Statistics Cheat Sheet- 1

POPULATION AND SAMPLE

Population: The entire group that you want to draw conclusions about.

- Example: All students in a university.

Sample: A subset of the population selected for analysis.

- Example: 100 students selected from the university.

- Why it Matters: Samples are used to make inferences about populations, as it’s often impractical to collect data from an entire population. Samples are easier to manage and analyze (Time and cost constraints).

Sampling Error: Arises because the sample is not a perfect

Converting Qualitative Data to Quantitative Data

- **Replace Method**: Manually replace categories with numbers.

 - Example: Gender: Male = 1, Female = 2.

```
df[''].replace([],[])
```

- **Dummy Method**: Convert categories into binary dummy variables (0 and 1).

 - Example:* Color: Red = 1, Not Red = 0.

```
 pd.get_dummies(df[])
```

- **Label Encoder**: Assign numerical values to categories automatically.

```
from sklearn.preprocessing
import LabelEncoder
le = LabelEncoder()
le.fit_transform(df[])
```

In alphabetical order it give labels of 0,1,2..

TYPES OF DATA

Qualitative (Categorical) Data: Data that describes categories or groups.

✓ **Ordinal Data**: Data with a clear ordering (comparable).
 - Example: Ratings (1 to 5 stars).

✓ - **Nominal Data**: Data without any order or noncomparable.
 - Example: Types of fruit (apple, banana, cherry).

- **Quantitative (Numerical) Data:** Data that represents quantities and can be measured.

✓ - **Continuous Data:** Numerical data that can take any value within a range.
 - Example: Temperature in degrees Celsius.

✓ - **Discrete Data:** Numerical data that can only take specific values.
 - Example: Number of children in a family.

Machine Learning models cannot work on categorical variables as strings, so we need to change them into numerical form.

Outlier:

An outlier is a data point that differs significantly from other observations.

It may indicate variability in measurement or experimental errors.

Measures of Central Tendency

- **Mean** (Average): Sum of all data points divided by the number of points.

```
import statistics as stats
stats.mean()
```

- **Median**: The middle value when data is sorted in order.

If the number of data values is even, it returns the average of the two middle values.

```
import statistics as stats
stats.median()
```

- **Mode**: The most frequently occurring value in a dataset.

 - Useful for categorical data.

```
import statistics as stats
stats.mode(data)
```

A distribution with more than one mode is called multimodal.

Measures of Dispersion

Measures of dispersion describe how spread out or clustered data points are in a dataset. They help us understand the variability or consistency within the data.

Variance:

Measures how much each data point differs from the mean.

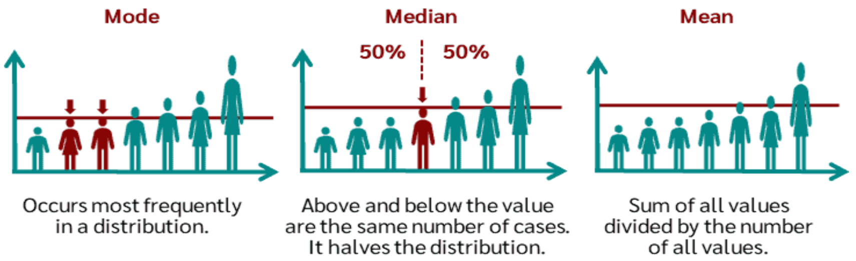
Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

```
import statistics as stats
stats.variance() #sam var
stats.pvariance() #pop var
```

Standard Deviation:

The square root of the variance, providing dispersion in the same units as the data.

```
import statistics as stats
stats.stdev() #sample stdev
stats.stdev()  #pop stdev
```



Occurs most frequently in a distribution.

Above and below the value are the same number of cases. It halves the distribution.

Sum of all values divided by the number of all values.

Errors

Error measures the difference between predicted values and actual values in a dataset.

Mean Absolute Error (MAE)

MAE is the average of all absolute errors.

Absolute error preserves the same units of measurement as the data.

```
from sklearn.metrics
import mean_absolute_error
mean_absolute_error(actual
, predicted)    #mae
```

Mean Square Error (MSE)

MSE, measures the average of the squares of the errors.

Due to the square, large errors are emphasized and have a relatively greater effect.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

```
from sklearn.metrics
import mean_squared_error
mean_squared_error(actual,
predicted)      #mse
```

Normal Distribution

- A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme.

- Key Properties:

- **A bell curve** is the graph that represents a normal probability distribution.
- Mean, median, and mode are equal and located at the centre of the distribution.
- The bell curve is perfectly symmetrical.
- It is concentrated around the peak and decreases on either side.
- The standard deviation defines the width of the graph.
- The area under the whole curve is equal to 1, or 100%.

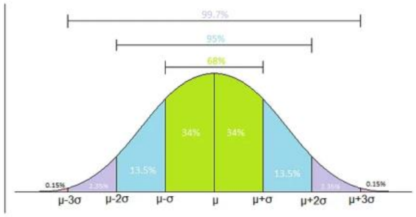
Empirical rule

An empirical rule in statistics states almost all of the observations in a normal distribution lie within three standard deviations from the mean.

- 68% of data falls within 1 standard deviation from the mean.

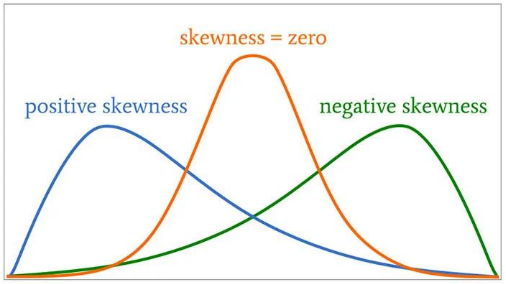
- 95% within 2 standard deviations from the mean.

- 99.7% within 3 standard deviations from the mean.



Skewness:

A measure of the asymmetry of the distribution of values.



- Positive Skew (Right Skew):

Tail on the right side; mean > median > mode.

- Negative Skew (Left Skew):

Tail on the left side; mean < median < mode.