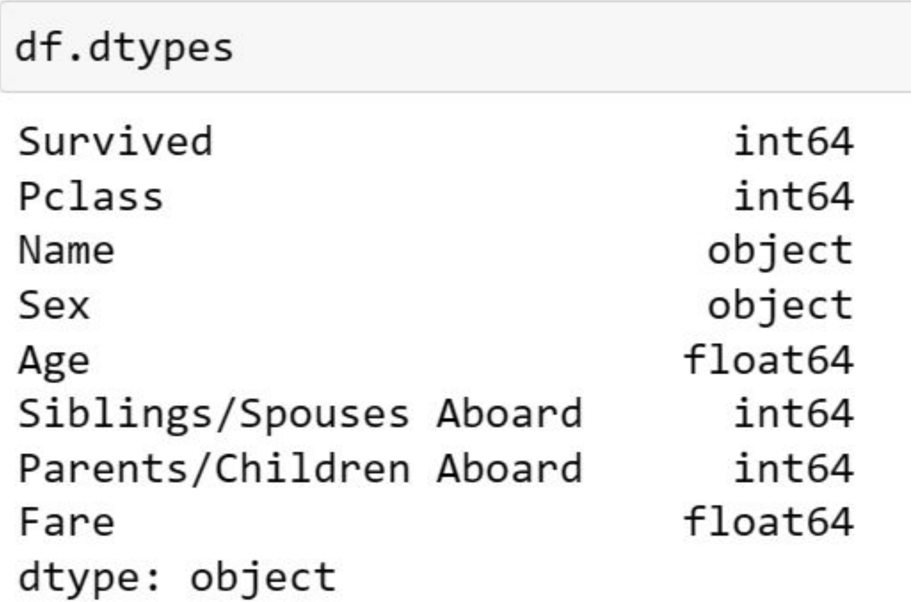
Discussion cont. HINT



```
# Data type and missing values
df.info()

# Summary statistics for numerical columns
df.describe()
```

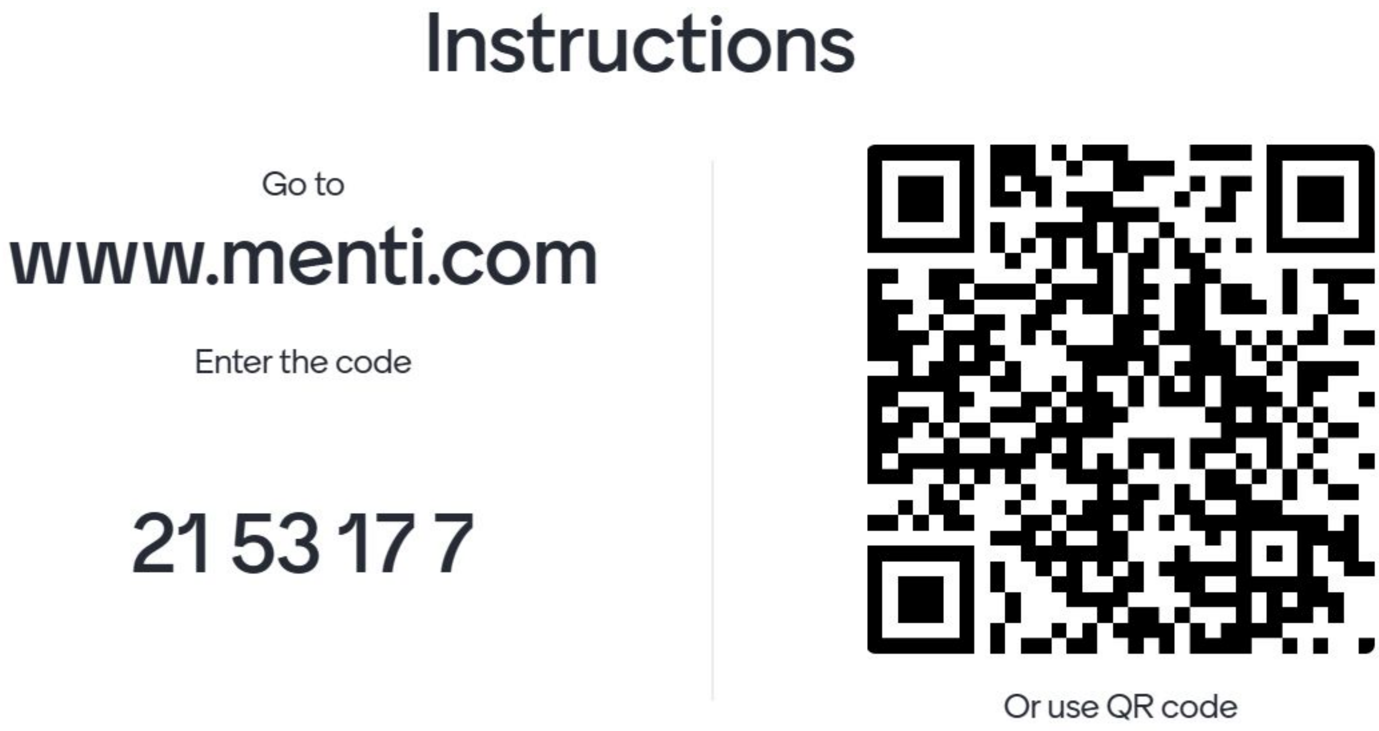
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 887 entries, 0 to 886
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Survived                 887 non-null    int64  
 1   Pclass                   887 non-null    int64  
 2   Name                     887 non-null    object 
 3   Sex                      887 non-null    object 
 4   Age                      887 non-null    float64
 5   Siblings/Spouses Aboard  887 non-null    int64  
 6   Parents/Children Aboard  887 non-null    int64  
 7   Fare                     887 non-null    float64
dtypes: float64(2), int64(4), object(2)
memory usage: 55.6+ KB
```



```
df.dtypes
```

```
Survived                     int64
Pclass                       int64
Name                        object
Sex                         object
Age                        float64
Siblings/Spouses Aboard      int64
Parents/Children Aboard      int64
Fare                       float64
dtype: object
```

Quiz!



Instructions

Go to

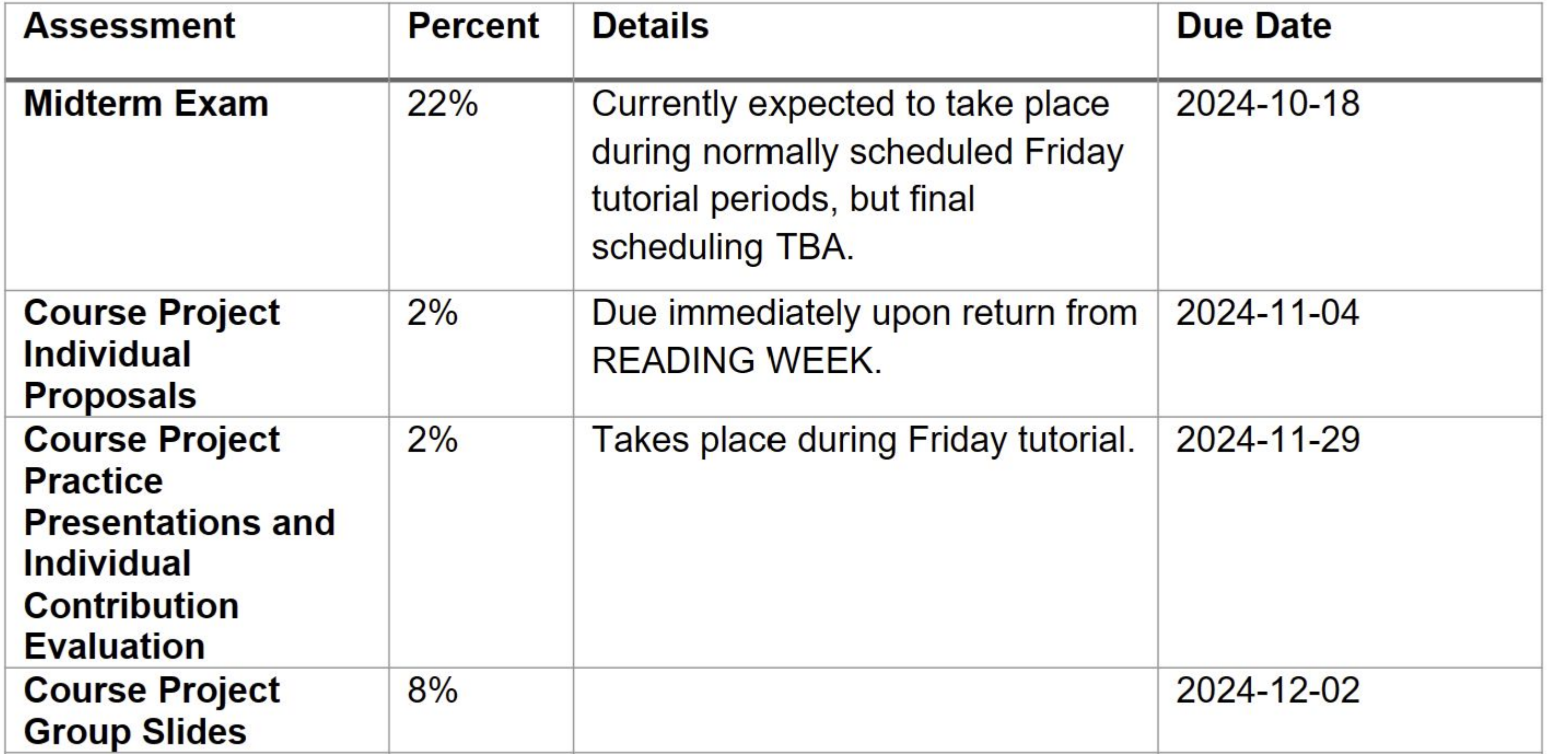
www.menti.com

Enter the code

21 53 17 7

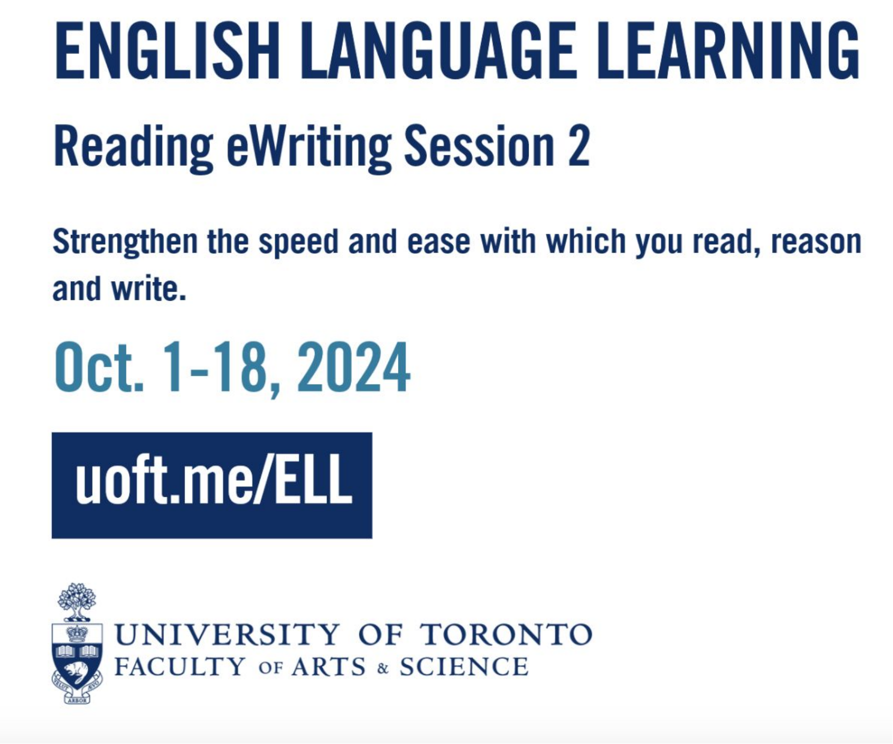
Or use QR code

Announcement



Assessment	Percent	Details	Due Date
Midterm Exam	22%	Currently expected to take place during normally scheduled Friday tutorial periods, but final scheduling TBA.	2024-10-18
Course Project Individual Proposals	2%	Due immediately upon return from READING WEEK.	2024-11-04
Course Project Practice Presentations and Individual Contribution Evaluation	2%	Takes place during Friday tutorial.	2024-11-29
Course Project Group Slides	8%		2024-12-02

Announcement



ENGLISH LANGUAGE LEARNING

Reading eWriting Session 2

Strengthen the speed and ease with which you read, reason and write.

Oct. 1-18, 2024

uoft.me/ELL

UNIVERSITY OF TORONTO
FACULTY OF ARTS & SCIENCE

Demo

https://colab.research.google.com/drive/1FCm24jj5s5PGWeOq-NhqzEyGDC_hd7fV?usp=sharing

Demo

For which countries do you think we can most accurately estimate the average 'points' score of cups of coffee?

Demo

How does the variability/uncertainty of means of simulated samples change as a function of sample size?

Beta Distribution PDF Grapher (eurekastatistics.com)

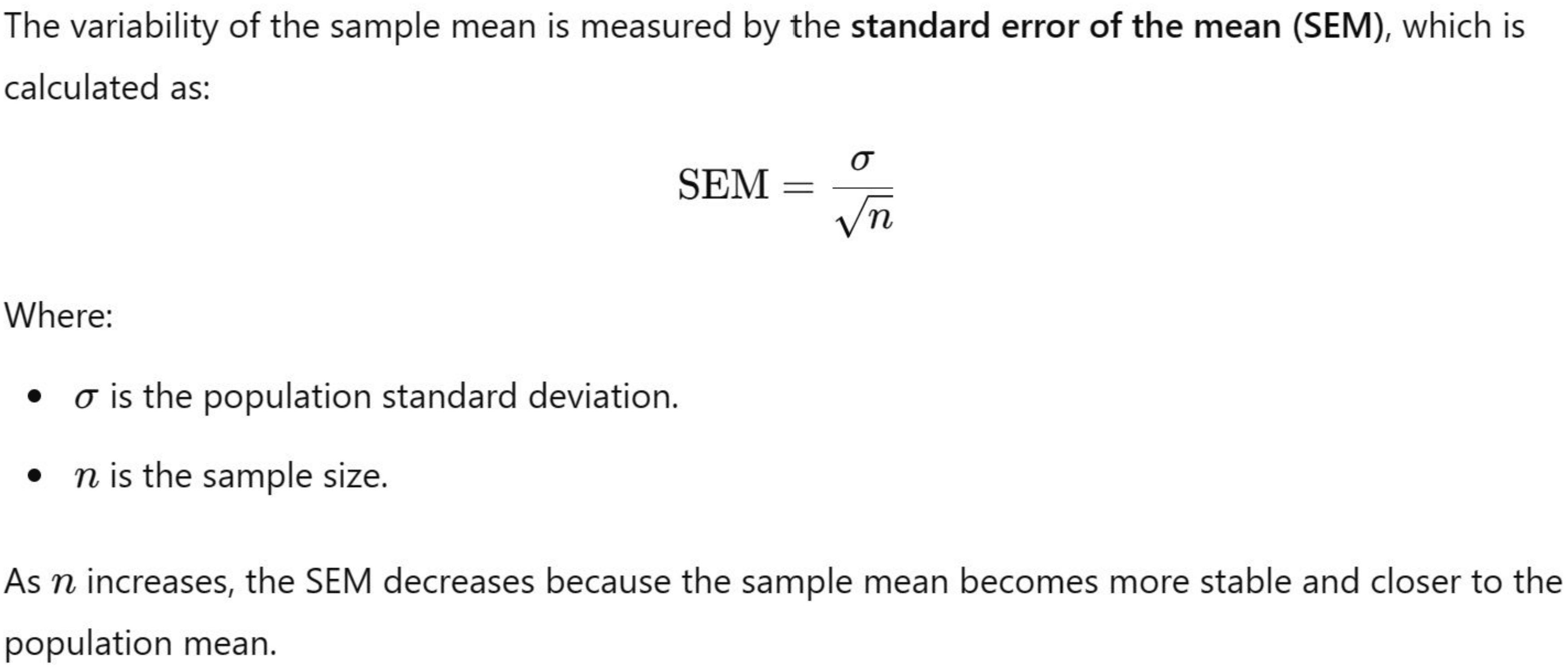
Demo- sampling

```
my_theoretical_sample =
my_theoretical_population.rvs(size=sample_size)
```

Demo - bootstrapping

```
my_bootstrapped_sample =
np.random.choice(penguins_noNaN.body_mass_g,
size=sample_size, replace=True)
```

Demo



The variability of the sample mean is measured by the **standard error of the mean (SEM)**, which is calculated as:

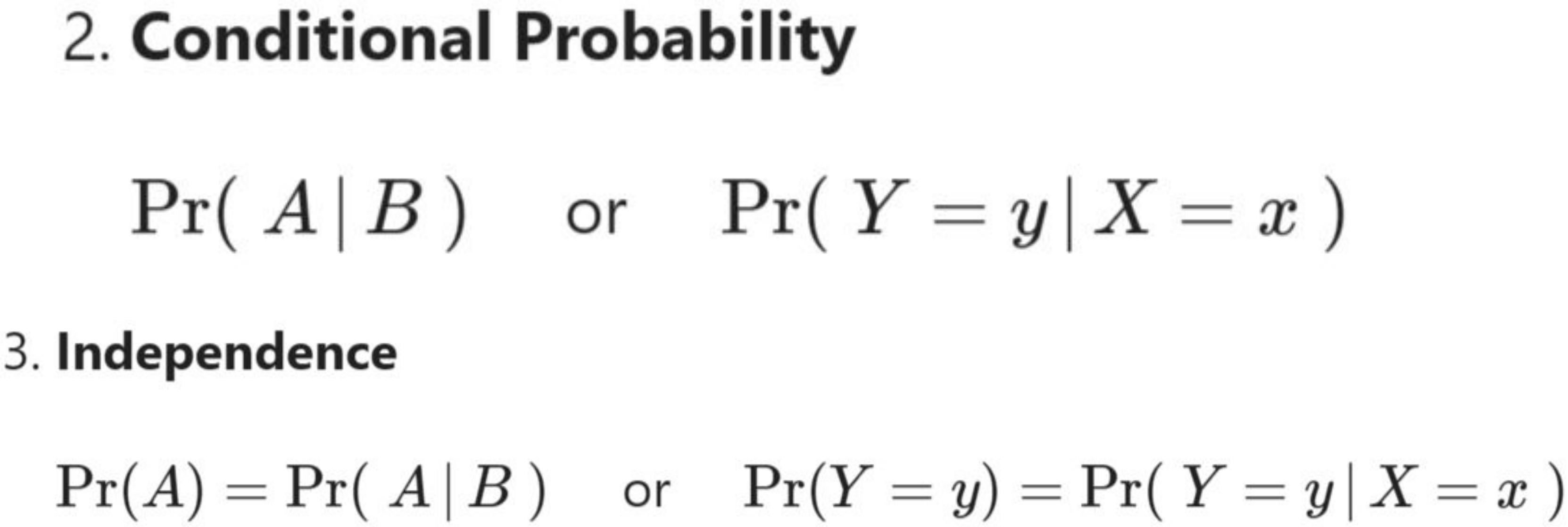
$$\text{SEM} = \frac{\sigma}{\sqrt{n}}$$

Where:

- σ is the population standard deviation.
- n is the sample size.

As n increases, the SEM decreases because the sample mean becomes more stable and closer to the population mean.

Midterm review



2. **Conditional Probability**

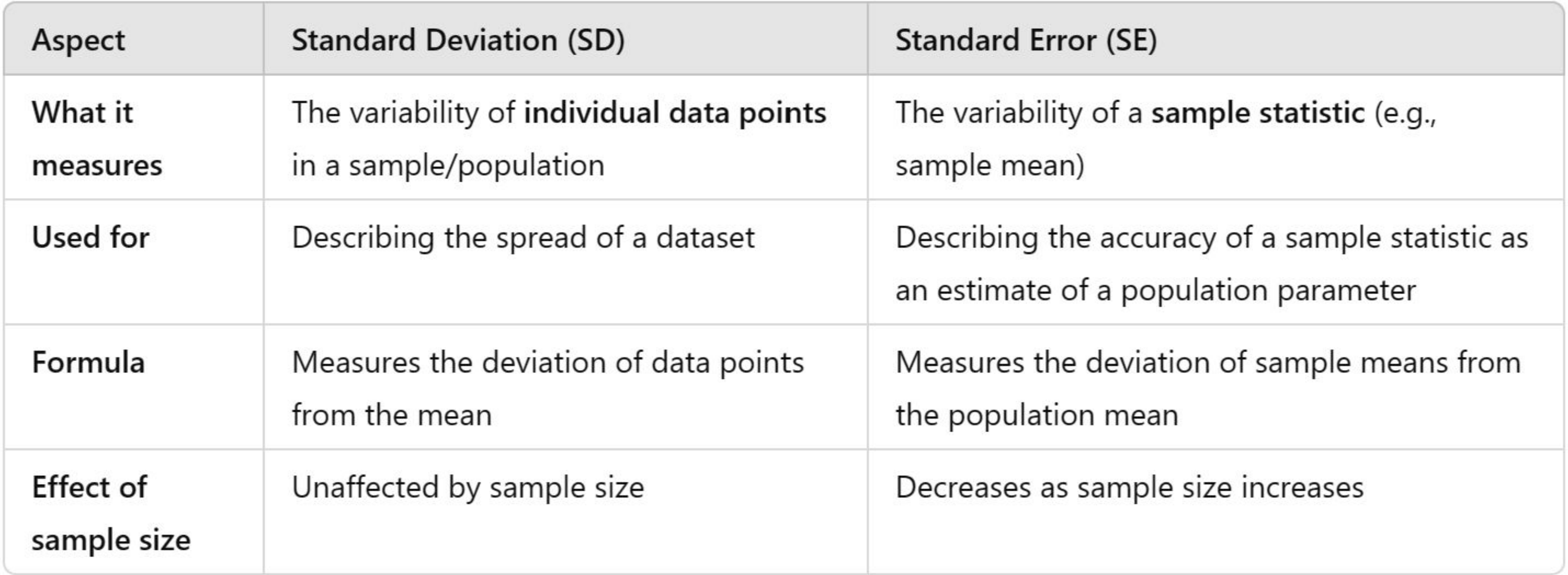
$$\Pr(\ A \,|\, B\ ) \quad \text{or} \quad \Pr(\ Y = y \,|\, X = x\ )$$

3. **Independence**

$$\Pr(A) = \Pr(\ A \,|\, B\ ) \quad \text{or} \quad \Pr(Y = y) = \Pr(\ Y = y \,|\, X = x\ )$$

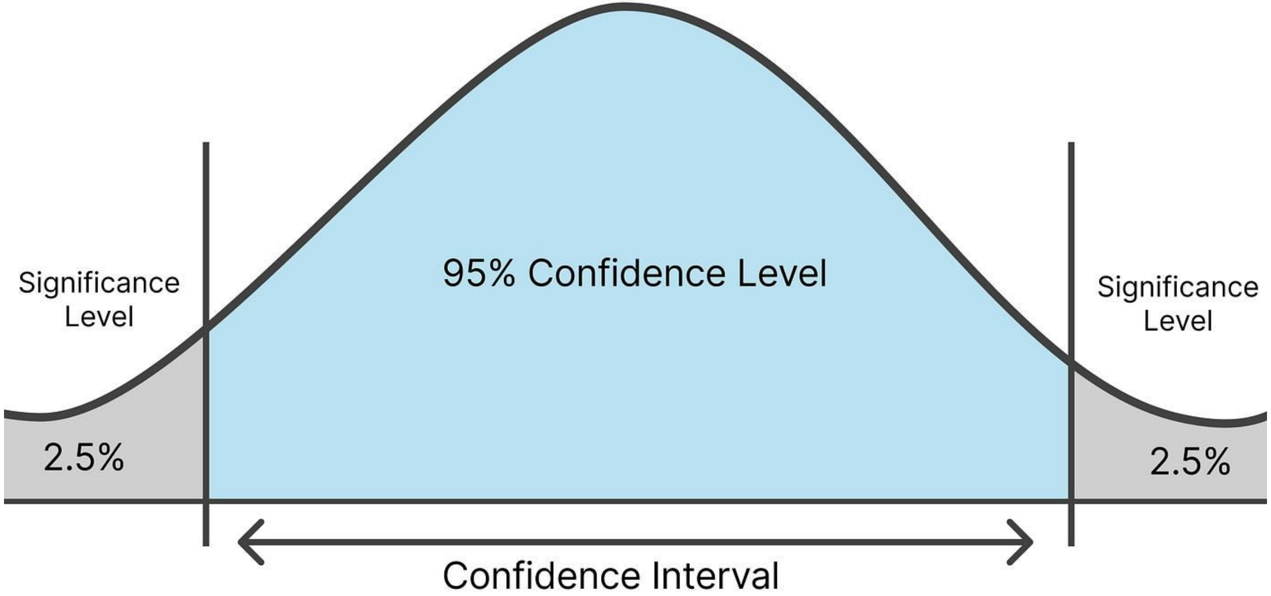
Midterm review

1. [4] In three or four complete English sentences, explain what a P-value is, and what it is used for.

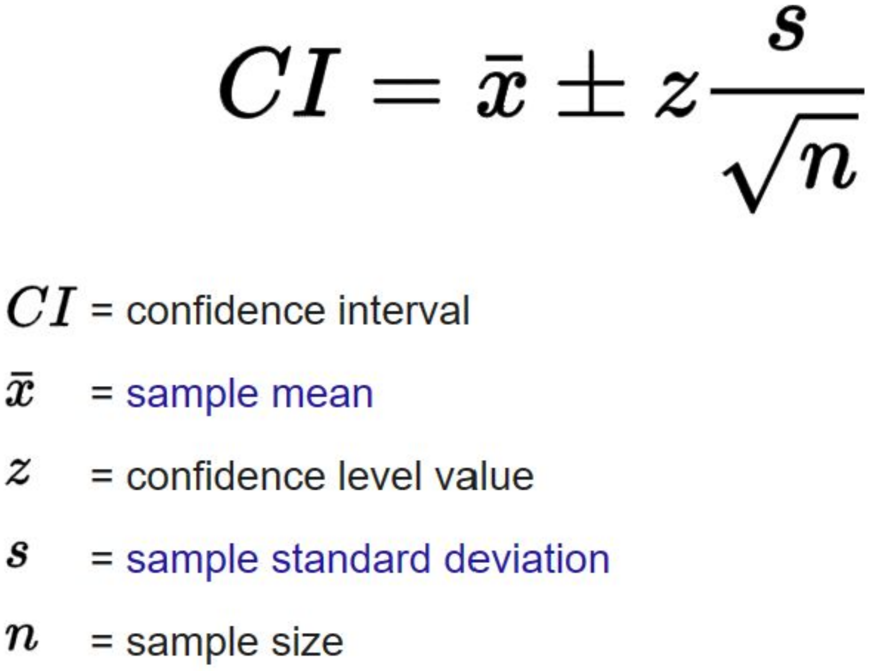


Aspect	Standard Deviation (SD)	Standard Error (SE)
What it measures	The variability of **individual data points** in a sample/population	The variability of a **sample statistic** (e.g., sample mean)
Used for	Describing the spread of a dataset	Describing the accuracy of a sample statistic as an estimate of a population parameter
Formula	Measures the deviation of data points from the mean	Measures the deviation of sample means from the population mean
Effect of sample size	Unaffected by sample size	Decreases as sample size increases

Confidence interval



Confidence interval



$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

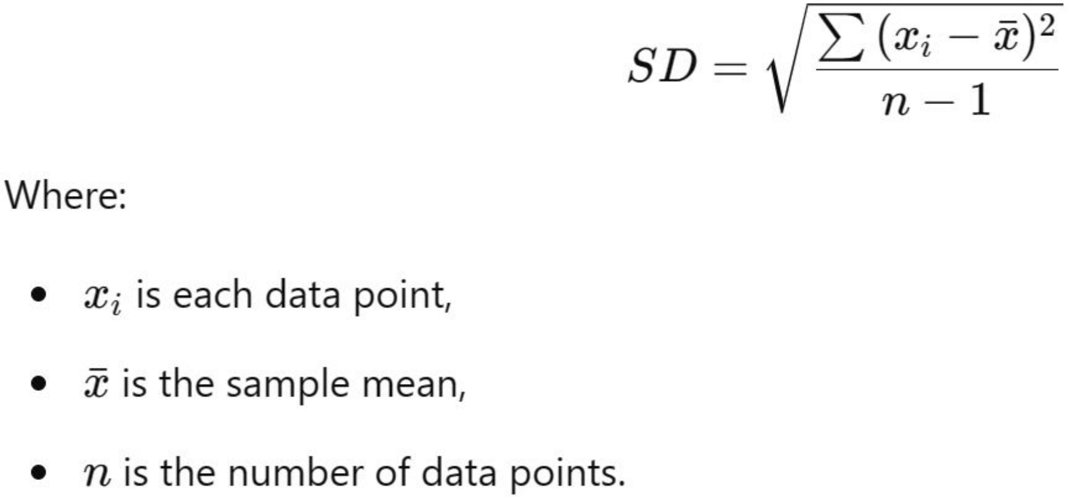
$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

SD vs SE

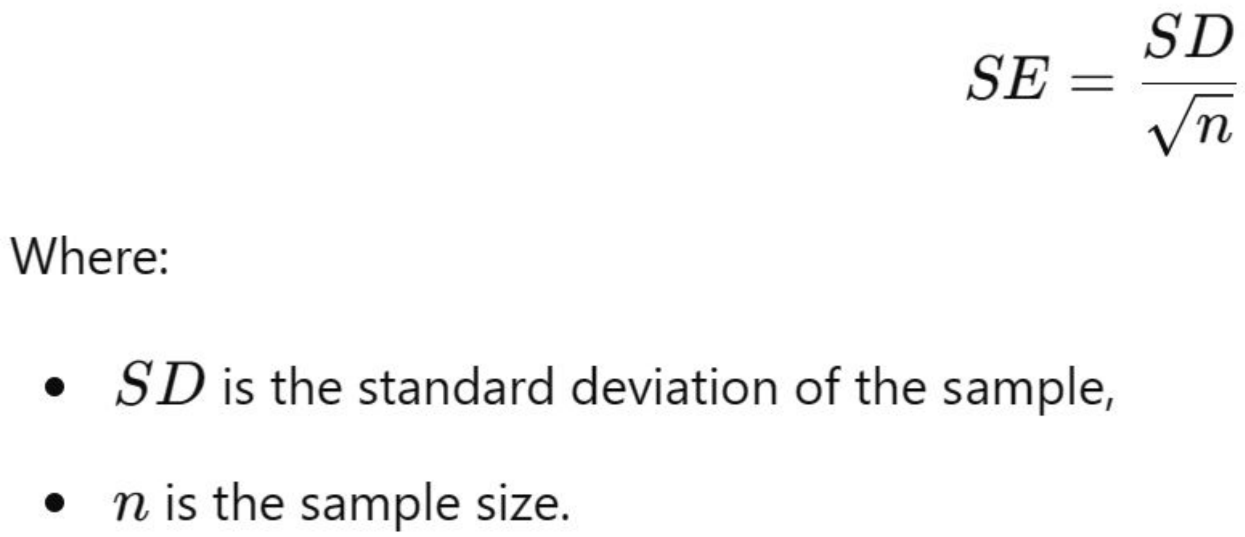


$$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Where:

- x_i is each data point,
- $\bar{x}$ is the sample mean,
- n is the number of data points.

SD vs SE



$$SE = \frac{SD}{\sqrt{n}}$$

Where:

- SD is the standard deviation of the sample,
- n is the sample size.