

Statistics for Data Science

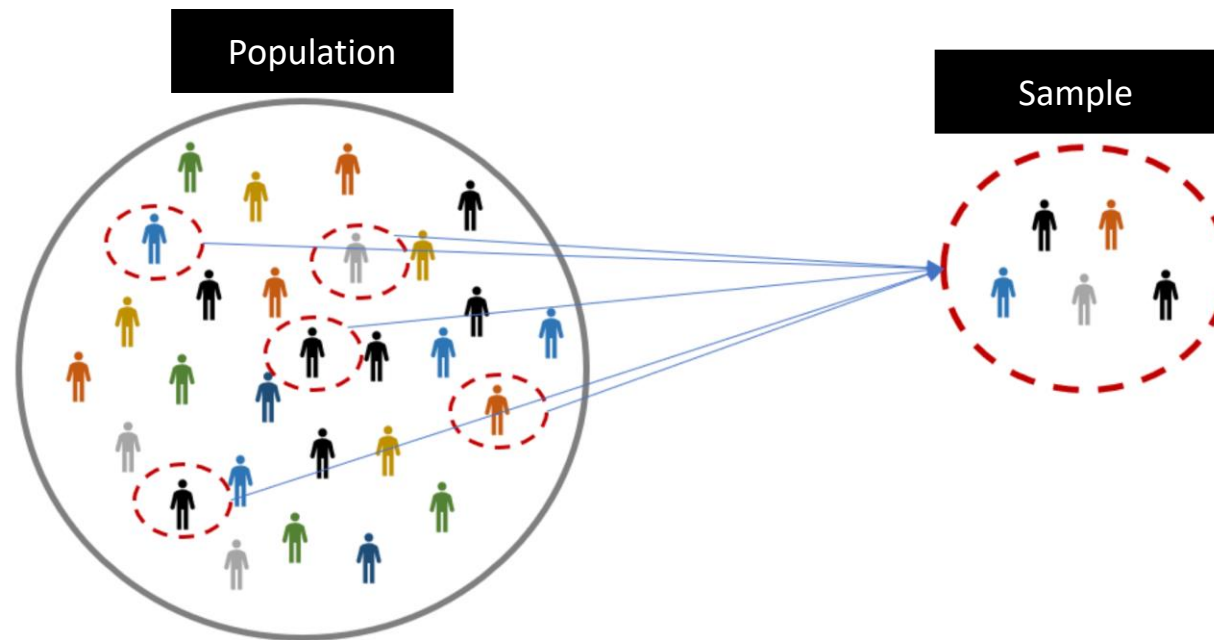
- Population
- Sample
- Standard Deviation
- Histogram
- BoxPlot
- Skewness
- Normal Distribution / Gaussian Distribution
- Log Normal Distribution
- Z Scores
- Outlier Removal using Python



Population & Sample

- A population is an entire group that you want to draw conclusions about.
- A sample is a specific group that you will collect data from. The size of the sample is always less than the total size of the population.





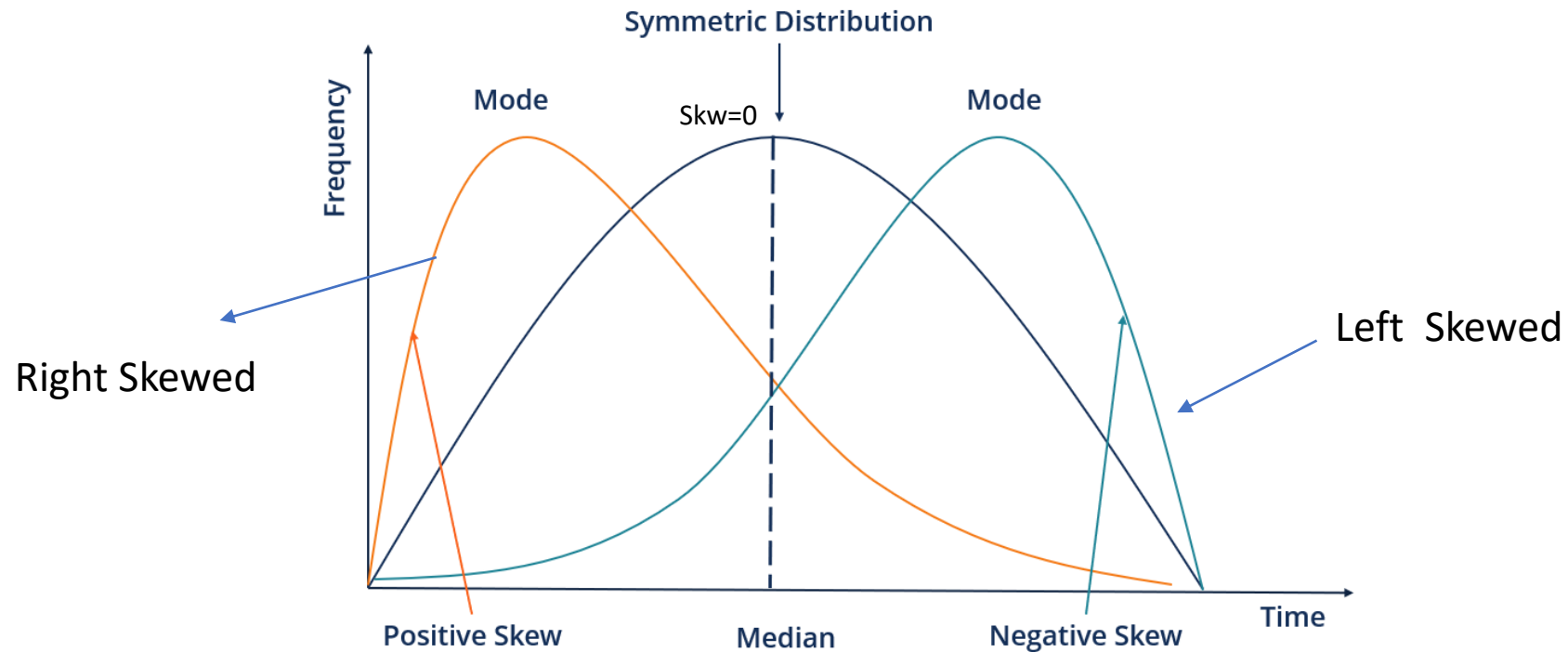
Standard deviation

In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

$$SD = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

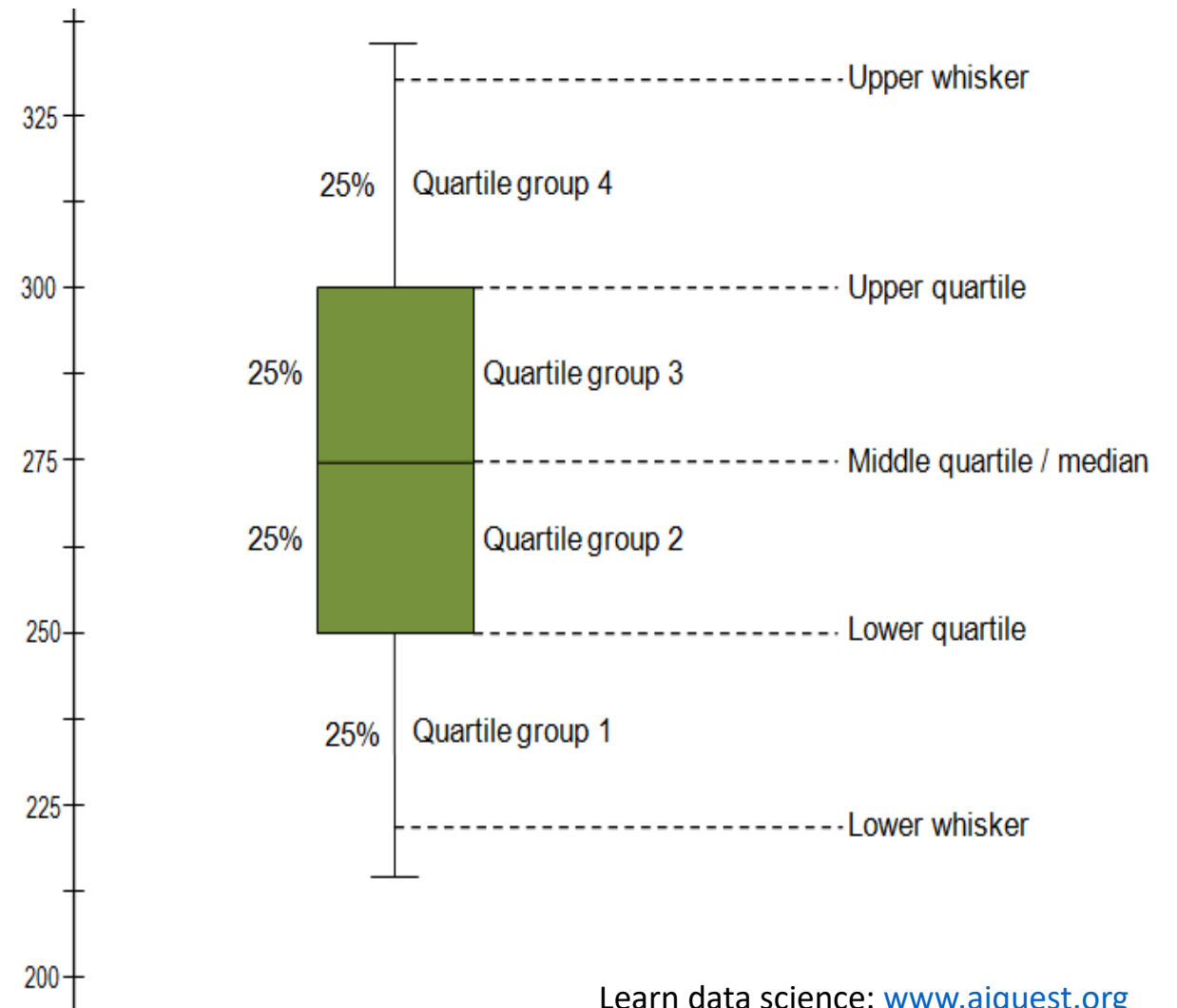
Skewness

Skewness refers to a distortion that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed



Boxplot

Typically, any data point that falls **below $Q1 - 1.5 * IQR$** or **above $Q3 + 1.5 * IQR$** is considered an **outlier**. This is a general guideline, and the threshold of 1.5 can be adjusted based on the specific requirements of your analysis or the characteristics of your dataset.



Quartile & Percentile

The first quartile (Q1) corresponds to the 25th percentile, which means 25% of the data falls below Q1. The second quartile (Q2) is the median and corresponds to the 50th percentile, indicating that 50% of the data falls below Q2. The third quartile (Q3) corresponds to the 75th percentile, where 75% of the data falls below Q3.

- 1st Quartile (Q1) or 25 percentile = $((25/100) * (N+1)) = P_{th}$ index
 - Q2 = Median
 - 3rd Quartile (Q3) or 75 percentile = $((75/100) * (N+1))$
 - IQR = Q3 – Q1
-
- Lower Whisker = $Q1 - (1.5 * IQR)$
 - Upper Whisker = $Q3 + (1.5 * IQR)$

Here,
N = Total Data
P = Position

Construct Boxplot:

Data = [1,2,3,3,5,7,7,8,9,10,20]

- Min = 1
- $Q1 = (25/100) * (11+1) = 3^{\text{rd}} \text{ value} = 3$
- $\text{Median} = (50/100) * (11+1) = 6^{\text{th}} \text{ value} = 7$
- $Q3 = (75/100) * (11+1) = 9^{\text{th}} \text{ value} = 9$
- $IQR = 9 - 3 = 6$
- Max = 10

- Lower Whisker = $Q1 - (1.5 * IQR)$
 - $= 3 - 9$
 - $= -6$

- Upper Whisker = $Q3 + (1.5 * IQR)$
 - $= 9 + 9$
 - $= 18$

Boxplot



To create a boxplot, you can follow these steps:

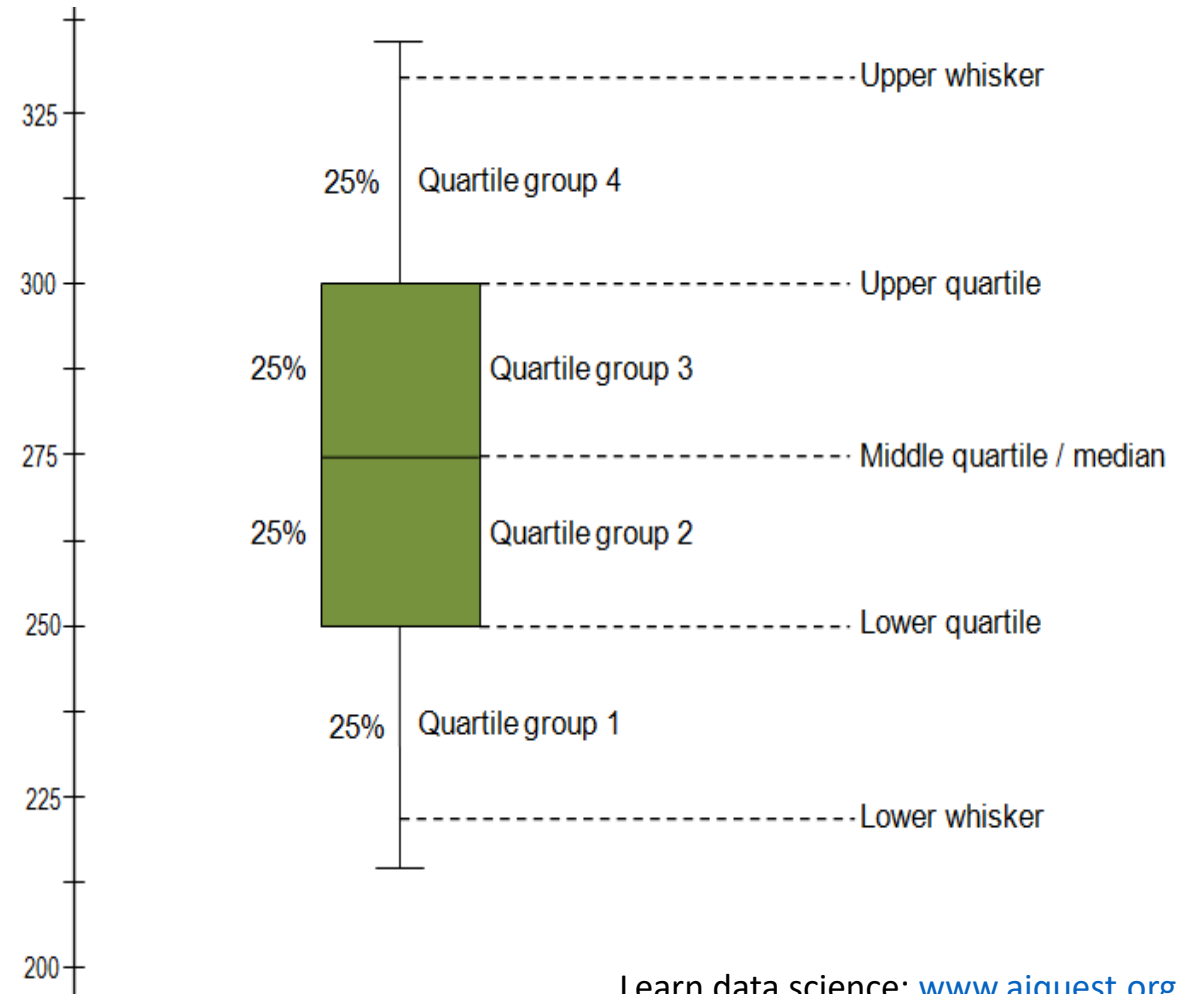
1. Gather your dataset: Collect the numerical data that you want to visualize using a box plot.
2. Determine the key components: Identify the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values of your dataset.
3. Calculate the interquartile range (IQR): Subtract Q1 from Q3 to obtain the IQR.
4. Identify any outliers: Determine if there are any values that fall below $Q1 - 1.5 * IQR$ or above $Q3 + 1.5 * IQR$. These values are considered outliers and may be plotted individually as points.
5. Set up the box plot: Draw a number line or axis to represent the range of your dataset. Place a box that spans from Q1 to Q3 on the number line. Draw a line within the box to represent the median (Q2).
6. Add whiskers: Extend lines, known as "whiskers," from the box to the minimum and maximum values that are not considered outliers.
7. Plot outliers: If there are any outliers, plot them individually as points outside the whiskers.
8. Label and title: Add appropriate labels to the number line, box, whiskers, and outliers. Provide a title that describes the dataset or the purpose of the box plot.

Boxplot with matplotlib

```
import matplotlib.pyplot as plt
# Sample dataset
data = [1,2,3,3,5,7,7,8,9,10,20]
# Create a figure and axis
fig, ax = plt.subplots()

# Create a box plot
ax.boxplot(data)

# Add labels and title
ax.set_xlabel('Data')
ax.set_ylabel('Values')
ax.set_title('Box Plot Example')
# Display the plot
plt.show()
```



Boxplot with Seaborn

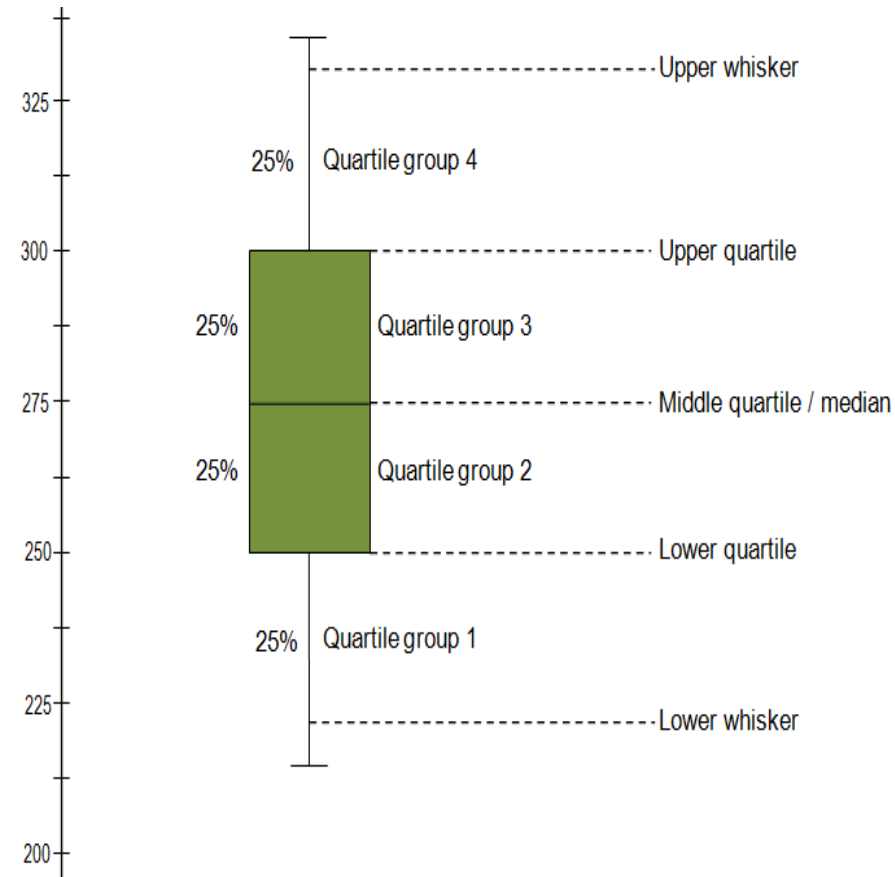
```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# Sample dataset
data = [1,2,3,3,5,7,7,8,9,10,20]
```

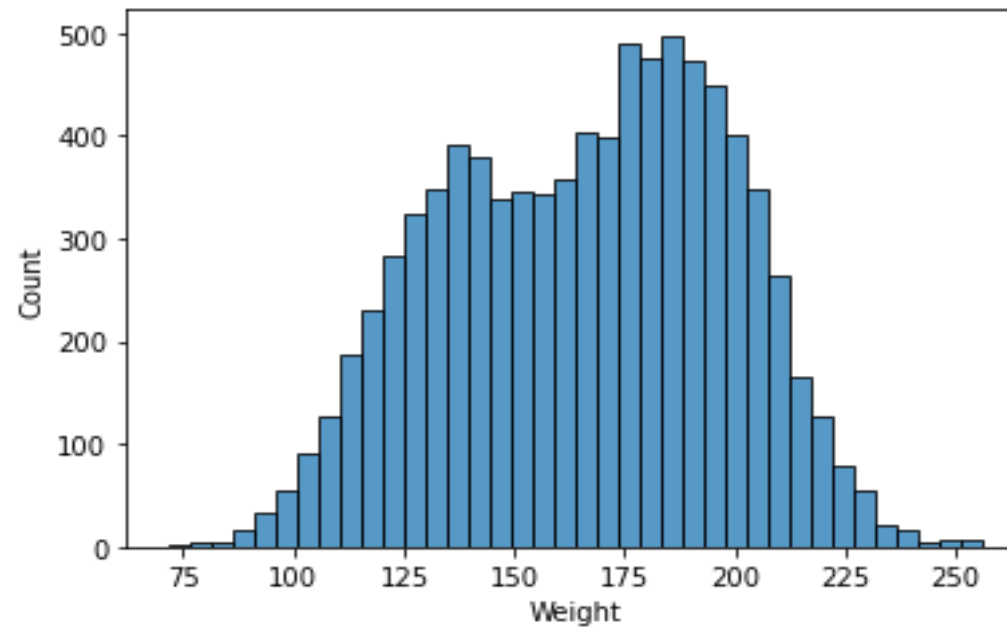
```
# Create a box plot
sns.boxplot(data=data)
```

```
# Add labels and title
plt.xlabel('Data')
plt.ylabel('Values')
plt.title('Box Plot Example')
```

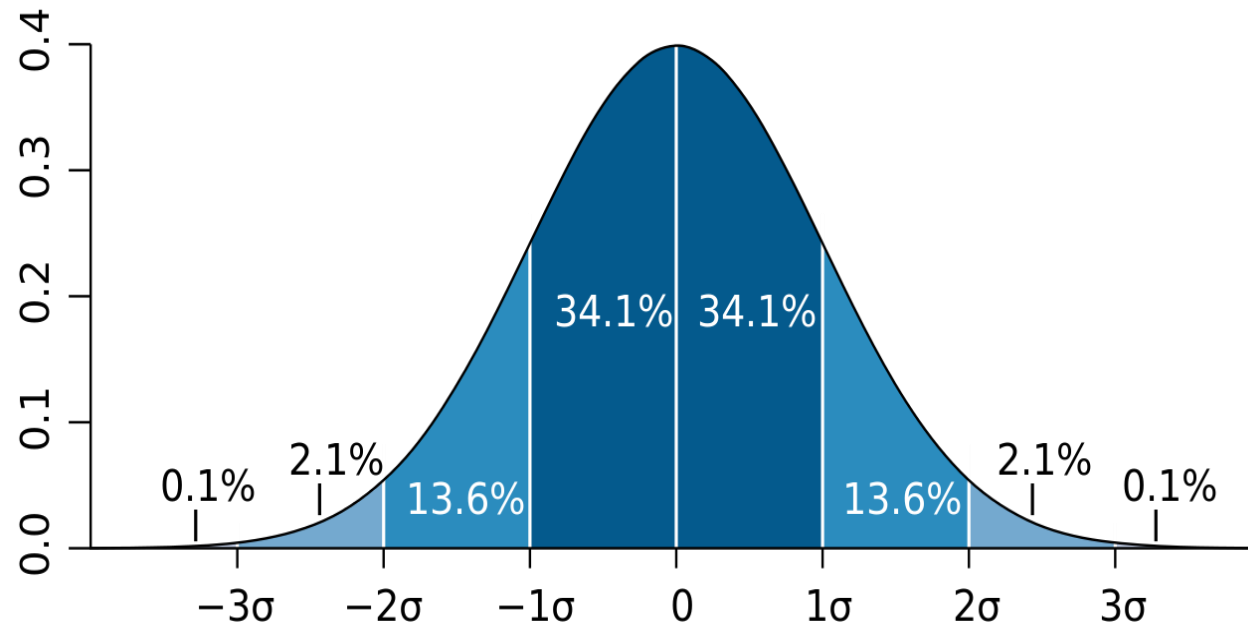
```
# Display the plot
plt.show()
```



Histogram

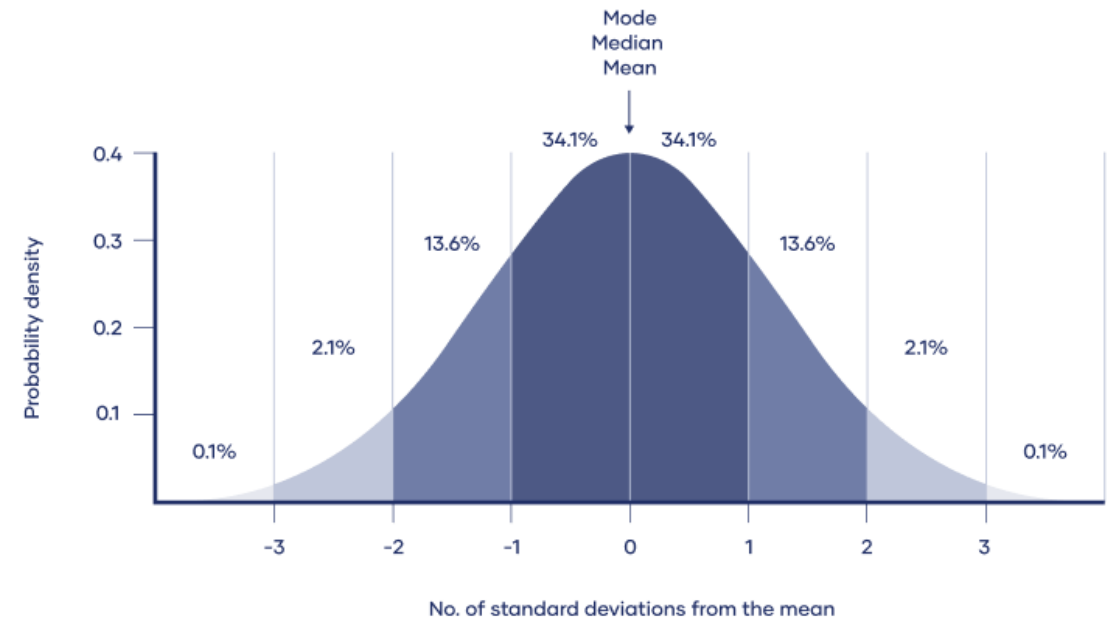


Standard deviation

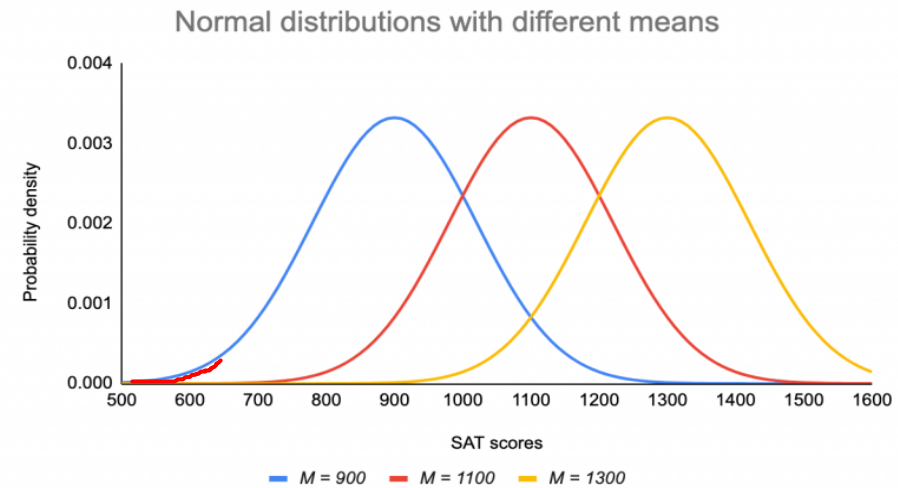


What are the properties of normal distributions?

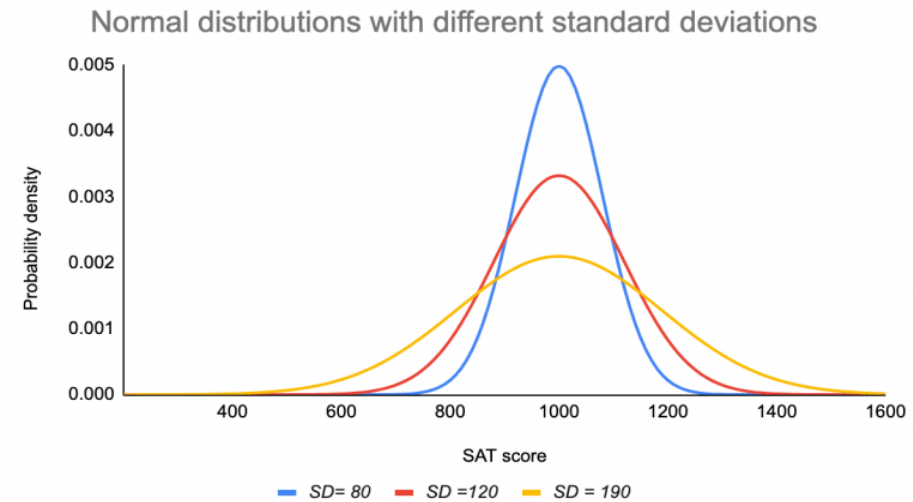
- The mean, median, and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.



The mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.

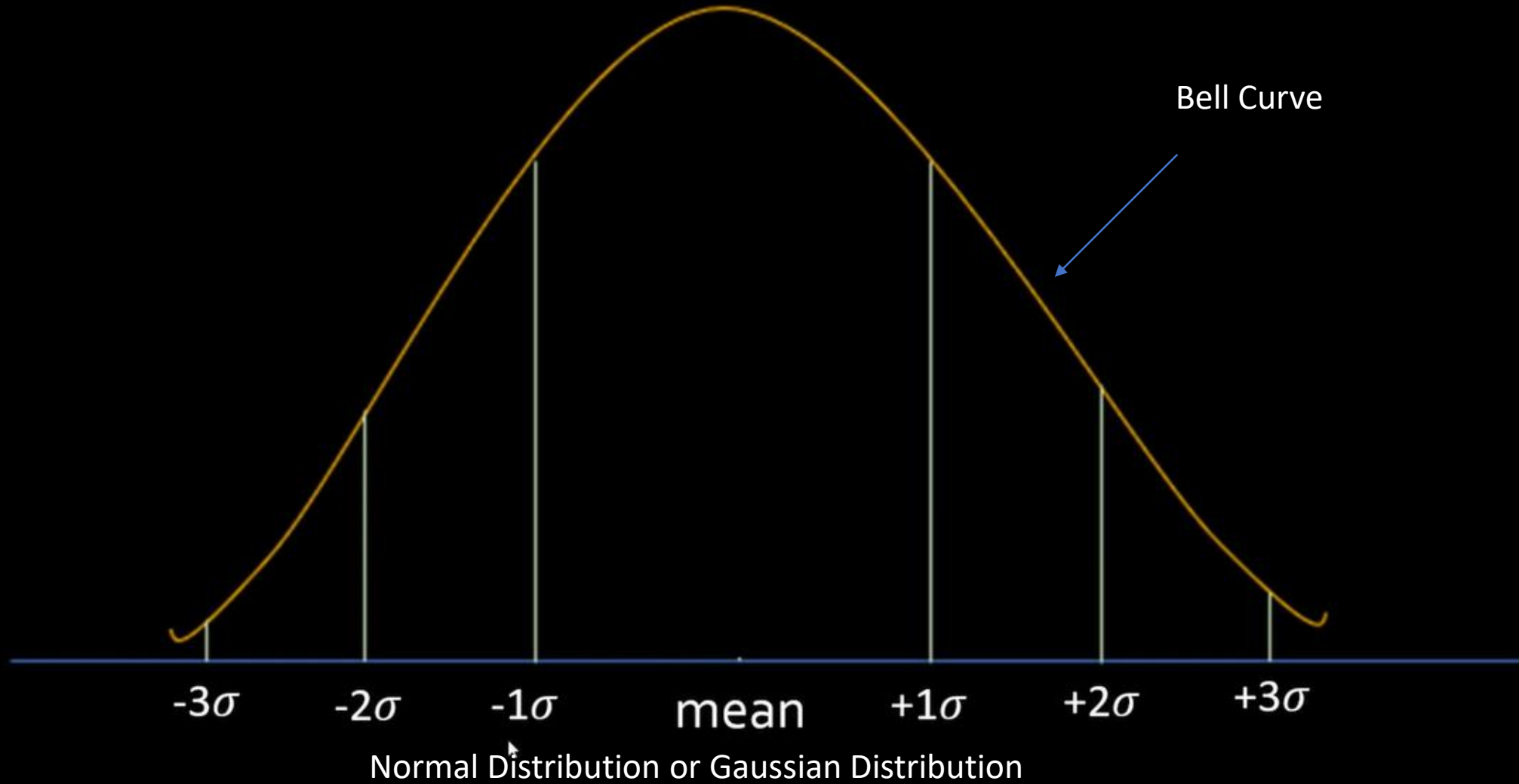


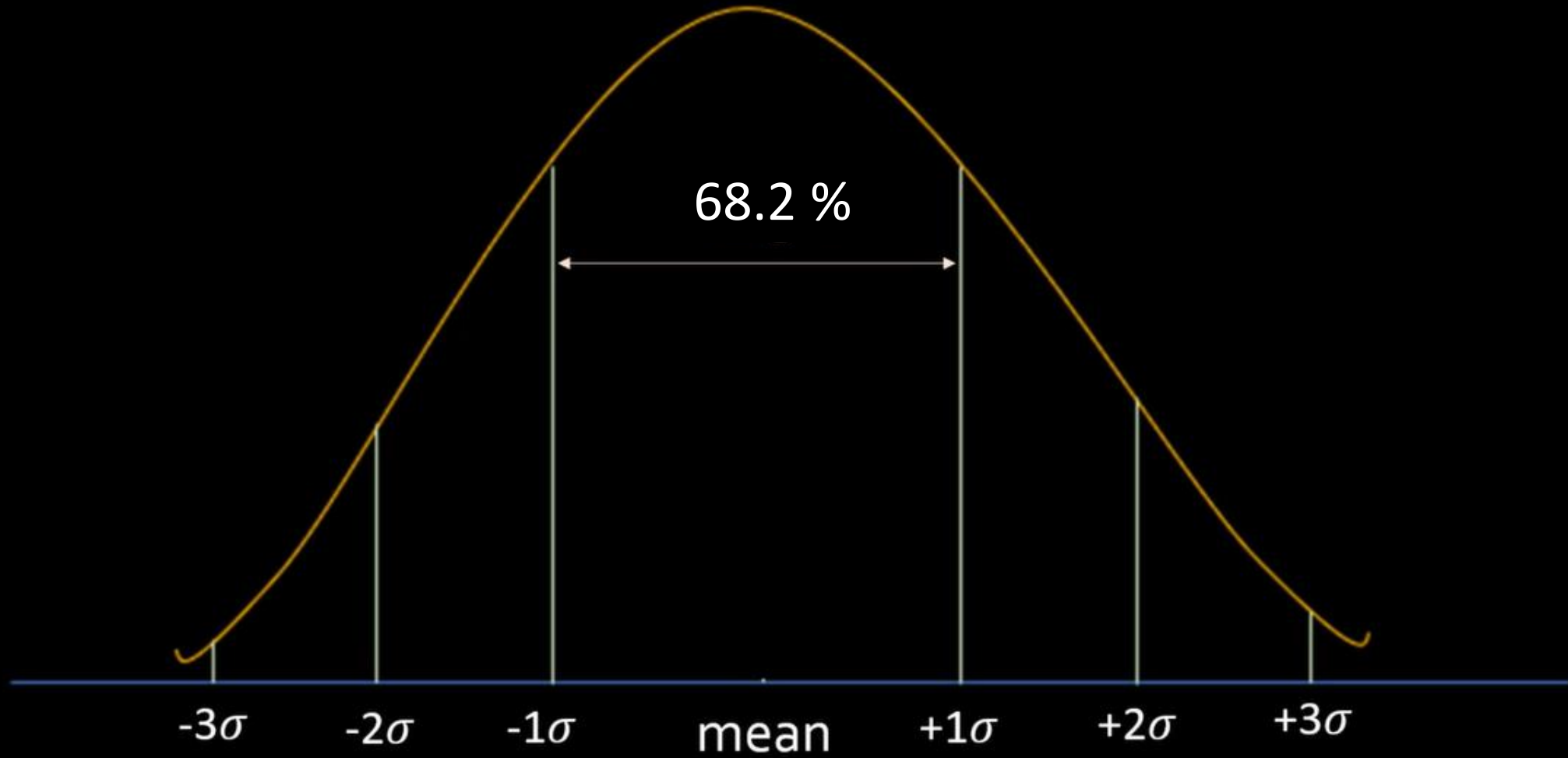
-
- The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.



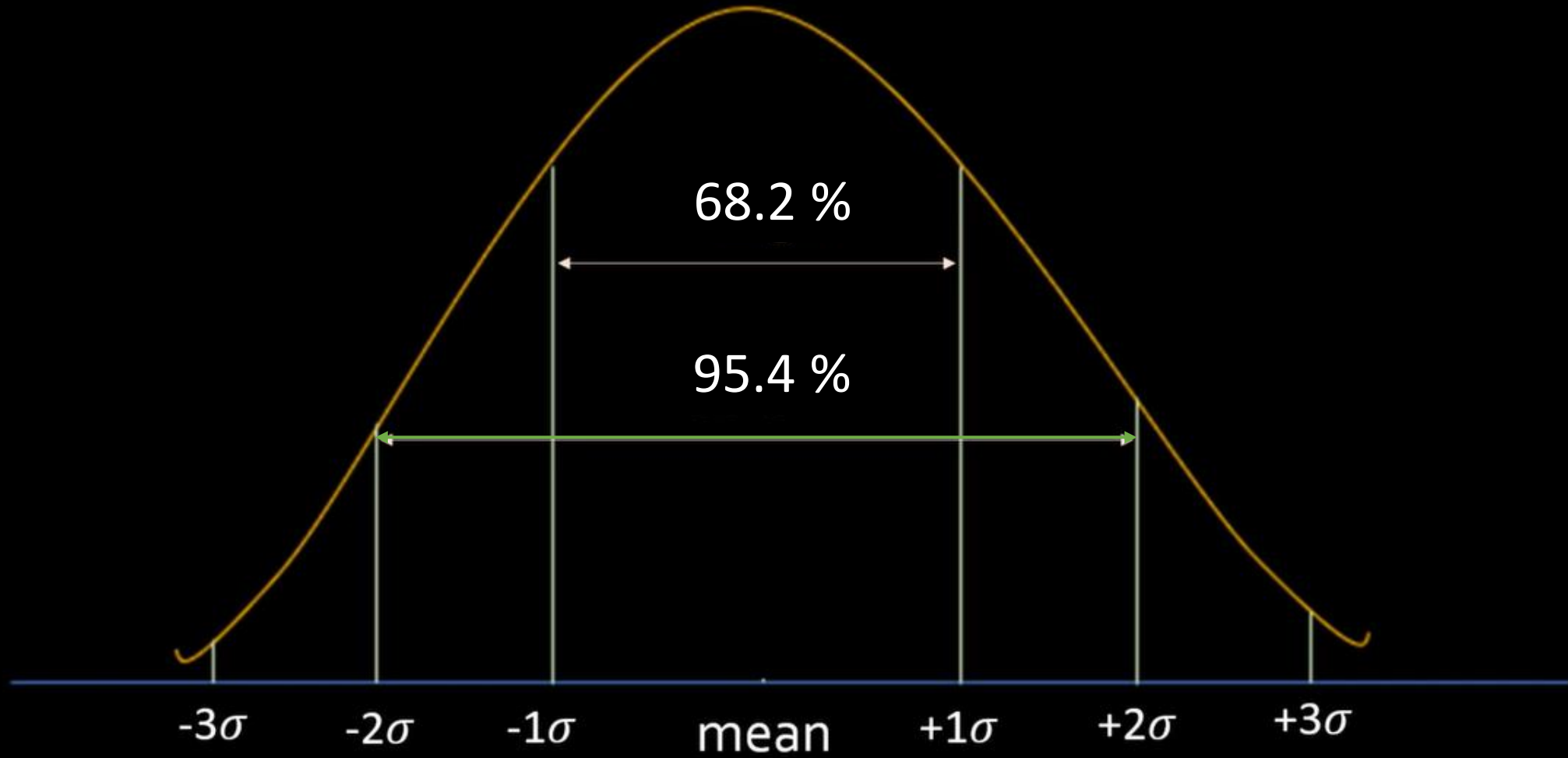
Normal Distribution

- Income Distribution In Economy
- Shoe Size
- Birth Weight
- Spending Days in Hospital

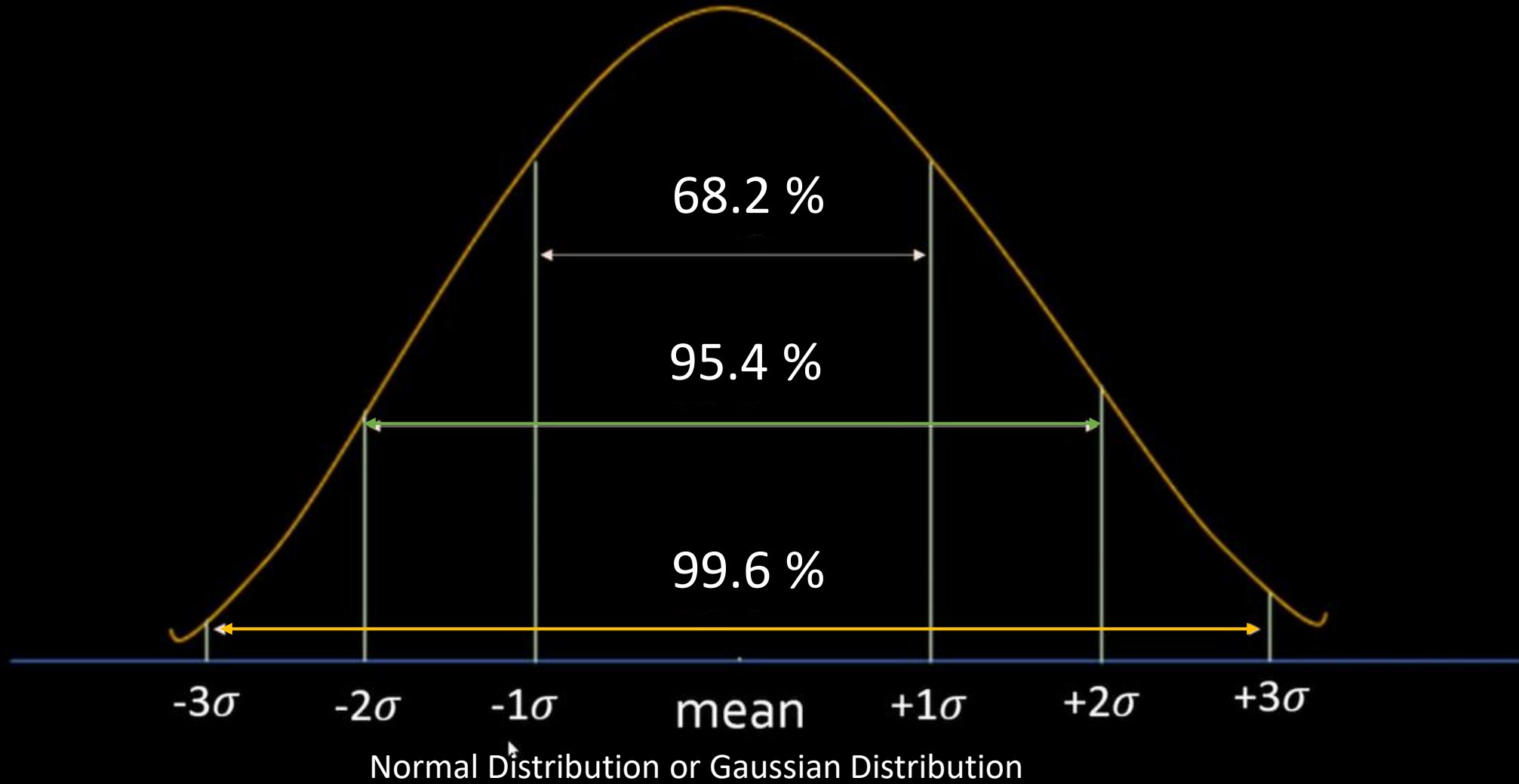




Normal Distribution or Gaussian Distribution



Normal Distribution or Gaussian Distribution



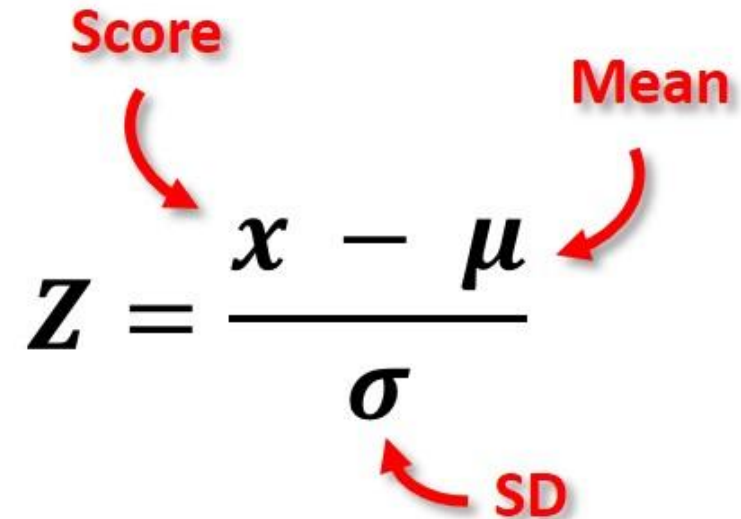
Z Scores

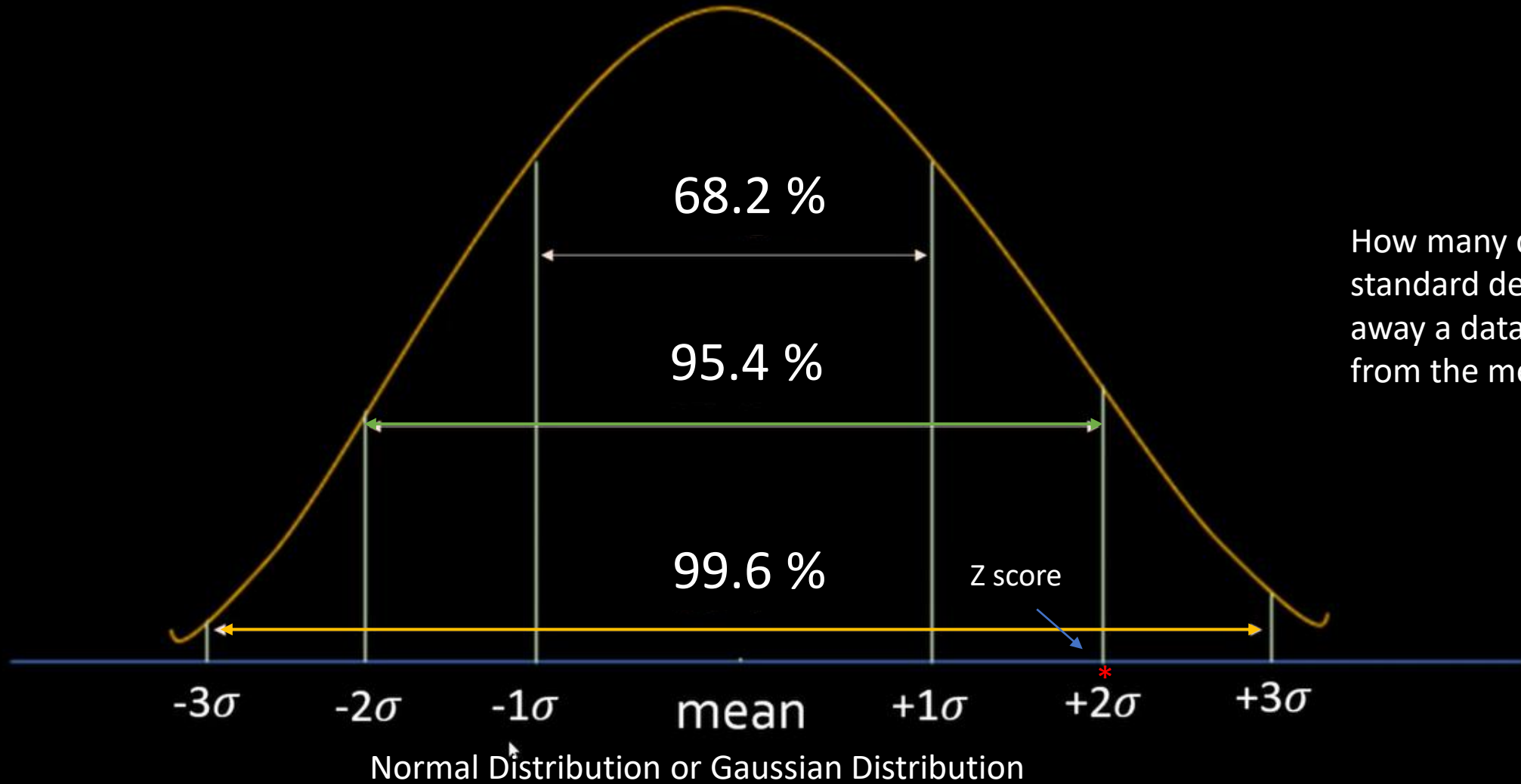
How many distances standard deviation away a data point is, from the mean value?

How to Calculate a Z-Score,

$$Z = \frac{x - \mu}{\sigma}$$

Score Mean
SD

The diagram shows the Z-score formula with red arrows pointing to each part: 'Score' points to 'x', 'Mean' points to 'μ', and 'SD' points to 'σ'.



How many distances standard deviation away a datapoint is, from the mean value.